Statistical Analysis in Climate Research

Hans von Storch
and Francis W. Zwiers

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1999

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Times 10/12pt [DBD]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Storch, H. V. (Hans von), 1949–
Statistical analysis in climate research / Hans von Storch and
Francis W. Zwiers.
p. cm.
Includes index.
ISBN 0 521 45071 3
1. Climatology – Statistical methods. I. Title.
QC981.S735 1998
551.5'072–dc21 98-17416 CIP

ISBN 0 521 45071 3 hardback

Contents

1 Introduction

1.1 The Statistical Description and Understanding of Climate

Climatology was originally a sub-discipline of geography, and was therefore mainly descriptive (see, e.g., Brückner [70], Hann [155], or Hann and Knoch [156]). Description of the climate consisted primarily of estimates of its mean state and estimates of its variability about that state, such as its standard deviations and other simple measures of variability. Much of climatology is still focused on these concerns today. The main purpose of this description is to define 'normals' and 'normal deviations,' which are eventually displayed as maps. These maps are then used for regionalization (in the sense of identifying homogeneous geographical units) and planning. The paradigm of climate research evolved from the purely descriptive approach towards an understanding of the dynamics of climate with the advent of computers and the ability to simulate the climatic state and its variability. Statistics plays an important role in this new paradigm.

The climate is a dynamical system influenced not only by immense external factors, such as solar radiation or the topography of the surface of the solid Earth, but also by seemingly insignificant phenomena, such as butterflies flapping their wings. Its evolution is controlled by more or less well-known physical principles, such as the conservation of angular momentum. If we knew all these factors, and the state of the full climate system (including the atmosphere, the ocean, the land surface, etc.), at a given time in full detail, then there would not be room for statistical uncertainty, nor a need for this book. Indeed, if we repeat a run of a General Circulation Model, which is supposedly a *model* of the real climate system, on the same computer with exactly the same code, operating system, and initial conditions, we obtain a second realization of the simulated climate that is identical to the first simulation.

Of course, there is a 'but.' We do not know all factors that control the trajectory of climate in its enormously large phase space.[1] Thus it is not possible to map the state of the atmosphere, the ocean, and the other components of the climate system in full detail. Also, the models are not deterministic in a practical sense: an insignificant change in a single digit in the model's initial conditions causes the model's trajectory through phase space to diverge quickly from the original trajectory (this is Lorenz's [260] famous discovery, which leads to the concept of chaotic systems).

Therefore, in a strict sense, we have a 'deterministic' system, but we do not have the ability to analyse and describe it with 'deterministic' tools, as in thermodynamics. Instead, we use probabilistic ideas and statistics to describe the 'climate' system.

Four factors ensure that the climate system is amenable to statistical thinking.

- The climate is controlled by innumerable factors. Only a small proportion of these factors can be considered, while the rest are necessarily interpreted as background noise. The details of the generation of this 'noise' are not important, but it is important to understand that this noise is an *internal* source of variation in the climate system (see also the discussion of 'stochastic climate models' in Section 10.4).

- The dynamics of climate are nonlinear. Nonlinear components of the *hydrodynamic* part include important advective terms, such as $u\frac{\partial u}{\partial x}$. The *thermodynamic* part contains various other nonlinear processes, including many that can be represented by step functions (such as condensation).

[1] We use the expression 'phase space' rather casually. It is the space spanned by the state variables x of a system $\frac{dx}{dt} = f(x)$. In the case of the climate system, the state variables consist of the collection of all climatic variables at all geographic locations (latitude, longitude, height/depth). At any given time, the state of the climate system is represented by one point in this space; its development in time is represented by a smooth curve ('trajectory').
This concept deviates from the classical mechanical definition where the phase space is the space of generalized coordinates. Perhaps it would be better to use the term 'state space.'

- The dynamics include linearly unstable processes, such as the baroclinic instability in the midlatitude troposphere.
- The dynamics of climate are dissipative. The hydrodynamic processes transport energy from large spatial scales to small spatial scales, while molecular diffusion takes place at the smallest spatial scales. Energy is dissipated through friction with the solid earth and by means of gravity wave drag at larger spatial scales.[2]

The nonlinearities and the instabilities make the climate system *unpredictable* beyond certain characteristic times. These characteristic time scales are different for different subsystems, such as the ocean, midlatitude troposphere, and tropical troposphere. The nonlinear processes in the system amplify minor disturbances, causing them to evolve irregularly in a way that allows their interpretation as finite-amplitude noise.

In general, the dissipative character of the system guarantees its 'stationarity.' That is, it does not 'run away' from the region of phase space that it currently occupies, an effect that can happen in general nonlinear systems or in linearly unstable systems. The two factors, noise and damping, are the elements required for the interpretation of climate as a stationary stochastic system (see also Section 10.4).

Under what circumstances should the output of climate models be considered stochastic? A major difference between the real climate and any climate model is the size of the phase space. The phase space of a model is much smaller than that of the real climate system because the model's phase space is truncated in both space and time. That is, the background noise, due to unknown factors, is missing. Therefore a model run can be repeated with identical results, provided that the computing environment is unchanged and the same initial conditions are used. To make the climate model output realistic we need to make the model unpredictable. Most Ocean General Circulation Models are strongly dissipative and behave almost linearly. Explicit noise must therefore be added to the system as an explicit forcing term to create statistical variations in the simulated system (see, for instance [276] or [418]). In dynamical atmospheric models (as opposed to energy-balance models) the nonlinearities are strong enough to create their own unpredictability. These models behave in such a way that a repeated run will diverge quickly from the original run even if only minimal changes are introduced into the initial conditions.

1.1.1 The Paradigms of the Chaotic and Stochastic Model of Climate. In the paradigm of the chaotic model of the climate, and particularly the atmosphere, a small difference introduced into the system at some *initial* time causes the system to diverge from the trajectory it would otherwise have travelled. This is the famous *Butterfly Effect*[3] in which infinitesimally small disturbances may provoke large reactions. In terms of climate, however, there is not just *one* small disturbance, but myriads of such disturbances at all times. In the metaphor of the butterfly: there are millions of butterflies that flap their wings all the time. The paradigm of the stochastic climate model is that this omnipresent noise causes the system to vary on all time and space scales, independently of the degree of nonlinearity of the climate's dynamics.

1.2 Some Typical Problems and Concepts

1.2.0 Introduction. The following examples, which we have subjectively chosen as being typical of problems encountered in climate research, illustrate the need for statistical analysis in atmospheric and climatic research. The order of the examples is somewhat random and it is certainly not a must to read all of them; the purpose of this 'potpourri' is to offer a flavour of typical questions, answers, and errors.

1.2.1 The Mean Climate State: Interpretation and Estimation. From the point of view of the climatologist, the most fundamental statistical parameter is the mean state. This seemingly trivial animal in the statistical zoo has considerable complexity in the climatological context.

First, the computed mean is not entirely reliable as an estimate of the climate system's true long-term mean state. The computed mean will contain errors caused by taking observations over a limited observing period, at discrete times and a finite number of locations. It may also be affected by the presence of instrumental, recording, and

[2] The gravity wave drag maintains an exchange of momentum between the solid earth and the atmosphere, which is transported by means of vertically propagating gravity waves. See McFarlane et al. [269] for details.

[3] Inaudil et al. [194] claimed to have identified a Lausanne butterfly that caused a rainfall in Paris.

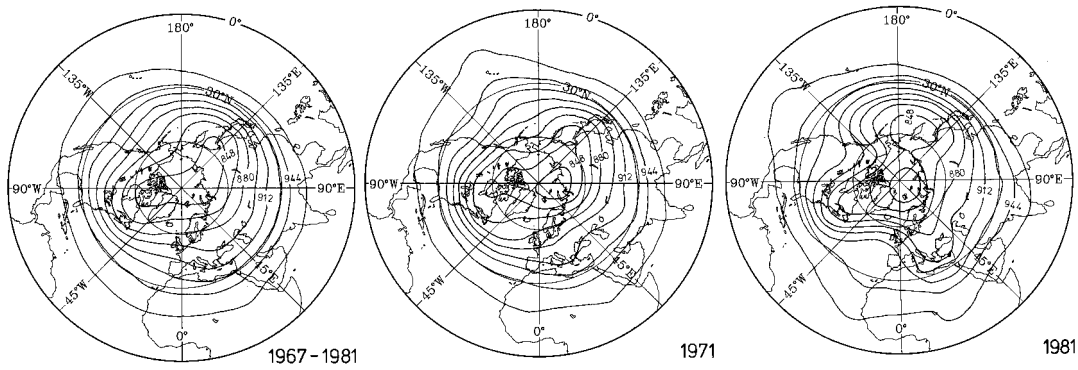

Figure 1.1: *The 300 hPa geopotential height fields in the Northern Hemisphere: the mean 1967–81 January field, the January 1971 field, which is closer to the mean field than most others, and the January 1981 field, which deviates significantly from the mean field. Units: 10 m [117].*

transmission errors. In addition, reliability is not likely to be uniform as a function of location.

Reliability may be compromised if the data has been 'analysed', that is, interpolated to a regular grid using techniques that make assumptions about atmospheric dynamics. The interpolation is performed either *subjectively* by someone who has experience and knowledge of the shape of dynamical structures typically observed in the atmosphere, or it is performed *objectively* using a combination of atmospheric and statistical models. Both kinds of analysis are apt to introduce biases not present in the 'raw' station data, and errors at one location in analysed data will likely be correlated with those at another. (See Daley [98] or Thiébaux and Pedder [362] for comprehensive treatments of objective analysis.)

Second, the mean state is *not* a typical state. To demonstrate this we consider the January Northern Hemisphere 300 hPa geopotential height field[4] (Figure 1.1). The mean January height field, obtained by averaging monthly mean analyses for each January between 1967 and 1981, has contours of equal height which are primarily circular with minor irregularities. Two troughs are situated over the eastern coasts of Siberia and North America. The Siberian trough extends slightly farther south than the North American trough. A secondary trough can be identified over eastern Europe and two minor ridges are located over the northeast Pacific and the east Atlantic.

Some individual January mean fields (e.g., 1971) are similar to the long-term mean field. There are differences in detail, but they share the zonal wavenumber 2 pattern[5] of the mean field. The secondary ridges and troughs have different intensities and longitudinal phases. Other Januaries (e.g., 1981) 300 hPa geopotential height fields are very different from the mean state. They are characterized by a zonal wavenumber 3 pattern rather than a zonal wavenumber 2 pattern.

The long-term mean masks a great deal of interannual variability. For example, the minimum of the long-term mean field is larger than the minima of all but one of the individual January states. Also, the spatial variability of each of the individual monthly means is larger than that of the long-term mean. Thus, the long-term mean field is not a 'typical' field, as it is very unlikely to be observed as an individual monthly mean. In that sense, the long-term mean field is a rare event.

Characterization of the 'typical' January requires more than the long-term mean. Specifically, it is necessary to describe the dominant patterns of spatial variability about the long-term mean and to say something about the range of patterns one is likely to see in a 'typical' January. This can be accomplished to a limited extent through the use of a technique called *Empirical Orthogonal Function analysis* (Chapter 13).

Third, a climatological mean should be understood to be a moving target. Today's climate is different from that which prevailed during the Holocene (6000 years before present) or even during the Little Ice Age a few hundred years ago.

[4]The *geopotential height field* is a parameter that is frequently used to describe the dynamical state of the atmosphere. It is the height of the surface of constant pressure at, e.g., 300 hPa and, being a length, is measured in metres. We will often simply refer to 'height' when we mean 'geopotential height'.

[5]A zonal wavenumber 2 pattern contains two ridges and two troughs in the zonal, or east–west, direction.

We therefore need a clear understanding of our interpretation of the 'true' mean state before interpreting an estimate computed from a set of observations.

To accomplish this, it is necessary to think of the 'January 300 hPa height field' as a *random field*, and we need to determine whether the observed height fields in our 15-year sample are representative of the 'true' mean state we have in mind (presumably that of the 'current' climate). From a statistical perspective, the answer is a conditional 'yes,' provided that:

1 the time series of January mean 300 hPa height fields is stationary (i.e., their statistical properties do not drift with time), and

2 the memory of this time series is short relative to the length of the 15-year sample.

Under these conditions, the mean state is representative of the random sample, in the sense that it lies in the 'centre' of the scatter of the individual points in the state space. As we noted above, however, it is not representative in many other ways.

The characteristics of the 15-year sample may not be representative of the properties of January mean 300 hPa height fields on longer time scales when assumption 1 is not satisfied. The uncertainty of the 15-year mean height field as an estimator of the long-term mean will be almost as great as the interannual variability of the individual January means when assumption 2 is not satisfied. We can have confidence in the 15-year mean as an estimator of the long-term mean January 300 hPa height field when assumptions 1 and 2 hold in the following sense: the *law of large numbers* dictates that a multi-year mean becomes an increasingly better estimator of the long-term mean as the number of years in the sample increases. However, there is still a considerable amount of uncertainty in an estimate based on a 15-year sample.

Statements to the effect that a certain estimate of the mean is 'wrong' or 'right' are often made in discussions of data sets and climatologies. Such an assessment indicates that the speakers do not really understand the art of estimation. An estimate is by definition an approximation, or guess, based on the available data. It is almost certain that the exact value will never be determined. Therefore estimates are never 'wrong' or 'right;' rather, some estimates will be closer to the truth than others on average.

To demonstrate the point, consider the following two procedures for estimating the long-term mean January air pressure in Hamburg (Germany). Two data sets, consisting of 104 observations each, are available. The first data set is taken at one minute intervals, the second is taken at weekly intervals, and a mean is computed from each. Both means are estimates of the long-term mean air pressure in Hamburg, and each tells us something about our parameter.

The reliability of the first estimate is questionable because air pressure varies on time scales considerably longer than the 104 minutes spanned by the data set. Nonetheless, the estimate does contain information useful to someone who has no prior information about the climate of locations near sea level: it indicates that the mean air pressure in Hamburg is neither 2000 mb nor 20 hPa but somewhere near 1000 mb.

The second data set provides us with a much more reliable estimate of long-term mean air pressure because it contains 104 almost independent observations of air pressure spanning two annual cycles. The first estimate is not 'wrong,' but it is not very informative; the second is not 'right,' but it is adequate for many purposes.

1.2.2 Correlation. In the statistical lexicon, the word *correlation* is used to describe a *linear statistical* relationship between two random variables. The phrase 'linear statistical' indicates that the mean of one of the random variables is linearly dependent upon the random component of the other (see Section 8.2). The stronger the linear relationship, the stronger the correlation. A correlation coefficient of $+1$ (-1) indicates a pair of variables that vary together precisely, one variable being related to the other by means of a positive (negative) scaling factor.

While this concept seems to be intuitively simple, it does warrant scrutiny. For example, consider a satellite instrument that makes radiance observations in two different frequency bands. Suppose that these radiometers have been designed in such a way that instrumental error in one channel is independent of that in the other. This means that knowledge of the noise in one channel provides no information about that in the other. However, suppose also that the radiometers drift (go out of calibration) together as they age because both share the same physical environment, share the same power supply and are exposed to the same physical abuse. Reasonable models for the total error as a function of time in the two radiometer

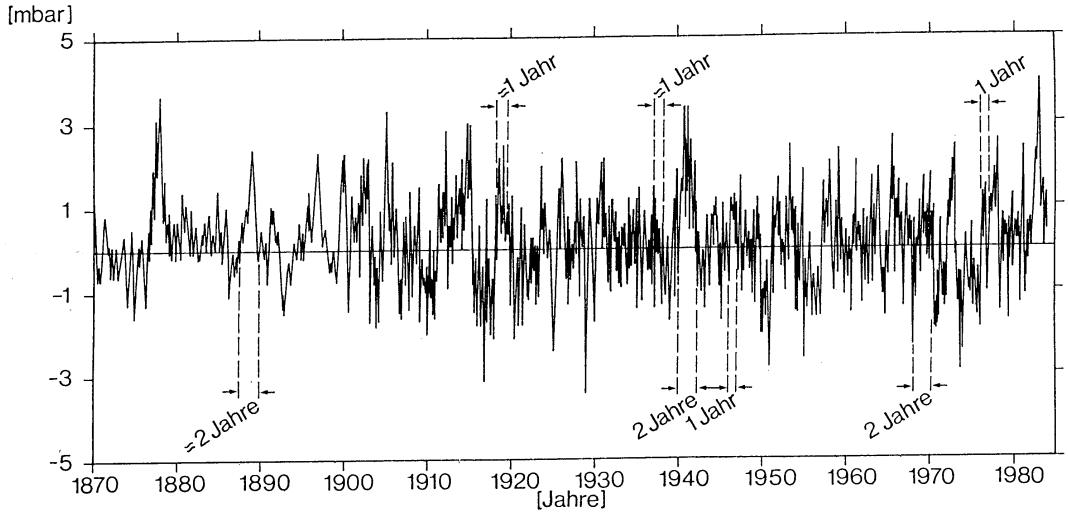

Figure 1.2: *The monthly mean Southern Oscillation Index, computed as the difference between Darwin (Australia) and Papeete (Tahiti) monthly mean sea-level pressure ('Jahr' is German for 'year').*

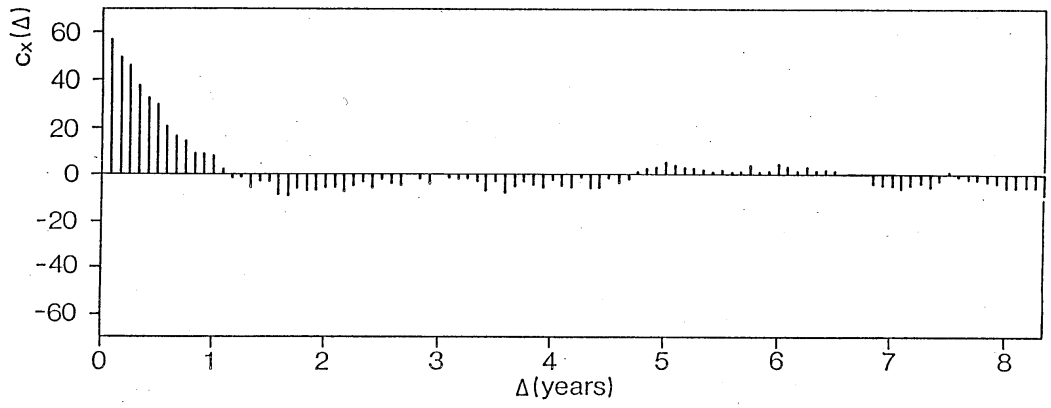

Figure 1.3: *Auto-correlation function of the index shown in Figure 1.2. Units: %.*

channels might be:

$$e_{1t} = \alpha_1(t - t_0) + \epsilon_{1t},$$
$$e_{2t} = \alpha_2(t - t_0) + \epsilon_{2t},$$

where t_0 is the launch time of the satellite and α_1 and α_2 are fixed constants describing the rates of drift of the two radiometers. The instrumental errors, ϵ_{1t} and ϵ_{2t}, are statistically independent of each other, implying that the correlation between the two, $\rho(\epsilon_{1t}, \epsilon_{2t})$, is zero. Consequently the total errors, e_{1t} and e_{2t}, are also statistically independent even though they share a common systematic component. However, simple estimates of correlation between e_{1t} and e_{2t} that do not account for the deterministic drift will suggest that these two quantities are correlated.

Correlations manifest themselves in several different ways in observed and simulated climates. Several adjectives are used to describe correlations depending upon whether they describe relationships in time (serial correlation, lagged correlation), space (spatial correlation, teleconnection), or between different climate variables (cross-correlation).

A good example of *serial correlation* is the monthly Southern Oscillation Index (SOI),[6] which

[6]The Southern Oscillation is the major mode of natural climate variability on the interannual time scale. It is frequently used as an example in this book.
It has been known since the end of the last century (Hildebrandson [177]; Walker, 1909–21) that sea-level pressure (SLP) in the Indonesian region is negatively correlated with that over the southeast tropical Pacific. A positive SLP anomaly

is defined as the anomalous monthly mean pressure difference between Darwin (Australia) and Papeete (Tahiti) (Figure 1.2).

The time series is basically stationary, although variability during the first 30 years seems to be somewhat weaker than that of late. Despite the noisy nature of the time series, there is a distinct tendency for the SOI to remain positive or negative for extended periods, some of which are indicated in Figure 1.2. This persistence in the sign of the index reflects the serial correlation of the SOI.

A quantitative measure of the serial correlation is the *auto-correlation function*, $\rho_{SOI}(t, t + \Delta)$, shown in Figure 1.3, which measures the similarity of the SOI at any time difference Δ. The auto-correlation is greater than 0.2 for lags up to about six months and varies smoothly around zero with typical magnitudes between 0.05 and 0.1 for lags greater than about a year. This tendency of *estimated* auto-correlation functions not to converge to zero at large lags, even though the real auto-correlation is zero at long lags, is a natural consequence of the uncertainty due to finite samples (see Section 11.1).

A good example of a *cross-correlation* is the relationship that exists between the SOI and various alternative indices of the Southern Oscillation [426]. The characteristic low-frequency variations in Figure 1.2 are also present in area-averaged Central Pacific sea-surface temperature (Figure 1.4).[7] The correlation between the two time series displayed in Figure 1.4 is 0.67.

Pattern analysis techniques, such as Empirical Orthogonal Function analysis (Chapter 13), Canonical Correlation Analysis (Chapter 14) and Principal Oscillation Patterns (Chapter 15), rely upon the assumption that the fields under study are *spatially correlated*. The Southern Oscillation Index (Figure 1.2) is a manifestation of the negative correlation between surface pressure at Papeete and that at Darwin. Variables such as pressure, height, wind, temperature, and specific humidity vary smoothly in the free atmosphere and consequently exhibit strong spatial interdependence. This correlation is present in each weather map (Figure 1.5, left). Indeed, without this feature, routine weather forecasts would be all but impossible given the sparseness of the global observing network as it exists even today. Variables derived from moisture, such as cloud cover, rainfall and snow amounts, and variables associated with land surface processes tend to have much smaller spatial scales (Figure 1.5, right), and also tend not to have normal distributions (Sections 3.1 and 3.2). While mean sea-level pressure (Figure 1.5, left) will be more or less constant on spatial scales of tens of kilometres, we may often travel in and out of localized rain showers in just a few kilometres. This dichotomy is illustrated in Figure 1.5, where we see a cold front over Ontario (Canada). The left panel, which displays mean sea-level pressure, shows the front as a smooth curve. The right panel displays a radar image of precipitation occurring in southern Ontario as the front passes through the region.

1.2.3 Stationarity, Cyclo-stationarity, and Non-stationarity.

An important concept in statistical analysis is *stationarity*. A random variable, or a random process, is said to be stationary if all of its statistical parameters are independent of time. Most statistical techniques assume that the observed process is stationary.

However, most climate parameters that are sampled more frequently than one per year are not stationary but *cyclo-stationary*, simply because of the seasonal forcing of the climate system. Long-term averages of monthly mean sea-level pressure exhibit a marked annual cycle, which is almost sinusoidal (with one maximum and one minimum) in most locations. However, there are locations (Figure 1.6) where the annual cycle is dominated by a *semiannual* variation (with two maxima and minima). In most applications the mean annual cycle is simply subtracted from the data before the remaining *anomalies* are analysed. The process is *cyclo-stationary in the mean* if it is stationary after the annual cycle has been removed.

Other statistical parameters (e.g., the percentiles of rainfall) may also exhibit cyclo-stationary behaviour. Figure 1.7 shows the annual cycles

(i.e., a deviation from the long-term mean) over, say, Darwin (Northern Australia) tends to be associated with a negative SLP anomaly over Papeete (Tahiti). This seesaw is called the Southern Oscillation (SO). The SO is associated with large-scale and persistent anomalies of sea-surface temperature in the central and eastern tropical Pacific (El Niño and La Niña). Hence the phenomenon is often referred to as the 'El Niño/Southern Oscillation' (ENSO). Large zonal displacements of the centres of precipitation are also associated with ENSO. They reflect anomalies in the location and intensity of the meridionally (i.e., north–south) oriented Hadley cell and of the zonally oriented Walker cell.

The state of the Southern Oscillation may be monitored with the monthly SLP difference between observations taken at surface stations in Darwin, Australia and Papeete, Tahiti. It has become common practice to call this difference the Southern Oscillation Index (SOI) although there are also many other ways to define equivalent indices [426].

[7] Other definitions, such as West Pacific rainfall, sea-level pressure at Darwin alone or the surface zonal wind in the central Pacific, also yield indices that are highly correlated with the usual SOI. See Wright [427].

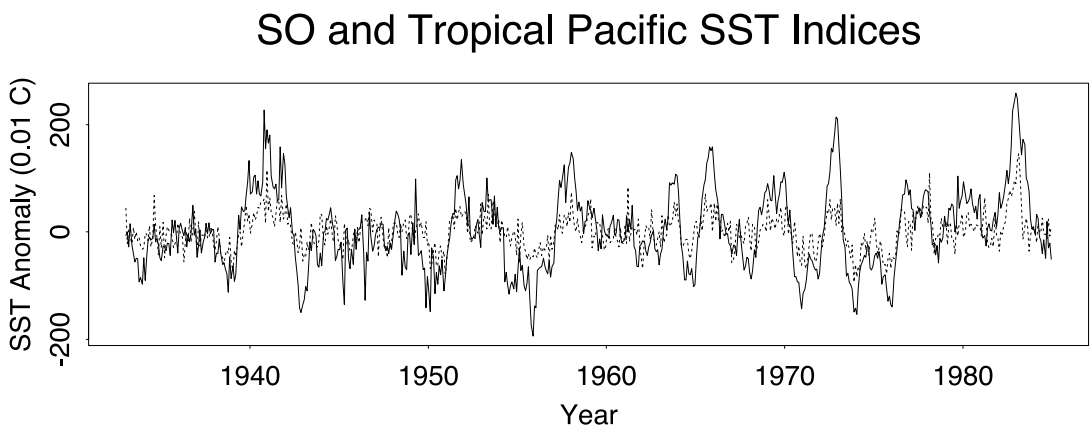

Figure 1.4: *The conventional Southern Oscillation Index (SOI = pressure difference between Darwin and Tahiti; dashed curve) and a sea-surface temperature (SST) index of the Southern Oscillation (solid curve) plotted as a function of time. The conventional SOI has been doubled in this figure.*

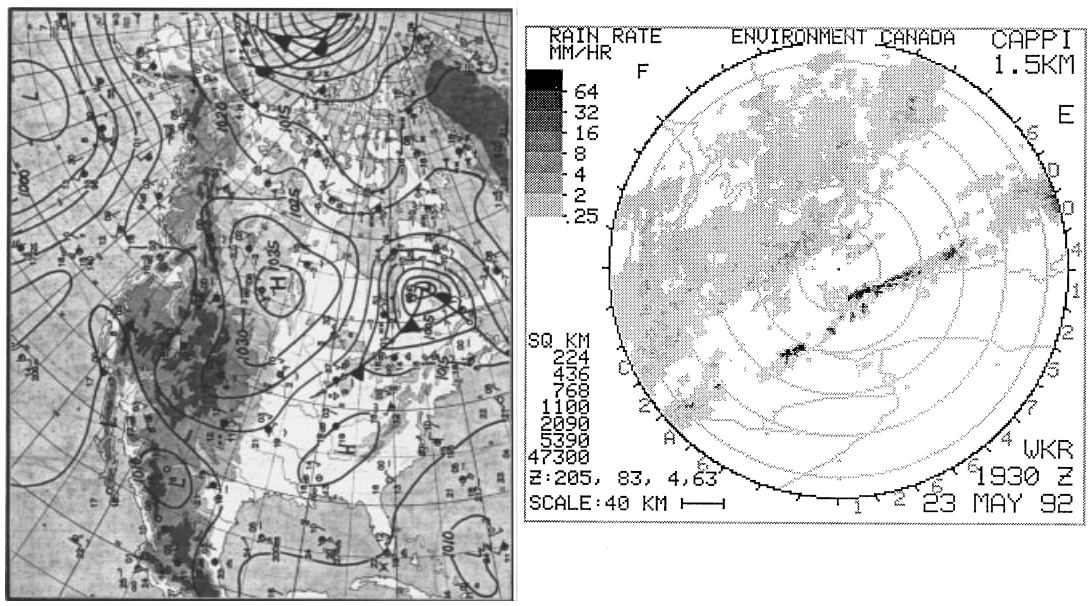

Figure 1.5: *State of the atmosphere over North America on 23 May 1992.*
Left: Analysis of the sea-level pressure field (12:00 UTC (Universal Time Coordinated); from Europäisher Wetterbericht 17, Band 144; with permission of the Deutsher Wetterdienst).
Right: Weather radar image, showing rainfall rates, for southern Ontario (19:30 local time; courtesy Paul Joe, AES Canada [94].)
Note that the radar image and the weather map refer to different times, namely 12:00 UTC on 23 May and 00:30 UTC on 24 May.

of the 70th, 80th, and 90th percentiles[8] of 24-hour rainfall amounts for each calendar month at Vancouver (British Columbia) and Sable Island (off the coast of Nova Scotia) [450].

The Southern Oscillation Index is not strictly stationary. Wright [427] showed that the linear serial correlation of the SOI depends upon the time

[8]Or 'quantiles,' that is, thresholds selected so that 70%, 80%, or 90% of all 24-hour rainfall amounts are less than the respective threshold [2.6.4].

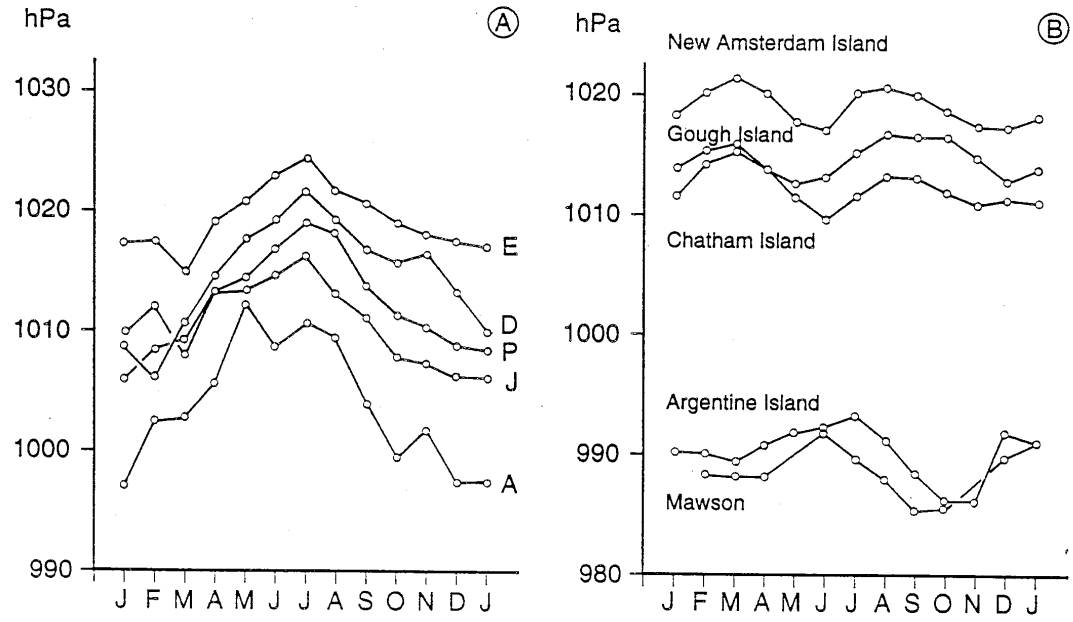

Figure 1.6: *Annual cycle of sea-level pressure at extratropical locations.*
a) Northern Hemisphere Ocean Weather Stations: A = 62° N, 33° W; D = 44° N, 41° W; E = 35° N, 48° W; J = 52° N, 25° W; P = 50° N, 145° W.
b) Southern Hemisphere.

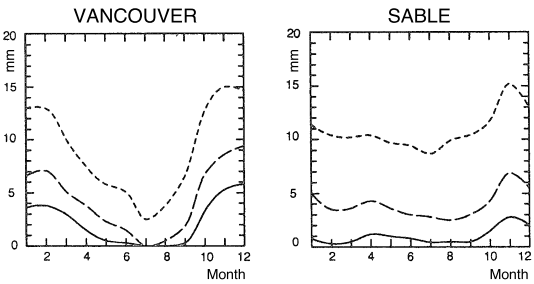

Figure 1.7: *Monthly 90th, 80th, and 70th percentiles (from top to bottom) of 24-hour rainfall amounts at Vancouver and Sable Island [450].*

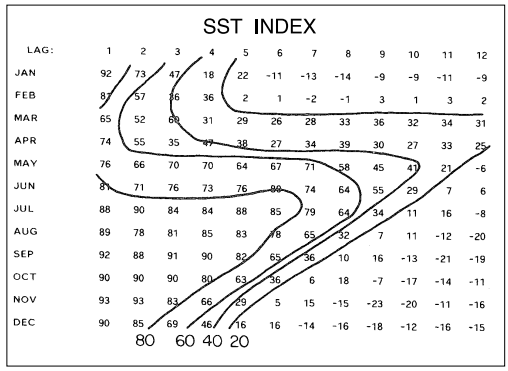

Figure 1.8: *Seasonal dependence of the lag correlations of the SST index of the Southern Oscillation. The correlations are given in hundreds so that isolines represent lag correlations of 0.8, 0.6, 0.4, and 0.2. The row labelled 'Jan' lists correlations between January values of the index and the index observed later 'lag' months [427].*

of the year. The serial correlation is plotted as a function of time of year and lag in Figure 1.8. Correlations between values of the SOI in May and values in subsequent months decay slowly with increasing lag, while similar correlations with values in April decay quickly. Because of this behaviour, Wright defined an ENSO year that begins in May and ends in April.

Regular observations taken over extended periods at a certain station sometimes exhibit changes in their statistical properties. These might be abrupt or gradual (such as changes that might occur when the exposure of a rain gauge changes slowly over time, as a consequence of the growth of vegetation or changes in local land use). Abrupt changes in the observational record may take place if the instrument (or the observer) changes, the site is moved,[9] or recording practices are changed. Such non-natural or artificial changes are

[9]Karl et al. [213] describe a case in which a precipitation gauge recorded significantly different values after being raised one metre from its original position.

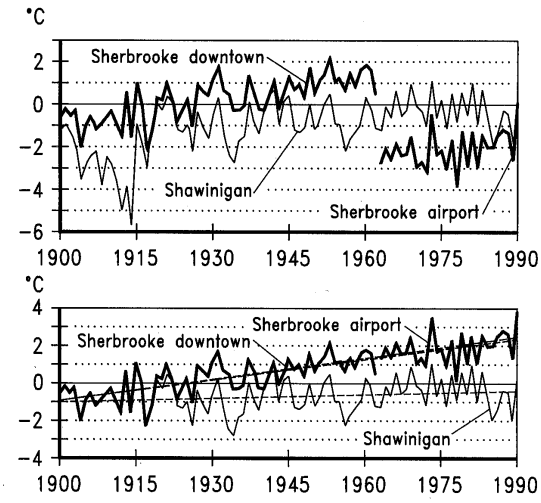

Figure 1.9: *Annual mean daily minimum temperature time series at two neighbouring sites in Quebec. Sherbrooke has experienced considerable urbanization since the beginning of the century whereas Shawinigan has maintained more of its rural character.*
Top: The raw records. The abrupt drop of several degrees in the Sherbrooke series in 1963 reflects the move of the instrument from downtown Sherbrooke to its suburban airport. The reason for the downward dip before 1915 in the Shawinigan record is unknown.
Bottom: Corrected time series for Sherbrooke and Shawinigan. The Sherbrooke data from 1963 onward are increased by $3.2\,^{\circ}C$*. The straight lines are trend lines fitted to the corrected Sherbrooke data and the 1915–90 Shawinigan record.*
Courtesy L. Vincent, AES Canada.

called *inhomogeneities*. An example is contained in the temperature records of Sherbrooke and Shawinigan (Quebec) shown in the upper panel of Figure 1.9. The Sherbrooke observing site was moved from a downtown location to a suburban airport in 1963—and the recorded temperature abruptly dropped by more than $3\,^{\circ}$C. The Shawinigan record may also be contaminated by observational errors made before 1915.

Geophysical time series often exhibit a trend. Such trends can originate from various sources. One source is urbanization, that is, the increasing density and height of buildings around an observation location and the corresponding changes in the properties of the land surface. The temperature at Sherbrooke, a location heavily affected by development, exhibits a marked upward trend after correction for the systematic change in 1963 (Figure 1.9, bottom). This temperature trend is much weaker for the neighbouring Shawinigan, perhaps due to a weaker urbanization effect at that site or natural variations of the climate system. Both temperature trends at Sherbrooke and Shawinigan are real, not observational artifacts. The strong trend at Sherbrooke must not be mistaken for an indication of *global warming*.

Trends in the large-scale state of the climate system may reflect systematic forcing changes of the climate system (such as variations in the Earth's orbit, or increased CO_2 concentration in the atmosphere) or low-frequency internally generated variability of the climate system. The latter may be deceptive because low-frequency variability, on short time series, may be mistakenly interpreted as trends. However, if the length of such time series is increased, a metamorphosis of the former 'trend' takes place and it becomes apparent that the trend is a part of the natural variation of the system.[10]

1.2.4 Quality of Forecasts. The *Old Farmer's Almanac* publishes regular outlooks for the climate for the coming year. The method used to prepare these outlooks is kept secret, and scientists question the existence of skill in the predictions. To determine whether these skeptics are right or wrong, measures of the skill of the forecasting scheme are needed. These *skill scores* can be used to compare forecasting schemes objectively.

The Almanac makes *categorical* forecasts of future temperature and precipitation amount in two categories, 'above' or 'below' normal. A suitable skill score in this case is the number of correct forecasts. Trivial forecasting schemes such as persistence (no change), climatology, or pure chance can be used as reference forecasts if no other forecasting scheme is available. Once we have counted the number of correct forecasts made with both the tested and the reference schemes, we can estimate the improvement (or degradation) of forecast skill by computing the difference in the counts. Relatively simple probabilistic methods can be used to make a judgement about the

[10] This is an example of the importance of time scales in climate research, an illustration that our interpretation of a given process depends on the time scales considered. A short-term trend may be just another swing in a slowly varying system. An example is the Madden-and-Julian Oscillation (MJO, [264]), which is the strongest intra-seasonal mode in the tropical troposphere. It consists of a wavenumber 1 pattern that travels eastward round the globe. The MJO has a mean period of 45 days and has significant memory on time scales of weeks; on time scales of months and years, however, the MJO has no temporal correlation.

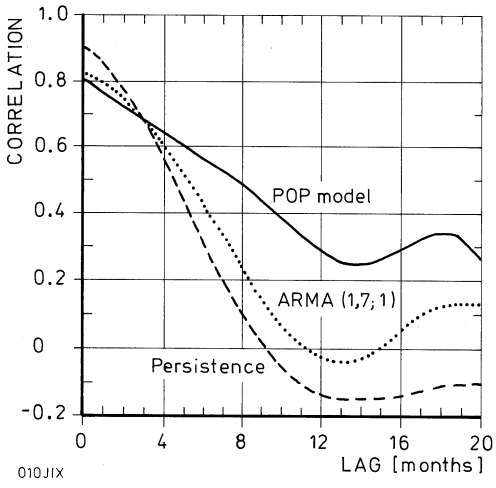

Figure 1.10: *Correlation skill scores for three forecasts of the low-frequency variations within the Southern Oscillation Index (Figure 1.2). A score of 1 indicates a perfect forecast, while a zero indicates a forecast unrelated to the predictand [432].*

significance of the change. We will return to the *Old Farmer's Almanac* in Section 18.1.

Now consider another forecasting scheme in which *quantitative* rather than categorical statements are made. For example, a forecast might consist of a statement such as: '*the SOI will be x standard deviations above normal next winter.*' One way to evaluate such forecasts is to use a measure called the *correlation skill score* ρ (Chapter 18). A score of $\rho = 1$ corresponds with a perfect forecasting scheme in the sense that forecast changes exactly mirror SOI changes even though the dynamic range of the forecast may be different from that of the SOI. In other words, the correlation skill score is one when there is an exact linear relationship between forecasts and reality. Forecasts that are (linearly) unrelated to the predictand yield zero correlation.

The correlation skill score for several methods of forecasting the SOI are displayed in Figure 1.10. Specifically, persistence forecasts (Chapter 18), POP forecasts (Chapter 15), and forecasts made with a univariate linear time series model (Chapters 11 and 12). Forecasts based on persistence and the univariate time series model are superior at one and two month lead times. The POP forecast becomes more skilful beyond that time scale.

Regretfully, forecasting schemes generally do not have the same skill under all circumstances. The skill often exhibits a marked annual cycle (e.g., skill may be high during the dry season, and low during the wet season). The skilfulness of a forecast also often depends on the low-frequency state of the atmospheric flow (e.g., blocking or westerly regime). Thus, in most forecasting problems there are physical considerations (state dependence and the memory of the system) that must be accounted for when using statistical tools to analyse forecast skill. This is done either by conducting a statistical analysis of skill that incorporates the effects of state dependence and serial correlation, or by using physical intuition to temper the precise interpretation of a simpler analysis that compromises the assumptions of stationarity and non-correlation.

There are various pitfalls in the art of forecast evaluation. An excellent overview is given by Livezey [255], who presents various examples in which forecast skill is overestimated. Chapter 18 is devoted to the art of forecast evaluation.

1.2.5 Characteristic Times and Characteristic Spatial Patterns. What are the temporal characteristics of the Southern Oscillation Index illustrated in Figure 1.2? Visual inspection suggests that the time series is dominated by at least two time scales: a high frequency mode that describes month-to-month variations, and a low-frequency mode associated with year-to-year variations. How can one objectively quantify these characteristic times and the amount of variance attributed to these time scales? The appropriate tool is referred to as time series analysis (Chapters 10 and 11).

Indices, such as the SOI, are commonly used in climate research to monitor the temporal development of a process. They can be thought of as filters that extract physical signals from a multivariate environment. In this environment the signal is masked by both spatial and temporal variability unrelated to the signal, that is, by spatial and temporal noise.

The conventional approach used to identify indices is largely subjective. The characteristic patterns of variation of the process are identified and associated with regions or points. Corresponding areal averages or point values are then used to indicate the state of the process.

Another approach is to extract characteristic patterns from the data by means of analytical techniques, and subsequently use the coefficients of these patterns as indices. The advantages of this approach are that it is based on an objective algorithm and that it yields the characteristic patterns explicitly. *Eigentechniques* such as Empirical Orthogonal Function (EOF)

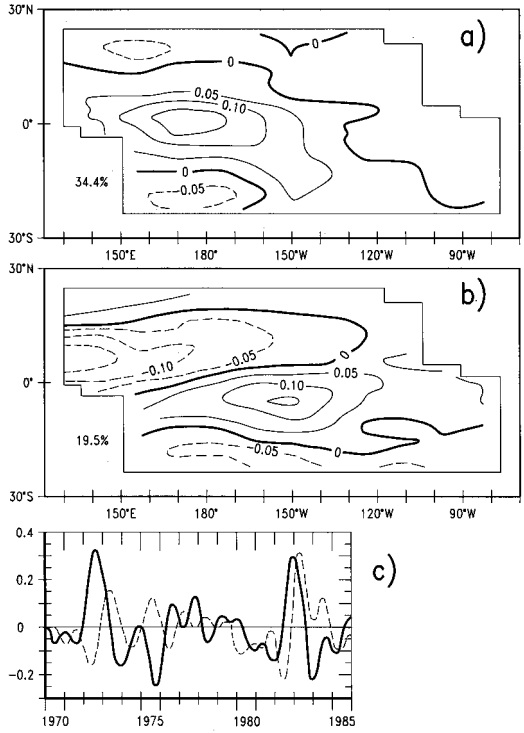

Figure 1.11: *Empirical Orthogonal Functions (EOFs; Chapter 13) of monthly mean wind stress over the tropical Pacific [394].*
a,b) The first two EOFs. The two patterns are spatially orthogonal.
c) Low-frequency filtered coefficient time series of the two EOFs shown in a,b). The solid curve corresponds to the first EOF, which is displayed in panel a). The two curves are orthogonal.

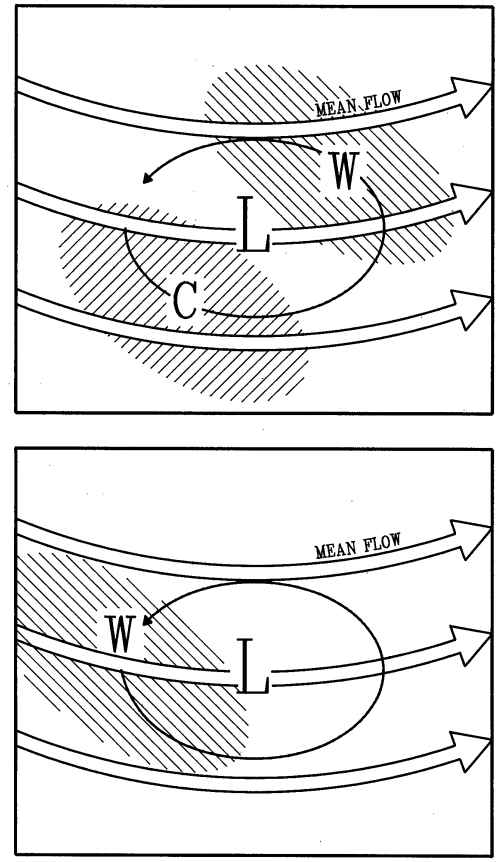

Figure 1.12: *A schematic representation of the spatial distributions of simultaneous SST and SLP anomalies at Northern Hemisphere midlatitudes in winter, when the SLP anomaly induces the SST anomaly (top), and when the SST anomaly excites the SLP anomaly (bottom).*
The large arrows represent the mean atmospheric flow. The 'L' is an atmospheric low-pressure system connected with geostrophic flow indicated by the circular arrow. The hatching represents warm (W) and cool (C) SST anomalies [438].

analysis and Principal Oscillation Pattern (POP) analysis are tools that can be used to define patterns and indices objectively (Chapters 13 and 15).

An example is the EOF analysis of monthly mean wind stress over the tropical Pacific [394]. The first two EOFs, shown in Figure 1.11a and Figure 1.11b, are primarily confined to the equator. The two fields are (by construction) orthogonal to each other. Figure 1.11c shows the time coefficients of the two fields. An analysis of the coefficient time series, using the techniques of cross-spectral analysis (Section 11.4), shows that they vary coherently on a time scale $T \approx$ 2 to 3 years. One curve leads the other by a time lag of approximately $T/4$ years. The temporal lag-relationship of the time coefficients together with the spatial quadrature leads to the interpretation that the two patterns and their time coefficients describe an eastward propagating signal that, in fact, may be associated with the Southern Oscillation.

1.2.6 Pairs of Characteristic Patterns. Almost all climate components are interrelated. When one component exhibits anomalous conditions, there will likely be characteristic anomalies in other components at the same time. The relative shapes of the patterns in related climate components are often indicative of the processes that dominate the coupling of the components.

To illustrate this idea we consider large-scale air–sea interactions on seasonal time scales at midlatitudes in winter [438] [312]. Figure 1.12

illustrates the two mechanisms that might be involved in air–sea interactions in the North Atlantic. The lower panel illustrates how a sea-surface temperature (SST) anomaly pattern might induce a simultaneous sea-level pressure (SLP) anomaly pattern. The argument is linear so we may assume that the SST anomaly is positive. This positive SST anomaly enhances the sensible and latent heat fluxes into the atmosphere above and downstream of the SST anomaly. Thus SLP is reduced in that area and anomalous cyclonic flow is induced.

The upper panel of Figure 1.12 illustrates how a SLP anomaly might induce an anomalous SST pattern. The anomalous SLP distribution alters the wind stress across the region by creating stronger zonal winds in the southwest part of the anomalous cyclonic circulation and weaker zonal winds in the northeast sector. This configuration induces anomalous mixing of the ocean's mixed layer and anomalous air–sea fluxes of sensible and latent heat (cf. [3.2.3]). Stronger winds intensify mixing and enhance the upward heat flux whereas weaker winds correspond to reduced mixing and weaker vertical fluxes. The result is anomalous cooling of the sea surface in the southwest sector and anomalous heating in the northeast sector of the cyclonic circulation.

One strategy for finding out which of the two proposed mechanisms dominates air–sea interaction is to identify the dominant patterns in SST and SLP that tend to occur simultaneously. This can be accomplished by performing a *Canonical Correlation Analysis* (CCA, Chapter 14). In the CCA two vector variables $\vec{\mathbf{X}}$ and $\vec{\mathbf{Y}}$ are considered, and sets of orthogonal patterns $\vec{p}\,^{i}_{X}$ and $\vec{p}\,^{i}_{Y}$ are constructed so that the expansion coefficients α^{x}_{i} and α^{y}_{j} in $\vec{\mathbf{X}} = \sum_i \alpha^{x}_{i}\,\vec{p}\,^{i}_{X}$ and $\vec{\mathbf{Y}} = \sum_j \alpha^{y}_{j}\,\vec{p}\,^{j}_{Y}$ are optimally correlated for $i = j$ or uncorrelated for $i \neq j$.

Zorita, Kharin, and von Storch [438] applied CCA to winter (DJF) mean anomalies of North Atlantic SST and SLP and found two pairs of CCA patterns $\vec{p}\,^{i}_{SST}$ and $\vec{p}\,^{j}_{SLP}$ that were associated with physically significant correlations. The pair of patterns with the largest correlation (0.56) is shown in Figure 1.13. The SLP pattern represents 21% of the total DJF SLP variance whereas the SST pattern explains 19% of the total SST variance.[11] Clearly the two patterns support the hypothesis that the anomalous atmospheric circulation is responsible for the generation of SST anomalies off the North American coast. Peng and Fyfe [312] refer to this as the 'atmosphere driving the ocean' mode. See also Luksch [261].

Canonical Correlation Analysis is explained in detail in Chapter 14 and we return to this example in [14.3.1–2].

1.2.7 Atmospheric General Circulation Model Experimentation: Evaluation of Paired Sensitivity Experiments and Verification of Control Simulation.

Atmospheric General Circulation Models (AGCMs) are powerful tools used to simulate the dynamics of the atmospheric circulation. There are two main applications of these GCMs, one being the simulation of the present, past (e.g., paleoclimatic conditions), or future (e.g., climate change) statistics of the atmospheric circulation. The other involves the study of the simulated climate's sensitivity to the effect of different boundary conditions (e.g., sea-surface temperature) or parameterizations of sub-grid scale processes (e.g., planetary boundary layer).[12]

In both modes of operation two sets of statistics are compared. In the first, the statistics of the simulated climate are compared with those of the observed climate, or sometimes with those of another simulated climate. In the second mode of experimentation, the statistics obtained in the run with anomalous conditions are compared with those from the run with the *control* conditions. The simulated atmospheric circulation is turbulent as is that of the real atmosphere (see Section 1.1). Therefore the true signal (excited by the prescribed change in boundary conditions, parameterization, etc.) or the true model error is masked by random variations.

Even when the modifications in the experimental run have no effect on the simulated climate, the difference field will be nonzero and will show structure reflecting the random variations in the control and experimental runs. Similarly, the mean difference field between an observed distribution and its simulated counterpart will exhibit, possibly large scale, features, even if the model is perfect.

[11] The proportion of variance represented by the patterns is unrelated to the correlation.

[12] Sub-grid scale processes take place on spatial scales too small to be resolved by a climate model. Regardless of the resolution of the climate model, there are unresolved processes at smaller scales. Despite the small scale of these processes, they influence the large-scale evolution of the climate system because of the nonlinear character of the climate system. Climate modellers therefore attempt to specify the 'net effect' of such processes as a transfer function of the large-scale state itself. This effect is a forcing term for the resolved scales, and is usually expressed as an expected value which is conditional upon the large-scale state. The transfer function is called a 'parameterization.'

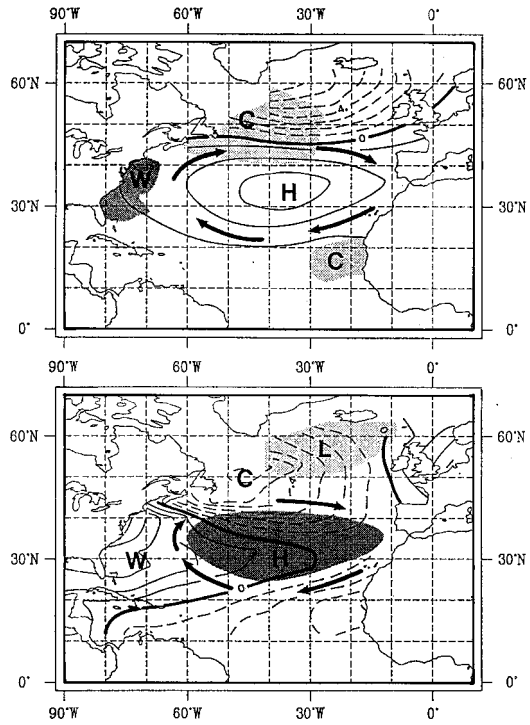

Figure 1.13: *The dominant pair of CCA patterns that describe the connection between simultaneous winter (DJF) mean anomalies of sea-level pressure (SLP, top) and sea-surface temperature (SST, bottom) in the North Atlantic. The largest features of the SLP field are indicated by shading in the SST map, and vice versa. See also [14.3.1]. From Zorita et al. [438].*

Therefore, it is necessary to apply statistical techniques to distinguish between the deterministic signal (or model error) and the internal noise.

Appropriate methodologies designed to diagnose the presence of a signal include the use of interval estimation methods (Section 5.4) or hypothesis testing methods (Chapter 6). Interval estimation methods use statistical models to produce a range of signal estimates consistent with the realizations of control and experimental mean fields obtained from the simulation. Hypothesis testing methods use statistical models to determine whether information in the realizations is consistent with the null hypothesis that the difference fields, such as in Figures 1.14 and 1.15, do not contain a deterministic signal and thus reflect only the effects of random variation.

We illustrate the problem with two examples: an experiment in which there is no significant signal, and another in which modifications to the model result in a strong change in the atmospheric flow.

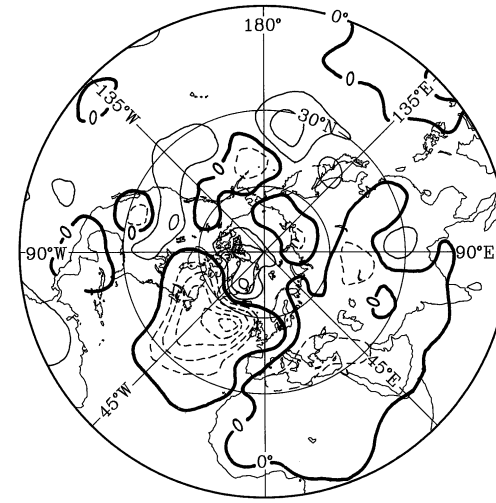

Figure 1.14: *The mean SLP difference field between control and experimental atmospheric GCM runs. Evaporation over the Iberian Peninsula was artificially suppressed in the experimental run. The signal is not statistically significant [402].*

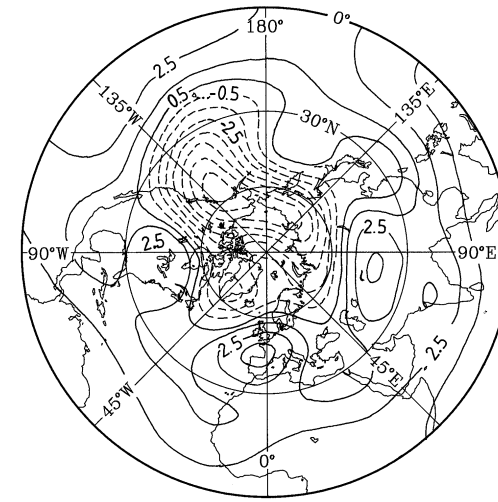

Figure 1.15: *The mean 500 hPa height difference field between a control run and an experimental run in which a positive (El Niño) SST anomaly was imposed in the equatorial Central and Eastern Pacific. The signal is statistically significant. See also Figures 9.1 and 9.2 [393].*

In the first case, the surface properties of the Iberian peninsula were modified so as to turn it into a desert in the experimental climate. That is, evaporation at the grid points representing the Iberian peninsula was arbitrarily set to zero. The response, in terms of January Northern Hemisphere sea-level pressure, is shown in Figure 1.14 [402]. The statistical analysis revealed

that the signal, which appears to be of very large scale, is mainly due to noise and is not statistically significant.

In the second case, anomalously warm sea-surface temperatures were prescribed in the tropical Pacific, in order to simulate the effect of the 1982/83 El Niño event on the atmosphere. The resulting anomalous mean January 500 hPa height field is shown in Figure 1.15. In this case the signal is statistically distinguishable from the background noise.

Before using statistical tests, we must account for several methodical considerations (see Chapter 6). Straightforward statistical assessments that compare the mean states of two simulated climates generally use simple statistical tests that are performed locally at grid points. More complex *field tests*, often called *field significance tests* in the climate literature, are used less frequently.

Grid point tests, while popular because of their simplicity, may have interpretation problems. The result of a set of statistical tests, one conducted at each grid point, is a field of decisions denoting where differences are, and are not, *statistically significant*. However, statistical tests cannot be conducted with absolute certainty. Rather, they are conducted in such a way that there is an *a priori* specified risk $1-\tilde{p}$ of rejecting the null hypothesis: 'no difference' when it is true.[13]

The specified risk $(1 - \tilde{p}) \times 100\%$ is often referred to as the *significance level* of the test.[14]

A consequence of setting the risk of false rejection to $1 - \tilde{p}$, $0 < \tilde{p} < 1$, is that we can expect approximately $(1 - \tilde{p}) \times 100\%$ of the decisions to be *reject* decisions when the null hypothesis is valid. However, many fields of interest in climate experiments exhibit substantial spatial correlation (e.g., smooth fields such as the geopotential heights displayed in Figure 1.1).

The spatial coherence of these fields has two consequences for hypothesis testing at grid points. The first is that the proportion of the field covered by reject decisions becomes highly variable from one realization of the climate experiment to the next. In some problems a rejection rate of 20% may still be globally consistent with the null hypothesis at the 5% significance level. The second is that the spatial coherence of the studied fields also leads to fields of decisions that are spatially coherent: if the difference between two mean 500 hPa height fields is large at a particular point, it is also likely to be large at neighbouring points because of the spatial continuity of 500 hPa height. A decision made at one location is generally not statistically independent of decisions made at other locations. This makes regions of significant change difficult to identify. Methods that can be used to assess the field significance of a field of reject/retain decisions are discussed in Section 6.8. Local, or *univariate*, significance tests are discussed in Sections 6.6 and 6.7.

Another approach to the comparison of observed and simulated mean fields involves the use of classical *multivariate statistical tests* (Sections 6.6 and 6.7). The word *multivariate* is used somewhat differently in the statistical lexicon than it is in climatology: it describes tests and other inference procedures that operate on vector objects, such as the difference between two mean fields, rather than scalar objects, such as a difference of means at a grid point. Thus a multivariate test is a field significance test; it is used to make a single inference about a field of differences between the observed and simulated climate.

Classical multivariate inference methods can not generally be applied directly to difference of means or variance problems in climatology. These methods are usually unable to cope with fields under study, such as seasonal geopotential means, that are generally 'observed' at numbers of grid points one to three orders of magnitude greater than the number of realizations available.[15]

[13] The standard, rather mundane statistical nomenclature for this kind of error is *Type I* error; failure to reject the null hypothesis when it is false is termed a *Type II* error. Specifying a smaller risk reduces the chance of making a Type I error but also reduces the sensitivity of the test and hence increases the likelihood of a Type II error. More or less standard practice is to set the risk of a Type I error to $(1 - \tilde{p}) \times 100\% = 5\%$ in tests of the mean and to $(1 - \tilde{p}) \times 100\% = 10\%$ in tests of variability. A higher level of risk is usually felt to be acceptable in variance tests because they are generally less powerful than tests concerning the mean state. The reasons for specifying the risk in the form $1-\tilde{p}$, where $\tilde{p}$ is a large probability near 1, will become apparent later.

[14] There is some ambiguity in the climate literature about how to specify a 'significance level.' Many climatologists use the expression 'significant at the 95% level,' although standard statistical convention is to use the expression 'significant at the 5% level.' With the latter convention, which we use throughout this book, rejection at the 1% significance level indicates the presence of stronger evidence against the null hypothesis than rejection at the 10% significance level.

[15] A typical climate model validation problem involves the comparison of simulated monthly mean fields obtained from a 5–100 year simulation, with corresponding observed mean fields from a 20–50 year climatology. Such a problem therefore uses a combined total of $n = 25$ to 150 realizations of mean January 500 hPa height, for example. On the other hand, the horizontal resolution of typical present day climate models is such that these mean fields are represented on global grids with $m = 2000$ to 8000 points. Except on relatively small regional scales, the dimension of (or number of points in) the difference field is greater than the combined number of realizations from the simulated and observed climates.

One solution to this difficulty is to reduce the dimension of the observed and simulated fields to less than the number of realizations before using any inference procedure. This can be done using pattern analysis techniques, such as EOF analysis, that try to identify the climate's principal modes of variation empirically. Another solution is to abandon classical inference techniques and replace them with ad hoc methods, such as the 'PPP' test (Preisendorfer and Barnett [320]).

Both grid point and field significance tests are plagued with at least two other problems that result in interpretation difficulties. The first of these is that the word *significance* does not have a specific physical interpretation. The statistical significance of the difference between a simulated and observed climate depends upon both location and sample size. Location is a factor that affects interpretation because variability is not uniform in space. A 5 m difference between an observed and a simulated mean January 500 hPa height field may be statistically very significant in the tropics, but such a difference is not likely to be statistically, or physically, significant at midlatitudes where interannual variability is large. Sample size is a factor because the sensitivity of statistical tests is affected by the amount of information about the mean state contained in the observed and simulated realizations. Larger samples have greater information content and consequently result in more powerful tests. Thus, even though a 5 m difference at midlatitudes may not be physically important, it will be found to be significant given large enough simulated and observed climatologies. The statistical strength of the signal (or model error) may be quantified by a parameter called the *level of recurrence*, which is the probability that the signal's signature will not be masked by the noise in another identical but statistically independent run with the GCM (Sections 6.9–6.10).

The second problem is that objective statistical validation techniques are more honest than modellers would like them to be. GCMs and analysis systems have various biases that ensure that objective tests of their differences will reject the null hypothesis of no difference with certainty, given large enough samples. Modellers seem to have an intuitive grasp of the size and spatial structure of biases and seem to be able to discount their effects when making climate comparisons. If these biases can be quantified, statistical inference procedures can be adjusted to account for them (see Chapter 6).