ON THE ROLE OF STATISTICS IN CLIMATE RESEARCH

FRANCIS W. ZWIERS and HANS VON STORCH

a Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, Victoria, BC, Canada
b Institute for Coastal Research, GKSS Research Center, Geesthacht, Germany

Received 22 March 2003
Revised 6 January 2004
Accepted 17 January 2004

ABSTRACT

We review the role of statistical analysis in the climate sciences. Special emphasis is given to attempts to construct dynamical knowledge from limited observational evidence, and to the ongoing task of drawing detailed and reliable information on the state, and change, of climate that is needed, for example, for short-term and seasonal forecasting. We conclude with recommendations of how to improve the practice of statistical analysis in the climate sciences by drawing more efficiently on relevant developments in statistical mathematics. Copyright © 2004 Environment Canada. Published by John Wiley & Sons, Ltd.

KEY WORDS: statistical methods; statistical climatology; climate diagnostics; climate change; weather forecasting

1. INTRODUCTION

The study of the climate system is, to a large extent, the study of the statistics of weather; so, it is not surprising that statistical reasoning, analysis and modelling are pervasive in the climatological sciences. Statistical analysis helps to quantify the effects of uncertainty, both in terms of observation and measurement and in terms of our understanding of the processes, that govern climate variability. Statistical analysis also helps us to identify which of the many pieces of information derived from observations of the climate system are worthy of synthesis and interpretation.

Statistical methods are needed for a whole gamut of activities that contribute to the ultimate synthesis of climate knowledge, ranging from the collection of primary data, to the interpretation and analysis of the resulting high-level data sets. Statistical procedures are fundamental components of procedures for data retrieval from remotely sensed observations (e.g. satellite and radar), data retrieval from proxy records (e.g. tree rings, coral records, ice and sediment cores), quality control (primarily to ensure the homogeneity of observational data), determination of the representativity of data (e.g. urban warming effects), data assimilation, the identification of predictive information in a given state (forecasts) and climate-change detection. Statistical procedures are also integral to the large majority of efforts that seek physically meaningful interpretations of observed climate variability. This includes the identification of ‘modes’ in the climate record, such as the Arctic oscillation (AO) pattern (Thompson and Wallace, 1998), the identification of subsystems such as the Madden–Julian oscillation (MJO; Madden and Julian, 1971) and the formulation of conceptual models explaining the variability of climate (e.g. Hasselmann, 1976; Da Costa and Vautard, 1997; Burgers, 1999; Dobrovolski, 2000; Timmermann and Lohmann, 2000).

In the remainder of this paper we will briefly survey some of the aspects of climate science in which statistics plays a central role. We will not, however, provide a total overview. The climate sciences, which, among
others, include atmospheric and oceanic physics, remote Earth observation, and palaeoclimatic reconstructions from proxy data, are simply too broad for us to be exhaustive. Furthermore, our survey is strongly constrained by our own research experience and interests. We hope that readers will not be offended by our selection of topics. Although we discuss only a few areas in the climate sciences, the message that we hope to convey is that the use of statistics is pervasive in the climate sciences, not only for the extraction and quality control of data, but also for the synthesis of knowledge and information from that data. Thus, it is surprising that there are only a very few comprehensive books on statistics for climatologists (e.g. Wilks, 1995; von Storch and Zwiers, 1999) and only an additional few that deal with specialized topics (e.g. Daley, 1991; von Storch and Navarra, 1999; Jolliffe, 2002; Jolliffe and Stephenson, 2003).

Some readers may be disappointed to see that we have not delved into fashionable topics such as neural networks (e.g. Hsieh and Tang, 1998), cluster analysis (e.g. Mimmack et al., 2001) or wavelet analysis (e.g. Farge et al., 1993) to any great extent. These techniques are no doubt very useful. For example, neural nets have proven to be very useful as emulators of complex physical models. Such emulators can substantially reduce the cost of the operational processing of high volumes of remotely sensed data (e.g. Schiller and Doerffer, 1999). However, our perception is that the application of these kinds of techniques in diagnostic studies has not generally improved our ability to synthesize knowledge, either from the observational record or from the output of physically based dynamical models such as coupled global climate models. Rather, much of the work that has had a large impact on climate research has used relatively simple techniques that allow transparent interpretation of the underlying physics (e.g. Walker and Bliss, 1932; Madden and Julian, 1971; Wallace and Gutzler, 1981; Thompson and Wallace, 1998). A point that we will emphasize below is that a state space approach, in which statistical models are used to describe what we have observed and how this is related to the underlying dynamical system, is very appropriate for climate applications. Such an approach is often used implicitly in climate research, in the sense that climate scientists do their work in a framework that is defined by dynamical knowledge.

Some readers may also be disappointed that we have not surveyed recent developments in the statistical sciences that may pertain to climate research. This is not to say that statistical science has nothing to offer. Indeed, quite the opposite is true, as, for example, can be seen by reading the recent literature on extreme-value analysis (e.g. IPCC, 2002; Katz et al., 2002; Kharin and Zwiers, 2004) or forecast calibration and verification (e.g. Jolliffe and Stevenson, 2003; Coelho et al., 2004). Rather, this choice reflects our experience that new technology developed in the statistical sciences cannot simply be transferred to the climate sciences by informing climatologists of its existence. Effective technology transfer across disciplines requires cross-fertilization, such as has occurred in geostatistics, and as is being fostered by efforts such as the National Center for Atmospheric Sciences (NCAR) Geophysical Statistics Project (http://www.cgd.ucar.edu/stats/).

The remainder of this paper is organized as follows. Section 2 deals with the use of statistical methods to learn about the dynamics of the climate system. In Section 3 we discuss some of the roles of statistics in the acquisition of data, and in Section 4 we discuss the role of statistics in forecasting. We continue in Section 5 with a discussion of climate-change detection, and complete the paper in Section 6 with some additional discussion, a summary, and a few recommendations.

2. CONSTRUCTING KNOWLEDGE ABOUT THE DYNAMICS OF CLIMATE

When we construct knowledge, i.e. produce a structured, reasoned judgment about the functioning of the climate system, we implicitly or explicitly assume a state space model (e.g. Honerkamp, 1994). These statistical models clearly discriminate between a system of state variables that define a theoretical construct and a collection of observables that contain information about the system. More formally, we describe the system with a state vector \( \phi \), that is continuous in space and time. Almost always, this state vector cannot be observed, sometimes because of lack of suitable sensors, but also because continuous space–time observations are generally not available. The dynamics \( F \) of this state variable are often described by a system of differential equations:

\[
\dot{\phi} = F(\phi, \alpha, t) + \epsilon
\]

The dynamics depend on a set of free parameters \( \mathbf{\alpha} = (\alpha_1, \alpha_2, \ldots) \). The ‘noise’ term \( \mathbf{\epsilon} \) that appears on the right-hand side of Equation (1) arises because the dynamics can only approximately describe the behaviour of the system regardless of the choice of the parameters, and because of the fact that seemingly random effects act upon the system. Of course, the dynamics may generate internal noise as well, and noise may enter the system in more subtle ways than suggested by the simple additive model in Equation (1). The dynamics \( \mathbf{F} \) may be derived from theoretical arguments, such as the conservation of momentum or mass, or less frequently from an empirical fit.

Even if \( \mathbf{\phi} \) is not observable in its entirety, observations containing information about \( \mathbf{\phi} \) may be available, for instance at some locations and at some times. Alternatively, indirect evidence may be available from proxy information contained in media such as tree rings, ice cores, corals, and lake warves (e.g. Fritts, 1991; Cook, 1995; Bradley, 1999; Mann et al., 1999; Briffa, 2000; Crowley and Lowery, 2000; Jones et al., 2001a; Esper et al., 2002). When these observations are combined into an observation vector \( \mathbf{\omega} \), an observation equation can be used to relate the state variable to the observed variables

\[
\mathbf{\omega}_t = P( \mathbf{\phi}_t ) + \mathbf{\delta}_t \tag{2}
\]

through an operator \( P \). The noise term \( \mathbf{\delta} \) on the right-hand side of Equation (2) indicates that the observation equation is not exactly satisfied; there may be measurement uncertainties with respect to the value, location or timing of \( \mathbf{\omega} \). As with the dynamics, noise may enter into the observations more subtly (e.g. perhaps as a multiplicative term) than indicated by Equation (2). Also, the link between \( \mathbf{\phi} \) and \( \mathbf{\omega} \) may not be fully known, as is the case with proxy data.

The state space approach is the core of Hasselmann’s principal interaction pattern (PIP) concept (Hasselmann, 1988), with the PIPs being a basis of an optimally determined subspace that contains the relevant dynamics. This concept is general and difficult to implement. Examples are provided by Achatz et al. (1995), Achatz and Schmitz (1997), Kwasiok (1996) and Selten (1995). An important aspect of this approach is that it provides a means for estimating (an upper limit of) the number of degrees of freedom of the dynamical phenomenon and to characterize the system’s attractor (in the sense of a manifold within which the system may be considered approximately closed) by specifying suitable coordinates. Selten (1995) and Kwasiok (1996), for instance, found that less than 20 degrees of freedom were enough to describe the Northern Hemisphere equivalent barotropic dynamics. Reduced models based on about 30–50 degrees of freedom were found to be able to capture the essential features of the long-term behaviour of a spectral T42 quasi-geostrophic climate model.

Conceptually simpler analyses formulate a model, in a given space, with a few unknown parameters. The task of the statistician is then to estimate these parameters and to determine the skill of the fitted model in reproducing the variability of the observables. Examples include the delayed action oscillator model of the El Niño–southern oscillation (ENSO) phenomenon (with local wind and sea-surface temperature (SST) observations as observables; Suarez and Schopf, 1988; Battisti and Hirst, 1989; Mechoso et al., 2003) and the stochastic climate model (with the spectra of, for instance, SST as observables; Hasselmann, 1976; Frankignoul, 1985). The analysis of observed data in the framework of a spatial wave-like phenomena, as in the case of principal oscillation patterns (POPs; e.g. analysis of the MJO by von Storch and Xu (1990)) or a frequency wavenumber analysis (analysis of extratropical storm track by Hayashi (1982) and Speth and Madden (1983)) are other examples. Questions regarding air–sea interaction (Does the ocean drive the atmosphere or vice versa?) can also be addressed successfully in this format (von Storch, 2000; Zorita et al., 1992; Goodman and Marshall, 2003).

A widely pursued line of research is the identification of ‘modes’ in the observational record or in extended quasi-realistic model simulations. A large variety of methods have been developed in the past several decades, with empirical orthogonal functions (EOFs) as the most frequently used. Other techniques include variants of EOFs, such as rotated EOFs (Richman, 1986), extended EOFs (Weare and Nasstrom, 1982), Hilbert EOFs (or, as they are often named, complex EOFs; Wallace and Dickinson, 1972), singular spectrum analysis (SSA; Vautard et al., 1992) and multi-channel SSA (MSSA; Vautard, 1995). Other techniques used to identify modes include the analysis of teleconnections (Wallace and Gutzler, 1981), empirical orthogonal teleconnections (van...

In all cases, the quest is not for information about the state of the system, but rather about the dynamics of the system. Initially, the goal will be modest, i.e. simply to gather information that will eventually allow one to articulate the dynamics encompassed in Equation (1). Ultimately, however, the question becomes more specific: Does Equation (1) successfully describe the observed variability?

The danger with ‘blind’ techniques, such as canonical correlation, EOF, POP and other types of analysis, is that they deliver dynamically interesting looking numbers and patterns, but reliable problem-independent tests are unavailable. Confirmatory analysis, whereby one tests a specific hypothesis developed from exploration with such techniques on independent data, is hardly possible in this context. This is not to say that descriptive, exploratory analysis with these tools is not useful. The systems considered are complex and non-linear, and often do not yield to analysis with first principles. In that case, exploratory analysis is a valuable aid that may lead to recognition and identification of dynamical mechanisms of variability. One should, however, keep in mind that the patterns obtained from methods such as EOF analysis are not a priori constructed to extract information about dynamical behaviour: they simply provide an efficient representation of variability. There is no guarantee that the often tacitly made assumption that EOFs represent dynamical ‘modes’ of variability will hold. The recent discussion about the dynamical interpretation of the AO and the North Atlantic oscillation is an indication of this problem (Baldwin and Dunkerton, 1999; Deser, 2000; Monahan et al., 2000, 2001, 2003; Wallace, 2000; Ambaum et al., 2001; Thompson and Wallace, 2001). Also the discussion about the ‘Indian Ocean dipole’ (Saji et al., 1999) is sometimes drawn in this direction (Dommenger and Latif, 2002, 2003; Behera et al., 2003; Jolliffe, 2003). Another such example concerns the interpretation of the coupled modes that are obtained by performing a singular vector decomposition on a cross-covariance matrix (Bretherton et al., 1992; for subsequent caveats on the usage of this technique see Newman and Sardeshmukh (1995) and Cherry (1996)).

A symptom of the difficulty with confirmatory analysis is the often heard, ill-posed, question about the significance of EOFs. If ‘significance’ is meant to stand for ‘non-degenerate’ (in the sense that the eigenvectors are not primarily mixed combinations of the true underlying eigenvectors), then tools such as the North et al. (1982) ‘rule of thumb’ appear to be useful, even though the inferences made with these rules may not be precise. For example, whereas the North et al. (1982) rule has its basis in asymptotic statistical theory, we inevitably apply it in circumstances when samples are very far from being asymptotically large. But simple ‘rules’ are not available if the question concerns the uncertainty of the pattern itself. Here, it is necessary to build confidence and understanding through the analysis of simplified examples, to articulate the particular question carefully regarding the pattern uncertainty that we wish to assess, and then to apply resampling tools carefully to the full pattern estimation procedure (e.g. Michaelson, 1987; Barnston and van den Dool, 1993; Efron and Tibshirani, 1993). In most cases, the only credible way in which to discriminate between the effects of the sampling properties of the observables and truly dynamical features of the system is dynamical plausibility and reproducibility in detailed dynamical models.

3. ACQUIRING HIGH-QUALITY, LOW-LEVEL INFORMATION

Traditionally, from the 19th to the middle of the 20th century, climatology was bookkeeping for meteorology. The problem was to record in an objective and representative way typical climatic variables such as near-surface minimum, maximum and mean temperature, precipitation amounts, surface pressure, wind speeds, wave heights and the like. These data were, and are, used for planning and safety standards of houses and, for instance, offshore operations. As such, the data need to be comparable — in space and time.

In many applications the data are assumed to fit a stationary (if we disregard the annual cycle) probability distribution for describing the ‘normal’ range of variability, or to fit an extreme value distribution for assessing risks of rare events. This task does not really pose a challenge for modern climatology as long as the assumption of stationarity is satisfied. The challenge is with temporally changing statistics. This is not a new challenge:
Brückner (1890) described a systematic approach for making data comparable in space and time in order to describe large-scale climatic variations objectively over centuries. Brückner (1890), among other variables, attempted to homogenize water-level records from the Caspian Sea by comparing simultaneous observations from different locations.

Often, records exhibit creeping or sudden changes. Sometimes these changes are not due to changes in the climatic variables, but to changes in the instruments used to measure the variables, the physical characteristics of the instrument’s immediate environment (its exposure), and recording and observing procedures. Easy-to-understand effects relate to a change in the exposure of instruments, as described by Karl et al. (1993), Peterson et al. (1998) and Vincent and Gullett (1999) among others. Also, observation practices, such as changes in the observation time (already noted by Ellis (1890) and Donnell (1912); a modern reference is Karl et al. (1986)) or changing ship routing affect the climate record. More examples are listed in Folland et al. (2001). Less obvious effects are found in remotely sensed data, e.g. related to the changing flight altitude (orbital drift), which played an important role in the recent controversy about the apparent inconsistency between the warming at the surface and the much smaller rate of warming in the lower troposphere (e.g. Christy et al., 2000, 2001, 2003; NRC, 2000; Santer et al., 2000, 2003; Folland et al., 2001; Mears et al., 2003).

Other unwanted effects appear if the data are not representative of what they are assumed to be. Growing populated areas produce changes in the local micro-climate (urban warming) that are real and, when studying the climate of cities, very relevant. However, such local changes cannot be interpreted as being representative of climate change on the continental or even global scale. Climatologists understand how to avoid these problems (Jones et al., 1990; Gallo and Owen, 1999; Peterson and Vose, 1997; Peterson et al., 1999), but misunderstanding about this effect has been a cause for continuing doubt about global warming among lay people.

Statistics helps in these cases in first discriminating such changes as being either ‘inhomogeneities’, i.e. related to changes in the instrument, observing procedure, the local environment, or a true climatic variation, and to correct for inhomogeneities if needed. Sudden changes are sometimes detected with classical ‘change-point’ analysis (e.g. Busuioc and von Storch, 1996). Sudden and creeping changes are found by systematically comparing neighbouring stations, a technique that was already in use by Brückner in 1890. Peterson et al. (1998) give an extensive overview about the state of the art. Also, for indirect climate data, like tree rings, such techniques to correct for non-climatic influences have been developed (e.g. Cook, 1995; Fritts, 1976, 1991). In addition, dendro-climatologists have developed several data-adjustment techniques that better preserve low-frequency and secular variability when several records are joined together to form a very long temperature time series (e.g. Briffa et al., 2001; Esper et al., 2002).

Another modern challenge is the global, dynamically consistent analysis of the synoptic state of the atmosphere or ocean. Geostatistical techniques, which include ‘optimal interpolation’ in meteorology, have been used successfully; also, Kalman-filter-type techniques have been developed, tested and applied (Evensen, 1992, 1994; Burgers et al., 1998; Houtekamer and Mitchell, 2001). Such combined statistical–dynamical analysis is in routine use at the weather services. In case of the European Centre for Medium-range Weather Forecasting (ECMWF) and the National Centers for Environmental Prediction (NCEP), these schemes have been used to analyse retrospectively weather observations for periods of 15 years (ECMWF ERA-15, 1979–94; Gibson et al., 1997) and 50 years (NCEP, 1958–2001; Kalnay et al., 1996). The ECMWF has recently completed an extended, high-resolution global 40 year reanalysis (see http://www.ecmwf.int/research/era). These retrospective products have been immensely useful to climate research because they provide data with complete global coverage that are as consistent as possible with our archives of in situ and remotely sensed observations, and that are not affected by the frequent changes that occur in operational data analysis systems that assimilate data into weather forecasting models. However, there remains an urgent need for very high resolution (<50 km) regional retrospective analyses for the study of climate on scales where the impacts of climate variability and change can be assessed.
4. FORECASTING

Statistics, and statistical reasoning, has also played key roles in the forecasting of weather and climate on time scales ranging from hours to several seasons. Statistical methods have been used to assess the potential predictability of climate and weather, to develop schemes for initializing dynamical forecasting models, for post-processing dynamical forecasts (both to remove biases and to add additional skill), and to forecast future weather and climate states empirically without the aid of dynamical models. Statistical concepts have also been key to the evaluation of forecasting systems and have provided objective forecast skill scores that cannot be manipulated by the forecaster.

4.1. Weather forecasting

The multiply chaotic evolution of weather, with rapidly growing differences between any two rather similar states (analogues), makes weather forecasting a field for statistical analysis. First, the problem of forecasting is cast in a statistical framework: the future weather state is understood as a random variable conditioned upon the initial state and, in certain cases, external parameters such as SST or soil moisture. Of course, the forecast lead, i.e. the time in the future that the forecast is supposed to predict, is an important parameter as well. In the following we will briefly touch upon four problems associated with weather forecasting that have been dealt with via statistical methods: initial conditions, predictability, forecast improvement via the perfect prognosis (PP) or model output statistics (MOS) approach, and the assessment of forecast skill.

Early weather forecasts, whether subjective or objective, were essentially statistical. Forecasts were often based on a process called weather typing (e.g. Köppen, 1923), which is similar to modern-day cluster analysis of weather states. In contrast, modern weather forecasts are produced by numerical weather prediction (NWP) systems, with sophisticated data assimilation systems for determining the initial state plus advanced high-resolution dynamical global or regional models of the atmosphere. (Note that ‘frozen’ versions of such systems are used to produce the retrospective analyses (reanalysis) discussed in Section 3). Statistics continues to play an important role in NWP. At the ‘front end’ of the forecasting system, data assimilation systems are used to specify the initial conditions for NWP models. More broadly, data assimilation is closely related to objective analysis, the spatial (and sometimes temporal) interpolation of data that are scattered irregularly in space (and time) to a regular grid in space (and a fixed point in time). The methods used range from techniques that are primarily statistical to methods that are fully integrated with the dynamics of the system that is being analysed.

The primarily statistical techniques, which are often termed objective analysis, include ad hoc methods such as Cressman (1959) successive correction (see Given and Ray (1994) for an application; see Trapp and Doswell (2000) and Askelson et al. (2000) for discussion of the use of this technique in the analysis of radar data), to optimal interpolation techniques (Gandin, 1963; Thiébaux and Pedder, 1987; see Sokolov and Rintoul (1999) for a recent inter-comparison of techniques). In the latter, care is often taken to model the spatial correlation structure of forecast errors so as to maintain the dynamical balance between forecast parameters, such as geopotential and winds (e.g. Thiébaux, 1976). At the opposite end of the spectrum are modern three-dimensional (space only) and four-dimensional (space and time) variational techniques (e.g. Daley, 1991) that incorporate so-called adjoint models (e.g. Le Dimet and Talagrand, 1986; Thépaut and Coultrier, 1991; Tanguay et al., 1995) of the atmospheric dynamics. These techniques are now also being coupled with state space (Kalman filtering) approaches that include explicit representation of the dynamics of the system that is to be predicted and the statistical nature of the observables (e.g. Daley, 1991; Rabier et al., 1998; Houtekamer and Mitchell, 2001). An overview of methods used in oceanography is given by Robinson et al. (1998).

Statistics also plays an important role at the output end of NWP systems. The performance of these NWP systems is far better than that of any statistical forecast system on scales of a few hundred kilometres and greater, but they often fail to deliver skilful forecasts of local, impact-relevant variables, such as the danger of freezing or sunshine duration in a physiographically structured landscape. Statistical methods have long been used to correct for systematic errors, particularly biases, in NWP products and to derive transfer functions that relate well the forecast tropospheric variables to impact-relevant surface variables.

Two classes of forecast-improvement methods have been developed, namely PP (Klein et al., 1959) and MOS (Glahn and Lowry, 1972; Klein and Glahn, 1974). In the former, a statistical model, often a multiple regression, is derived from simultaneous observations of the tropospheric variables and the variable of interest, whereas MOS ‘consists of determining a statistical relationship between a predictand and variables forecast by a numerical model at some projection time’ (Glahn and Lowry, 1972). The latter is more efficient, and is able to adjust for forecast model biases and other kinds of systematic error, but it needs to be updated whenever a component of the forecast system is modified. The PP approach has received renewed interest in the past 10 years for ‘downscaling’ large-scale climate-change information to the local level (e.g. Wilby et al., 1998; Widman et al., 2003). The MOS approach has continued to be used for weather forecasting, and considerable research has been conducted on updatable MOS systems that more easily adapt to the frequent changes that are made in weather forecasting systems (Ross, 1987; Wilson and Vallée, 2002; Yuval and Hsieh, 2003). MOS approaches are now also being used extensively for seasonal forecasting (see below).

In addition, statistics plays important roles in the study of predictability (Lorenz, 1982) and in the assessment of the skill of forecasting systems (e.g. Mason and Graham, 1999). The predictability question relates to the forecast lead for which all forecast skill is lost, i.e. that lead at which the distribution of the verifying observations, conditional upon the forecast, is as wide as the unconditional distribution. Related to this is the question of whether the uncertainty of the forecast itself may be predicted. Advancement in all of these areas (the specification of initial conditions, predictability studies, and dynamical and statistical forecast improvements) is measured by means of forecast skill scores (Murphy and Epstein, 1989; Gandin and Murphy, 1992; Murphy and Wilks, 1998; Mason and Graham, 1999; Wilks, 2000). Overviews are given by Stanski et al. (1989), Livezey (1999) and von Storch and Zwiers (1999).

4.2. Seasonal forecasting

The instantaneous state of the atmosphere is not generally predictable beyond about 2 weeks because of its inherent chaotic nature. Nonetheless, in many parts of the world there has been gradual progress in forecasting seasonal mean conditions with leads of up to about a year. This has been achieved by conditioning on parts of the climate system that evolve more slowly than the atmosphere and communicate with the atmosphere (e.g. see Shukla (1998) and references cited therein). Most effort to date has focused on the tropical ocean as the source of predictable climate signals on time scales of months to seasons (e.g. Livezey et al., 1997; Hoerling and Kumar, 2002, 2003) although it is recognized that there may also be other signal sources, such as soil moisture (Fennesy and Shukla, 1999).

Considerable work has been done to assess the potential predictability of climate on seasonal to inter-annual time scales (e.g. Zwieters et al., 2000 and references cited therein), and decadal time scales (e.g. Latif and Barnett, 1996; Rowell and Zwieters, 1999; Saravanan et al., 2000; Collins and Allen, 2002). Much of this work has used the potential predictability concept that was first outlined and applied by Madden (1976). This is essentially a classical analysis of variance (ANOVA; e.g., Scheffé, 1959) conducted in the time domain so that one looks for evidence of excess variability on time scales longer than the scale of interest (e.g. seasonal; see Zwieters (1996)). This evidence is interpreted as an indication of potential predictive skill.

Observational studies (e.g. Madden, 1976) have been supplemented with perfect model studies in which the time evolution of a slow part of the climate system (typically SST) is prescribed. Two-way ANOVA is often applied to ensembles of such simulations to quantify the proportion of inter-annual or decadal variability that is ascribable to lower boundary forcing of the climate system (e.g. Zwieters, 1996; Rowell, 1998; Zwieters et al., 2000).

The most reliable seasonal forecasting tools are presently statistical, prevalently canonical correlation analysis (e.g. Barnston, 1994; Shabbar and Barnston, 1996; Hwang et al., 2001). This approach is moderately successful in parts of the world where there are teleconnections between the ocean surface and the atmosphere. Many other approaches have been used, including POP analysis (Tang, 1995), SSA, MSSA, neural networks (Tangang et al., 1998).

While statistical methods are prominent, the widely accepted goal is to be able to produce reliable dynamical forecasts. Most dynamical systems are presently two tier (Bengtsson et al., 1993), meaning that lower
boundary conditions (usually SST) are first forecast, followed by a calculation to determine the atmospheric response to the boundary forcing forecast. Lower boundary forecasts may be statistical or dynamical in nature. Most systems presently require MOS to adjust for biases in the forecast system (e.g. Derome et al., 2001) and to link surface parameters to the generally more skillfully forecast large-scale circulation. This is an area of active research (e.g. Krishnamurti et al., 1999, 2000; Doblas-Reyas et al., 2000; Kharin et al., 2001; Kharin and Zwiers, 2002, 2003a,b; Yun et al., 2003). Anderson et al. (1999) assess statistical and dynamical seasonal forecasting methods. Kharin and Zwiers (2003a) describe several methods for producing probability forecasts of seasonal conditions, and Kharin and Zwiers (2003b) discuss some properties of skill scores that are often used to evaluate seasonal forecasts. In connection with seasonal probability forecasts, see also Buzzia and Palmer (1998), Mason and Graham (1999) and Wilks (2001).

Skill is considerably more difficult to assess in climate forecasting than in NWP because of the increased time scale of the forecasts, and hence the reduced frequency with which forecasts can be made and evaluated. This substantially slows progress in this area. Kharin and Zwiers (2002) discuss the limitations of some seasonal forecasting improvement techniques. Wilks (2000) demonstrates an evaluation of the skill of seasonal forecasts produced at the NCEP and discusses diagnostics of the consistency of seasonal forecasts that are issued monthly.

5. CLIMATE-CHANGE ASSESSMENT

Human activity is altering the composition of the Earth’s atmosphere through the addition of greenhouse gases such as carbon dioxide and chlorofluorocarbons, and various species of aerosol. Science has been aware of the potential for global warming by greenhouse gases at least since Arrhenius (1896). This prospect has been the focus of much physical research and data analysis during the past 20 years. See, for example, the assessments of the Intergovernmental Panel on Climate Change (e.g. Houghton et al., 2001), and the United States National Assessment (NAST, 2001; MacCracken et al., 2003).

Several external forcing factors are thought to affect the mean state of the climate on decadal and longer time scales (Houghton et al., 2001). These include anthropogenically caused changes in radiative forcing that result from the emission of greenhouse gases and the creation of aerosols from fossil fuel and biomass burning. In addition, there are natural external forcing factors, such as changes in solar irradiance and volcanic activity. However, the climate system does not vary because of external factors alone. The climate system, even when not perturbed by external factors, produces substantial amounts of natural internal variability (von Storch et al., 2001) on large spatial scales and long temporal scales. Hence, detection and attribution of the effects of external forcing are statistical signal-to-noise problems (Hasselmann, 1979, 1993).

The detection part of this problem is the process of demonstrating that an observed change is not likely to have been entirely the result of natural internal variability. The attribution aspects of the problem are more difficult because it is not possible to conduct controlled experiments with the climate system. The practical approach that has been taken in the climate research community involves statistical analysis and the assessment of multiple lines of evidence to demonstrate that: (a) observed changes are consistent with forcing of the climate by a combination of anthropogenic and natural external factors; and (b) the changes are inconsistent with alternative, physically plausible explanations.

The detection technique that has been used in most studies performed to date has several equivalent representations (Hegerl and North, 1997; Zwiers, 1999; Hegerl and Allen, 2002). It can be cast as a multiple regression problem (Hasselmann, 1993, 1997; Allen and Tett, 1999) in which a field of \( n \) ‘observations’ \( \mathbf{y} \) is represented as a linear combination of signal patterns \( \mathbf{g}_1, \ldots, \mathbf{g}_m \) plus residual climate noise \( \mathbf{n} \):

\[
\mathbf{y} = \sum_{i=1}^{m} a_i \mathbf{g}_i + \mathbf{n} = \mathbf{G} \mathbf{a} + \mathbf{n}
\]  

(3)

where \( \mathbf{G} = (\mathbf{g}_1 | \ldots | \mathbf{g}_m) \) is the matrix composed of the signal patterns and \( \mathbf{a} \) is the vector composed of the unknown amplitudes.
The signal patterns are derived by using physically based models to simulate the response to estimated changes in external forcing that are thought to have occurred during the past two centuries. These physical models generally also simulate internal variability. Therefore, the signals (responses to external forcing) are usually estimated by averaging a small ensemble of runs with the same model, each run being started from different initial conditions. This averaging reduces contamination of the signal by the internal noise that is simulated by the physical model. The amplitudes $a$ are estimated either by means of a standard least-squares method, or by the total least-squares method (Ripley and Thompson, 1987). The latter method is generally preferred because it takes account of the fact that the signal estimates are not completely free of the effects of internal noise.

The detection issue is dealt with by testing the null hypothesis $H_0$: $a = 0$ that the signals have zero amplitude in the observations. The test that is used in most cases is a variant of the Hotelling $T^2$-test, although Bayesian approaches to assess the presence of the signals in the data have also been used (e.g. Berliner et al., 2000; Schnur and Hasselmann, 2004).

Research on climate-change detection and attribution over the last 10–20 years has searched for evidence of anthropogenic and natural external signals primarily in the observed global temperature record. Some recent studies have also considered some other climate elements, such as the global ocean heat content (Barnett et al., 2001; Reichert et al., 2002) or even the propagation of ocean waves (Pfizenmayer and von Storch, 2001).

Statistical evidence typically contributes to an attribution assessment by testing the null hypothesis $H_A$: $a = 1$ that the model simulated signals all have the correct amplitude. If this can be demonstrated, then there is evidence that the physical model has responded to the historical changes in external forcing in the same way as the observed system, which in turn contributes to a comprehensive attribution assessment, as described in Mitchell et al. (2001). A difficulty with this approach is that the analyst must interpret a failure to reject $H_A$ as evidence in support of an attribution assessment. This situation is not entirely acceptable, because it means that the likelihood of attribution cannot be controlled by acquiring more information or better delineating the signals $G$. In fact, when more data become available, the power of the test increases, making the rejection of $H_A$ more likely. Levine and Berliner (1999) have pointed out that a more satisfactory approach would be to conduct a test of inconsistency in which consistency is the alternative hypothesis. Such a test has, apparently, not yet been applied in the climate literature. Bayesian methods (e.g. Hasselmann, 1998; Berliner et al., 2000; Schnur and Hasselmann, 2004) approach the problem in a more satisfactory way by using evidence from the observations to estimate the $a$ posteriori probability of attribution for a suitable defined attribution criterion.

A difficulty with the detection and attribution analysis is that an estimate of the covariance matrix $C_{nn}$ of the residual noise field is required to make statistical inferences about the amplitudes $a$. However, the instrumental record gathered during the past 150 years (e.g. Jones, et al., 2001b; Jones and Moberg, 2003) cannot provide a reliable estimate of residual noise covariability. This is because the length of the observed record is not sufficient to estimate variability on the decadal and longer time scales that are important for detection and attribution. Also, natural internal variability is confounded with the effect of anthropogenic and natural forcing during the instrumental period. Palaeo-reconstructions of past climate are a possible future source of information for this purpose. There is growing confidence that these records provide realistic representations of hemispheric-scale decadal variability (e.g. Hegerl et al., 2003), but they presently do not adequately resolve spatial variability on the scales needed for detection and attribution. Thus, the covariance matrix is generally estimated from long control simulations performed with a climate model in which concentrations of greenhouse gases and aerosols are fixed at present or pre-industrial levels.

The climate models used in detection and attribution studies are likely not to be able to simulate climate variability correctly on all spatial and temporal scales. This problem is circumvented, in part, by performing the detection and attribution analysis in a reduced dimension space that is spanned by a small number of EOFs of the estimated internal variability. Thus, a constraint on the choice of the number of EOFs is that the variability of the residual noise should be consistent with the variability of the control simulation in the dimensions that are retained. An approximate chi-squared test can be used for this purpose (Allen and Tett, 1999). Detection and attribution studies typically use approximately 10 EOFs, although some studies (e.g.
North and Wu, 2001) have used a much higher level of truncation. Experience has shown that results are typically not very sensitive to the level of truncation (e.g. see Zwiers and Zhang (2003)).

Multiple linear regression is not the only technique that has been used in detection and attribution studies. There is also a substantial body of work (see Santer et al. (1996) and Zwiers (1999)) that relies on pattern correlation methods. These methods are closely related to optimal detection with one signal pattern.

6. SUMMARY AND DISCUSSION

We have tried to survey some of the ways in which statistics pervades the climate sciences. Obviously, a total overview is impossible: the climate sciences, which include atmospheric and oceanic physics, remote Earth observation, and palaeoclimatic reconstructions from proxy data, are simply too broad for us to be exhaustive. We have covered only a limited part of the field, and have not used this paper to attempt to transfer new technology to climate science from the statistical sciences. Instead, our purpose has been to demonstrate that statistical concepts and methods are necessary in all facets of the climate science enterprise, ranging from the gathering of data to the derivation of knowledge from carefully synthesized data products.

Statistical analysis is needed to interpret observations, from either the real world or the artificial world of a climate model, in a proper framework. Statistical reasoning allows analysts to extract information about some part of the climate system by making a number of simplifying assumptions about the way in which the system generates information (e.g. an observed process is stationary and ergodic) and about the way in which the data analysed have been observed. Within the context of these assumptions, statistical reasoning imposes an important element of rigour when extracting information from data. Again, within the context of the explicit, and implicit, approximating assumptions, statistical methods allow one to deal explicitly with the effects of uncertainty on inferences and to quantify its effects on forecasts, projections, etc.

This need for statistical thinking stems foremost from two particularities of the climate system, which are that the climate system has a large number of components (degrees of freedom) and that it is impossible to conduct laboratory experiments with the Earth system. Consequently, there is considerable scope for the misinterpretation of statistical evidence. Flawed results can be avoided, not through the use of a ‘silver bullet’, but by clearly articulating all assumptions required to apply a given analysis technique. This suggestion applies to both the concrete analysis of specific data and to the conceptual framework within which we build our theories and knowledge (Petersen, 1999; Sarewitz and Pielke, 1999; von Storch, 2001a,b). For the latter, the concept of state space models is particularly useful, as it helps us to discriminate between our hypothetical construction of dynamics and the observational process. Flawed results, when they occur, are not easy to identify, and significant work is sometimes required to uproot them. Examples of the successful disclosure of such problems are given by Allen and Smith (1996) and Nitsche et al. (1994).

When dealing with environmental systems, different types of uncertainty arise (Sarewitz and Pielke, 1999; Risbey et al., 2000). One type of uncertainty is inherent, like the weather in 50 days or greenhouse gas emissions in 50 years. No improved understanding of atmospheric or social dynamics will provide us with information; the only way to reduce the uncertainty is to wait for 40 days, or 40 years, until we enter the period prior to the future date of interest in which the weather or greenhouse gas emissions are actually predictable. However, there is another type of uncertainty, which is temporary or malleable. By collecting additional data and by improving our conceptual understanding of the phenomenon under consideration, we may in some cases reduce uncertainty. One example is the additional information provided by the Tropical Atmosphere Ocean buoy array in the equatorial Pacific, which allows a more complete analysis of the state of the tropical Pacific and has thus provided knowledge to improve ENSO forecasting (see the extensive bibliography available at http://www.pmel.noaa.gov/tao/). A second example is the improved understanding of the climate system that helps us to constrain our estimates of the climate sensitivity to increased greenhouse gases (e.g. Gregory et al., 2002).

The role of statistics is, again, twofold. It helps to quantify the degree of ‘inherent’ uncertainty, and it helps to assess the value of learning. The Bayesian approach seems to be more successful in dealing explicitly with both types of uncertainty, as is demonstrated for instance by Risbey et al. (2000) in an analysis of the
inherent and malleable uncertainties of global warming. Another example is Hasselmann’s (1998) study of the role of preconceptions in detecting climate change (also see Schnur and Hasselmann (2004)). An account, mainly seen from the standpoint of theoretical statisticians, on the potential of Bayesian methods in climate analysis is provided by Berliner et al. (1998).

In frequentist analysis, the approach that most readers are familiar with, assumptions about uncertainty are often hidden. In general, we condition our inferences on many things without being explicit. An example is climate-change detection: the methodology used in most studies (see Section 5) implicitly assumes that our current assessment of potentially important external forcings (greenhouse-gas emissions, sulphur dioxide emissions, solar variability, and volcanic variability) is exhaustive. To be fair, however, we should hasten to add that most authors of climate-change-detection papers do clearly state the limitations of their analyses. Clearly, inferences will only be as good as our assumptions about uncertainty.

We have the impression that the discussion about statistical methodology in the climate sciences is generally not very deep and that straightforward craftsmanship is pursued in many cases. As a consequence, much of the statistical practice in climate science is of a home-grown nature. It is often ad hoc, and is sometimes not well informed or supported by statistical theory. Clearly, the link between climate and the statistical sciences should continue to be improved with additional efforts such as the Geophysical Statistics Project (http://www.cgd.ucar.edu/stats/) at the NCAR. However, such links cannot consist of simple unidirectional transfers of knowledge from the statistical sciences into the climate sciences: statisticians need to collaborate closely with climate scientists to ensure that the advanced methods that they develop will yield new information that can provide new, clear, insights about the climate system and its dynamics.

We feel that the cooperation between the statistical and climate sciences does not function as well as that between, for example, statistical and biomedical science. Among other reasons, this is due to the two particularities mentioned above: confounded dynamics with a large number of degrees of freedom, so that it is not always easy to identify isolated problems amenable to statistical analysis, and the impossibility of generating additional independent data in experiments (apart of numerical experiments with physically based models). Thus, better communication between statisticians and climatologists requires a better understanding by statisticians of the specifics of climate science, and a greater effort by climatologists to communicate the specifics of open problems to statisticians.

One way to overcome these communication problems between the different scientific cultures of statistical and climate sciences is to arrange many more occasions where statisticians can meet climatologists, including meteorologists, oceanographers and other geo-scientists, in a constructive environment. Successful activities along these lines include the International Meetings on Statistical Climatology (Murphy and Zwiers, 1993; also see http://imsc.seos.uvic.ca/), whose ninth meeting will take place in May 2004, in Capetown, South Africa, or the NCAR Geophysical Statistics Project cited above. Useful activities in this respect in the past include the 1993 and 1997 ‘Aha Huliko’a Hawaiian Winter Workshops on ‘Statistical Methods in Physical Oceanography’ and ‘Monte Carlo Simulations in Oceanography’ (Müller and Henderson, 1993, 1997), the 1993 Autumn School ‘Analysis of Climate Variability — Applications of Statistical Techniques’ organized by the Commission of the European Community (von Storch and Navarra, 1999) and the ‘Statistics and Physical Oceanography Report’ (Chelton, 1994; Panel on Statistics and Oceanography, 1994). Many more such activities are required. Greater opportunities for joint research, which means a greater emphasis by funding agencies on cross-disciplinary research that specifically links climatologists and statisticians, is also required.

REFERENCES


Ellis W. 1890. On the difference produced in the mean temperature derived from daily maxima and minima as dependent on the time at which the thermometers are read. *Quarterly Journal of the Royal Meteorological Society* 16: 213–218.


