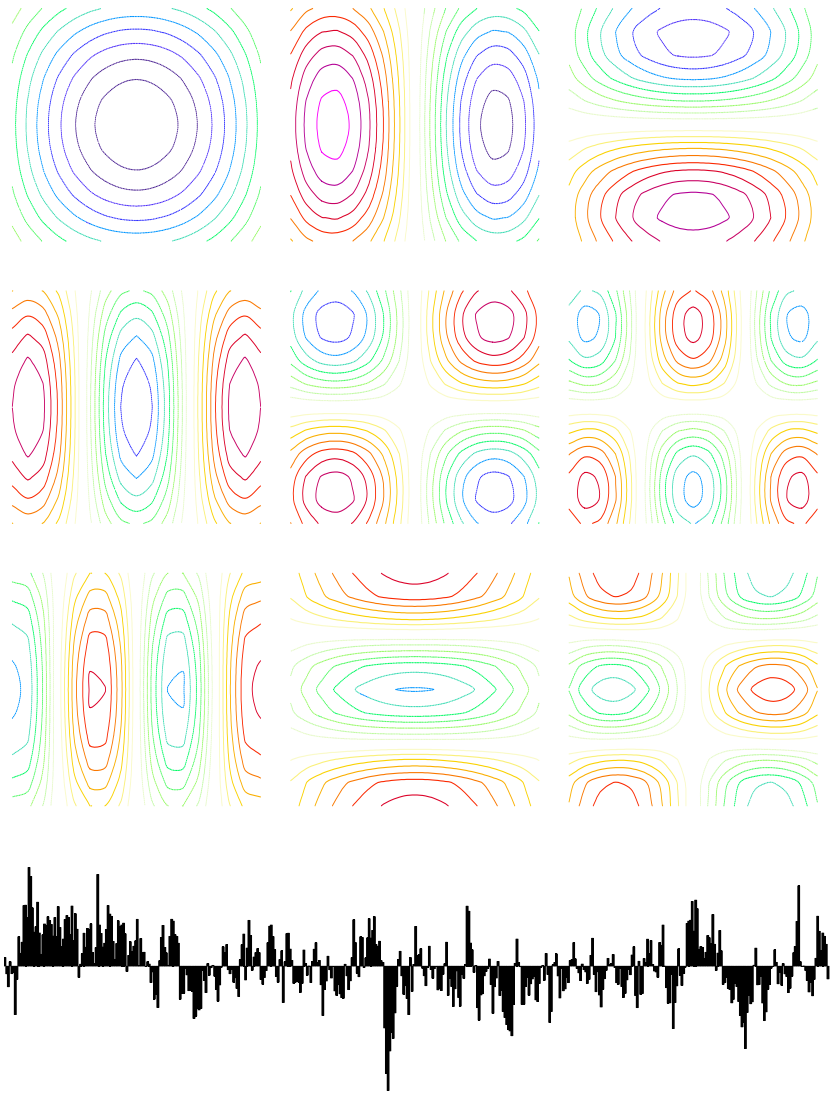
An Introduction to Statistical Analysis in Climate Research

Lecture notes (textbook) for the statistic lecture
revised, but unfinished.

Version May 27, 2015

Dietmar Dommenget[1]

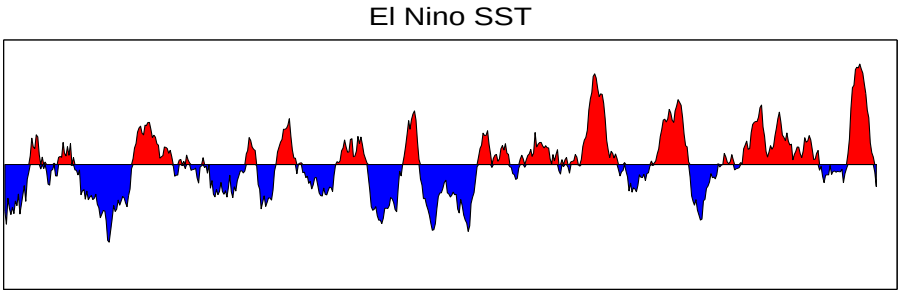


[1]Corresponding Author:
Dietmar Dommenget, School of Mathematical Sciences, Building 28, Room 212, Clayton Campus, Monash University, VIC 3800, Australia;
email: dietmar.dommenget@monash.edu

Contents

Chapter 1

Introduction



These notes are an introduction to statistical analysis in climate dynamics. Statistical analysis is essential in the discovery of new findings if based on observation data or experiments. However, in school and university we start with learning about science or climate dynamics based on equations, relationships and dynamical laws of physics, but statistical methods are not needed for that. Indeed by learning about the laws of physics or climate dynamics as students in schools/university we have very little or no contact with statistical analysis or inferences. We learn the laws of physics by logical reasoning and by presentations of the observed values.

This however changes dramatically when you move from being a student that learns the textbooks to becoming a researcher figuring out new science. In research we most often have to deal with observational or experimental (simulation) data. First these data will present themselves with apparently very little structure in it. They appear to be chaotic and seem to make no sense. Or sometimes this data will have structure in it but no physical explanation can be given for it. Good examples are El Nino and the Southern Oscillation, the Madden Julian Oscillation, the North Atlantic Oscillation or some Hurricane statistics. These observations first of all them to make no sense, but with the help of statistical analysis we gain more understanding and with our physical understanding we can eventually make sense of the data. So in research statistical description of the observation often comes long before the physical understanding comes. The statistical analysis if therefore often our starting point for the physical understanding of the system.

The statistical methods introduced in these notes are the basis for many different aspects of statistical analysis. We can in principal put these into two categories: Measurement uncertainties and stochastic variability. In measurements we need to deal with statistical analysis, because ur measurement have errors. These may be instrumental errors, they can be errors resulting from transformation of indirect measurements to the variables of interest (e.g. estimating heat transport in the ocean on the basis of temperature and salinity profiles). Errors aslo result from interpolations or proxy data or assimilation of observations into models. The statistical methods introduced in these notes are a good basis to deal with all these measurement uncertainties.

However, these notes will not focus on measurement uncertainties, but will focus on stochastic variability. The climate system is a chaotic system. The non-linear dynamics of the atmosphere cause the system to vary on all time and spatial scales (e.g. The lorenz model). Describing this stochastic variability with statistical methods is the focus of these notes.
The lectures notes are organised in five chapters. In the first chapter we will introduce some basic probability theory, with introducing probability, probability density functions and some important concepts such as independence, conditional probability and covariance/correlation.

- statistic is made by mathematicians. they have a different point of view than physicist.
- only a small fraction of the statistical method
- i will keep definitions/developments short and try to put the statistics onto examples and relate them to physics
- it is next to impossible to understand statistical method without applying them
- it is important to view statistics for m different perspectives, textbooks, authors different applications

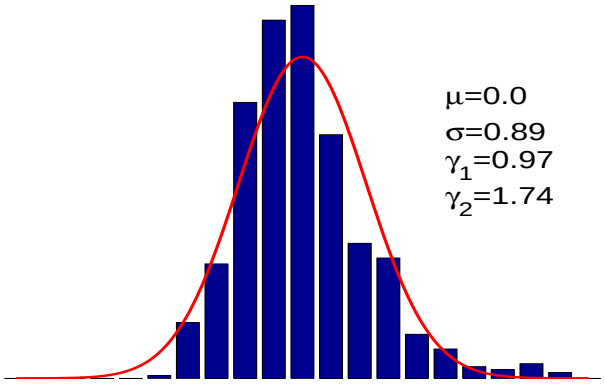
1.1 Literature

Below is a short list of the literature that helped me to develop this lecture. Much of the chapters I, II and IV have been developed on the basis of the textbook by Von Storch and Zwiers. Many problems in climate research are somehow unique to the climate community and are only discussed in articles of scientific journals, some helpful discussions of statistical methods can be found on websites.
The following list is in order of relevance:

- "Statistical Analysis in Climate Research" a lecture script by Dietmar Dommenget
- "Statistical Analysis in Climate Research" by Hans von Storch and Francis W. Zwiers.
- "Statistische Methoden der Datenanalyse" a lecture script by Juergen Willebrand
- "Taschenbuch der Mathematik "by Bronstein et al., 5. Edition.
- "Principal Component Analysis" by Ian T. Jolliffe
- "Zeitreihenanalyse" by Schlittgen and Streitberg
- "Time Series Analysis" by George E.P. Box, Gwilym M. Jenkins and Gregory C. Reinsel

Part I

Fundamentals



Here we will have a short discussion of probability theory and we will define some of the most basic parameters which are important for statistical analysis in general and are most important for of the subsequent sections.
Statistical analysis is strongly related to probability theory, especially the statistical inference and tests of hypothesis needs a strong background in probability theory and logics. Miss/understandings of logics/probability theory is also very important for the interpretation of statistics, which we will discuss in the final section.

Chapter 2

Probability

The Probability Theory deals with uncertain events of stochastic processes, in contrast to events that are deterministic (with well-defined initial values) and therefore certain in the outcome.

Examples of uncertain/stochastic events:
Coin tossing,
Lottery numbers,
Rainfall at certain time/place,
Any kind of physical quantity of a stochastic or thermodynamical process.

Examples of certain/deterministic events:
Length of a body (table, room or ship)
time of the sun rise,
wave propagation,
Any kind of physical quantity of a deterministic process (with well-defined initial values).

2.1 Probability of an Event

The space of all possible values that an uncertain event can take is the sample space ,$\mathcal{S}$.

Examples:

- Tossing dices; $\mathcal{S} = [1, 2, 3, 4, 5, 6]$ with an event $\mathcal{A} = 1$ or the event $\mathcal{B} = \textit{odd number} = [1, 3, 5]$.
- Temperature: $S = R^+$,
- Wind direction: $S \in [0, 2\pi]$

Often the rules of probability a more easily understood, if we think of the theory of sets and how the number of elements relate to different subsets.
Some basic rules for the probability:

- Probabilities are always $\in [0, 1]$.
- When an experiment is conducted, one of all possible events in $\mathcal{S}$ must occur, so

$$P(\mathcal{S}) = 1 \tag{2.1}$$

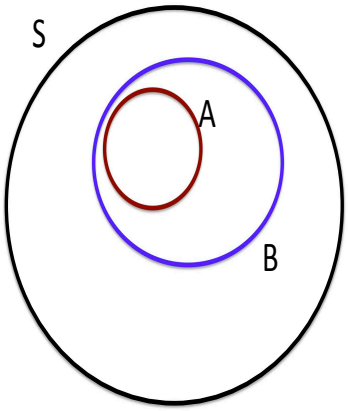


Figure 2.1: Sketch to illustrate the sample space and two events. The circled areas should represent the set of values belonging to the sample space (S), event A and event B.

- Sometimes it is easier to compute the probability of the complement of an event than that of an event itself. If $\mathcal{A}$ denotes an event, than $\neg\mathcal{A}$ denotes its complement, the collection of all events in $\mathcal{S}$ that are not contained in $\mathcal{A}$. Thus, $\mathcal{S} = \mathcal{A} \cup \neg\mathcal{A}$, and $\mathcal{A} \cap \neg\mathcal{A} = \emptyset$. Therefore,

$$P(\mathcal{A}) = 1 - P(\neg\mathcal{A}) \tag{2.2}$$

  Example: The event $\mathcal{A}$ of tossing the number six at least once with three dices. The probability of complement event, not tossing the number six with three dices, is $\neg\mathcal{A} = (\frac{5}{6})^3 \Rightarrow \mathcal{A} = 1 - (\frac{5}{6})^3$.

- It is often useful to divide an event into two mutually exclusive events. Two events $\mathcal{A}$ and $\mathcal{B}$ are mutually exclusive if they do not contain any common sample space elements, that is $\mathcal{A} \cap \mathcal{B} = \emptyset$. An experiment can not produce two mutually exclusive event at the same time. Hence,

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) \tag{2.3}$$

- In general, the expression for the probability of observing one of two events $\mathcal{A}$ and $\mathcal{B}$ is

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) \tag{2.4}$$

  or

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cup \mathcal{B}) \tag{2.5}$$

  The common part of the two events, $\mathcal{A} \cap \mathcal{B}$, is included in both $\mathcal{A}$ and $\mathcal{B}$ and thus $P(\mathcal{A} \cap \mathcal{B})$ is included in the calculation of $P(\mathcal{A}) + P(\mathcal{B})$ twice.

  Example with tossing a dice: Event $\mathcal{A}$ is tossing an even number and $\mathcal{B}$ is tossing a number < 4. So, $P(\mathcal{A}) = \frac{1}{2}$, $P(\mathcal{B}) = \frac{1}{2}$ and $P(\mathcal{A} \cup \mathcal{B}) = \frac{5}{6}$, it follows that $P(\mathcal{A} \cap \mathcal{B}) = \frac{1}{2} + \frac{1}{2} - \frac{5}{6} = \frac{1}{6}$

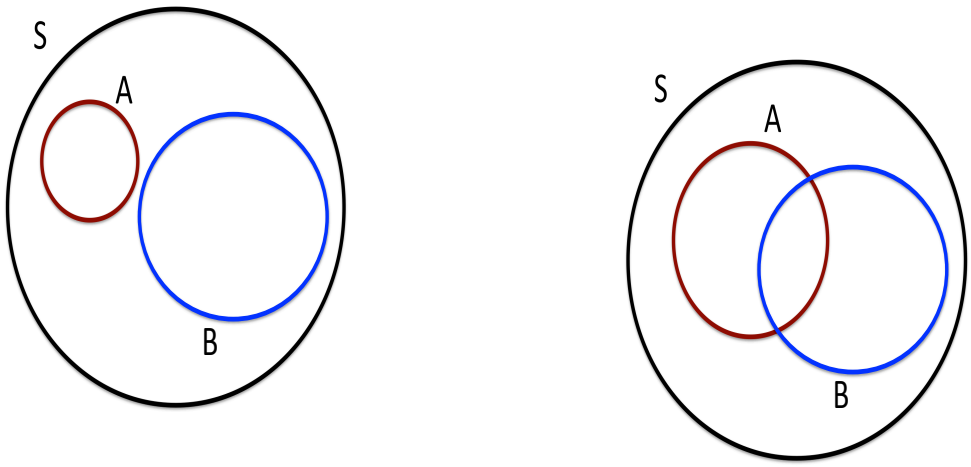


Figure 2.2: Sketchs to illustrate the sample space and two events. Left: mutually exclusive events. Right: two events, general.

2.2 Conditional Probability

Often we are dealing with the evaluation of probability of some result/event, but we do not realize, that we have some additional 'hidden' condition for this result/event. In many cases these additional condition will change probability outcome.
Example **Storms over seasons**: Consider the probability of a weather event $\mathcal{A}$, such as a storm (strong winds), and suppose that the climatology probability of such an event is $P(\mathcal{A})$. Now consider that beginning in one season of the year is an event $\mathcal{B}$. If $P(\mathcal{A})$ is different at different seasons, the perception of likelihood of $\mathcal{A}$ will change. That is, the probability of $\mathcal{A}$ conditional upon the season $\mathcal{B}$, which is written $P(\mathcal{A}|\mathcal{B})$, will not be the same as the climatology probability $P(\mathcal{A})$.
Example **ship measurements**: Consider the climatology of temperature of the North Atlantic. The climatological mean can be regarded (approx.) as the most likely value. If this climatology of temperature of the North Atlantic is based on ship measurements the Probability of different temperature events maybe skewed towards 'good' weather, simply because ship do not measure in bad weather and ship change routes due to bad weather, so that the region with bad weather is not measured. Thus a climatological temperature of the North Atlantic based on ship measurements is a Conditional Probability estimate. It is not the real Probability distribution.
The conditional probability of event $\mathcal{A}$, given an event $\mathcal{B}$ for which $P(\mathcal{B}) \neq 0$, is

$$P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B})/P(\mathcal{B}) \tag{2.6}$$

The interpretation is that only the part of $\mathcal{A}$ that is contained within $\mathcal{B}$ can take place and thus the probability that this restricted version of $\mathcal{A}$ takes place must be scaled by $P(\mathcal{B})$ to account for the change of context. The sample space $\mathcal{S}$, of all possible events, is replaced by $\mathcal{B}$. Note that all conditional probabilities are $\in [0, 1]$.
Note: That both $P(\mathcal{A}|\mathcal{B}) < P(\mathcal{A})$ and $P(\mathcal{A}|\mathcal{B}) > P(\mathcal{A})$ are possible.
Note: $P(\mathcal{A}|\mathcal{B}) \neq P(\mathcal{A} \cap \mathcal{B})$. The probability A and B is different from $P(\mathcal{A}|\mathcal{B})$, because in $P(\mathcal{A}|\mathcal{B})$ we assume that B has happend: $\mathcal{B} = \mathcal{S} \Rightarrow P(\mathcal{B}) = 1$

Note: $P(\mathcal{A}|\mathcal{B}) > P(\mathcal{A}) \Rightarrow P(\mathcal{A}|\neg\mathcal{B}) < P(\mathcal{A})$. If, for instance, the temperature, T at time interval of a time series is large than average, than there most be a part of the time series where T is smaller than average. It seems trivial to point this out, but often you find studies in which the statement $\Rightarrow P(\mathcal{A}|\neg\mathcal{B}) < P(\mathcal{A})$ is made as if unexpect, when $P(\mathcal{A}|\mathcal{B}) > P(\mathcal{A})$ was already noted.

2.3 Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are said to be independent of each other if

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})P(\mathcal{B}) \tag{2.7}$$

It follows from (2.6) that if $\mathcal{A}$ and $\mathcal{B}$ are independent, then $P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A})$. That is, restriction of the sample space to $\mathcal{B}$ gives no additional information about whether or not $\mathcal{A}$ will occur.

Examples:

1.) Two dices with the same number: $\mathcal{A}$: tosing number six with die A.
$\mathcal{B}$: tosing number six with die B.

Since we know that the result of die A is independent from the result of die B: $P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})P(\mathcal{B}) = 1/6 \times 1/6 = 1/36$. And we also know that $P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A})$. Thus the likelyhood of $\mathcal{A}$: (tosing number six with die A) is independent on the result of die B, $\mathcal{B}$:.
2.) Weather events:
$\mathcal{A}$: Rainfall in Melbourne
$\mathcal{B}$: Rainfall in New York
$\mathcal{B}$: Rainfall in Geelong

Are the events $\mathcal{A}$ and $\mathcal{B}$ independent? If so: $P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A})$. Thus the probability of Rainfall in Melbourne does not depend on the Rainfall in New York.

Are the events $\mathcal{A}$ and $\mathcal{C}$ independent? Probably not, so: $P(\mathcal{A}|\mathcal{C}) \neq P(\mathcal{A})$. Thus the probability of Rainfall in Melbourne does depend on the Rainfall in Geelong.

3.) Suppose $\mathcal{A}$ represents a weather event and $\mathcal{B}$ its forecast by a weather forecast service. If $\mathcal{A}$ and $\mathcal{B}$ are independent, then the forecasting system does not produce skillful weather forecasts: a 'bad' weather forecast does not change out perception of the likelihood of this weather event. So a skillful weather forecast depend on the weather, but of cause the weather does not depend on the forecast. So dependence that imply a causality direction.

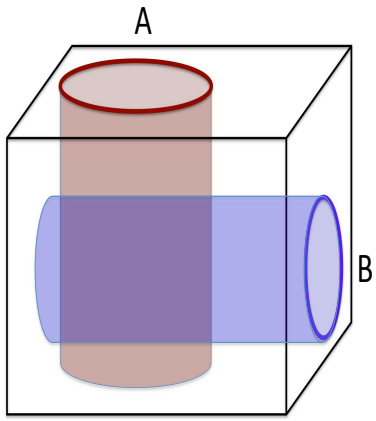


Figure 2.3: Sketchs to illustrate independence.

Chapter 3

Probability Density and Distribution

3.1 Continuous Random Variables

A continuous random variable is a statistical variable, those value are continuously changing. The dimension along which the random variable is changing can be different and depends on the problem studied. The temperature, for example, is a continuous random variable, that can be studied either along the time dimension or along a spatial dimension.
In general we will sample continuous random variable in discrete samples, which makes the continuous random variable a discrete random variable. In most cases this simply changes most definitions below by replacing the integrals against sums.

3.2 The Probability Density and Distribution Functions

Let $\mathbf{X}$ be a continuous random variable that takes values in the interval Ω. The *probability density function* (*pdf*) for $\mathbf{X}$ is a continuous function $f_X(\cdot)$ defined on R with the following properties:

1. $f_X(x) \geq 0 \; for \; all \; x \; \epsilon \; \Omega$,
2. $\int_\Omega f_X(x)dx = 1$,
3. $P(\mathbf{X} \; \epsilon \; (a,b)) = \int_a^b f_X(x)dx \; for \; all \; (a,b) \subseteq \Omega$

Thus the *pdf* has the unit $1/unit(x)$, since it is a density. Some *pdf*s of observed daily mean climate variables are shown in Fig. 3.1.
The *cumulative distribution function* for $\mathbf{X}$ is a non-decreasing differentiable function $F_X(\cdot)$ defined on R with the following properties

$$\lim_{x \to -\infty} \; F_X(x) = 0$$

,

$$\lim_{x \to \infty} \; F_X(x) = 1$$

,

$$\frac{d}{dx} F_X(x) = f_X(x)$$

.

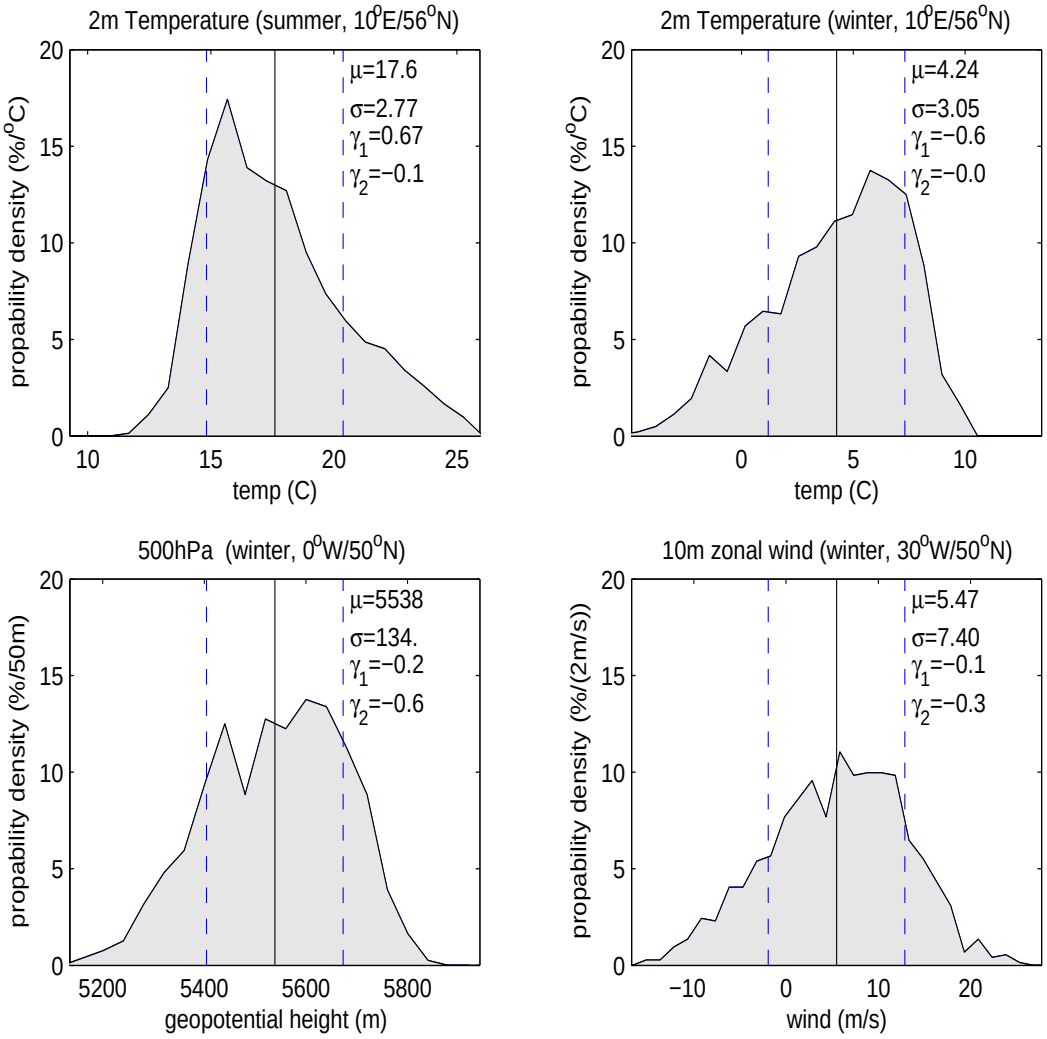


Figure 3.1: Some *pdf*s of observed daily mean climate variables.

Thus,

$$F_X(x) = \int_{-\infty}^{x} f_X(r)dr \tag{3.1}$$

The cumulative distribution function is non-dimensional and is often useful for computing probabilities because

$$P(\mathbf{X} \,\epsilon\, (a,b)) = F_X(b) - F_X(a) \tag{3.2}$$

3.3 Expectation

The expected value of a continuous random variable $\mathbf{X}$ is given by

$$\mathcal{E}(\mathbf{X}) = \int_{\Omega} x f_X(x)dx \tag{3.3}$$

If $g(\cdot)$ is a function then the definition of the expected value of $g(\mathbf{X})$ is

$$\mathcal{E}(g(\mathbf{X})) = \int_{\Omega} g(x) f_X(x)dx \tag{3.4}$$

Further rules apply

$$\mathcal{E}(g_1(\mathbf{X}) + g_2(\mathbf{X})) = \mathcal{E}(g_1(\mathbf{X})) + \mathcal{E}(g_2(\mathbf{X})) \tag{3.5}$$

$$\mathcal{E}(ag(\mathbf{X}) + b) = a\mathcal{E}(g(\mathbf{X})) + b \tag{3.6}$$

3.4 The Central Moments: Location, Scale, and Shape Parameters

The kth moment $\mu^{(k)}$ of a continuous random variable $\mathbf{X}$ is given by

$$\mu^{(k)} = \mathcal{E}(x^k) = \int_\Omega x^k f_X(x) dx \tag{3.7}$$

The kth central moment $\mu'^{(k)}$ of a continuous random variable $\mathbf{X}$ is the expectation of $(\mathbf{X} - \mu)^k$, given by

$$\mu'^{(k)} = \int_\Omega (x - \mu)^k f_X(x) dx \tag{3.8}$$

Note, that the central moments are the moments of the anomalies $x' = x - \mu$, where the mean μ is often defined as a seasonal cycle, $\mu = \mu(season)$.
Most characteristics of a distribution can be summarized through the use of simple functions of the first four moments. These slightly modified parameters are the mean, variance (standard deviation), skewness and kurtosis. Note, observed continuous random variables will in general be more complex, with non-zero moments $\mu^{(k>4)}$.

Mean

The mean also known as the location parameter is given by the first moment

$$\mu = \mu^{(1)} \tag{3.9}$$

Thus it is the expectation as in eq. [3.3]. It can also be considered as the center of mass of the *pdf*, since the definition is similar.

Variance and Standard Deviation

The variance is given by the second central moment

$$Var(\mathbf{X}) = \int_\Omega (x - \mu)^2 f_X(x) dx \tag{3.10}$$

For a linear function of $\mathbf{X}$ the variance is

$$Var(a\mathbf{X} + b) = a^2 Var(\mathbf{X}) \tag{3.11}$$

Hence, if a random variable is shifted by a constant, its variance does not change. On the other hand, multiplying a random variable with a constant does change the variance.
A more practical parameter for the scale of the variability of the random variable is the standard deviation $\sigma_X = \sqrt{Var(X)}$, because this measure has the same unit length as $\mathbf{X}$. The standard deviation is a measure of the 'width' of the *pdf*, see Fig. 3.1 for some examples.

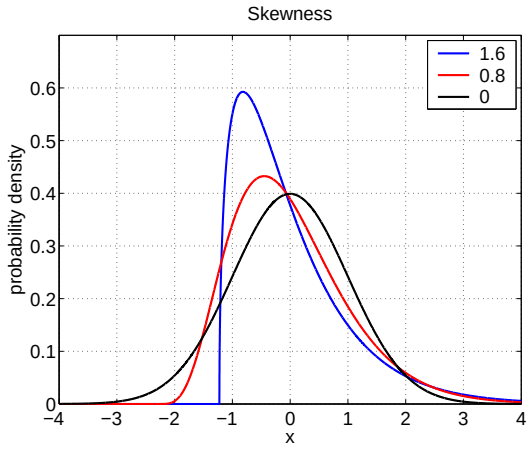


Figure 3.2: Illustration of differently skewed (see legend) distributions, with a $\chi^2(k = 3)$ (doted),a $\chi^2(k = 10)$ (dashed) and a normal (solid) distribution. Note the χ^2-distributions are shifted and scaled to have $mean = 0$ and $\sigma = 1$.

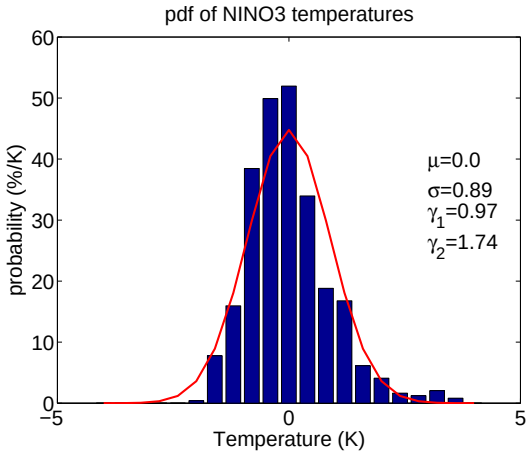


Figure 3.3: *pdf* of El Niño monthly mean SST.

It can be very limited in its interpretation, if the *pdf*s is very different from a normal distribution. The Chebyshev's inequality gives, however, a upper limit for the probability of $\mathbf{X}$ to take values far away from the mean:

$$P(|\mathbf{X} - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2} \tag{3.12}$$

The probability $P(|\mathbf{X} - \mu| \geq \lambda\sigma)$ is much smaller than $\frac{1}{\lambda^2}$ for normal *pdf*s, see section 3.7.
If a random variable, such as rainfall or wind speed, takes only positive values a scale parameter called the coefficient of variation

$$C_X = \sigma_X/\mu_X \tag{3.13}$$

is sometimes used. The standard deviation of such variables is often proportional to the mean and it is therefore useful to describe the scale parameter relative to the mean. See the log-normal distribution for instance, Section 3.9.

Skewness

The skewness is a scaled version of the third central moment that is given by

$$\gamma_1 = \int_\Omega \left(\frac{x - \mu}{\sigma}\right)^3 f_X(x)dx \tag{3.14}$$

The scaling of γ_1 by σ makes the skewness on non-dimensional shape parameter; γ_1 is independent of the size of σ.
Symmetric distributions have $\gamma_1 = 0$. Hence the skewness is a measure of the asymmetry of the *pdf*. Distributions with $\gamma_1 > 0$ are said to be skewed to the right, which means that positive extreme values $(x - \mu) > 0$ are more likely than negative extreme values and vice versa for $\gamma_1 < 0$. Fig. 3.2 illustrates some *pdf*s which are positively skewed. We can see that a skewness of 0.8 is already leading to quite different likelihoods for the extreme values and the skewness of 1.6 is very different

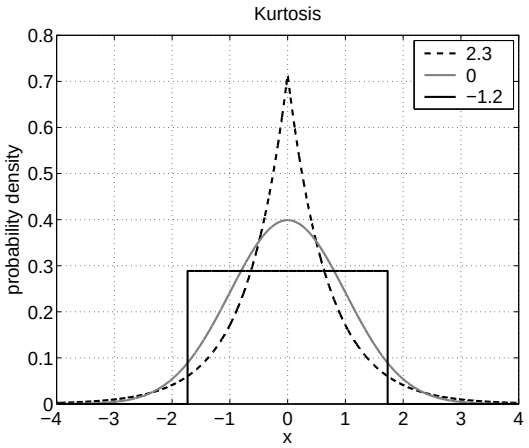


Figure 3.4: Illustration of distributions with different kurtosis (see legend), but with identical means and variance based on a $e^{-|x|}$ (dashed), a normal (gray solid) and a uniform (black solid) distribution.

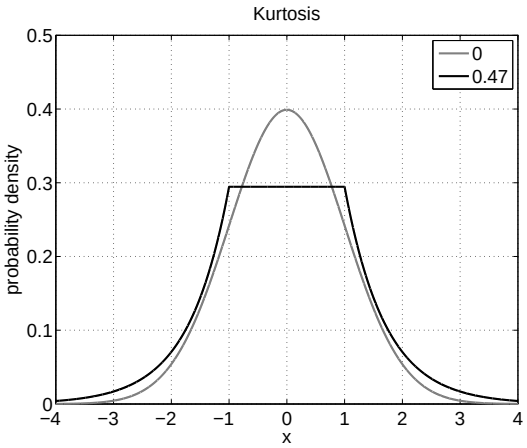


Figure 3.5: As in 3.4, but the $e^{-|x|}$ distribution is modified to have a uniform (flat) distribution for near zero values (black solid line).

form the normal, symmetric, distribution. For most practical problems we may consider the *pdf* as skewed if the absolute values of the skewness are larger than 0.2.
Fig 3.1 shows that 24hrs mean 2m temperature in Germany is skewed negatively in winter and positively in summer. The skewness is related to changes in winds, where no or north-east winds in winter lead to extreme cold days and in summer no or southern winds lead to extreme hot days.
The best known skewed climate variability is the El Niño mode (Fig.3.3. Here the SST is positively skewed, with strong warm events called 'El Niño' and the weaker negative events are called 'La Niñia'.

Kurtosis

The kurtosis, a scaled and shifted version of the fourth central moment, is given by

$$\gamma_2 = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma} \right)^4 f_X(x) dx - 3 \tag{3.15}$$

The scaling makes the kurtosis a non-dimensional shape parameter. The kurtosis is also shifted by 3 for reference to the normal (gaussian) distribution, those unshifted kurtosis is 3. Note that some statistical programs (e.g. MATLAB) does not shift the kurtosis, thus the kurtosis of the normal *pdf* is 3.
The kurtosis γ_2 is usually a measure of the peakedness of the distribution, see fig. 3.4. A $\gamma_2 < 0$ refers to a distribution which is less peaked than the normal (gaussian) distribution and vice versa for $\gamma_2 > 0$. A uniform distribution is an example for $\gamma_2 < 0$ and $f_X(x) = 1/|x-\mu|$ for $\gamma_2 > 0$. But, note that the kurtosis parameter is more sensitve to the extrem values than to the near mean values, which follows from its definition ($\sim (x-\mu)^4$). In some unusual cases the pdf will be less peaked near the mean than the normal pdf, but the kurtosis will be positive, due to larger probabilities for extrem values, see Fig. 3.5.
The skewness and kurtosis are shape parameter which are useful in the analysis of extreme values where debate of the merits of various distributions is intense. However, skewness and kurtosis are

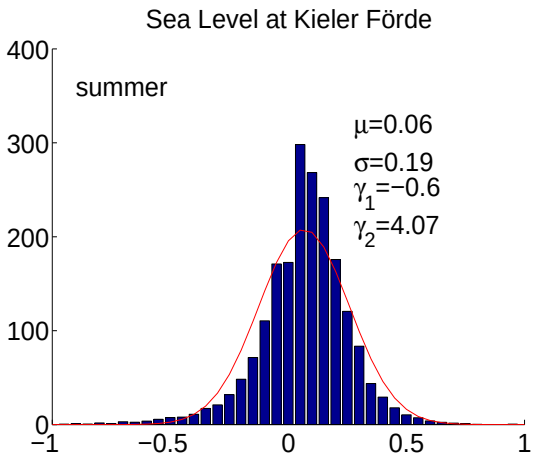


Figure 3.6: *pdf* of daily Kieler Foerde gauge heights summer month [in meter].

often difficult to estimate and the *pdf* of the extreme values is often not sufficiently determined by skewness and kurtosis.
Fig. 3.6 shows that the sea level in at mouth of the Kiel Foerde has a large kurtosis, with much more frequent extreme lows and high, than a normal *pdf*.

3.5 Median and Quantiles

For many physical quantities the mean and variance are effected by the tail ends (likelihood of extreme values) of the *pdf* and are sometimes bad measures of the variability and the peak of the *pdf*.
The median and quantiles are in general the more robust parameter of the *pdf*, because they are insensitive to the tail ends of the *pdf*. The median ,m_{50}, is the solution of

$$F_X(m_{50}) = 0.5 \tag{3.16}$$

It presents the middle of the distribution in the sense that

$$P(x \leq m_{50}) = P(x \geq m_{50}) = 0.5 \tag{3.17}$$

Exactly 50% of all random values will be below the median and Exactly 50% will be above. Note that the median is different from the mean if $\mu^{(k\geq 3)} \neq 0$; for odd values of k. The median is, in contrast to the mean, insensitive to the extreme value distributions and is therefore a much more robust estimate of the *pdf* if the sample size is small.
The median is a special case of a p-quantile, the point x_p for which

$$P\left(\mathbf{X} \in (-\infty, x_p)\right) = p \tag{3.18}$$
$$P\left(\mathbf{X} \in (x_p, \infty)\right) = 1 - p \tag{3.19}$$

The p-quantile is the solution of

$$F_X(x_p) = p \tag{3.20}$$

In Fig. 3.7 the median, 10%-quantile and the 90%-quantile are illustrated for the log-normal distribution.

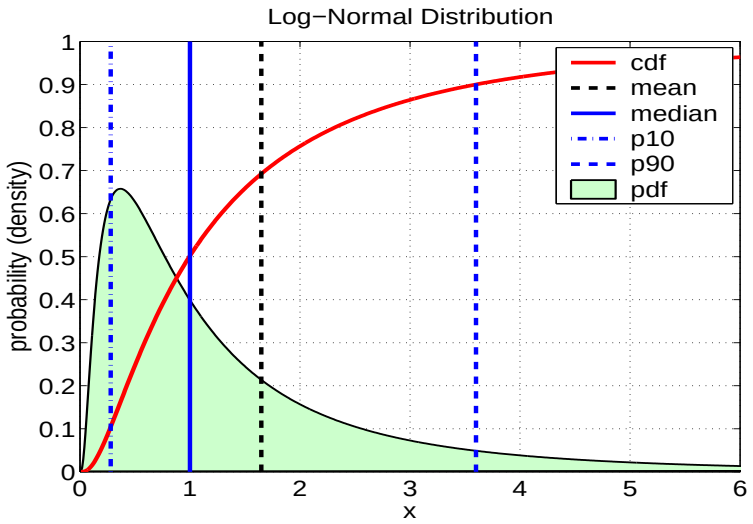


Figure 3.7: The Log-Normal distribution and its cumulative distribution function, mean, median, 10%-quantile and the 90%-quantile .

3.6 The Uniform Distribution

A random variable that takes values in an interval $[a, b]$ is said to be uniform if it has a *pdf* that is constant inside the interval and 0 outside. Thus,

$$f_X(x) = \begin{cases} 1/(b-a) & \forall\, x \ \in [a,b] \\ 0 & elsewhere \end{cases} \tag{3.21}$$

and the cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & for\ x \leq a \\ (x-a)/(b-a) & \forall\, x \ \in [a,b] \\ 1 & for\ x \geq b \end{cases} \tag{3.22}$$

The uniform distribution is a function of a, b: $\mathbf{X} \sim \mathcal{U}(a, b)$. The central moments are:

$$\begin{aligned} \mu(\mathcal{U}(a,b)) &= \tfrac{1}{2}(a+b) \\ Var(\mathcal{U}(a,b)) &= \tfrac{1}{12}(b-a)^2 \\ \sigma(\mathcal{U}(a,b)) &= \sqrt{\tfrac{1}{12}}(b-a) \\ \gamma_1(\mathcal{U}(a,b)) &= 0 \\ \gamma_2(\mathcal{U}(a,b)) &= -1.2 \end{aligned} \tag{3.23}$$

The uniform distribution is symmetric ($\gamma_1 = 0$) und less peaked than the normal distribution ($\gamma_2 = -1.2$).

3.7 The Normal (Gaussian) Distribution

The normal distribution is of fundamental importance in statistical analysis, because most physical quantities are nearly normal distributed (see also the central limit theorem, section 3.8) and most

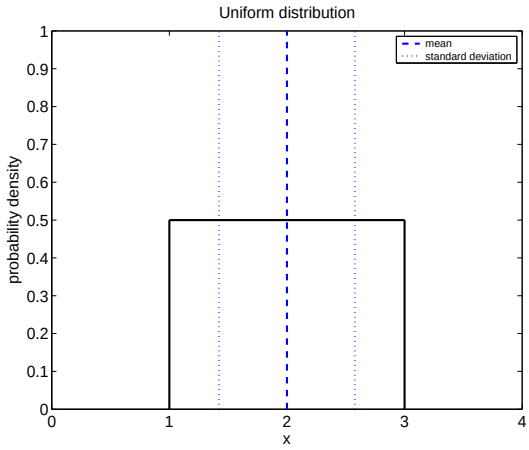


Figure 3.8: A uniform *pdf*, with $a = 1, \quad b = 3$.

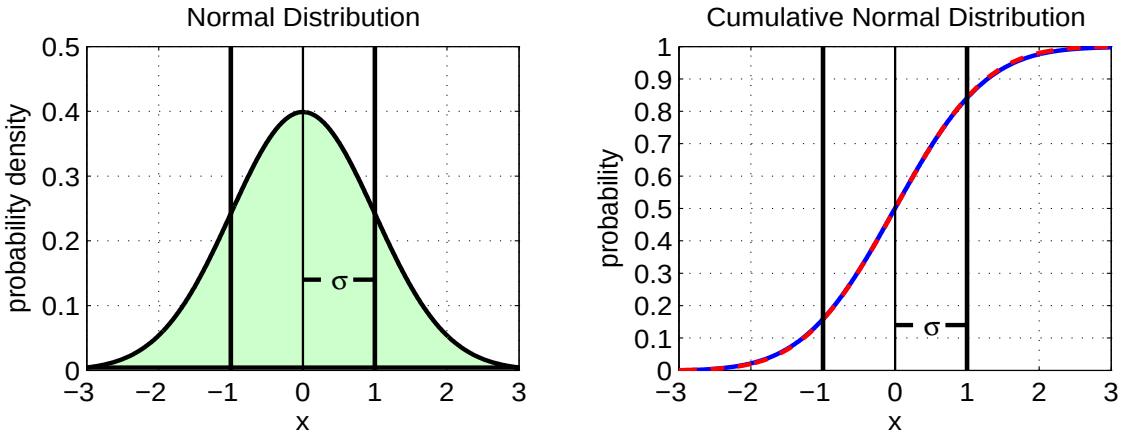


Figure 3.9: Normal probability density function and cumulative density function for $\mu = 0$ and $\sigma = 1$. The blue line in the right panel is a numerical estimate of $F_{\mathcal{N}}$ and the dash red line the approximation of $F_{\mathcal{N}}$ after eq. [3.25]

statistical analysis assume a normal distribution of the operators. Some statistical analysis will fail to produce useful results if the operators are not normal distributed.
The form of the normal distribution is defined by the mean and variance. Thus, we write $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that $\mathbf{X}$ has a normal distribution with parameters μ and σ.
The normal density function is given by

$$f_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \forall\, x \;\in\; R \tag{3.24}$$

see Fig. 3.9. The skewness and kurtosis vanish for the normal distribution.
The cumulative density function $F_{\mathcal{N}}$ cannot be given explicitly because the analytical form of $F_{\mathcal{N}}$ does not exist. $F_{\mathcal{N}}$ which is usually tabulated in statistical text books (see Storch and Zwiers), can also be evaluated by numerical integration or by using a simple approximation. For most purposes

$$F_{\mathcal{N}}(x) \approx \left(1 + sign(\frac{x-\mu}{\sigma})\sqrt{1 - e^{-2(\frac{x-\mu}{\sigma})^2/\pi}}\right)/2 \tag{3.25}$$

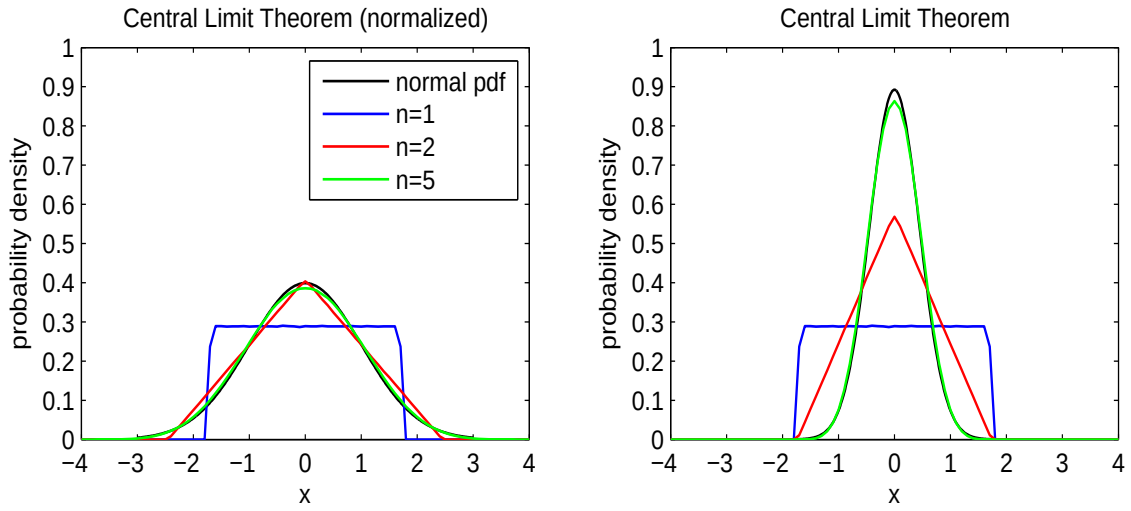


Figure 3.10: Illustration of the central limit theorem with a uniform distribution for $n = 1, 2, 5$. The normal distribution is shown for comparison. (a) all pdfs are normalized (scaled) for better comparison of the shape; illustrating eq. [3.26]. (b) pdfs are not normalized; illustrating eq. [3.27].

is a good approximation (compare red dashed line with the solid blue line Fig. 3.9).

Note that about 68% of normal distributed random values are within the interval $[\mu - \sigma, \mu + \sigma]$. The likelihood decreases fast with only 95% of the random values $|x - \mu| < 2\sigma$ and with 99.99% random values $|x - \mu| < 4\sigma$. Thus it is unlikely to have a value $|x - \mu| > 4\sigma$ if less than 10^4 samples are taken. The 4σ limit is therefore often chosen to eliminate false data. It is also important to note that a 4σ signal is not only unusual, it may in many physical system be in different regime and could therefor follow different physical laws. That means if you do an experiment in which you disturb the system by a 4σ signal, we are in a regime which is far away from normal, and it may in many cases not give as any meaning full results about the nature of the normal system. So make sure to test models if signals no larger than 4σ of the undisturbed/observed *pdf*.

3.8 Central Limit Theorem

The central limit theorem is of fundamental importance for statistics because it establishes the dominant role of the normal distribution.

If $\mathbf{X}_k, k = 1, 2, ...,$ is an infinite series of independent and identically distributed random variables with $\mathcal{E}(\mathbf{X}_k) = \mu$ and $Var(\mathbf{X}_k) = \sigma^2$ then the average $\frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k$ is asymptotically normal distributed. That is,

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_k - \mu)}{\frac{1}{\sqrt{n}}\sigma} \sim \mathcal{N}(0, 1) \tag{3.26}$$

Note that the central limit theorem holds regardless of the *pdf* of $\mathbf{X}_k$ According to the central limit theorem, the distribution of a mean (or sum) of independent and identically distributed random variables converges towards a normal distribution as the number, n of independent random variables increases and the standard deviation of the mean decreases by $1/\sqrt{n}$. That of the sum would increase by $\sqrt{n}$.

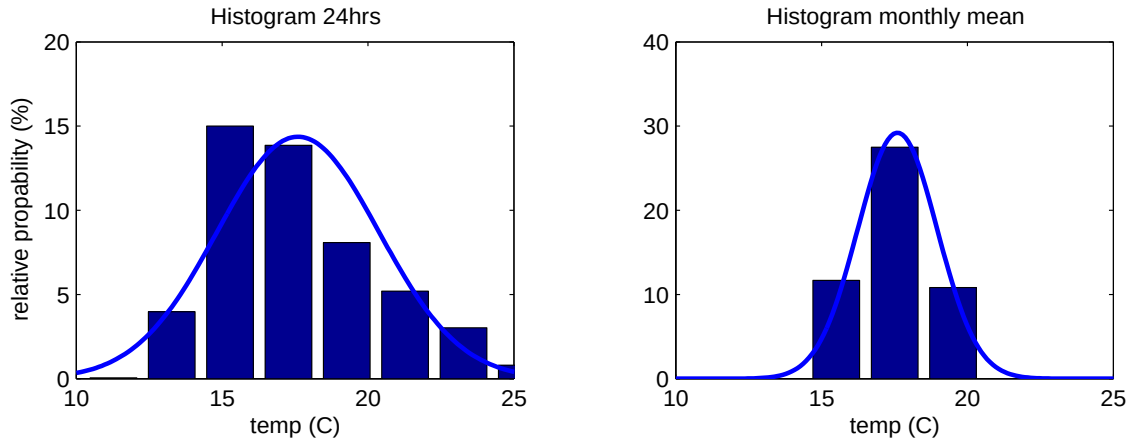


Figure 3.11: Illustration of the central limit theorem on the basis of 24hrs and monthly mean surface temperatures at $56^oN/10^oE$ in July/August. The normal distributions with identical mean and variance are shown for comparison (solid lines).

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_k - \mu) \sim \mathcal{N}(0, \frac{\sigma^2}{n}) \Rightarrow \quad \sigma_\Sigma = \frac{\sigma}{\sqrt{n}} \tag{3.27}$$

However, nothing is known about when the convergence has made substantial progress. Fig. 3.10 illustrates how uniform distributed random variable converge to the normal distribution. We can see that in this case the convergence is already very near the normal distribution with $n = 5$.
The central limit theorem has of cause some implication for physical climate variables. The temporal or spatial average of physical climate variable can sometimes be considered as the average of independent and identical distributed random variables, if, of cause the physical climate variable are temporally or spatially independent and identical distributed. Fig. 3.11 illustrates this on the basis of 24hrs surface temperatures and monthly mean surface temperatures. The 24hrs values are significantly skewed, while the monthly mean values are nearly normal distributed with a smaller standard deviation.

3.9 The Log-Normal Distribution

Some physical quantities are positive definite, such as rainfall or wind speed. Such quantities often follow the Log-Normal Distribution.
A random variable $\mathbf{X}$ has a log-normal distribution with the median θ if $ln(\mathbf{X}) \sim \mathcal{N}(ln(\theta), \sigma)$. The density function is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}e^{\frac{-(ln(x)-ln(\theta))^2}{2\sigma^2}} \quad for\ x > 0 \tag{3.28}$$

The moments are given by

$$\mathcal{E}(\mathbf{X}^k) = \theta^k e^{(k\sigma)^2/2} \tag{3.29}$$

Therefore

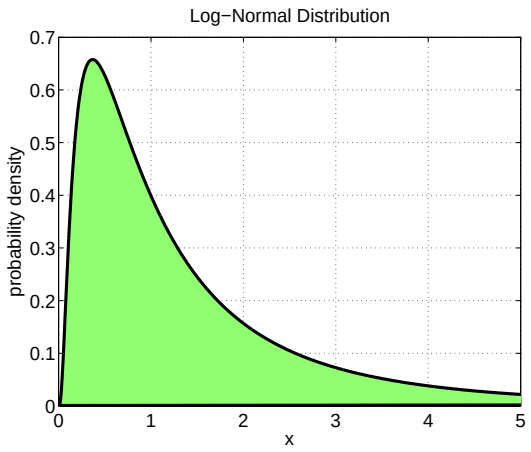


Figure 3.12: The log-normal-distribution for $ln(\theta) = 0$ and $\sigma = 1$.

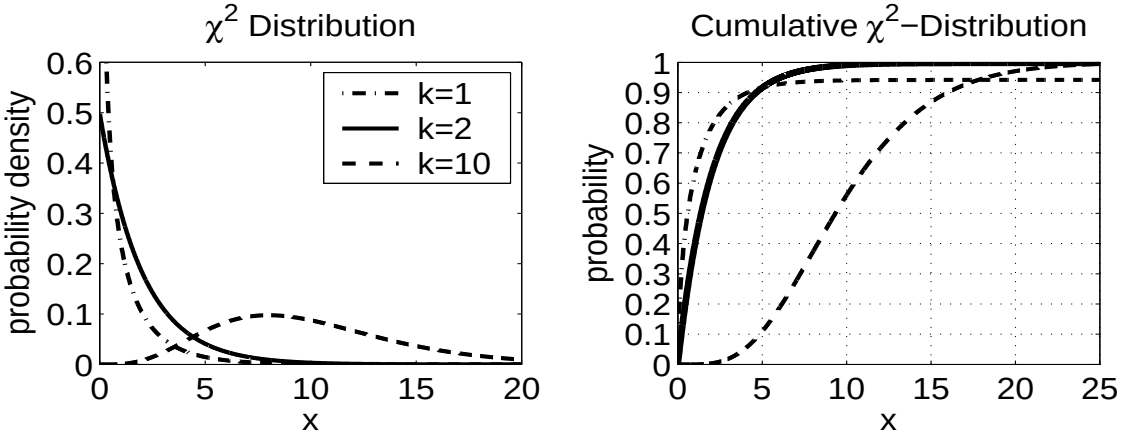


Figure 3.13: The χ^2 distribution for different degrees of freedom $k = 1, 2, 10$.

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &= \theta e^{\sigma^2/2} \\ Var(\mathbf{X}) &= \theta^2 e^{\sigma^2}(e^{\sigma^2} - 1) \\ \gamma_1 &= \sqrt{e^{\sigma^2} - 1}(e^{\sigma^2} - 1) \end{aligned} \tag{3.30}$$

3.10 Distribution Related to the Normal Distribution

It is often important to know the uncertainty of an estimate of the mean, variance or correlation. For such significant tests the tail ends of the relevant *pdf*s are important to estimate the cumulative distributions. In these cases the χ^2, t-, F-distributions are often needed. See also section ??? for significant tests.

3.11 The χ^2 Distribution

The χ^2 distribution is defined as the sum of k independent squared $\mathcal{N}(0, 1)$ random variables. The most important purpose of the χ^2 distribution is for the *pdf*s of estimates of the variance. The

spectral variance estimates are χ^2 distributed (see section 9). Squared wind speed with $k = 2$ is another example. The form of this distribution depends only on one parameter, k, referred to as the number degrees of freedom (dgf).
The probability of a $\chi^2(k)$ random variable is given by

$$f_X(x) = \frac{x^{(k-2)/2}e^{-x/2}}{\Gamma(k/2)2^{k/2}} \quad if\ x > 0 \tag{3.31}$$

where Γ denotes the Gamma function. [1]
We write $\mathbf{X} = \chi^2(k)$ to indicate that a random variable $\mathbf{X}$ is χ^2 distributed with k degrees of freedom. The distribution is tabulated in statistical text books (see Storch and Zwiers).
Some important characteristics of the χ^2 distribution:

- The χ^2 distribution is additive for independent $\mathbf{X}_1$ and $\mathbf{X}_2$ with dgf k_1 and k_2, than $\mathbf{X}_1 + \mathbf{X}_2$ is a $\chi^2(k_1 + k_2)$ distribution. Thus the χ^2 distribution can be thought of as a sum of $\chi^2(1)$ distributions.

- The χ^2 distribution is skewed, where distributions with smaller k are more skewed. This is important to note, because χ^2 distributions with small ks tend to underestimate the expected value, as you can see from Fig. 15.7.

- It follow from the central limit theorem that the χ^2 distribution converges to the normal distribution; $\chi^2(30)$ is very near the normal distribution.

- $\chi^2(1)$ and $\chi^2(2)$ have their modes (their most likely values; the peak in the pdf) at the origin.

- The spread of the distribution depends strongly on k. the moments are:

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &= k \\ Var(\mathbf{X}) &= 2k \end{aligned} \tag{3.32}$$

  We see that both $\mathcal{E}(\mathbf{X})$ and $Var(\mathbf{X})$ are dimensionless. If we include the dimensions and express $Var(\mathbf{X})$ as a function of $\mathcal{E}(\mathbf{X})$ we find:

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &= k \cdot c \\ Var(\mathbf{X}) &= 2k \cdot c^2 = 2\frac{\mathcal{E}(\mathbf{X})^2}{k} \end{aligned} \tag{3.33}$$

  So we see that σ, the spread of the pdf, decreases when k increases.

- In statistical tests we often need the p-quantiles, x_p, for defining confidence intervals. The dimensionless x_p can be read from the cumulative χ^2-distribution, by scaling it with $\frac{\mathcal{E}(\mathbf{X})}{k}$ we can include the physical dimensions.

3.12 The Students t Distribution

The Students t distribution is of fundamental importance for testing the significance of the differences in the means or the significant of a correlation value (see sections ???). In both cases we need to assume a normal distribution and a χ^2 distribution.

[1] The Gamma function takes the factorial function onto the real space; $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ for $x > 0$.
Thus $\Gamma(1) = 1$ and $\Gamma(x + 1) = x\Gamma(x)$

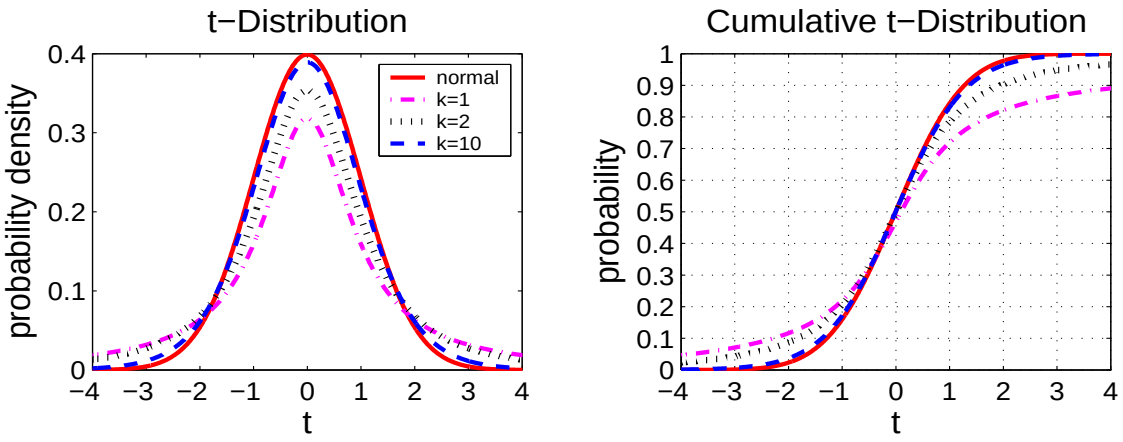


Figure 3.14: The t-distribution for different degrees of freedom $k = 1, 2, 10$.

In particular we write for the Students t Distribution $\mathbf{T} = t(k)$ ($\mathbf{T}$ instead of $\mathbf{X}$ to indicate that it is a test variable), if $\mathbf{A}$ and $\mathbf{B}$ are independent random variables such that

$\mathbf{A} \sim \mathcal{N}(0,1)$ and $\mathbf{B} \sim \chi^2(k)$

then

$$\frac{\mathbf{A}}{\sqrt{\mathbf{B}/k}} \sim t(k). \tag{3.34}$$

This kind of relation between $\mathbf{A}$ and $\mathbf{B}$ is used in test of differences in the mean, for instance. With the mean:$\mathbf{A} =\sim \mathcal{N}(0,1)$ and $\mathbf{B} = Var(A) \sim \chi^2(k)$. The test variable is $\frac{\mu_1-\mu_2}{\sigma} = \frac{\mathbf{A}}{\sqrt{\mathbf{B}/k}}$.
The probability density function is given by

$$f_T(t) = \frac{\Gamma((k+1)/2)(1+t^2/k)^{-(k+1)/2}}{\sqrt{k\pi}\Gamma(k/2)} \tag{3.35}$$

The distribution and especially the cumulative t-distribution is tabulated in statistical text books (see Storch and Zwiers).
Some important characteristics of the t-distribution:

- $\mathcal{E}(\mathbf{T}) = 0 \quad for\ k \geq 2$
- $Var(\mathbf{T}) = \frac{k}{k-2} \quad for\ k \geq 3$
- The t-distribution is symmetric.
- It follow from the central limit theorem that the t-distribution converges to the normal distribution; $t(30)$ is very near the normal distribution.

3.13 The Fisher F-Distributions

The F Distributions is needed if $\mathbf{X}$ and $\mathbf{Y}$ are independent random variables such that $\mathbf{X} \sim \chi^2(k)$ and $\mathbf{Y} \sim \chi^2(l)$, then

$$\frac{\mathbf{X}/k}{\mathbf{Y}/l} \sim F(k,l). \tag{3.36}$$

The probability density function is given by

$$f_F(f) = \frac{(k/l)^{k/2}\Gamma((k+l)/2)}{\Gamma(k/2)\Gamma(l/2)} f^{(k-2)/2}\left(1+\frac{k}{l}f\right)^{-(k+l)/2} \tag{3.37}$$

The distribution and especially the cumulative F-distribution is tabulated in statistical text books (see Storch and Zwiers). Some important characteristics of the F-distribution:

- It is positive definite.
- It is positively skewed.
- It converges to the χ^2 distribution for $l \to \infty$.

3.14 Summary of Theoretical Distributions

pdf	parameters	purpose	example
Uniform	$\mathcal{U}(a,b)$ interval boundaries	$X \in [a,b]$	wind direction, ice cover
Normal(Gauss)	$\mathcal{N}(\mu,\sigma^2)$ mean, variance	$X \in \mathbb{R}$	Temperature, Pressure
Log-normal	$\mathcal{N}(ln(\theta),\sigma^2)$ median, variance($ln(X)$)	$X \in \mathbb{R}^+$ or μ near boundary	Rain, wind speed
χ^2	$\chi^2(k)$ Number degree of freedom	$X = \sum_{i=1}^{k} X_i^2$ $X_i \sim \mathcal{N}(0,1)$	Variance, Spectral variance
Students t	$t(k)$ Number degree of freedom	$\frac{\mathbf{A}}{\sqrt{\mathbf{B}/k}}$ $\mathbf{A} \sim \mathcal{N}(0,1)$ and $\mathbf{B} \sim \chi^2(k)$	test of mean change
Fisher F	$F(k,l)$ Number degree of freedom	$\frac{\mathbf{X}/k}{\mathbf{Y}/l}$ $\mathbf{X} \sim \chi^2(k)$ and $\mathbf{Y} \sim \chi^2(l)$	test of variance change

3.15 Continuous Random Vectors / Multi-Variate Data

Until now we discussed the *pdf* of single continuous random variables, such as physical scalar values (e.g. temperature, wind speed or rainfall at one location). Often we need to evaluate the probability density function of bivariate (e.g. such as temperature and sea level pressure at the same time) or multivariate data, such as global temperature fields.
Multivariate data are continuous random vectors or fields for which we can carryover much of the structures we defined for continuous random variables. Instead of the probability of a scalar event we describe the probability to find a pair. e.g. the *pdf* of the co-variability of $\mathbf{X}$ and $\mathbf{Y} \sim f_{X,Y}(x,y)$ (e.g. NAO-index and temperature in Kiel) or between fields/vectors $\vec{\mathbf{X}}$ and $\vec{\mathbf{Y}} \sim f_{\vec{X},\vec{Y}}(\vec{x},\vec{y})$ (e.g. such as temperature and sea level pressure at the same time).
For details on how to define the *pdf* of continuous random vector see Storch and Zwiers. The calculation of the statistical parameters of the higher-dimensional *pdf*s is similar to those of the scalar functions. For instance:

e.g.:

$$\vec{\mu} = \int_{R^M} \vec{x} f_{\vec{X}}(\vec{x}) d\vec{x}$$

An example of a 2-dimensional *pdf* is shown in Fig. 4.1. It illustrates three different cases of 2-dimensional normal *pdf*s. In the uncorrelated case (left) the *pdf* is simply the product of the two scalar normal *pdf*s and the likelyhoods of X2 does not depend on the values of X1. But in the cases when X1 and X2 are correlated, the likelyhoods of X2 depend on the values of X1. For this characteristic we need to define new parameters unique to multivariate data, which will be discussed in this section.

Chapter 4

The Covariance Matrix

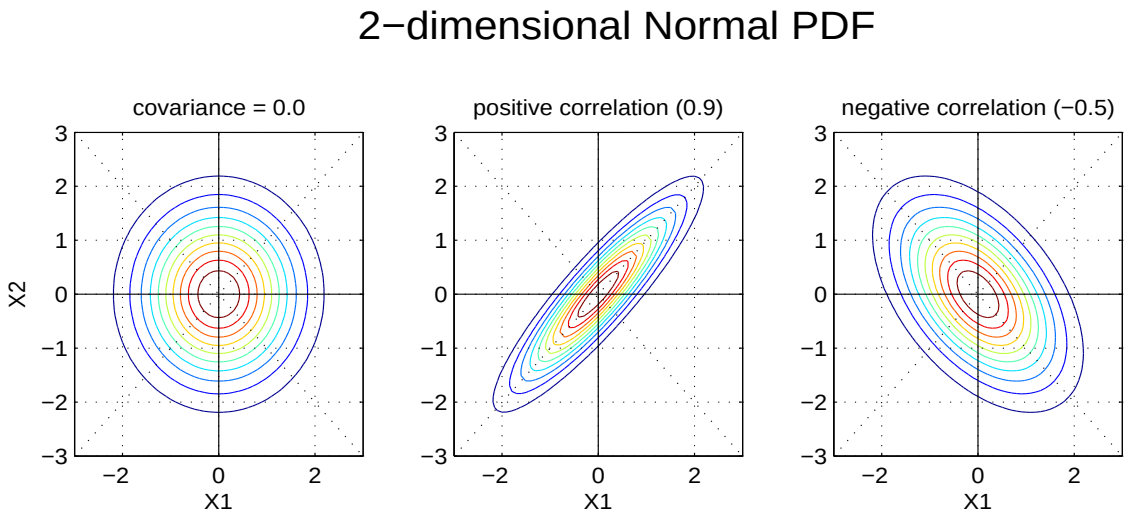


Figure 4.1: Three examples of 2-dimensional normal distributions for uncorrelated variables(left) and for correlated variables (middle and right). Note that the covariance is proportional to the correlation.

The concept of variance can be carried over to define a covariance between two continuous random variables. The covariance between a continuous random variable $\mathbf{X}$ and $\mathbf{Y}$ is

$$\sigma^2_{XY} = \int\int_{R^2} (x - \mu_x)(y - \mu_y) f_{X,Y}(x, y) dx dy \tag{4.1}$$

The characteristics of this equation and the covariance as such may best be illustrated by the examples of the 2-dimensional normal distribution in Fig. 4.1. The Integral becomes large when the product of $x'y'$ is large where the *pdf* is large too. In the first case (left) the *pdf* is as large for $x'y'$ being positive as it is for $x'y'$ being negative, so the integral/covariance becomes zero. In the second case (middel) the *pdf* is larger when $x'y'$ is positive (upper right and lower left quarter of the diagram), than when $x'y'$ is negative (upper left and lower right). So the integral/covariance becomes positive. Similar but with opposite sign in the last (right) case. So the covariance between X1 and X2 becomes large in amplitude if the *pdf* is 'focussed' on the diagonals. Or in other words the covariance between X1 and X2 becomes large in amplitude if the likelyhoods of X1 depend on the values of X2 and vice versa. So incontrast to the variance, the covariance is not just a measure of the spread of the distribution, but it is a measure of the syncronized spread.

If the continuous random variables $\mathbf{X}$ and $\mathbf{Y}$ are elements of vectors/fields $\vec{\mathbf{X}}$ and $\vec{\mathbf{Y}}$ we may be interested in the covariance between all possible pairs of $\mathbf{X_i}$ and $\mathbf{Y_j}$, namely the covariance matrix

$$\Sigma^2_{\vec{X},\vec{Y}} = \int_{R^m}\int_{R^n} (\vec{x}-\vec{\mu_x})(\vec{y}-\vec{\mu_y})^T f_{\vec{X},\vec{Y}}(\vec{x},\vec{y})d\vec{x}d\vec{y} \tag{4.2}$$

The covariance matrix Σ^2 of n,m-dimensional continuous random vector $\vec{X},\vec{Y}$ is a (m,n) matrix. The (i,j)th element of Σ^2 contains the covariance

$$\sigma^2_{x_i,y_j} = \int\int_{R^2} (x_i-\mu_{x_i})(y_j-\mu_{y_j}) f_{x_i,y_j}(x_i,y_j)dx_i dy_j \tag{4.3}$$

Some important characteristics of the covariance matrix:

- The covariance describes the tendency of jointly continuous random variables to vary in concert. If deviations of X_i and X_j from their respective means tend to be of the same sign, the covariance between X_i and X_j will be positive, if the deviations tend to be of opposite sign the covariance will be negative.

- X_i and X_j are said to be independent if the covariance is zero. Note that independence is sometimes used in a more restrict sense, where two continuous random variables can be dependent from each other even if the covariance is zero, see section ???.

- Note that the variance of a *pdf* is only a good measure of the spread of the distribution if the *pdf* is a near normal distribution. In analogy, the covariance is only a good measure of the joint variability of two continuous random variables if each of them is nearly normal distributed, see also section ???.

If $\vec{\mathbf{X}} = \vec{\mathbf{Y}}$ we have the covariance matrix Σ^2_{XX} of the vector/field with itself, which is called the auto-covariance matrix. Otherwise we call Σ^2_{XY} the cross-covariance matrix.
Some important characteristics of the auto-covariance matrix:

- The auto-covariance matrix is symmetric.

- The diagonal elements are the variance of the continuous random variable X_i, such that $\sigma^2_{ii} = Var(X_i)$. Thus the square root of the diagonal elements is the standard deviation vector/field.

An example of the auto-covariance matrix is shown in Fig. 4.2. The figure shows the square root of the diagonal elements of the auto-covariance matrix of global monthly mean sea surface temperatures (SST), namely the standard deviation of global anomalous monthly mean SST.
Examples of single columns/rows will be discussed in the following section, on the basis of the correlation matrix.
The analysis of the cross-covariance matrix plays an important role in the discussion of turbulence. If we are interested in the mean value of a product of two physical quantities $\overline{XY}$; the heat transport for instance ($X = \vec{v},\ Y = T$):

$$\overline{XY} = \overline{X}\cdot\overline{Y} + \overline{X'Y'} \tag{4.4}$$

The turbulent or transient part $\overline{X'Y'}$ is often the dominating term and it is the covariance between X and Y.

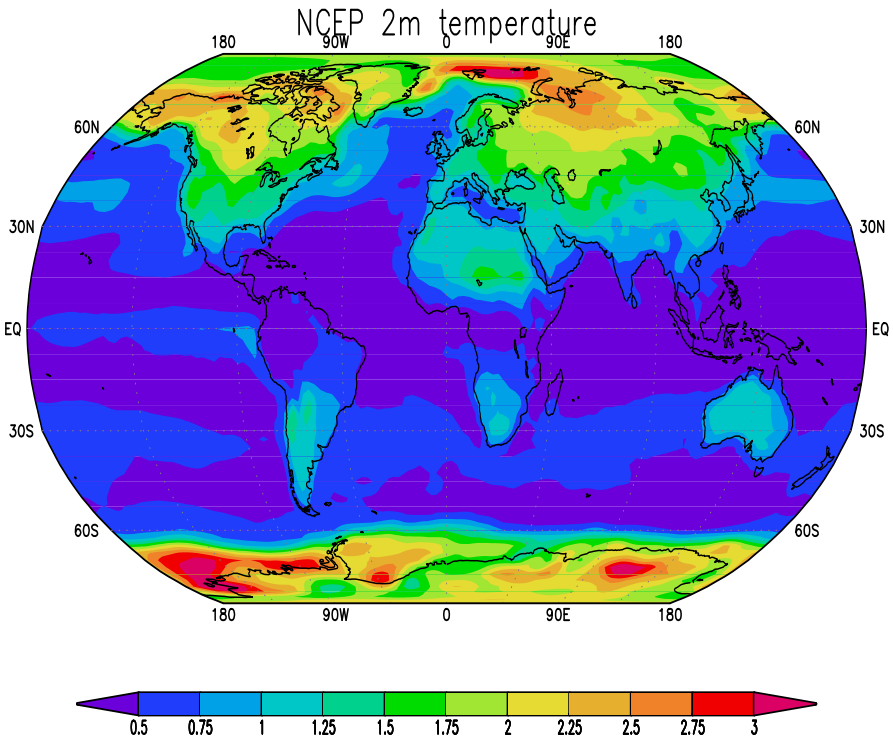


Figure 4.2: The global standard deviation of the NCEP monthly mean 2m temperature field. An Illustration of the square root of the diagonal elements of the covariance matrix. Units are in Kelvin.

4.1 The Correlation

A problem with the covariance as a measure of the covariability is that it has a squared ($[x]^2$ or $[x][y]$) unit scale, for which it is sometimes difficult to get a felling for. The measure that is scale invariant is the correlation.
The correlation between two random variables X and Y is given by

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{4.5}$$

Some characteristics of the correlation ρ_{xy}:

- The correlation coefficient takes values in the interval $[-1, 1]$
- $\rho_{x_i y_j}$ builds (i, j)th element of the correlation matrix between $\vec{\mathbf{X}}$ and $\vec{\mathbf{Y}}$.
- As for the covariance, the correlation coefficients are an indication of the extent to which the two variables X and Y are linearly related; that is , $Y = a + bX$.
- ρ_{xy}^2 can be interpreted as the explained variance. It is the proportion of the variance of one of the variables that can be represented by a linear model of the other.
- Note, that two variables with zero correlation can still be related by a non-linear relation.
- Note, that two variables with non-zero correlation, are not necessarily directly related to each other. Both can depend on a third variable.

- As for the covariance, the correlation is only a good measure to covariability if both variables are nearly normal distributed.
- $\rho_{X_i X_j}$ refers to the auto-correlation if X_i and X_j are variables of the same quantity (e.g. temperature). The cross-correlation otherwise.
- We refer to lag/lead correlations if the indices i, j refer to different points in time, see section (???).

The best way to understand correlation values is to study some examples:

Example 1: The Fig. 4.3 illustrates some correlation values by a scatter and time-series plot. We can see that the correlation value 0.9 refers to a very close agreement between the two random variables. The linear relation between the two random variables is still clear with a correlation of 0.6. For the 0.3 and 0.0 correlation examples it is difficult to subjectively (be visual inspection of the plots) decide if the two random variables are uncorrelated or not.

Example 2: The correlation is a measure of the linear relationship between two variables. Fig. 4.4 illustrates that two variables can have a strong non-linear relation, while the correlation coefficient is zero.

Example 3: An example of a correlation matrix is shown in Fig. 4.5. The Figure shows the diagonal elements of the correlation matrix between monthly mean surface temperature and the 1000hPa geopotential height. Such a plot is often called a "point to point" correlation map, since it represents the correlation between the two fields at identical spatial locations.

Example 4: The columns and rows of the covariance matrix are the correlation vectors/fields of one location with the rest of the field, see Fig. 4.6. These correlation fields are often called "box" correlations.

Example 5: The Fig. 4.7 illustrates a correlation of a random variable (the NINO3 SST index) with a random vector/field, which is a $1 \times m$ matrix, thus it is a vector or field. In this case the complete correlation matrix is presented as a global field. The global correlation field is often called the teleconnections of the index time-series.

Example 6: The Fig. 4.8 illustrates a correlation of the NAO-Index time series with the 1000hPa geopotential height field. In this case the complete correlation matrix is presented as a global field. Note that the NAO-index is constructed by Azores minus Iceland. So it is by construction positively (negatively) correlated with Azores (Icland). We can estimate the correlation of Island with the NAO-Index by: $\rho(NAOvsIsland) = \mathcal{E}()$.
... to be continued

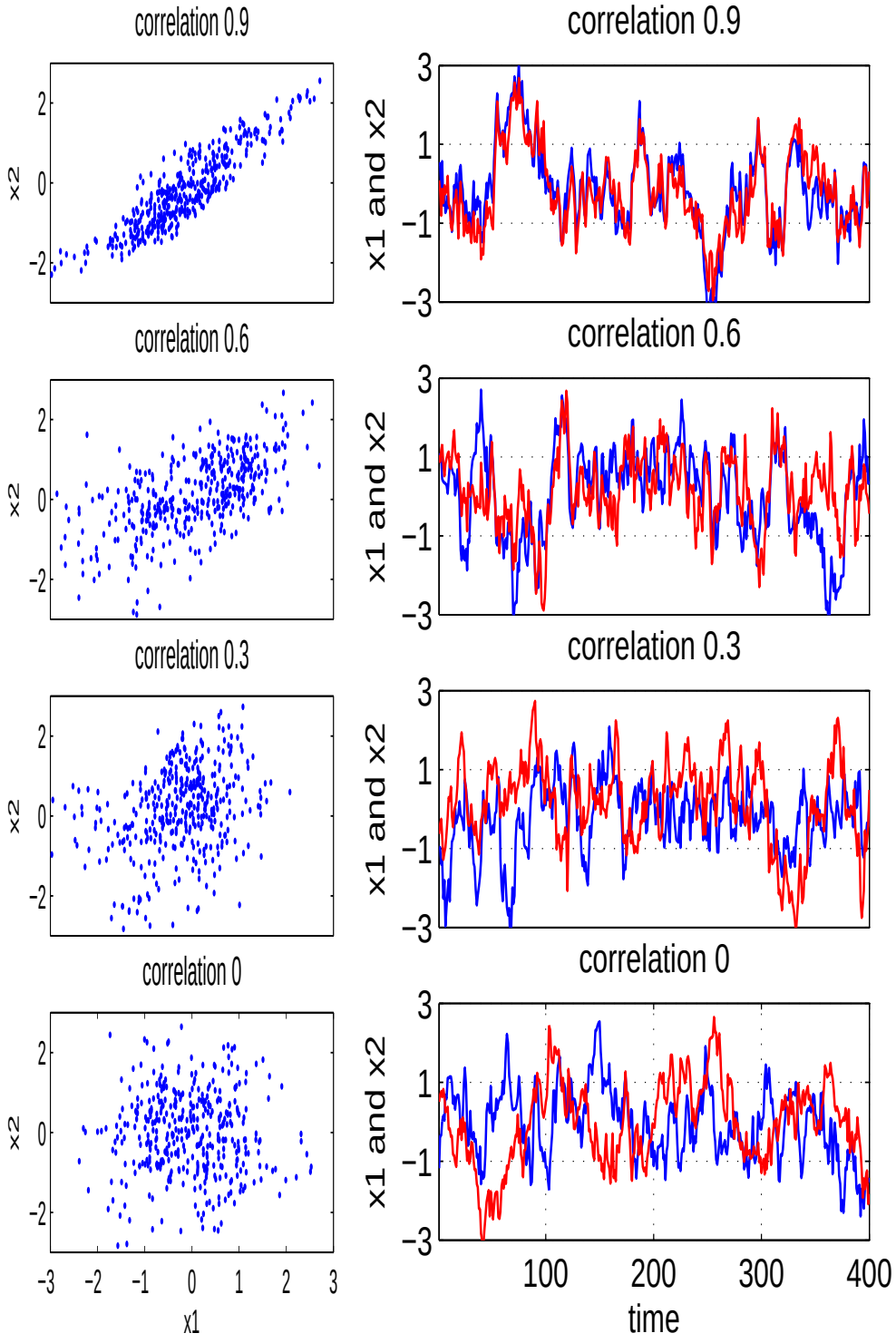


Figure 4.3: Illustration of correlation values by a scatter and time-series plot.

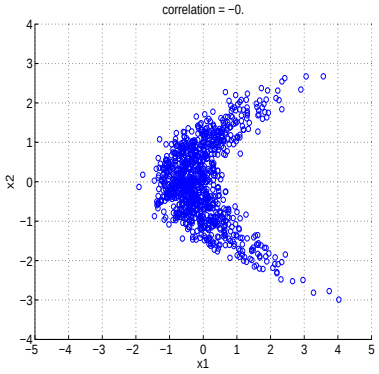


Figure 4.4: Illustration of a non linear relation between $x1$ and $x2$ with zero correlation coeffient.

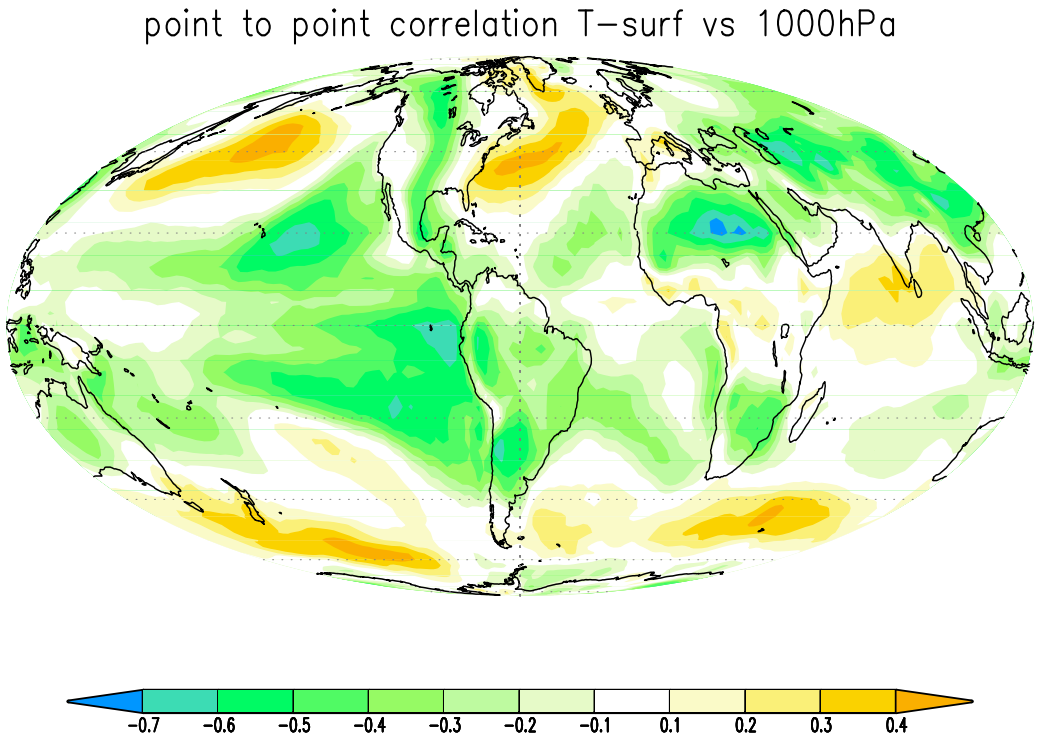


Figure 4.5: Correlation of monthly mean 2m-temperature with 1000hPa geopotential height. Data from the NCEP period.

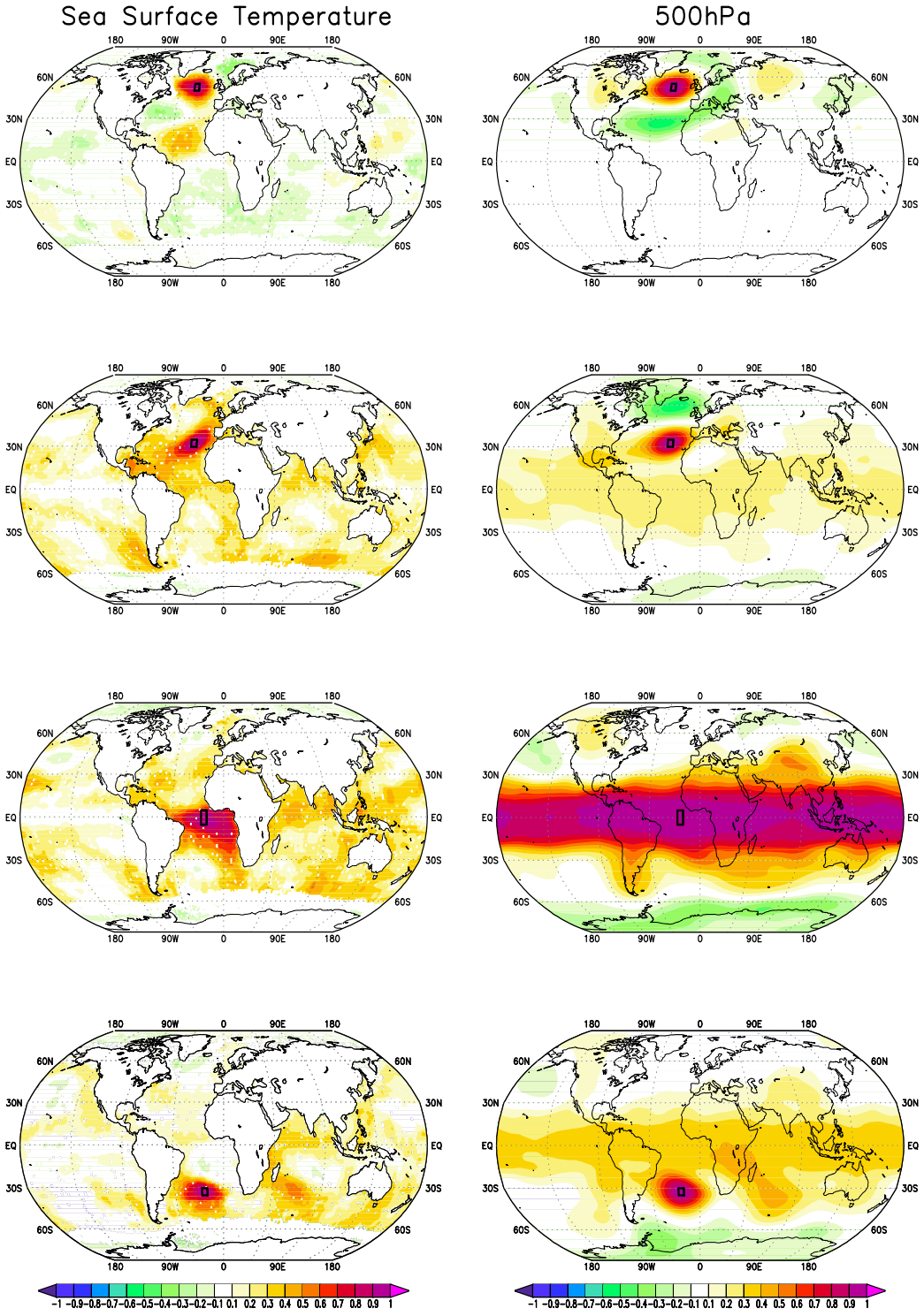


Figure 4.6: Global auto-correlation fields of one point(box) with the global field of monthly mean sea surface temperature (left) and NCEP 500hPa geopotential height (right).

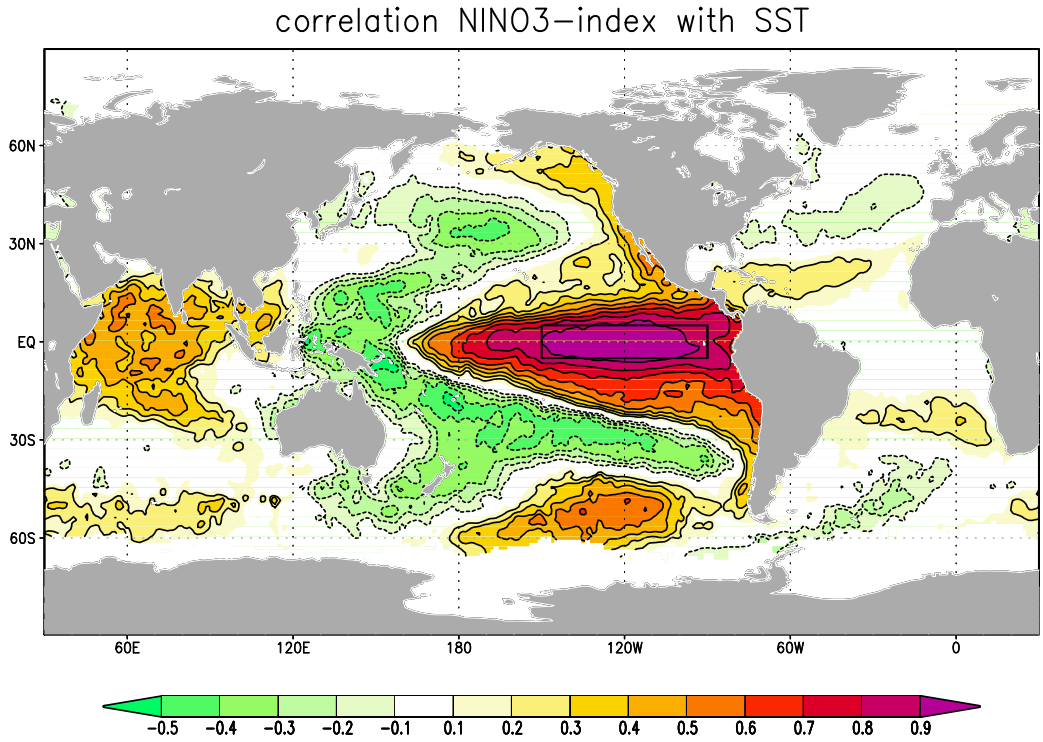


Figure 4.7: Correlation of monthly mean sea surface temperature with NINO3-index region. Data from HADISST 1950 to 2010; linearly detrended.

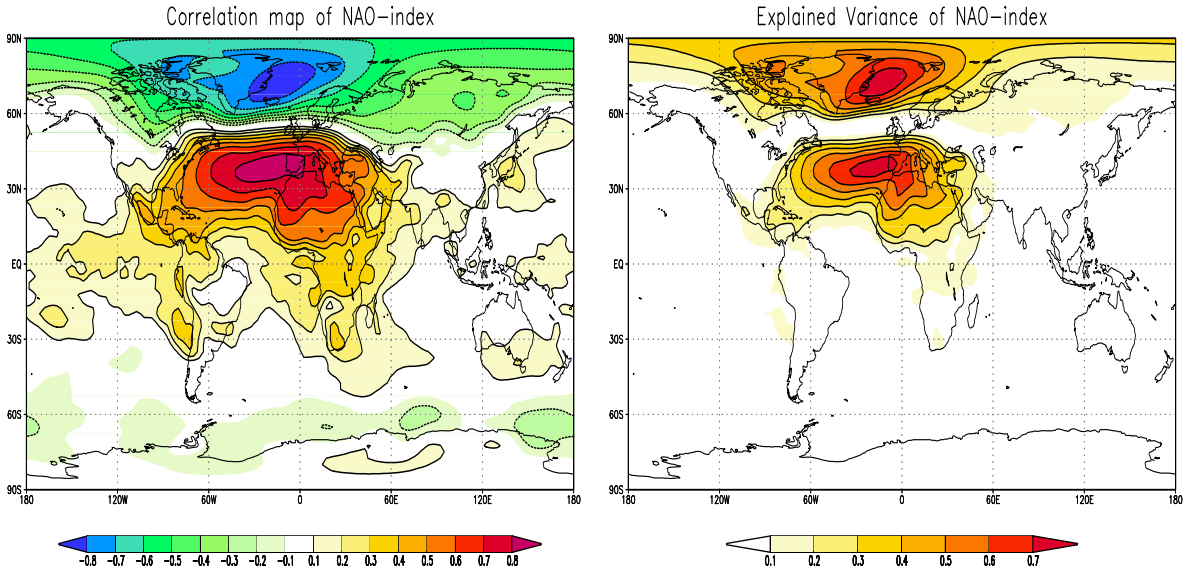


Figure 4.8: Correlation of monthly mean 1000hPa geopotential height field with the NAO-index (Azores - Island).

4.1.1 The interpretation of Correlation

The correlation seem to be a fairly straight forward concept, which may not be considered a complex idea or problem. But if you think about some of the implications in more details you will realize, most likely, that it indeed does have some potentials for confusion. We will now discuss a bit more in detail how you can interprete the correlation values and what potential pitfalls there are.

Consider the following statement:

Highly intelligent women tend to marry men who are less intelligent than they are.

You may first of all doubt this, but it actually is true (in statistical average). You would then, possibly, think about the character of highly intelligent women and may not have too many nice things to say about them. Maybe you would argue that they like to dominate their husbands? When you do this you start to interprete the statistical finding. You interprete this as something like:

$$I_{husband} = \rho * I_{wife} \tag{4.6}$$

With $0 < \rho < 1$. Thus the husbands are not as intelligent as the wives. So you build a linear model as discussed in the previous section, when we introduced the correlation ρ. But this model is neglecting something. First, consider now the following, also, true (observed) statement:

Highly intelligent men tend to marry women who are less intelligent than they are.

So we also would have:

$$I_{wife} = \rho * I_{husband} \tag{4.7}$$

Now the eqs. [4.6 and 4.8] cant be true both at the same time. And they are both missing something: Randomness. We have neglected the stochastic nature of the problem in eqs. [4.6 and 4.8]. We can first of all note a few more observations:

Highly stupid women tend to marry men who are more intelligent than they are.

Highly stupid men tend to marry women who are more intelligent than they are.

The intelligence of husband and wife has a positive correlation .

Lets assume the correlation is 0.7. This is illustrated in Fig. 4.9. Lets assume $X1 = I_{wife}$ and $X2 = I_{husband}$. We can take the objective point of view and say:

$$I_{wife} + \xi_w = I_{husband} + \xi_h \tag{4.8}$$

So Husband and wife have (in average) the same intelligence, but some additional noise (randomness) spreads out the one-to-one relationship (see Fig. 4.9). Now if you take the point of view for a highly intelligent woman:

$$I_{husband} = \rho * I_{wife} + \xi_{hw} \tag{4.9}$$

Now the intelligence of the husband is (in average) smaller than that of the wife due to the noise (see Fig. 4.9 upper right). It is 0.7 in mean. This is because we cut through the distribution

along a diagonal relative to the main axis of the distribution (one-to-one relation). This effect can be understand by the fact we start out from assuming a very intelligent woman, which is already unusual. Finding an equally intelligent man is challenged by the randomness, which simply makes it much more likely to be closer to the mean. This effect is called **Regression to the Mean**.
So the important conclusion that we have to take here is that: Correlations is not just a deterministic linear model, but it also models the mean effect of randomness. So when the correlation between A and B is 0.7, then that means that A and B may have a linear relation larger than 0.7, assuming you could get rid of the stochastic noise.
reducing the noise ... correlation increases ...spread reduced.
Lets look at two more examples to better understand this:

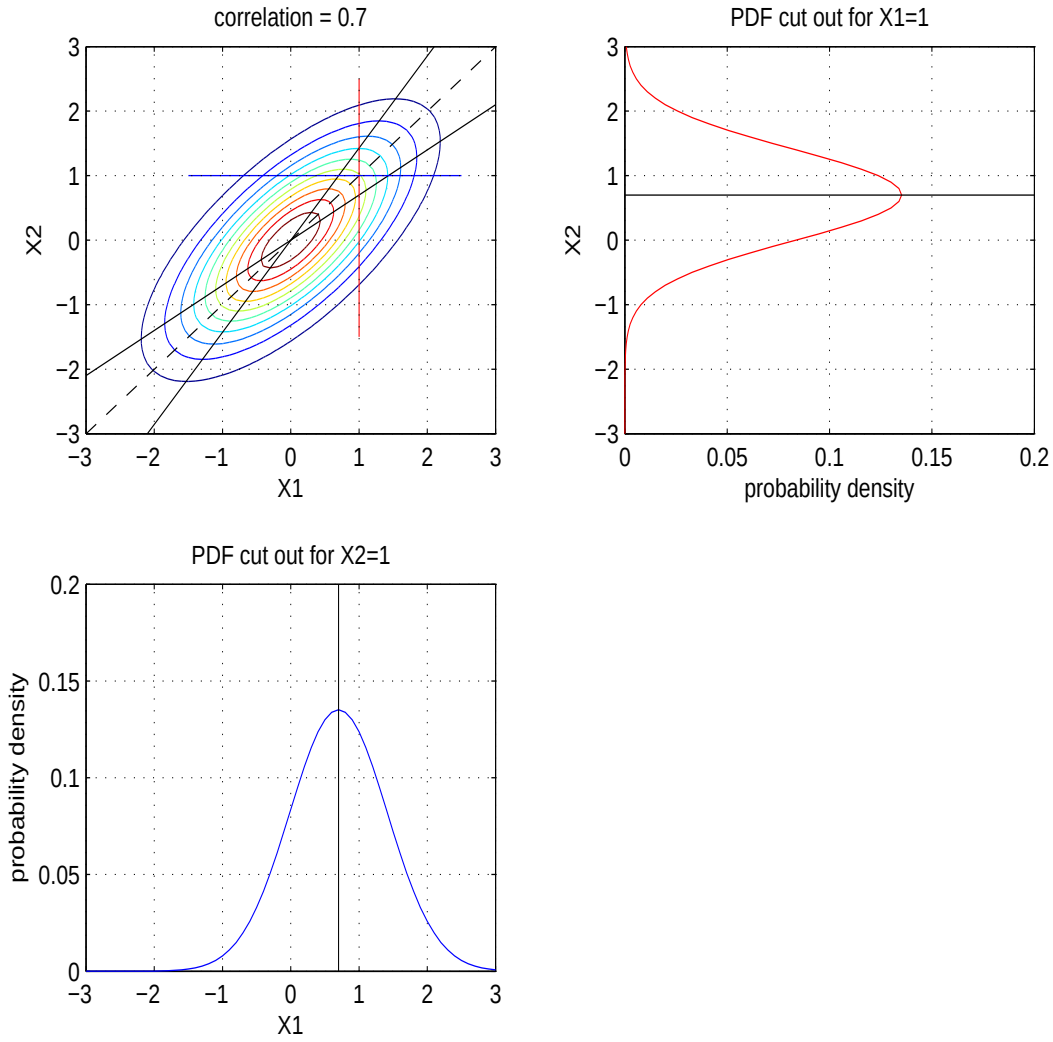


Figure 4.9: (a) Two dimensional normal distribution (pdf) for a correlation of 0.7. The dashed diagonal line marks the linear one-to-one relation (X2=X1); the two solid black lines mark the linear models $X_2 = 0.7X_1$ and $X_1 = 0.7X_2$, respectively. (b) cutout of the 2-dimensional pdf in (a) along the red line ($X_1 = 1.0$) in (a) .(c) cutout along the blue line ($X_2 = 1.0$) in (a). The solid black lines in (a) and (b) mark the means of the pdfs.

4.1.2 The Uncentered (spatial) Correlation

If consider to evaluate the similarity in spatial patterns, an uncentered correlation may often be used. Consider the following example: We try to estimate the global warming pattern or the teleconnections of ENSO. If we have two warming or teleconnection patterns we may like to know how similar the two patterns are. For this we would compute the correlation between the two patterns. Using the normal definition of the correlation (eq. [4.5]) would require to subtract the mean of the pattern first. However, the mean warming or teleconnection may actually be the main part of the pattern. The normal definition of the correlation (eq. [4.5]) would only evaluate the similarity in the pattern that are different from the overal mean.
An uncentered correlation may be more useful for defining the spatial correlation. In the uncentered correlation the mean of the two random variables X and Y is not subtracted. Thus eq. [4.5] is not evaluated based on anomalies, but on the whole values of X and Y. Whether such an uncentered spatial correlation is more meaningful depends on the problem addressed.

Chapter 5

Estimation of Statistical Parameters

We assume in this section that the result of a sampling process can be represented by a sample of n independent and identically distributed (iid) random variables $\{X_1, X_2, ..., X_n\}$. In general, we use X to represent any of iid random variables in the sample and assume that the common probability density function of X is $f_X(\cdot)$. Furthermore, we do not yet assume a specific form for f_X.

5.1 Discrete Conditional Samples of Continuous Random Variables

Continuous Random Variables are usually sampled in discrete intervals.
Note that the above estimate of the *pdf* assumes that the samples are taken randomly or in fixed intervals. The decision to take a sample or not, must not depend on the state of the system itself, which is for observation often not the case.

Examples:

- **Ship-measurements**: Assume we want to determine the mean state and variance of the wind field over the North Atlantic based on Ship-measurements. Ship-measurements often depend on the weather. Sometimes the measurements are not carried out if the weather is bad or the measurements are taken at a different location (ship tracks are depending on the weather). The resulting *pdf* or parameters of the *pdf* are biased or conditioned by the weather. It is likely, for instance, that the wind speed will be under estimated and the *pdf* of temperatures may be shifted towards warmer in high-latitudes.

- **Satellite-measurements**: The same as in the previous example holds for Satellite-measurements. Satellites can not always measure due to weather conditions (e.g. Clouds, day light, no day ligth).

- **Post-processing of data**: Often data are post-processed to eliminate false measurements. This is often done by some simple statistical considerations. Large deviations from the mean are, for instance, often considered to false measurements. This of cause will reduce the probability to find extreme events in post-process observed data and can change the shape of the *pdf*.

- **The broken cloud effect** (BCE): One may want to know under which cloudiness the broken cloud effect (more than 100% incoming sun light due to additional cloud reflections) is strongest. For this one may plot the *pdf* of the BCE as a function of cloudiness. However,

this histogram does not say if the effect is related to cloudiness, because the samples may be taken under specific cloudy conditions. If we have only measured a certain type of cloudiness, than we will have most BCE for this cloudiness. We have to fold the *pdf* with the *pdf* of the cloudiness itself. A small probability of BCE may be due to only few measurements for this cloudiness. See also section 17.5.2.

5.2 Histograms: An Estimator for the Probability Density Function

The frequency histogram is a crude estimator for the probability density function, f_X of X. To obtain a frequency histogram the phase space (e.g. R) is divided into K subsets Θ_k such that

$$\bigcup_{k=1}^{K} \Theta_k = R \quad and \tag{5.1}$$

$$\Theta_k \cap \Theta_j = \emptyset \quad for\ k \neq j \tag{5.2}$$

The number of observation that fall into each Θ_k is counted and divided be the total number of observations so we obtain

$$\mathbf{H}(\Theta_k) = \frac{|\{\mathbf{X_k} : \mathbf{X_k} \in \Theta_k\}|}{n} \tag{5.3}$$

where $|\mathcal{S}|$ denotes the number of elements in set $\mathcal{S}$. $\mathcal{H}(\Theta_k)$ is an estimator of

$$P(\mathbf{X_k} \in \Theta_k) = \int_{\Theta_k} f_X(x)dx \tag{5.4}$$

Consequently, the random step function

$$\widehat{f_X}(x) = \frac{\mathbf{H}(\Theta_k)}{\int_{\Theta_k} dx} \quad if\ x\ \in \Theta_k \tag{5.5}$$

is a crude estimator of the true density function, with $\int_{\Theta_k} dx$ being the length of the interval $\mathbf{H}(\Theta_k)$. The empirical cumulative distribution function can be estimated by

$$\widehat{F_X}(x) = \mathbf{H}([-\infty, x]) \tag{5.6}$$

Note that empirical estimations of the higher order moments or the tails of the *pdf* usually require many observations, while the mean and variance (unless time scale dependent) are often well estimated with small numbers of observations (e.g. 10). Thus the analysis of extreme values requires large data sets.

5.3 Estimating the Mean

The best estimate of the mean μ is

$$\widehat{\mu} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X_k} \tag{5.7}$$

It may not be straight forward to see that this is the best estimate of eq. [3.3], that is

$$\mu = \int_{-\infty}^{\infty} x f_X(x) dx, \tag{5.8}$$

but it may easier to comprehend if we sort the sum in eq. [5.7] into a histogram, which is an estimate of $f_X(x)$.

5.4 Estimating the Central Moments

In general the moments based on integrals of a continuous functions $f_X(\cdot)$ transform into sums of discrete samples as given by

$$\widehat{\int_{\Omega} g(x) f_X(x) dx} = \frac{1}{n} \sum_{k=1}^{n} g(X_k) \tag{5.9}$$

However, if the central moments are estimated relative to the mean, $\widehat{\mu}$, which is estimated by the same data, then the estimate changes slightly. The best estimate of the variance is

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \widehat{\mu})^2 \tag{5.10}$$

Note that he sum is divided by $n-1$ instead of n, as for the mean. This reduced degree of freedom in the estimate of the variance is related to the fact that the same samples are used to estimate the mean value (note that the mean value in eq. 5.10] is also an estimate. Consider, for instance, that the exact mean is known a priori: It is easy to see that the sum in eq. [5.10] would be larger compared to the one that uses the estimated mean, $\widehat{\mu}_i$. The version with the a priori known mean is often used in forecast skills and is called the "root mean squared error" (RMS-error or RMSE):

$$\varepsilon_{RMS} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (X_k - \mu)^2} \tag{5.11}$$

5.5 Estimating the Covariance and Correlation

The estimate of the covariance is slightly different from the estimate of the mean. The covariance is given by

$$\widehat{\sigma_{ij}^2} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{k;i} - \widehat{\mu}_i)(X_{k;j} - \widehat{\mu}_j) \tag{5.12}$$

and the related estimate of the correlation is

$$\widehat{\rho}_{ij} = \frac{\widehat{Cov}(X_i, X_j)}{\sqrt{\widehat{Var}(X_i)\widehat{Var}(X_j)}} \tag{5.13}$$

5.6 The Rank Transformation (Spearman Rank Correlation)

In the discussion of the parameters of a *pdf* we have seen that the variance is only a good measure for the scale of the distribution, if the *pdf* is nearly normal distributed. The same is true for the covariance and correlation. A non-parametric approach based on ranks can be used when the observations are thought not to be normal.

The sample $\{(\mathbf{X}_i, \mathbf{Y}_i) : \mathbf{i} = \mathbf{1}, \ldots, \mathbf{n}\}$ is replaced by the corresponding sample of ranks $\{(\mathbf{R}_{\mathbf{X}_i}, \mathbf{R}_{\mathbf{Y}_i}) : \mathbf{i} = \mathbf{1}, \ldots, \mathbf{n}\}$, where $\mathbf{R}_{\mathbf{X}_i}$ is the rank of $\mathbf{X}_i$ amongst the $\mathbf{X}_\mathrm{S}$ and $\mathbf{R}_{\mathbf{Y}_i}$ is defined similarly.
An Example: $\mathbf{X} = 1, 9, 5, 8, 3, 7$ und $\mathbf{Y} = 5, 9, 2, 8, 78, 0 \Rightarrow R(\mathbf{X}) = 1, 6, 3, 5, 2, 4$ und $(\mathbf{Y}) = 3, 5, 2, 4, 6, 1$.
The dependence between $\mathbf{X}$ and $\mathbf{Y}$ is then estimated with the *Spearman rank correlation coefficient* $\hat{\rho}^S_{XY}$

$$\hat{\rho}^S_{XY} = \frac{\sum_{i=1}^{n} \mathbf{R}_{X_i}\mathbf{R}_{Y_i} - \mathbf{N}}{\sqrt{(\sum_{i=1}^{n} \mathbf{R}^{\mathbf{2}}_{X_i} - \mathbf{N})(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{R}^{\mathbf{2}}_{Y_i} - \mathbf{N})}} \tag{5.14}$$

where $N = n(\frac{n+1}{2})^2$. This is just the ordinary sample correlation coefficient[1] in eq.[5.12] of the ranks. Thus the Spearman ranks are a transformation of a variable that is non-normal distributed into a variable that is uniformly distributed. Note that the interpretation of this correlation coefficient may still be problematic, because a non-normal *pdf* indicates some non-linear behavior of the variable which can probably not be compared with another variable by a measure of linear relation.

5.7 Sample Vectors

Much of the algebra for the series of samples of continuous random variables can be interpreted better if we write the series of samples as a sample vector:

$$\hat{\mathbf{X}} = \{x_1, x_2, ..., x_n\} = \vec{X} \tag{5.15}$$

Each sample is independent of the other. Therefore each can be interpreted as a component of n-dimensional vector. The variance can then be written as

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \widehat{\mu})^2 = \frac{1}{n-1} < \vec{X} - \widehat{\mu}|\vec{X}^T - \widehat{\mu} > \tag{5.16}$$

or with $\widehat{\mu} = 0$

$$\widehat{\sigma^2} = \frac{1}{n-1} < \vec{X}|\vec{X}^T > \tag{5.17}$$

The covariance:

$$\widehat{\sigma^2_{xy}} = \frac{1}{n-1} < \vec{X}|\vec{Y}^T > \tag{5.18}$$

The correlation:

$$\widehat{\rho}_{xy} = \frac{< \vec{X}|\vec{Y}^T >}{\sqrt{< \vec{X}|\vec{X}^T >< \vec{Y}|\vec{Y}^T >}} \tag{5.19}$$

[1] Also known as *Pearson's r*

Part II

Time Series Analysis

Unlike in other research areas, where we could do experiments and study the results of many realizations, in climate research we have only one world (one realization) and therefore have to relay on the analysis of the time series of events.
In time series analysis we are interested in the characteristics of the time evolution of a physical quantity, which is a central aspect of climate variability analysis. By studying the characteristics of the time series of a physical quantity we learn something about the dynamical system that drives the physical quantity.

Chapter 6

Basic Definitions and Examples

The time series analysis is often central to statistical analysis in climate research. The following examples illustrate some of the methods discussed in the following subsections.

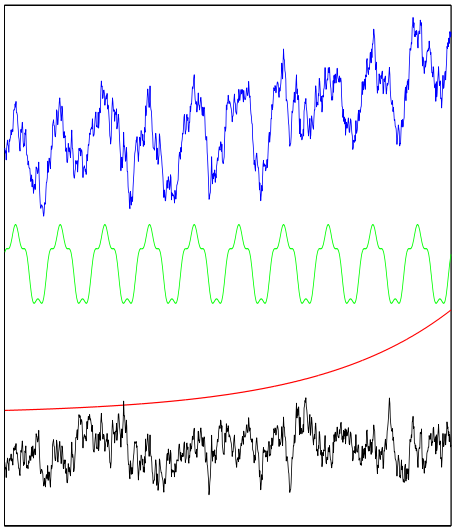


Figure 6.1: A time series (blue line) deconstructed into a desterministic cycle (green line), a desterministic non-linear trend (red line) and into the residual red-noise variability (black line).

In general a time series of a climate variable will be a combination of a deterministic part, like a cycle (e.g. seasonal, diurnal) or trend, and a stochastic part. Fig. 6.2 illustrates the decomposition of a time series. In the following sections we will focus on the stochastic part, assuming that the deterministic cycle or trend has been removed prior to the analysis of the time series.
Fig. 6.2 shows 50 years long time series of monthly mean temperatures at different locations. We can see that the time scale in which the temperature values change are different in the three time series. In the atmosphere we find rapid changes will in the SST we find both rapid and long time changes, and in the ocean the time evolution of the variability is the slowest. The following sections will define parameters, such as the auto-correlation, decorrelation time and spectra, which will help to quantify these differences in the time scale behavior.
In Fig. 6.3 we see the El Niño SST time series and an estimate of the spectra together with a theoretical spectra. The theoretical spectra is that of an auto regressive process of the first order, where the parameters of the process has been fitted to those of the El Niño SST time series. It is for the understanding of climate variability of interested to know what kind of stochastic process is driving the variable presented in the time series. In the following section we will give some

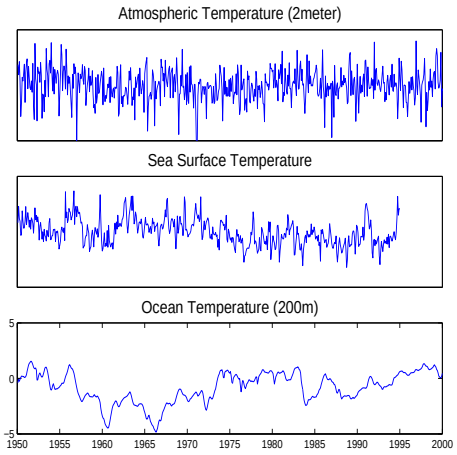


Figure 6.2: 50 years long time series of monthly mean temperatures of the atmosphere in 2meter height at location in central northern Asia (upper), of the sea surface in the North Atlantic (middle) and in 180meters depth in the North Atlantic Ocean (lower).

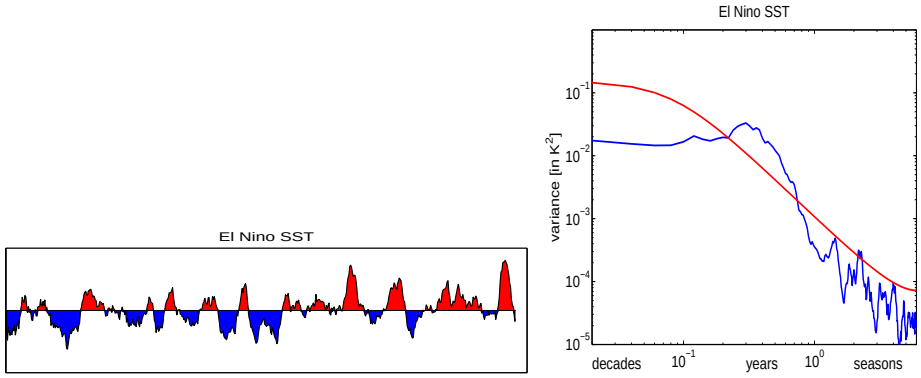


Figure 6.3: El Niño SST time series (left) and an estimate of the spectra together with a theoretical spectra (rigth). The theoretical spectra is that of an auto regressive process of the first order, where the parameters of the process has been fitted to those of the El Niño SST time series.

discussion of stochastic processes in general and of auto regressive process in particular. We will also discuss how these processes are compared with time series.
Often we want to compare the variability of two time seriess with each other, in doing so we are interested in the time scales on which the two time series correlate well with each other, see Fig. 6.4 for instance. We will discuss the cross correlation function and the cross-spectrum of two time series.

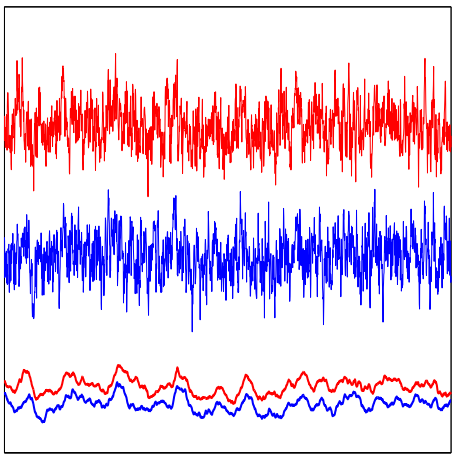


Figure 6.4: Two time series (upper lines) that are uncorrrelated, but are correlated on longer time scales (lower lines).

6.1 Stationary Processes

A stochastic process $\mathbf{X_t} : t \in Z$ is said to be stationary if all stochastic properties (e.g. mean, variance, correlation) are independent of the index t, which can be an index of time or a spatial dimension.
It follows that if $\mathbf{X_t}$ is stationary then:

1. $\mathbf{X_t}$ has the same distribution function F for all t, and
2. for all t and s the parameters of the joint distribution function of $\mathbf{X_t}$ and $\mathbf{X_s}$ depend only on $|t - s|$.

Most climate variables are not stationary in this restrict definition, mostly due to externally forced deterministic cycles such as the diurnal or annual; e.g. the mean temperature in Kiel at midnight in January is different from the mean at noon in August.
However, many climate variables can be considered as cyclo-stationary processes. If a stochastic process $\mathbf{X_t}$ is cyclo-stationary then:

1. $\mathcal{E}(\mathbf{X_t}) = \mu_{t|m}$. The mean is a function of the time within the external cycle, where $t|m$ refers to the phase in the external cycle
2. $\forall t, s$ is $f_{X_t X_s} = f_{X_t X_s}(|t - s|, t|m)$. For all t and s the parameters of the joint distribution function, $f_{X_t X_s}$ of $\mathbf{X_t}$ and $\mathbf{X_s}$ depend only on $|t - s|$ and the phase $t|m$ in the external cycle.

Unfortunately, even this less restricted version of stationarity often does not apply to climate variables due to long term trends in the boundary conditions (e.g. CO2 increase). However, we usually assume that these trends are negligible or we remove an estimate of the trend and analysis the residual under the assumption of cyclo-stationarity.

6.2 Ergodicity

Definition: Physical systems are ergodic, if the estimate of the statistical parameters based on averages in time,

$$\widehat{\mu_t} = \frac{1}{n} \sum_{t=1}^{n} g(X_t)$$

are identical to the estimate of the statistical parameters based on ensemble means,

$$\widehat{\mu_i} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

Thus $\widehat{\mu_t} = \widehat{\mu_i}$.

Which is usually the case in climate research. We can observe time series of climate variables, but we cannot observe ensembles of the climate system. We therefore study time series to estimate the statistical parameters.

Chapter 7

Stochastic Climate Models

A physical system is usually described by the dynamical equation that govers the time evolution of the system:

$$\frac{dX(t)}{dt} = A(X(t), Y_i(t)) \tag{7.1}$$

Here $X(t)$ is a function of many other physical variables Y_i, which themselves depend in time and from each other. In princple the time evolution is given by the initial state of all variables and by the formulation of eq. [7.1].
In practice, however this is an impossible endeavor, due to three significant limitations:

i.) The initial state, $X(t_0), Y_i(t_0)$, is not exactly known.
ii.) The time derivative, $\frac{d}{dt}$, is not exactly known.
iii.) The operator, $A(\cdot)$, is not exactly known.

The problem with points i.) and ii.) is nicely demonstrated by the Lorenz model. The Lorenz model is a low dimensional problem of convection cells, with only three variables interacting. The time evolution of the system can not be predicted beyond a certain time period if the initial state has a finite error. Further the time derivative has to be estimated by discrete time steps, which contributes to the uncertainty in predicting the time evolution.
The climate system is a much more complicated thermodynamical system with a very high dimensional state space. In practice the dynamical system of the climate cannot be simulated by pure dynamical equations, although general circulation models are trying to do so.
A concept that helps to understand the most important dynamics of the system and therefore the main statistical characteristics of the system is the use of stochastic climate models. In general we may assume that any climate variable X is a result of a stochastic process:

$$\frac{dX(t)}{dt} = f_m(t) * A(X(t), Y_i(t)) + f_a(t) \tag{7.2}$$

$A(\cdot)$ = slow dynamics relative to Δt
f_m, F_a = fast dynamics relative to Δt
f_m = multiplicative noise (mulitplikatives Rauschen)
f_a = additive noise (Additives Rauschen)

Here $A(X, Y_i)$ denotes any kind of function depending on the climate variable X, other climate variables Y_i and some fast changing variable f_m = multiplicative noise and f_a =additive noise. The basic concept of a stochastic climate model is the introduction of the fast changing variable f, which represents physical process that happen so fast and may be on smaller spatial scales that

the climate model can not resolve the physical processes for f. Thus f represents some white noise for the climate system which is independent of the system. The most important effect of the noise f is that it can generate variability in X that are much larger than the variability of f. Prior to the introduction of the stochastic climate models by Klaus Hasselmann (1976) the general believe was that low-frequency variability in X must be introduced by some kind of external forcing, such as fluctuation of the solar radiation or vulcanos.

7.1 Example: Slab ocean model

A simple example of the stochastic climate model is the heat balance of the oceans surface layer (the mixed layer). If we only consider the atmospheric heat flux as source of heat for the ocean mixed layer, than the sea surface temperature (SST) follows the equation:

$$\frac{dSST}{dt} = \frac{1}{c_p h} F_{atmos}, \quad cp = 4 * 10^6 J/Km^3 \tag{7.3}$$

The depth of the mixed layer h is in average about 50 meter (therefore 'slab'). The atmospheric heat flux is in this model given by

$$F_{atmos} = c_A(T_{atmos} - SST), \quad c_A = \frac{40W}{m^2K} \tag{7.4}$$

Thus,

$$\frac{dSST}{dt} = -cSST + cT_{atmos} \tag{7.5}$$

The SST is an AR(1)-process if the T_{atmos} is white noise and the constant c is indeed constant.

7.2 The Probability Distribution Function of some Stochastic Processes

In the following we will discuss the probability distribution function of some simple stochastic processes to illustrate that the parameters of the *pdf* are a reflection of the physical processes driving the system.

A linear damping model / A normal *pdf*: One of the simplest stochastic processes is linear damping, so that

$$\frac{dX}{dt} = -c * X + f \tag{7.6}$$

with c a positive contant and f a random noise forcing. The *pdf* of this process is a normal distribution, with σ only depending on the strength of c and f, see Fig. 7.1.

A asymetric feedback model / A skewed *pdf*: A physical process that leads to a skewed distribution is a process that responses asymmetrically (non-linear) to deviations from its equilibrium state. If, for instance, the damping is stronger for deviations to the lower side of the equilibrium state than for deviations to the larger side, than the *pdf* will be positively skewed and deviations to larger values are more likely than deviations to smaller. Such as,

$$\frac{dX}{dt} = c_0 - c_1 X - c_2 X^2 + f \tag{7.7}$$

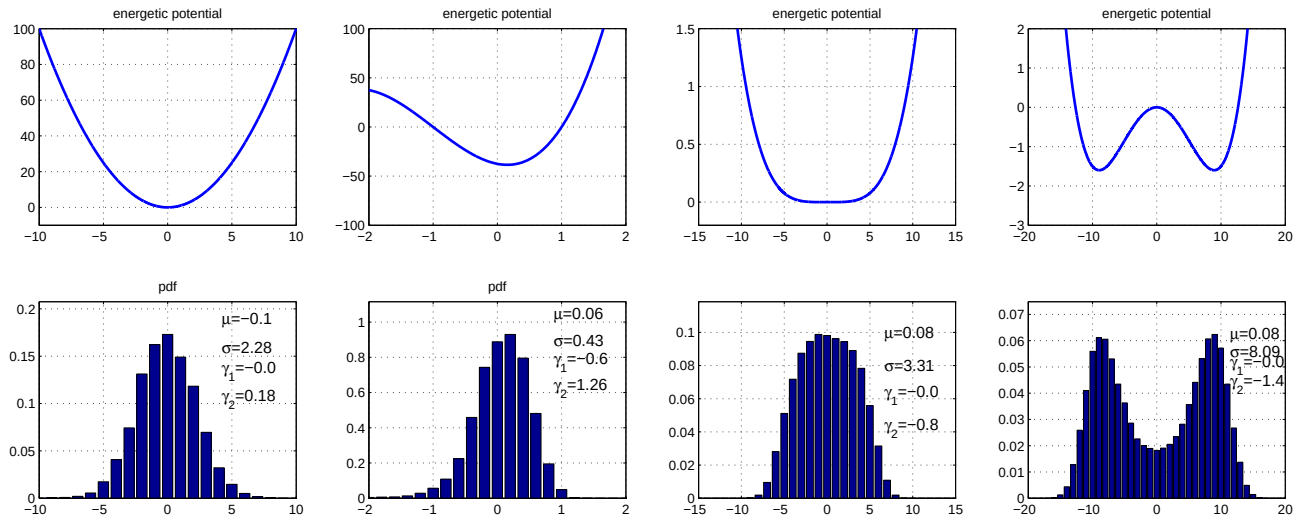


Figure 7.1: The potential functions(upper row) and *pdf*(lower) of the processes discussed in section 7.2.

for $c_i > 0$. A simple way to illustrate the skewness of this physical process is to evaluate the potential function of the forces that drive the physical variable. If the potential function is asymmetric, than the *pdf* of this process will be skewed towards the less stepper slope. An examples is shown in Fig. 3.1.
A well known example is the El Niño index of the east equatorial SST in the tropical Pacific, which is skewed to positive values. Thus extreme El Niño events (positive deviations) are more likely as equally strong La Niña events (negative deviations), see Fig. 3.3.

A non-linear damping model / A *pdf* with kurtosis: A stochastic process with a non-linear but symmetrical damping is an example of a *pdf* with kurtosis. Such as

$$\frac{dX}{dt} = -cX^3 + f \tag{7.8}$$

Here the damping is very small for small deviations, but it increases much faster for large deviations from the equilibrium than a linear damping the resulting *pdf* has negative kurtosis. The system is therefore nearly free to evolve within a certain interval, but fell is strong boundary at the limits of this interval. See the potential function in Fig. 3.1.

A non-linear model / A bimodal *pdf*: A stochastic process that has two equilibria is an example of a bimodal process. Such as

$$\frac{dX}{dt} = c_1X - c_3X^3 + f \tag{7.9}$$

However, bimodality can result from many different processes. In the climate system the deterministic cycles (daily,annual) are the main source for bimodality, because the *pdf* of a sine function is bimodal.

Note that neither the mean state nor the median of the *pdf* of a stochastic processes must fall together with the equilibrium of the unforced system.

7.3 Autoregressive (Markov) Processes

The dynamics of many physical processes can be approximated by first-, second or sometimes higher-order ordinary linear differential equations, for example,

$$a_2\frac{d^2x(t)}{dt^2} + a_1\frac{dx(t)}{dt} + a_0x(t) = z(t)$$

where z is some external (independent of x) forcing function. Time discretization yields

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \xi_t \tag{7.10}$$

where

$$\alpha_1 = \frac{a_1 + 2a_2}{a_0 + a_1 + a_2}$$

$$\alpha_2 = \frac{-a_2}{a_0 + a_1 + a_2}$$

$$\xi_t = \frac{1}{a_0 + a_1 + a_2} z_t$$

If z_t is a white noise process (an uncorrelated time series of normally distributed random values), then eq. [7.10] defines an auto-regressive process of the second order or AR(2) process.
An auto-regressive process pf order p, or an AR(p) process, is generally defined as follows:
$\mathbf{X_t} : t \in Z$ is an auto-regressive process p of order p if there exist real constants $\alpha_k, k = 0, 1, ..., p$, with $\alpha_k \neq 0$ and a white noise process $\mathbf{Z_t} : t \in Z$ such that

$$\mathbf{X_t} = \alpha_0 + \sum_{k=1}^{p} \alpha_k \mathbf{X_{t-k}} + \mathbf{Z_t} \tag{7.11}$$

AR(p) Processes are also called Markov Processes. Note that $\mathbf{X_t}$ is dependent of the present and all pasted $\mathbf{Z_t}$, while it is independent of $\mathbf{Z_t}$ in the future. The AR(p) processes are part of a larger class of auto-regressive and moving-average (ARMA) processes. The moving-average processes are, however, of little relevance in climate research.
The above description of an auto-regressive process is based on a discrete time series. We can also formulate the auto-regressive processes as a continuous differential equation:

$$a_0x(t) + \sum_{k=1}^{p} a_k \frac{d}{dt^k} x(t) = z(t) \tag{7.12}$$

Note that the relation between the a_k of the continuous differential equation and the α_k of the discrete eq. [7.11] is somewhat complex for higher order AR(p) processes.

7.3.1 Variance of AR(p) Processes

We will in general assume that the mean $\mu = 0$ and therefore $\alpha_0 = 0$, thus discuss only anomalies. The variance is given by

$$Var(\mathbf{X_t}) = \frac{Var(\mathbf{Z_t})}{1 - \sum_{k=1}^{p} \alpha_k \rho_k} \tag{7.13}$$

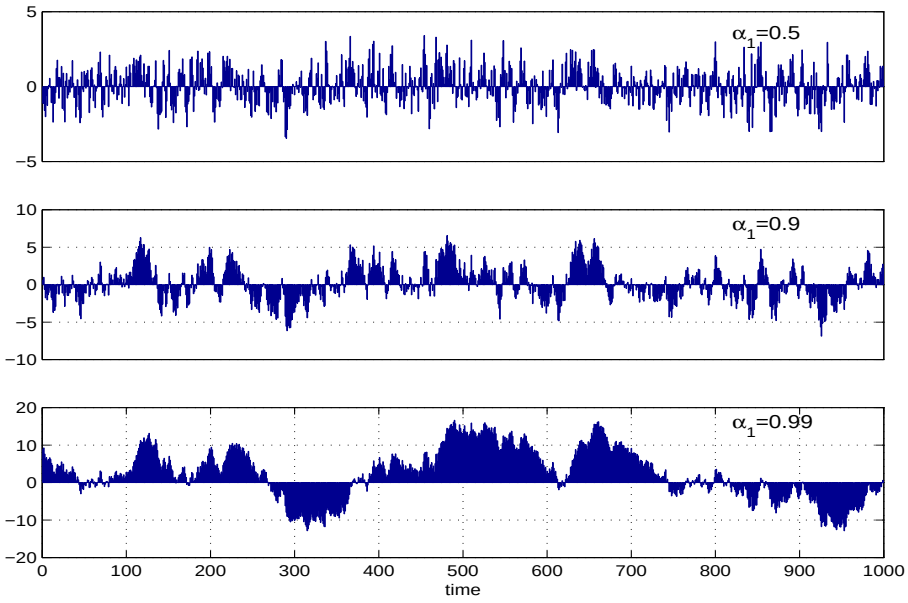


Figure 7.2: Different realization of AR(1) processes with different $\alpha = 0.5, 0.9, 0.99$ but with identical unit variance normal white noise processes $\mathbf{Z_t}$. Note the different y-axis scaling.

where

$$\rho_k = \frac{Covar(\mathbf{X_{t-k}X_t})}{Var(\mathbf{X_t})} \tag{7.14}$$

is the auto-correlation function of $\mathbf{X_t}$.

7.3.2 Examples of AR(1) Processes

A discrete AR(1)-process follows the equation:

$$\mathbf{X_t} = \alpha_1 \mathbf{X_{t-1}} + \mathbf{Z_t} \tag{7.15}$$

Fig. 7.2 illustrates some AR(1) Processes with $\alpha_1 \in]0, 1[$. Note that α_1 can take all real values, but only $\alpha_1 \in [0, 1[$ are of physical relevance. For $\alpha_1 < 0$ the time series flips sign in every time step, while for $|\alpha_1| \geq 1$ the process becomes non-stationary.
Note that the different AR(1) Processes are integrated with identical white noise processes, we can see that all AR(1) Processes exhibit the same fluctuation, but that the AR(1) Process acts as a low-frequency amplifier of the white noise time series. Following eq.[7.13] we find that the variance is given by

$$Var(\mathbf{X_t}) = \frac{\sigma_z^2}{1 - \alpha_1^2} \tag{7.16}$$

which uses the fact that the lag(1) auto-correlation $\rho_1 = \alpha_1$. See Fig.7.2 for increase in variance with increase in α_1. The auto correlation of an AR(1) process is decreasing exponentially with $\rho_{t-k} = \alpha_1^k$, but is never zero (theoretically). Note, that in this discrete AR(1)-process the variance of $\mathbf{X_t}$ is larger than that of $\mathbf{Z_t}$, which is somewhat misleading if we think of continuous physical processes.
Physical model are usually written as differential equations. The differential equation for an AR(1) process is

$$\frac{dx(t)}{dt} = a_1 x(t) + z(t) \tag{7.17}$$

with

$$a_1 = \frac{\alpha_1 - 1}{\alpha_1} \tag{7.18}$$

The parameter a_1 is the damping of x. The physical interpretation of an AR(1) process with $\alpha_1 \in]0, 1[$ is a damped system that responses to every disturbance from its mean value by a linear negative feedback, acting to rebuild the equilibrium state, where the damping is proportional to the amplitude of the disturbance. Note that in this notation the physical units of $x(t)$ and $z(t)$ are different.
We can rewritten this equation to find a form in which the amplitudes (variances) of the forcing, $z(t)$, and $x(t)$ can be directly be compared. A linear damped system may be presented with a newtonian damping:

$$c\frac{dx(t)}{dt} = \gamma(z(t) - x(t)) = -\gamma x(t) + \gamma z(t) \tag{7.19}$$

$c =$ the inertia of the $x(t)$ (e.g. for temperature it is the heat capacity)
$\gamma =$ a damping.

In this formulation we can compare the variance of the forcing $z(t)$ with $x(t)$.
to be continued ...

The AR(1)-process $\frac{dX}{dt} = cX + f$ is the simplest case of a stochastic climate model, which is a good approximation for many physical processes as, for instance, the glacier grows, lake or river water levels, grows of plants or the heat balance of the upper ocean. It is therefore often chosen as the null hypothesis for climate variability. The most important effect of the noise f is that it can generate variability in X on longer time scales.

7.3.3 Examples of AR(2) Processes

A discrete AR(2)-process follows the equation:

$$\mathbf{X_t} = \alpha_1 \mathbf{X_{t-1}} + \alpha_2 \mathbf{X_{t-2}} + \mathbf{Z_t} \tag{7.20}$$

The AR(2) Process has a second additional parameter, which can, but does not have to, be an indication of oscillating behavior. Note that not all combinations of α_1, α_2 are stationary only if

$$\alpha_2 + \alpha_1 < 1$$

$$\alpha_2 - \alpha_1 < 1$$

$$|\alpha_2| < 1$$

and not all are AR(2) Processes, only if

$$\alpha_2 \neq \alpha_1^2$$

and AR(2) Process can only show oscillating behavior if $\alpha_2 < 0$, but do not necessarily show oscillating behavior, see Storch and Zwiers 10.3.5 for details.
Fig. 7.3 illustrates some AR(2)-Processes with different values for α_1 and α_2. We can clearly see that the three examples have very different characteristics in the time scale behavior. The first

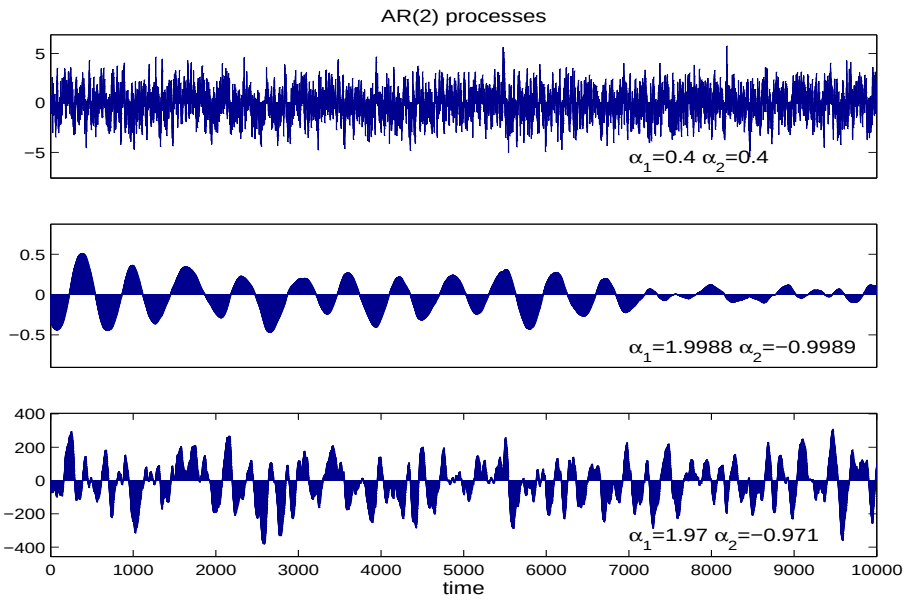


Figure 7.3: Different realization of AR(2) processes with different α_1 and α_2, but with identical unit variance normal white noise processes $\mathbf{Z_t}$. Note the different y-axis scaling.

example with $\alpha_1 = 0.4$ and $\alpha_2 = 0.4$ is dominated by large random high-frequency variability, but has also a low-frequency part of variability. The second example with $\alpha_1 = 1.9988$ and $\alpha_2 = -0.9989$ has a relative regular oscillation on longer time scales, whose amplitude is variable of time. The third is example with $\alpha_1 = 1.97$ and $\alpha_2 = -0.971$ is similar to the second, but the oscillation is on shorter time scales and less dominant.
The differential equation for an AR(2) process is:

$$a_0 x(t) + a_1 \frac{dx(t)}{dt} + a_2 \frac{dx(t)}{dt^2} = z(t) \tag{7.21}$$

Note that fluctuations of $x(t)$ can grow in the absence of any driving forcing $z(t)$, due to the term $a_2 \frac{dx(t)}{dt^2}$. The physical interpretation of an AR(2)-process is a damped oscillation system, which can, unlike the AR(1)-process, have a preferred time scale at which the system oscillates if driven with noise $z(t)$.
An example of a damped oscillation in climate is the El Niño / Southern Oscillation, which was shown by Jin (1997) and Burgers et al. (2005) to behavior much like a damped oscillation.

Chapter 8

The Auto-Covariance Function

The autocovariance or correlation function is the statistical parameter that describes the time scale behavior of a time series. It presents the covariance or correlation of the time series relative to the time series in a lag or lead of a time interval. The Fourier transform of the auto-covariance is the spectra of the time series, which is an alternative statistical parameter (presentation) of the time series which is more appropriate if one is interested in the variance as a function of frequencies or periods. The spectra will be discussed in the next section.

8.1 Estimating the Auto-correlation/-covariance Function

A non-parametric estimator of the auto-correlation function $\rho(\tau)$ is given by

$$\rho(\tau) = \gamma(\tau)/\gamma(0) \tag{8.1}$$

where $\gamma(\tau)$ is the sample auto-covariance function

$$\gamma(\tau) = \frac{1}{T} \sum_{t=|\tau|+1}^{T} \mathbf{X}'_{t-|\tau|} \mathbf{X}'_t. \tag{8.2}$$

The sample auto-covariance(correlation) function $\gamma(\tau)$ is simply the covariance(correlation) of $\mathbf{X}$ with itself at a time lag of τ. It is set to zero for $|\tau| \geq T$. As a repetition of the fundamentals section:

- γ ist symetrisch
- $\gamma(0) = Var(\mathbf{X})$
- $\gamma(\tau) <= \gamma(0)$
- $|\rho(\tau)| <= 1$
- $\rho(\tau) < 0 \Rightarrow$ oscillation.

A statistical interpretation: $\rho(\tau)$ illustrates the mean time evolution a unit 1 signal. Small absolute values indicate little (in average) relation of $\mathbf{X}$ to past/future value.

A physical interpretation: $\rho(\tau)$ illustrates how the dynamical system responses to a disturbance from the equilibrium.

The auto-correlation function of the time series can be interpreted as the persistence forecast skill.

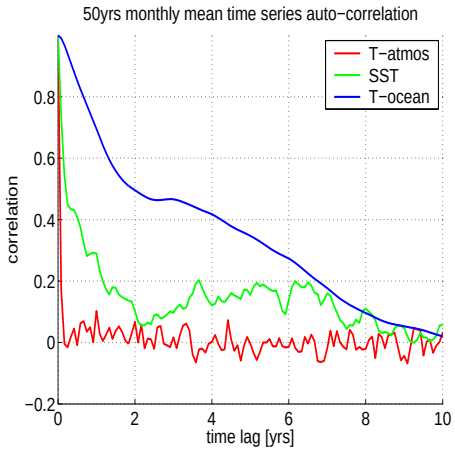


Figure 8.1: The auto correlation function of the monthly mean temperature time series as shown in Fig. 6.2. The atmosphere in 2meter height at location in central northern Asia (upper), of the sea surface in the North Atlantic (middle) and in 180meters depth in the North Atlantic Ocean (lower).

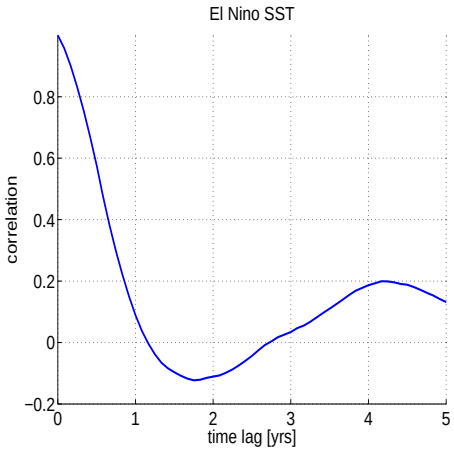


Figure 8.2: The auto correlation function of the observed time series of El Niño monthly mean SST, as shown in Fig. 6.3.

8.2 Examples of the auto-correlation function

Example 1: The auto correlation function of observed time series of monthly mean Temperatures at different locations are shown in Fig.8.1. We see that all auto-correlation functions decrease more or less monotonically to zero, without any significant crossing of the zero line, which indicates that no oscillations are present in all three time series. Further we see that the atmospheric temperature decreases very fast, within month, the SST decreases fast in the first year, but after the first year it decreases more slowly to zero. The ocean temp decreases relatively slowly to zero, in which it reaches near zero correlation after about 10 years.

Example 2: The observed time series of El Niño shows a significantly different auto correlation function, if compared to those of the first example, see Fig.8.2. Here find a significant oscillation of the auto correlation function around zero, indicating an oscillation of the El Niño time series with a period of about 4 years (indicated by the first minima by about 2years, half a period, and the second local maxima in the auto correlation function at about 4years the full period).

Example 3: The auto correlation function of Observed time series of 24hrs 500hpa together with the time series themselves are shown in Fig.8.3. We can see that auto correlation function of 500hpa in the higher latitudes (Northern Germany) decreases fast to zero, with near zero correlation after 10 days. In the tropics we find an initial fast decrease of the auto correlation function, but it remains on a level of about 0.3 after 20 days. It indicates that the variability in the tropics has some significant low-frequency variability. This is already visible in the time series where we can see that the 500hpa anomalies have one sign for over one year (positive over most of 1983, negative for most of 1984), a characteristic we can not find in the higher latitudes. This is related to the Ocean-Atmosphere interaction in the tropics that also cause the El Niño / Southern Oscillation mode.

Example 4: Forecast skill of Models. The forecast skills of models are often quantified by using the correlation with the observed values at different lead times as a skill value, see Fig.8.4. The simplest assumption for the time evolution of a climate variable is the auto correlation function,

which reflects the mean time evolution. It is often simplified to *damped persistence* (an AR(1)-process). A dynamical model should be able to 'beat' the damped persistence (auto-correlation) skill values. Thus the correlations of the models forecast with the observed values should be large for dynamical models than the auto-correlation.

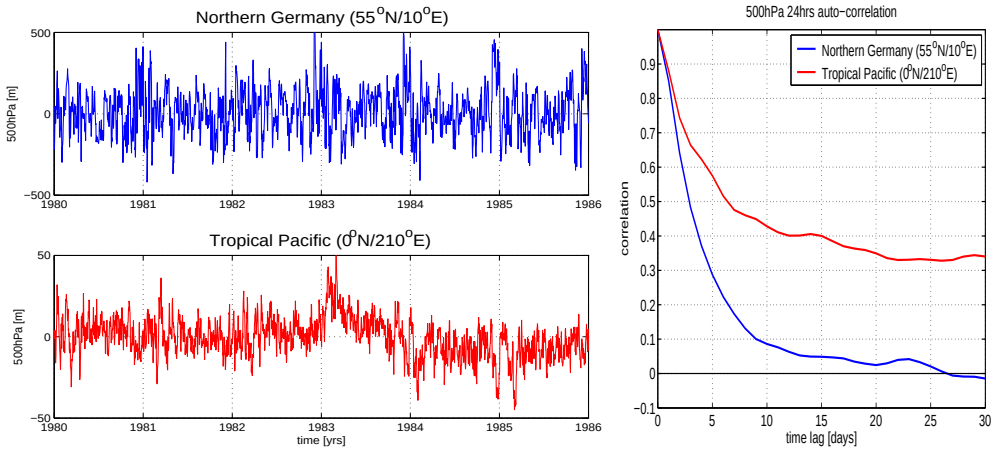


Figure 8.3: Left: Time series of 24hrs mean geopotential heights in Northern Germany (upper) and at the equatorial east Pacific. right: The auto-correlation functions corresponding to the time series in the left panels.

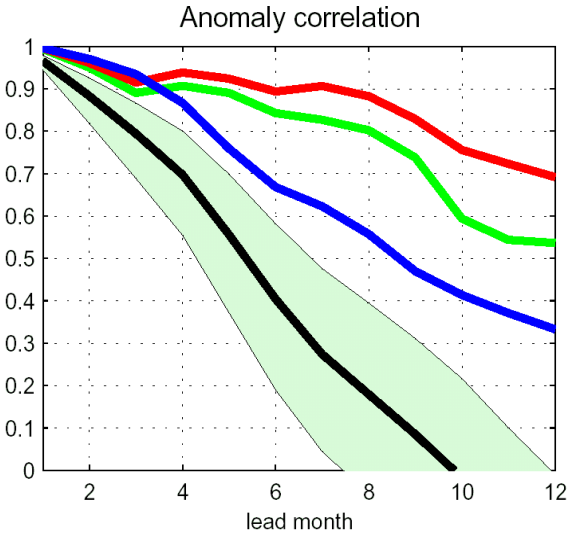


Figure 8.4: The black line is the auto-correlation function of EQ2-region of the equatorial Pacific SST. The colored lines are different model forecast correlations with the EQ2-region SST for different lead times. From Dommenget and Stammer (2004).

8.3 11.1.6 The Yule-Walker Equations for an AR(p) process.

If we multiply a zero mean AR(p) process $\mathbf{X}_t$, eq.[7.11], by $\mathbf{X}_{t-\tau}$, for $\tau = 1, \ldots, p$,

$$\mathbf{X}_t \mathbf{X}_{t-\tau} = \sum_{i=1}^{p} \alpha_i \mathbf{X}_{t-i} \mathbf{X}_{t-\tau} + \mathbf{Z}_t \mathbf{X}_{t-\tau}, \tag{8.3}$$

and take expectations, we obtain a system of equations

$$\vec{\gamma}_p = \Sigma_p \vec{\alpha}_p \tag{8.4}$$

that are known as the *Yule-Walker equations.* The equations relates the auto-covariances

$$\vec{\gamma}_p = \Big(\gamma(1), \gamma(2), \ldots, \gamma(p)\Big)^T$$

at lags $\tau = 1, \ldots, p$ to the process parameters

$$\vec{\alpha}_p = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$$

and the auto-covariances $\gamma(\tau)$ at lags $\tau = 0, \ldots, p-1$ through the $p \times p$ matrix

$$\Sigma_p = \begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \ldots & \gamma(0) \end{pmatrix}$$

This system of equations has two applications. First, if $\gamma(0), \ldots, \gamma(p)$ are known (or have been estimated from a time series), the parameters of the AR(p) process can be determined (or estimated) by solving eq.[8.4] for $\vec{\alpha}_p$. Once the parameters have been estimated, both the auto-covariance function for lags $\tau > p$ and the spectrum of the unknown process can be estimated by the corresponding characterizations of the fitted AR(p) process. Second, if $\vec{\alpha}_p$ is known, then (11) can be recast as a linear equation with unknowns $\gamma(1), \ldots, \gamma(p)$, given the variance of the process $\gamma(0)$. Thus the Yule-Walker equations can be used to derive the first $p+1$ elements $1, \rho(1), \ldots, \rho(p)$ of the auto-correlation function. The full auto-covariance or auto-correlation function can now be derived by recursively extending equations (11). This is done by evaluating equation (10) for $\tau \geq p$ and taking expectations to obtain

$$\gamma(\tau) = \sum_{k=1}^{p} \alpha_k \gamma(k-\tau) \tag{8.5}$$

and

$$\rho(\tau) = \sum_{k=1}^{p} \alpha_k \rho(k-\tau) \tag{8.6}$$

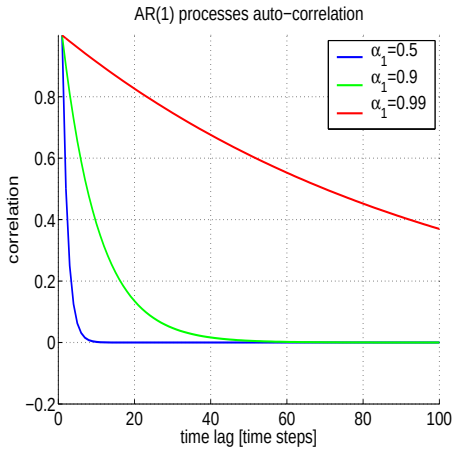


Figure 8.5: The auto-correlation function of different AR(1)-processes. They correspond to the time series shown in Fig. 7.2.

8.4 The Auto-correlation Functions of AR(1)- and AR(2)-Processes

AR(1)-process: The Yule-Walker equation (8.4) for an AR(1) process is

$$\gamma(1) = \alpha_1 \gamma(0)$$

Hence $\rho(1) = \alpha_1$. Applying (8.6) recursively we see that

$$\rho(\tau) = \alpha_1^{|\tau|}. \tag{8.7}$$

The Fig. 8.5 illustrates the auto-correlation function of different AR(1)-processes. They correspond to the time series shown in Fig. 7.2.

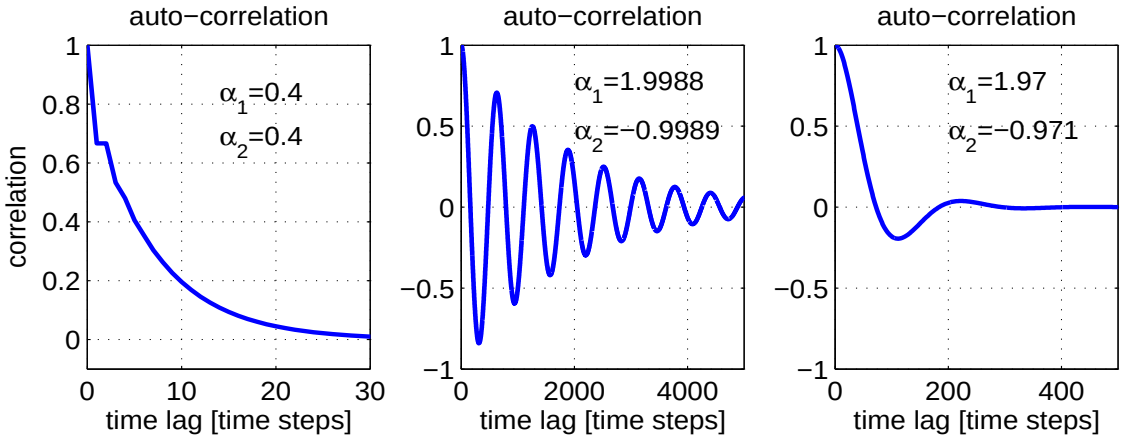


Figure 8.6: The auto-correlation function of different AR(2)-processes. They correspond to the time series shown in Fig. 7.3

AR(2)-process: The Yule-Walker equations (8.4) for an AR(2) process are

$$\alpha_1\gamma(0) + \alpha_2\gamma(1) = \gamma(1)$$
$$\alpha_1\gamma(1) + \alpha_2\gamma(0) = \gamma(2)$$

Using the first equation, we see that

$$\rho(1) = \frac{\alpha_1}{1-\alpha_2}. \tag{8.8}$$

Recursion (8.6) can be used to extend the auto-correlation function to higher lags. For example, the auto-correlation at lag-2 is

$$\rho(2) = \frac{\alpha_1^2 - \alpha_2^2 + \alpha_2}{1-\alpha_2}$$

The Fig. 8.6 illustrates the auto-correlation function of different AR(2)-processes. They correspond to the time series shown in Fig. 7.3.

8.5 The Characteristic Time Scales of Stochastic Processes (The Decorrelation Time)

In many statistical analysis we need to know the number degree of freedom, n_X, of the time series. e.g. the χ^2-pdf or the tests of mean,variance, correlation. Initially we had the definition for n_X, that $\mathbf{X_i}$ and $\mathbf{X_j}$ have to be independent, meaning uncorrelated. This definition is for stochastic processes not helpful, because the auto-correlation of an AR(1) process, for instance, does not go to zero at all.
However, n_X of $\mathbf{X}$ can be estimated by using the statistical relation between the statistical parameter of interest and n_X. For the mean we know from the central limit theorem that:

$$Var(\bar{\mathbf{X}}) = \frac{\sigma_X^2}{n_X} \tag{8.9}$$

If we know σ_X and $Var(\bar{\mathbf{X}})$ we can get n_X. The relation between the true number of time steps used and n_X is the decorrelation time :

$$\tau_D = \frac{n}{n_X} \tag{8.10}$$

For the mean we find:

$$\tau_D = 1 + 2\sum_{k=1}^{\infty}\rho(k) \tag{8.11}$$

For the variance we find:

$$\tau_D = 1 + 2\sum_{k=1}^{\infty}\rho(k)^2 \tag{8.12}$$

So why do we have different characteristic time scales for the mean and variance? Examples of AR(1) and AR(2) process can illustrate why the characteristic time scales must be different.
To be continued *ldots*

8.6 The Auto-Correlation Function of a Cyclo-Stationary Process

The auto-covariance function of a cyclo stationary process relative to a phase of the cycle is not symmetric. In order to illustrate the cyclo stationary behavior of the time series, it is helpful to present the auto-correlation function as a function of the time lag relative to a phase of the cycle, $\rho(t/m - \tau)$, with t/m being a phase of the cycle.
Fig. 8.7 illustrates an example: The auto-correlation of the monthly mean SST in the North Pacific as a function of the lag τ has no unusual features (ignoring the fact that is does not go zero, which is due to long time variations), but the auto-correlation as a function calendar month shows an asymmetry and it has a significant 'oscillation'. The oscillation of the auto correlation is not indicating an oscillation of the time series, which would have been an oscillation of the auto correlation around. This oscillation of the auto-correlation as a function calendar month is an indication of the reemergence of SST anomalies.

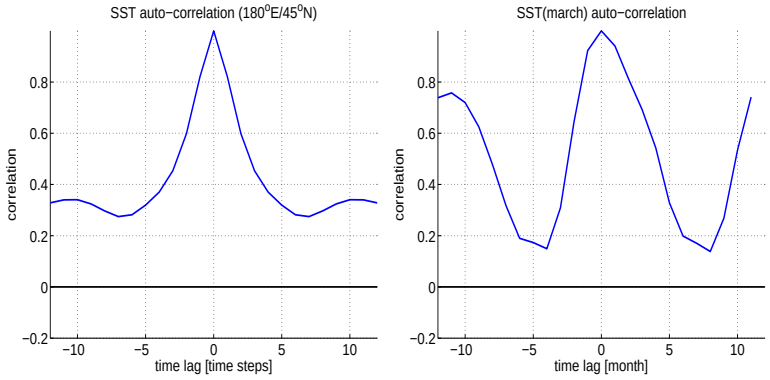


Figure 8.7: The auto-correlation of the monthly mean SST in the North Pacific as a function of the lag τ (left) for all month and the auto-correlation of march as a function calendar month relative to march (right). Positive lags indicate correlations to the future.

Chapter 9

The Spectrum

Power spectra of physical quantities are used in many different fields of research, e.q. light-, sound-waves. The spectra of a time series is the Fourier analysis of the time series, and it is the Fourier transform of the auto-covariance function of the time series. It presents the variance per frequency of the time series as a function of frequencies and therefore distributes the variance onto different frequencies. Studying the power-laws by which the variance is distributed over the frequencies is often one of the best ways to understand the underlaying physical processes. The analysis of the spectrum is therefore a central part of statistical analysis of time series.
We will first define the spectra as the Fourier transform of the auto-covariance function and discuss some properties. A more colorful description will be given by the Fourier analysis of the time series, namely the periodogram, in the final subsection, which discusses how the spectra is estimated from a time series. We will further discuss how the spectrum shall be presented and what the caveats of the different presentations are. The information contained in the spectrum or how it could be interpreted will be discussed based on some examples. The theoretical spectrum for AR(1)- and AR(2)-process will be discussed. Finally we give a short description of how the spectrum can be estimated from a time series, based on the periodogram.

9.1 Definition of the Spectrum

Let $\mathbf{X}_t$ be an ergodic weakly stationary stochastic process with auto-covariance function $\gamma(\tau), \tau = 0, \pm 1, \ldots$. Then the spectrum (or power spectrum) Γ of $\mathbf{X}_t$ is the Fourier transform $\mathcal{F}$ of the auto-covariance function γ. That is

$$\begin{aligned} \Gamma(\omega) &= \mathcal{F}\{\gamma\}(\omega) \\ &= \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-2\pi i \tau \omega}. \end{aligned} \tag{9.1}$$

for all $\omega \in [-1/2, 1/2]$. Note that the largest frequency that a time series with time step of 1.0 can resolve is $\omega = 1/2$. Therefore the spectrum is only defined for $\omega \in [-1/2, 1/2]$. The spectrum should not be extend beyond $\omega = 1/2$, this will in general make no sense (see, for instance, the spectrum of an AR(1)-process in section 9.6).

Note that since γ is an even function of τ,

$$\Gamma(\omega) = \gamma(0) + 2 \sum_{\tau=1}^{\infty} \gamma(\tau) \cos(2\pi\tau\omega)$$

This definition of the spectrum is similar to the definition of the covariance between $\gamma(\tau)$ and $\cos(2\pi\tau\omega)$ (compare with eq. 5.12). The spectrum can therefore be interpreted as the covariance between the auto-correlation function and the cosine function at different frequencies. so if the auto-correlation function project well onto a specific cosine-function, than this frequency will be the dominant frequency of the spectrum.

Characteristics of the spectrum:

- The spectrum ,$\Gamma(\omega)$, is continuous and differentiable everywhere in the interval $[-1/2, 1/2]$. So unlike the discrete time series or auto-correlation function, the spectrum is always continuous.
- The spectrum describes the distribution of variance across the time scales. In particular,

$$\mathrm{Var}(\mathbf{X}_t) = \gamma(0) = 2\int_0^{\frac{1}{2}} \Gamma(\omega)d\omega \tag{9.2}$$

  The spectra,$\Gamma(\omega)$, is a variance per frequency as a function of frequency, in units: $[\Gamma(\omega)] = [\mathbf{X}_t^2] \cdot [time]$. Note that is often unclear if a spectral estimate considers the right frequency unit, if the time step is not 1.0. Thus one should check if the integral of the spectrum fits to the total variance of the time series.

- $\gamma(\tau) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \Gamma(\omega)e^{2i\pi\omega\tau}d\omega$. The auto-covariance function can be reconstructed from the spectrum.
- $\frac{d}{d\omega}\Gamma(\omega)|_{\omega=0} = 0$. The spectra must be flat at long time scales for stationary processes.
- $\Gamma(\omega) \sim \chi^2 \quad \rightarrow \quad \sigma(\Gamma(\omega)) \sim E(\Gamma(\omega))$ The spectral coefficient, $\Gamma(\omega)$, are variances, those *pdf* is χ^2-distributed. Thus the statistical uncertainty of an estimated spectrum is proportional to expectation value, which is important for discussion of the significance of peaks in the spectrum.

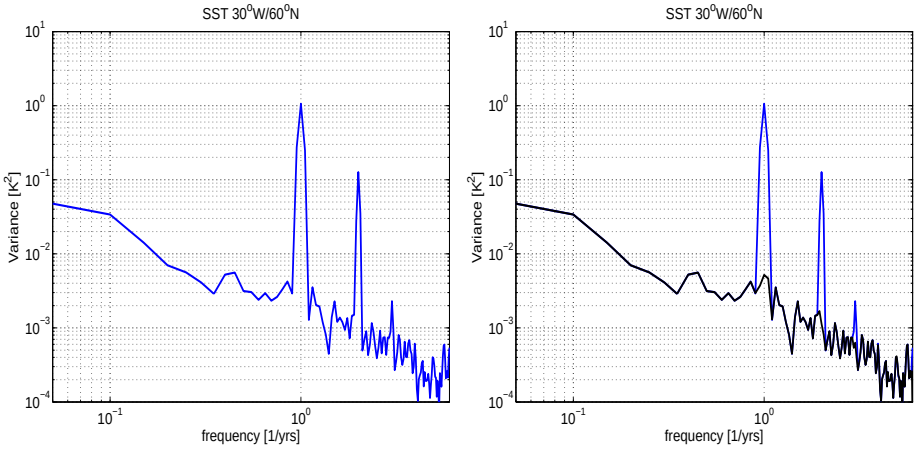


Figure 9.1: Spectral estimates of SST time series in the northern North Atlantic. Left is with seasonal cycle and right is also with the mean seasonal cycle removed.

9.2 Presentation of the Spectrum

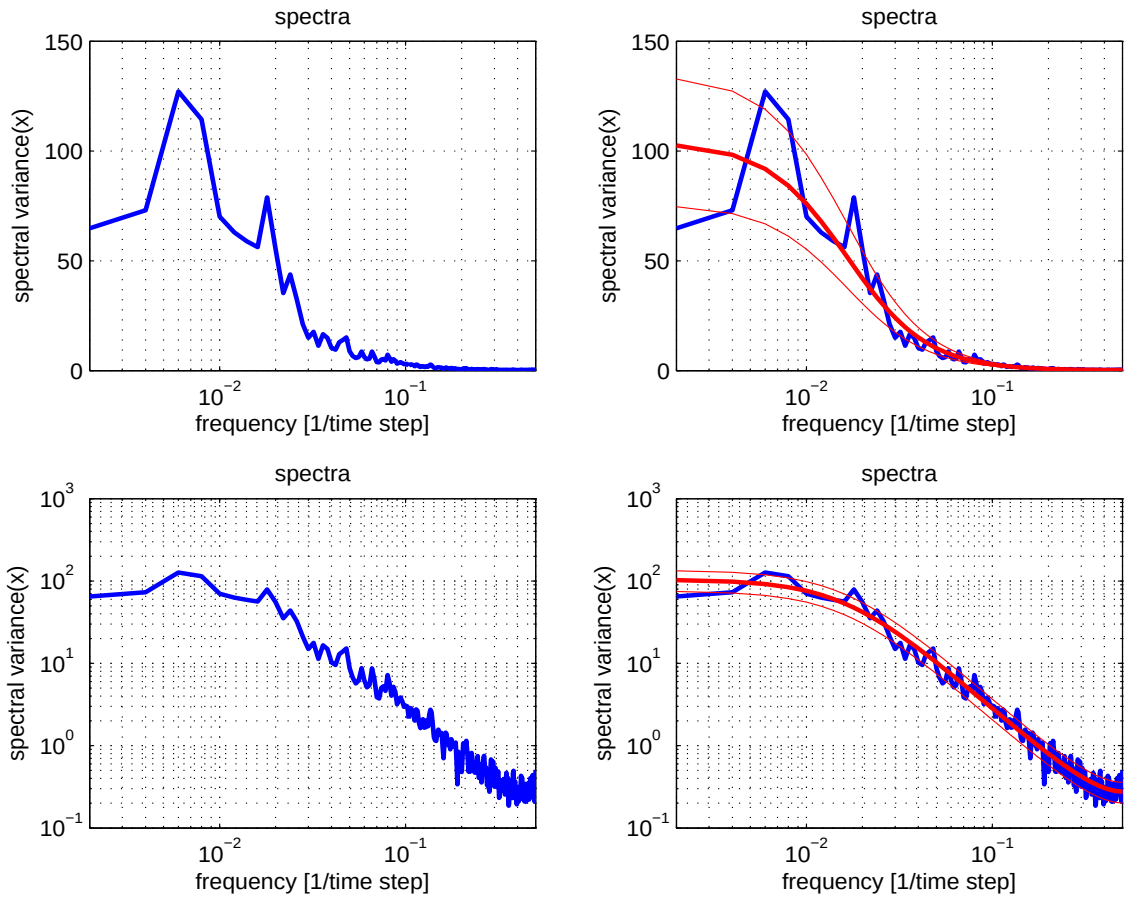


Figure 9.2: The spectrum of an AR(1)-process time series presented in log-linear scaling (upper) and in log-log scaling (lower). In addition the 10% and 90% quantiles of the spectral coefficients estimate are plotted (right).

By representing/plotting the spectra one should consider, some important characteristic of the spectra. We have basically three option to present the spectra, with each of it has its advantages and its drawbacks:

1. **Linear:** This seems the obvious way to present the spectrum, but is in many cases not optimal. The Advantage: The area between the x-axis and the $\Gamma(\omega)$-curve is the total variance. So we can easily evaluate how the total variance is distributed over the frequency domain.

   The drawbacks are: It is difficult to highlight low-freq. variances within the full spectrum. The spectral coefficients are equidistant on the frequency axis, but we tend to think of the time scales in terms of orders of magnitudes, e.g. 1, 10^1 , 10^2 . . . . Thus we like to present the x-axis in log-scale.

   Further the uncertainties of $\Gamma(\omega)$ are a function of frequency. The spectral coefficients are variances, those *pdf* is χ^2-distributed. Thus the statistical uncertainty is proportional to expectation value. If plotted linearly the uncertainty of each spectral coefficients has different length. If plotted on log-scale the uncertainty of each spectral coefficients has equal length. The example in Fig. 9.2 illustrates that the linear presentation of $\Gamma(\omega)$ emphasis spectral

peaks, which are actually just fluctuations due to the limited time series for estimating the spectrum. In the log-scaling the peak $5 \cdot 10^{-3}$ appears to much less 'significant'.

Another drawback is that theoretical models for the stochastic nature of the time series tent to follow simple power-laws in log-log scaling, which are not as easily verified in linear presentations. See spectrum of AR-p processes.

2. **LogLog-scale**: In log-log scaling we can present the entire spectrum over all timescales and all scales of variance. The advantage is that the error of each spectral estimate is now a constant length. Simple power laws, such as the AR(1)-process, have often a linear relation over a large part of the log-log presentation. The drawback is that it unclear how the total variance is distribution over the frequency, because now the area between the x-axis and the $\Gamma(\omega)$-curve is not the total variance anymore. One tents to over estimate the importance of low-frequency variance. Compare, for instance the different presentations of the El Niño spectrum in Fig. 9.3.

   Note that sometimes the log-scaling unit *decibell* [db] is used. This is note a real unit, it means: $1db = 10^{0.1}$, $10db = 10^{1}$ and , $25db = 10^{2.5}$. However, this unit is not used in climate research. We right the real units (e.g. $K^2 \cdot yrs$).

3. **Linear-Log(f)-scale**: A trade off between linear and log-presentation is to plot $\Gamma(\omega) \cdot \omega$ as function of $log(\omega)$. Note: $dlog(\omega) = \frac{1}{\omega} d\omega$. By presenting $\Gamma(\omega) \cdot f$ as function of $log(f)$ we maintain the option how the total variance is distributed over the frequency domain, by the area between the x-axis and the $\Gamma(\omega)$-curve. The drawback is still that the errors are a function of frequency and theoretical power-laws are still not as easily evaluate as in the log-log presentation.

Fig. 9.3 illustartes the different presentations for the El Niño spectrum.

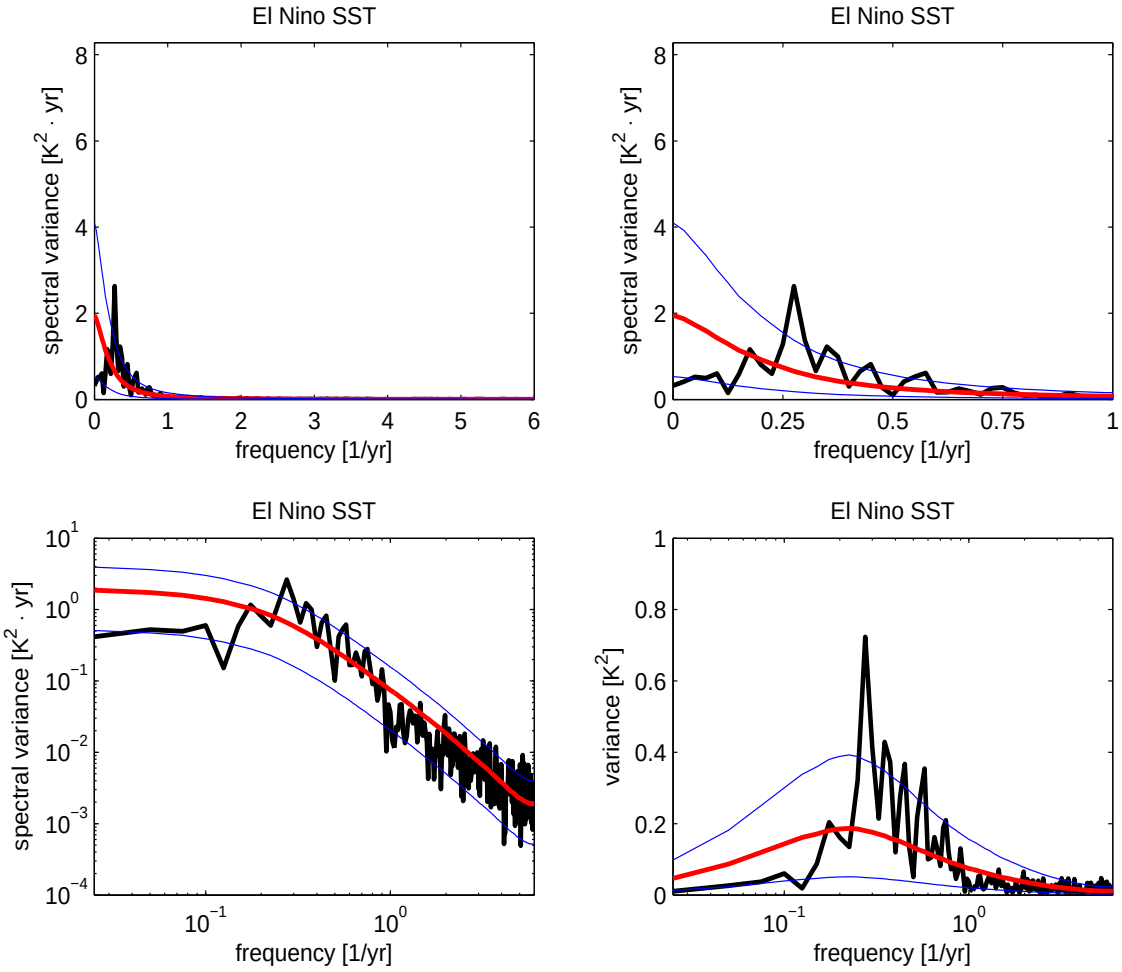


Figure 9.3: The spectrum of the monthly mean El Niño time series in four different presentations of the same spectrum. Note the lower right panel shows $\Gamma(\omega) \cdot \omega$. For all spectra the fitted AR(1)-process spectrum with the 90% confidence interval is plotted for comparison.

9.3 Interpretation of the Spectrum

The Spectrum is: $[\Gamma(\omega)] = \frac{[Var(\omega)]}{[\omega]}$. It gives a variance density along the frequency-scale. The amplitude of a sine-function at frequency ω fitted to the time series would have the amplitude:

$$\mathcal{E}(\sigma(sin(2\pi\omega))) = \sqrt{\int_{\omega-\Delta\omega}^{\omega+\Delta\omega} \Gamma(\omega)d\omega} \approx \sqrt{\Gamma(\omega)\cdot 2\Delta\omega} \approx \sqrt{\Gamma(\omega)\cdot\omega}$$

if we assume that we resolve ω by $\Delta\omega \approx 0.5\omega$ (it seems reasonable to assume that the resolution is proportional to ω). Note, that in this estimate you have to multiply by the frequency, ω, to get the variance. So a constant spectrum will have sine-functions with decreasing amplitude for low-frequencies, although the variance density is constant.

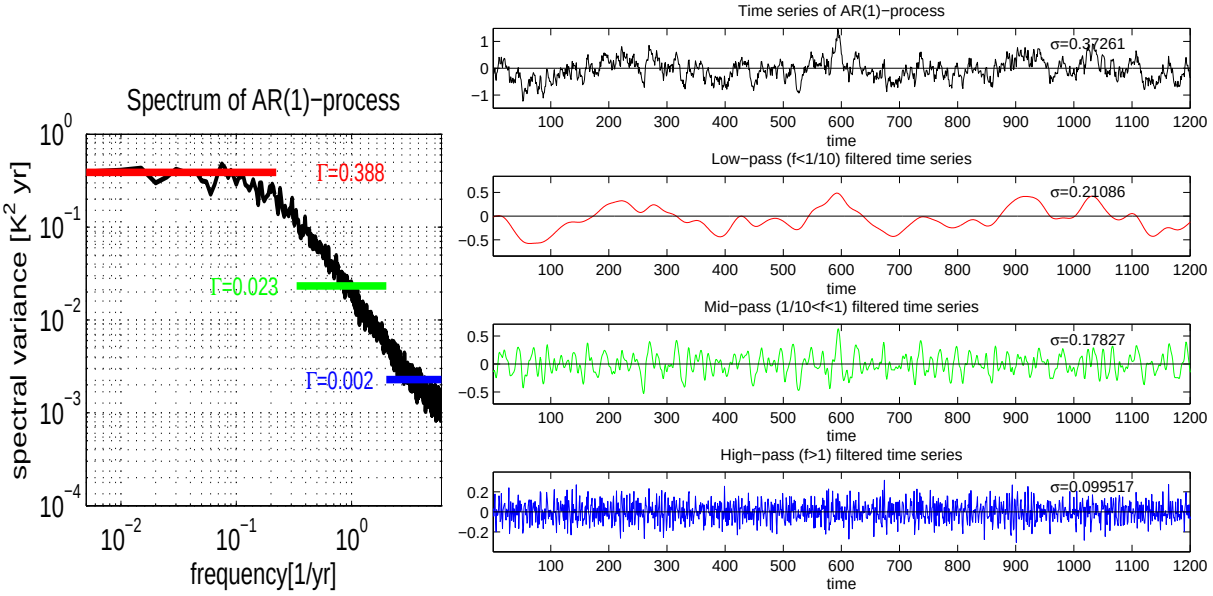


Figure 9.4: Comparison of the time series (upper right), band-pass filtered time series (lower right) and its relation to the spectrum of the time series. The colored line in the spectrum (left) illustrate the frequency-band for which the time series on the rigth were filtered. The numbers next to it are the mean Γ for the frequency-band.

Fig. 9.4 illustrates how the spectrum can be 'read'. The figure shows a time series of an AR-1 process and its spectrum. In addition three band-pass filtered time series of the original time series are shown. The standard deviations, σ, of the four time series are different, with the largest standard deviation for the original time series. The standard deviations of the band-pass filtered time series can be estimated by the looking at the spectrum. We can estimate the frequency-band to which we have filtered the time series and we can read the mean Γ for the frequency-band. Thus we estimate the integral or area underneath the curve (if plotted in linear scale; not as shown in Fig. 9.4) by the length of the $\delta\omega$ and the mean $\Gamma(\omega)$ over the interval. The standard deviation of the filter time series would than be:

$$\sigma(low-pass) \approx \sqrt{0.39 * 0.114} = 0.21$$

$$\sigma(mid-pass) \approx \sqrt{0.0232 * 1.167} = 0.16$$

$$\sigma(high-pass) \approx \sqrt{0.0023 * 4.0} = 0.10$$

A comparison with standard deviation estimated from band-pass filtered the time series (Fig. 9.4) shows that the estimates are quite close. Note that one may be mislead by the spectrum if we just compare the variance at low-freq. with those at high-freq.. $\Gamma(low - freq.)$ is 100 times larger than $\Gamma(high - freq.)$, but the low-pass time series standard deviations is only twice as large. We have to account for the different length of the frequency-band, which is about a 100-times shorter for the low-pass filtered time series.

9.4 The Spectra of AR(p) Processes

The spectrum of an AR(p) process with process parameters $\{\alpha_1, \ldots, \alpha_p\}$ *and noise variance* $Var(\mathbf{Z}_t) = \sigma_Z^2$ *is*

$$\Gamma(\omega) = \frac{\sigma_Z^2}{|1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i k \omega}|^2}. \tag{9.3}$$

The structure of this equation is similar to the total variance of AR(P), see eq.[7.13].

9.5 The Spectrum of a white noise process.

A white noise process is an AR(0) process:

$$\Gamma(\omega) = \sigma_Z^2 \tag{9.4}$$

A white noise process distributes the variance on all frequencies equally.

9.6 The Spectrum of an AR(1) Process.

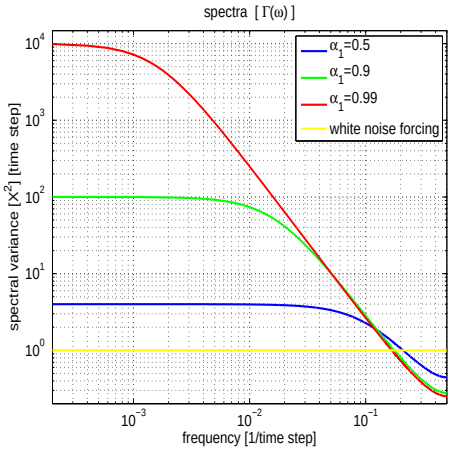


Figure 9.5: Spectra of different AR(1)-processes.

The power spectrum of an AR(1) process with lag-1 correlation coefficient α_1 is

$$\Gamma(\omega) = \frac{\sigma_Z^2}{|1 - \alpha_1 e^{-2\pi i \omega}|^2} = \frac{\sigma_Z^2}{1 + \alpha_1^2 - 2\alpha_1 \cos(2\pi\omega)} \tag{9.5}$$

We can approximate the cosine for low values of $\omega \in [0, 0.5]$ with the first two terms of the cosine series:

$$cos(2\pi\omega) = 1 - \frac{(2\pi\omega)^2}{2!} + ...$$

thus the AR(1) spectra is approximately:

$$\Gamma(\omega) \approx \frac{\sigma_Z^2}{1 + \alpha_1^2 - 2\alpha_1(1 - \frac{(2\pi\omega)^2}{2!})} = \frac{\sigma_Z^2}{1 + \alpha_1^2 - 2\alpha_1 + \alpha_1(2\pi\omega)^2} = \frac{c_1\sigma_Z^2}{c_2 + \omega^2} \tag{9.6}$$

The spectra of an AR(1) process is therefore approximately following a linear gradient of -2 in loglog-scale for frequencies $\omega >> 0$. This spectrum has no extremes in the interior of the interval $[0, 1/2]$ because, everywhere inside the interval, the derivative

$$\frac{d}{d\omega}\Gamma(\omega) = -2\alpha_1\Gamma_1(\omega)^2 \sin(2\pi\omega) \neq 0$$

The sign of the derivative is determined by α_1. Thus the spectrum has a minimum at one end of the interval $[0, 1/2]$ and a maximum at the other end. When $\alpha_1 > 0$, the 'spectral peak' is located at frequency $\omega = 0$, where it reaches a plateau. Such processes are often referred to as *red noise processes*.

In Fig. 9.5 we see the spectra of AR(1) processes with different α_1 compared with the spectrum of the driving white process. We can see that all spectra of AR(1) processes have more variance on the low-frequencies while they have less variance for the high-frequencies if compared with the spectrum of the driving white process. In the intermediate frequencies all spectra of AR(1) processes have a linear decreasing spectrum with a gradient of -2, while the AR(1) process with the largest α_1 has the longest increase in variance.

Note that this presentation may be somewhat misleading, since it indicates that an AR(1)-process amplifies the forcing. But a linear damping system, the physical process associated with an AR(1)-process (see section7.3.2), is always damping the forcing signal. For more details see 9.9.

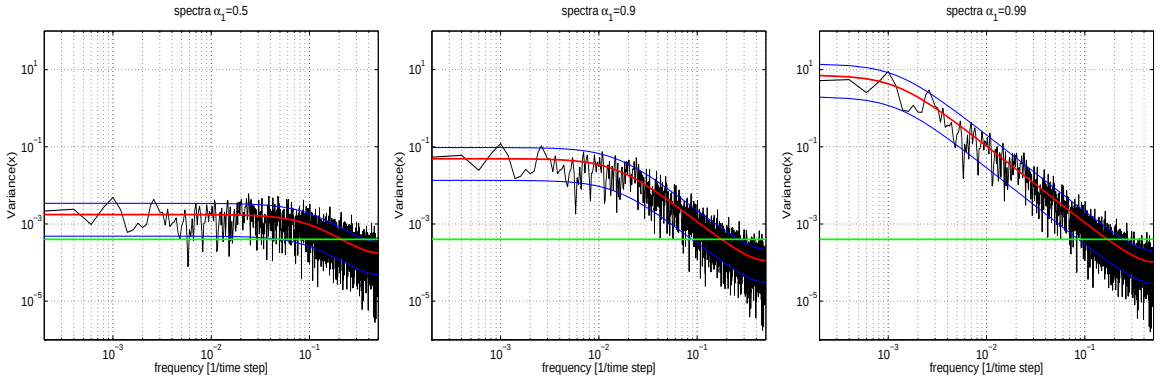


Figure 9.6: The spectra of AR(1) processes with different α_1 compared with the spectrum of the driving white process. The spectra correspond to the time series in Fig. 7.2.

In Fig. 9.6 we see the spectra of different AR(1)-process, as estimated from time series and the theoretical spectrum.

9.7 Fitting the AR(1)-Process to a time series.

The AR(1)-process, red noise, is often chosen as the null hypothesis for the time scale characteristics of a climate variables variance. It is therefore a practice to compare the spectra of time series with

the fitted AR(1) process. The AR(1)-process is well defined by the standard deviation of the time series, σ_{X_t} and the lag-1 correlation. The spectrum of the fitted AR(1) process results from eq.[9.5], using the relation between σ_{X_t} and σ_Z from eq.[7.16]:

$$\sigma_Z^2 = (1 - \alpha_1^2)\sigma_{X_t}^2 \tag{9.7}$$

Some examples are shown in Fig. 9.7 and 9.8.

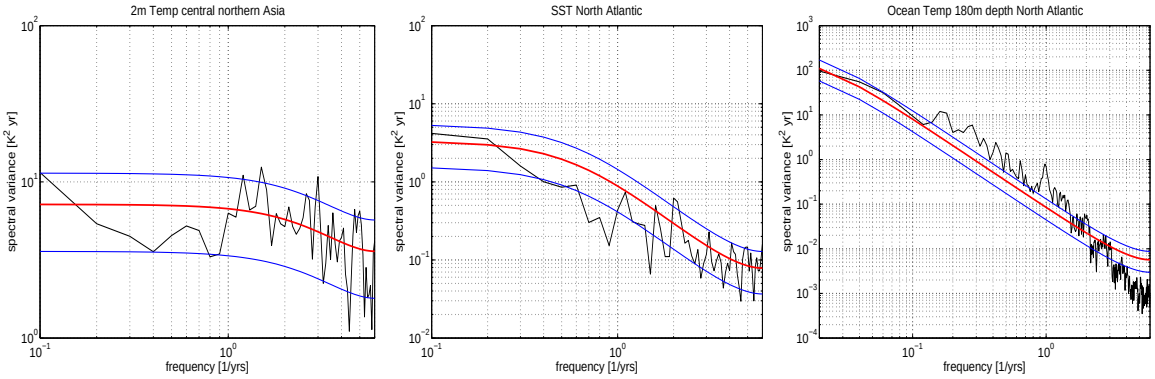


Figure 9.7: Spectra of observed 24hrs mean temperature time series. For comparison the fitted AR(1)-process spectra are shown together with the 95% confidence interval.

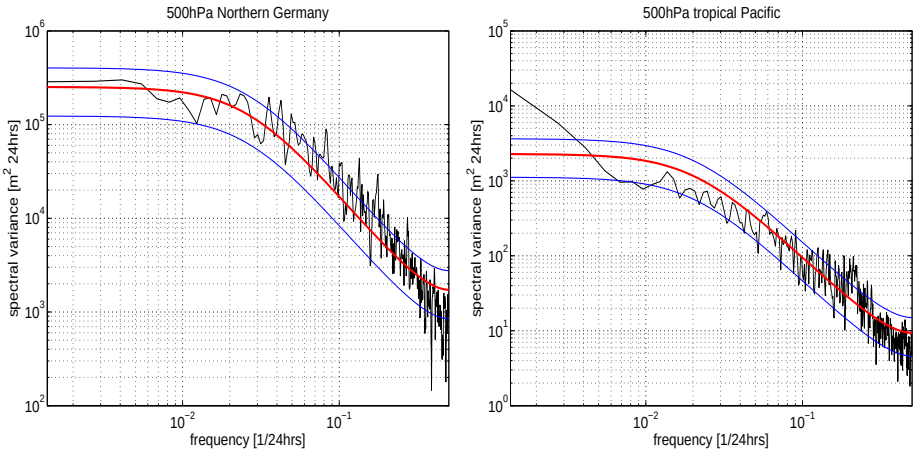


Figure 9.8: Spectra of observed 24hrs mean 500hPa geopotential heights time series. For comparison the fitted AR(1)-process spectra are shown together with the 95% confidence interval.

Note, that comparing these spectra of observed time series with the fitted AR(1)-process has some limitations. The spectrum following eq.[9.5] is the spectrum of a discrete process with the time step of the time series. The observed time series are in general continuous processes that are sampled on some time interval (days, hours, etc.) and than averaged onto the time step of the time series (e.g. monthly means, annual means).
Fig. 9.10 illustrates that the spectrum of a time series resulting from a discrete AR(1)-process with the time step of one month is different from the spectrum of a discrete AR(1)-process with the time step of one day averaged to a monthly mean time series. This effect has to be considered if deviations from the AR(1)-process are discussed.

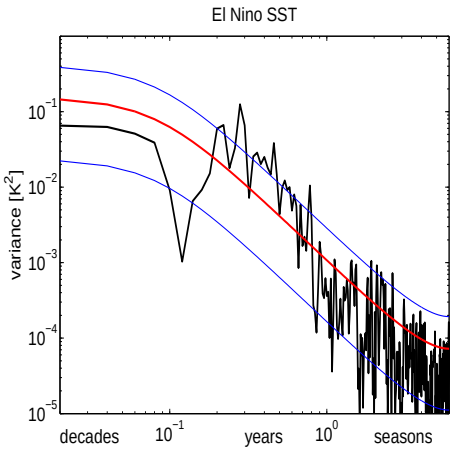


Figure 9.9: Spectra of the observed monthly mean El Niño time series (NINO3 region average). For comparison the fitted AR(1)-process spectra are shown together with the 95% confidence interval.

9.8 The Spectrum of an AR(2) Process.

The power spectrum of an AR(2) process with parameters (α_1, α_2) (24) is given by

$$\Gamma(\omega) = \frac{\sigma_Z^2}{1 + \alpha_1^2 + \alpha_2^2 - 2g(\omega)}$$

where

$$g(\omega) = \alpha_1(1 - \alpha_2)\cos(2\pi\omega) + \alpha_2\cos(4\pi\omega)$$

Depending upon the parameters, the spectrum can have a minimum or maximum in the interior of the interval $[0, 1/2]$. When its derivative is zero, $\Gamma(\omega)$ has a maximum or minimum, and we note that $\Gamma'(\omega) = 0$ whenever $g'(\omega) = 0$. By using the identity $\sin(4\pi\omega) = 2\sin(2\pi\omega)\cos(2\pi\omega)$, we find that

$$\begin{aligned} g'(\omega) &= -2\pi\alpha_1(1 - \alpha_2)\sin(2\pi\omega) \\ &\quad -4\pi\alpha_2\sin(4\pi\omega) \\ &= (-2\pi)\sin(2\pi\omega) \\ &\quad \times\Big(\alpha_1(1 - \alpha_2) + 4\alpha_2\cos(2\pi\omega)\Big). \end{aligned}$$

Since $\sin(2\pi\omega) \neq 0$ for all $\omega \in (0, 1/2), \Gamma'(\omega) = 0$ it follows the condition for a maxima and minima:

$$\cos(2\pi\omega) = -\alpha_1(1 - \alpha_2)/(4\alpha_2). \tag{9.8}$$

The last equation has a solution $\omega \in (0, 1/2)$ when $|\alpha_1(1 - \alpha_2)| < 4|\alpha_2|$. This solution represents a spectral maximum when $\alpha_2 < 0$ and a spectral minimum when $\alpha_2 > 0$.
Note also, that the log-slope of an AR(2) is variable. It can be smaller or larger than the slope of an AR(1)-process. Examples of spectra of AR(2)-processes are shown in Fig.9.11.

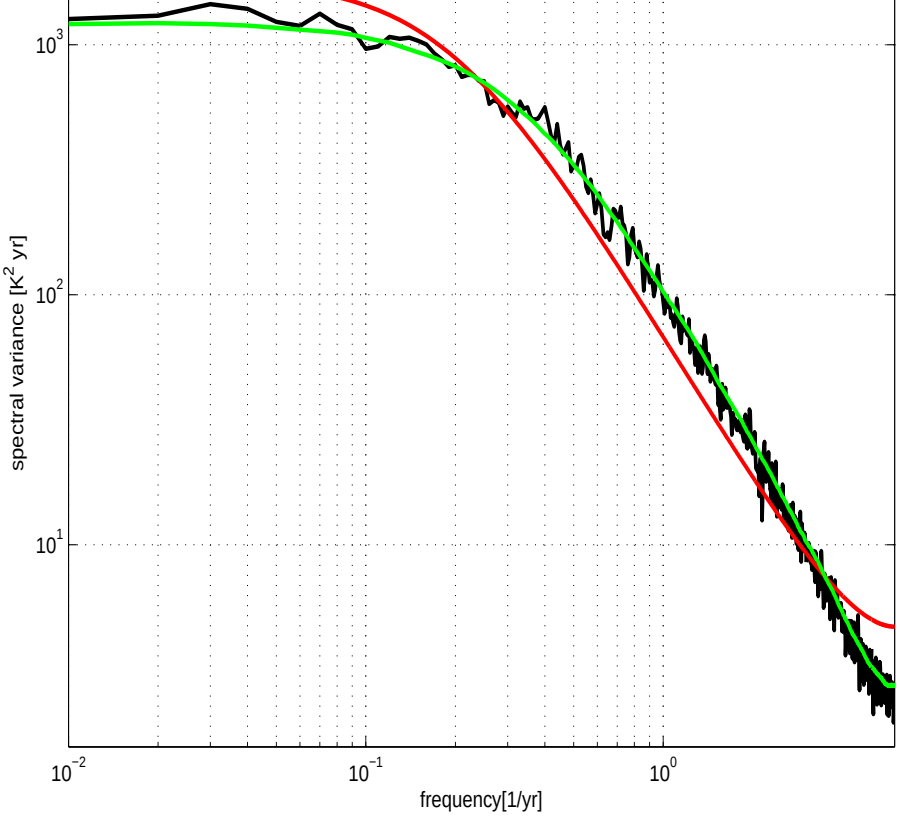


Figure 9.10: The spectrum of a discrete AR(1)-process with the time step of one day averaged to a monthly mean time series (black, green). For comparison the spectrum resulting from eq.[9.5] of the AR(1)-process fitted to monthly mean time series is shown (red).

9.9 The Spectra of some continuous physical processes (differential equations)

The discussion of the spectra of AR-p processes can be somewhat misleading, because it focus on discrete time series and some statistical parameters. The spectrum of a simple differential equation, which describe some simple physical processes, has sometimes different characteristics. The variance of an AR-1 process, for instance, is larger than that of the forcing. However, the variance of a damped system (eq.7.17), which is a physical interpretation of an AR-1 process, is always smaller than that of the forcing.
We therefore will discuss the spectrum of some simple physical processes to illustrate some important characteristics.
... to be continued!

9.10 Estimating the Spectra (The Periodogram)

The variance of a time series $\{X_1, X_2, \ldots, X_T\}$ of *finite* length may be attributed to different time scales by expanding it into a finite series of trigonometric functions. We have assumed, for mathematical convenience, that T is odd. The expansion is slightly more complex when T is even.

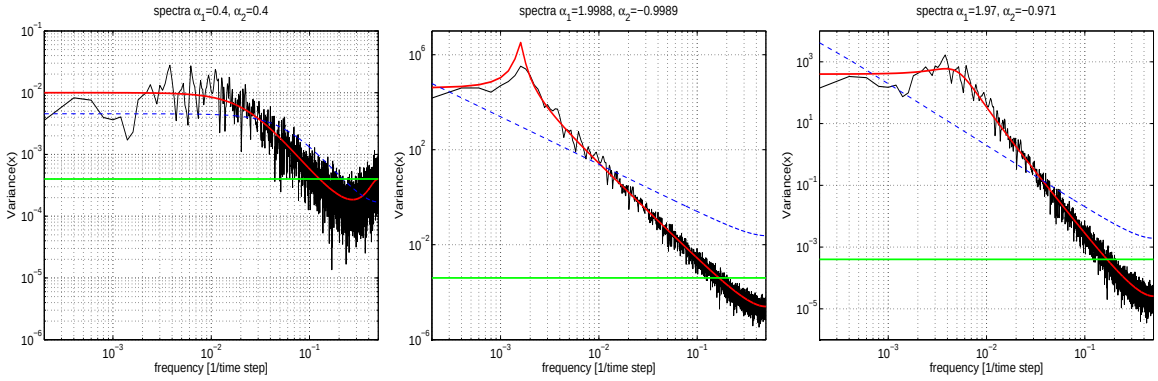


Figure 9.11: The spectra of AR(2) processes with different α_1, α_2 (red line) and the estimated spectra from the time series in Fig. 7.3 (black line), compared with the spectrum of the driving white process(green line) and a fitted AR(1)-process(dashed blue line).

$$X_t = A_0 + \sum_{k=1}^{(T-1)/2} \left(a_k \cos \frac{2\pi k t}{T} + b_k \sin \frac{2\pi k t}{T} \right). \tag{9.9}$$

So the T values of X_t are reorganized into a mean A_0 and (T-1)/2 pairs of spectral coefficients. We will see below that the spectrum is a continuous function of frequency. In contrast, the periodogram is always discrete. Thus the number of non-trivial coefficients a_j and b_j is always T.
The coefficients are estimated over the mean and covariances:

$$A_0 = \hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \tag{9.10}$$

and

$$a_j = \frac{2}{T} \sum_{t=1}^{T} \mathbf{x}_t \cos(2\pi\omega_j t) \tag{9.11}$$

$$b_j = \frac{2}{T} \sum_{t=1}^{T} \mathbf{x}_t \sin(2\pi\omega_j t), \tag{9.12}$$

for $j = 1, \ldots, q$. Note that, for even T,

$$a_q = \frac{1}{T} \sum_{t=1}^{T} (-1)^q \mathbf{x}_t, \tag{9.13}$$

$$b_q = 0 \tag{9.14}$$

The equation splits the time series into a number of frequencies:

$$\omega_i = \frac{i}{T} \quad i = 1, ..., (T-1)/2 \tag{9.15}$$

with $\Delta\omega = \frac{1}{T}$. So the longer the time series the smaller the $\Delta\omega$, and the more independent spectral coefficients are estimated. The frequency interval is constant. it follows from the fact that this choice of frequencies leads to uncorrelated estimates over the time series length:

$$\begin{aligned} \text{a)} \quad & \sum_{t=1}^{T} \cos(2\pi\omega_k t)\cos(2\pi\omega_l t) &&= \frac{T}{2}\delta_{kl} \\ \text{b)} \quad & \sum_{t=1}^{T} \sin(2\pi\omega_k t)\sin(2\pi\omega_l t) &&= \frac{T}{2}\delta_{kl} \\ \text{c)} \quad & \sum_{t=1}^{T} \cos(2\pi\omega_k t)\sin(2\pi\omega_l t) &&= 0, \end{aligned}$$

where $\delta_{kl} = 1$ if $k = l$ and 0 otherwise. Equation (9.9) distributes the variance in the time series

$$Var(X_t) = \frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2 = \frac{1}{2}\sum_{k=1}^{(T-1)/2}\left(a_k^2 + b_k^2\right) \tag{9.16}$$

to the periodic components in the expansion shows in (9.9). The elements $(a_k^2 + b_k^2)$ are collectively referred to as the *periodogram* of the finite time series $\{X_1, \ldots, X_T\}$. Unfortunately, it is not readily apparent that the expansion in (20) is related to the spectrum of an *infinite* time series or a stationary process, although it is true.
The *periodogram* is defined in terms of the coefficients a_j and b_j as

$$I_{Tj} = \frac{T}{4}(a_j^2 + b_j^2) \tag{9.17}$$

The periodogram is the basis for most estimates of the spectra. However, the periodogram has some bad characteristics:

- $I_{Tj} \sim \chi^2(2)$. The Distribution of the Periodogram is therefore relatively wide, strongly skewed to larger values and it peaks at zero, see Fig. 15.7.
- The uncertainty of the spectra coefficients is independent of the length of the time series. Thus increasing the length of the time series does not improve the estimate of single spectral coefficients, which is very unfortunate.

9.11 Better estimates of the spectra based on the Periodogram

each estimate of a spectra based on finite time series has an uncertainty in the variance and in the frequency:

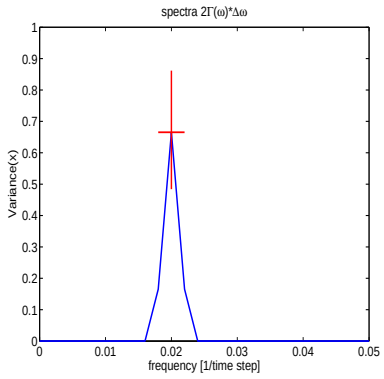


Figure 9.12: Spectral estimate of a sine function time series with period 50 time steps. The red lines indicate uncertainties in the variance and period estimation.

The different methods try to minimize the combination of both in different ways.

The basic idea two ideas are to split and tapper the time series

The chunk estimator is computed as follows.

1. Divide the time series into m chunks of length $M = \lceil \frac{T}{m}$.

2. Compute a periodogram

   $$I_{Tj}^{(\ell)}, \quad j = 0, \ldots, q, \quad q = \lceil \frac{M}{2}$$

   from each chunk $\ell = 1, \ldots, m$.

3. Estimate the spectrum by averaging the periodograms:

$$\hat{\Gamma}(\omega_j) = \frac{1}{m} \sum_{\ell=1}^{m} I_{Tj}^{(\ell)}. \tag{9.18}$$

The result is an estimator with approximately $\sim \chi^2(2m)$

The estimate at each frequency is representative of a special *bandwidth* of approximately $1/M$.

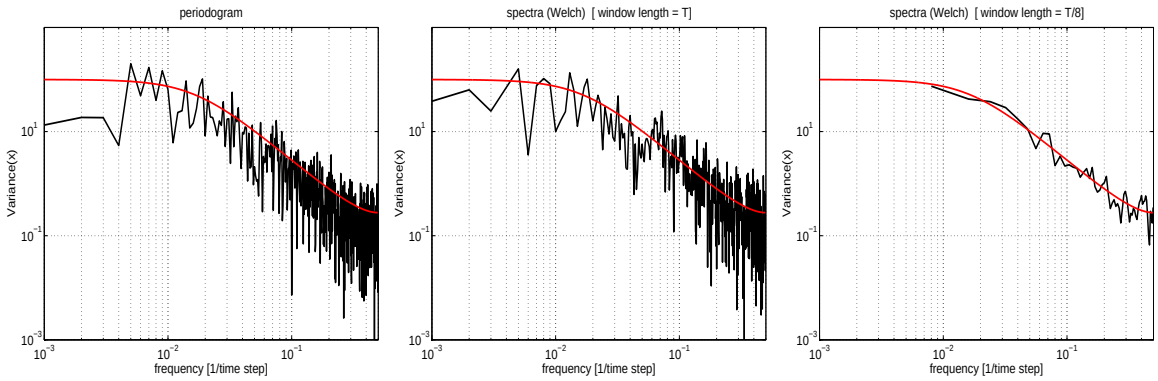


Figure 9.13: Different estimates of the spectrum of a AR(1)-process time series, based on the Periodogram.

9.11.1 Filter of a Time Series (The Running Mean)

Often we like to filter the time series to highlight the time evolution of some specific frequency band.

... to be continued

Chapter 10

The Cross-Covariance Function

If we want to analyze the covariability between two time series of random variables $\mathbf{X_t}$ and $\mathbf{Y_t}$ we study the cross-covariance function and the Fourier transform of it, the cross-spectrum. Much of the math is similar to the auto-covariance analysis in the previous sections.
A non-parametric estimator of the cross-correlation function $\rho_{xy}(\tau)$ is given by

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_X \sigma_Y} \tag{10.1}$$

where $\gamma_{xy}(\tau)$ is the sample cross-covariance function is for $\tau \geq 0$

$$\gamma_{xy}(\tau) = \frac{1}{T} \sum_{t=1}^{T-\tau} \mathbf{X}'_t \mathbf{Y}'_{t+\tau}. \tag{10.2}$$

and for $\tau < 0$

$$\gamma_{xy}(\tau) = \frac{1}{T} \sum_{t=1-\tau}^{T} \mathbf{X}'_t \mathbf{Y}'_{t+\tau}. \tag{10.3}$$

The sample cross-covariance function is set to zero for $|\tau| \geq T$. As a repetition of the fundamentals section:

- $\tau > 0 \Rightarrow$ the time evolution of $\mathbf{X_t}$ leads those of $\mathbf{Y_t}$ and vice versa for $\tau < 0$
- γ_{xy} can be asymmetric
- $\gamma_{xy}(\tau)$ can be larger than $\gamma_{xy}(0)$
- $|\rho_{xy}(\tau)| <= 1$

10.1 Some Examples

- **$\mathbf{Y_t}$ is a linear function of $\mathbf{X_t}$:**

$$\mathbf{Y_t} = a\mathbf{X_t} \tag{10.4}$$

the auto-correlation of $\mathbf{Y_t}$ is

$$\gamma_{yy}(\tau) = \mathcal{E}(Y_t \cdot Y_t) = \mathcal{E}(\alpha X_t \cdot \alpha X_t) = \alpha^2 \mathcal{E}(X_t \cdot X_t) = \alpha^2 \gamma_{xx}(\tau) \tag{10.5}$$

then the cross-covariance function is simply

$$\gamma_{xy}(\tau) = \mathcal{E}(X_t \cdot Y_t) = \mathcal{E}(X_t \cdot \alpha X_t) = \alpha \cdot \mathcal{E}(X_t \cdot \alpha X_t) = \alpha \cdot \gamma_{xx}(\tau) \tag{10.6}$$

and the cross-correlation function is

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_X \sigma_Y} = \frac{\alpha \gamma_{xx}(\tau)}{\sigma_X \alpha \sigma_X} = \frac{\gamma_{xx}(\tau)}{\sigma_X \sigma_X} = \rho_{xx}(\tau) \tag{10.7}$$

- **$\mathbf{Y_t}$ is a linear function of $\mathbf{X_t}$, but some additional white noise is added.**

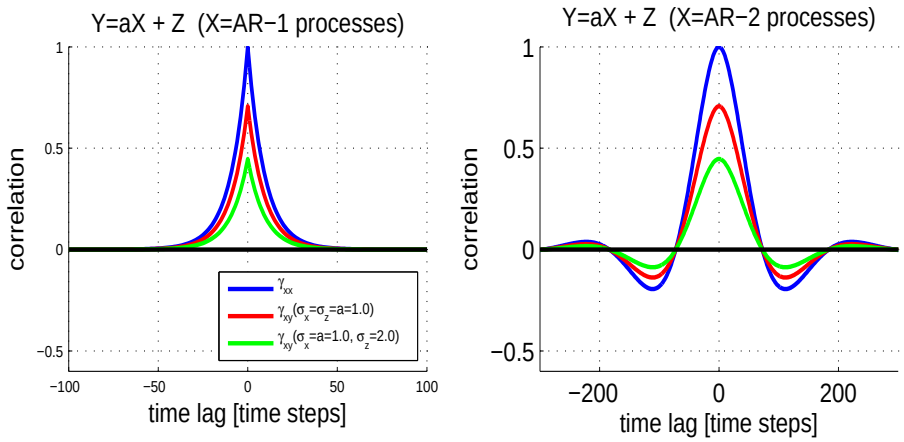


Figure 10.1: The auto/cross-correlation for the process $\mathbf{Y_t} = \alpha \mathbf{X_t} + \mathbf{Z_t}$. For $\mathbf{X_t}$ following an AR(1)-process (left) and an AR(2)-process (right).

$$\mathbf{Y_t} = \alpha \mathbf{X_t} + \mathbf{Z_t} \tag{10.8}$$

with $\mathbf{Z_t}$ as white noise. It follows for the auto-correlation of $\mathbf{Y_t}$

$$\gamma_{yy}(\tau) = \mathcal{E}((\alpha x + z)(\alpha x + z)) = \mathcal{E}((\alpha^2 xx + 2\alpha xz + zz))$$

with $\mathcal{E}(xz) = 0$,

$$\gamma_{yy} = \mathcal{E}((\alpha^2 xx + zz))$$

for $\tau = 0$

$$\gamma_{yy}(0) = \alpha^2 \gamma_{xx}(0) + \sigma_z^2$$

and $\forall \tau \neq 0$

$$\gamma_{yy}(\tau) = \alpha^2 \gamma_{xx}(\tau)$$

The covariance is

$$\gamma_{xy}(\tau) = \alpha\gamma_{xx}(\tau)$$

For the cross-correlation function we find

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_X\sigma_Y} = \frac{\alpha\gamma_{xx}(\tau)}{\sqrt{\sigma_X^2(\alpha^2\sigma_X^2+\sigma_z^2)}} = \frac{\alpha\gamma_{xx}(\tau)}{\alpha\sqrt{\sigma_X^2\sigma_X^2(1+\frac{\sigma_z^2}{\alpha^2\sigma_X^2})}} = \frac{\rho_{xx}(\tau)}{\sqrt{1+\frac{\sigma_z^2}{\alpha^2\sigma_X^2}}} \tag{10.9}$$

thus the cross-correlation function between $\mathbf{Y_t}$ and $\mathbf{X_t}$ is just the auto-correlation of $\mathbf{X_t}$, but scaled down in relation to $\frac{\sigma_z^2}{\alpha^2\sigma_X^2}$. So the larger the relative influence of the white noise onto $\mathbf{Y_t}$ the smaller the cross-correlation function with $\mathbf{X_t}$.

- **$\mathbf{Y_t}$ is a lagged time series of $\mathbf{X_t}$,**

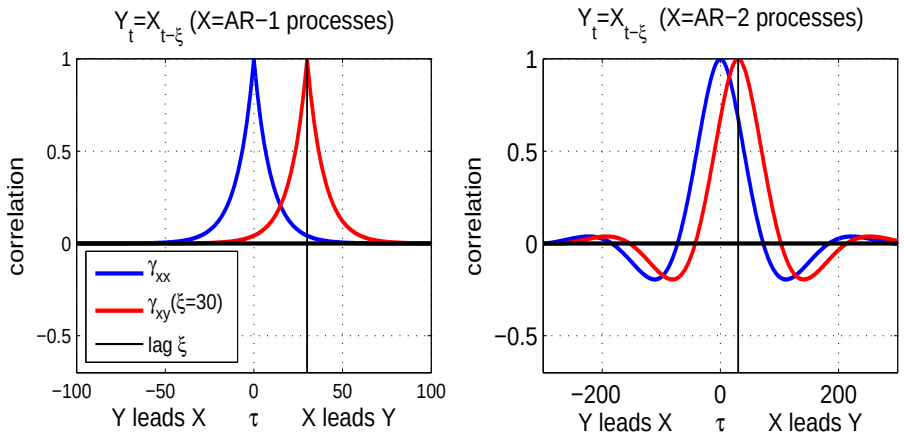


Figure 10.2: The auto/cross-correlation for the process $\mathbf{Y_t} = \mathbf{X_{t-\xi}}$. For $\mathbf{X_t}$ following an AR(1)-process (left) and an AR(2)-process (right). For $\tau > 0$ the time evolution of anomalies in $\mathbf{X_t}$ leads those of $\mathbf{Y_t}$.

$$\mathbf{Y_t} = \mathbf{X_{t-\xi}} \tag{10.10}$$

Any variability that happens in $\mathbf{X_t}$ at time step t will occur ξ time steps later in $\mathbf{Y_t}$. Thus the auto-correlation is the same

$$\gamma_{yy}(\tau) = \gamma_{xx}(\tau)$$

The cross-covariance is

$$\gamma_{xy}(\tau) = \gamma_{xx}(\tau - \xi)$$

The cross-correlation is

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_X\sigma_Y} = \frac{\gamma_{xx}(\tau-\xi)}{\sigma_X\sigma_Y} = \rho_{xx}(\tau-\xi) \tag{10.11}$$

The cross-correlation of $\mathbf{X_t}$ with $\mathbf{Y_t}$ is just the auto-correlation of $\mathbf{X_t}$ but shifted to positive τ values.

See Fig. 10.2. For $\tau > 0$ the time evolution of anomalies in $\mathbf{X_t}$ leads those of $\mathbf{Y_t}$.

- **$\mathbf{Y_t}$ is the derivative of $\mathbf{X_t}$,**

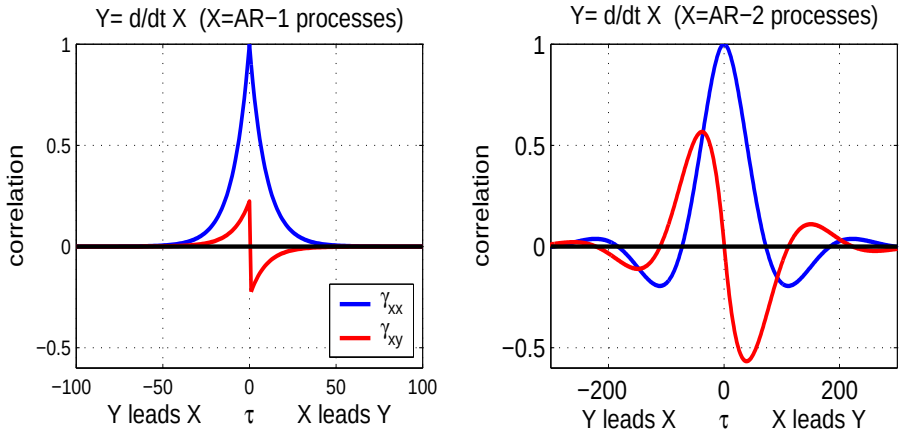


Figure 10.3: The auto/cross-correlation for the process $\mathbf{Y_t} = \mathbf{X_t} - \mathbf{X_{t-1}}$. For $\mathbf{X_t}$ following an AR(1)-process (left) and an AR(2)-process (right). For $\tau > 0$ the time evolution of anomalies in $\mathbf{X_t}$ leads those of $\mathbf{Y_t}$.

$$\mathbf{Y_t} = \mathbf{X_t} - \mathbf{X_{t-1}} \approx \frac{d}{dt}\mathbf{X_t} \tag{10.12}$$

The auto-covariance of $\mathbf{Y_t}$ is:

$$\gamma_{yy}(\tau) = 2\gamma_{xx}(\tau) - (\gamma_{xx}(\tau-1) + \gamma_{xx}(\tau+1)) \approx \frac{d^2}{dt^2}\gamma_{xx}(\tau) \tag{10.13}$$

$\frac{d^2}{dt^2}\gamma_{xx}(\tau) > 0$ for τ near zero.

The cross-covariance is

$$\gamma_{xy}(\tau) = \mathcal{E}(\mathbf{X_t}(\mathbf{X_{t+\tau}} - \mathbf{X_{t-1+\tau}})) = \mathcal{E}(\mathbf{X_t}\mathbf{X_{t+\tau}} - \mathbf{X_t}\mathbf{X_{t-1+\tau}}) = \gamma_{xx}(\tau) - \gamma_{xx}(\tau-1) \approx \frac{d}{dt}\gamma_{xx}(\tau) \tag{10.14}$$

Since $\gamma_{xx}(\tau) <= \gamma_{xx}(0) \Rightarrow$ local maximum at zero.

$\Rightarrow \gamma_{xy}(\tau) \approx \frac{d}{dt}\gamma_{xx}(\tau) < 0$ for τ near zero but $\tau > 0$
$\Rightarrow \gamma_{xy}(\tau) \approx \frac{d}{dt}\gamma_{xx}(\tau) > 0$ for τ near zero but $\tau < 0$.

So since the auto-correlation/covariance peaks at zero, the cross correlation function is changing sign at zero, with positive cross-correlation when the forcing (derivative) leads X_t.

- **$\mathbf{Z_t}$ as the white noise that drives $\mathbf{X_t}$ in an AR(1)-process,**

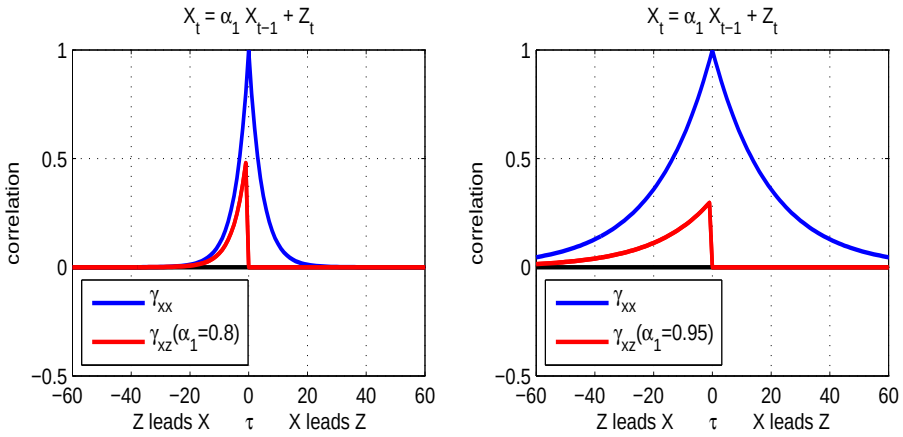


Figure 10.4: The auto/cross-correlation for the process $\mathbf{X_t} = \alpha 1 \mathbf{X_{t-1}} + \mathbf{Z_t}$. For an AR(1)-process with $\alpha_1 = 0.8$ (left) and $\alpha_1 = 0.95$ (right). For $\tau > 0$ the time evolution of anomalies in $\mathbf{X_t}$ leads those of $\mathbf{Z_t}$.

$$\mathbf{X_t} = \alpha 1 \mathbf{X_{t-1}} + \mathbf{Z_t} \tag{10.15}$$

The cross-covariance between the AR(1)-process, $\mathbf{X_t}$ and its driving noise $\mathbf{Z_t}$ is,

$$\gamma_{xz}(\tau) = \alpha_1^\tau \sigma_Z^2 \qquad \forall \quad \tau \leq 0 \tag{10.16}$$

$$\gamma_{xz}(\tau) = 0 \qquad \forall \quad \tau > 0 \tag{10.17}$$

Thus the cross-correlation $\forall \quad \tau \leq 0$ is

$$\rho_{xz}(\tau) = \frac{\gamma_{xz}(\tau)}{\sigma_X \sigma_Z} = \frac{\alpha_1^\tau \sigma_Z^2}{\sigma_X \sigma_Z} \tag{10.18}$$

with $\sigma_X^2 = \frac{\sigma_z^2}{1-\alpha_1^2}$ we find

$$\rho_{xz}(\tau) = \frac{\alpha_1^\tau \sigma_Z^2}{\frac{\sigma_z}{\sqrt{(1-\alpha_1^2)}} \sigma_Z} = \alpha_1^\tau \sqrt{1-\alpha_1^2} \tag{10.19}$$

and $\rho_{xz}(\tau) = 0 \qquad \forall \quad \tau > 0$

10.2 The Cross-Correlation of Cyclo-Stationary Time Series

Note that cyclo-stationary time series, with seasonal changes in the standard deviation for instance, can lead to apparent lag-lead cross-correlation, which are not 'real', see Fig. 10.5.

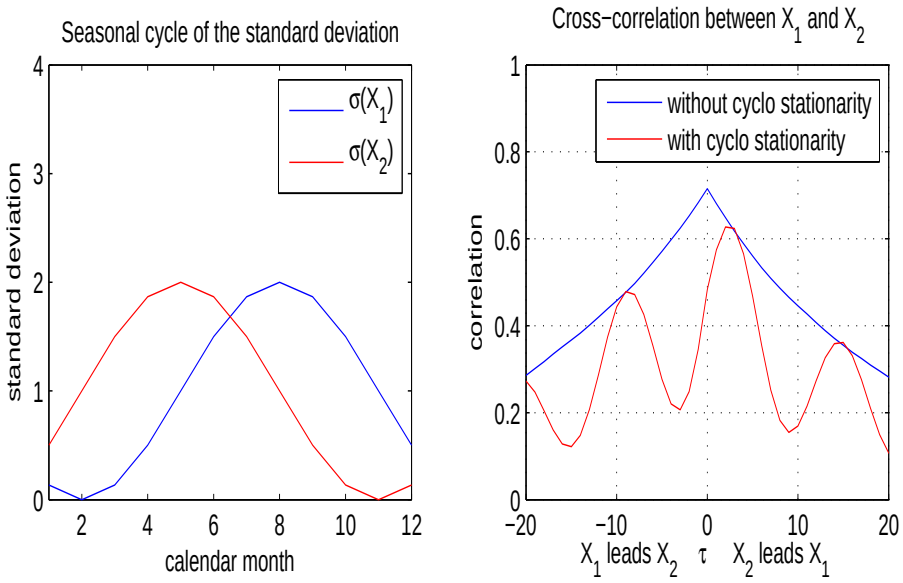


Figure 10.5: Left: The standard deviation of two cyclo stationary AR(1)-process. Right: The Cross-correlation of the two AR(1)-processes with and without cyclo stationarity in the standard deviation.

10.3 The Cross-spectrum

Let $\mathbf{X}_t$ *and* $\mathbf{Y}_t$ *be two weakly stationary stochastic processes with covariance functions* γ_{xx} *and* γ_{yy}*, and a cross-covariance function* γ_{xy}*. Then the cross-spectrum* Γ_{xy} *is defined as the Fourier transform of* γ_{xy}*:*

$$\begin{aligned} \Gamma_{xy}(\omega) &= \mathcal{F}\{\gamma_{xy}\}(\omega) \\ &= \sum_{t=-\infty}^{\infty} \gamma_{xy}(\tau) e^{-2\pi i \tau \omega} \end{aligned} \tag{10.20}$$

for all $\omega \in [-1/2, 1/2]$.
The cross-spectrum is generally a complex-valued function since the cross-covariance function is, in general, neither strictly symmetric nor anti-symmetric.
The cross-spectrum can be represented in a number of ways.

1. The cross-spectrum can be decomposed into its real and imaginary parts as

   $$\Gamma_{xy}(\omega) = \Lambda_{xy}(\omega) - i\Psi_{xy}(\omega).$$

   The real and imaginary parts Λ_{xy} and Ψ_{xy} are called the *co-spectrum* and *quadrature* spectrum. Note that we define the quadrature spectrum as the negative imaginary part of the cross-spectrum, as it is done in the MATLAB routines, but it is contrary to the notation in Storch and Zwiers. In Storch and Zwiers it is defined as the positive imaginary part. This choice is arbitrary, but may cause a great deal of confusion in the definition of the phase, for instance.

2. The cross-spectrum can be written in polar-coordinates as

   $\Gamma_{xy} = A_{xy}(\omega)e^{i\Phi_{xy}(\omega)}$.

   Then A_{xy} and Φ_{xy} are called the *amplitude spectrum* and *phase spectrum* respectively. The amplitude spectrum is given by

   $A_{xy}(\omega) = \left(\Lambda_{xy}(\omega)^2 + \Psi_{xy}(\omega)^2\right)^{1/2}$.

   The phase-spectrum is given in three parts

   $$\Phi_{xy}(\omega) = \tan^{-1}\left(\Psi_{xy}(\omega)/\Lambda_{xy}(\omega)\right) \tag{10.21}$$

   when $\Psi_{xy}(\omega) \neq 0$ and $\Lambda_{xy}(\omega) \neq 0$,

   $$\Phi_{xy}(\omega) = \begin{cases} 0 & \text{if} \quad \Lambda_{xy}(\omega) > 0 \\ \pm\pi & \text{if} \quad \Lambda_{xy}(\omega) < 0 \end{cases} \tag{10.22}$$

   when $\Psi_{xy}(\omega) = 0$, and

   $$\Phi_{xy}(\omega) = \begin{cases} -\pi/2 & \text{if} \quad \Psi_{xy}(\omega) > 0 \\ \pi/2 & \text{if} \quad \Psi_{xy}(\omega) < 0 \end{cases} \tag{10.23}$$

   when $\Lambda_{xy}(\omega) = 0$.

3. The (squared) *coherency spectrum*

   $$\kappa_{xy}(\omega) = \frac{A_{xy}^2(\omega)}{\Gamma_{xx}(\omega)\Gamma_{yy}(\omega)} \tag{10.24}$$

   expresses the amplitude spectrum in dimensionless units. It is formally similar to a conventional (squared) correlation coefficient.

10.4 Presentation of Cross spectra

The cross spectra between a variable $\mathbf{X_t}$ and $\mathbf{Y_t}$ is usually presented by the amplitude of the cross spectra, $A_{xy}(\omega)$, the coherency, $\kappa_{xy}(\omega)$, and the phase, $\Phi_{xy}(\omega)$. $A_{xy}(\omega)$ is plotted in loglog-scale as normal spectra are plotted. It makes sense to compare $A_{xy}(\omega)$ with $\Gamma_{xx}(\omega)$ and $\Gamma_{yy}(\omega)$. However, $A_{xy}(\omega)$ is the least interesting parameter of the cross spectra. The most important is the coherency, $\kappa_{xy}(\omega)$, which one should examine first. If the $\kappa_{xy}(\omega)$ indicates coherence, it is useful to study the phase, $\Phi_{xy}(\omega)$. The phase is usually given in degrees $[0^o, 360^o]$ or $[-180^o, 180^o]$

10.5 Some Simple Theoretical Examples

We described the cross-covariance functions of a number of simple processes in [11.3.3]. We present the cross-spectra of these processes here.

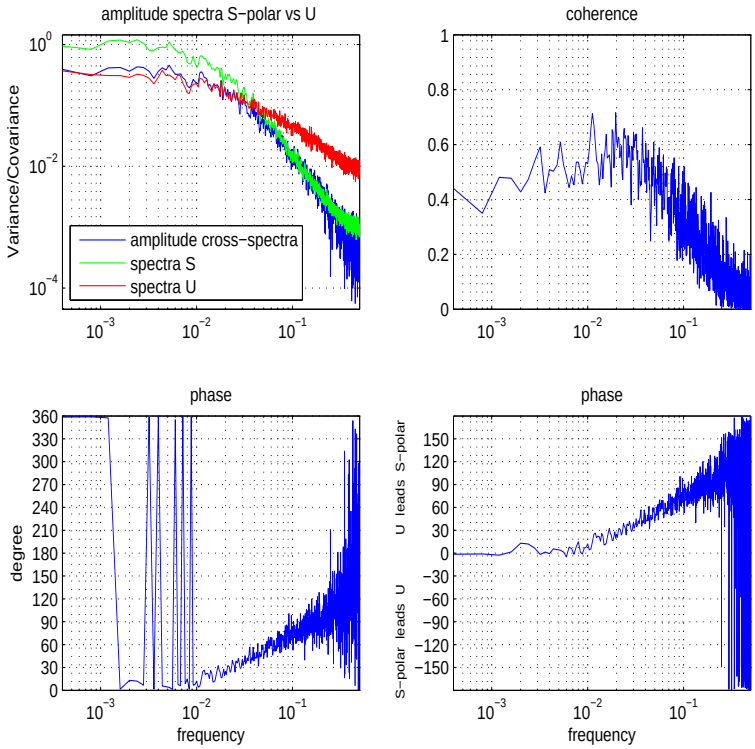


Figure 10.6: Cross-Spectrum of Salt ,S and advection, U from the Stommel model forced with a time series of white noise.

- **$\mathbf{Y_t}$ is no function of $\mathbf{X_t}$:**

$$\mathbf{Y_t} \neq F(\mathbf{X_t}) \tag{10.25}$$

It is obvious what the expected values of the cross-spec parameters are:

$$\begin{aligned} \Gamma_{xy}(\omega) &= 0 \\ \Gamma_{yy}(\omega) &= \Gamma_{yy}(\omega) \\ \Lambda_{xy}(\omega) &= 0 \\ \Psi_{xy}(\omega) &= 0 \end{aligned}$$

$$\begin{aligned} A_{xy}(\omega) &= 0 \\ \Phi_{xy}(\omega) &= 0 \\ \kappa_{xy}(\omega) &= 0 \end{aligned}$$

But! The estimated values from a finite time series will have:

$$\begin{aligned} A^2_{xy}(\omega) &\sim \Gamma_{xx}\Gamma_{yy} \\ \Phi_{xy}(\omega) &\neq 0 \\ \kappa_{xy}(\omega) &> 0 \end{aligned}$$

See Fig. 10.7

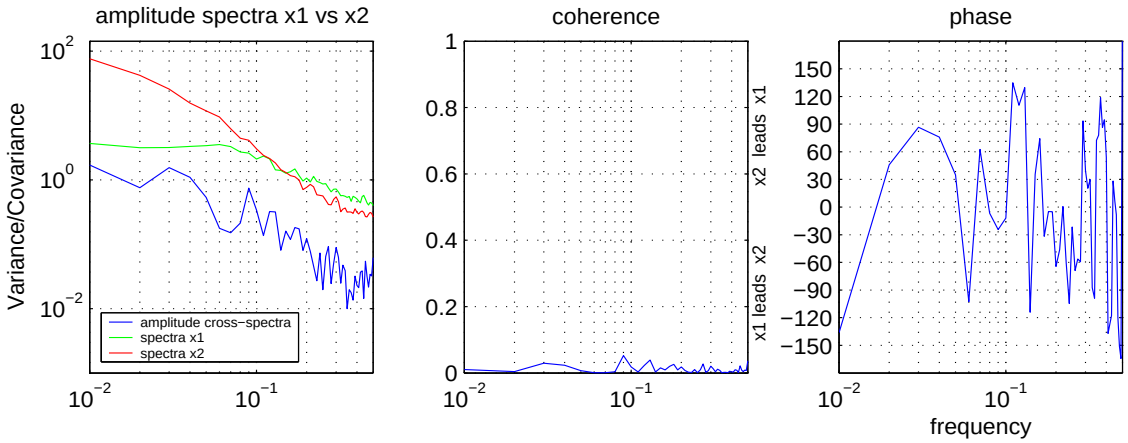


Figure 10.7: The cross-spectrum of two independent AR(1)-processes, $\mathbf{Y_t} \neq F(\mathbf{X_t})$.

- **$\mathbf{Y_t}$ is a linear function of $\mathbf{X_t}$:**

$$\mathbf{Y_t} = a\mathbf{X_t} \tag{10.26}$$

because $\gamma_{\alpha x,x} = \alpha\gamma_{xx}$ (see eq.[**??**]), the cross-spectrum is a simple function of $\mathbf{X}$:

$$\begin{aligned}
\Gamma_{xy}(\omega) &= \alpha\Gamma_{xx}(\omega) \\
\Gamma_{yy}(\omega) &= \alpha^2\Gamma_{xx}(\omega) \\
\Lambda_{xy}(\omega) &= \alpha\Gamma_{xx}(\omega)
\end{aligned}$$

$$\begin{aligned}
A_{xy}(\omega) &= \alpha\Gamma_{xx}(\omega) \\
\Phi_{xy}(\omega) &= \begin{cases} 0 & \text{if} \quad \alpha > 0 \\ \pm\pi & \text{if} \quad \alpha < 0 \end{cases} \\
\kappa_{xy}(\omega) &= 1.
\end{aligned}$$

These are intuitively reasonable results. All events in the two time series occur synchronously, thus the phase spectrum is zero everywhere and the coherency spectrum is one for all ω.

- **$\mathbf{Y_t}$ is a linear function of $\mathbf{X_t}$, but some additional white noise is added.**

$$\mathbf{Y}_t = \alpha\mathbf{X}_t + \mathbf{Z}_t$$

The cross-, co-, quadrature, amplitude, and phase spectra are unaffected by the added noise. However, the power spectrum of $\mathbf{Y}$, and therefore the coherency spectrum, do change. Specially,

$$\begin{aligned}
\Gamma_{yy}(\omega) &= \alpha^2\Gamma_{xx}(\omega) + \sigma_Z^2 \\
A_{xy}(\omega) &= \alpha\Gamma_{xx}(\omega) \\
\kappa_{xy}(\omega) &= \frac{\alpha^2\Gamma_{xx}(\omega)}{\sigma_z^2 + \alpha^2\Gamma_{xx}(\omega)} < 1.
\end{aligned}$$

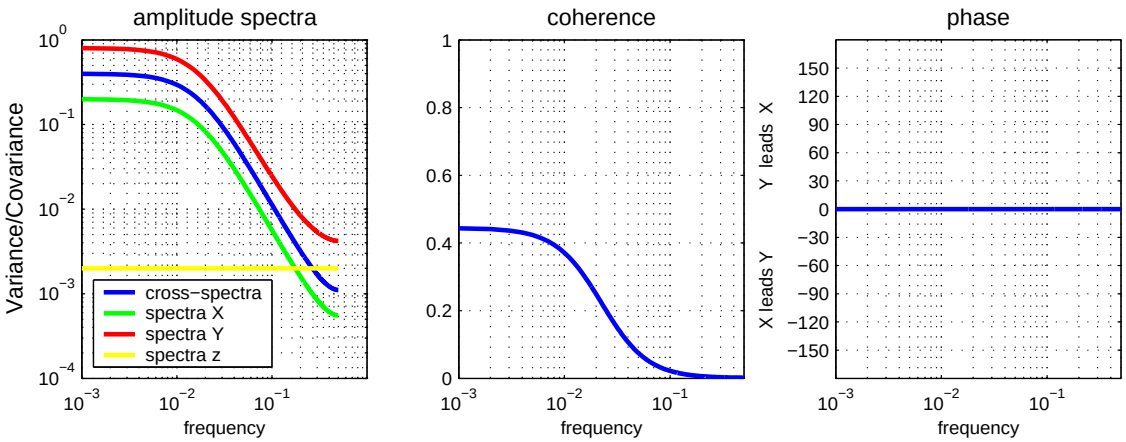


Figure 10.8: The cross-spectrum between $\mathbf{X}_t$ and $\mathbf{Y}_t$ for the process $\mathbf{Y}_t = \alpha\mathbf{X}_t + \mathbf{Z}_t$.

The coherency is now less than 1 at all time scales, indicating that knowledge of the sequence of the events is $\mathbf{X}$ is no longer enough to completely specify the sequence of events in $\mathbf{Y}$. The impact of the noise is small if its variance is small relative to that of $\alpha\mathbf{X}_t$ (and vice versa).

- **$\mathbf{Y_t}$ is a lagged time series of $\mathbf{X_t}$,**

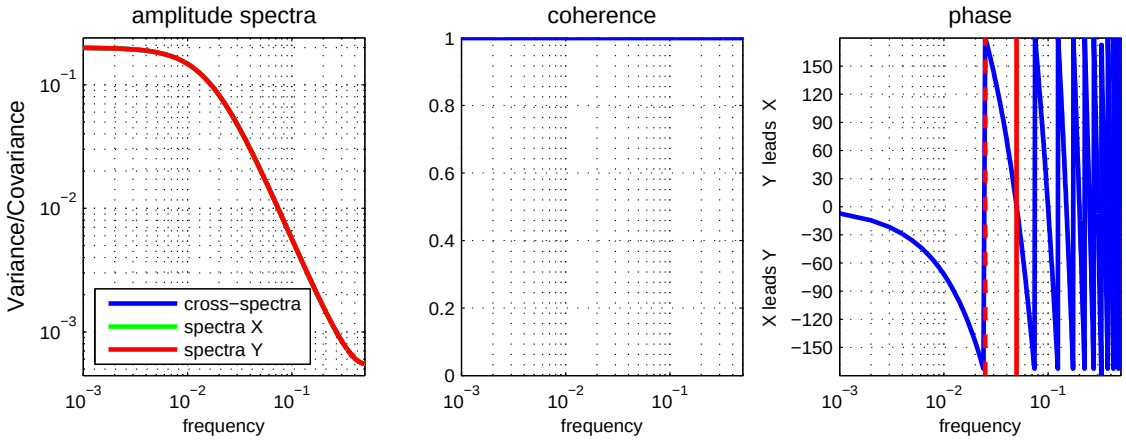


Figure 10.9: The cross-spectrum between $\mathbf{X}_t$ and $\mathbf{Y}_t$ for the process $\mathbf{Y}_t = \mathbf{X}_{t-\zeta}$, with $\zeta = 20$. The solid red line in the phase plot (right) indicates $1/\zeta$ and the dashed red line $1/2\zeta$

$$\mathbf{Y}_t = \mathbf{X}_{t-\zeta}$$

we find that

$$\begin{aligned}
\Gamma_{xy}(\omega) &= e^{i2\pi\zeta\omega}\Gamma_{xx}(\omega) \\
\Gamma_{yy}(\omega) &= \Gamma_{xx}(\omega) \\
\Lambda_{xy}(\omega) &= \cos(2\pi\zeta\omega)\Gamma_{xx}(\omega)
\end{aligned}$$

$$\begin{aligned}\Psi_{xy}(\omega) &= -\sin(2\pi\zeta\omega)\Gamma_{xx}(\omega)\\ \\ A_{xy}(\omega) &= \Gamma_{xx}(\omega)\\ \Phi_{xy}(\omega) &= -2\pi\zeta\omega\\ \kappa_{xy}(\omega) &= 1.\end{aligned}$$

When we shift $\mathbf{X}_t$ a fixed number of lags we obtain the same coherency spectrum as when $\mathbf{X}_t$ is simply scaled. It is 1 for all time scales meaning that the sequence of events in $\mathbf{Y}$ is completely determined by $\mathbf{X}$ In contrast, the phase spectrum has changed from being zero for all ω to a linear function of ω. This type of linear dependency is characteristic of shifts that are independent of the time scale.

Note that if the process $\mathbf{X}$ *lags*the process $\mathbf{Y}$ (i.e., if $\zeta > 0$), then the phase spectrum Φ_{xy} is *positive* for positive frequencies.

- **$\mathbf{Y_t}$ is the derivative of $\mathbf{X_t}$,**

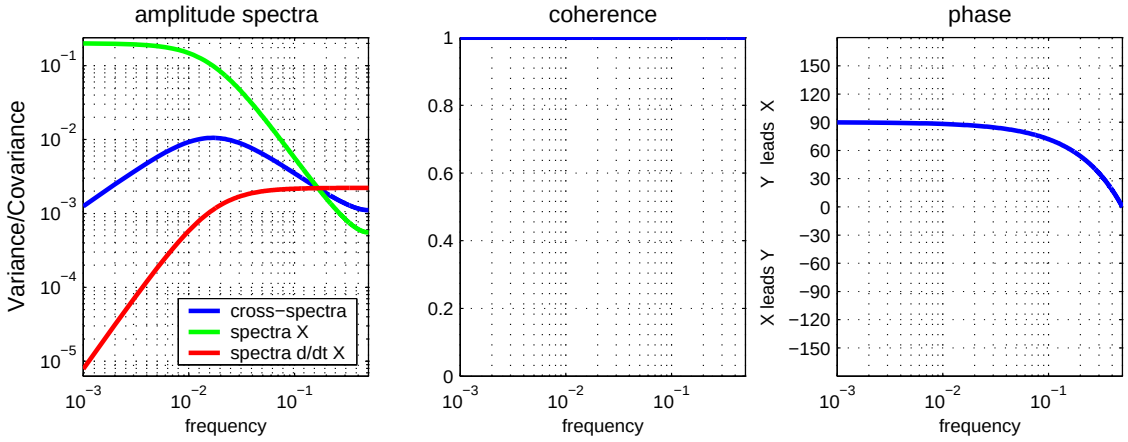


Figure 10.10: The cross-spectrum between $\mathbf{X}_t$ and $\mathbf{Y}_t$ for the process $\mathbf{Y}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$.

$$\mathbf{Y}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$$

that approximates a discretized time derivative. Recall that

$$\begin{aligned}\gamma_{xy}(\tau) &= \gamma_{xx}(\tau) - \gamma_{xx}(\tau - 1)\\ \gamma_{yy}(\tau) &= 2\gamma_{xx}(\tau)\\ &\quad -(\gamma_{xx}(\tau - 1) + \gamma_{xx}(\tau + 1).\end{aligned}$$

Thus, again using (C.8),

$$\begin{aligned}\Gamma_{xy}(\omega) &= (1 - e^{-2\pi i\omega}\Gamma_{xx}(\omega)\\ \Gamma_{yy}(\omega) &= 2(1 - \cos(2\pi\omega))\Gamma_{xx}(\omega)\end{aligned}$$

$$\begin{aligned}
\Lambda_{xy}(\omega) &= (1-\cos(2\pi\omega))\Gamma_{xx}(\omega) \\
\Psi_{xy}(\omega) &= -\sin(2\pi\omega)\Gamma_{xx}(\omega)
\end{aligned}$$

$$\begin{aligned}
\Gamma_{yy}(\omega) &= 2(1-\cos(2\pi\omega))\Gamma_{xx}(\omega) \\
A^2_{xy}(\omega) &= 2(1-\cos(2\pi\omega))\Gamma_{xx}(\omega)^2 \\
&= \Gamma_{xx}(\omega)\Gamma_{yy}(\omega) \\
\Phi_{xy}(\omega) &= \tan^{-1}\left(\frac{-\sin(2\pi\omega)}{1-\cos(2\pi\omega)}\right) \\
&= \tan^{-1}(-\cot(\pi\omega)) \\
&= \pi\left(\frac{1}{2}-\omega\right) \geq 0 \quad \text{for} \quad \omega \geq 0 \\
\kappa_{xy} &= 1 \quad \text{for} \quad \omega \neq 0.
\end{aligned}$$

Several things can be noted here.

i) The coherancy is 1 at all time scales except 0. This is reasonable since integration can undo differentiation up to a constant.

ii) The spectrum of the differences process $\mathbf{Y}$ has more short time scale variability than the spectrum of the original process $\mathbf{X}$. Indeed, subtracting acts as a high-pass filter that dampens long time scale variability and eliminates the time mean($\Gamma_{yy}(0) = 0$). For example, Figure 11.9 displays the spectrum of an AR(1) process $\mathbf{X}_t$ with $\alpha = 0.3$ and that of the differences process $\mathbf{Y}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$. The $\mathbf{X}$-spectrum is 'red with a maximum at zero frequency whereas the $\mathbf{Y}$-spectrum is 'blue' with a maximum at frequency 1/2.

iii) 'Physical reasoning' suggests that the forcing should lead the response[1] in the sense that the phase lag Φ_{yx} between the 'forcing' $\mathbf{Y}$ and the 'response' $\mathbf{X}$ is $\pi/2$. This is approximately the case for the long time scales near $\omega = 0$, since $\Phi_{xy}(0) = -\pi/2$. The phase converges towards zero on shorter time scales. This ect occurs because the time derivative is only approximated by the time difference, and the accuracy of this approximation increases with the time scale.

- **$\mathbf{Z_t}$ as the white noise that drives $\mathbf{X_t}$ in an AR(1)-process,**

$$\mathbf{X_t} = \alpha 1 \mathbf{X_{t-1}} + \mathbf{Z_t} \tag{10.27}$$

Unfortunately I could not derive all ellements of the cross spectra analytically, but here the incomplete list:

$$\begin{aligned}
\Gamma_{xz}(\omega) &= \textit{unknown}!!! \\
\Gamma_{zz}(\omega) &= \sigma_z^2 \\
\Lambda_{xz}(\omega) &= \textit{unknown}!!!
\end{aligned}$$

[1] The 'physical' argument is as follows. Suppose $dX/dt = Y$ where $Y = A\cos(\omega t)$. Then $X = A/\omega\cos(\omega t + \Phi_{xy})$ where $\Phi_{xy} = -\frac{\pi}{2}$.

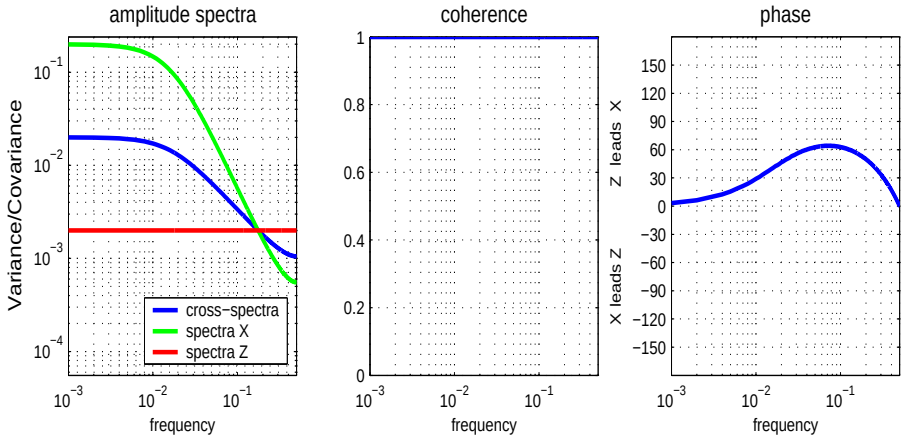


Figure 10.11: The cross-spectrum between $\mathbf{X}_t$ and $\mathbf{Z}_t$ for the process $\mathbf{X_t} = \alpha 1 \mathbf{X_{t-1}} + \mathbf{Z}_t$.

$$\begin{aligned} \Psi_{xz}(\omega) &= unknown!!! \\ \\ A^2_{xz}(\omega) &= \Gamma_{xx}(\omega)\Gamma_{zz}(\omega) = \sigma_z^2 \Gamma_{xx} \\ \Phi_{xz}(\omega) &= unknown!!! \\ \kappa_{xy} &= 1 \end{aligned}$$

However, we can estimate the missing parts with a time series realization of an AR(1)-process.

10.6 Some Examples with Climate Observations

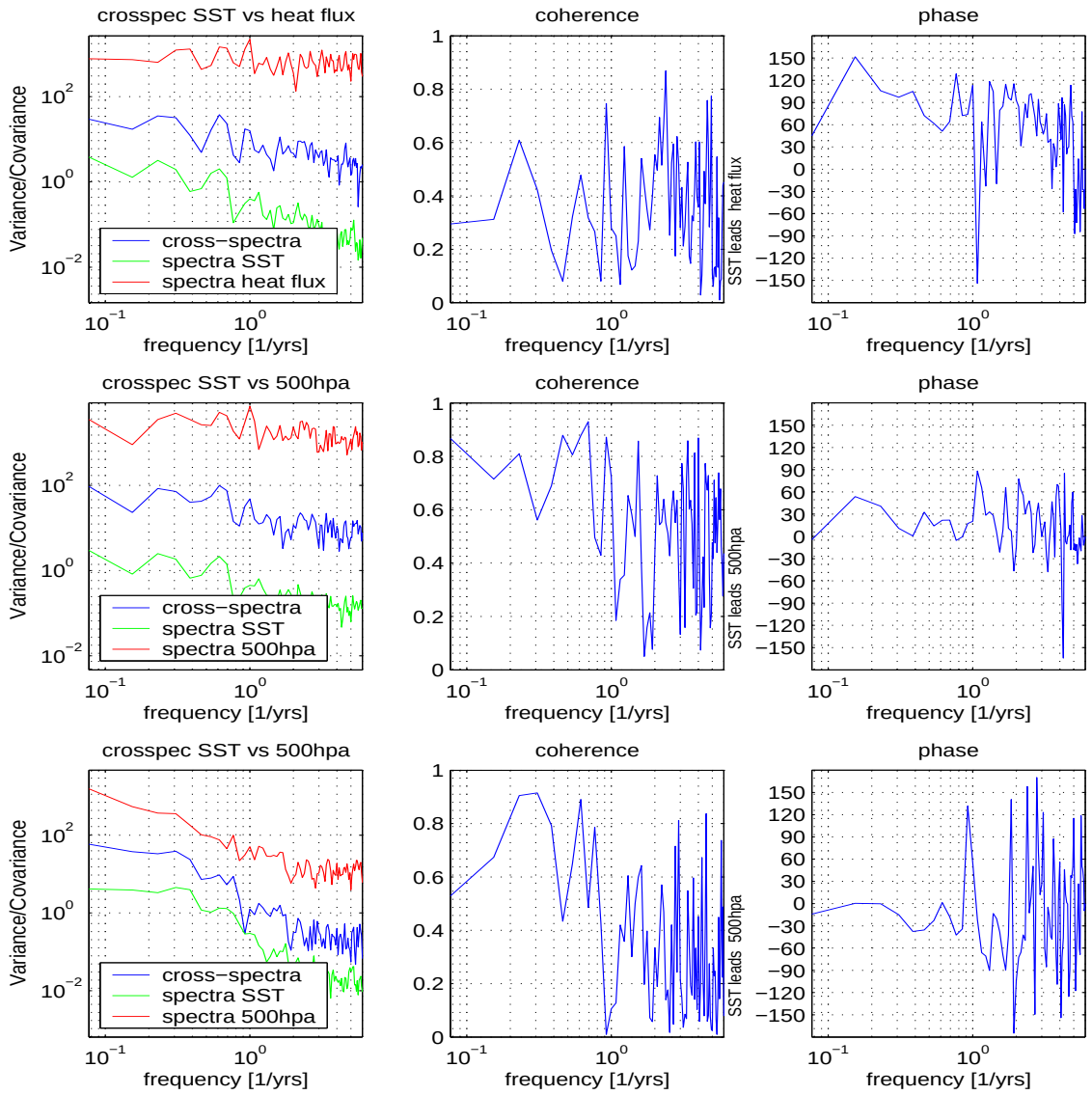
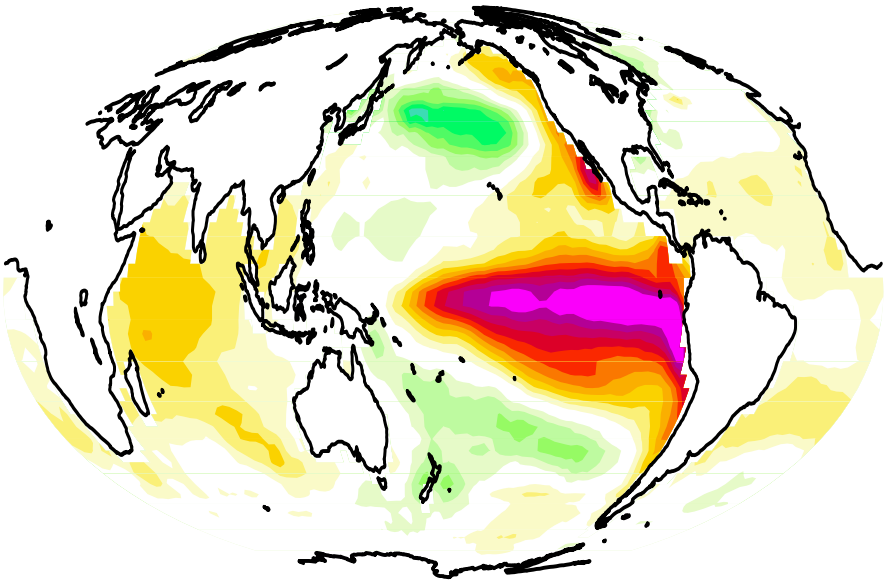


Figure 10.12: Cross-spectra of observed monthly mean time series. Upper: Between SST and heat flux in the northern central Pacific. Middle: Between SST and 500hPa geopotential height in the northern central Pacific. Lower: Between SST and 500hPa geopotential height in the equatorial east Pacific.

Part III

Multivariate Data Analysis



In Multivariate Data Analysis we are interested in the covariability of multivariate data, which could be field of a physical quantity (e.g. global temperature) or it could be fields of many physical quantities.
In time series analysis we analyzed the variance and the covariance in terms of the auto/cross-correlation function or spectra. In Multivariate Data Analysis we continue to analyze the auto/cross-correlation function, namely the covariance matrix.
In time series analysis we were concerned with peaks in the spectra, indicating some interesting oscillating variability, which may indicate predictable climate variability. We need to note that time scale variability is a continuous spectra, which even in the absence of peaks can have different characteristics. In multivariate data analysis the analog to the spectra are the principal components. Here we are interested in dominant spatial modes of variability.
As for the spectra of time series, the multivariate pattern-modes are a continuous series of modes, we will see that it is difficult to define something that is similar to a peak in the spectrum, a dominant coherent mode.
This part starts with some basic linear algebra, which is followed by the presentation of the principal component (also named Empirical Orthogonal Function (EOF)) analysis, which is the fundament of multivariate data analysis. Some other and alternative methods will be discussed shortly. Some discussion of the interpretation of the EOFs will be discussed and a null hypothesis for the EOF-modes will be presented.

Chapter 11

Principal Component Analysis (Empirical Orthogonal Functions (EOFs))

11.1 Basics

Much of the Multivariate Data Analysis is linear algebra; Vector space analysis. Some important basic elements shall be discussed first:

Data Sample Vectors: As in the fundamentals section introduced, a time series of a continues random Variable can be written as a vector:

$$\hat{\mathbf{X}} = \{x_{t=1}, x_{t=2}, ..., x_{t=T}\} = \vec{X} \tag{11.1}$$

The vector $\vec{X}$ has the dimension T. It may be good to note that the dimension T means that none of the x_t can be expressed by the other components of $\vec{X}$.

Using this vector notation we find that the covariance between two sampling vectors $\vec{X}$ and $\vec{Y}$ is

$$\widehat{\sigma^2_{xy}} = \frac{1}{n-1} < \vec{X}|\vec{Y}^T > \equiv \frac{1}{n-1} X \cdot Y' \tag{11.2}$$

we skip the vector sign if is obvious that X and Y are vectors. The correlation:

$$\widehat{\chi}_{xy} = \frac{X \cdot Y'}{\sqrt{X \cdot X' Y \cdot Y'}} \tag{11.3}$$

The Data Sample Matrix: In Multivariate Data Analysis we analyze continues random vectors or fields, χ. These can be represented in terms of a data matrix, $\mathbf{D}$. We assume here that the continues random vectors or fields, , χ is sampled at different times and at different locations, in which we assume that , χ is sampled at each location at the same time. We can than represent χ as a martix $\mathbf{D}$:

$$\mathbf{D} = \begin{pmatrix} \chi_{x=1,t=1} & \chi_{x=1,t=2} & \chi_{x=3,t=1} & \cdots & \chi_{x=S,t=1} \\ \chi_{x=1,t=2} & \chi_{x=2,t=2} & & & \\ \chi_{x=1,t=3} & & \chi_{x=3,t=3} & & \\ \vdots & & & \ddots & \\ \chi_{x=1,t=T} & & & & \chi_{x=S,t=T} \end{pmatrix} \tag{11.4}$$

where each row is the data field at one time and each column represents the time series of one data point. The index x refers to an index in the spatial dimension which can include the x,y and z dimension at the same time, so that the row vector can represent a three dimensional spatial field. The data matrix $\mathbf{D}$ is a $T \times S$ matrix. A global climate data set is in the order of $S = 360 \cdot 180 \cdot 20 \approx 10^6$ and $T = 100 * 12 \approx 10^3$, thus $\mathbf{D}$ has about 10^9 (Giga) elements.

Following this notation we can write the estimated auto-covariance matrix of χ as sampled by $\mathbf{D}$ as:

$$\widehat{\Sigma_{\chi\chi}} = \frac{1}{T-1}\mathbf{D}' \cdot \mathbf{D} \tag{11.5}$$

The auto-covariance matrix $\widehat{\Sigma_{\chi\chi}}$ is a $S \times S$ matrix. The diagonal elements of the auto-covariance matrix $\widehat{\Sigma_{\chi\chi}}$, are the variances of the time series of each spatial point. The other elements are the covariance between two spatial points.
We can also estimate the covariance or correlation of a time series, $\vec{\psi}$ with the data matrix:

$$\widehat{\gamma_{\psi\chi}} = \frac{1}{T-1}\psi \cdot \mathbf{D} \tag{11.6}$$

Here the result is a covariance vector or field, $\widehat{\gamma_{\psi\chi}}$. It represents the spatial pattern that is associated if the time series $\vec{\psi}$.
We can also estimate the covariance between a spatial pattern, represented by a vector π of dimension S, and the data matrix $\mathbf{D}$:

$$\widehat{\gamma_{\pi\chi}} = \frac{1}{S-1}\pi \cdot \mathbf{D}' \tag{11.7}$$

Here the result is a covariance vector, $\widehat{\gamma_{\pi\chi}}$ of dimension T, representing a time series. The time series of the spatial pattern π for the data set $\mathbf{D}$.

11.2 Estimation of the Principal Components

Deconstruct the data matrix $\mathbf{D}$ into a more efficient structure. $\mathbf{D}$ has usually highly redondant data, that means that most of the variance in $\mathbf{D}$ can be represented with much less data. A single column of $\mathbf{D}$, for instance, is usually highly correlated to the neighboring column. The Principal Component Analysis finds the most efficient representation of the data, which represents the largest amount of variance with the least number of data points.

The Principal Component Analysis, can also be interpreted as a more efficient way to analyze the covariance matrix.

An efficient way to represent the data is to deconstruct the data matrix $\mathbf{D}$ into a number of patterns, where each pattern has an associated time series.

$$\mathbf{D}(T \times S) = \Psi(T \times N) \cdot \Pi^T(N \times S) \tag{11.8}$$

The matrix Π is a set of N spatial pattern, π_i, with the associated matrix Ψ, of N time series, ψ_i. Since the total number of data points in $\mathbf{D}$ is $S \cdot T$, we find for $N = \frac{S \cdot T}{S+T}$. If one of S or T is much larger than the other N will converge towards the smaller one.
We can estimate the total variance that a single pair of a pattern, π_i with the time series, ψ_i explains in $\mathbf{D}$:

$$\lambda_i = (\psi_i \cdot \mathbf{D})'(\psi_i \cdot \mathbf{D}) \tag{11.9}$$

The explained variances are λ_i scalar values. It makes sense to further deconstruct the data set, by separating the variance, λ_i, from the pair of pattern, π_i and time series, ψ_i:

$$\mathbf{D} = \Psi \cdot \Lambda \cdot \Pi \tag{11.10}$$
$$\mathbf{D}(T \times S) = \Psi(T \times N) \cdot \Lambda(N \times N) \cdot \Pi(N \times S)$$

where Λ is a diagonal matrix with the diagonal elements $\sqrt{\lambda_i}$. So we have deconstructed our data into a set of modes where each mode consists of three parts: a spatial pattern, a time series of the evolution of this spatial pattern and an explained variance value for this mode. We can reconstruct our data field at any time step with the linear combination of the modes,

$$\mathbf{D}(t) = \sum_{i=1}^{N} \psi_i(t) \cdot \sqrt{\lambda_i} \cdot \pi_i \tag{11.11}$$

In the literature we can find a nearly infinite number of methods on how to estimate this modes of variability. The by far most widely used method is the Principal Component Analysis as described in the following. Although, this method may from the statistical/ mathematical point of view be the most attractive one, it is from a physical point of view often of little interest.
For defining the modes it seems efficient to have a representation of the data where the time series of each pattern is uncorrelated with the time series of all other patterns. This is not the case in many empirical deconstructions of data sets such as in cluster analysis (weather-regimes, Grosswetterlagen).
Further it may seem efficient to have a representation of the data in which the modes are ordered by the amount of total variance, λ_i the modes explain. Thus we have a representation of the data in which the leading modes explain the largest amount of variance. We can therefore often simplify the data by just analyzing a few leading modes.
In the Principal component analysis we seek for the one and only mode which explains the largest amount of variance. This mode is by definition our mode-1. The next mode shall have an uncorrelated time series with the mode-1 and it shall be the mode which explains the largest amount of variance of the residual of $\mathbf{D}$ if mode-1 is subtracted from $\mathbf{D}$. This structure shall be used for all following modes.
This ansatz for our modes leads to a maximization problem, which leads to the eigenvalues of the covariance matrix Σ:

$$(\Sigma - \lambda_1 \cdot \mathbf{I_S}) \cdot \pi_1 = 0 \tag{11.12}$$

with Σ the covariance matrix of $\mathbf{D}$, λ_1 the eigenvalue, $\mathbf{I_S}$ a $S \times S$ identity matrix, and π_1 the eigenvector. The eigenvalues λ_i are the variances of the eigenvectors π_i. We order the eigenvalues, with λ_1 being the largest eigenvalue and so on. The eigenvectors, the patterns, π_i are orthogonal to each other, meaning they are uncorrelated:

$$\pi_i' \cdot \pi_j = 0 \quad \forall i \neq j \tag{11.13}$$

We find the time evolution of π_i, ψ_i, by projecting π_i onto the data matrix $\mathbf{D}$, see eq. [11.7]. As for the eigenvectors, the ψ_i are orthogonal to each other, meaning they are uncorrelated:

$$\psi_i \cdot \psi_j' = 0 \quad \forall i \neq j \tag{11.14}$$

The time evolution of ψ_i is usually refered to as the Principal component. The eigenvalue is usually referred to as the explained variance of the mode or just the order number. The eigenvector is often called the pattern, or EOF-pattern.

11.3 A Simple Example

A Simple Example shall illustrate how the Principal components are calculated. We use a two dimensional problem, since it is the simplest of all possible problems. For this we define our data set as the time series of SLP at the Azores, point 1 and at Iceland as point 2. Both time series have the length $T = 486(month)$ Thus we have a data matrix $\mathbf{D}(2 \times 486)$. The number of patterns we can define is therefore $N = \frac{S \cdot T}{S+T} \approx 2$.
The covariance matrix for this problem is:

$$\Sigma = \begin{pmatrix} 10.2 & -10.5 \\ -10.5 & 32.5 \end{pmatrix} \tag{11.15}$$

the diagonal elements are the variances of the time series. We can see that the variance at Iceland (point 2) is much larger than at the Azores. The cross-covariance between the Azores and Iceland is negative indicating that both are anti-correlated, which is well known.
The eigenvalue problem is:

$$det(\Sigma - \lambda \cdot \mathbf{I_S}) = 0 \tag{11.16}$$

$$\Rightarrow det \begin{pmatrix} 10.2 - \lambda & -10.5 \\ -10.5 & 32.5 - \lambda \end{pmatrix} = (10.2 - \lambda)(32.5 - \lambda) - 10.5^2 = 0 \tag{11.17}$$

$$\Rightarrow \lambda_1 = 36.7 \quad \lambda_2 = 6.0 \tag{11.18}$$

The eigenvectors are given by

$$(\Sigma - \lambda_i \cdot \mathbf{I_S}) \cdot \pi_i = 0 \tag{11.19}$$

which leaves a freedom of amplitude which we chose to be 1

$$\pi_1 = \begin{pmatrix} -0.37 \\ 0.93 \end{pmatrix} \quad \pi_2 = \begin{pmatrix} 0.93 \\ 0.37 \end{pmatrix} \tag{11.20}$$

The structure of the 2 dimensional problem is illustrated in Fig. 11.1. we see that the EOF find the vectors that point into the direction of the main axis of the 2 dimensional ellipsoid.

11.4 The Effective Spatial Number Degrees of Freedom

The eigenvalues of the Principal component analysis are ordered so that they decrease with increasing order number. The characteristics of the decrease in eigenvalues gives some insight into the Effective Spatial Number Degrees of Freedom. Bretherton et al .(1999) found that the Effective Spatial Number Degrees of Freedom is given by a simple sum of the eigenvalues:

$$N_{spatial} = \frac{1}{\sum_{i=1}^{N} \lambda_i^2} \tag{11.21}$$

with the eigenvalues normalized to $\sum_{i=1}^{N} \lambda_i = 1$. This equation can somehow intuitively be understood. Assume that we have just two eigenvalues with each explaining 1/2 of the total variance we would expect the Spatial Number Degrees of Freedom to be two, as it is in eq.[11.21]. Thus we have a simple way to estimate how many modes of variability we need to represent the data set $\mathbf{D}$. it should be in the order of $N_{spatial}$. See Fig. 11.2 for an illustration of $N_{spatial}$ in different sets of eigenvalues.

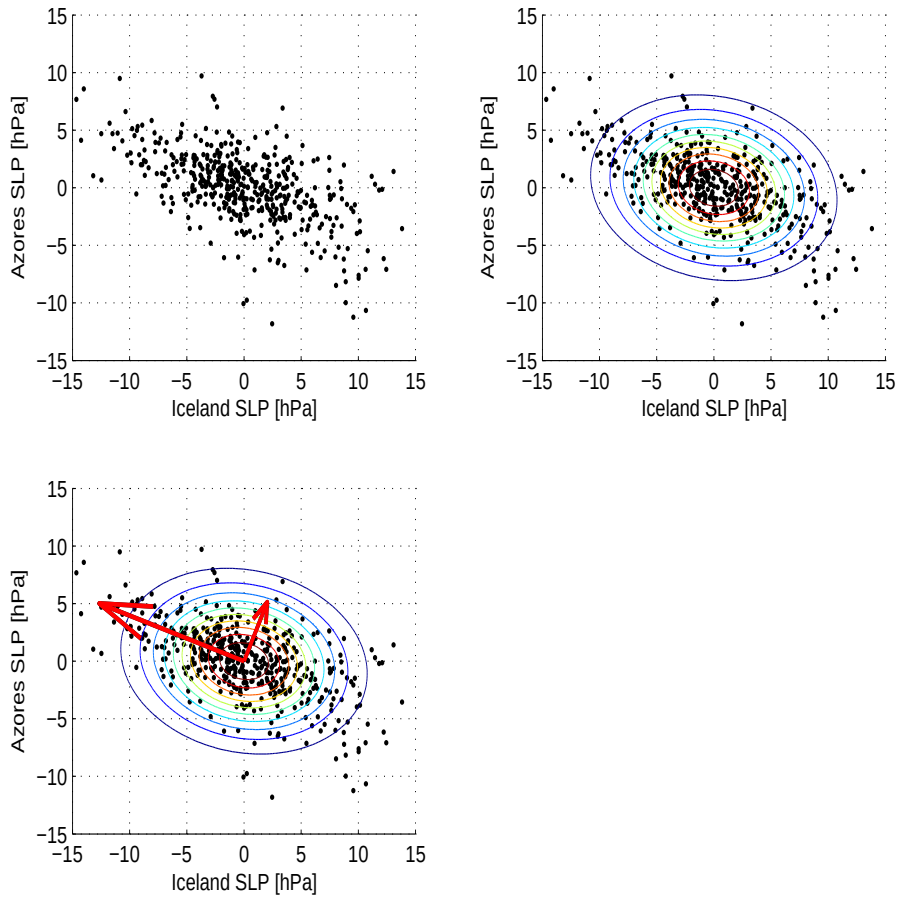


Figure 11.1: The scattered dots are the values of the 2 dimensional SLP field with the x-axis representing Iceland and the y-axis representing Azores. The contour line represented a fitted 2 dimensional normal *pdf*. The vectors are the eigenvectors (EOF-modes) scaled by $2.5\sqrt{(}\lambda_i)$.

11.5 Presentation of EOFs

The EOF-modes have three components, the eigenvalues, the eigenvectors (EOF-patterns) and the PC time series. The eigenvectors (EOF-patterns) are usually of primary interest since they hold the information on how the data set is spatially organized. We have essentially three different ways of presenting a spatial pattern associated with an EOF-mode:

- $\sqrt{\lambda_i}\pi_i$: We simply present the eigenvector as a spatial pattern. However, the eigenvector has no dimension and it may be instructive to present the eigenvector scaled by $\sqrt{\lambda_i}$. The EOF-pattern amplitudes are now in values of the field itself and can be compared more easily with anomalies of the data field. See example in Fig. 11.5.

- $\rho_{\pi_i \mathbf{D}}$: The correlation field of the ψ_i with the data set is a presentation of the EOF-mode which can be quite different from the eigenvector itself if the variance field of the data set is very inhomogenuous. See example in Fig. 11.6.

- $\rho^2_{\pi_i \mathbf{D}}$: An EOF-mode that has an eigenvalue of 20%, for instance does not explain 20% at all points of the domain, but only where the amplitude of the eigenvector is larger compared to the standard deviation. The squared correlation field is the explained variance field of the

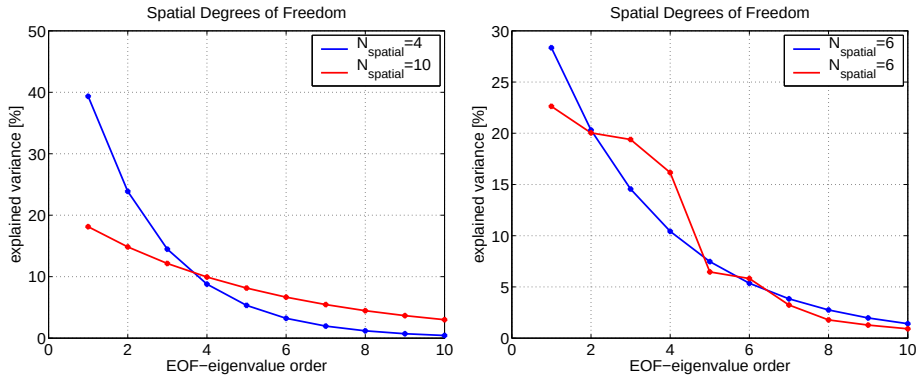


Figure 11.2: An illustration of $N_{spatial}$ in different sets of eigenvalues.

EOF-mode. It essentially puts the eigenvalue of the EOF-mode into a local perspective. See example in Fig. 11.7.

The eigenvalues are usually presented in relation to the total variance, as percentages. It sometimes is useful to presented the integrated eigenvalues against the eigenvalue number to find the number of eigenvalues needed to explain a certain amount of variance.
The principal component time series, ψ_i are analyzed as discussed in the time series analysis chapter.

11.6 Examples

The winter time monthly mean SLP of the Northern Hemisphere is a good example to demonstrate that the different presentations of the leading EOF-modes can lead to different interpretations, Figures 11.5, 11.6 and 11.7. The spatial domain has 4176 points and the time series is 227 month. Therefore the vector space has the dimension $N = 227$.

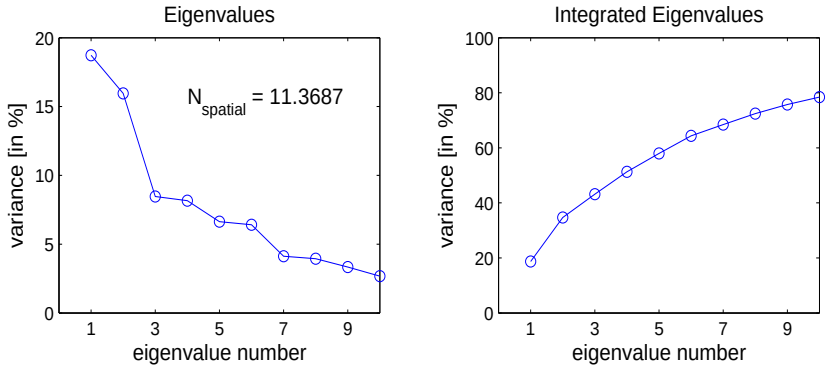


Figure 11.3: The leading eigenvalues of winter time monthly mean SLP of the Northern Hemisphere

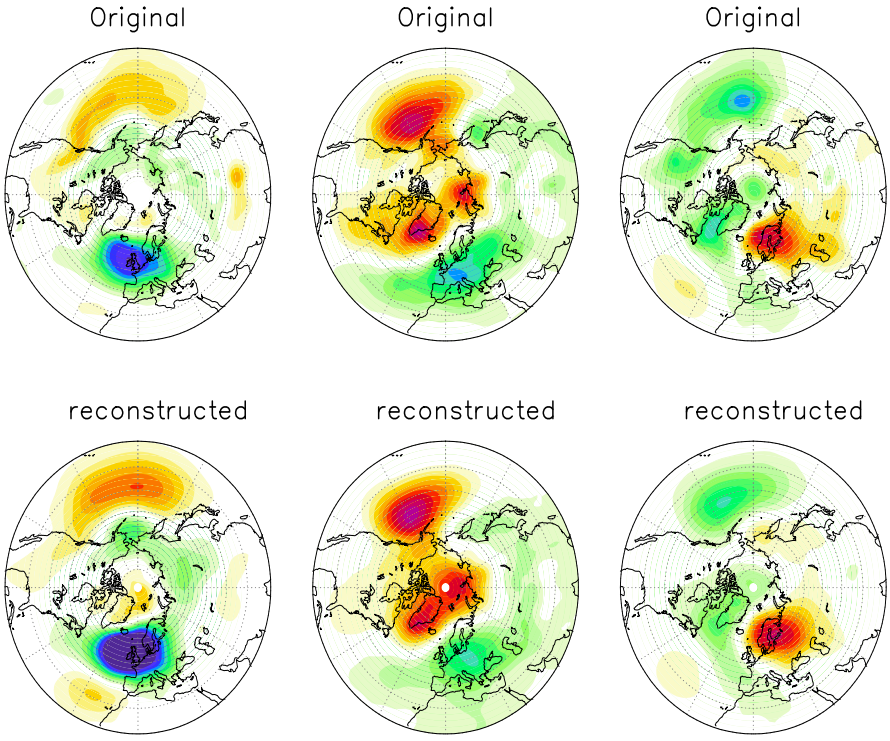


Figure 11.4: Comprison of original fields at three random time steps (upper) and a reconstrution by 10 EOF-modes of winter time monthly mean SLP of the Northern Hemisphere. In units of hPa.

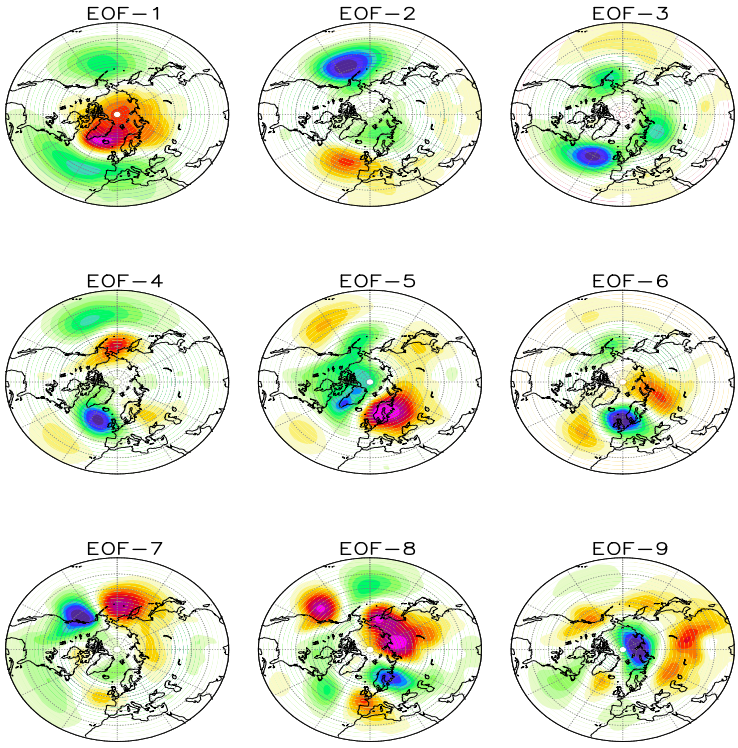


Figure 11.5: The leading eigenvectors (EOF-modes) of winter time monthly mean SLP of the Northern Hemisphere, scaled by $\sqrt{\lambda_i}$. In units of hPa.

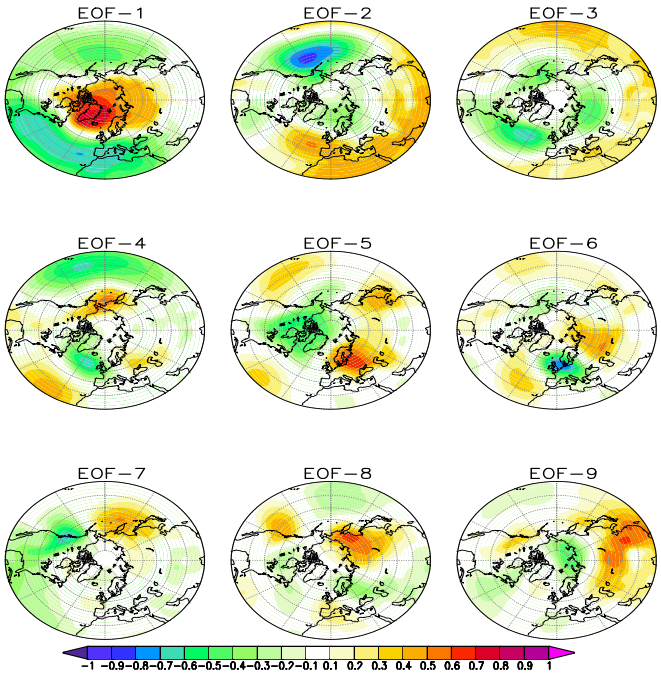


Figure 11.6: The correlation fields of the leading EOF-modes of winter time monthly mean SLP of the Northern Hemisphere.

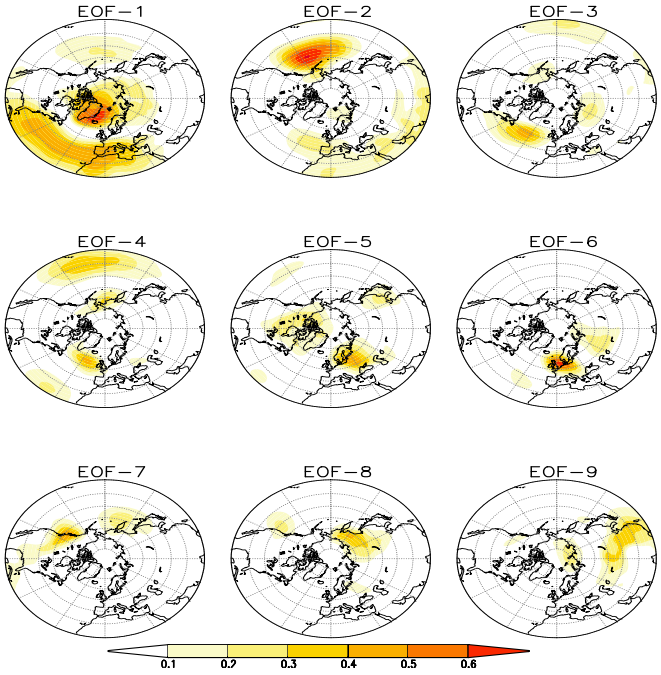


Figure 11.7: The explained variance fields of the leading EOF-modes of winter time monthly mean SLP of the Northern Hemisphere. In relative values

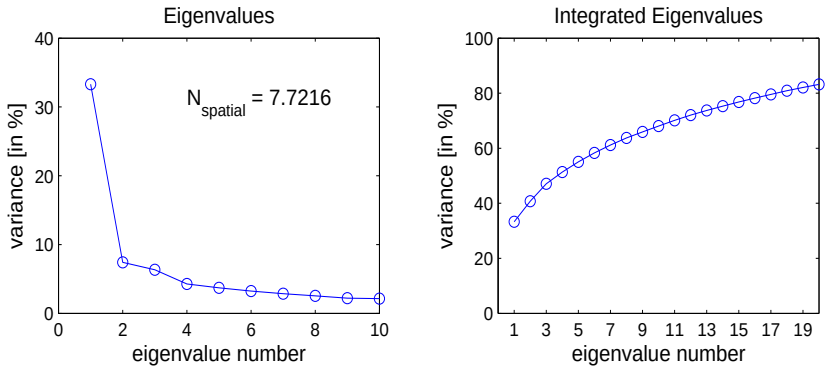


Figure 11.8: The leading eigenvalues of monthly mean SST in the Pacific.

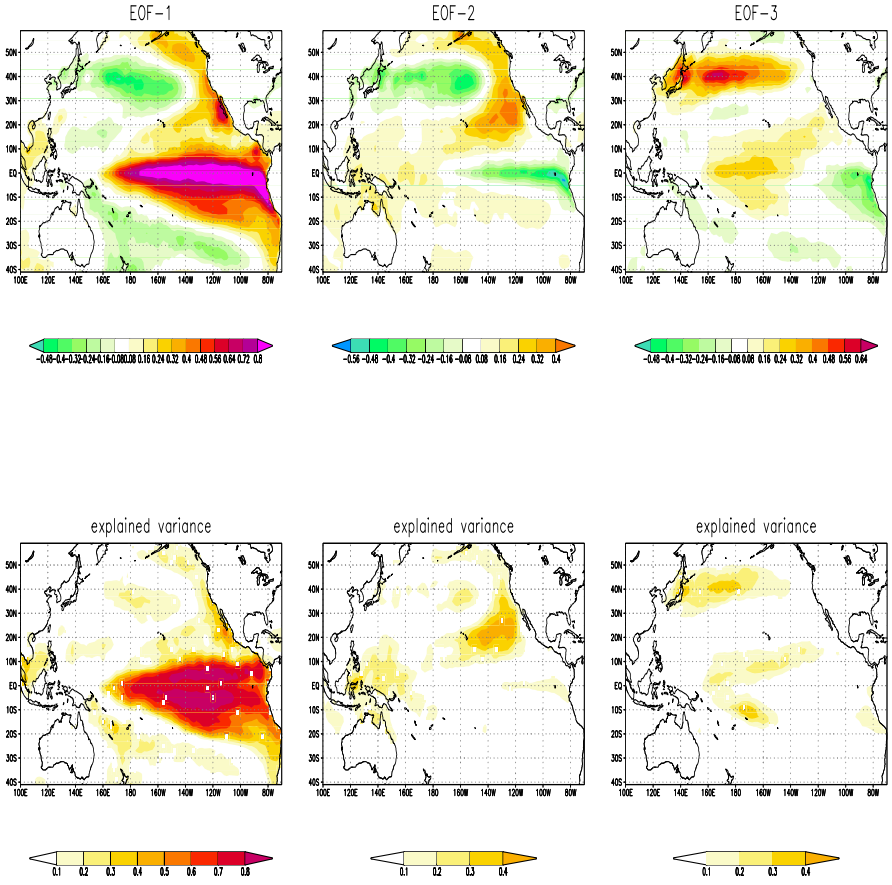


Figure 11.9: The leading eigenvectors (EOF-modes) of monthly mean SST of the Pacific, scaled by $\sqrt{\lambda_i}$. In units of K. In addition the corresponding explained variance fields are shown

Chapter 12

Alternative Analysing Techniques

The number of techniques and names for defining the leading modes of variability is nearly infinite. We can divid this methods into two categories: the ones based on eigenvalues of a covariance matrix, maximizing some covariance and those that do not. In the following only a few will be discussed, starting we those based on eigenvalues of a covariance matrix.

12.1 Canonical Correlation Analysis (CCA)

Is identical to Principal Component Analysis, but the eigenvalue problem of the correlation matrix is solved, instead of the covariance matrix. The resulting CCA-modes maximize the correlation between the points of the domain. This method is sometimes preferred over the Principal Component Analysis if the coherency of the patterns is more important than the variance. In data sets of large spatial inhomogenity of the variance fields it is often more instructive to study the CCA-modes instead of the EOF-mode. Rainfall data sets, for instance, have often same small regions with heavy rainfall, thus very large variances. The EOF-modes will gather around these spots, while large regions of the domain may be ignored due to the small variance at these regions. The CCA-modes do not know about variance they only care about coherency, thus they will compared to EOF-modes have larger spatial extend. Whether or not the CCA-modes or the EOF-modes are the *'better'* presentation depends of cause on the point of view. CCAs are good for the large scale patterns of rainfall, where EOF-modes are good for heavy rainfall events,floods where the total amount counts.

12.2 Singular Value Decomposition (SVD)

Is essentially another word for Principal Component Analysis, although in a strictly mathematical point of few they may be different. And another word for Singular Value Decomposition is Maximum Covariance Analysis (MCA).
The main difference is that SVD analysis we can not only find the eigenvalues of the auto-covariance matrix, but we can find the eigenvalues of the cross-covariance matrix. That means we find the patterns of covariance between two different fields. The structure of the problem is identical to the Principal Component Analysis, but we know have for each SVD-mode two patterns, one for each field, and two time series, one for each field. the eigenvalue of a SVD-mode is a covariance value.

12.3 Rotation of Principal Components

A multivariate data set $\mathbf{D}$ represents an N-dimensional vector spaces that evolves in time (or other dimension). As for any vector space we can chose different basis for the presentation of the space.

The number of possible basis is infinite. The basis of the Principal Components analysis is found by maximising the total variance of one basis vector, the EOF-1.
The literature is full of criteria to define other basis usually guided by some other statistical or mathematical criteria to maximize. Most methods have in common that the algorithm for finding the alternative basis is initialised with the EOF-basis. This concept we can sketch as: In PCA we analyse the data matrix $\mathbf{D}$ in the following steps:

$$\mathbf{D} \quad \rightarrow \Sigma \qquad \rightarrow det(\Sigma - \lambda \cdot \mathbf{I_S}) = 0 \quad \rightarrow \Pi_{eigenvectors} \tag{12.1}$$

So we have a data matrix $\mathbf{D}$ from which we get our covariance matrix Σ. This gets us to the eigenvalue problem, which we solve and than find the eigenvectors (patterns) $\Pi_{eigenvectors}$.
Methods that are based on EOF-basis rotation do not analyse the data matrix $\mathbf{D}$ itself, but indirectly estimate some optimisation criteria by assuming that the leading EOF-modes are a good approximation of the data matrix $\mathbf{D}$. So Rotated analyse starts from the PCA results and then do the following steps:

$$\Pi_{eigenvectors} \quad \rightarrow max(criteria(\Pi_{rotated})) \quad \rightarrow \Pi_{rotated} \tag{12.2}$$

The new basis is found by pair wise rotating the EOF-modes, until the maxima of the criteria of the alternative method is reached with sufficient accuracy. Thus the name rotated EOF is a totally insufficient description of the basis, since it does not say what criteria the basis of the vector space is maximising. Unfortunately, it seems that many researcher are not aware that rotation of EOFs is only describing the calculation algorithm, it is not the essential statistical or mathematical criteria underlying the new basis.This is probably because the common criteria in climate research is the VARIMAX criteria, which is basically the default way in rotating the EOF-modes. Thus rotated EOFs are the VARIMAX basis, if not otherwise specified.
The new set of patterns, $\Pi_{rotated}$, do not need to be orthogonal to each other ($\pi_i' \cdot \pi_j = 0 \quad \forall i \neq j$; eq. [11.13]). Indeed in many criteria in the literature the new set of patterns will have patterns that are similar to each other:

$$\pi_i' \cdot \pi_j \neq 0 \quad \forall i, j \tag{12.3}$$

The time series of these new patterns $\Pi_{rotated}$, also do not need to be orthogonal to each other ($\psi_i \cdot \psi_j' = 0 \quad \forall i \neq j$; eq. [11.14]). Indeed in many criteria in the literature the new set of patterns will have time series that are correlated to each other:

$$\pi_i' \cdot \pi_j \neq 0 \quad \forall i, j \tag{12.4}$$

Quit often it is unclear what exactly the constraints are in the patterns defined in the literature. In the VARIMAX criteria, for instance, the new patterns are often not orthogonal to each other and this may also be the case for the time series. It how depends on how the VARIMAX opitimisation has been applied. In principle the VARIMAX criteria or most other optimisation criteria can be apply with the constraints of eqs. [11.13] and [11.13]. For VARIMAX it is at least common in climate research to have relax the pattern orthogonality (VARIMAX pattern are often chosen to be not orthogonal to each other).

12.3.1 VARIMAX (Simplicity)

The VARIMAX criteria is based on maximising the simplicity of all spatial patterns. The simplicity of a pattern is defined as:

$$S_{var}(\pi) = \frac{1}{S}\sum_{i=1}^{S}(\frac{\chi_i}{\sigma_i})^4 - \frac{1}{S^2}\left(\sum_{i=1}^{S}(\frac{\chi_i}{\sigma_i})^2\right)^2 \tag{12.5}$$

with σ_i a possible normalisation for each spatial point of the domain, which is not used in the 'raw' VARIMAX criteria. The simplicity compares the amplitudes of the patterns to the power 4 with the squared amplitudes. This maximizes if only a few point have large amplitudes. Thus we find more compact (simple) patterns than the EOF-patterns.

$$\Pi_{eigenvectors} \quad \rightarrow max(S_{var}(\Pi_{rotated})) \quad \rightarrow \Pi_{VARIMAX} \tag{12.6}$$

Note that the rotation has different options, where you can for instance chose if you want to relax the orthogonality in the spatial patterns or the in time series. It is, I think more common to relax the orthogonality in the spatial patterns, so that VARIMAX patterns are spatially correlated to each other, but the time evolutions different VARIMAX-modes are uncorrelated.

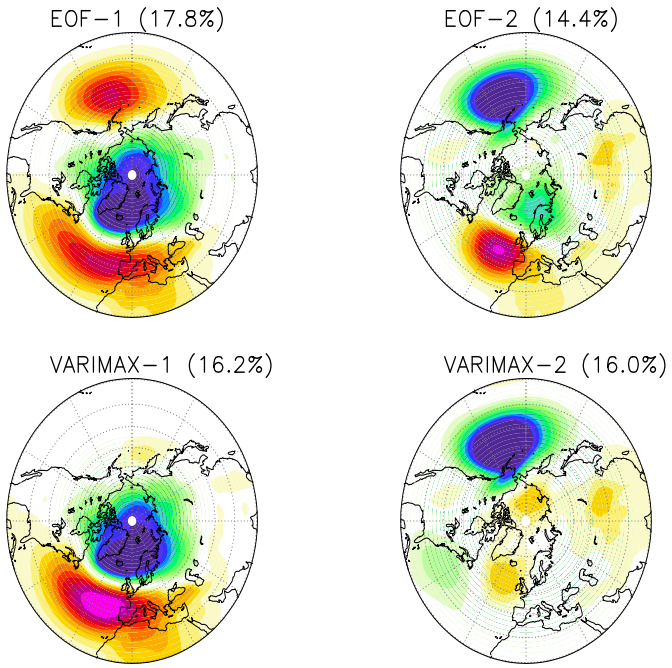


Figure 12.1: The leading EOFs and VARIMAX patterns of winter time monthly mean SLP of the Northern Hemisphere.

12.4 ... and a Million Other Methods or Names

12.5 Cluster Analysis (Methods not based on Eigenvalues of the Covariance Matrix)

The set of modes or patterns that we use to describe the data matrix can be driven in many different ways. We already discuss methods that use the results of the PCA as the starting point for their optimisation criteria. these are often called 'rotated EOFs'. However, many different methods exist to define patterns in the data set that do not depend on PCA at all. Those are of called 'Cluster Analysis' or just "Clusters".
As for the PCA and the rotated EOFS these methods will have an optimisation criteria. The concept of cluster analysis can be sketched as follows:

$$\mathbf{D} \quad \rightarrow max(algorithm(\mathbf{D})) \quad \rightarrow \Pi_{cluster} \tag{12.7}$$

So unlike rotated EOFs, a cluster analysis applies, in general, some maximisation algorithm directly onto the data matrix **D**. In principle a cluster analysis criteria may be optimised directly on the basis of the data matrix **D** or by rotation of EOF-modes. The resulting patterns should be about the same, but do not have to be (depending on the method/criteria).
"K-means" is one out of many examples for cluster analysis. Another example is self organising maps which is based on training neural networks. The number of methods to define the clusters is virtually infinite, it seems.

12.6 Detection of Propagating Structures

To be continued ...

Chapter 13

Interpretation of Principal Component Analysis

In many studies EOF-analysis or alternative pattern definition methods are carried out in order to understand the physical processes that drive the spatial variability of data field. Therefore the analysis aims on trying to understand what the structure of the leading EOF-modes may tell us about the teleconnections between different regions of the data domain. In most of these studies the principal component analysis is interpreted in terms of Factor Analysis, see next section. As an alternative we may think of the EOF-patterns as a continuous spectra of different spatial scales, similar to the power spectra of time series. In chapter 13.2.1 we will discuss an alternative approach, which considers the multi-variate data as a high dimensional stochastic process similar to the approach used in time series analysis.
In this chapter here we will discuss the factor analysis approach and the problems in interpreting the EOF-modes that go long with it. The next section will introduce the factor analysis approach. In section 13.1.2 we will discuss some exmples of the interpretation of EOF-modes as it has been discussed in the literature. As comment on this discussion a simple artificial example of an EOF-analysis is presented in section 13.1.3, that shall illustrate the problems in the interpretation of individual EOF-modes. In final section of this chapter the Indian Ocean SST is discussed in more detail.
Most of this chapter is based on two of my own publications (Dommenget and Latif 2002 and 2003). So be warned, that this my not be the common textbook point of view.

13.1 The Deterministic Mode View

13.1.1 Factor Analysis

The way EOFs modes are discussed in most statistical analysis (e.g. Dommenget and Latif (2002) and references therein) is based on factor analysis, as pointed out by Jolliffe (2003). It is assumed that the multivariate data $\mathbf{X}$ is a result of the evolution of a set of factors ,π_i, (often called teleconnections), and some residual unstructured noise Ξ:

$$\mathbf{X} = \Psi\Pi + \Xi \tag{13.1}$$

Π is a matrix of factors π_i, where each factor π_i is interpreted as a coherent spatial pattern (teleconnections) with a time evolution of ψ_i. The factors are the dominating influences for $\mathbf{x}$ (for details on factor analysis see textbook by Jolliffe 2002). Here we usually assume that $\mathbf{X}$ is a high dimensional space $N_X >> 1$, while the the number of factors is small $N_\Pi \approx 1$.

It may be important to note here that neither in EOF nor in VARIMAX analysis any evidence is given that the leading modes represents a factor λ_i, and is not just a fancy representation of the residual unstructured noise **e**. Dommenget and Latif (2002) argue that most likely neither EOF nor VARIMAX will find the leading teleconnection factors in climate data sets.

13.1.2 Some Examples of EOF-Analysis in Recent Publications

EOF-analysis is currently very popular in the climate community, which we can easily quantify be the citation index of some recent publications (numbers update 2005):

Thompson and Wallace 1999: "The Arctic Oscillation signature in the wintertime geopotential height and temperature fields". *times cited: 605.*

Saji et al. 1999: "A dipole mode in the tropical Indian Ocean". *Times cited: 227*

Chang et al. 1997: "A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air-sea interactions".*Times cited: 140*

MOURA AD, SHUKLA J 1981: ON THE DYNAMICS OF DROUGHTS IN NORTHEAST BRAZIL - OBSERVATIONS, THEORY AND NUMERICAL EXPERIMENTS WITH A GENERAL-CIRCULATION MODEL. *Time cited: 195.*

These are outstanding numbers considering that the citation index of the leading scientific climate journals are in the order of 2-3 (citations over the first two years).
Thus the climate modes motivated by EOF-analysis are the most popular topics of our times in climate research. In the following we will present the EOF-modes of this three examples and highlight some of the controversial aspects of these examples.

The Tropical Atlantic SST

The first two EOFs and VARIMAX patterns of the tropical Atlantic SST anomalies are shown in Fig. 13.1. The EOF-1 pattern is more or less uniform over the entire domain, while the EOF-2 is an inter-hemispheric dipole pattern. In contrast two the two EOF-patterns, the two leading VARIMAX patterns are more localized, while each of the two leading VARIMAX pattern covers just one hemisphere and the two pattern do not overlap significantly. Two regression patterns between the box averaged SST of the centers of the dipole pattern and the SST field are shown additionally for comparison in Fig. 13.1.
The inter-hemispheric dipole pattern in the EOF-2 has received a lot of attention, in terms of whether this pattern represents a potential physical mode of SST variability on decadal time scales (Weare (1977), Servain (1991), Nobre and Shukla (1996), Chang et al. (1997) and Tourre et al. (1999)), or is only an artifact of the EOF analysis (Houghton and Tourre 1992, Enfield et al. 1999, Dommenget and Latif 2000). Dommenget and Latif (2000) basically argue on the basis of coupled model results and observations that the dipole in the tropical Atlantic does not represent a physical mode.

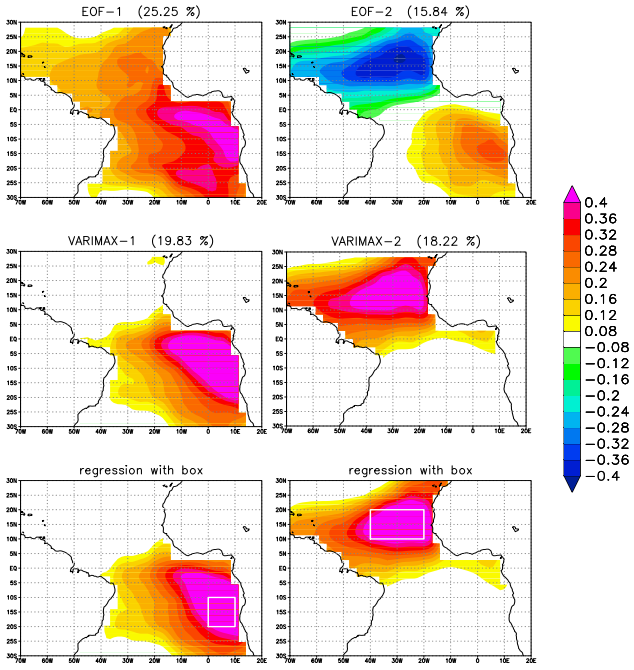


Figure 13.1: The EOFs, VARIMAX patterns and regressions of box averaged monthly mean sst in the tropical Atlantic Ocean. The amplitudes are in Kelvin.

SST in the tropical Indian Ocean

A similar analysis is now repeated for the tropical Indian Ocean. The first two EOF and VARIMAX patterns and two regression patterns between box averaged SST and the SST field are shown in Fig. 13.2.

Again, the EOF-2 of the SST variability is characterized by a dipole. However, there are some significant differences compared to the tropical Atlantic. First, the EOF-1 of the Indian Ocean explains much more variance than the EOF-1 of the tropical Atlantic, and second, the EOF-1 explains also much more variance than the EOF-2 of the tropical Indian Ocean. Furthermore, the VARIMAX patterns do not pick up the two centers of EOF-2. The eastern center of the dipole does not show up in any of the four most dominant VARIMAX patterns (patterns 3 and 4 are not shown).

The first two EOF patterns have been interpreted in terms of potential physical processes by Saji et al. (1999). They point out that the EOF-1 has a strong correlation with the El Niño in the tropical Pacific and can therefore be interpreted as the Indian Ocean response to El Niño. A response of the Indian Ocean to ENSO is well known and has also been pointed out by others (e.g. Venzke et al. 2000 and Reason et al. 2000). Since the EOF-2 has an orthogonal time evolution to EOF-1, they argue that the EOF-2 can be interpreted as an El Niño independent mode of variability, which is unique to the tropical Indian Ocean. However, the VARIMAX representation and the regressions provide no indication for the existence of a dipole mode, as suggested by EOF-2.

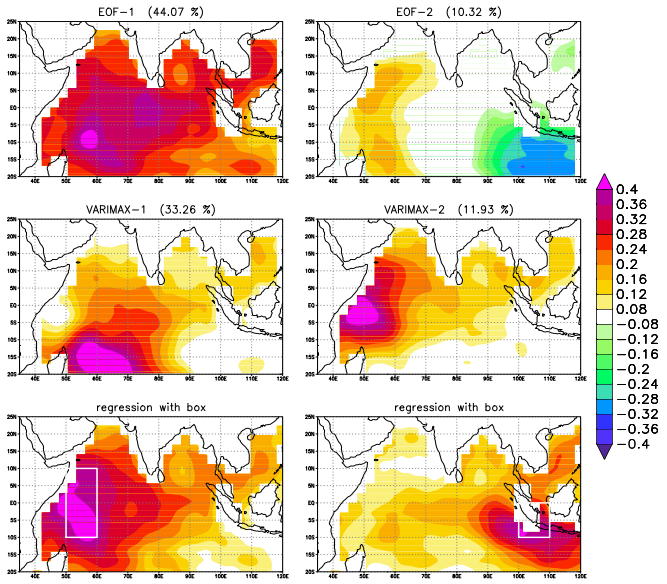


Figure 13.2: The EOFs, VARIMAX patterns and regressions of box averaged monthly mean sst in the tropical Indian Ocean. The amplitudes are in Kelvin.

SLP variability in the Northern Hemisphere

We shall now analyze the Northern Hemisphere winter SLP variability. The following example is different in many aspects compared to the ones described above. In contrast to SST anomalies, SLP anomalies in one region are usually compensated by SLP anomalies of opposite sign in a nearby region at the same time. Therefore the patterns of SLP have in general a dipole or multipole structure.
Furthermore, the standard deviation of the SLP is very inhomogeneous, with much stronger variance in higher latitudes compared to lower latitudes. In data sets with inhomogeneous standard deviations, the covariance-matrix based EOF can be very different from a correlation-matrix based EOF analysis. It is therefore instructive to additionally calculate the correlation-matrix based EOF-analysis.
In Fig. 13.3, the first two covariance-matrix based EOFs, correlation-matrix based EOFs, VARIMAX modes and two regression patterns are shown. Again, the different methods of representing the SLP variability in the Northern Hemisphere give quite different results with respect to the teleconnections. This may be one of the reasons why there is a scientific debate about which of these patterns describe best the dominant modes of SLP variability. For an overview of this controversy see Ambaum et al. 2001 (see also Barnston and Livezey 1987, Thompson and Wallace 2000, Wallace 2000).

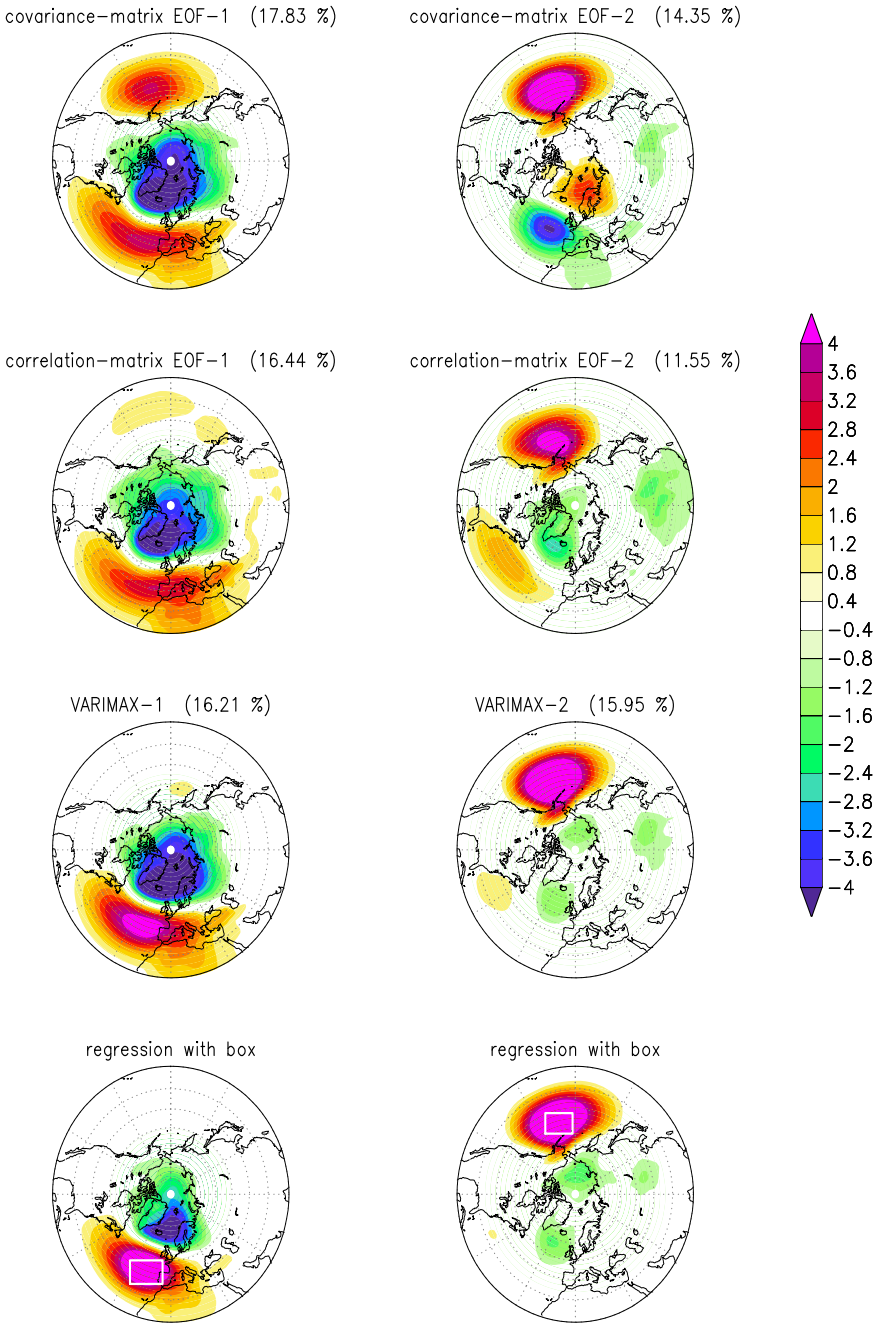


Figure 13.3: The leading EOFs, VARIMAX patterns and regressions of box averaged monthly mean winter time (November to April) SLP in the Northern Hemisphere. The amplitudes are in Pascal.

13.1.3 A Simple Artificial Example of EOF-Analysis

A simple 3-dimensional example might help to understand the difficulty in interpreting the patterns of the former examples. The advantage of the following artificial example compared to the ones described above is that we discuss a low-dimensional problem which is well-defined and in which statistical uncertainties do not exist.

We assume that our domain can be divided into three regions. We then define three different modes of variability, which are shown in the upper panel of Fig. 13.4. We have one mode which only acts in the left region, one only in the right region and one which covers all three regions. The explained variance of each mode is shown in the titles of each plot in Fig. 13.4. We assume that the time evolutions of theses modes are uncorrelated and that the standard deviation of all time series of these modes amount to unity.

The structures of the physical modes are motivated by the analyses of the SST in the tropical Atlantic and Indian Oceans. The three modes may therefore yield some further insight into the modal structure in these regions.

For the SLP in the Northern Hemisphere, Mode-1 could be interpreted as the North Atlantic Oscillation (NAO) of the Atlantic-European region (similar to VARIMAX-1 in Fig. 13.3), Mode-2 as the Pacific North America pattern (PNA) (similar to VARIMAX-2 in Fig. 13.3) and the Mode-3 would be an annular mode (similar to EOF-1 in Fig. 13.3, but much weaker and more zonal). The three regions of the simple example would then be interpreted as the Atlantic-European region (the left region in Fig. 13.4), the Pacific domain (the right region) and the rest of the Northern Hemisphere (the central region).

However, to keep the problem as simple as possible we represent each region by one point only. The values at these points are printed on top of the mode (see Fig. 13.4). We can therefore interpret each physical mode as a three dimensional vector, π_i, where each component of this vector represents the variability of one region. The Variance of π_i ist given by,

$$\gamma(\pi_i) = \pi_i \pi_i' \tag{13.2}$$

The set of the three vectors defines a matrix Π,

$$\Pi = \begin{pmatrix} 5 & 0 & 2 \\ 0 & 0 & 2 \\ 0 & 4.5 & 2 \end{pmatrix} \tag{13.3}$$

Each column represents one of the modes shown in Fig. 13.4. The actually observed variability in the three regions defines a matrix Y which is related to Π by:

$$Y = \Psi\Pi \tag{13.4}$$

$$\Psi := (\psi_1(t), \psi_2(t), \psi_3(t)) \tag{13.5}$$

The coordinates ψ_i of Vector Ψ describe the time evolutions (PCs) of the basis Modes. The construction of our example allows us to calculate the covariance matrix exactly, since our example has been constructed such that the characteristics of the physical modes are known exactly. Therefore all structures that appear in the following statistical analysis are well-defined. The covariance matrix is given by,

$$\Sigma_{YY} = \Pi\Pi' = \begin{pmatrix} 29 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 24.25 \end{pmatrix} \tag{13.6}$$

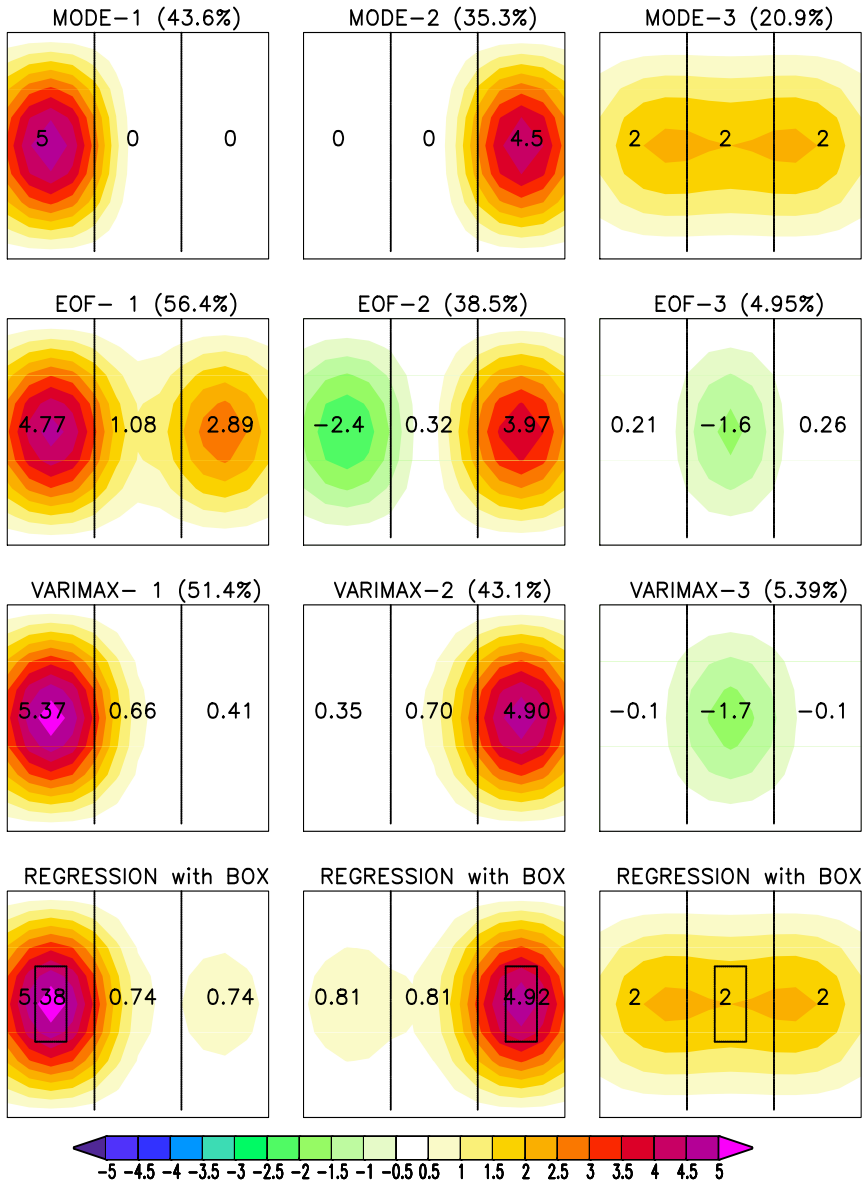


Figure 13.4: The physical modes (1st panel from top), EOF (2nd panel from top), VARIMAX (3rd panel from top) pattern and the regressions patterns of each coordinate with all coordinates (bottom panel) of the simple low-dimensional example are shown. The values plotted on top of the patterns represent the associated vectors and are identical to the amplitudes of the patterns in the respective region. The amplitudes are in arbitrary units.

The square root of the covariance matrix yields the regressions of one coordinate of the vector space (one region) with all coordinates (regions) of the vector space. The regression patterns and values are shown in the lower panel of Fig. 13.4.
Based on the covariance matrix we can also calculate the EOF vectors exactly. We therefore do not have to consider the sampling error problem, which can lead to unstable estimations of the EOF vectors (North et al. 1982). The EOFs are also shown in Fig. 13.4. The EOF vectors are not degenerated, since all eigenvalues of the covariance matrix are different (see explained variances of the EOFs in Fig. 13.4).
The set of the three EOF vectors define a matrix Q. Similar to equation [13.4] the observed Vector Y is related to Q by:

$$Y = QP_Q \tag{13.7}$$

The Vector P_Q describes the time evolutions (PCs) of the EOFs. Using the equations [13.4] and [13.7] we can show that the Vector P_Q can be presented by as linear combination of the Vector ψ_i:

$$QP_Q = \Psi\Pi \tag{13.8}$$

$$=> Q^T QP_Q = Q^T\Psi\Pi \tag{13.9}$$

with $\Lambda = Q^T Q =$ the diagonal matrix of the eigenvalues of the EOFs we find:

$$=> \Lambda P_Q = Q^T\Pi\Psi \tag{13.10}$$

$$=> P_Q = \Lambda^{-1}Q^T\Pi\Psi \tag{13.11}$$

$$A := \Lambda^{-1}Q^T\pi \tag{13.12}$$

Thus the matrix A describes the linear combination of the Vector Ψ, which constructs the vector P_Q.
The coefficients of A are listed in Table 13.1. A row in Table 13.1 describes the relative influence of the basis modes onto a single EOF-mode. For example, it can easily be seen (in Fig. 13.4 and corresponding in Table 13.1.) that the EOF-2 includes the time evolutions of Mode-2 with positive loadings and Mode-1 with slightly smaller negative values (Table 13.1). Please note that the EOF-2 represents a pattern which does not really exist in our simple example, so that it is completely artificial.

principal component	Mode-1	Mode-2	Mode-3
PC-1	0.74	0.40	0.54
PC-2	-0.56	0.81	0.6
PC-3	0.37	0.43	-0.82

Table 13.1: The matrix A, by which the PCs of the EOF vectors are constructed.

Usually the VARIMAX representation is calculated by using the EOF-patterns. Here we can directly calculate the VARIMAX representation from our basis vectors, since the basis vectors are already given with orthogonal time evolutions, which is usually not the case in climatological data sets. Therefore the VARIMAX vectors are well-defined. The VARIMAX patterns and their explained variances are also shown in Fig. 13.4 and the corresponding transformation matrix A for the PCs of the VARIMAX vectors are listed in Table 13.2.
Our simple 3-dimensional example has an inhomogeneous distribution of the local standard deviation, with larger variability in the left and right region and less variability in the center region. It

principal component	Mode-1	Mode-2	Mode-3
VPC-1	0.94	-0.06	0.33
VPC-2	-0.07	0.93	0.35
VPC-3	0.33	0.36	-0.87

Table 13.2: The matrix A, by which the PCs of the VARIMAX vectors are constructed.

is therefore similar to the inhomogeneous standard deviation of Northern Hemisphere winter SLP variability.
In data sets with inhomogeneous standard deviations, the covariance-matrix based analysis can be very different from a correlation-matrix based analysis. We have therefore calculated the correlation-matrix based EOFs and VARIMAX modes, and correlations between the different regions (Fig. 13.5). The correlation-matrix based VARIMAX analysis is equivalent to the "normal" VARIMAX as stated in Kaiser (1958).
The transformation matrix A for the PCs of the correlation-matrix based EOFs and VARIMAX vectors are listed in Table 13.3 and 13.4.

principal component	Mode-1	Mode-2	Mode-3
PC-1	0.38	0.39	0.84
PC-2	0.75	-0.66	-0.03
PC-3	0.55	0.64	-0.54

Table 13.3: The matrix A, by which the PCs of the correlation-matrix based EOFs vectors are constructed.

principal component	Mode-1	Mode-2	Mode-3
VPC-1	0.98	-0.04	0.19
VPC-2	0.00	0.98	0.21
VPC-3	-0.19	-0.21	0.96

Table 13.4: The matrix A, by which the PCs of the correlation-matrix based (normal) VARIMAX vectors are constructed.

Discussion: We used three different statistical methods (EOF, VARIMAX and regression analysis) to identify the different variability modes in different multi-variate data sets.
In the following discussion we compare the results from the simple low-dimensional example with those from the other three examples using observed data. We do not consider statistical uncertainties, since the problems due to statistical uncertainties in EOF analysis have already been discussed elsewhere (e.g. North et al. (1982) and Richman (1986)). Furthermore, the points that we make here are not related to statistical uncertainties.
Although the discussion will be mostly focused on the differences in the spatial patterns, one has to take into account that each pattern is related to a specific time series. Patterns that do show large differences in the spatial structures will in general also have large differences in the corresponding time series.
In the simple low-dimensional example we consider three variability modes. The three modes can be interpreted as the 'real physical modes' of the domain. From a mathematical point of view all representations (e.g. EOF, VARIMAX) of the simple low-dimensional example are equally valid, but from a physical point of view we would like to find the representation, which is most clearly

pointing towards the 'real physical modes' of the problem.
We constructed the simple low-dimensional example by two local and spatially orthogonal modes, which should represent some simple internal modes (see Fig. 13.4). The third mode in this example represents a domain-wide mode, which may be regarded, for instance, as the response of the domain to some kind of external influence. The third mode is not orthogonal in space with the other two, which will be important in the following discussion. By construction the simple low-dimensional example does not contain any statistical uncertainties, which allows us to determine the EOF and VARIMAX patterns exactly.
Although the Mode-3 is the weakest one in the simple low-dimensional example, the EOF-1 is very similar to it (see Fig. 13.4). Despite the fact that it captures some features of the two other basis modes, it may be interpreted as the domain response to some kind of external influence, similar to how Saji et al. (1999) have interpreted the EOF-1 of the tropical Indian Ocean. Although the EOF-1 in the simple low-dimensional example is very similar to the Mode-3, the PC-1 is a superposition of all three basis modes (see Table 13.1).
In the tropical Indian and Atlantic Ocean SSTs this kind of weak external influence may be the ENSO response or a greenhouse warming trend, as expressed by the leading EOFs (see Fig. 13.1 and Fig. 13.2). In the Northern Hemisphere SLP such an external influence might manifest itself as an annular mode like the EOF-1 of Northern Hemisphere SLP (see Fig. 13.3).
On the other hand, we would like to clarify that the EOF-1 does not need to be a superposition of many modes. If we would have chosen the Mode-3 as the most dominant mode in our simple example then the patterns of the EOF and VARIMAX analyses would not look much different compared to the ones shown in Fig. 13.4, but the PC-1 would clearly be dominated by the Mode-3. It is, for example, well known that the EOF-1 of the tropical Pacific SST is really representing the El Niño mode.
The orthogonality constraint in space forces the EOF-2 of the simple low-dimensional example to be a domain-wide dipole, although the two centers of the dipole are not anti-correlated by construction (see Fig. 13.4). It can therefore be concluded that in a domain which has an EOF-1 pattern with a shape of a domain-wide monopole must have a dipole in the EOF-2. The dipole, however, is totally an artifact of the orthogonality constraint.
The EOF-3 and VARIMAX-3 patterns in Fig. 13.4 are interesting, since they indicate a kind of central mode which does not really exist. Interestingly, the time evolution of this mode is a superposition of all three basis modes. This leads to the fact that the PC-3 includes variability from the basis Mode-1 and Mode-2 which actually are not influencing this region at all (see Table 13.1 and 13.2).
By construction the EOF analysis maximizes the explained variance in the leading EOFs. This will in general lead to the fact that only a few EOF patterns are needed to explain a large amount of variability. In the artificial example the two leading EOFs explain more than 95% of the total variance (see Fig. 13.4). However, our artificial example has three modes. This indicates that the EOF analysis will in general underestimate the complexity of the problem. This is also indicated in the tropical Indian Ocean SST analysis, in which the two leading EOFs explain much more total variance than the two leading VARIMAX pattern (see Fig. 13.2).
Sometimes maps of explained local variances are shown in order to highlight certain regions in which a relatively high amount of variance is explained, indicating that these regions should be analyzed in greater detail. This approach will in general favor the VARIMAX method, since VARIMAX optimizes the simplicity and therefore produces local patterns. Although, the VARIMAX representation is often a very instructive representation of the data, it may often fail to find global modes, like the Mode-3, due to the optimization of the simplicity which favors localized modes.
In Fig. 13.5, we have repeated the analysis of our simple example but with correlation-matrix based EOFs and VARIMAX analysis, and by computing the correlations between the different regions. The patterns are presented in terms of correlation values. These representations look quite different

from the covariance-matrix based analyses. Here the VARIMAX analysis and the correlations are in very good agreement with the original modes, but the EOF patterns are again very different from the original modes.

This example and the example of the SLP variability in the Northern Hemisphere (see Fig. 13.3) may indicate that correlation-matrix based analyses are more instructive than covariance-matrix based analyses. However, we believe that this cannot be generalized. Whether correlation- or a covariance-matrix based analysis gives a better representation of the 'physical modes' depends strongly on the spatial structure of the 'physical modes'. Imagine, for instance, that the Pacific and the Atlantic pole in the covariance-matrix based EOF-1 in Fig. 13.3 would have the same spatial structure, but the Pacific pole would have a larger amplitude than the Atlantic Pole. In this case a correlation-matrix based analysis would not be able to focus on one of the poles, as in Fig. 13.3, since the correlation-matrix does not know anything about the larger amplitude of the Pacific Pole. In this case the covariance-matrix based analyses would be a better representation, and the EOF-1 would be focused on the stronger Pacific pole.

In the artificial example the regression patterns seem to be most instructive in representing the dominant modes of variability. However, the disadvantage of the regression analysis is that the choice of the index region is highly subjective and it is much easier to choose an index that is not instructive at all, than to choose an adequate index. For the SLP in the Northern Hemisphere, for instance, we could have chosen an index region over the North Pole and the regression would look very much like the covariance-matrix based EOF-1 (the regression pattern is not shown, but see Fig. 13.3 for the EOF-1). Thus, the disadvantage of the regression analysis is its subjectivity so that one always needs to argue why a certain index has been chosen.

Often regression indices are motivated by EOF analysis (e.g. the tropical Atlantic or Indian Ocean dipole indices), which seem to make the regression indices more objective. However, one has to consider that these indices are as limited in the interpretations as the EOF patterns themselves from which these indices are derived.

In our simple example both covariance-matrix based EOF and VARIMAX analysis somehow fail to adequately represent the weak global mode (Mode-3) and one can imagine that in many practical problems the correlation-matrix based EOF and VARIMAX analysis will also fail to identify the weak global mode. It may therefore be a good approach to eliminate the weak global mode prior to the EOF analysis.

However, there is no simple way to determine the pattern and time series of such a weak global, since one can not derive these structures by analyzing the domain itself. This would again lead to a superposition of the local and weak global modes into one mode, as in EOF-1 of Figs. 13.4 and 13.5. The structure of the weak global mode has to be determined by some additional knowledge about external influences like global warming or ENSO.

Conclusions: We have shown that EOF and rotated EOF analyses have problems in identifying the dominant centers of action or the teleconnections between these centers of action in multivariate data sets. We therefore have to be very careful in interpreting the EOF or rotated EOF modes as potential physical modes.

The problems in interpreting the patterns derived from EOF and VARIMAX analyses arise from the basic assumptions that are made by these statistical methods, which are not identical to the assumptions that we make to derive the so called 'physical modes' of the problem. The EOF analysis always represents modes of variability that are orthogonal in time and space. The constraint of the orthogonality in space is often not consistent with the real nature of the problem, like in the simple example, in which the basis modes are not orthogonal in space (see Fig. 13.4). The VARIMAX analysis is looking for localized modes, which is also not adequate for our simple example, since the Mode-3 is highly non-local.

A good strategy for statistical analysis of climate data is to look at the data with different statistical tools, like regressions, VARIMAX or EOF analysis, and develop a hypothesis for the potential

'physical modes' which is consistent with all representations of the data, instead of developing a hypothesis for the potential 'physical modes' based on only one representation, which is often in contradiction with other representations.
We would like to conclude our discussion with the following caveats for the interpretation of the results of the EOF or VARIMAX methods:

- The teleconnection patterns derived from the orthogonal analysis cannot necessarily be interpreted as teleconnections which are associated with a potential physical process (e.g. the dipole pattern Fig. 13.4).
- The centers of action derived from the EOF or VARIMAX methods do not need to be the centers of action of the real physical modes (see EOF-3 or VARIMAX-3 in Fig. 13.4).
- The PCs of the dominant patterns are often a superposition of many different modes, which are uncorrelated in time and which are often modes of remote regions that have no influence on the region in which the pattern of this PC has its center of action.

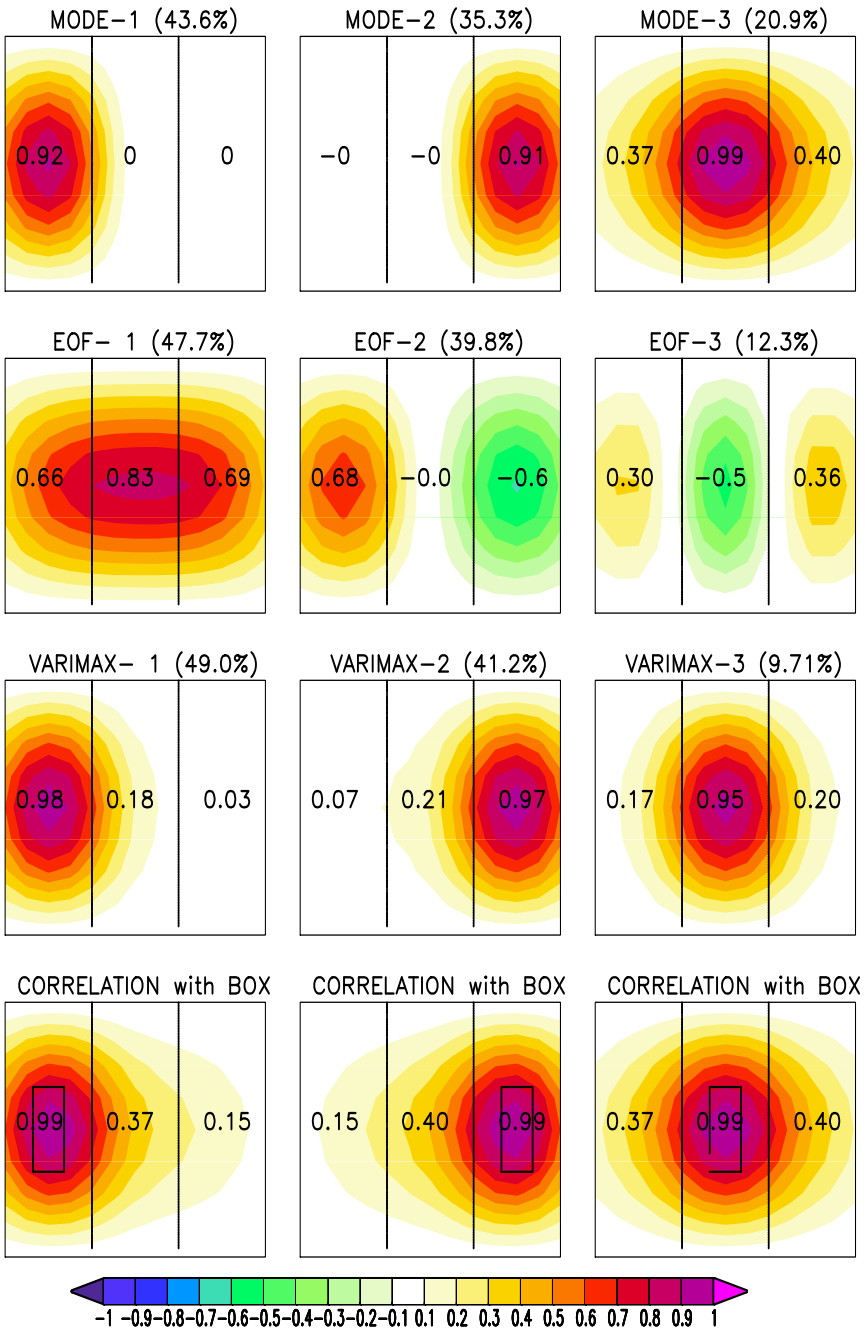


Figure 13.5: Same as in Fig. 13.4 but all analysis are based on the correlation matrix and the values are in terms of correlation.

13.1.4 A dispute about modes (tropical Indian Ocean SST Variability)

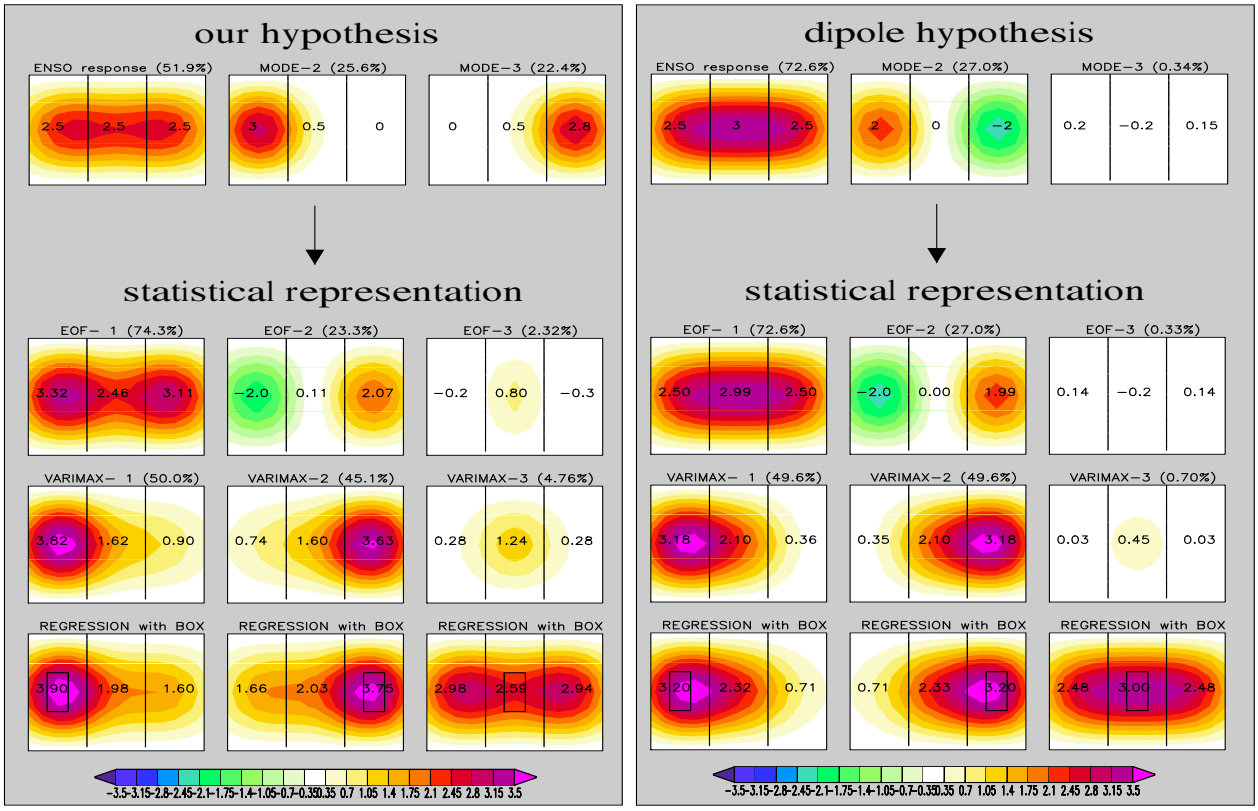


Figure 13.6: The hypothetical modes and their statistical representations by EOFs, VARIMAX patterns and regressions of box averaged SST in an artificial 3-dimensional domain are shown for two different hypotheses. The amplitudes are in arbitrary units.

The discussion of the previous artifical example caused some response by the coauthors of the Indian Ocean Dipole paper Behera et al. (2002). In their introduction Behera et al. (2002) say: *"DL question the existence of the IOD as a physical mode. The argument put forward by DL, however, suffers from misinterpretation of statistical analysis as well as misuse of the statistical techniques..."* (DL = Dommenget and Latif, 2002). This is of cause a statement that can not be left without a reply. So parts of the reply by DL:

Behera et al. (2002) report in their comments that several statistical characteristics or relationships are consistent with their dipole-hypothesis. We can only agree to this, but we would like to point out that all the characteristics or relationships they present are indeed also consistent with the alternative hypothesis that a dipole mode does not exist. These characteristics can therefore not be put forward as evidence for the existence of the dipole mode.

In the following, we would like to outline how these apparently contradictory hypotheses can be tested. Therefore we would like to use the concept of the artificial 3-dimensional example (described in DL) as a simplification of the Indian Ocean SST variability.

We assume that the response to ENSO is the most dominant mode of SST variability in the Indian Ocean, which can be simplified as a domain wide monopole (see left panels of Fig. 13.6). We assume further that the remaining ENSO-unrelated SST variability can be explained by local

air-sea interaction, which will most likely favor localized SST modes. In the framework of our 3-dimensional example we simplify the remaining variability by two local modes, as shown in the left panels of Fig. 13.6.
The dipole-hypothesis from Behera et al. (2002) is treated in a similar way in the right panels of Fig. 13.6
The two apparently contradictory hypotheses lead to essentially the same large-scale SST statistics as shown by the EOFs, VARIMAX and box-regressions in Fig. 13.6. Since both hypotheses describe the same SST statistics, analyses based on these statistics cannot be used to either support or reject one of the two hypotheses. It therefore does not make sense to count the number of dipole events, as Behera et al. (2002) do, in order to support the existence of a dipole mode. The numbers of expected dipole events are the same in both hypotheses.
According to Behera et al. (2002) the linear relationship between the IOD index and the zonal wind anomalies along the Indian Ocean equator as shown in their Fig. 1 should support the ocean-atmosphere coupled nature of the IOD mode. We do not see why this relationship should not exist in our "local" hypothesis. Thus the apparent *"physical and dynamical understanding of various ocean-atmosphere parameters"* in Behera et al. (2002) or Saji et al. (1999) cannot be put forward as evidence for the dipole mode, since they are also consistent with the alternative "local" hypothesis, which does not include a dipole mode.
Furthermore, Behera et al. (2002) say, in their section *"4. Dipole Mode in EOF and VARIMAX ": "The question, which arises here, is that whether it is possible to identify the dipole mode in the real SST data using these two methods? This can be achieved by filtering out the monopole mode related to ENSO (Fig.4)."*

principal component	ENSO-response	Mode-2	Mode-3
PC-1 (our hypothesis)	0.83	0.42	0.37
PC-1 (dipole hypothesis)	0.9999	0.0001	0.0128

Table 13.5: The table shows the contributions of the hypothetical modes to the PCs of the EOF-1 vector for both hypotheses.

Removing the *"monopole mode related to ENSO"*, which is the ENSO-response mode, would indeed be a good strategy for testing the hypothesis of a dipole mode. However, it cannot be done by removing EOF-1, since this already assumes that EOF-1 is identical to the ENSO-response, which is not necessarily true. In Table 13.5, the contributions of different hypothetical modes to the time evolution (PC) of EOF-1 are shown. For the dipole-hypothesis of Behera et al. (2002), the EOF-1 is essentially identical to the ENSO-response. In our hypothesis the EOF-1 is a superposition of all modes, although it is dominated by the ENSO-response.
In section 2. *"IOD as a climate mode"* Behera et al. (2002) point out that: *"The correlation coefficient peaks at 0.75 when the Niño-3 index leads the EOF mode by 4 months."*
This is indicates that at most 56 % (0.75^2) of the variance of EOF-1 is related to the ENSO response, while the remaining 44% is not related to ENSO. This appears to be consistent with our local hypothesis, but it is not consistent with the dipole hypothesis, in which EOF-1 has to be identical to the ENSO response.
Thus removing EOF-1 does not verify the existence of the dipole mode, since it already assumes that the dipole-hypothesis is valid. Instead of removing EOF-1, the real ENSO-response should be removed.
In Figure 13.7 we removed the ENSO-response from the data prior to the statistical analyses. Now the statistical representations of the two hypotheses are very different. Similar to Figure 4 in Behera et al. (2002), all statistical representations of the dipole-hypothesis would now clearly point towards a dominant dipole-mode. In our hypothesis, none of the statistical representations

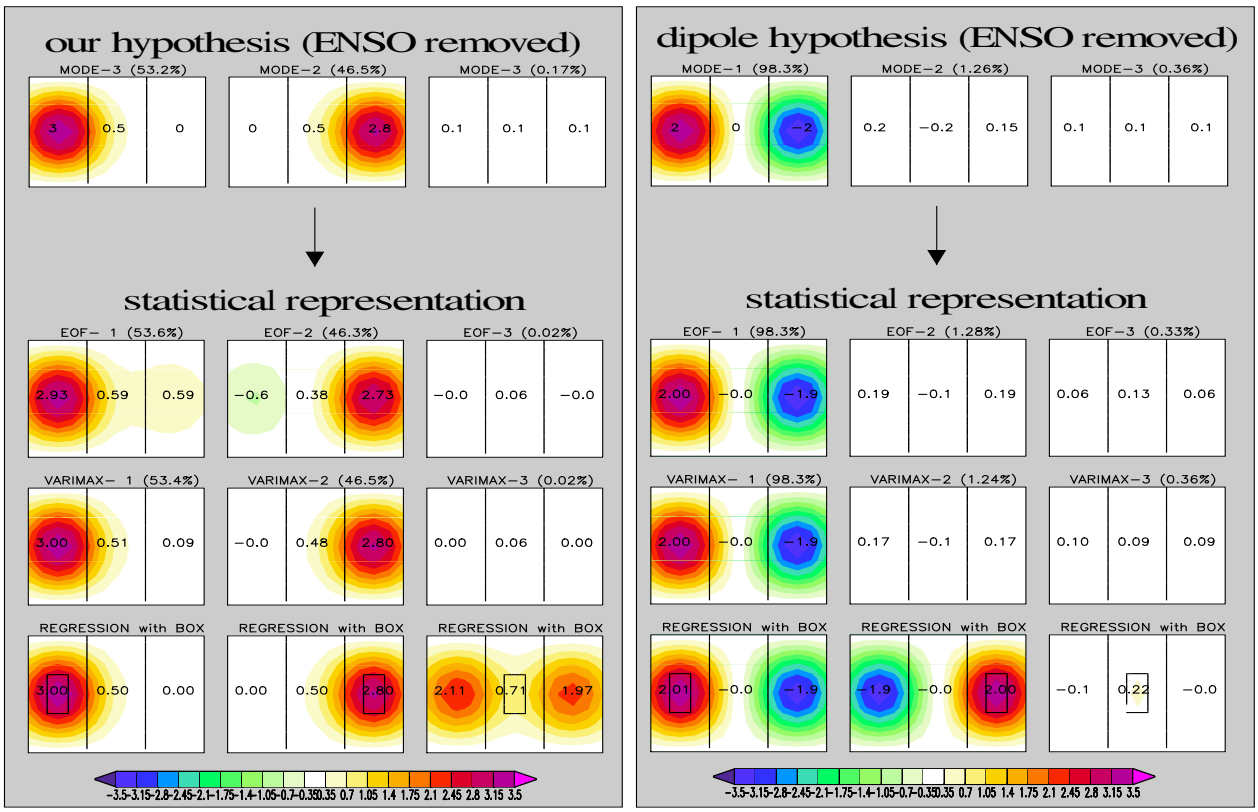


Figure 13.7: As in Figure 1, but the ENSO-response mode has been removed from both hypotheses.

would show a dipole mode as the most dominant mode. Thus removing the ENSO-response mode properly would clearly show which of the two hypotheses is true.

Baquero et al. (2002), followed the strategy outlined above. In their analysis of observed SST variability in the Indian Ocean they removed the ENSO-response statistically using the leading POP mode of the tropical Pacific. In addition they also analyzed two global coupled general circulation models, one in which a realistic ENSO mode is present and one in which the ENSO mode is suppressed physically.

None of the analyses presented in Baquero et al. (2002) can support the dipole-hypothesis. Moreover, all results seem to be consistent with our "local" hypothesis that the SST variability in the Indian Ocean is dominated by the ENSO-response alone and that the remaining ENSO-independent SST variability is consistent with local air-sea interaction.

13.2 The Stochastic Continuum View

13.2.1 A Null Hypothesis for Teleconnections (EOF-modes)

In the previous chapter the problems with interpreting the EOF-modes was discussed and it illustrated that statistical inferences from EOF-modes is far from trivial. In this chapter an approach will be presented that may help understanding the physical processes causing the structure of the leading EOF-modes. The basic idea is to compare the spectrum of EOF-modes with a simple stochastic null hypothesis, similar to the approach used for the time series analysis.
In following section some concepts from time series analysis and a definition of climate modes or teleconnections are presented. In section 13.2.3 a stochastic null hypothesis for the spatial structure of climate variability if formulated. The concept on how this stochastic null hypothesis can be compared with observed EOF-modes is presented in section 13.2.4 and some artificial and real data examples are discussed in the subsequent section. The chapter concludes with a discussion section.

13.2.2 Concepts

It is helpful to first discuss how climate modes could be defined and how limited such definitions may be. It is also instructive to illustrate how the concept of testing a stochastic null hypothesis is performed in time series analysis, which will be a guide for the subsequent analysis of the spatial structures of climate variables.

The null hypothesis in time series analysis

It is common in time series to evaluate the spectra of time series against an first order auto-regressive process (AR(1)-process), which goes back to the stochastic climate model of Hasselmann (1976). In its simplest form, Hasselmann's model is an AR(1)-process, which is defined by the following differential equation for time evolution of any physical variable Φ:

$$\frac{d}{dt}\Phi = c_{damp} \cdot \Phi + f \tag{13.13}$$

with $c_{damp} < 0$ being a constant damping and f white noise. The auto-correlation function in time, $c(\tau)$, of Φ is:

$$c(\tau) = e^{-\tau/t_0} \tag{13.14}$$

with the time lag τ and the e-folding time $t_0 =!/c_{damp}$. One can derive the analytical form of the spectral distribution of the null hypothesis of Φ from eq.[13.14]. In time series analysis, this null hypothesis is often used to evaluate the temporal behavior of Φ, by simply comparing the spectrum of Φ with that of a fitted AR(1)-process. The parameters of the fitted AR(1)-process are derived from the auto covariance function of Φ (e.g. Reynolds 1978, Dommenget and Latif 2002b).
In the case of the El Niño SST time series, for instance, the spectrum shows some characteristic enhanced variance (peak) in the interannual frequency range, which is usually interpreted as an indication for the oscillating nature of El Niño SST. The spectrum of the midlatitudes SST time series shows no peak, but a different overall slope of the spectrum, which indicates deviations from the AR(1)-process null hypothesis (Dommenget and Latif 2002b).

Definitions of teleconnection/climate modes and their limitations

The way EOFs modes are discussed in most statistical analyses (e.g. Dommenget and Latif (2002a) and references therein) is based on a factor analysis approach, as pointed out by Jolliffe (2003). It is implicitly assumed that the multivariate data $\mathbf{X}$ is a result of the time evolution of a set of K

fixed factors ,π_i, (often called teleconnections, modes or patterns), and some residual unstructured noise ξ.

$$\mathbf{X}(t) = \Psi(t)\Pi + \xi(t) \tag{13.15}$$

Π is a matrix of factor loadings π_i, where each π_i is interpreted as a coherent spatial pattern (teleconnection). These patterns are the dominating influence for $\mathbf{X}$ (for details on factor analysis see textbook by Jolliffe 2002). The time evolution of Π is given by a matrix of time series Ψ.
The idea is to assume that the high-dimensional system can be approximated by a low-order state space model, with the number of modes, K, much smaller than the dimension of $\mathbf{X}$ (von Storch and Zwiers 1999 section 15.5). The patterns Π in this approach are a reflection of the underlying low-order physical model. This approach, however, depends strongly on how the patterns Π are estimated. In the recent literature it seems a popular approach to associate the leading EOFs or other statistical modes with the leading teleconnections (e.g. Thompson and Wallace 1998 or Saji et al. 1999). It is, however, important to note that it is in general unclear if any teleconnections exist in the data set and how they can be estimated (e.g. Jolliffe 2002 section 7 and Dommenget and Latif 2002a). Dommenget and Latif (2002a) argue that most likely neither EOF nor VARIMAX will find the leading teleconnection factors in climate data sets. The inherent problem in this approach is that a criteria or algorithm needs to be formulated by which the empirical patterns Π are chosen. Thus the resulting modes may in many cases be a reflection of the statistical method used, but are not a good representation of the underlying physical processes.
An alternative method, which avoids to formulate any criteria for the structure of teleconnection modes, is to formulate a null hypothesis for the structure of spatial variability, which can be regarded as a model for the noise. Any pattern that is very distinct from the patterns of the null hypothesis is a good starting point for the estimation of teleconnection modes. This concept is similar to the time series analysis, in which the time scale behavior of El Niño, for instance, is simplified into a distinct oscillation mode on interannual time scales and a background red noise. In analogy the teleconnection modes are defined as the modes that stick out of the background noise, as define by the null hypothesis.

13.2.3 A stochastic null hypothesis for the spatial structure of climate variability

Before a stochastic null hypothesis for the spatial structure of climate variability can be fromulated it helps to have a look at some data to get an impression on how the spatial structure in climate variability looks like.
In Fig. 13.8 and 13.9 the NCEP global monthly mean SLP from 1948 to 1999 (see Kalney et al. 1996) and the global monthly mean SST anomalies, based on the Reynolds data set from 1950 to 2000 (Reynolds and Smith 1994) are represented by 5 simple box-correlations as well as by the 5 leading EOF-modes. The box-correlations have been chosen somewhat randomly with some guidance by the structure of the EOF-modes. From the two different representations of the data we find several interesting characteristics of both the data and the statistical methods we used:

- In all box-correlations the correlations decrease relatively fast with distance to the center of the box, clearly indicating that most of the SST and SLP variability is related to nearby regions only, while remote regions have in general a much weaker influence (correlation). This clearly represents a main characteristic of climate variability and can not be inferred from the EOF-modes.

- In all box-correlations positive correlations reach beyond the box boundaries, which indicate that nearby regions are in general linked by some kind of diffusive processes. Again this is not obvious from the EOF structure.

- The EOF-modes tend to be global modes with long distance teleconnection patterns. Some of these teleconnection patterns can be seen in the box-correlations as well, but others are not reproducible by box-correlations. For instance, the EOF-2 of the SLP presents a dipole over the North Atlantic, which is similar to the NAO-pattern, but has as well large correlation values in other regions of the globe. The box-correlation (see box-3 in Fig. 13.8) shows similar correlations in the North Atlantic, but very different correlations outside this region. Thus, teleconnection of EOF-modes can not be verified by box-correlations. However, this does not imply that the EOF teleconnection is not valid, but there seem to be no objective way to decide which teleconnection of the EOF modes is indeed reflecting a physical mode of variability.

- A few EOF-modes explain a lot of variability, while the same number of local box-correlation tend to explain much less variability. Therefore, EOF-analysis implies that a few modes of variability are responsible for a large fraction of the variability covering the entire domain. On the other hand, the same number of modes based on local box-correlations explains a much smaller fraction of the variability, which only covers the nearby regions.

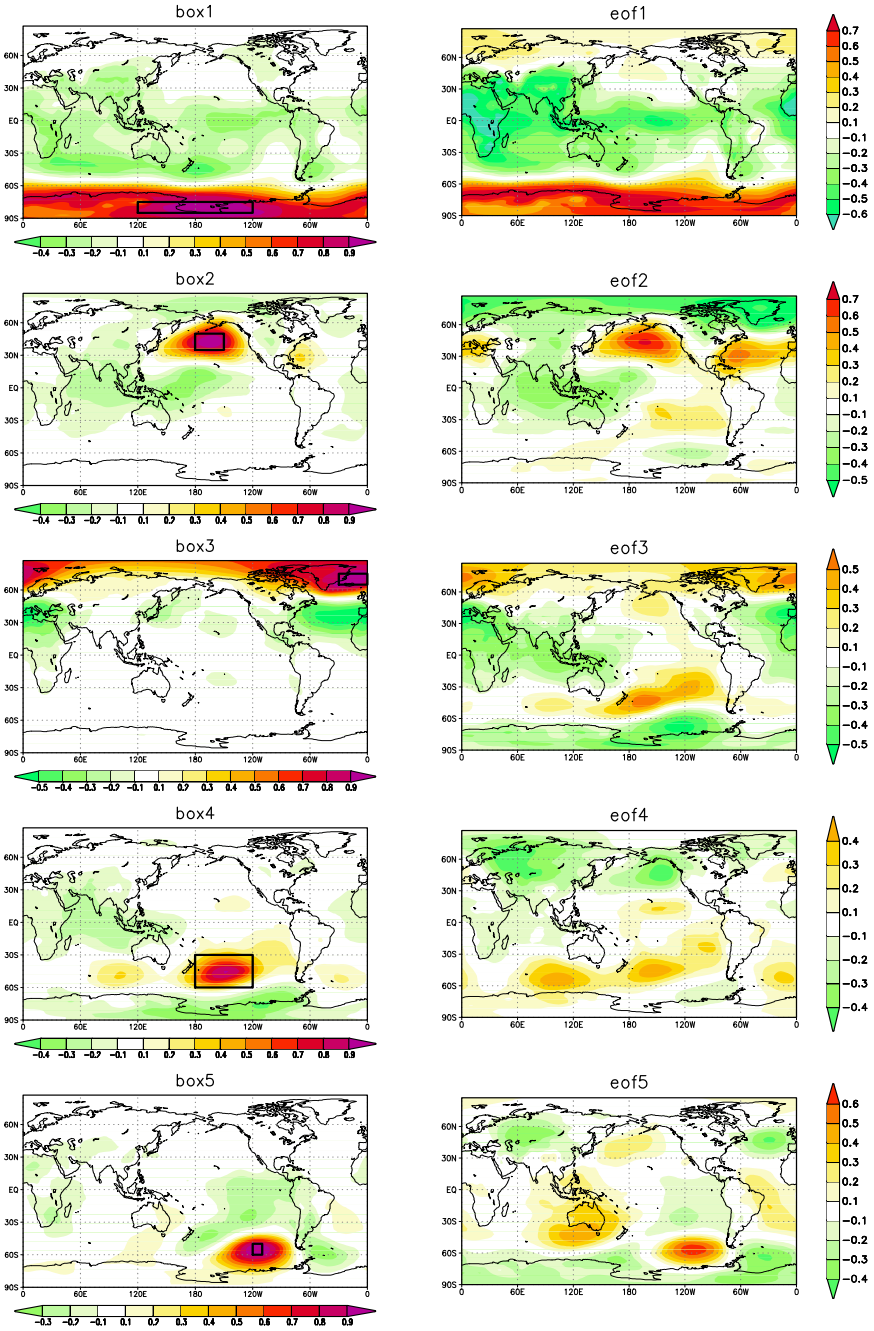


Figure 13.8: Box correlations and EOF-analysis of global monthly mean SLP anomalies. The left column shows 5 different box correlations. The right column shows the correlation patterns of the 5 leading EOF-modes of the domain. Values are non-dimensional.

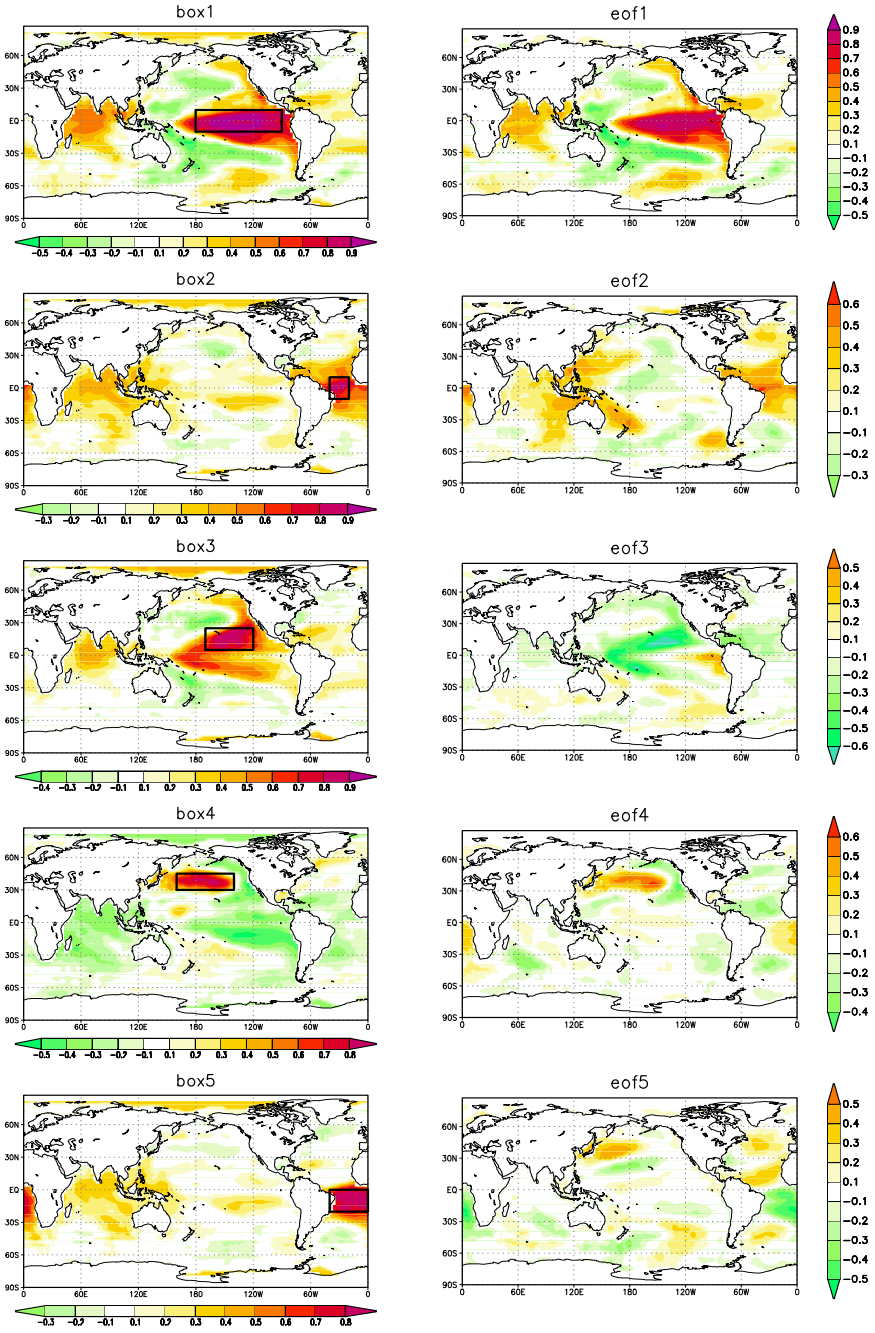


Figure 13.9: Same as Fig. 13.8 but for global monthly mean SST anomalies.

The stochastic model of Calahan et al. (1996) is essentially given by the correlation between two spatial locations of the data field, Φ:

$$c(r) = e^{-r/d_0} \tag{13.16}$$

Here r is the distance between the two locations and d_0 is the decorrelation length. Note that eq.[13.16] is the equivalent to eq.[13.14]. Thus the stochastic model of Calahan et al. (1996) is an AR(1)-process in the spatial domain dimension.
The simple physical model in eq.[13.13] can be extended to include diffusion for the relation between two locations:

$$\frac{d}{dt}\Phi = c_{damp} \cdot \Phi + c_{diffuse}\nabla^2\Phi + f \tag{13.17}$$

$c_{diffuse}$ is a diffusion coefficient and f now represents spatial and temporal white noise. In this equation the diffusion is just introduced in a statistical sense. This diffusion model is often refered to as a simple energy balance models of the climate system (see e.g. North et al. 1981 and 1983 and references there in). Leung and North (1991) discussed some statistics of this model for the atmospheric variability of a zonally symmetric planet. North (1984) finds that the EOFs of this model driven by homogenous forcing f (spatially white noise with the standard deviation of f constant over the domain) coincide with the eigen modes of the dynamic operator of the system.
Note that for an isotropic diffusive process (neither c_{damp} nor $c_{diffuse}$ are a function of the location) driven by a homogenous forcing f, the model in eq.[13.17] is an AR(1)-process in the spatial domain. We can derive the covariance matrix of Φ:

$$\Sigma_{ij} = \sigma_i\sigma_j e^{-d_{ij}/d_0} \tag{13.18}$$

where σ_i is the standard deviation of Φ at point i and d_{ij} the spatial distance between the two points i and j. If the standard deviation field σ or d_0 exhibit spatial variations (e.g. $\sigma_i \neq \sigma_j \; for \; i \neq j$), than the model in eq.[13.17] is not a spatial AR(1)-process any more and eq.[13.18] does not exactly represent the covariance matrix of Φ. However, eq.[13.18] should be a good approximation if the spatial variations of σ_i and d_0 are small. The effect of spatial variations of σ will be discussed in section 13.2.6 by means of a realistic example.
An isotropic diffusive process in equations [13.17] and [13.18] is the null hypothesis for the spatial characteristics of a climate variable Φ. In this formulation, Φ has no teleconnections other than the exponential decay of its auto-correlation function. In analogy, the spectrum of a time series of an AR(1)-process, is not considered to have a significant time scale (peak in the spectrum) other than a damping time scale.
We can find the EOF modes and eigenvalues of the null hypothesis numerically. In Fig. 13.10 the leading EOF modes of a domain defined by 17x11 points with constant $\sigma = 1$ and $d_0 = 4.6$ points is shown. The eigenvalues of the leading EOF modes are also shown in Fig. 13.10.
Based on this example and a few other examples with variations in the domain dimensions and d_0 (not shown), a few important characteristics of the EOF modes and eigenvalues of the diffusion null hypothesis can be formulated as follows:

- The EOF modes are a hierarchy of multi poles, starting with a monopole as EOF-1, followed by a dipole, and then by higher order multi poles. The order and structure of the multi poles is a result of the domain dimensions and the decorrelation length d_0. Note that this kind of structure of the observed leading EOF modes of most climate data sets is also discussed in Richman (1986), but not in the context of the simple stochastic model in [13.17].

- The EOF-1 peaks in the center of the domain, because the center point is the point which is in average closest to all other points and has therefore a larger covariance with all other

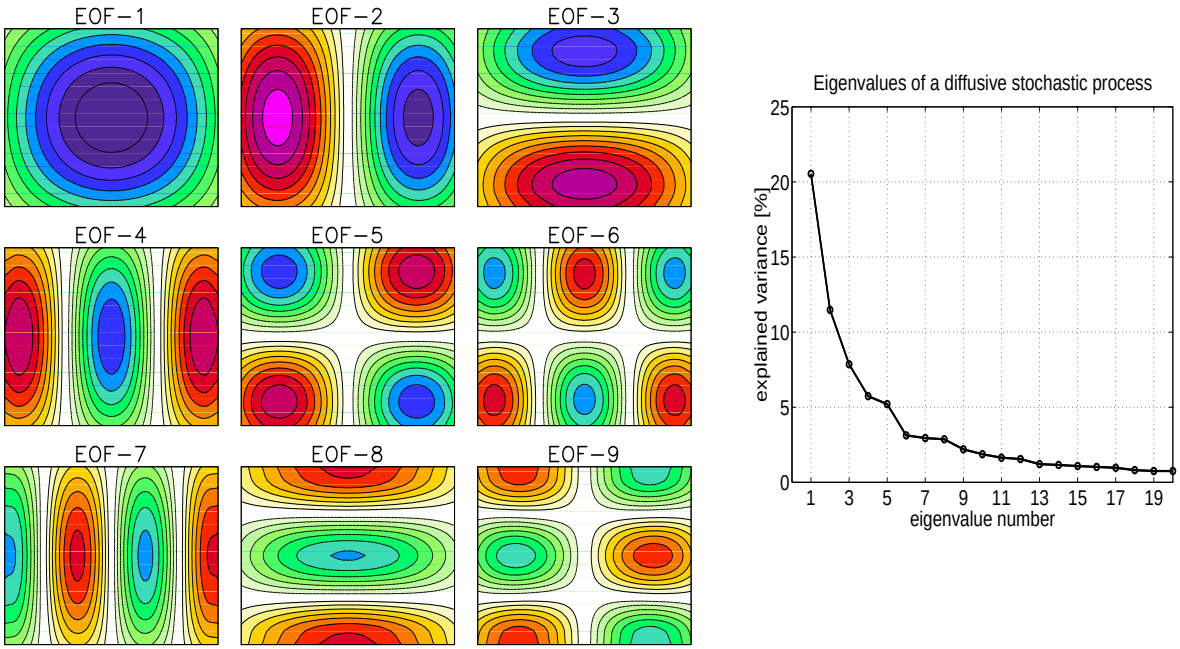


Figure 13.10: The leading EOF modes (left panels) and eigenvalues (right panel) of a spatial AR(1)-process in a 17×11 points domain.

points. Note that σ and d_0 are identical for all points, so that the statistics of all points of the domain are identical. The EOF-1 mode is therefore only a reflection of the domain geometry. It simply reflects that there is no structure in the variability other than exponential decay of covariance with distance.

- None of the EOF modes represent teleconnections (factors), since no teleconnections exist in this simple model. In the simple model of an spatial AR(1)-process the spatial variability is a continuous spectrum of spatial patterns, where no spatial pattern is dominating over the other patterns. The EOF modes should be interpreted as a reflection of different spatial scales. In analogy, the spectral coefficients of a continuous spectrum of an AR(1)-process are a reflection of the different time scales, but not a representation of an oscillating behavior. The domain wide monopole of the leading EOF-1 represents the largest spatial scale of variability in the domain, which in an AR(1)-process has the largest variance. EOF-2 and EOF-4, for instance, should be interpreted as spatial variability along the x-axis with a spatial length scale of about 1/2 of the domain size along the x-axis. They do not represent an anti correlation between the centers. The same holds for all other EOF modes.

- The decrease of the eigenvalues to higher order EOFs is only a function of the domain size and the decorrelation length d_0. None of the 20 leading eigenvalues is degenerated (equal to another eigenvalue), reflecting the different length of the domain along the x- and y-axis. Note that in this example the number of points in each direction was chosen as a prime number to avoid degenerated eigenvalues, which in real domains, such as ocean basins, would not occur. Note also that the numerical precision of the EOF analysis in this example is much better than the line (dot) thickness in Fig. 13.10.

  An important quantity that quantifies the degree of complexity in the domains spatial vari-

ability is the effective spatial number of degrees of freedom N_{eff} (Bretherton et al. 1999). It essentially estimates the effective dimension of the multivariate variability:

$$N_{eff} = \frac{1}{\sum e_i^2} \ , \quad with \qquad \sum e_i = 1 \tag{13.19}$$

with e_i the eigenvalues as derived from the EOF analysis. The number N_{eff} corresponds to the number of independent spatial modes. It also quantifies the decrease of the eigenvalues and is a monotonic function of the decorrelation length d_0. It can therefore also be used as an estimate for the decorrelation length.

13.2.4 Evaluating EOF-vectors and eigenvalues against a stochastic null hypothesis

The stochastic model allows an evaluation of EOF modes. In many studies only the leading EOF modes of an observed data set are discussed. Here the focus is often on the spatial structure of the observed EOF patterns, which are interpreted as teleconnection patterns. It is therefore important to discuss to which degree the leading EOF modes are consistent with the simple null hypothesis or, in a more objective approach, to find those patterns, that are most distinguished from those of the null hypothesis.

Fitting an isotropic diffusion process to data

The null hypothesis as formulated in the previous section can be fitted to any data set by estimating the standard deviation field σ and the average decorrelation length d_0. Given σ and d_0 the covariance matrix of the null hypothesis in eq.[13.18] is defined and the EOF modes of the null hypothesis can be calculated.
Note that the estimation of d_0 can have large uncertainties in a limited gridded domain (see e.g. Storch and Zwiers, 1999). However, d_0 is a monotonic function of the spatial number of degrees of freedom, N_{eff}, which is estimated by the sum of eigenvalues. The estimation of d_0 will usually depend on the correlation of neighboring points, which is a function of the variability on all spatial scales. The estimation of N_{eff} is essentially a function of the leading EOF-modes only, while the small scale variability has little effect on this quantity. Hence, the agreement between the leading eigenvalues of the observations and the fitted null hypothesis appears to be better if the observed N_{eff} is used to estimate the fitted d_0 in (13.18). An analytical relation between N_{eff} and d_0 may exist for some simple domain geometries, such as a sphere for instance. However, it will be difficult to write down an analytical relation for complicated geometry and boundary conditions and it may therefore be most practical to estimate these quantities numerically. Thus, d_0 is varied until N_{eff} of the fitted null hypothesis agrees with the N_{eff} of the observational data set within the uncertainty range of N_{eff}, which should be given by the statistical uncertainties of the eigenvalues due to sampling errors (North et al. 1982).

Comparing the observed EOF modes with a null hypothesis

An EOF eigenvector (mode) of an observed data set, $\vec{E}_i^{obs}$, and the corresponding eigenvalue e_i^{obs} can be compared to the eigenvectors $\vec{E}_j^{null}$ and eigenvalues e_j^{null} of a null hypothesis by projecting the eigenvectors $\vec{E}_j^{null}$ onto the eigenvector $\vec{E}_i^{obs}$.

$$c_{ij} = \frac{\vec{E}_i^{obs}\vec{E}_j^{null}}{|\vec{E}_i^{obs}||\vec{E_j^{null}}|} \tag{13.20}$$

c_{ij} is the uncentered pattern correlation coefficient between the two EOF-patterns. The variance that the mode $\vec{E}_i^{obs}$ would have under the null hypothesis can be estimated by the linear combination of all eigenvalues e_j^{null} of the null hypothesis using c_{ij}:

$$e_i^{obsnull} = \sum_{j=1}^{N} c_{ij}^2 e_j^{null} \tag{13.21}$$

The variance $e_i^{obsnull}$ is the expected variance of $\vec{E}_i^{obs}$ if the data follows the diffusive process of the null hypothesis. Note that while the eigenvalues e_i^{obs} decrease monotonically with higher order numbers, the $e_i^{obsnull}$ values does not need to decrease with higher order number. A pattern that

explains a lot of variance in the observations (large e_i^{obs}) may explain little variance under the null hypothesis (small $e_i^{obsnull}$ values).

13.2.5 Statistical inferences about the nature of EOF modes

The uncertainties of the eigenvalues e_i^{obs} of the observed data due to sampling errors are given by North et al. (1982). When the observed data follows the null hypothesis we expected the $e_i^{obsnull}$ value to be within the uncertainties of the eigenvalues e_i^{obs}. A comparison of the eigenvalue spectrum e_i^{obs} with the spectrum of $e_i^{obsnull}$ allows to quantify the deviations of the observed data from the null hypothesis, which can be the basis for statistical inferences about the nature of EOF modes. The concept is in analogous to the comparison of the spectrum of an observed time series with the spectrum of the fitted AR(1)-process.
e_i^{obs} and $e_i^{obsnull}$ are variances, which tend to be χ^2-distributed. Statistical inferences about χ^2-distributed random variables are usually obtained on the basis of the ratio, $e_i^{obs}/e_i^{obsnull}$, as in time series analysis (e.g. Reynolds 1978, Dommenget and Latif 2002b). However, as mentioned in Calahan et al. (1996) , the strongest deviations of the ratio, $e_i^{obs}/e_i^{obsnull}$, are found in the low (higher order) eigenvalues, which are in most studies of little interest. It therefore may be more instructive for large-scale teleconnections to base the statistical inferences on the difference between e_i^{obs} and $e_i^{obsnull}$. However, the choice of the right test variable depends on the focus of the analysis. The method of projecting the null hypothesis onto observed patterns can be used for all kind of patterns, like box-averages or more sophisticated indices. The explained variance of the index compared to the explained variance $e_{obsnull}$ could reveal whether the index indeed presents an unexpected structure and thus can be used to justify a specific choice of indices.

13.2.6 DEOFs: An estimate of teleconnection modes

If the $\vec{E}_{obs}$ appear to be different from the null hypothesis one may be interested in the spatial pattern that maximizes the difference in explained variance between the data and the null hypothesis. These are named distinct EOFs (DEOFs or $\vec{D}^{obs}$) and distinct PCs for the time series (DPCs), respectively. The leading $\vec{D}^{obs}$ is defined as the pattern that maximizes the differences in explained variance Δ_{var}:

$$\Delta_{var} = Var_{obs}(\vec{D}^{obs}) - Var_{null}(\vec{D}^{obs}) \tag{13.22}$$

Where Var_{obs} denotes the variance that the pattern $\vec{D}^{obs}$ explains in the observed data and Var_{null} denotes the variance that the pattern $\vec{D}^{obs}$ explains under the null hypothesis following [13.21]. The leading $\vec{D}^{obs}$ can be found by pairwise rotation of the leading EOFs, as it is done for determining the VARIMAX modes (Kaiser 1958), until the maximum of Δ_{var} is found. By iterating this procedure we can define a complete set of orthogonal DEOFs, building a complete representation of the data. The patterns that are most distinguished from the null hypothesis, the leading DEOFs, are, from a statistical point of view, a good first guess for the teleconnections. They should in general be a good starting point for the understanding of the underlying physical processes. Identifying the DEOFs with the teleconnections, however, depends strongly on the formulation of the null hypothesis.
The DEOF have, however, some limitation in the interpretation, that are similar to those pointed out for EOFs in section 13.2.3. The concept of teleconnection patterns may not always be helpful in the understanding of the multivariate data. In some systems, the DEOFs may be a reflection of deviations from the isotropic diffusion, that are better described by physical process parameter, such as anisotropy in the diffusion or advection. The DEOFs focus on the deviations in the leading (large) eigenvalues, while differences in the higher order (small) eigenvalues are neglected, which in some systems may be important for the understanding of the underlying physical processes (Crommelin and Majda 2004).

The DEOF are defined in an orthogonal system, which, similar to EOFs, maximize some variance criteria. Therefore it can be difficult to interpret the DEOFs in systems where many DEOFs explain more variance than expected under the null hypothesis (see also the discussion of the observed Northern Hemisphere and tropical SLP variability in sections 13.2.6, 13.2.6 and 13.2.7). Note that due to the limited length of the time series the expected value of $\Delta_{var} \neq 0$ when the data follows the null hypothesis. For statistical inference about the significance of Δ_{var} of the leading DEOF we have to estimate the probability distribution function (PDF) of Δ_{var}. The simplest way to estimate the PDF of Δ_{var} of the leading DEOF is by means of a bootstrapping approach (see von Storch and Zwiers 1999).

Examples

We start with some artificial examples, in which the true nature of the problem is well defined. The two artificial examples illustrate two different ways in which a multivariate data set can differ from a pure isotropic diffusion process. We shall then discuss several examples of observed climate variability, some of which have led to some controversy in the recent literature. In the discussion of all examples, the null hypothesis of the climate variability is an isotropic diffusion process as formulated in section 13.2.3 and the parameters are fitted to the data as described in section 13.2.4. The discussion of the observed climate variability modes will be brief and focuses on the new technique. A more detailed physical analysis of the observed climate variability modes may be desirable, but would be beyond the scope of this paper.

An isotropic diffusive field with inhomogeneous standard deviation

The following example is a numerical stochastic realization of eq.[13.17]., which should be similar in its statistics to typical monthly mean time series of ocean basin SSTs. The model in eq.[13.17] was integrated on a grid with 18×18 points with a daily time step. The diffusion coefficient was chosen to produce a decorrelation length of about 3 points. For the statistical analysis, 30 time steps were averaged to build a monthly mean and the damping c_{damp} in eq.[13.17] was chosen to create a one month lag correlation of about 0.6. The resulting time series has a length of 1000 months with about 500 degrees of freedom. The standard deviation of the spatially uncorrelated forcing f was increased in two regions, with one peak in the northeast and one in the southwest. The resulting standard deviation of Φ varies between 1.0 and 4.0 (at the peaks) in arbitrary units. The EOF analysis was performed on only the central 10×10 domain to avoid boundary effects.
In Fig. 13.11a-c, the leading EOFs of the stochastic simulation are shown. In addition, the leading EOFs of the covariance matrix Σ based on eq.[13.18] with parameters σ_i and d_0 fitted to the statistics of the simulation are shown for comparison (Fig.13.11d-f). The explained variance of the EOF modes of the simulation, e_i^{obs}, are compared with the fitted isotropic diffusion process by projecting the EOF modes of the null hypothesis onto the EOF modes of the stochastic simulation as outlined in section 13.2.4 using eqs. [13.20] and [13.21] (see Fig.13.11g). Note that while the eigenvalues e_i^{obs} decrease monotonically with higher order numbers, the variance under the null hypothesis, $e_i^{obsnull}$, does not need to decrease with higher order number, because these are not eigenvalues (see section 13.2.4).
In a first comparison we find that the EOF modes and eigenvalues of the stochastic simulation are in good agreement with the fitted null hypothesis. If we rotate towards the leading vectors $\vec{D}^{obs}$, which point towards the largest differences, the two leading modes (see Fig. 13.11h,i) reveal some significant structures (no other higher order mode shows any significant differences). These two structures represent some differences in the spatial scale of variability near the two centers of the leading EOF mode. They reflect that eq.[13.18] is only an approximation if structures in σ or d_0 exist. In this example the structure introduced in the standard deviation of the white noise forcing f leads to some structure in the standard deviation of Φ and it also creates some variations in d_0.

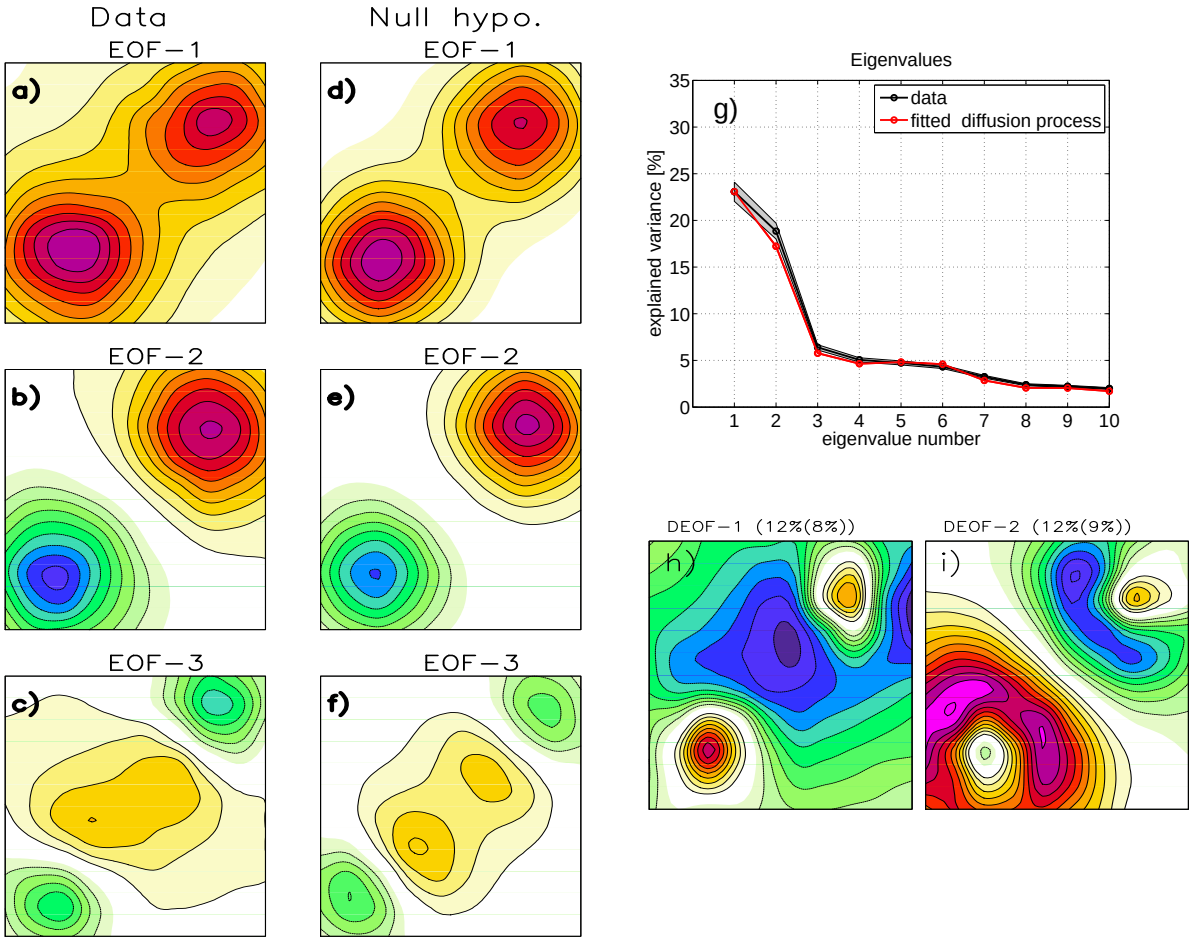


Figure 13.11: The leading EOFs of the stochastic simulation (panels a-c) and the EOFs of the fitted isotropic diffusion process (panels d-f) of the example in section 13.2.6 are shown. The eigenvalues e_i^{obs} (black line) are compared with the projection of the fitted diffusion process $e_i^{obsnull}$ (red line) in panel g). The shaded envelope around the black line is the statistical uncertainty of the eigenvalues e_i^{obs} due to sampling errors after North et al.(1982). The panels h and i show the leading DEOF-1 and DEOF-2. The first percentage value in the heading of the panels h and i give the explained variance of the DEOF in the stochastic simulation and the second value the explained variance of the DEOF under the null hypothesis. All spatial modes are in arbitrary units.

However, the difference between the stochastic simulation and the null hypothesis amounts to only 4% for the leading mode.
If we repeat the experiment with a homogenous standard deviation of the forcing f, the significant structure in the leading vectors $\vec{D}^{obs}$ is gone (the difference in explained variance is $< 2\%$, decreasing with the length of the time series).
In summary, the EOF modes and eigenvalues are close to those of the null hypothesis.

A diffusive field with a weak teleconnection pattern

Here the standard deviation of the forcing f is homogeneous through out the domain. In addition to the spatially and temporally white noise forcing f, a teleconnection forcing pattern π was introduced in eq.[13.17] leading to the following equation:

$$\frac{d}{dt}\Phi = c_{damp} \cdot \Phi + c_{diffuse}\nabla^2\Phi + \pi \cdot F + f \tag{13.23}$$

The spatial pattern of π is shown in Fig. 13.12a, where F is a white noise time series with a variance of about 12% of the variance of f. The teleconnection forcing pattern π is therefore relatively weak. Fig. 13.12b,c highlights the correlation between the two centers of the teleconnection forcing pattern by means of box correlations, showing only a weak correlation in Φ between the two centers. For the EOF analysis, the data were normalized so that each point has unity standard deviation. As a result the stochastic simulation reflects a domain which has no structure in the standard deviation of Φ, but it has a structure in the covariance matrix forced by a teleconnection pattern. Fig. 13.12

The EOF modes and eigenvalues are shown in Fig. 13.12d-k. The EOF modes are very similar to those of a purely diffusive process as discussed in section 13.2.3, but that the eigenvalue of EOF-2 is larger than expected by a diffusive process. The leading mode of the rotation towards the largest difference relative to the fitted isotropic diffusion process, DEOF-1, is very similar to the teleconnection forcing pattern. DEOF-1 explains 18% of the total variance, where this pattern would only explain 10% in the fitted AR(1)-process (see Fig. 13.12l). Thus the residual of about 8% of the total explained variance may be associated with a teleconnection following the spatial structure of DEOF-1. Note that none of the leading VARIMAX modes (not shown) have any similarity to the teleconnection π, because the structure of the teleconnection forcing pattern (a dipole) does not maximize the VARIMAX criteria 'simplicity'.

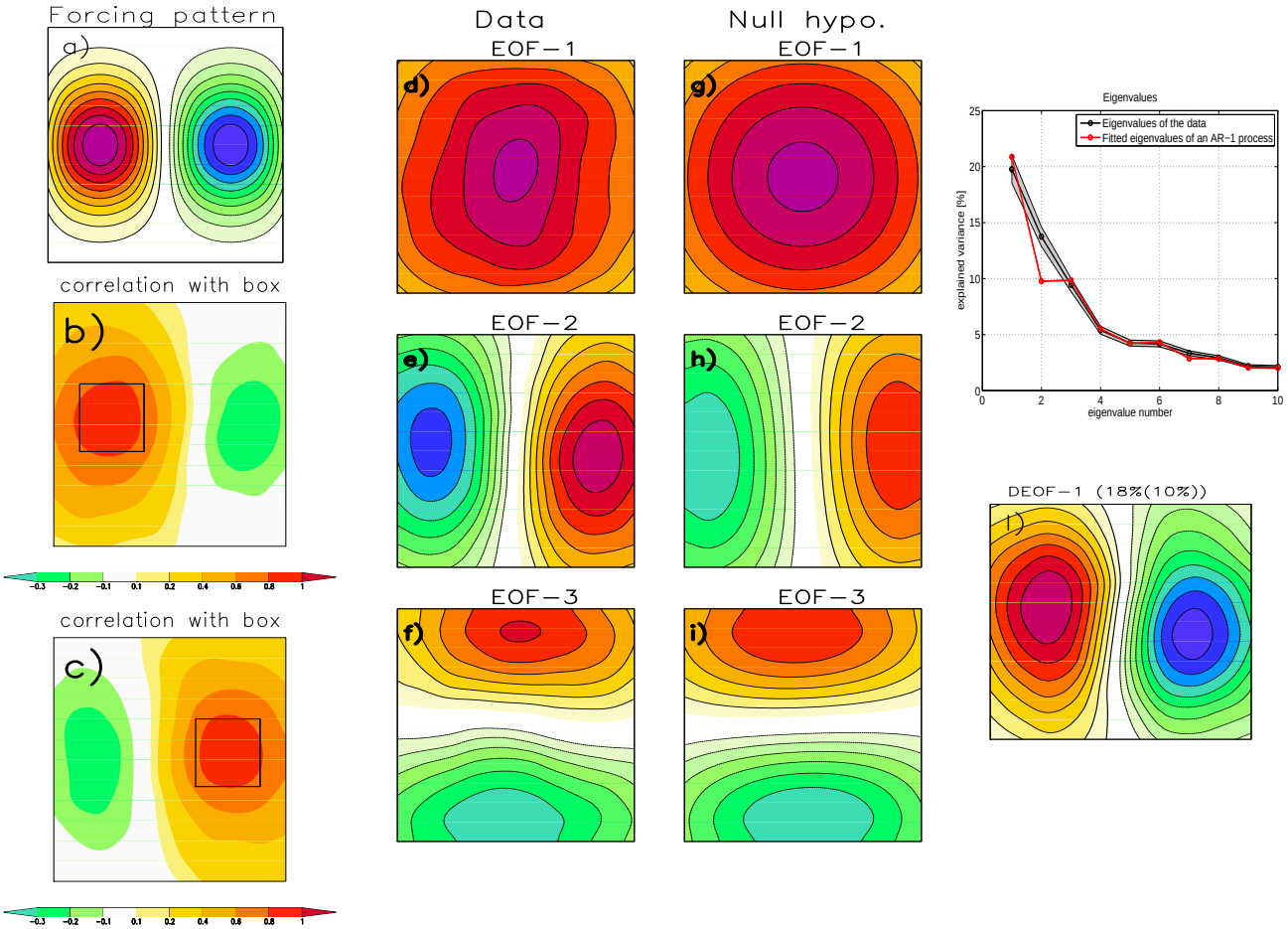


Figure 13.12: As in Fig. 13.11 but for the example in section 13.2.6. In addition the forcing pattern π is shown in panel a, and the box-average correlations of two regions with the rest of the domain are shown (panel b & c).

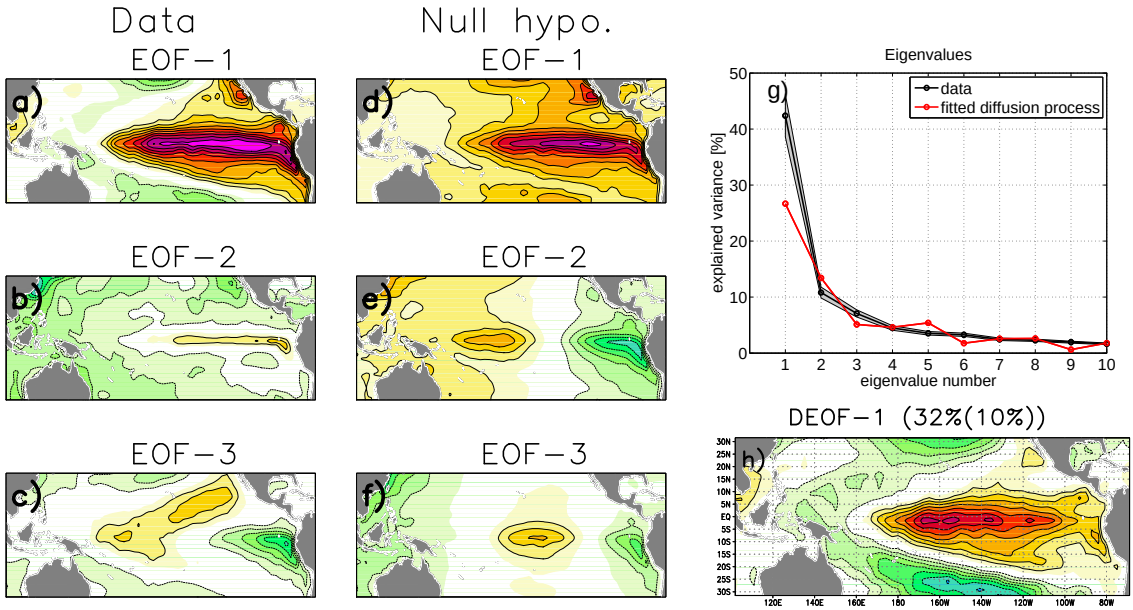


Figure 13.13: As in Fig. 13.11 but for the tropical Pacific SST as discussed in section 13.2.6.

Tropical Pacific SST

The first example of observed data is the tropical Pacific (from $30^oS - 30^oN$ to $100^oE - 70^oW$) monthly mean SST as presented by the HADISST data set from 1870-2003 (Folland et al. 1999). The ENSO mode in the tropical Pacific is probably the best understood teleconnection mode of natural global climate variability and is therefore a good example on which to apply the analysis introduced in this paper. Whether or not the ENSO mode is stochastically forced, as assumed by the null hypothesis, or due is to intrinsic chaotic behavior, will not be addressed in this work (Kirtman et al. 2005 and references therein).

The three leading EOFs are compared with the EOFs of the fitted null hpyothesis in Fig. 13.13a-f. Fig. 13.13 The structures of the leading EOFs of the observed SST are quite different to those of an isotropic diffusion process. The comparison of the variance of the eigenvalues shown in Fig. 13.13g clearly shows that nearly all leading EOFs are different from the null hypothesis.

If we maximize the difference between the observed EOF modes and the null hypothesis by rotation we find a pattern similar to EOF-1 (see Fig. 13.13h), with less explained variance (32%) but with a much larger difference relative to the null hypothesis of about 22%, which makes this mode more distinct to a diffusive process than EOF-1. It is also interesting to note that the leading teleconnection DEOF-1 is more focused on the central equatorial region than the EOF-1, a region often discussed in ENSO forecasting studies to be the most predictable region on seasonal to interannual time scales (e.g. Barnett et al. 1993, Dommenget and Stammer 2004).

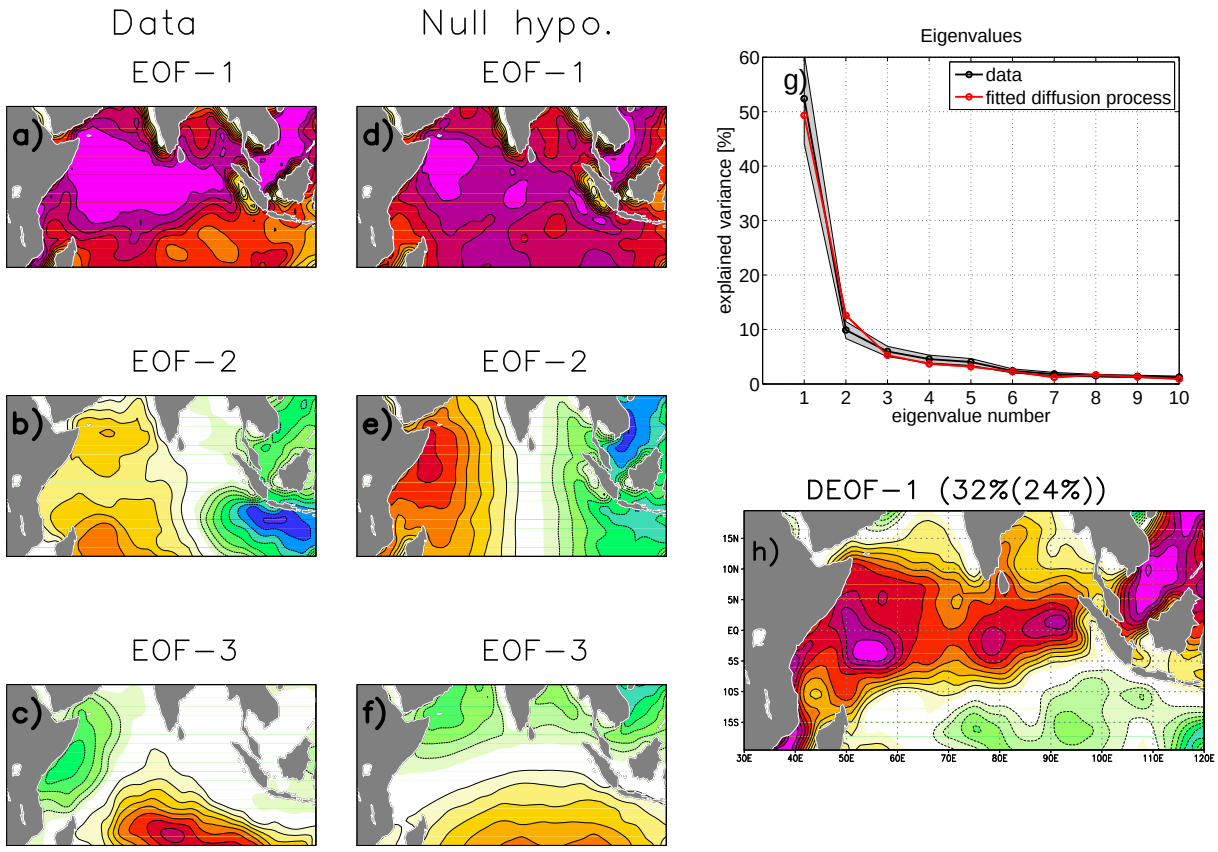


Figure 13.14: As in Fig. 13.11 but for the tropical Indian Ocean SST as discussed in section 13.2.6.

Tropical Indian Ocean SST

The EOF-1 of the monthly mean Indian Ocean SST ($20^oS - 30^oN$ to $30^oE - 120^oE$) from the HADISST data set over the period from 1870 to 2003 (Folland et al. 1999), as shown in Fig. 13.14a, has been identified as the response of the Indian Ocean to ENSO by Saji et al.(1999). They further identify the EOF-2 as a new mode of ocean-atmosphere interaction in the Indian Ocean. A discussion of whether or not this interpretation is justified can be found in Baquero and Latif (2002), Behera et al. (2003) and in Dommenget and Latif (2003).

The spatial structure of the leading EOFs appear to be very similar to the EOFs of the null hypothesis and the leading eigenvalues are in good agreement with the variance of the null hypothesis (see Fig. 13.14a-g). It therefore seems that the SST variability of the Indian Ocean is consistent with a purely diffusive process. In particular, EOF-2, the so called Indian ocean dipole mode, explains less variance than expected from the fitted AR(1)-process.

Although there is no indication for strong deviations from an isotropic diffusion process, a rotation towards the leading differences from the null hypothesis was performed (see Fig. 13.14h). However, DEOF-1 explaining 32% of the total variance (24% in the null hypothesis) is only slightly different from the null hypothesis.

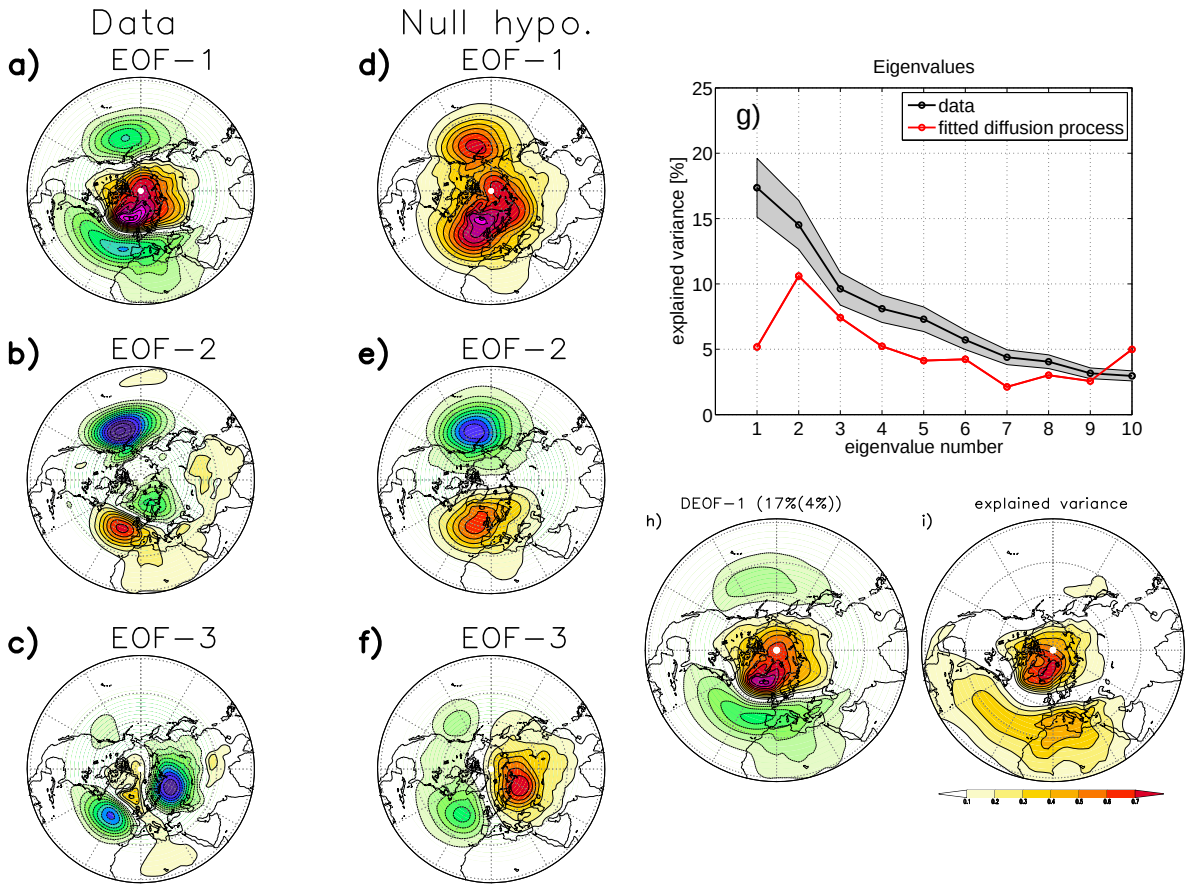


Figure 13.15: As in Fig. 13.11 but for the Northern Hemisphere wintertime SLP as discussed in section 13.2.6. In addition the explained variance field of DEOF-1 is shown (panel i).

Northern Hemisphere winter time SLP

Thompson and Wallace (1998) proposed the Arctic Oscillation (AO) or the annular mode as one of the leading modes of climate variability, which is defined by EOF-1 of Northern Hemisphere (from 10^oN) wintertime SLP. The leading EOFs are shown in Fig. 13.15a-c based on the NCEP SLP from 1948-1999 (Kalney et al. 1996). A teleconnection between the Pacific and Atlantic region is seen in EOF-1. Fig. 13.15

However, Deser (2000) and Ambaum et al. (2001) pointed out that the AO does not project onto local correlation patterns as well as the two more localized patterns of the North Atlantic Oscillation (NAO) and the Pacific North America pattern (PNA), which are the leading EOFs of the Atlantic and Pacific sub domains. The NAO and the PNA both project well onto the AO pattern. Further, they argue that the data does not give much support for strong interactions between the Atlantic and Pacific region as the AO pattern suggests.

In response to the lack of correlation between the two oceans, Wallace and Thompson (2002) argue that the EOF-2 may represent another inter-oceanic mode of variability, which leads to the apparent weak correlation between the SLP over the two oceans. In summary, Thompson and Wallace indicate that there is a relatively strong connection between the Atlantic and Pacific regions, whereas Deser (2000) and Ambaum et al. (2001) do not see evidence for this connection. Furthermore, the arguments of Deser (2000) and Ambaum et al. (2001) are similar to the arguments which lead to the null hypothesis. Assuming the leading modes of variability should be reflected in the local correlation patterns, as the two more localized patterns NAO and PNA, is in principle

the same as assuming that the data are dominated by diffusive processes and a few (one or two) teleconnections.

The leading EOFs of observed wintertime SLP are quite different from those of the null hypothesis in that each of the leading EOFs explains considerably more variance than it would under the null hypothesis (see Fig. 13.15a-g). The comparison therefore indicates that the wintertime SLP is inconsistent with diffusive processes. However, the leading teleconnection DEOF-1 is quite clearly represented by a NAO like structure explaining about 17% (4% in the null hypothesis) of the total variance (see Fig. 13.15h,i). Note that this pattern has a high correlation with the EOF-1 or AO mode, but it explains very little variance in the Pacific region (see Fig. 13.15i).

Note that one should resist in interpreting all the DEOFs, that explain more variance, than expected under the null hypothesis, as teleconnection patterns. In multivariate systems with many DEOFs explaining more variance than expected under the null hypothesis, the interpretation of the DEOFs can be very difficult and the concept of teleconnection modes may not be very helpful. It may in some cases be possible to identify some of the DEOFs with teleconnections, but one have to keep in mind that in a multivariate orthogonal system, rotation of the dominant DEOFs patterns may lead to a different presentation of the leading teleconnections. Moreover, the DEOF will in most cases not represent any coherent teleconnections, but be a reflection of dominant physical process that drive SLP in the extra tropics, such as mass and vorticity conservation.

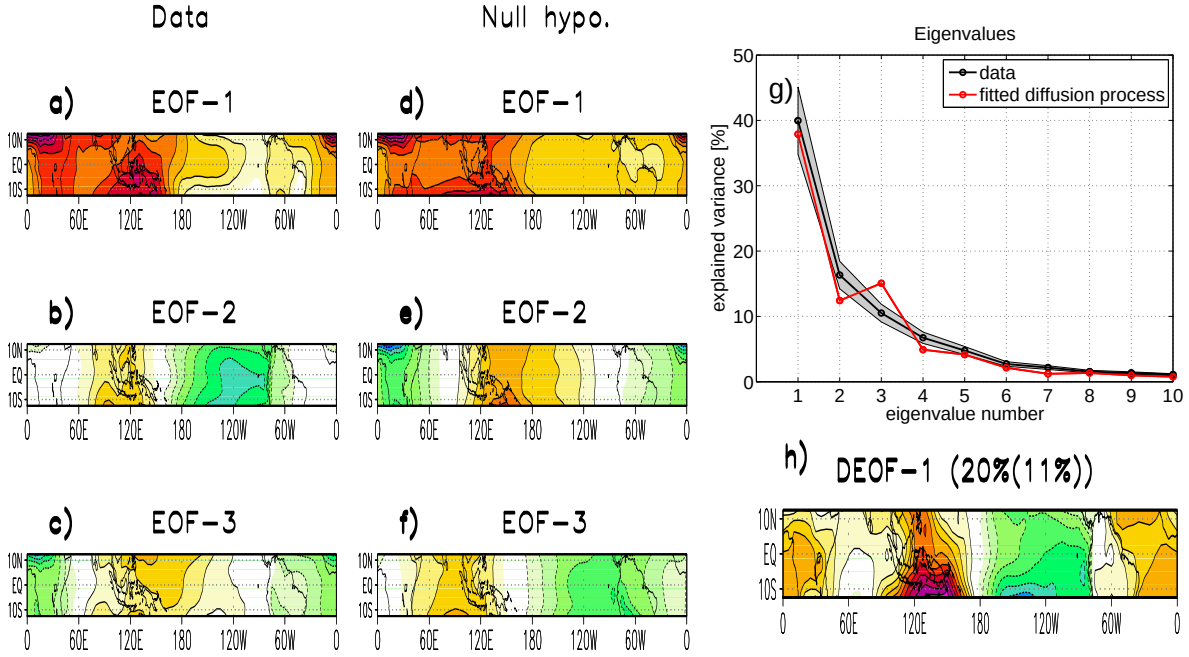


Figure 13.16: As in Fig. 13.11 but for the tropical SLP (from $15^oS - 15^oN$ to $0^oE - 360^oE$).

Tropical SLP

Tropical monthly mean SLP variability is strongly related to SST variability, which is dominated by the ENSO-mode. While the El Niño SST pattern is well represented by the leading EOF of the tropical Pacific SST, the Southern Oscillation mode is not well represented by the leading EOF of the tropical SLP, see Fig. 13.16a . The Southern Oscillation has some similarity with EOF-2, but is usually defined by the correlation with the NINO3 region ($5^oS - 5^oN/150^oW - 90^oW$) or as the pressure difference between the stations at Darvin and Tahiti. Fig. 13.16

The structure of the EOF-patterns is similar to what is assumed for a pure isotropic diffusion process (as discussed in section 13.2.3), with a monopole as EOF-1 followed by dipole patterns in EOF-2 and EOF-3. Thus the structure or the EOF-patterns does not suggest any characteristic teleconnection pattern. The important role of the EOF-2 (Southern Oscillation) becomes clear, when the eigenvalues of the EOFs are compared with the fitted null hypothesis (Fig.13.16g). Overall the eigenvalues of the EOFs are relatively close to those of the fitted null hypothesis, but the EOF-2 explains considerably more variance than expected, while EOF-3 explains much less variance than expected. The situation is similar to the artificial example with a zonal dipole teleconnection, as discussed in section 13.2.6.

The leading DEOF is similar to the EOF-2 (Southern Oscillation), but is more global, with larger amplitudes in the tropical Atlantic region. In the context of the ENSO mode we would expect the leading SLP in the tropical atmosphere to be correlated to the SST. The DPC-1 shows higher correlations with the PC-1 and DPC-1 of the SST in the tropical Pacific (as discussed in section 13.2.6) than the PC-2. It also shows larger correlations with global SST than the PC-2, including the North Pacific, tropical Indian Ocean and Atlantic, which are region known to be influenced by the ENSO mode.

The tropical SLP also shows some clear anisotropy in the decorrelation length. In the zonal direction the decorrelation length is much larger than in meridional directions. This deviation from the

isotropic diffusion process becomes more dominant in the leading DEOFs, if the analysis is repeated on a wider latitudes range (e.g.$30^oS - 30^oN$; not shown). The mismatches between the leading eigenvalues and the fitted null hypothesis become larger, but the EOF-2 (Southern Oscillation) remains to be the largest deviations from the null hpyothesis. The southward shift of the amplitudes in the DEOF-1 (Fig.13.16h)is also a reflection of the anisotropy in decorrelation length.

13.2.7 Discussion of the Evaluation of EOFs against a Stochastic Null Hypothese

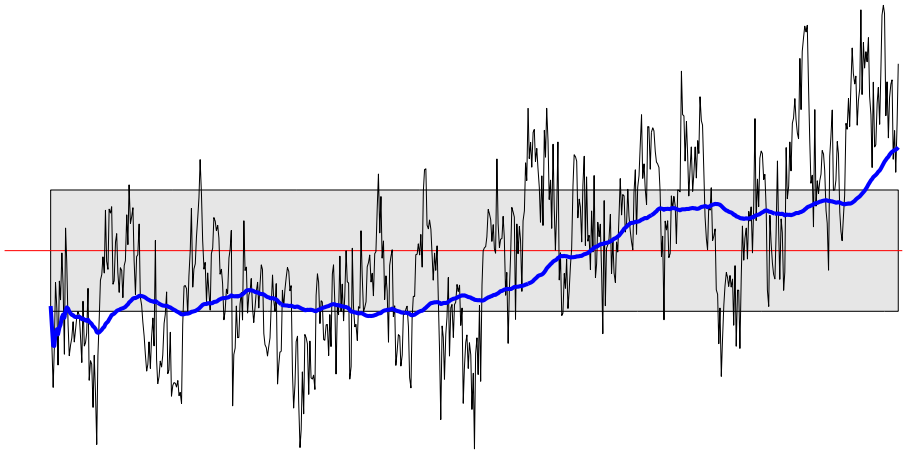
In this paper it is suggested that the leading EOF modes of observed data are compared with the EOF modes of a fitted stochastic null hypothesis in order to determine what the nature of the spatial structures of the data are. Calahan et al. (1996) formulated a simple stochastic model for rainfall data, which can be used as a general null hypothesis for the spatial structure of climate fields. The stochastic model of Calahan et al. (1996) is an AR(1)-process in the spatial dimension, which is the same as the null hypothesis for the temporal dimension (time series) as introduced by Hasselmann (1976). The spatial AR(1)-process can be described by a simple physical model, in which the relation between two spatial locations is only due to isotropic diffusion. The EOF modes of a spatial AR(1)-process are characterized by a hierarchy of multi poles with decreasing eigenvalues. In this simple model the spatial variability is a continuous spectrum of spatial patterns, where no spatial pattern is dominating over the other patterns.
Similar to time series analysis the formulation of this stochastic null hypothesis for the spatial structure of climate variability allows one to compare the EOF modes and eigenvalues of an observed data set with the EOF modes and eigenvalues of a fitted null hypothesis. It also allows to define a representation alternative to the EOF modes, the so called distinct EOFs (DEOFs or $\vec{D}^{obs}$). The leading DEOF is defined as the mode that is most distinguished from the modes of the null hypothesis. It represents the direction in the multivariate space, in which the observed data differs most from the null hypothesis, which may be called the "finger print" of the observed data. It is a good starting point for the understanding of underlying physical processes. However, one should be careful in interpreting the DEOF as a coherent teleconnection pattern. This will in many cases be a misleading interpretation.
Note that in VARIMAX or other criteria for rotation of the EOFs a simple equation, which reflects a predefined symmetry in the system (e.g. simplicity for VARIMAX), is maximized. The rotation analysis will therefore find patterns that follow the assumed symmetry. The DEOFs introduced in the present study are rotated by comparison with a stochastic null hypothesis, which reflects a physical model. The structure of the resulting DEOF-1 is therefore not predefined by any mathematical symmetry. It is only assumed that it is different from the null hypothesis. It can in some cases point to a coherent teleconnection pattern, but it may also be a reflection of physical processes, different from isotropic diffusion, driving the variability of the domain.
As an example the SST of the tropical Pacific was analyzed, which is known to contain the ENSO teleconnection pattern. The comparison with the fitted isotropic diffusion process clearly supports the idea that the El Niño pattern is the leading teleconnection. The rotation towards the leading differences finds a pattern similar to the EOF-1 but more focused in the central Pacific. It is interesting to note that the EOF-1 mode explains 41% and about 34% in the fitted null hypothesis. Thus about 4/5 of the variance of EOF-1 may be explained by the fitted isotropic diffusion process. The leading rotated mode DEOF-1 explains 32% and about 10% in the fitted null hypothesis. If we consider the diffusive part of the fitted null hypothesis as noise, then the leading DEOF-1 has a much better signal to noise ratio, which amounts to 3:1.
In the other example of the tropical Indian Ocean, the SST seems to be much closer to the fitted isotropic diffusion process.
Northern Hemisphere winter time SLP showed that SLP variability is not well described by a pure

isotropic diffusion process. Essentially the entire large-scale structure of Northern Hemisphere SLP deviates from the modes of the fitted null hypothesis. This is somehow not surprising since the large-scale SLP is driven by the quasi-geostrophic equations in which the conservation of absolute vorticity and mass plays an important role, forcing wave like structures (Navarra 1993, Metz 1994, Gerber and Vallis 2005). It is therefore inappropriate to assume that local box correlations should reflect the leading teleconnections, because this already assumes that the main characteristics of SLP is that of a diffusive process. A better strategy appears to be a formulation of a stochastic null hypothesis based not on the isotropic diffusion, but on the quasi-geostrophic equations or simple linearized models (Navarra 1993, Metz 1994, Gerber and Vallis 2005). Comparing the observed EOF modes against the EOF modes of a stochastic quasi geostrophic model will help to decide if the SLP variability has teleconnections with strong links between the Pacific and Atlantic region. The SLP variability of the tropical regions is much closer to the null hypothesis, which may reflect that mass and vorticity conservation are less important in the relatively narrow zonal band of the tropics.

In summary, one should compare the observed spatial patterns to those expected from a simple physical model to evaluate their significance. A good starting point is the isotropic diffusion process, which is the equivalent to the AR(1)-process used in time series analysis.

Part IV

Statistical Inference / Testing Hypothesis



In climate or statistical studies we often want to know thinks like:

- Is the value of $\mathbf{X}(\mathbf{t})$ significantly different from $\mathbf{X_0}$? (The global warming issue)
- Is $\mathbf{X}$ related to $\mathbf{Y}$, as estimated by the correlation? (Teleconnections)
- Is this Peak in the spectrum of $\mathbf{X}(\mathbf{t})$ significant? (Climate Modes)

The concepts needed to address these issues are discussed in this chapter. In this section we will present and discuss the statistical uncertainties of the parameter introduce in the previous sections and we will discuss the concepts of hypothesis testing and how statistical inferences can be made.

Chapter 14

Uncertainties in Statistical Analysis

In statistical analysis we analysis stochastic continuous random variables $\mathbf{X}$, which have a theoretical expected value and a *pdf*. Any sample of $\mathbf{X}$ will be different from another sample due to basically two uncertainties:

- Uncertainties in the estimate due measurement errors or uncertainties in the estimating algorithm (as for the spectrum for instance).
- Uncertainties due to the inherent stochastic nature of the variable. e.g. in throwing a dice the result is uncertain due to the stochastic nature, but not due to errors in the measurement of the value of the dice.

So in statistical analysis we have to deal with some uncertainties that is inherent to the problem and it is not caused by measurement errors. We will therefore always formulate the results of statistical analysis in terms of likelihoods.
Examples may illustrate this:

- Example 1: Lets assume we have a time series of $\mathbf{X}$ and $\mathbf{Y}$ with the length T for both time series. Further we find that the correlation, γ_{xy}, between both is 0.3. The uncertainties due to measurement errors or uncertainties in the estimating algorithm are negligible in general, the correlation for this time interval is 0.3.

  But statistical uncertainties due to the limited time series are more relevant. If we measure $\mathbf{X}$ and $\mathbf{Y}$ again for a time period T_2, independent of T , we do not expect γ_{xy} to be exactly 0.3, but only to be within some uncertainty range. If we estimate the statistical uncertainty of the correlation value due to the limited length of the time series, we can make inferences, about the theoretical expected value of γ_{xy} or the value we expect for γ_{xy} if we measure it again for a time period T_2, independent of T.

- Example 2: see Fig. 14.1. Here we have the spectrum of a time series of $\mathbf{X}$ with length T. We see that the spectrum has a maximum at a certain frequency with a relatively large uncertainty for the variance at this frequency. If we measure $\mathbf{X}$ again for a time period T_2, independent of T , we do not expect that the peak will remain at the frequency, but that it will vary within in some uncertainty range. We would say that the peak is statistically insignificant, but for the time interval at hand it is the frequency with maximum variance.

14.1 The Confidence Interval

We incorporate the likelyhoods of $\mathbf{X}$ into our results by a Confidence Interval. We assume that $\mathbf{X}$ is following a known *pdf*, which we either estimated somehow or we have made an assumption

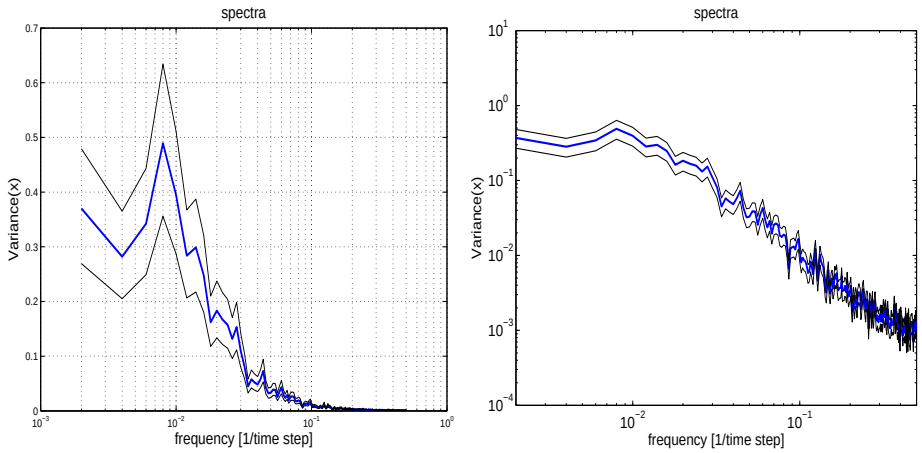


Figure 14.1: The spectrum of an AR(1)-process time series presented in log-linear scaling (left) and in log-log scaling (right). In addition the 10% and 90% quantiles of the spectral coefficients estimate are plotted.

about a theoretical *pdf* (Gauss for instance). We further assume that the sampled value of $\mathbf{X}$, x_t is the expected value of the *pdf* and we build a confidence interval around this value by either the standard deviation:

$$\Omega = [x_t - \sigma, x_t + \sigma] \tag{14.1}$$

if we assume near normal distribution or the quantiles (recall that the 90% quantile, x_{90} was define as $F_X(x_{90}) = 90$, with F_X as the *cdf*, see section 3.5):

$$\Omega = [F_X(x_{10}), F_X(x_{90})] \tag{14.2}$$

in all none-normal distributions. So we treat the confidence interval of $\mathbf{X}$ like error bars. See Fig. 14.4 for examples.

14.2 Uncertainties of the Correlation

When the sample $\{(\mathbf{X}_i, \mathbf{Y}_i)^{\mathrm{T}} : i = 1, ..., n\}$ consists of independent, identically distributed random vectors of length n, a good estimator of the correlation coefficient ρ_{XY} is

$$\hat{\rho}_{XY} = \frac{\widehat{\gamma_{XY}}}{\widehat{\sigma_X}\widehat{\sigma_Y}} = \frac{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^2 \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})^2}} \tag{14.3}$$

This is the maximum likelihood estimator when $(\mathbf{X}, \mathbf{Y})$ is bivariate normally distributed. Furthermore, eq.[14.3] is asymptotically normally distributed with mean ρ_{XY} and variance $(1 - \rho_{XY}^2)^2/n$. However, because $\hat{\rho}_{XY}$ converges slowly to its asymptotical distribution, this result is generally not used to make inferences about ρ_{XY}. Instead, inferences are based on Fisher's z-transform,

$$z = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{XY}}{1 - \hat{\rho}_{XY}} \right), \tag{14.4}$$

and inverse:

$$\widehat{\rho_{XY}} = \frac{e^{2z} - 1}{e^{2z} + 1} = tanh(z) \tag{14.5}$$

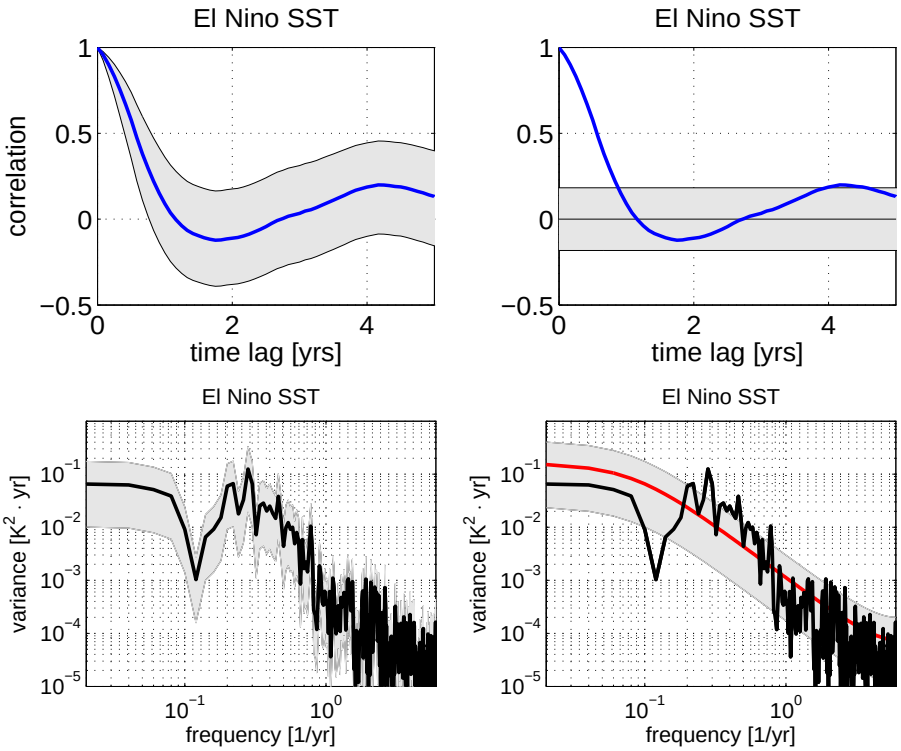


Figure 14.2: Auto-correlation (upper) and spectrum (lower) with the 90% confidence interval relative to the observed values (left) and relative to a null hypothesis (right) of the monthly mean El Niño SST time series.

which converges quickly to the normal distribution $\mathcal{N}\left(\frac{1}{2}\log\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right), \frac{1}{n-3}\right)$ when ρ_{XY} is nonzero. It is easily demonstrated that an approximate $\tilde{\rho} \times 100\%$ confidence interval for ρ_{XY} is given by

$$(\tanh(z_L),\ \tanh(z_U)), \tag{14.6}$$

where

$z_L = z - Z_{(1+p)/2}/\sqrt{n-3}$

$z_U = z + Z_{(1+p)/2}/\sqrt{n-3}$

and $Z_{(1+p)/2}$ is the $(1+p)/2$-quantitle of the standard normal distribution (see Appendix D). David [100] (see also Pearson and Hartley [308]) gives tables for exact confidence intervals for ρ_{XY}.

As an example we can again discuss the global mean surface temperature, T_{surf}, time series, see Fig. 15.5. It seems that T_{surf} has a positive trend. This can be quantified by the correlation between the time and T_{surf}, which is

$$\widehat{\rho_{tT}} = 0.67 \quad \Rightarrow z = 0.7928$$

, the time series has 681 month long, while we assume that very 24month is an independent sample,

$$n = 681/24 \approx 28$$

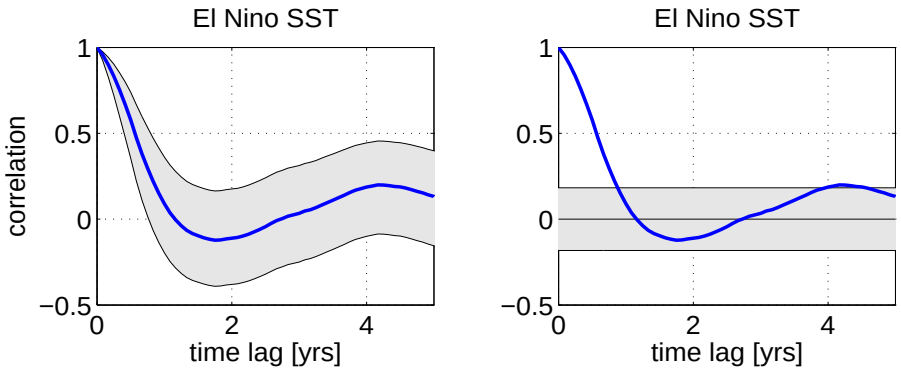


Figure 14.3: ...

for our confidence interval we choose $p = 90\%$, thus the interval boundaries are

$$Z_{(1+p)/2} = Z_{95\%} = 1.66$$

thus we get

$$\Rightarrow z_L = z - Z_{(1+p)/2}/\sqrt{n-3} = 0.7928 - 1.66/\sqrt{25} = 0.46$$

$$z_U = z + Z_{(1+p)/2}/\sqrt{n-3} = 0.7928 + 1.66/\sqrt{25} = 1.125$$

$$\Rightarrow \widehat{\rho_{tT}} = [0.43, 0.81]$$

So it seems that this trend is very likely to be positive (see also Fig.).

14.3 Uncertainties of the Spectrum

The spectra $\Gamma(\omega)$ are Variances as a function of frequencies. So each $\Gamma(\omega)$ is $\chi^2(k)$ distributed, while the number degree of freedom k depend on the estimating method.
The periodogram is $\chi^2(k = 2)$ distributed. Smoothed estimates like the *Chunk* or *Welch* estimate which split the time series into chunks are $\chi^2(k)$ distributed with

$$k \approx 2\frac{T}{M} \tag{14.7}$$

depending on the window function. T is the length of the time series and M the length of the window. The different estimating methods will in general give the values of k as a function of the estimating parameters.
We see that for spectral estimates we do not need to estimate any number of independent samples, the spread of the *pdf* only depends on the ratio $\frac{T}{M}$. Or in other words it only dependence on the number of cycles that the time series includes from the longest period estimated. e.g. A time series with $T = 600$ and $M = 120$ includes 5 cycles of the longest period, $f = 1/120$. So the number degree of freedom $k = 10$. The 95%-quantile of a $\chi^2(k = 10)$ is $x_p \approx 18$. Scaling it with the physical dimensions we have

$$x_{95\%} \approx 18\frac{\Gamma(\omega)}{k} = \frac{9}{5}\Gamma(\omega) \tag{14.8}$$

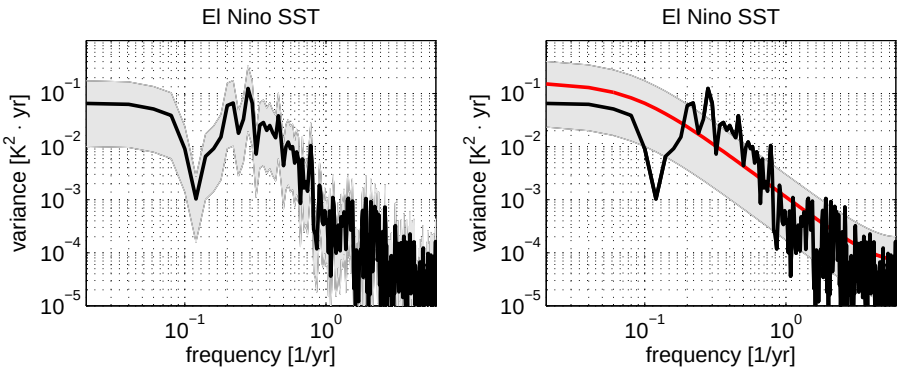


Figure 14.4: The Spectrum of monthly mean SST of the NINO3 region. the statistical uncertainties are given by the 90% confidence interval relative to the observed spectrum (left) and relative to the null hypothesis of a fitted AR(1)-process spectrum (right).

14.4 Uncertainties of the Cross Spectrum

14.5 Uncertainties for the Coherancy Spectrum.

The smoothed coherency spectrum can be thought of as a squared correlation coefficient that depends upon frequency.
This is most easily appreciated by considering the Daniell estimator, but the analogy applies equally to the other spectral estimators summarized in [12.3.19].
The Daniell cross-spectral estimator [12.3.11] is given by

$\hat{\Gamma}_{xy}(\omega_j) = \frac{1}{n}\sum_{k=j-(n-1)/2}^{j+(n-1)/2} I_{xy_{Tk}}$.
We can view $\hat{\Gamma}_{xy}(\omega_j)$ as an estimate of the (complex) covariance between processes $\mathbf{X}$ and $\mathbf{Y}$ at time scales between $\omega^{-1}_{j+(n-1)/2}$ and $\omega^{-1}_{j-(n-1)/2}$. To appreciate this, we substitute equation (12.56) for the cross-periodogram to obtain

$\hat{\Gamma}_{xy}(\omega_j) = \frac{T}{4n}\sum_{k=j-(n-1)/2}^{j-(n-1)/2} \mathbf{Z}_{xTk}\mathbf{Z}^*_{y_{Tk}}$.
Except for the factor T, this expression looks just like an estimate of the (complex) covariance between a pair of zero mean random variables $\mathbf{Z}_{xT}$ and $\mathbf{Z}_{y_T}$ that is computed from a sample $\{(\mathbf{Z}_{xTk}, \mathbf{Z}_{y_{Tk}} : k = j-(n-1)/2, \ldots, j+(n-1)/2\}$. This interpretation becomes even stronger when we assume that the cross-spectral density function is constant in the interval $(\omega_{j-(n-1)/2}, \omega_{j+(n-1)/2})$ because then the random pairs $(\mathbf{Z}_{xTk}, \mathbf{Z}_{y_{Tk}})$ are approximately independent and identically distributed.
We can estimate the correlation between the $\mathbf{X}$ and $\mathbf{Y}$ processes in the frequency range $(\omega_{j-(n-1)/2}, \omega_{j+(n-1)/2})$ by normalizing $\hat{\Gamma}_{xy}(\omega_j)$ with estimates of standard deviations of the $\mathbf{X}$ and $\mathbf{Y}$ in this frequency range. The latter are just the square roots of the estimated auto-spectra of $\mathbf{X}$ and $\mathbf{Y}$. Thus we have

$\hat{\rho}_{xy}(\omega_j) = \frac{\hat{\Gamma}_{xy}(\omega_j)}{\left(\hat{\Gamma}_{xx}(\omega_j)\hat{\Gamma}_{yy}(\omega_j)\right)^{1/2}}$.
Consequently the estimated coherency

$\hat{\kappa}_{xy}(\omega_j) = |\hat{\rho}_{xy}(\omega_j)|^2$

can be viewed as a measure of the squared correlation, or proportion of common variance that is

shared by $\mathbf{X}$ and $\mathbf{Y}$ in the $\omega^{-1}_{j-(n-1)/2}$ to $\omega^{-1}_{j+(n-1)/2}$ time scale range.
This interpretation of the coherency carries over to other periodogram-based spectral estimate sectionMonte Carlo Simulations Fisher's z-transform was used in [8.2.3] to construct confidence intervals for ordinary correlation coefficients. The same method can be used here for nonzero $\kappa_{xy}(\omega_j)$. Fisher's z-transform (8.5) of the square root of the coherency,

$\frac{1}{2}\ln\left(\frac{1+\hat{\kappa}_{xy}(\omega_j)^{1/2}}{1-\hat{\kappa}_{xy}(\omega_j)^{1/2}}\right) = \tanh^{-1}(\hat{\kappa}_{xy}(\omega_j)^{\frac{1}{2}})$,
is approximately normally distributed with mean $\tanh^{-1}(\kappa_{xy}(\omega_j)^{1/2})$ and variance $1/r$, where r is the equivalent degrees of freedom of the spectral estimator. Therefore approximate $\tilde{p}\times 100\%$ confidence limits for the squared coherency are

$$\left(\tanh\left(\tanh^{-1}\left(\hat{\kappa}_{xy}(\omega_j)^{1/2}\right)\pm\frac{\mathbf{Z}_{(1+\tilde{p})/2}}{\sqrt{r}}\right)\right)^2, \tag{14.9}$$

where $\mathbf{Z}_{(1+\tilde{p})/2}$ is the $(1+\tilde{p})/2$critical value of the standard normal distribution (Appendix D).[1]
The approximation that leads to interval (57) breaks down when $\kappa_{xy}(\omega_j)$ is zero. Then

$\frac{(r/2-1)\hat{\kappa}_{xy}(\omega_j)}{1-\hat{\kappa}_{xy}(\omega_j)}$
is approximately distributed as an $F(2, r-2)$ random variable. Thus

$\mathrm{H}_0 : \kappa_{xy}(\omega_j) = 0$ versus $\mathrm{H}_a : \kappa_{xy}(\omega_j) < 0$
can be tested at the $(1-\tilde{p})\times 100\%$ significance level by comparing $\hat{\kappa}_{xy}(\omega_j)$ with

$$\frac{2F_{\tilde{p}}}{r-2+2F_{\tilde{p}}} \tag{14.10}$$

where $F_{\tilde{p}}$ is the $\tilde{p}$ critical value of the $F(2, r-2)$ distribution.

14.6 Uncertainties for the Phase of the Cross Spectrum.

Hannan [157, p. 257] shows that approximate $\tilde{p}\times 100\%$ confidence limits for the phase spectrum Φ_{xy} are given by

$\hat{\Phi}_{xy}(\omega_j) \pm \sin^{-1}\left(\frac{t_{(1+\tilde{p})/2}}{r-2}\left((\hat{\kappa}_{xy}(\omega_j))^{-1}-1\right)\right)$
where $\hat{Phi}_{xy}(\omega_j)$ is the phase estimate obtained by substituting a periodogram-based estimator $\hat{\Gamma}_{xy}(\omega_j)$ of the cross-spectral density into equations (11.63)-(11.65), r is the equivalent degrees of freedom of the spectral estimator, and $t_{(1+\tilde{p})/2}$ is the $(1+\tilde{p})/2$ critical value of the $t(r-2)$ distribution (see Appendix F).

14.7 Uncertainties of EOF-Eigenvalues (Degenerated eigenvalues)

To be continued *ldots*

[1] Koopmans [229, p. 283] gives a slightly refined version of this interval. He also points out that the quality of the approximation depends upon the equivalent degrees of freedom r and $\kappa_{xy}(\omega_j)$, and that is best when $r > 40$ and $0.4 < \kappa_{xy}(\omega_j) < 0.95$. However, in our experience, interval (57) gives useful, although perhaps not precise, information when there are substantially fewer equivalent degrees of freedom.

Chapter 15

Test of a Hypothesis

15.0.1 The Logic of a Hypothesis test

In statistics you often like to test a hypothesis, hoping the observed data is verifying your hypothesis. To understand the logics in this we consider the probabilities of the following events:

H: Hypothesis is true
N: Null Hypothesis is true
D: Observed the data or characteristics (e.g. $\rho = 0.3$)

So what you like to know is: the probability of your hypothesis, H, under the condition of observing D:
Hypothesis Test: $P(H|D) =?$
In the following you will see that it is essentially impossible to estimate this probability

However, what we test is: $P(D|N) =?$
We assume that a null hypothesis is true and estimate what the likelihood of D is. Note: $P(D|N) \neq P(N|D)$. e.g. P(snow storm / winter) $\neq$ P(winter /snow storm)

So knowing how likely the null hypothesis is quite tricky.
$P(N|D) = P(N \wedge D)/P(D)$
$P(D|N) = P(N \wedge D)/P(N)$

$P(N) * P(D|N) = P(N \wedge D)$

$\Rightarrow P(N|D) = P(N) * P(D|N)/P(D)$
It is not possible to know P(N), similar for P(H). We would need to know all possible hypothesis.

Example:

hypothesis: El Nino influences Melbourne Rainfall
null hypo.: El Nino and Melbourne Rainfall are independent
obs.: correlation = 0.3

$P(D|N) = 0.001\% \rightarrow$ null hypo. rejected

El Nino influences Melbourne Rainfall? $\rightarrow$ You dont know this yet.

Many other hypo. are possible:

Example:

hypothesis: El Nino decreases Melbourne Rainfall
null hypo.: El Nino and Melbourne Rainfall are independent
obs.: Melbourne Rainfall for El Nino = -20%

$P(D|N) = 10\% \rightarrow$ null hypo. Not rejected?

El Nino does not influences Melbourne Rainfall? $\rightarrow$ You dont know this yet.

Signal is just not strong enough

Example:

hypothesis: EOF-2 is El Nino Modoki (physical mode)
null hypo.: EOF-2 is degenerated (North et al.1984 test) / Test of white noise for a pair of EV.
obs.: EV-1: 45% EV-2: 10% EV-3: 7%

$P(D|N) = 1\% \rightarrow$ null hypo. Rejected?

EOF-2 is El Nino Modoki? $\rightarrow$ NO! statistical significance is largely irrelevant for the interpretation.

EOF-2 can result from red noise.

15.1 The Null Hypothesis

The Null Hypothesis is a priori assumption about the expected value of the *pdf* of $\mathbf{X}$, independent of the samples of $\mathbf{X}$. We have a Null Hypothesis, H_0. We can than build a test to make inferences like 'H_0 is true' or 'H_0 is rejected' as discussed in the subsequent section or we can build a confidence interval for this null hypothesis. Here we build the confidence interval relative to the expected value in the null hypothesis, x_0, with the *pdf* of the null hypothesis:

$$\Omega = [-\sigma_{X_0} + x_0, \sigma_{X_0} + x_0] \tag{15.1}$$

if we assume near normal distribution for the null hypothesis or the quantiles:

$$\Omega = [F_{X_0}(x_{10}), F_{X_0}(x_{90})] \tag{15.2}$$

in all none-normal distributions. See Fig. 14.2 for example.

15.2 The structure of a Test

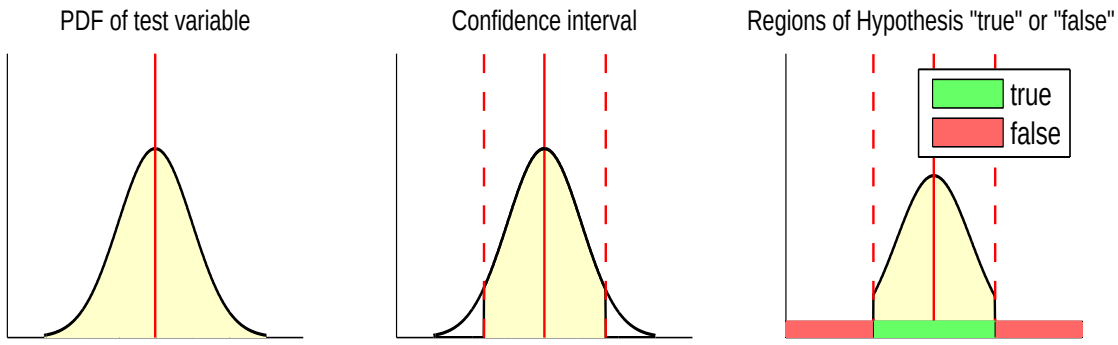
A test of a null hypothesis, H_0 has several elements that need to be specified:

- **The Test Variable:** We need to formulate a test variable, $\mathbf{T}$ that quantifies deviations from the null hypothesis in the most efficient way. That means for a given sample size, $\mathbf{T}$ is the variable that will have low likelihoods if H_0 is false and it will have large likelihoods if H_0 is true. No other variable will separate truth and falsehood of H_0 better. Finding the right

test variable is often not easy. Some examples for 'standard tests' are given in the following subsections.

- The *pdf* of the test variable: Obviously we need to know the *pdf* of **T** in order to quantify likelihoods of truth and falsehood of H_0. The theoretical *pdf* of **T** will in general make some assumption about the *pdf* of the sampled variables used for **T**. If these assumption are not valid than the test will not be valid, meaning the likelihood of rejecting H_0 when H_0 is right will increase, while at the time the likelihood of non-rejecting H_0 when H_0 is wrong will increase too. Which is not good! So it is important to know the assumptions made for a specific test situation. See also chapter 17.
- Confidence level: We need to decide about the probability limit of our Confidence level, which is a quantile of the cumulative distribution function of **T**. This level sets the strength or risk of the test. The Confidence level should be large, because small values (e.g. $< 80\%$) makes the sampling of observation point less, since the test will likely fail no matter what the observations are. if

So a hypothesis is tested as follows: We formulate the test variable, T and find its *pdf*. We verify that the assumption made for the *pdf* of T under the null hypothesis are indeed given. We than estimate the value of T from the data and check if the T value passes the Confidence level in the cumulative distribution of T under the null hypothesis. Some examples will be discussed in the subsequent sections.



caption Sketch illustrating the three elements of a test.

15.3 The strength and risk of a Test

A test is *strong* when it can reject H_0 whenever H_0 is wrong and it does not reject H_0 whenever H_0 is right. Although, this seems obvious, it is important to know that the some tests are not as strong as others.
If, for instance, **T**, the data under investigation, is not consistent with H_0, but it follows a *pdf* similar to that of H_0. Than the test will have problems to reject H_0. So we need to optimize the test variable.
The *risk* of a test is to reject H_0 when H_0 is true or to find H_0 true when H_0 is false. The risk of a test is usually defined by the Confidence level. The risk can never by zero due to the stochastic nature of the problem. In the test we assume, for instance, that if **T** is far away from the expected value (far in terms of passing a confidence level), than H_0 is false. But obviously there is a non-zero likelihood that **T** was just an unusual event of H_0. If we choose our Confidence level to be 95%, than we will reject H_0 when H_0 is true in 5% of all cases.

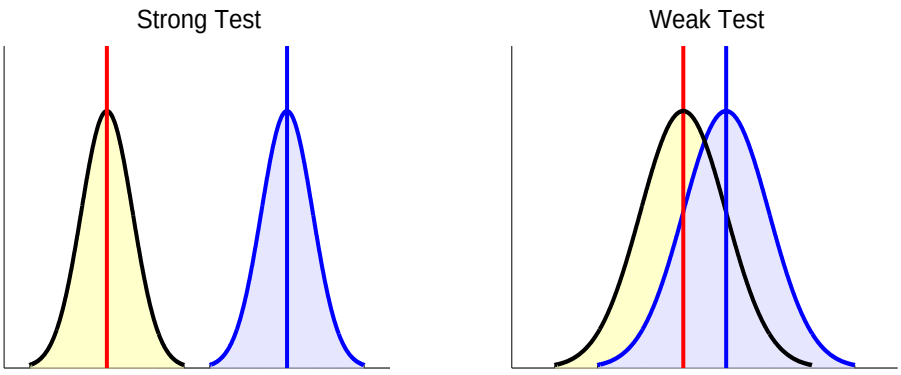


Figure 15.1: . Sketch illustrating strong and weak tests.

Minimizing the risk is a trade of between false rejection of H_0 when H_0 was true and false non-rejection when H_0 was wrong. Some guidance may be given by considerations of what we consider as 'physically' significant.

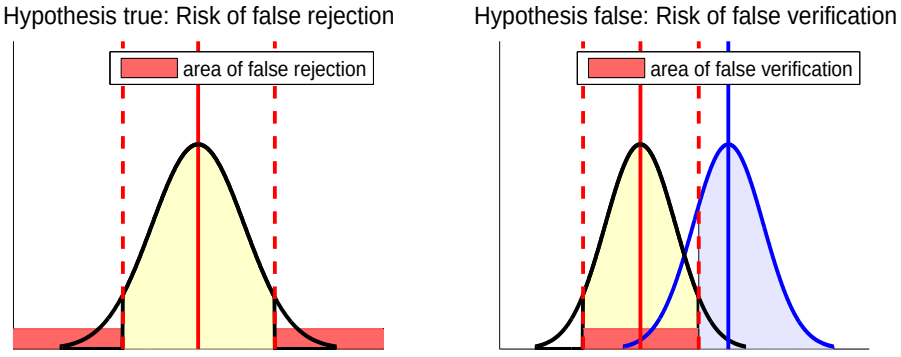


Figure 15.2:

15.3.1 Multiple use of Tests: Global tests

Note: The probability value of a test is only valid if it is used once:

$P(t - value >= 95\%) = 5\%$

but obviously:

P(1 out of 10 t-values $>= 95\%) = 1 - 0.95^{10} \approx 40\%$

Thus Multiple use of test changes the probabilities: the significance is lower. All failed tests need to be part of the result. To consider situations in which we use a single (local) test many times we need to have a global test that evaluates the probabilities of all test results together:

Local test: single-use test

Global test: evaluates many local tests → significance of multiple local test results.

Examples:

(1) Australia climate change:

- Null Hypothesis: Australia climate is stationary
- t-tests of 10 different climate indicies (e.g. Melbourne temp, rain, winds, heat waves, etc.)
- you find that 1 out of 10 passes the 95% level → its the rainfall in Vic.
- Conclusion: Australia climate in rainfall is changing? → Wrong!
- You need to do global test.

(2) Test global field of Tsurf:

- Null Hypothesis: Tsurf is stationary at all locations
- t-tests of all locations
- About 10% of all points will pass the 95% level
- Need Field significance test (global test)
- Need to consider spatial correlation

15.4 The Estimation of the Effective Sample Size

As a repetition of section 8.5.
In many statistical analysis we need to know the number degree of freedom, n_X, of the time series. e.g. the χ^2-pdf or the tests of mean, variance, correlation. Initially we had the definition for n_X, that $\mathbf{X_i}$ and $\mathbf{X_j}$ have to be independent, meaning uncorrelated. This definition is for stochastic processes not helpful, because the autocorrelation of an AR(1) process, for instance, does not go to zero at all.
However, n_X of $\mathbf{X}$ can be estimated by using the statistical relation between the statistical parameter of interest and n_X. For the mean we know from the central limit theorem that:

$$Var(\bar{\mathbf{X}}) = \frac{\sigma_X^2}{n_X} \tag{15.3}$$

If we know σ_X and $Var(\bar{\mathbf{X}})$ we can get n_X. The relation between the true number of time steps used and n_X is the decorrelation time :

$$\tau_D = \frac{n}{n_X} \tag{15.4}$$

For the mean we find:

$$\tau_D = 1 + 2\sum_{k=1}^{\infty} \rho(k) \tag{15.5}$$

For the variance we find:

$$\tau_D = 1 + 2\sum_{k=1}^{\infty} \rho(k)^2 \tag{15.6}$$

So why do we have different characteristic time scales for the mean and variance? see section 8.5. However, in general we would approximate τ_D simply by evaluating the auto correlation function, γ_{xx}.

$$\gamma_{xx}(\tau_D) \approx 0.2 \tag{15.7}$$

15.5 Test of the Mean

We want to test if the mean of a sample of $\mathbf{X}$ and the mean of a sample of $\mathbf{Y}$ are the same. We assume the following:

- We have a number of independent samples of $\mathbf{X}$, n_x and a number of independent samples of $\mathbf{Y}$, n_y. As mentioned earlier the number of samples is usually much larger than n_x, n_y, because not all samples are independent. We need to estimate the numbers n_x, n_y, see section 8.5.
- $\mathbf{X}$ and $\mathbf{Y}$ are realizations of the same normal distribution

An obvious test variable for testing the differences between two values is something like:

$$t = \frac{\widehat{x} - \widehat{y}}{\sigma} \tag{15.8}$$

Defining the test variable for the mean we start with the simpler case of testing a null hypothesis in which the mean is a priori known. So we start with just one set of samples of $\mathbf{X}$ and compare it to a known mean, $\mu_0 = constant$, the mean for the null hypothesis. The central limit theorem tells us that the standard deviation of the mean of $\mathbf{X}$ is:

$$\sigma_\mu^2 = \sigma_X^2 / n_x \tag{15.9}$$

The optimal test variable for testing if $\widehat{\mu_x} = \mu_0$, with $\widehat{\mu_x}$ as the sample mean of $\mathbf{X}$, is:

$$t = \frac{\mu_0 - \widehat{\mu_x}}{\widehat{\sigma_x}/\sqrt{n_x}} = \sqrt{n_x}\frac{\mu_0 - \widehat{\mu_x}}{\widehat{\sigma_x}} \tag{15.10}$$

Lets examine the characteristics of this equation under the assumption that $\widehat{\mu_x} = \mu_0$:

- The expected value of t is zero. Any sample of $\widehat{\mu_x}$ can deviate into both directions. The larger the deviation of $\widehat{\mu_x}$ from μ_0 the larger t. Thus the *pdf* of t must be symmetric with the expected value zero and decreasing likelihoods for large values of $|t|$. Therefore the larger t the more significant the differences in means are.
- t is proportional to $1/\sigma_x$, with σ being the natural variability of the process (time series). The larger σ_x the smaller is t for a given difference in the means.
- t is proportional to the square root of the number of independent observations, n_x. So with increasing numbers of observation t is increasing for a given difference in the means and the difference becomes more significant. Or otherwise, with increasing numbers of observations, a significant difference in the means becomes smaller; increase in number of observations makes smaller differences detectable.

The theoretical distribution of t is, for the given assumptions, the student's t-distribution, see section 3.12. Note that the t-*pdf* converges towards the standard normal *pdf*. This in turn means that the t-values are similar to the values of a standard normal *pdf*. Thus, absolutes values large than 1/2/3 have a probability of about 35%/5%/0.01%. So with t-values larger than 2 or 3 we usually have a significant difference in the means.
An Example illustrates the test of means: We test a change in the global mean temperature. Lets assume that the global mean temperature has been estimated over a very long time in the past:

$$\mu_0 = 13.8$$

Now we measure the global mean temperature over the last ten years and find:

$$\widehat{\mu_x} = 14.1$$

$$\widehat{\sigma_x} = 0.2$$

We assume that every 2 years is an independent sample, thus:

$$n_x = 5$$

The significance of the result is most strongly depending on our assumptions about n_x. Knowing about the independence of samples does make some assumptions about the auto-correlation function (section 8.5), which in turn makes sumption about the spectrum of the global mean temperature. Unfortunately we know little about the low-frequency (periods longer 100yrs) variance of global mean temperature.
So using eq.[15.10] we find:

$$|t| = |\sqrt{n_x}\frac{\mu_0 - \widehat{\mu_x}}{\widehat{\sigma_x}}| = |\sqrt{5}\frac{13.8 - 14.1}{0.2}| = 3.4$$

Looking at the cumulative distribution function of t in Fig. 3.14 we find that this change in mean is highly significant.
Now we want to test the mean of a sample of $\mathbf{X}$ not against an a priori known μ_0, but against the mean of an other sample, $\mathbf{Y}$. the test variable is:

$$t = \frac{\widehat{\mu_x} - \widehat{\mu_y}}{S\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{15.11}$$

$$S^2 = \frac{(n_x - 1)\widehat{\sigma_x^2} + (n_y - 1)\widehat{\sigma_y^2}}{n_x + n_y - 2} \tag{15.12}$$

with $n_x + n_y - 2$ as the number of independent observations for the parameter of the t-*pdf*. It may not be strait forward to see that this is related to eq.[eq.15.10], but note that $S \approx \sigma_x$ and for $n_x << n_y$ we can set $\mathbf{Y}$ as the null hypothesis and will end up with eq.[eq.15.10].
As an example we apply this test to the global surface temperature field, see Fig. 15.6. As we can see most of the earth has been warming significantly over in the last 10yrs.

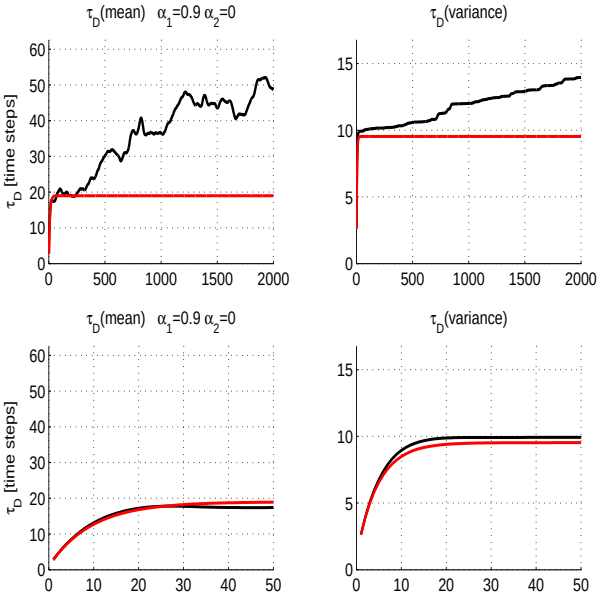


Figure 15.3: Estimated decorrelation times, τ_D for the mean (left) and variance (right) of an AR(1)-process as function of time lag used for the estimation. Lower panels are a blow up of the upper panels.

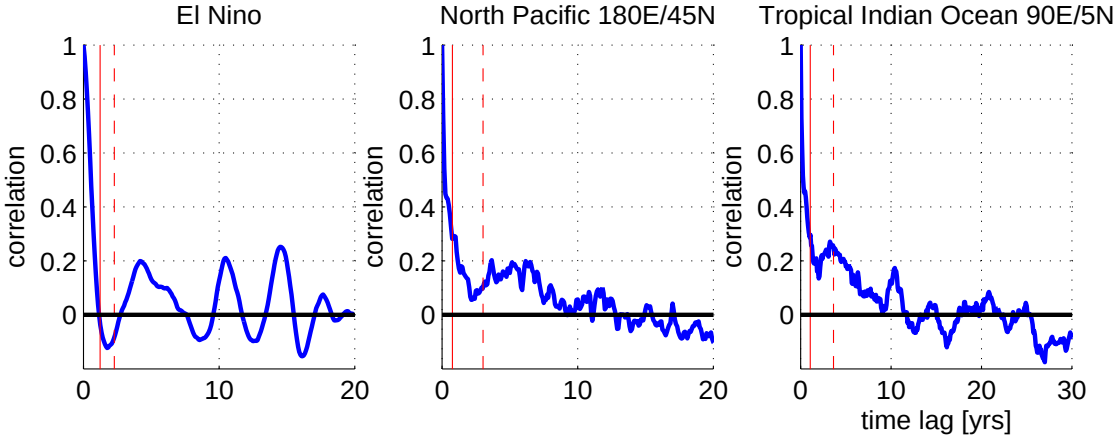


Figure 15.4: Observed auto-correlation functions (solid blue lines) of different SST time series and the estimated decorrelation times, τ_D for the mean (dashed red) and variance (solid red).

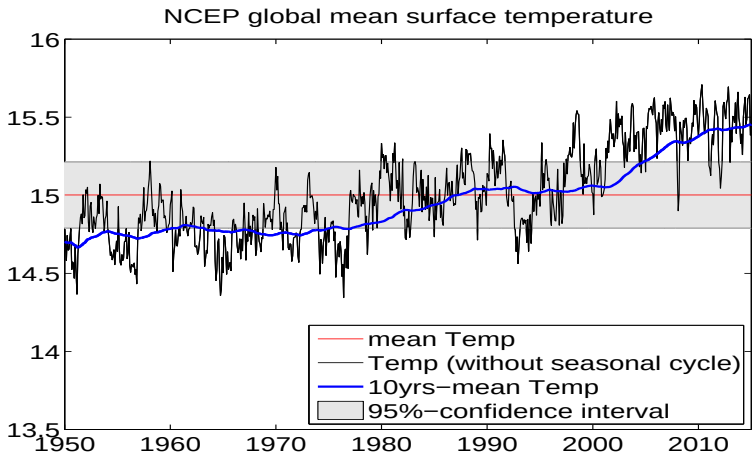


Figure 15.5: .

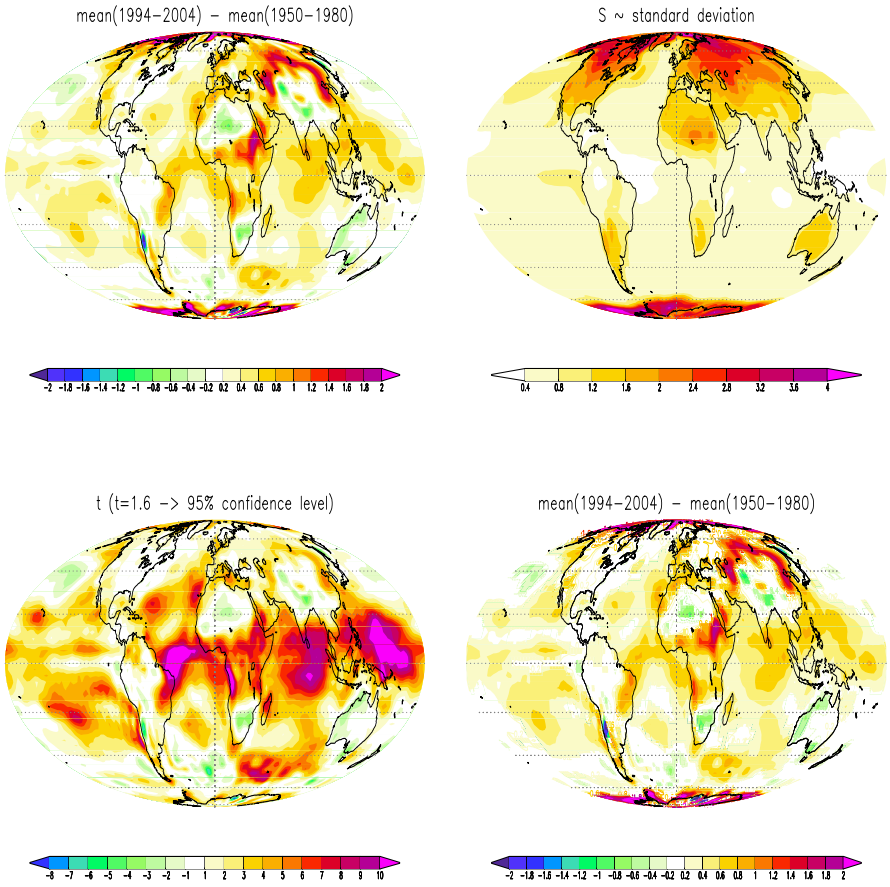


Figure 15.6: Illustration of the t-test, with the global 2meter temperature field from NCEP reanalysis data. Upper left: differences in the mean of two periods. Upper right: the S value of eq.[15.12]. Lower left: t-values and lower right is the same as upper left, but only points with t ¿ 1.6 are colored.

15.6 Test of Variances (Fisher F-test)

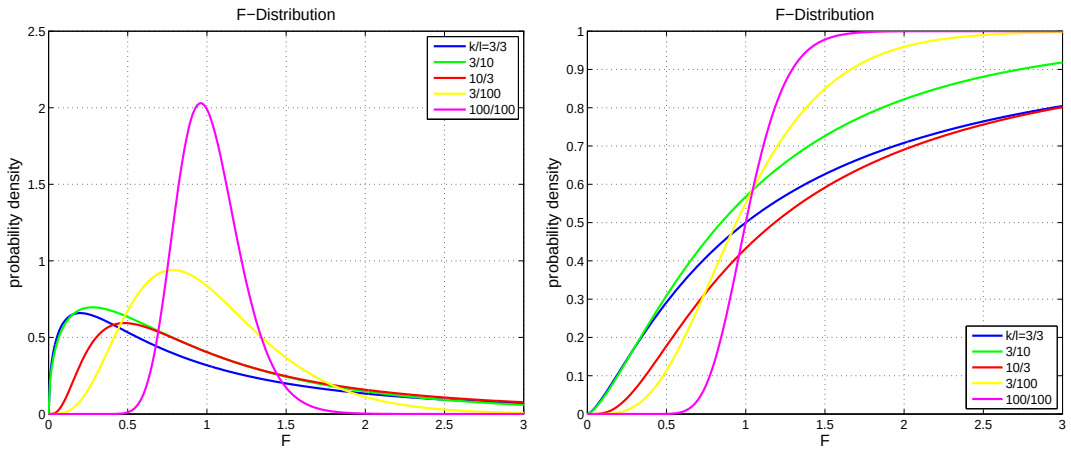


Figure 15.7: The χ^2 distribution for different degrees of freedom $k = 1, 2, 10$. Left pdf and right the cumulative distribution.

Variances are χ^2 distributed, see section 3.11. The spread of the χ^2-distribution depends strongly on k, the number degrees of freedom, which is the number of independent samples. The moments are:

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &= k \\ Var(\mathbf{X}) &= 2k \end{aligned} \tag{15.13}$$

We see that both $\mathcal{E}(\mathbf{X})$ and $Var(\mathbf{X})$ are dimensionless. If we include the dimensions and express $Var(\mathbf{X})$ as a function of $\mathcal{E}(\mathbf{X})$ we find:

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &= k \cdot c \\ Var(\mathbf{X}) &= 2k \cdot c^2 = 2\frac{\mathcal{E}(\mathbf{X})^2}{k} \end{aligned} \tag{15.14}$$

So we see that σ, the spread of the *pdf*, decreases when k increases. Further we find that $\sigma \sim \mathcal{E}(\mathbf{X})$. In statistical tests we often need the p-quantiles, x_p, for defining confidence intervals. The dimensionless x_p can be read from the cumulative χ^2-distribution, by scaling it with $\frac{\mathcal{E}(\mathbf{X})}{k}$ we can include the physical dimensions. If we want to compare the variance of a sample of $\mathbf{X}$ with the variance of a sample of $\mathbf{Y}$, the best test variable is

$$F = Var(\mathbf{X})/Var(\mathbf{Y}) \tag{15.15}$$

Here the test is based on the ratio, not the difference as for the test of means. This is because the spread (σ) of the variances is proportional to Variances itself. The test variable F follows the Fisher F-distribution. The probability density function is given by

$$f_F(f) = \frac{(k/l)^{k/2}\Gamma((k+l)/2)}{\Gamma(k/2)\Gamma(l/2)} f^{(k-2)/2} \left(1 + \frac{k}{l} f\right)^{-(k+l)/2} \tag{15.16}$$

The distribution and espacially the cumulative F-distribution is in MATLAB / or tabulated in statistical text books (see Storch and Zwiers). Some important characteristics of the F-distribution:

- It is positive definit.
- It is positively skewed.
- $Var(F)$ is much more sensitive to l than is is to k.
- It convergences to the χ^2 distribution for $l \to \infty$.

Examples:

(1) El Nino Variability:

$\sigma(1950-2000) = 1.0^oC \Rightarrow Var = 1[^oC]^2$

$\sigma(2000-2013) = 0.7^oC \Rightarrow Var \approx 0.5[^oC]^2$

$\chi^2(1950-2000) : k \approx 15$
$\chi^2(2000-2013) : k \approx 4$

$\to$ Variance fluctuations are likely if stationary

(2) Internal vs. External variability:

Model experiments: AGCM only vs. AGCM-coupled to Ocean

$\sigma(AGCM, internal) = 1.0^oC \Rightarrow Var = 1[^oC]^2$
$\sigma(extenral) = 0.5^oC \Rightarrow Var = 0.25[^oC]^2$

$Var(CGCM, total) = Var(internal) + Var(extenral) = 1.25[^oC]^2$

$\to$ you would need a long (about 50-100yrs) simulation to get the ocean signal to be 'significant'.

(3) Non-stationary Variance (Running means):

15.7 Test for Zero Correlation

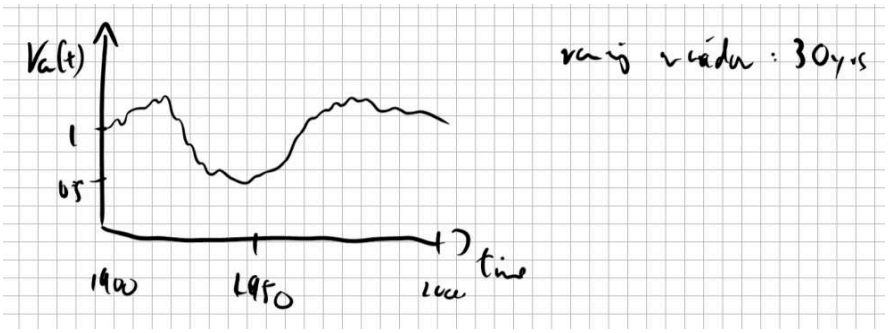
An approximate test of H_0: $\rho_{XY} = 0$ can be performed by computing

$$\mathbf{T} = |\hat{\rho}_{\mathbf{XY}}| \sqrt{\frac{\mathbf{n-2}}{\mathbf{1-\hat{\rho}^2_{XY}}}} \tag{15.17}$$

thus significant non-zero correlations are

$$\widehat{\rho_{XY}} = \pm \frac{T}{\sqrt{T^2+N}} \tag{15.18}$$

and comparing $\mathbf{T}$ with critical values from the t distribution with $N = n-2$ degrees of freedom (see Appendix F). The type of test, one sided or two sided, is determined by the form of the alternative



hypothesis.

As an example we can again discuss the correlation between the time and T_{surf}. The null hypothesis of no trend means H_0: $\rho_{XY} = 0$. The t value of t distribution with $N = 28 - 2$ and $p = 95\%$ is () Confidence interval (3) and test (4) both require the normal assumption.

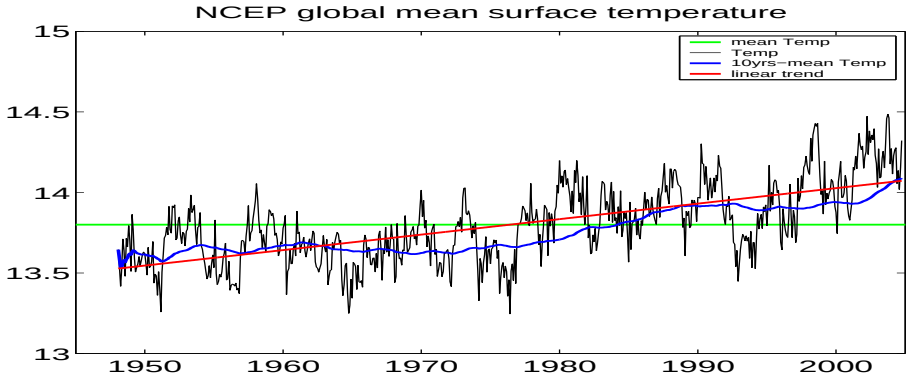


Figure 15.8: ...

15.8 Test for Distribution (Komolgorov Smirnov Test)

Often we need to know if a variable X is folowing a theortical *pdf* or if two variables X_1, X_2 are following the same *pdf*. So this is a more general question than to just test the mean or variance. This can be done with only a few samples, but it is of cause much more uncertain, than the simple tests of mean or variance, since there are now more degrees of freedom. The test variable is:

$$d = max(|F_{n1}(X) - F_{n2}(X)|) \tag{15.19}$$

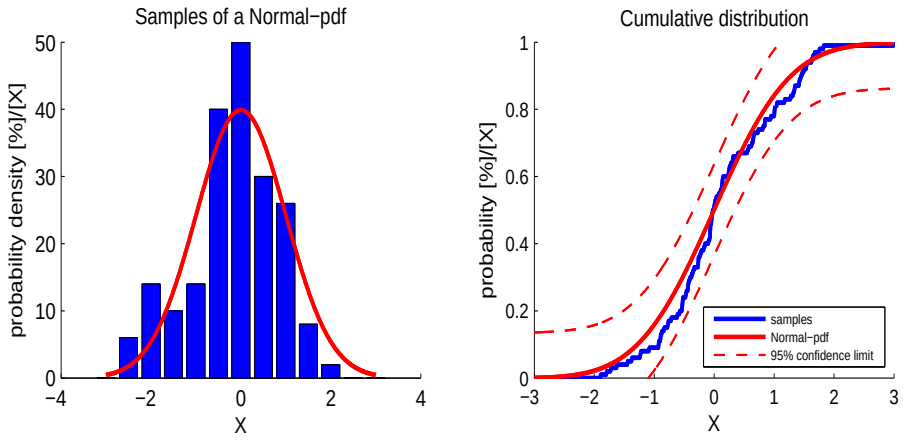


Figure 15.9: Illustration of the Komolgorov Smirnov Test for comparing two pdfs.

$F_{ni}(X)$ =observed cumulative distribution

d finds the maximum difference between two cumulative distributions$F_{n1}(X), F_{n2}(X)$. The inverse cumulative distribution of d,

$$D^{-1}(p) = (-\frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2})ln(\frac{p}{2}))^{1/2} \tag{15.20}$$

$D^{-1}(p)$ gives the confidence level for the probability level p.

Chapter 16

Monte Carlo Simulations


Monte Carlo Simulations are a general expression for simulation in which some variables are given as random numbers of a predefined *pdf* (just as the casino in Monte Carlo involves random numbers). Note that most climate models (AGCMs or OGCMs) are deterministic models where all variables have exact values.

The Monte Carlo is often used in stochastic problems, where deterministic models would be too complex or do not exist (e.g. in quantum mechanics). We assume that a random variable $\mathbf{Y}$ is a result of an operation A, such as

$$Y = A(X_1, X_2, ..., X_i) \tag{16.1}$$

where the X_i are random variables with known distribution. The operator can be of any complexity, a GCM-model for instance. For some simple operators, $A(\cdot)$, we can give the *pdf* of $\mathbf{Y}$, as done for the normal, χ^2, $t-$ or $F-$distribution, for those we do not need a Monte Carlo Simulation. But in most stochastic problems it is impossible to analytically find the *pdf* of $\mathbf{Y}$. It can only be estimated by the Monte Carlo approach.

Ensemble Forecast, or Models of the radiation transfer in the atmosphere are examples of Monte Carlo simulations. In Ensemble Forecast we are interested in the uncertainties, the *pdf*. In Models of the radiation transfer we are usually just interested in the expected value which can only be estimated with the Monte Carlo approach, the resulting *pdf* is usually normal.

16.1 Bootstrapping the Probability Density Functions

Bootstrapping (*'to pull oneself up by ones own bootstraps'*) is a Monte Carlo approach to estimate the *pdf* of any random variable. In many cases we do not know the *pdf* of a variable and analytical solutions are either inscrutable or frankly we do not really care about it. However, for statistical inferences we need to know the *pdf*. For any random variable we can estimate the *pdf*, if we can formulate $\mathbf{Y}$ is a result of an operation A, such as

$$Y = A(X_1, X_2, ..., X_i) \tag{16.2}$$

where the X_i are random variables with known distribution. The operator can be of any complexity, a GCM-model for instance. The problem usually is to formulate the operator $A(\cdot)$ for the given situation (null hypothesis). Once $A(\cdot)$ is formulated we can generate a distribution of $\mathbf{Y}$ with the computer, which is our best estimate of the *pdf*. The quality of the estimate depends only on the number of realizations used to define the *pdf*.
The elements of the Bootstraping approach are the following:

1. Hypothesis: (e.g. $H_0 : \widehat{\mu} = \mu_0$)
2. Stochastic/Test variable: (e.g. $t = \sqrt{n_x}\frac{\mu_0 - \hat{\mu}}{\widehat{\sigma_x}}$)
3. Stochastic Model/Test Situation: The test situation for the null hypothesis need to be specified: (e.g. for the test of means: $X \sim \mathcal{N}(\mu_0, \widehat{\sigma_x})$ and a time series of X_t with n_x independent samples.)
4. Numerical generation of many test situations: $\Rightarrow$ *pdf* of the test variable.

A few examples shall illustrate how a Bootstraping approach works:

Example: *pdf* for zero correlation

Lets assume you like to know the uncertainty in a correlation estimate. You are unshure about the pdf of the variables or the effective sampling size:

Hypo: $\rho(X1 \text{ vs. } X2) \neq 0$

Null Hypo.: zero correlation

Stochastic Model of reshuffling the order to test zero correlation (loop many times):

1) shuffle order in both time series

2) ρ(shuffled X1 vs. shuffled X2)

$\rightarrow$ result PDF for zero correl with same sample numbers and same PDF.

Note: MATLAB function bootstrp does something else: It resamples (??).

Example: *pdf* of spectral estimates

In MATLAB the uncertainy of the spectral estimates of the *psd*-function is not given. However, for statistical inferences we need to know the confidence intervals of the spectral estimate. In order

to estimate the *pdf* of the spectral estimates we generate a time series of a process from which we now the theoretical spectral distribution: An AR(1)-process for instance. Fig. 16.1 shows a spectral estimate of a time series, which is a realization of an AR(1)-process. For comparison the theoretical spectrum of the AR(1)-process is also shown. The non-dimensional variable for which we like to know the *psd* is:

$$\chi = \widehat{\Gamma(\omega)}/\Gamma_{AR(1)}(\omega)$$

Where $\widehat{\Gamma(\omega)}$ is the spectral estimate of a time series and $\Gamma_{AR(1)}(\omega)$ the theoretical spectrum of the AR(1)-process.
We can generate a large number of χ by the following bootstraping approach:

1. Generate a time series of the AR(1)-process of length T.
2. Compute the spectrum of the time series with the MATLAB-funcion *psd* using the window length M.
3. Compute χ for all ω (Spectral estimate $\widehat{\Gamma(\omega)}$ gives $M/2+1$ independent estimates of the χ)
4. Repeat step 1-3 many times to create a large number of realizations of χ.

The result for a large number of χ is shown in Fig. 16.1

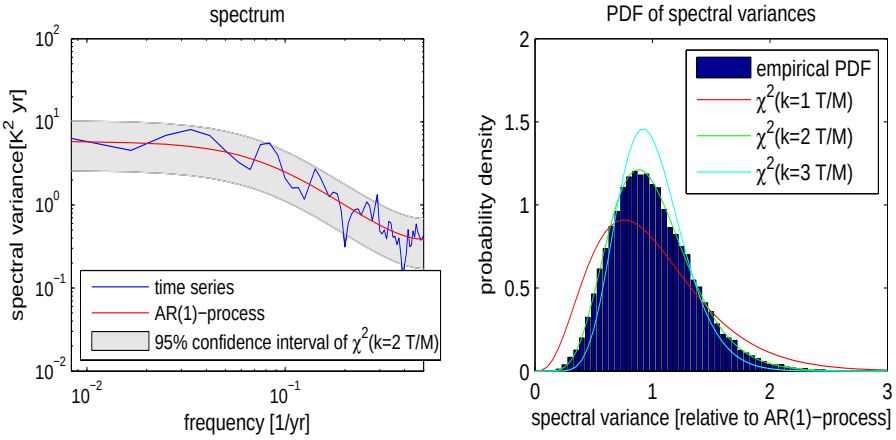


Figure 16.1:

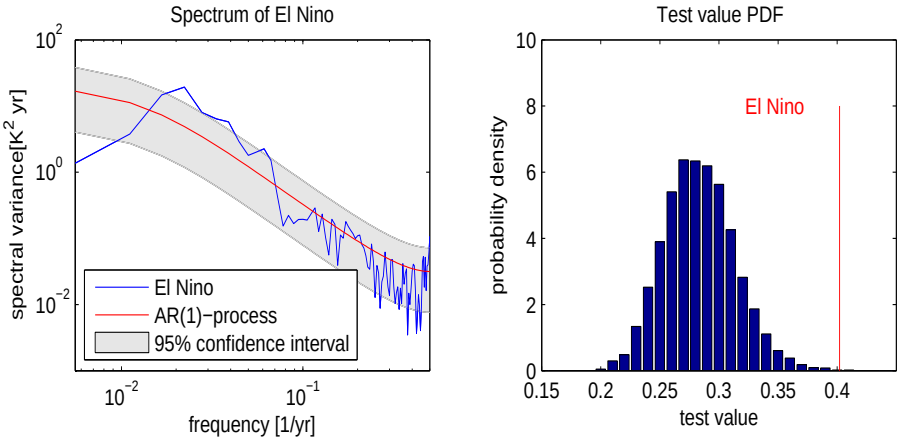
Example: Test if the El Niño time series is an AR(1)-process

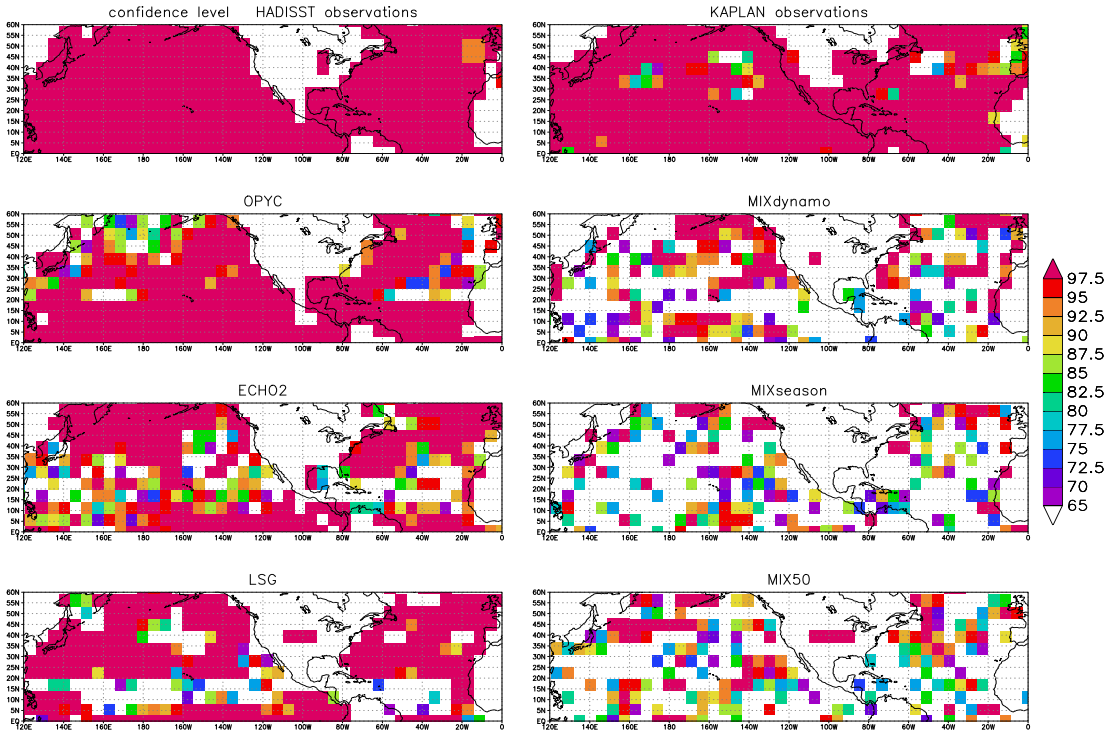
We may use the spectrum of the El Niño time series to test the hypothesis: The El Niño time series is an AR(1)-process. A good test Variable would be:
$t = \sum_{i=1}^{N_{freq.}} log(\widehat{\Gamma}(\omega_i)) - (\Gamma_{AR1}(\omega_i))$

Generate a large number of the test variable t for the hypothesis:

1. Time series of length T of the AR(1)-process with $\alpha_1 = \rho(1)$.
2. Compute the spectrum of the time series.
3. Compute t.

Repeat steps 1-3 many times to generate a large number of t, which builds the basis for the emperical estimate of the *pdf* of t. Fig. XXX shows the *pdf* of t and the t-values that the El Niño time series has, which is clearly outside the bulk of the distribution. Thus we may conclude that the El Niño time series is very likely different from an AR(1)-process.





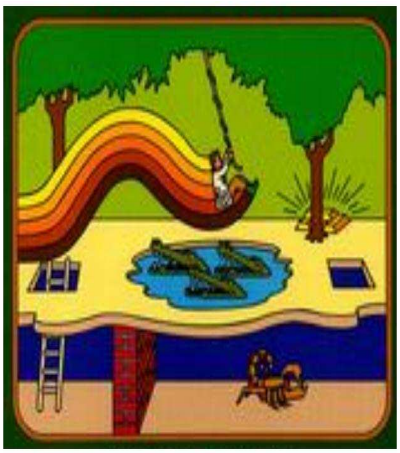
Part V

Strategies, Tactics and Pitfalls

In the previous section we learned the basics of statistical analysis. We know some basic parameters, fundamental statistical analysis methods and how to do statistical hypothesis testing. Although we now have the technical fundament for statistical analysis, we are still missing some guidance in terms of overall strategies, more specific tactis and tippical pitfalls. A typical problem in statistical analysis is to draw the right conclusions from the statistical analysis done. Most reserachers are very much capable in applying complex statistical analysis methods to data, but the conclusions they draw from these statistical analysis are often not supported by their statistical results. It is therefore helpful to get some more detailed advices beyond the definitions of methods or parameter. In analogy to a football player: we now have the technical skills to shoot and control the ball, we know how to head the ball and to run with it, but we miss all tactical skills to win a game. Probably you may have meet some every skillful football player that can do all tricks with the ball, but you still do not want him in your team, because he always looses the ball after the third super dooper trick with the ball. This guy is lagging some advices in team play and tactics. Which you will now get for statistical analysis.
In the first chapter ´Pitfalls´ we will look at some examples of statistical analysis or problems to explore how difficult it some times is to draw the right conclusions from statistical analysis or how difficult it is to do a meaningful statstical analysis. In the Second chapter ´Strategy´ we will discuss some main strategies in statistical ananlysis that helps in the development of meaninful statistical studies. In the final chapter ´Tactics´ we will discuss a number of helpful tactics to do successful statistical analysis.

Chapter 17

Pitfalls



Often the main problem in statistical analysis is to draw the right conclusions/inferences from the results. It is in most cases trivial to compute certain statistical parameters from the data, but interpreting the results is far from trivial. Most of the problems arise in understanding the probabilities and dealing with uncertainties. We have to acknowledge in many cases that our ability to think logically and to be objective in our analysis are limited.
We have seen in the very beginning of this course, that the probability of an event or the *pdf* of a random variable is changing if additional conditions are assumed. This is of cause also true for the *pdf* of test variables.
A short and simple example shall illustrate the problem: In the roulette game we have the option to set money on red or black numbers, both have a likelihood of about 50%. It is, however, unlikely to have five red numbers successively,

$$p(5reds) = (\frac{1}{2})^5 = 3\%$$

So some people think it is smart to go for black numbers if we already had 4 red numbers successively, since five red numbers successively are very unlikely. But of cause we know the likelihood of the next number is independent of the previous ones,

$$p(5reds|4\ reds\ already) = \frac{1}{2} = 50\%$$

So we see that under some conditions a very unlikely event becomes an ordinary event. So the $P(A)$, with A being "five time red in a row", is very different then the conditional probability

$P(A|B)$, with B being "four time red in a row". This is easy to see for us in the above example, but in many real world problem will will encounter such situations with conditional probabilities $P(A|B)$, but we may not even recognise that there is a conditional assumption and even if we know about it, we would find it difficult to deal with it.
In the following we will discuss several examples that shall illustrate some of the pitfalls in statistical inferences. The examples shall illustrate that additional conditions, constraints or knowledge can change the likelihoods of events drastically and thus change the statistical significance of the events. These example illustrate some of the most common pitfalls, which are important to known, because one is most likely doing the same mistakes if one is not aware of them. These pitfalls will guide us towards some general strategies and tactics in statistical analysis.
Many of these examples are taken from a number of textbooks that are not classical university textbooks. Several examples are from the books "Thinking Fast and Slow" from Daniel Kahneman (Nobel price in economics) and the "The Black Swan" from Nassim Nicholas Taleb (a non-academic stock trader). Both are books about financial markets, but still somewhat interesting for us. Daniel Kahneman made it quite clear that we tend to have problems in evaluating probabilities, dealing with randomness and in being objective. All every important aspects for climate researchers. The climate dynamics examples are from my own experience.

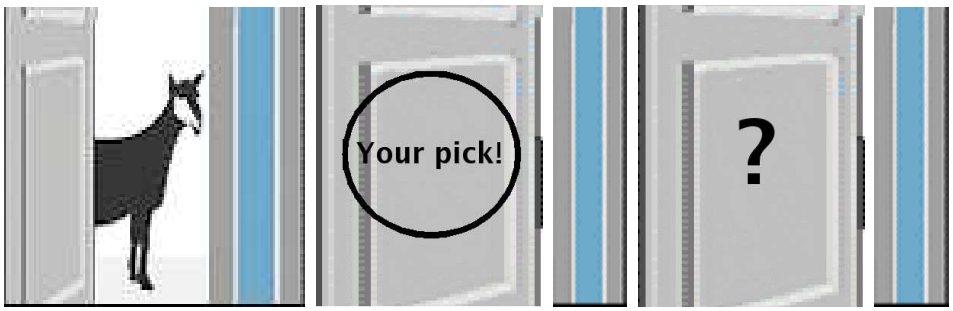
17.1 Additional or Hidden Assumptions

Lets have a look at a few examples in which additional assumption mess up the problem. We start with some examples that have nothing to do with climate dynamics, but are quite entertaining and helpful for our logical skills. We then discuss an example from climate dynamics research.

17.1.1 Example: The Goat Problem

Lets assume we are in a Quiz show. We shall select one out of three doors, to win a fancy car. So one of the three doors has a car behind it, which you would win, the other two have goats, which you will not win.

So you pick one door, but it will not jet be opened, instead the quiz master gives you a hint and opens one of the remaining doors, which uncovers one of the goats (this will always work). He asks you if you now want to change your mind and rather pick the remaining door? so what is the best strategy: go with your first pick, or change your mind and go with the advise of the quiz master?



Your intuition may tell you that it does not matter: all doors have the same likelihoods of showing the nice car. But this is wrong! your first pick has a 33% chance to find the car. The door suggested by the quiz master has a 66% chance to show the car, because it will only show a goat, if your first pick was the car. Think about it some time if you like to.

So again the additional informations have changed the likelihoods of an unusual event.

Usually the additional conditions of a specific problem will end up in a test variable those *pdf* is not listed in any textbook and therefore none of the standard test can be applied. So you are screwed? No, of cause not, you can find the *pdf* of any test variable with numerical approaches, see section Monte Carlo, or find the theoretical *pdf* by yourself. Well,... thinking about it ... you are screwed! Meaning you really need to put some work into finding the right *pdf*.

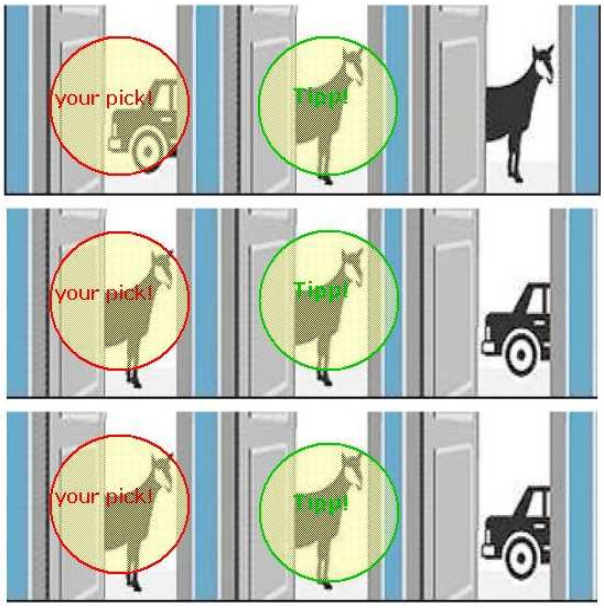


Figure 17.1:

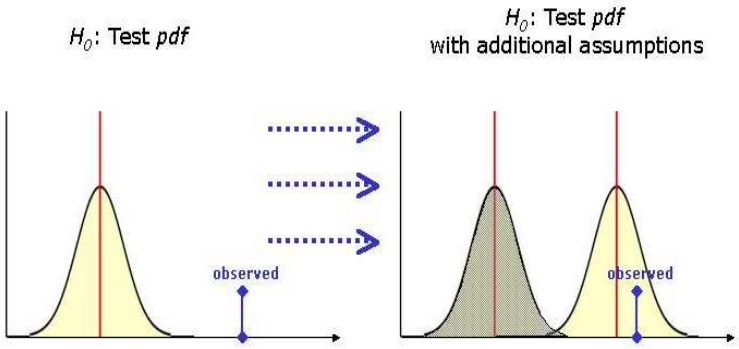


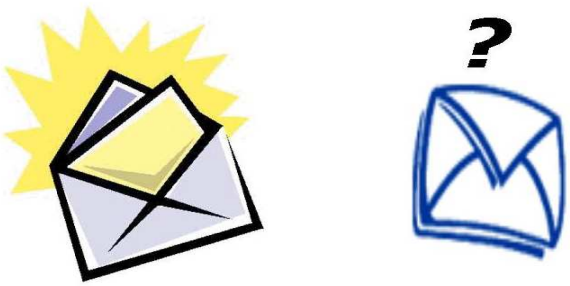
Figure 17.2:

17.1.2 Example: The Two Envelopes Paradox

The two envelopes paradox is a good example about statistics, where one thinks the statistical analysis is completely correct, but the result can simply not be true. Thus it is a paradox.

Lets assume we have two envelopes, each contains a certain amount of money, with the one envelope containing twice as much as the other. Now the quizmaster (again) lets you pick one of the envelopes. You can open the enevelope and see the amount of money, $X_\$$, that is in it. Now the quizmaster asks you, wether you would rather like to have the other envelope?

Now the question: Shall you go for the other envelope? Or stated differently, more in the context of this statistic lecture: Is the expected values larger when you take the other envelope? The question seems totally absurd: Why should there be any different in the expected values for the two envelopes? One may, obviously, think that the expected value of money in any of the two envelopes must be the same. So lets compute the expected value: Following eq.[3.3] the expected value for of a ramdon variable is:



$$\mathcal{E}(\mathbf{X}) = \int_{\Omega} x f_X(x) dx \tag{17.1}$$

The integral in this case reduces to a sum over all possibilities. In the case we keep our opened enevelope the expect values is of cause the alue we have in our envelope:

$$\mathcal{E}(\mathbf{X}) = \sum_{i=1}^{N=1} x_i f_X(x_i) = X_\$ 1.0 = X_\$ \tag{17.2}$$

So there is no uncertainty. The amount of money you get is what you already have, $X_\$$. In the case you want to take the other envelope you find that you have two possibilities: you either have half as much or twice as much money:

$$\mathcal{E}(\mathbf{X}) = \sum_{i=1}^{N=2} x_i f_X(x_i) = \frac{1}{2} X_\$ 0.5 + 2 X_\$ 0.5 = 1.25 X_\$ \tag{17.3}$$

So the surprising result is, that you will in average have more money when you take the other envelope. So it seems the grass is allways greener on the other side!

What is wrong here?

As the subject of this section may suggest, there is a hidden impossible assumption here: You can always double the amount of money $X_\$$, which is indeed impossible. The two enevelope paradox is set-up in a way that it is assumed that any value of $X_\$$ is possible, and therefor the expected value of $X_\$$, before you have opened any envelope is ∞. But in probabilities you cannot deal with ∞. For any realistic situation the the expected value of $X_\$$ finite, and therefore the probability of large $X_\$$ most become small. For any such *pdf* the expected value for both envelopes is always the same. A simple case is illustrated in Fig. 17.3. If the amount of money $X_\$$ is larger than 1/2 the budget limit the likelihood of doubled money becomes zero.

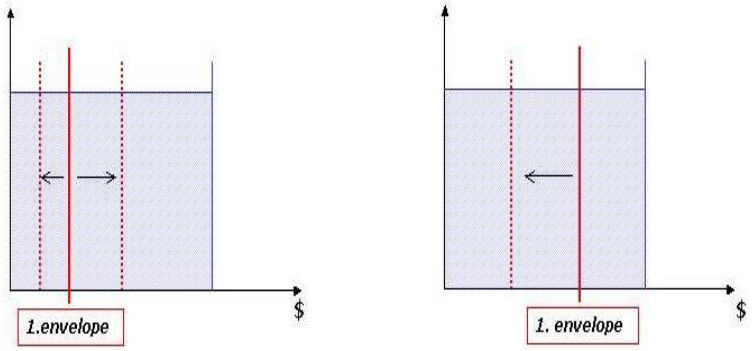


Figure 17.3: Illustration of the two envelope problem with a limited budget. In the first case (left) the amount of money $X_\$$ in the opened envelope was below 1/2 the budget limit, thus other envelope may have either 1/2 or twice as much money. In the second case (right) the opened envelope has more than 1/2 the budget limit and the other envelope must have 1/2 $X_\$$.

17.1.3 Example: Landfall of Hurricanes

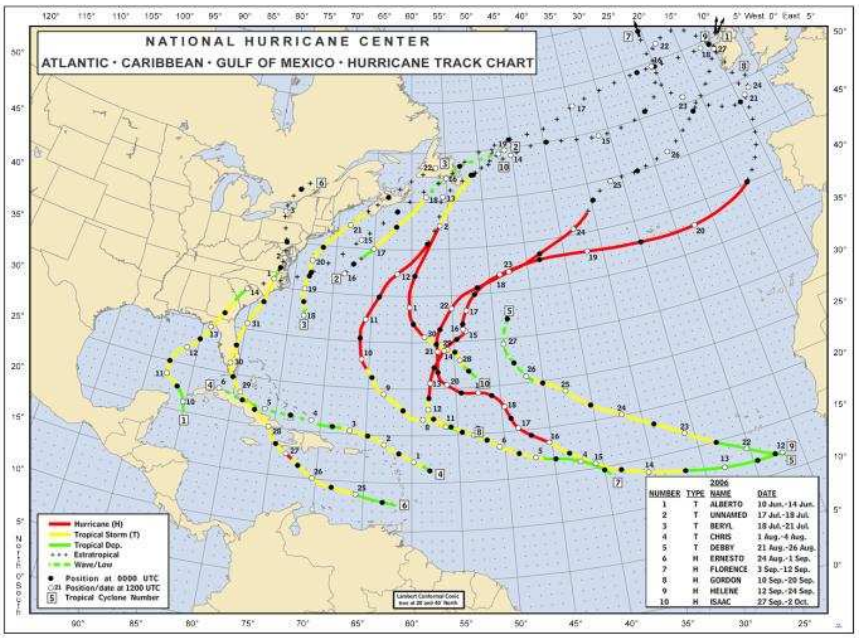


Figure 17.4: Path of Hurricanes in the 'calm' 2006 season.

An intersting study entitled: "Spatial Variations in Major U.S. Hurricane Activity: Statistics and a Physical Mechanism" by Elsner et al. 2000 in J. Climate finds a relationship between the number of hurricanes per year that make landfall along the Gulf coast versus those that makse landfall along the east coast of the USA. They find that the numbers are anti-correlationed, when more hurricanes in one year have struck the Gulf coast, then less hurricanes have struck the east coast. They further suggest that some climate variability (the NAO) may control the pathways of hurricanes. So some boundary condition may in one year lead most of the hurricanes in one direction and less in the other.

It seems strait forward to conclude, that if the number of hurricanes per year that make landfall along the Gulf coast versus those that make landfall along the east coast of the USA are anti-correlated, than there most be something that controls the direction of the hurricanes. But, surprisingly, this is not correct! There can be an anti-correlation even if there is no control on the direction of hurricanes, as will be demonstrated below by a simple model.

Fig. 17.4 shows the pathways of hurricanes in the year 2006. We can see some go into the direction of the Gulf, some to the east coast and some go to the North Atlantic. An important thing to note here is, that all hurricanes come more or less from the same region.

So the hidden assumption in Elsner et al. 2000 is that if the direction of hurricane tracks is uncontroled (random) than there would be no anti-correlation between the landfall numbers. In order to test this idea we can set up a simple Monte Carlo bootstraping model for the correlation. We assume the following for the hurricane statistics, which is roughly based on observed statistics:

1. All hurricanes originate from the same source region, see Fig. 17.5.

2. The number of hurricanes per year is nearly normal distributed with about 7 expected hurricanes per year, see Fig. 17.5 lower left.

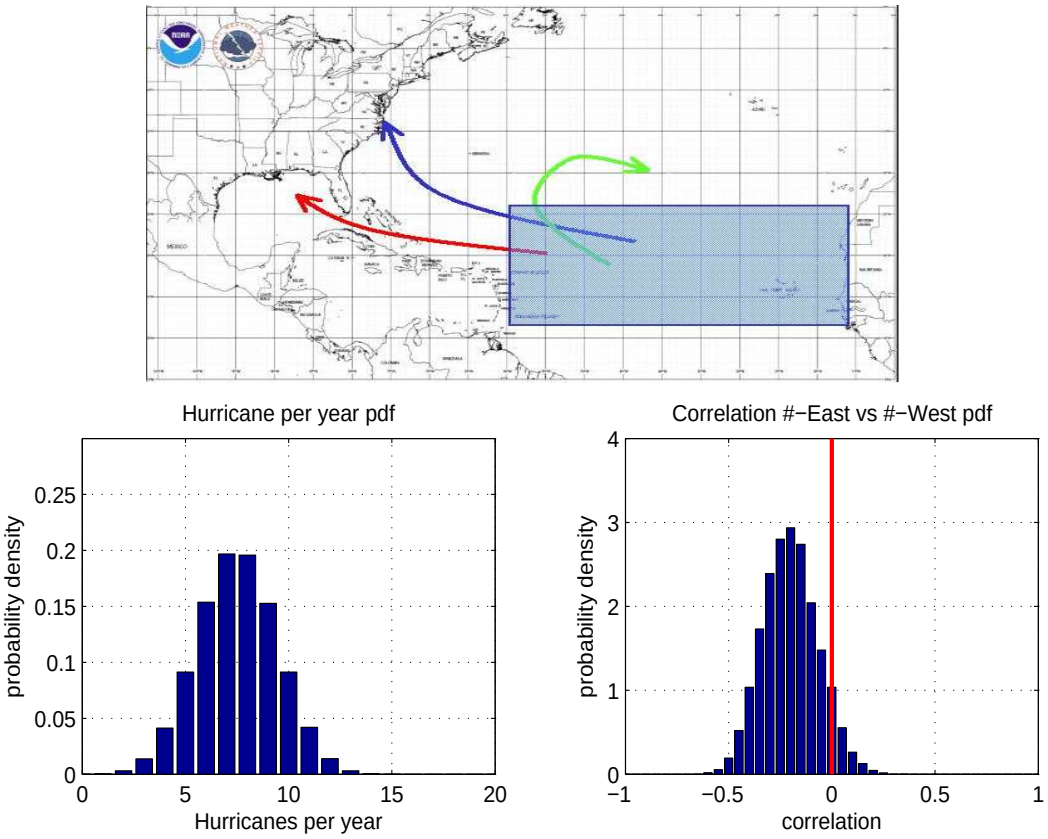


Figure 17.5: A Monte Carlo bootstrap model study of the distribution of correlations between numbers of hurricanes landfall at the East coast and the Gulf coast. Upper: Schematic of Hurricane paths; blue box is the source region; the red, blue and green line mark the three possible path ways. Left: Distribution of Hurricanes in source region. Right: *pdf* of the correlation between number of Hurricanes per year at the East coast and the Gulf region for a 100yr long time series.

3. The hurricanes leave the source region randomly into one of three possible directions: Gulf coast, east coast or the open North Atlantic Ocean, see Fig. 17.5. The probability for each direction is 1/3.

4. Create 100 years of statistics and compute the correlation between the number of hurricanes per year that make landfall along the Gulf coast versus those that make landfall along the east coast of the USA.

Creating some statistics with this model for about $10^4 \times 100$years, we find the *pdf* for the correlation between the number of hurricanes per year that make landfall along the Gulf coast versus those that make landfall along the east coast of the USA, see Fig. 17.5 lower right. The expected values is clearly below zero. Thus we expect an anti-correlation eventhough nothing controls the direction of the hurricanes. Why is that? The important point in this problem is that a limited number of hurricanes emerge from the same source region. If one of the hurricanes goes to the Gulf coast, than the number hurricanes left in this year, that can go to the east goes is reduced, because there is only a limited number of hurricanes per year, which all emerging from the same source region. If the problem is simplified to a fixed number of hurricanes per year (no variations any more) that have to go either to the gulf coast or to the east coast (no other direction posible), than the number of hurricanes per year at the two coasts must be perfectly anti correlated. The more the

number of haurricanes varries or other directions for the pathway of hurricanes exist the weaker the correlation.
This example demonstrates that it is sometimes not that easy to see what the expected value a statistical parameter of a Hyopthesis or anti-these is. In the above model there are of cause parameter regimes in which the expected value of the correlation is zero, but for the realistic parameters chosen, we get the unexpected result. So one always have to be careful in using unusual statistical parameters, such as the correlation between the landfall numbers per year from two regions. If you do, you should be aware of what you expect for the statistical parameter if your hypothesis is not true.

17.1.4 Example: The Role of the Indian Ocean for the ENSO mode

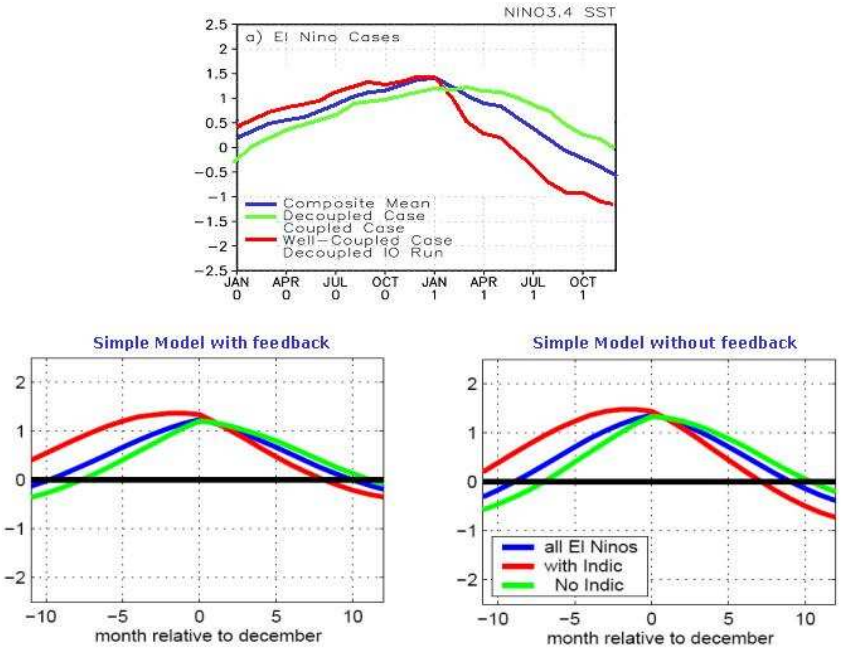


Figure 17.6: .

17.2 False Assumptions

examples: EOF-modes

17.3 Theory Bias

Academic researchers or students are obviously heavily theory biased. We spend much of our time being educated in the theories of our research areas. When we finally, after many years of theory training, get into real world problems we will have a tendency to overvalue our theories in situations where the empirical evidence should tell us to rethink the theory. Consider the following example.

17.3.1 Example: Fat Tony and Dr. John

Lets assume a fair coin is flipped 19 times and each time is comes up with a head. What are the odds that the 20th flip is head again?

The obvious answer for the well educated Dr. John (us) is: "The chance is 50/50, as the odds are not affected by the previous events."

However, a well experienced, street smart (lots of experience in real world problems, but no theoretical education), Fat Tony would say: "The chance is about 20/1. Screw you, no way that this is a fair coin tossing!"

Now we are in the dilemma that either the assumptions are wrong (fair coin) or we have to assume a 50/50 chance. If you consider that this a real world situation, then it seems much more likely that the assumptions are wrong than to assume that this is a fair coin tossing that result into 19 times with heads. The probability of 19 times head is $0.5^{19} = 2 \cdot 10^{-6}$, assuming a fair coin tossing. This only happens once in 500,000 tries. How likely is it that assumptions are wrong? This is hard to quantify, but it certainly happens sometimes, properly more often than once in 500,000 times. Thus we should conclude that the assumptions are wrong and not that we have indeed observed a fiar coin tossing that result into 19 times with heads, because a wrong assumption seems more likely.
We (the well educated Dr. John) tend to follow theories and be ignorant to empirical evidence. Fat Tony, in turn, ignores theories and considers empirical evidence. We can learn from this:

- Empirical evidence over rules theory (model).
- The real world if often more complex than the theory.

17.4 Confirmation Bias

We as humans tend to avoid rejecting theories, which is most likely in our genes. Back in the stone age we had an environment, which may not support good scientific behaviour. Consider a typical stone age situation (which is what is in our genes): All your mates come dashing away from something. This signals to you the following theory: Something very bad is coming your way that will kill you unless you start dashing away too. As a good scientist you would not just agree to this theory and start running away too, but you would like to first of all test this theory. Try to falsify it, by looking at what is really going to happen if you do not run away, but go look what is coming. ... you will not survive for long. Even today we have an environment that probably does not encourage to falsify theories. In the education system you have to have good marks and you dont spend much time on trying to challenge what you learn. As researchers you tend to focus on publishing a lot and new interesting results. There is not much time for rejecting theories.
Consider the following test as an example of our *Confirmation Bias*: Consider the following row of numbers:

2, 4, 6, ...

Now you should figure out the underlying rule. You can try a sequence of numbers to find the rule and I will tell you if your sequence fits the rule or not. You can try as often as you like and if you think you know the rule you can try to state the rule, but only once.
Typically the students will start and testing the rule by testing the hypothesis of :$X_{i+1} = X_i + 2$. they will figure out that this hypothesis works. They will have confirmed their theory every time and every time they have confirmed it they feel more certain that their theory is true. But you properly also note, that the only way to be sure about this is to test an alternative hypothesis that does not fit to your hypothesis. For instant, you may test: *1 4 7*. Then you will, to your surprise, figure out that this sequence of number is also consistent with the underlying rule, which is: $X_i < X_{i+1}$, and it is not consistent with your hypothesis. Thus falsifying your thesis is the only way to figure out if your thesis is correct. Unfortunately, students often tend to not do this, but guess the underlying rule, before they have tried to falsify it. So the important leassons we can learn from this is:

- We tend to be biased towards confirming theories.
- We neglect alternative theories.
- We are over confident: assuming verification of theory much too early.
- Considering alternatives and rejecting them is essential for verifying your hypothesis.

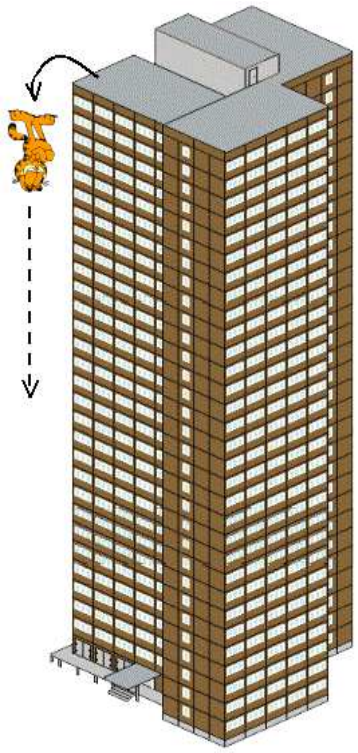
17.5 Biased Statistics

If we estimate statistical parameter from observed statistics we assume, that the samples are taken randomly or in fixed intervals. The decision to take a sample or not, must not depend on the state of the system itself, which is for observations often not the case. In many cases we will have baised statistics and the following examples may illustrate what can go wrong if you try to draw conclusions from such statistics. Other examples have also been discussed in sections 2.2 and 5.1.

17.5.1 Example: The Feline Multi-storey Building Syndrom

From the New York Times 22.August 1989 (translated from egnlish to german):

"Experten haben verblueffende Anhaltspunkte fuer die Ueberlebensfaehigkeit von Katzen gefunden, diesmal in New York, wo Katzen im Sommer haeufig aus den offenen Fenstern von Hochhaeusern fallen. Wissenschaftler nennen es 'das feline Hochhaussyndrom'.
Von 132 solcher Opfer, die in die Tierklinik aufgenommen wurden, ueberlebten die Mehrheit. Laut Experten ist dafuer die Physik und die 'Taktik der fliegenden Eichhoernchen' verantwortlich ... der Flug reichte von 1 bis 31 Etage ... 17 Katzen wurden von ihren Besitzern, vor allem aus Kostengruenden, eingeschlaefert. Von den uebrigen 115 starben 8 durch Schock oder Brustkorbverletzungen.
Noch erstaunlicher war, dass die Chance zu ueberleben umso groesser war, je laenger der Sturz dauerte. Nur eine der Katzen, die aus 7 oder mehr Stockwerken Hoehe gefallen, starb, und es gab nur einen Knochenbruch unter den 13, die mehr als 9 Stockwerke gefallen waren. Die Katze, die aus dem 31. Stock fiel, Sabrina, hatte nur leichte Verletzungen an Lunge und Gebiss.
Warum hatten Katzen, die aus grorsseren Hoehen gefallen waren, bessere ueberlebenschancen? Zum einen leigt die Terminalgeschindigkeit, die beim Menschen ca. 200 Km/h betraegt, bei Katzen nur bei ca. 100 km/h. Man vermuttet, Katzen wuerden, bevor sie diese Geschwindigkeit erreichen, ihre Extremitaeten ausstrecken. Wenn sie die Terminalgeschwindigkeit erreicht haetten, wuerden sie wie die fliegenden Eichhoernchen ihren Luftwiderstand maximieren und dadruch den Aufprall abfederen." (Also wie Fallschirmspringer, die erst ab einer gewissen Hoehe ueberleben)



The article may be summarized to the following:

- **Data source:** 132 cat brought into hospital after they jumped from a high building in Manhattan.
- **Phemomenon:** Cats have a high probability of surviving a jump from a high building, the high the level they jump from.
- **Explanation:** Cats use the tactic of a flying squirrel.
- **Observations:** 117 cats that jumped out from a tall building, with the following casualties statistics:

  - 8 died and only of them jumped from a level above the 7. floor.
  - From the 13 cats that jumped from the 9. floor or high, there was only one with a broken bone.
  - The cat Sabrina, which jumped from the 31. floor had only minor injuries at the denture and lungs.

Some minor problems with the authors interpretation of the statistics:

- **These is not supported by the data.** Even if we consider the data at hand, the authors conclusion is unsupported.
- **Small data basis.**
- **'Hand waveing' physical explanation.**
- **Selection of unusual stories.**

However, the main problem with this study is the use of biased statistics. In order to understand the casualties statistics of cats jumping from high-rising buildings, you need to take samples of jumping cats purely random, and not just those that were send into the hospital. You are missing some cats and they are missing for specific, non-random reasons:

- Cats jumping from a lower floor, not getting hurt at all and happily going to stroll through the streets with other cool cats. Those missing cats will cetainly dominate the statistics for lower level jumps and therefore make the casualty probability for lower level jumps very small.
- Nobody will bring a dead cat to the hospital. Cats that clearly do not need a doctor anymore, which are most likely those that jumped from high floors, will most likely 'spoil' casualty probability for higher level jumps.

17.5.2 Example: The Broken cloud effect

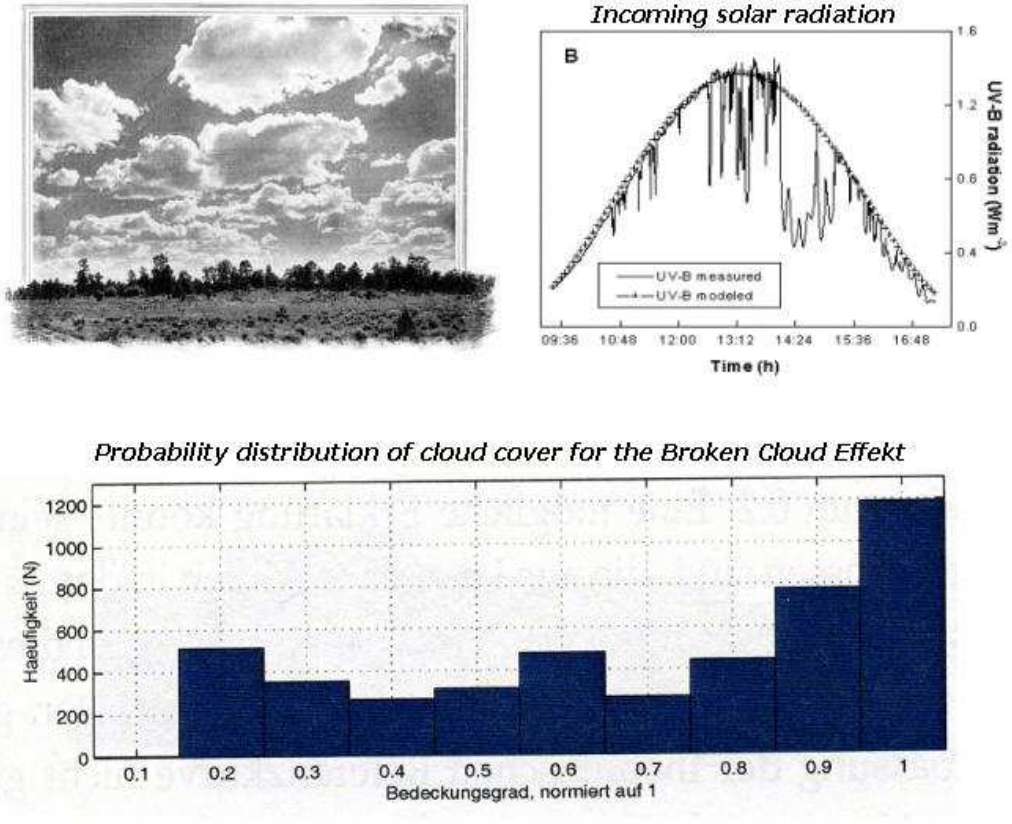


Figure 17.7:

The broken cloud effect (BCE) describes the effect that the incoming solar radiation at the surface is in partly cloudy conditions sometimes larger the clear sky incoming solar radiation, which is due to the effect that the boundary of nearby clouds scatter sometimes (over short time intervals) additional sun light onto a surface, see Fig.17.7.
One may want to know under which cloudiness the broken cloud effect (more than 100% incoming sun light due to additional cloud reflections) is strongest. For this one may plot the *pdf* of the BCE as a function of cloudiness (Fig.17.7). However, this histogram does not say if the effect is related to cloudiness, because the samples may be taken under specific cloudy conditions. If we have only measured a certain type of cloudiness, than we will have most BCE for this cloudiness. We have to fold the *pdf* with the *pdf* of the cloudiness itself. A small probability of BCE may be due to only few measurements for this cloudiness.

17.6 Framing ... the Opposite of Objectivity

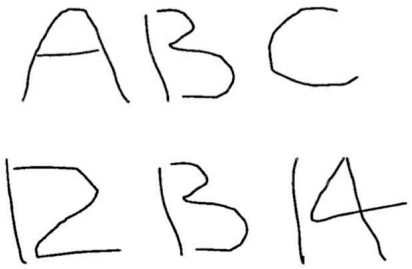


Figure 17.8: Note the "B" or is it a "13". Our interpretation depends on the "Frame", but for an objective researcher it should not.

Framing, in the context of this course, is similar to use of the word in the media and social sciences. It means how you present your work to the audience or how you put things into perspective. By framing your scientific results you want the audience to get a specific conclusions. This is problematic, as it may indicate that your research or analysis is not objective: You aim for a certain outcome, but this may only be because of the frame you set, but not because the outcome is in an 'objective' way achieved no matter how you look at it.
Have a look at the sketch 17.8. In the upper line you are forced (framed) to interpret the symbol in the centre as a 'B', but in the lower line you are framed to interpret it as a '13'. The interpretation of the symbol depends on the frame. A scientific result should not depend on the frame. It should be on objective truth. But for us as humans we find it very difficult to evaluate anything in an objective way. An objective point of view really does not exist. However, as a researcher we try to present an objective point of view.
In particular in high-profile science publication the authors seem to set a frame that make the audience interpret the result in a somewhat bold way, which, putting it mildly, would mean that they stretch the interpretation beyond a point where you may want to argue that this is objectively not a correct interpretation of the analysis. The following few cautionary point you should keep in mind:

- We consider things within its frame work
- We do not have an objective view point
- A scientific approach aims for an objective approach
- There is no such thing as an objective view

The next example is about a high-impact study from the literature. It indicates a 'framing' problem.

17.6.1 Example: The Trend in the North Atlantic Oscillation

The study from Hurrel et al., (1995) in science is one of the most cited publication in the field of climate variability (4000+ citation by 2015). This is on the same level as Nobel price studies. Fig. 17.9 shows on of the important figures in this study. Here is a citation of one of the important points from Hurrel et al., (1995):

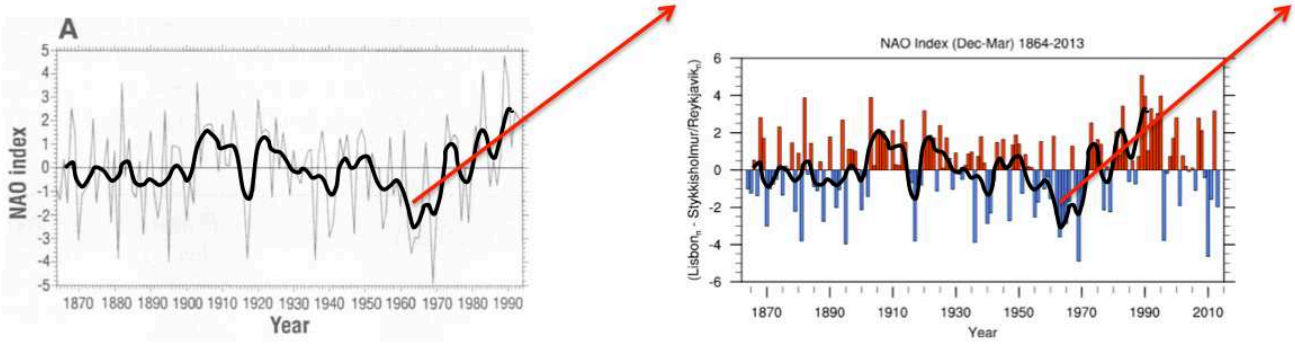


Figure 17.9: Left: Time series of North Atlantic Oscillation (NAO) as in the Hurrel et al. 1995. A suggestive trend line is plotted over it. Right: an updated version of the NAO time series with the same trend line.

... the past decade resemble some results obtained by coupled atmosphere-ocean models forced with steadily increasing atmospheric greenhouse gases, ... Decadal variability in the NAO has become especially pronounced since about 1950, but the causes for such variability in the Atlantic are not clear. The relation of the NAO to greenhouse gas forcing and possible links to coherent variations in tropical Atlantic SSTs need to be examined, ...

I think one of the reasons why this paper had such a big impact is, that we tend see the study in the context of climate change (e.g. trends) and decadal variability. The study is 'framed' this way. In Fig. 17.9 we tend to interpolate a linear trend (see red arrow), because the study is framed in a way that we 'see' a trend at the end of the time series. It is very disappointing in this context to see how the real world behaved the two decades after the publication of the paper (see right panel). It is fiar to say that including the two additional decades into the study would have weakened the story and the framing would not have worked. This illustrates that 'framing' of scientific results can mislead the research community, which should objectively not have 'seen' a strong linear trend in this time series, as it objectively was not there.

17.7 Problems with Probability

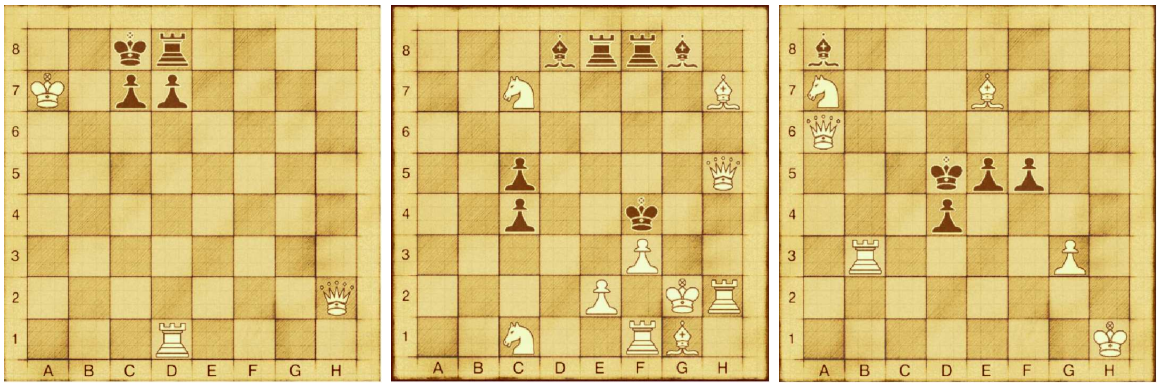


Figure 17.10: White to move. Mate in two.

17.7.1 Example: Regression to the Mean

In section 4.1.1 we discussed problems in interpreting correlation. In this context we discussed the *regression to the mean* problem, which indicated that we interpret correlation as a deterministic model, e.g.:

$Y = \rho \cdot X$

This models complete ignores the noise (randomness, stochastic nature). We tend to neglect the interpretation of the randomness. So the important conclusion was that:

- We interpret correlation as a deterministic model, ignoring that is also models the randomness.
- Correlation does quantify a stochastic process.
- We over interpret noise as a deterministic model.
- We find if difficult to include the effect of randomness (noise, stochastic process) into our model/explanation.

17.7.2 Example: A Logical Chain with Probabilities

In a mathematical proofs you can argue if $A \Rightarrow B \Rightarrow C \Rightarrow D$ then it follows $A \Rightarrow D$. Thus a very popular approach in statistical analysis is building up logical chains like: **A** is correlated to **B** and **B** is correlated to **C** ... correlated to **Z**, concluding **A** is correlated to **Z**. Unfortunately this is almost certainly not the case. Two examples to illustrate this.
A statement of a mathematician Emil ????? (name unkonwn) was something like the following (translated from german):

"Statistic is the same as probability theory.
Probability theory is nothing else than correlations.
Correlation is simply the cosine(angle).
Thus Statistic is trivial!"

So we have a number of statements where each has some trues in it, but there is also some uncertainty in it. So we could assign a probability for each statement to be true. First we have the following variables:
A = Statistic
B = probability theory
C = correlation
D = cosine(angle)
E = trivial

Now we assign each statement of Emil ???? a probability:

$$P(A = B) = 0.7$$

$$P(B = C) = 0.7$$

$$P(C = D) = 0.7$$

$$P(D = E) = 0.7$$

So in a optimistic linear model we would evaluate Emil ????? conclusion with,

$$P(A = E) = \textit{Statistic is trivial}$$

$$P(A = E) = P(A = B) \cdot P(B = C) \cdot P(C = D) \cdot P(D = E) = 0.7^4 = 0.24$$

Thus *statistic* has not much to do with *trivial*. But it gets worth for Emil ?????. Even if A is correlated to B and B correlated to C, we do not know anything about the correlation between A and C. So this logical chain does not work in statistics, where very single statement has some uncertainty. Such a logical chain is only good for an post priori explanation of the observed fact that A is correlated to E, which in Emil ???? example ("*Statistic is trivial!*") is certainly not the case.

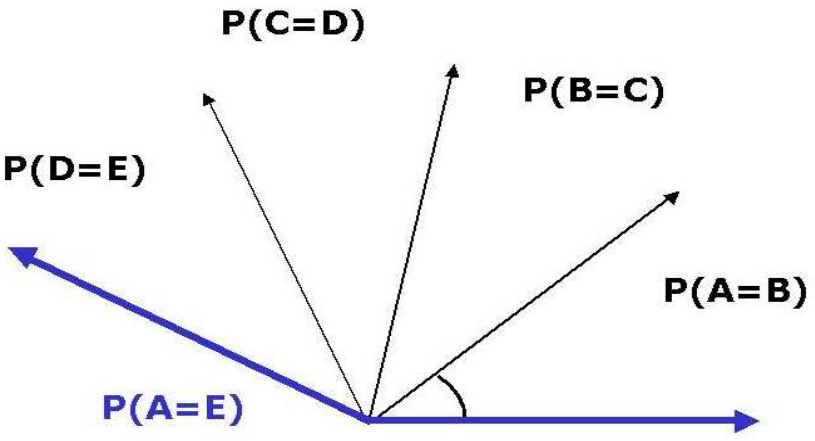


Figure 17.11:

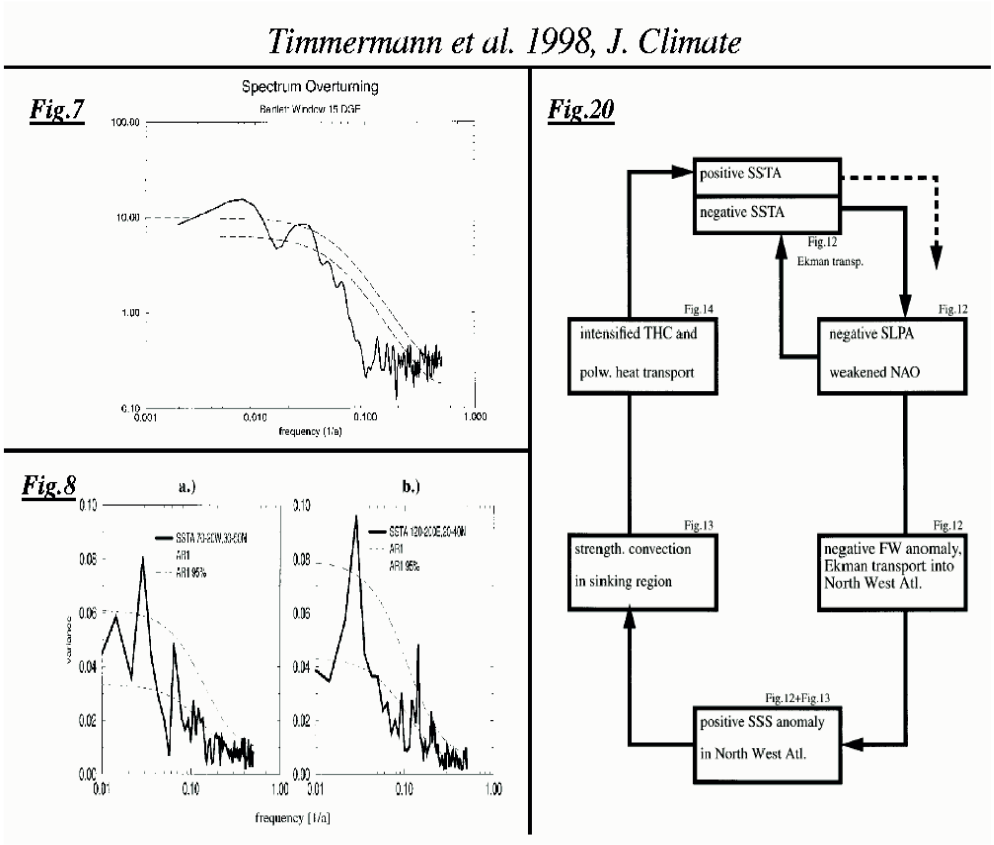


Figure 17.12:

17.7.3 Example: A Decadal Climate Mode

Timmermann et al. 1998, J. Climate is a good example of a logical chain of statistical inferences. The paper is a highly cited documentation of a decadal oscillation in a coupled GCM simulation, see Fig. 17.12. We see that the authors first of all find evidence for oscillations in the overturning of the North Atlantic and in the SST. As an explanation for this oscillation they build up a feedback loop, while they present evidences for each element of the feedback chain. Note that even if all elements are found to be true, the SST in the North Atlantic does not need to oscillate. The feedback loop only illustrates what could explain oscillations, if the oscillation is existing.
The above examples together with the little chess riddles illustrate that logical chains, even short ones, are difficult to see through.

17.8 Fishing for Something

Often the data at hand is analyzed with no particular aim. The only aim is to find some interesting structure in the data, and interestingly you often find something interesting. But it often turns out to not that intersting after all. The following examples shall illustrate the problem. See also Section 15.3.1 where we discuss the multiple use of local test and the need for global tests.

17.8.1 Example: The Mexican Hut

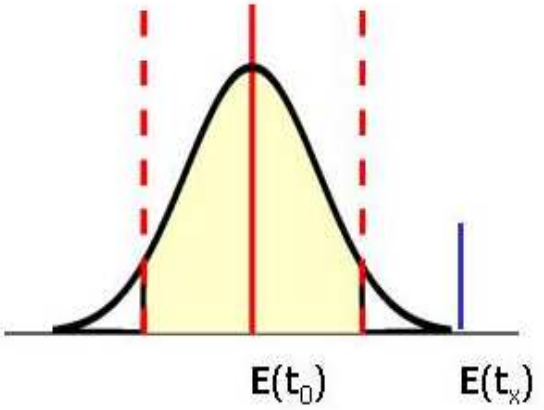


Figure 17.13: Left: The Mexican Hut rock formation. Right: Schematic of a hypothesis test with the blue line indicating the testvalue for the Mexican Hut.

The Mexican Hut (Fig. 17.13) is an unusual rock formation, which may raise the question: Is this a natural rock formation? Being a statistician we test the null hypothesis: The rock formation is natural. We therefore need to define a test variable, which quantifies the rock formation. We would come up with a variable T_{rock}, that would most likely evaluate the size of the rock relative to the height above the ground and the area that actually holds the rock in place or something similar. No matter how you would try to quantify this test, you would find that the null hypothesis must be rejected, thus: This rock formation is not natural.
Any geologist would, however, tell you that this rock formation is caused by an eruded soft bedrock below a solid rock formation. So the scientists with some physical understanding of this system

would dispute the statistician finding, no matter how strongly the statistician claims that their finding is highly significant. So what is the problem with the statistical approach?
The main problem here is that we build a test around an unsual finding. Or stated differently: We assume we made only one test, but in fact we have made (although not that obviously) many tests and only discuss the one that finds the significant result.
Building a test around an observation, which we think is interesting, will of cause lead to a rejection of the null hypothesis. So statistical tests should not be build after the data has been analyzed. In fact the second point of view, that we actually made many tests, is important here to understand the outcome. If we apply a statistical test, than the significance level of our test (e.g. 95%) only refers to a test situation where we have applied the test only once. If we have applied the test to many different samples, than the probability of passing the confidence level will strongly increase, just as throwing a 6 with one dice is less likely as throwing at least one 6 with 10 dices. Thus in statistical test you need to consider all failed tests in your statistics. In the case of the Mexican Hut we reconized the unusual formation, because we have already seen many other 'normal' rock formations. The likelyhood that one out of many thousand rock formations is as unusual as the Mexican Hut seems much less 'unusual'.

17.8.2 Example: A Decadal Climate Cycle in the North Atlantic

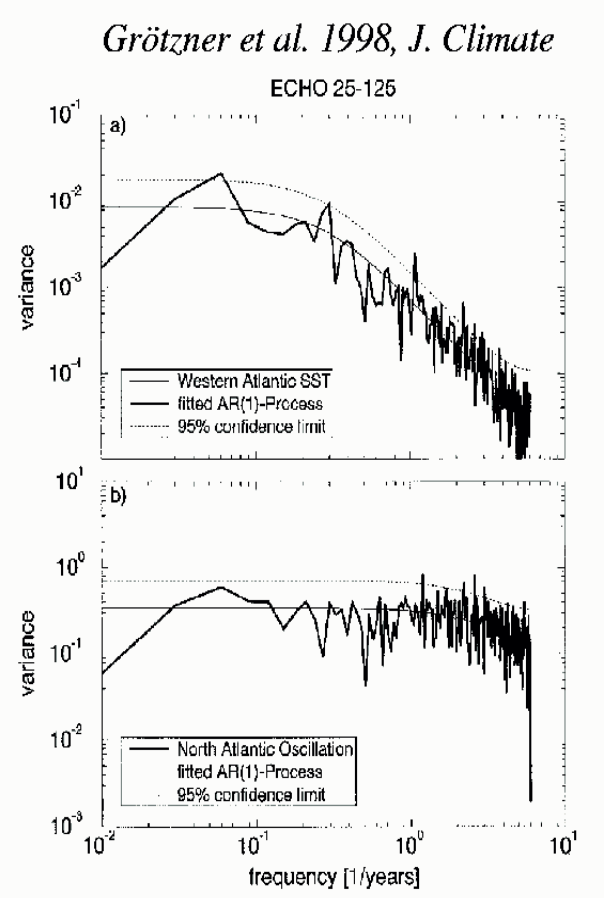


Figure 17.14: The Fig. 5 from Groetzner et al. 1998, J.Climate. "*Fourier spectra of anomalous western subtropical North Atlantic SST (*$30^o - 40^oN$, 45^o90^oW*) and of an index describing the model's North Atlantic Oscillation (*$60^o - 70^oN$ $0^o - 45^oW$ *minus* $35^o - 45^oN$, $O^o - 45^oW$*).* "

Groetzner et al. 1998 "A Decadal Climate Cycle in the North Atlantic Ocean as Simulated by the ECHO Coupled GCM" is an interesting example on how the confidence interval of spectral estimates are interpreted, it is also one of the highly cited papers on decadal variability. The Fig. 17.14 is the Fig. 5 from Groetzner et al. 1998. The authors say the following about this figure:

"*The existence of a dominant decadal timescale in the simulation, ... , is supported by two selected Fourier spectra shown in Fig. 5.*"

further they say:

"*At a period of 17 yr the spectrum exhibits a clear peak that is significant at the 95% level above the red noise background. "*

"*The dominant 17-yr timescale is related to an oscillation that is characterized by propagating temperature anomalies within the upper layers of the North Atlantic Ocean. "*

The discussion of this spectra by Groetzner et al. 1998 is typical for many other publication in the literature. So what is the problem with Groetzner et al. interpretation? I think from the statements one may assume that the probability that a red noise spectra would have a similar peak is only 5%. This is not true! The confidence level for the spectral estimate describes the probability that one of the spectral coefficients passes the confidence level. If the I like to know what is the probabilty that at least one out of N spectral coefficient passes the confidence level than the probability is (see section 2.1):

$$1 - 0.95^N$$

With $N \approx 20$ for the spectra in Fig. 17.14 for periods longer than one year, we find that the probability is larger than 60%. Thus a peak above the confidence level in this spectra will most likely appear.
The problem is again that the authors of this study have indeed done many tests, but they interpret there statistical significance level as if they have only done one test. A more appropriate way of doing such a test is illustrated in section 16.1. In section 15.3.1 we also discuss that is many local tests are done, then we need a global test that evaluates the likelihoods of all test results together.

17.9 Summary of Common Problems in Statistical Inferences

Chapter 18

Strategy

Before you start a statistical analysis it is helpful to develop some strategy of how you would like to approach your statistical analysis. In this chapter we like to take a look at a few concepts that help you to develop meaningful statistical study.

In the first section we take a look at the general limitations of statistical analysis, illustrating that statistical analysis can in general not proof anything, but can only give some circumstantial evidence. In the second section we will outline the two main approaches in statistical analysis and discuss the different philosophies behind them and what their advantages and drawbacks are. In the final section of this chapter we outline a simple method that should be a basic principle for most statistical analysis.

18.1 Empirical Proofs (A world full of non-elefants)

Quite often statistical analysis is carried out to 'proof' a theory. However, you have to keep in mind that statistical analysis can never proof any theory in a mathematical sense. An examples helps to understand the problem: Let assume we want to proof that all ravens are black. In mathematical terms we want to shows that from A (raven) it follows B (black): $A \Rightarrow B$. So practically we have to observe all ravens and verify that they are black. This is of cause impossible. We may now consider a mathematically alternative approach: If we can not proof $A \rightarrow B$, we may follow the approach to proof if not B than it does not follow A: $\neg B \quad \Rightarrow \quad \neg A$. If we can proof this, than we have indirectly proven $A \Rightarrow B$. So we observe everything that is not black and show that none of these non-blacks are ravens. So every yellow car proofs that ravens are black? ...

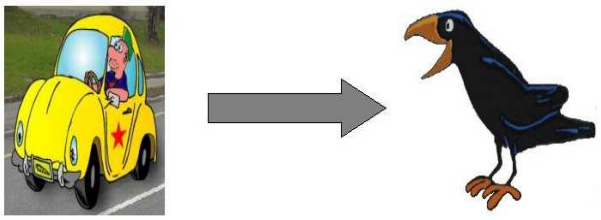


Figure 18.1: Illustrate that in statistical analysis observations of yellow cars can proof that all ravens are black!

18.2 Confirmatory and Exploratory analysis

Scientific discoveries are often driven by two different approaches: First you may have developed a theory or model from which you draw some conclusions, that you like to verify with observations. Example 1: Einstein's relative theory claimed that the space bends around heavy objects like the sun. This was verified by the observations of stars bending around the sun during a solar eclipse. Example 2: Global warming is assumed to follow if the CO_2 concentration in the atmosphere is increasing. Which is currently verified by the collective human activities. In statistical analysis we call this approach 'Confirmatory Analysis'.

The second approach is to observed the real world and try to understand what you see. So you typically have observational findings that are first of all not understood in theory; they contradict or reject current theories or null hypothesis. Thesis 'discoveries drive the development of new theoretical frame works. Example 1: Again the basis for Einstein's relativity theory was the empirical finding that the speed of light is a constant. Example 2: ENSO was an observational finding that initially had no theoretical framework and still does not have a complete theory.

We can summarise the main characteristics of Confirmatory and Exploratory analysis in a short table:

Confirmatory analysis (Hypothesis)	**Exploratory analysis** (Null Hypothesis)
approach	
Theory/Model → confirm with observations Hypothesis → test with observations	Observe phenomenon → develop model/theory reject Null Hypothesis → build better model
advantage	
process understanding	objective approach; discussion of complete observations
drawback	
subjective approach; ignorant towards observations; role of alternative hypothesis unclear	No physical model/theory; Sensitive to pitfalls in statistics; over-interpretation of noise

Example: What is causing climate variability/change in the North Atlantic?

Confirmatory analysis (Hypothesis)	**Exploratory analysis** (Null Hypothesis)
approach	
Theory: THC controls climate Experiment: Water-hosing Observations: study relation between THC/MOC indicies and the climate	Observations: study structures of patterns (EOFs) and time series of dominant modes. Model: associate leading modes with known physical processes (THC, NAO) → develop model
drawback	
role of alternatvie hypothesis unclear; role of THC is overstated	Statistical indicies are over-interpreted Sensitive to pitfalls in statistics

18.3 Toy models

18.3.1 Example: The Delayed Action Oscillator for El Niño

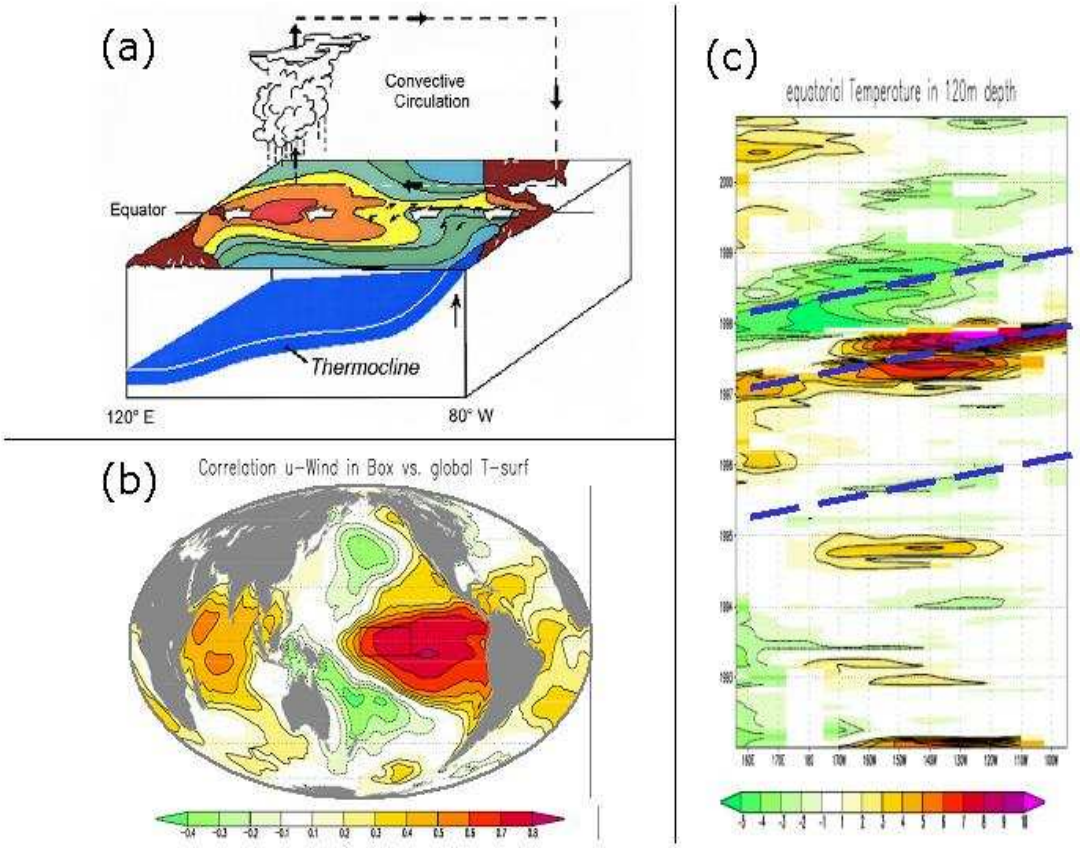

(a)
Convective
Circulation
Equator
Thermocline
120° E
80° W
(b)
Correlation u-Wind in Box vs. global T-surf
(c)
equatorial Temperature in 120m depth

Figure 18.2:

Chapter 19

Tactics

In the following we like to like point out a few concepts that help in statistical analysis. The ideas presented here should help to organise the way your draw conclusions from statistical analysis and should help you in avoiding the most common pitfalls in statistical analysis. Not all of the following advices are helpful in all statistical analysis, but most of them will help in most common statistical analysing methods. The section starts with describing some tactics that at the end of this section will be summarised in a list of all concepts.

19.1 Independent Verifications

In statistical analysis you often develop a model or theory on the basis of some data. It is important that you verify your model or theory with some independent data to improve the significance of the result and in order to void statistical artefacts that got into the statistical analysis with out having noticed it.
Independent Verifications are based on data that is statistically independent from the original data from which the model or theory has been developed. Such data in climate analysis is usually from another time period, a different region (if possible) or from a model simulation. If possible one could split the original data into two parts: one for the development of the model/theory and one for it s verification. Note that you can only use your verifications data once. If you apply your model to the verification data and then change the model to improve it, then your verification data is no longer verification data, but it is the data on which you build your model on and you no longer have verification data.
Note that Independent Verifications is *NOT*:

- Applying a different statistical method on the same data. In spectral power analysis, for instants, one may want to use different spectral estimates to test if a peak in the power spectrum is indeed 'significant', meaning it appears in all different spectral power estimation methods. This approach only estimates if this peak is a clear signature in the time series, but does not test if this peak in the power spectrum is a significant feature of the stochastic process from which the time series originates. Thus it does not verify that this peak in the power spectrum would appear again in an independent time series.

- Using a different data set of the same time period for the same variable. Using satellite data instead of in situ observations, for instance, is not a statistical independent verification. In this example your are still analysing the same realisation of the stochastic process, but you only test for errors in the observations. Independent Verifications is not about avoiding errors in the observations. Independent Verifications is about avoiding to model unusual events, that are not 'typical' for the stochastic process you analyse.

- Splitting the data into two half after the model/theory has been developed.

19.2 Handwaving Physical Explanations

The term 'hand waving' is used in mathematics and physics to describe arguments that are not mathematically rigorous. This is quite common in studies based on statistical analysis. The studies of cats jumping from high levels in New York (section 17.5.1) illustrates this very nicely: The study apparently (not really as we have seen) finds an interesting phenomena (higher chance of surviving for jumps from higher levels) entirely based on statistical analysis of a single data set. The author argue that this can be explained by the 'flying squirrel' tactics of the cats, but no evidence what so ever is given in this study to support the idea that cats indeed follow this tactic. This is quite common in statistical analysis: A statistical result is presented, with statistical evidences, which are then explained by some physical process, which are however not supported in this study at all, but seem plausible. If you start speculating like this, you should clearly mark this as unsupported speculation, or much better, you should provide additional evidence by either citing studies that do support this physical explanation or by providing additional theoretical or experimental results.

19.3 Definition of Statistical Measures/Thresholds

In statistical analysis it is important that any inference from the results are drawn under well defined assumptions. Only if the test situation is well understood, we can apply standard test methods and understand the significance of the result (e.g. passing a confidence level). It is therefore that we need to define our statistical measures, methods or parameter before we apply it to the data, because our standard test in general assume that we have applied the test only once and unconditionally. If, on the other hand, we optimise our statistical methods after we applied it to data to 'improve' (optimise) the significance, we actually have created a test situation that does not fit to any standard textbook test any more. Indeed we have now done several tests that where unsuccessful. In statistical inference unsuccessful tests have to be included in the consideration of significance (e.g. in throwing dice we have to account for all dice that we have thrown, not only that have the desired outcome; trowing one 6 with one dice has a lower probability than one 6 out of 10 dice). Indeed if we start optimising the significance by adjusting the statistical method we will most likely fool ourselves by misinterpreting the significance of the result by incorrectly applying standard tests. See also section 17.8 "Fishing for something".

19.4 Optimal Presentation

The aim of statistical analysis or research in general is to understand nature. Understanding nature to some degree means that you make your self a picture it; you aim for presenting a clear structure in the otherwise chaotic data. It is therefore central to statistical analysis that you present the outcome of your analysis in the best and most easily understandable way. Some items to consider here:

- Graphs and figures that present the results of your analysis should be optimised for clarity. This could include right scaling of axes, optimising the contour intervals and range, optimising the choice of colours to highlight the important structures (e.g. contrasting regions of positive values from those with negative values). Even though this sounds trivial, but often studies are lazy in optimising the presentation of graphs and figures arguing that the main results are visible even though the presentation is not optimal. However, if you optimise your presentation you will quite often 'discover' some important structures. Indeed optimising the presentation is part of the statistical analysis.

- Tables and formulas need to be optimised in presentations just as much as graphs and figures.
- An optimal presentation does not mean that you tune your statistical methods to increase the significance of a signal. See previous section.

19.5 A Language Barrier between Statistics and Physics

Language barrier is a figurative phrase used primarily to indicate the difficulties faced when people who have no language in common attempt to communicate with each other (e.g. mathematicians and physicist). One may argue that mathematicians and physicist should have a common language, but this turns out to be the language of mathematicians. It is important to note that not everything that from a mathematical point of view is the best approach is also the best approach for developing a physical understanding of the data. Quite often we have to ignore the statistics textbook (not this one of cause!!) and have to use methods that from a mathematical point of view are not as optimal or elegant. Some examples:

- Example 1: 22 football players: From a mathematical point of view $2 \cdot 11 = 11 \cdot 2$. But from a football coach of view $2 \cdot 11 \neq 11 \cdot 2$: $2 \cdot 11$ are two teams, whereas $2 \cdot 11 \neq 11 \cdot 2$ are just a bunch of 22 players, which do not make two teams. So the coach has the concept of a 'team', which is more than the sum of 11 players. This may not fit to the concepts of the mathematician. In principle we may be able to define the concept of a 'team' in a mathematical way, but we have to recognise that quite often such concepts are not available.
- Example 2: EOF-modes: From a mathematical point of view the EOF-modes are the most efficient and most elegant way of representing the data. However, from a physical point of view EOF-modes are quite often of little help in understanding the physical processes that cause these structure in the data set, because they are a very complex (chaotic) superposition and over simplification of the data with almost no value for developing physical models of the data.

It is often tempting to develop a new or significantly modified statistical method to get an improved presentation of the problem/data. It is important to note that one may be able to develop a statistical method that indeed may be the 'optimal' way of presenting the problem at hand. However, in this consideration of what is 'optimal' one also has to take into account the readership or community one is addressing. There is no point in presenting a statistical method that the largest part of the readership or community can not understand or are not familiar with. This will in most cases undermine the significant of the results and will let the readership doubt the results. The community will tend to: Question if this results is only an artefact of the method or they will tend give little significant to the result as it is only a gimmick of an unusual approach.

19.6 Simple vs. Complex Methods

A general principle in research is to make things a simple as possible, but not simpler. This is also true for statistical analysis. You should always aim to use the most simplest methods to produce your results. The simpler your approach the more researchers will be able to understand your results and the more significant you results will appear. If you on the other hand apply a very complex method that in the context of your study does not seem to be well motivated (at least for m the point of view of the readership), than this will undermine you outcome. It will be interpreted as less significant or it may be interpreted as an artefact of your complex method and the 'real' structure of the data does 'really' support your outcome.

Figure 19.1:

However, in climate research it appears that there is a tendency to apply more complex methods than it would be necessary to make the point. It often seems that the researchers fall in love with a particular method and are more interested in applying a 'cool' and sophisticated method than in understanding the data. It seems they are often more interested in intellectual entertainment than in scientific results. A good analog is professional football: Clearly football is for entertainment and sometimes it seems the football player more interested in showing off a la Ronaldinho than winning the game (shooting) goals. Professional football player that do not entertain as much but are very efficient for the team to make it win the game a la Micheal Ballack are not as popular.
In research you should not entertain but aim for an efficient presentation and avoid complex methods whenever you can. Complex method are difficult to understand and it is difficult to understand the limitations of these methods and often statistical significance of the results are difficult to evaluate. An Example should help to understand this problem:

19.6.1 Example: MSSA Analysis of El Niño Period Shift

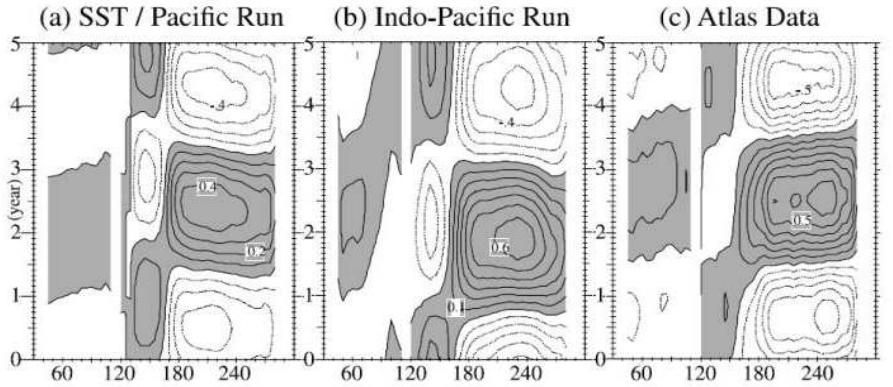


Figure 4. The eigenvector structures of SST anomalies along the equator for the leading oscillatory modes obtained by applying the combined 3-variable M-SSA analysis to (a) the Pacific Run, (b) the Indo-Pacific Run, and (c) the Atlas of Surface Marine Data 1994. The coordinate is the 61-month lag used in M-SSA. Contour intervals are 0.1 for SST. Positive values are shaded.

Figure 19.2:

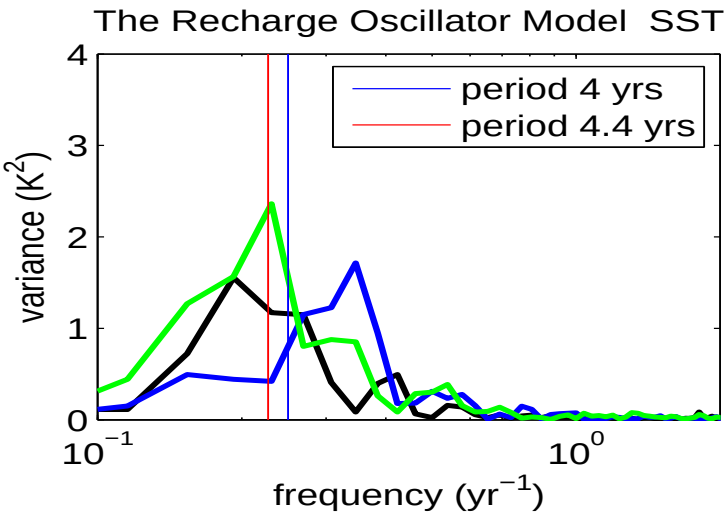


Figure 19.3: The power spectra of three independent realisations of an AR(2)-process with an oscillation period of about 4yrs. Each of the realisations is a 52yrs long time series similar as in Yu et al. [2002]. The log linear presentation is used ($\Gamma \cdot \omega$).

19.7 Hypothesis, Null Hypothesis and Anti-thesis

In complex statistical analysis it is important to formulate a hypothesis or null hypothesis as it helps to understand the outcome of the statistical methods applied.

19.8 Hierarchy of Methods

In statistical analysis we usually deal with a highly complex system. If we look at this highly complex system with just one statistical method we may ...

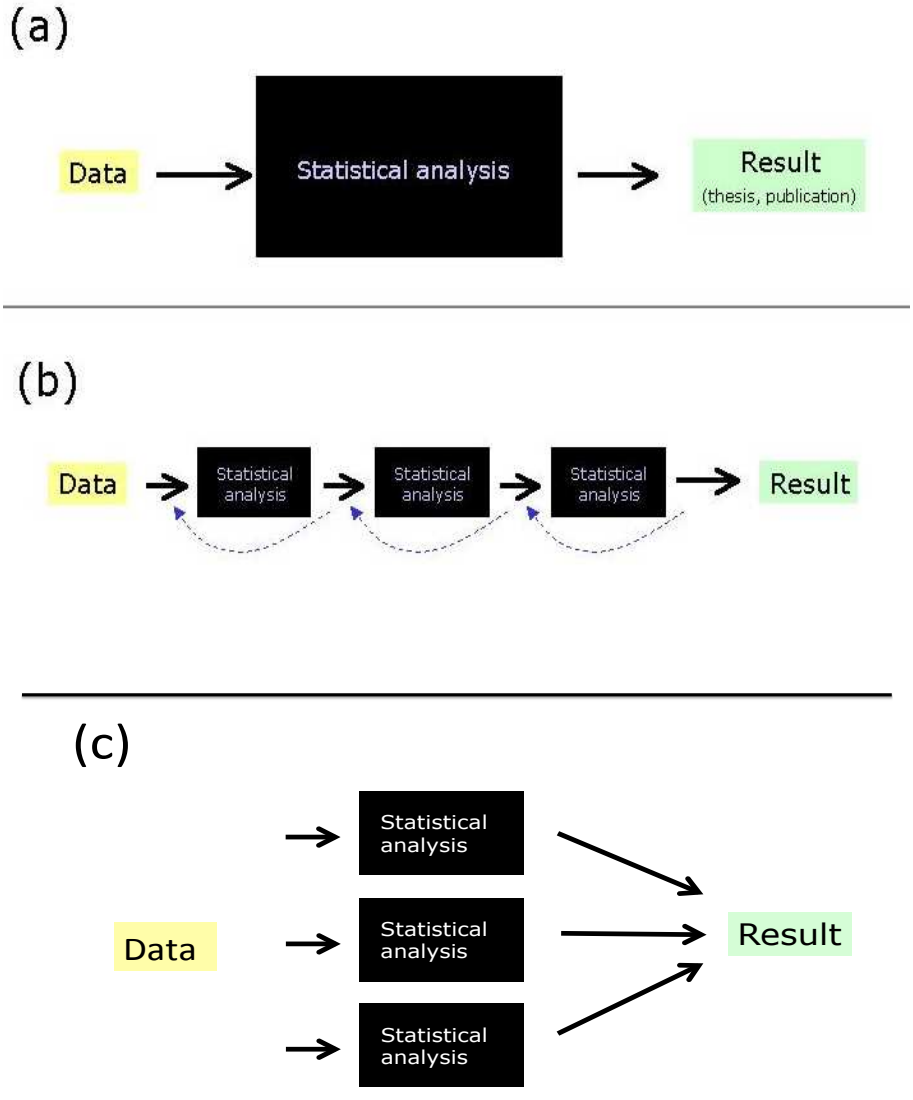


Figure 19.4:

19.9 Parametric and Non-Parametric Methods

19.10 Summary of Tactics in Statistical Analysis

Thinks to consider in statistical inferences / hypothesis testing:

- Be aware of all conditions of the test and estimate the *pdf* for these conditions.
- Failed test must be included in the discussion of the results in order to evaluate the statistical significance. If, for instance, you find A to be "*significantly*" correlated with B, than you should also note that you tested ten other variables with none being "*significantly*" correlated to B, which makes your "*significant*" correlation look much less "*significant*".
- Note that analysis guided by the statistics of the data will tend to result into apparently unusual signals (boost the signal), that are, however, totally ordinary. The apparently unusual signals will most likely not be statistically significant, given the right test *pdf*. Consequently the apparently unusual signal will not be found in a second independent data set.

Use a hierarchy of methods