Statistics and Climate

Peter Guttorp[1,2]

[1]Department of Statistics, University of Washington, Seattle, Washington 98195; email: peter@stat.washington.edu

[2]Norwegian Computing Center, NO-0373 Oslo, Norway

Annu. Rev. Stat. Appl. 2014. 1:87–101

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115648

Keywords

homogenization, nonstationarity, shift function, ranking

Abstract

For a statistician, climate is the distribution of weather and other variables that are part of the climate system. This distribution changes over time. This review considers some aspects of climate data, climate model assessment, and uncertainty estimation pertinent to climate issues, focusing mainly on temperatures. Some interesting methodological needs that arise from these issues are also considered.

1. INTRODUCTION

This review contains a statistician's take on some issues in climate research. The point of view is that of a statistician versed in multidisciplinary research; the review itself is not multidisciplinary. In other words, this review could not reasonably be expected to be publishable in a climate journal. Instead, it contains a point of view on research problems dealing with some climate issues, problems amenable to sophisticated statistical methods and ways of thinking. Often such methods are not current practice in climate science, so great opportunities exist for interested statisticians.

Guttorp (2011) gives some of the background as to why the field of statistics was not involved in the early development of the Intergovernmental Panel on Climate Change (IPCC). Basically, members of the International Statistical Institute (ISI) working in spatial statistics did not consider the field of climate change sufficiently exciting to be worth the bureaucratic effort needed to involve ISI. An increasing number of statisticians are now working on climate problems, but the number participating in writing IPCC reports is not growing. A drastic example is the recent IPCC report on extremes (Field et al. 2012), in which four out of approximately 750 authors and reviewers are statisticians (Georg Lindgren of Sweden, Sylvie Parey of France, David Stephenson of the United Kingdom, and Francis Zwiers of Canada). Extreme value theory is mentioned in just one place in 542 pages of text, whereas 35 out of more than 500 references deal with statistics or probability in any form. Rick Katz at the National Center for Atmospheric Research (NCAR) maintains a website of references to papers on statistics of extremes in climate change (**http://www.isse.ucar.edu/extremevalues/spellbib.html**). Out of 64 papers, 53 are in climate, geophysics, or hydrology journals. According to climate scientists, the reason that more statisticians are not involved in the IPCC reports is that statisticians do not tend to publish in climate science journals, but this claim is not believable.

To discuss climate issues, one should first answer the question, "What is climate?" In the preface to Peixoto & Oort (1992, p. xvii), Edward Lorentz writes,

> Early in the present century a local climate was often considered to be little more than the annual course of the long-term averages of temperature and precipitation. The existence of extensive regions of the globe with reasonably uniform local climates led to the concept of climatic zones. . . .
>
> By the middle of the century some meteorologists had extended the scope of climate to include not simply temperature and precipitation but virtually all atmospheric properties, at upper levels as well as near the earth's surface. To these investigators, climate consisted of the set of all long-term atmospheric statistics, and thus was almost synonymous with the general circulation of the atmosphere. . . .
>
> Within more recent years the concept of a *climate system* has become firmly established. The basis for this view is the realization that the underlying ocean and land surfaces (and the ice, snow, lakes, rivers, and living things that are often found between these surfaces and the atmosphere) are not mere inert boundary conditions, to be taken for granted in seeking explanations for the atmosphere's behavior. On the contrary, they possess their own internal dynamics, and for them the atmosphere is one of the boundary conditions. Together with the atmosphere they form a larger system that may logically be studied as a single entity.

From a statistical point of view, it is appropriate to view the climate as the distribution (changing over time) of climate variables. These include, but as the quotation from Lorentz points out above, are not limited to, weather variables such as temperature and precipitation. The view of the climate as a distribution (and the weather as a random draw from this distribution) allows a statistician to utilize a substantial body of methodology and also indicates some directions of theoretical investigation (of empirical processes of multivariate nonstationary and temporally

dependent observations). I take this point of view throughout this review. It is not one with which my climatology colleagues necessarily agree, as they tend to be empiricists and use definitions such as "climate is . . . the statistical description in terms of the mean and variability of relevant quantities over a period of time" (Solomon et al. 2007, p. 943).

This review discusses data issues, model assessment, and propagation of uncertainty, mainly in the context of temperature data. In the final discussion section, some other problems in climate science that could benefit from statistical work are considered. In another article in this volume, Rougier & Goldstein (2014) discuss climate models.

2. DATA

2.1. Homogenization

Data used in climate research have usually not been collected for climate purposes. Rather, they are intended for weather forecasting, air or sea vessel support, etc. However, using high-quality data to assess changes in climate is important. One would not want to have climate policy decisions seriously affected by nonclimatic circumstances, such as changes in the surroundings of a weather station, changes in instrumentation, changes in instrument location, and urban heat island effects (cf. Trewin 2010).

In the climate community, the approach to dealing with nonclimatic data issues has been to homogenize data sets. For example, if a station has changed site, one estimates a location change (typically a step change) and adjusts the data for this change. The reason for doing so is that climate is thought to be essentially local, so even a moved station would be measuring the same local climate. The intent is to obtain a series of data that is as long and as homogeneous as possible at each site. Homogenization of variance is not generally performed in the climate context. From a statistical point of view, if the intent is to use the data to estimate quantities such as global or regional average temperature, combining many data series, all of different lengths, will be necessary, so having two station locations, one collecting the data until the change and the other the data after the change, is just as reasonable. In fact, this method may provide a better sense of the smoothness of the spatial field. Similarly, if the station has changed instrumentation, the statistical approach would be to change the measurement characteristics associated with the station, rather than trying to homogenize the data. So far, to my knowledge, the only group performing global temperature analysis that has taken this approach is the Berkeley Earth group (Rohde et al. 2013). Such a practice is, however, not uncommon in statistical paleoclimatology (e.g., Li et al. 2010, Tingley et al. 2012), which uses various proxies for, e.g., temperature and greenhouse gases to study the natural variability of climate in the past.

Lund et al. (2007) demonstrate the dangers of using change point methods based on independent and identically distributed (i.i.d.) data when the data are actually autocorrelated. False detection is common, indicating the importance of accepting only shifts that are documented in the station metadata. Unpublished simulation studies indicate that the consequences of long-term memory are even worse, in that the proper critical value for a test for a shift in a time series with moderate long-term memory is an order of magnitude larger than for i.i.d. data.

2.2. Comparison of Databases

The core measurement used to illustrate global warming is the mean daily temperature. The Berkeley Earth data set (**http://berkeleyearth.org/dataset/**) contains measurements from some 36,000 stations, gathered from 16 different databases (see **Table 1** for the main ones). There are three main global average temperature series, each calculated somewhat differently and using

Table 1 Climate databases

Database	Acronym	Source	Number of stations
Colonial Era Weather Archive	CA	US National Climatic Data Center	1,243
Global Historical Climatology Network: Daily	GHCN	US National Climatic Data Center	15,069
Global Historical Climatology Network: Monthly	GHCN	US National Climatic Data Center	7,280
Global Climate Observing System (GCOS) Surface Network	GSN	US National Oceanic and Atmospheric Administration	1,018
Global Summary of the Day	CLIMVIS	US National Oceanic and Atmospheric Administration	20,000
Hadley Centre Climate Research Unit	HadCRU	UK Met Office	5,583
Monthly Climate Data for the World	MCDW	US National Oceanic and Atmospheric Administration	2,646
Scientific Committee on Antarctic Research	SCAR	British Antarctic Survey	46
US Cooperative Summary of the Day	UCSD	US National Climatic Data Center	8,500
US Cooperative Summary of the Month	UCSM	US National Climatic Data Center	23,000
US First Order Summary of the Day	N/A	US National Climatic Data Center	1,200
US Historical Climatology Network	USHCN	US National Center for Atmospheric Research	1,218
World Monthly Surface Station Climatology	WMSSC	US National Center for Atmospheric Research	4,700
World Weather Records	WWR	US National Climatic Data Center	1,930

Abbreviation: N/A, not available.

different sets of stations. The current series are HadCRUT4 from the UK Met Office Hadley Centre and the University of East Anglia Climate Research Unit, covering 1850 to the present; the Goddard Institute for Space Studies (GISS) surface temperature analysis (GISTEMP) version 3 series, covering 1880 to the present; and the National Climatic Data Center (NCDC) global annual anomalies, also covering 1880 to the present. Each of these analyses uses ad hoc approaches to estimate global averages and somewhat dubious approaches (if any) to estimate the uncertainty of the global mean estimates. The Berkeley Earth project produces only a land average temperature series, covering 1753–2011, using spatial statistics and unhomogenized data, and the group's estimates of uncertainty are by far the best from a statistical point of view.

Some of these stations have measured minimum and maximum temperature, whereas others have measured hourly temperature (sometimes taking three measurements per day) or used alternative observational schemes. Somehow, these measurements will need to be combined into an estimate of mean daily temperature at the site. The World Meteorological Organization (WMO 2010) suggests averaging the minimum and maximum daily temperature, if available. If a continuous record of daily temperature were well described by a shifted sine curve [which is an approximate model of the solar radiation that reaches the surface on a cloud-free day (Sampson & Guttorp 1992)], the average of the minimum and the maximum would indeed be the daily mean, modulo measurement error. But as **Figure 1** shows, a sine curve is not a particularly good description of the monthly average of one-minute-resolution temperature measurements taken at the air traffic control tower at Visby airport, Sweden, in January 2010.

Different countries use different approaches to estimate daily mean temperature from observed data, usually taken more than once a day. Sweden, for example, uses a weighted average of

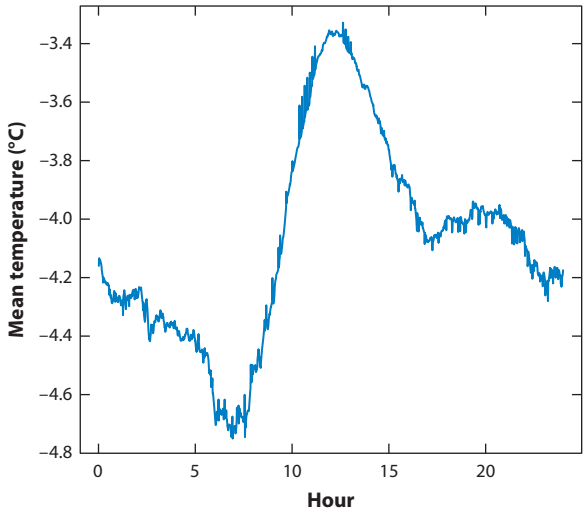


Figure 1

Average temperature by minute of the day for the Visby air traffic control tower. Data are taken from the Swedish Meteorological and Hydrological Institute.

observations taken at three hours during the day, as well as the minimum and the maximum. The weights depend on longitude and month. The other Nordic countries use other linear combinations of measurements, sometimes including the minimum and maximum. The United States, the United Kingdom, and Australia are among the countries reporting the average of the minimum and maximum. There are issues with what hours are used to define a climate day. This definition varies from country to country, and sometimes within a country. Ma & Guttorp (2013) make some uncertainty comparisons among different methods. Perhaps databases should contain estimates of uncertainty for each measurement (so to speak) of daily mean temperature. At the very least, the databases should indicate what method of estimation has been used.

The International Surface Temperature Initiative (**http://www.surfacetemperatures.org/**) is a WMO-sponsored project aiming to produce a transparent data bank in which the raw data and every modification of the data, as well as all available metadata, are clearly documented. In addition, the project aims to benchmark and assess methods of, e.g., homogenization and spatial estimation. The initiative involves statisticians, metrologists, and climate scientists. The first issue of the database, published in May 2013, contains only unhomogenized temperature data, but the plan is to extend the database to other climate variables, such as precipitation and sea level pressure. The intent is to produce a common high-quality database that everyone will use.

The ongoing refinement of climate data procedures (e.g., cleaning and homogenization) leads to changes in databases, which in turn yield changes in often-used products. For example, the NCDC database US Historical Climatology Network (USHCN) used to estimate monthly mean temperature for the continental United States (CONUS) was updated in October 2012. **Figure 2** shows the difference (color coded by month) between the mean monthly continental temperature estimates based on version 2.5 compared with version 2.0 of the USHCN data set. This difference emphasizes the importance of being careful in stating which version of the data is used in an analysis. To make research reproducible, keeping previous versions of data sets available for users is also important.

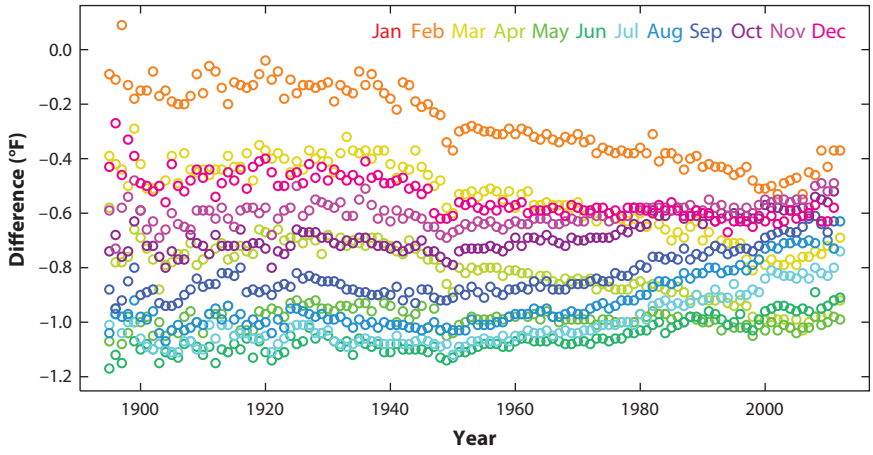


Figure 2

Differences between two versions of US Historical Climatology Network data on continental US monthly mean temperature, color coded by month. The later version (2.5) tends to be colder than the earlier version (2.0).

3. UNCERTAINTY

3.1. Ranking

The idea of ranking years based on temperature is quite common, particularly in the press. The uncertainty of ranks, however, is rarely mentioned. But if we rank years on their annual global average temperature, we must take into account the uncertainty in the estimate of global average temperature used in the ranking. How can we do that? The International Surface Temperature Initiative suggested that a reasonable way of describing uncertainty in data products such as global average temperature is to somehow generate an ensemble of possible realizations of these estimates. Users can then utilize these realizations to estimate, for example, the rank in each path. The spread in the distribution of ranks (over the realizations in the ensemble) is then a way to describe the uncertainty in these ranks.

For a statistician, simulation is an easy way to generate this ensemble. The simplest approach is to generate paths with annual means equal to the estimated values and standard deviations equal to the standard error of the estimate. Of course, temporal dependence likely exists in the temperature series, and one can model that before simulating the ensemble. In Guttorp & Kim (2013), this kind of modeling is done for CONUS annual mean temperature, using uncertainty estimates and data from Shen et al. (2012) for 1897–2008 (based on the USHCN database). **Figure 3** shows the results for 1993–2000: 1996 and 1997 have very uncertain ranks, whereas 1993 is among the bottom third, and 1998 is one of the warmest years, with a 0.6 probability of being the warmest. Most years have a very low probability of being the warmest. The time series is simulated using a permutation bootstrap of the estimated innovations from an ARMA(3,1) (autoregressive moving-average) model.

3.2. Trends

Changes in the mean climate are generally described using trend lines. Although fitting a straight line to the data (the thick black line in **Figure 5**) may be a somewhat dubious exercise because of the

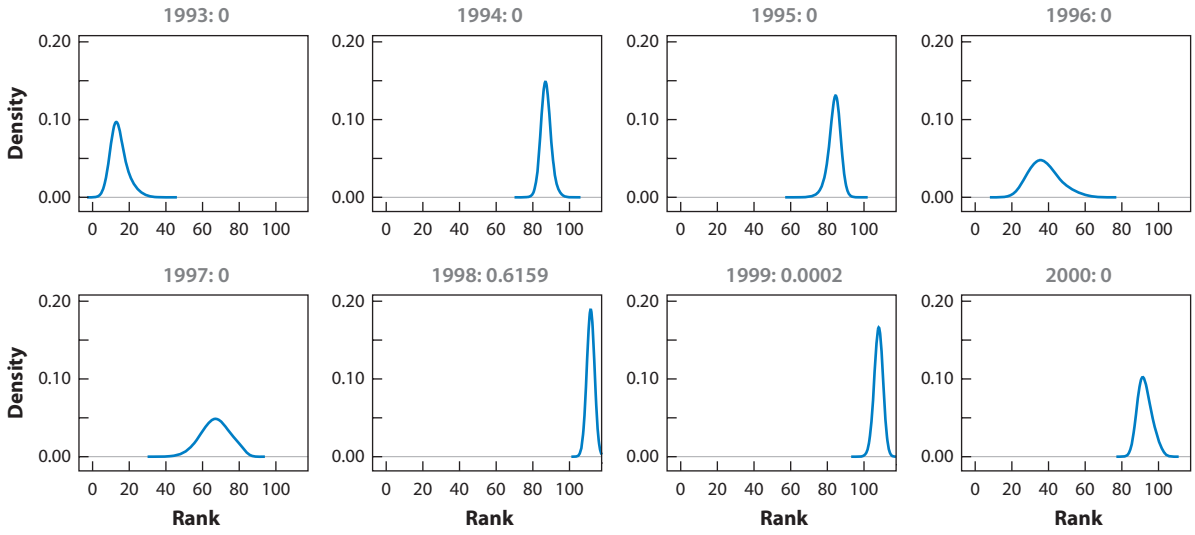


Figure 3

Rank distribution (in which a higher rank means warmer temperatures) by year from 100,000 simulations of a time series model estimated from the Shen et al. (2012) data and standard errors for continental US annual average temperatures. The numbers in the headings are the year followed by the frequency of the highest rank.

apparent nonlinearity, a fitted slope has the advantage that it can be described in observation units per century (or whatever time period is of interest). However, the statistical significance of the slope is often a quantity of discussion. For example, in a live interview, a climate scientist was asked if the slope of a line to the HadCRUT3 annual global temperature series for the period 1995–2009 significantly differed from zero. The answer was no (in fact, this particular time period was the longest for which this lack of significant difference was the case). Some interpreted this finding to mean that global warming was no longer taking place [the prosecutor's fallacy (Thompson & Schumann 1987)]. Of course, fitting a line to 16 data points is a rather uncertain enterprise. Furthermore, if one accepts a linear fit but attempts it for a longer time period, one still has to take account of the temporal dependence in the series as well as the varying uncertainty (which partly depends on how many weather stations, ships, buoys, and sensor floats are available to estimate the global average temperature). Thus, ordinary least squares should be replaced by generalized least squares, typically yielding a higher uncertainty of the slope estimate. As an example, fitting a trend to the CONUS annual average temperature for 1897–2008 (using the same data as in the previous subsection) yields the results in **Table 2**. As the modeling becomes more sophisticated, the significance of the trend decreases. Commonly, climate scientists model dependence using an

Table 2 Regression slopes

Method	Slope (°C/year)	Standard error	Sign
OLS	0.0055	0.0012	***
WLS	0.0048	0.0014	***
GLS (AR1)	0.0048	0.0018	*
GLS [ARMA(3,1)]	0.0059	0.0032	–

Abbreviations: AR, autoregressive; ARMA, autoregressive moving-average; GLS, generalized least squares; OLS, ordinary least squares; WLS, weighted least squares.

AR(1) model. In this case, an ARMA(3,1) model is a substantially better fit according to Akaike information criterion (AIC) and residual white-noise tests, at the cost of no longer having a significant trend (Guttorp & Kim 2013).

Temperature data tend to have a rather complicated correlation structure, owing to a variety of influences ranging from the short-term influence of cold fronts and warm fronts, to the multiannual influence of the El Niño–Southern Oscillation (ENSO), to the decadal influences of the Pacific Decadal Oscillation and the North Atlantic Oscillation, to the slow centennial-scale movement of water masses from the deep ocean to the surface. Smith (1993) notes that a time series with long-term memory can exhibit long stretches of increasing data without having an actual trend. He fits a regression model in which the noise exhibits long-term memory to a global temperature series and finds that the slope is still significant, indicating that the observed increase in global temperature cannot be explained simply as a spurious trend due to the long-term memory character of the data. Foster & Rahmstorf (2011) use ENSO, volcanic eruptions, and solar variability as covariates to explain part of the variability of a global monthly time series, thereby getting stronger evidence of trends. The covariates do not explain all the temporal dependence, as some correlation still remains in the residuals, which the authors describe with an ARMA(1,1) model.

When looking at a network of temperature stations, a common approach is to fit each station separately, either using ordinary least squares or employing a correction for first-order auto-correlation. The estimate, normalized by its estimated standard error, is then compared to a t-distribution to assess significance. A few researchers have realized the need to model the spatial correlation and to consider the multiple comparison problem. Sometimes data are put on a grid, and trend estimates for each grid square are used separately to assess the spatial coherence of the test statistics. A more natural approach is to use a hierarchical model that simultaneously estimates trend and seasonality and allows for complex spatiotemporal correlation structures. One example is Craigmile & Guttorp (2011), in which a combination of short- and long-term memory is used, and nonlinear trends are estimated in wavelet space.

3.3. Comparing Models with Data

An atmospheric climate model is a numerical solution of a set of coupled partial differential equations, describing the fluid dynamics of the atmosphere (Peixoto & Oort 1992). In addition, the model may be coupled to a similar description of oceans, land use, and the cryosphere (ice and snow). One can compare climate models with data in many ways. One approach, which is popular in the atmospheric science community, is to do a principal component analysis of both climate model fields and data fields and try to regress one on the other. In this section, I stick to the simpler problem of looking just at global mean temperature (not the most sensitive indicator of climate change, but one that is used a lot). Climate models are not trying to recreate weather data, although one might get that impression from the literature. In fact, **Figure 4** shows the actual output of 38 climate models in the Coupled Model Intercomparison Project Phase 5 (CMIP5) experiment (Taylor et al. 2012). The models simulated the historical record (up until 2000), all using the same set of historical forcings from, e.g., solar radiation, greenhouse gases, and volcanic eruption. The CMIP5 models also run simulations for the time after the year 2000 (up to 2100 or, in some cases, 2300), using a set of scenarios called representative concentration pathways to produce projections of future values of climate variables (estimates of future values are conditional on such scenarios of the future behavior of Earth's population and are therefore called projections instead of predictions). This review does not consider projections.

The different models in **Figure 4** do not yield the same temperatures for any given year. This discrepancy has to do with how the model runs are initiated, how long they run in prehistorical

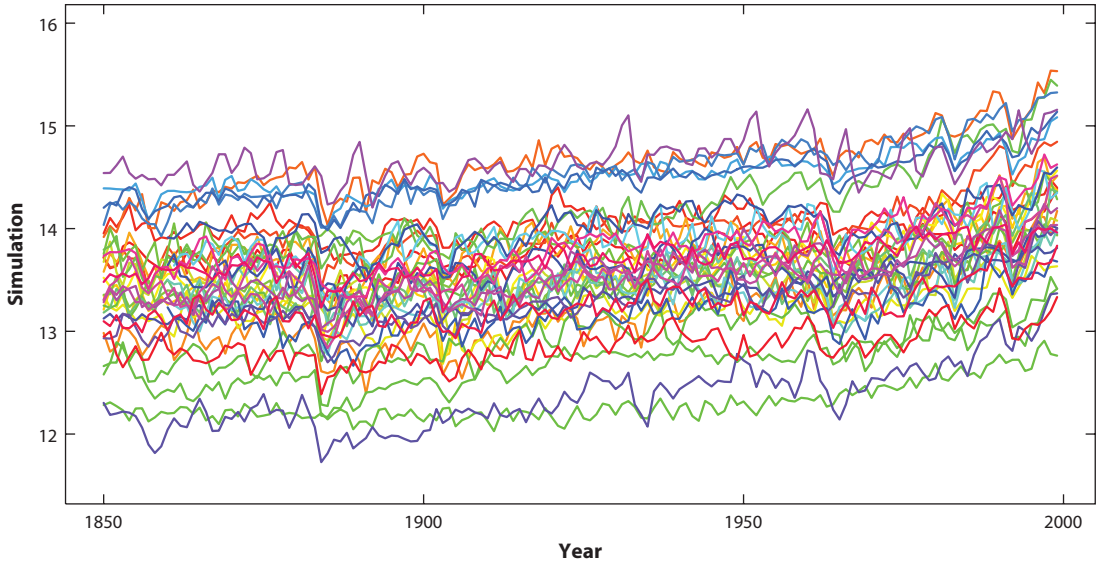


Figure 4

Global average temperatures for 1850–1999 from 38 models in the Coupled Model Intercomparison Project Phase 5 (CMIP5).

mode, and how precisely they model the interaction of the atmosphere with oceans. To make them more comparable, researchers commonly calculate anomalies, i.e., residuals compared with some fixed time period (here 1951–1980). **Figure 5** shows anomalies of the same 38 models, as well as the GISTEMP version 3 observed global annual mean temperature anomalies as estimated by GISS (Hansen et al. 2010), using the same baseline period as for the models. The vertical lines correspond to the two time periods used in **Figure 6** to compare distributions.

Based on the idea of using the distribution of weather data as an estimate of climate, a natural alternative would be to compare reasonably long stretches of the distribution of the two types of data, instead of comparing time series of anomalies. The WMO generally suggests that 30 years is reasonable, so I do comparisons on that timescale. **Figure 6** shows quantile-quantile (Q-Q) plots (Wilk & Gnanadesikan 1968) for two 30-year periods (1930–1959, during which temperatures were reasonably stable, and 1970–1999, during which they were increasing). The 95% confidence intervals are obtained from the asymptotic distribution of the Kolmogorov–Smirnov statistic and are simultaneous under the (dubious) assumption of independence in time. In the left column, the entire ensemble of model values is used. In the middle and right columns, two particular models, the CCSM4 model from NCAR in the United States (Gent et al. 2011) and the HadCM3 model from the Hadley Centre in the United Kingdom (Gordon et al. 2000), are chosen because the former seemed to fit the middle two quartiles of the data best in the earlier period (before substantial warming had started to appear), whereas the latter was one of the worst fits. Interestingly, the roles of the models are reversed in the second period, after substantial warming appeared in the data (cf. **Figure 7**). Both models overestimate the warming compared with the observations, whereas the whole ensemble of climate models seems to agree well with the data, although the entire range of their output values exceeds the range of the data (which should not be surprising, as the sample size is 39 times larger).

Can we see a change in climate between the two time periods? **Figure 7** uses Doksum's (1974) shift function to investigate this issue, both in the models and in the data. If the horizontal line at level 0 falls entirely within the confidence bands (Doksum & Sievers 1976), then no significant

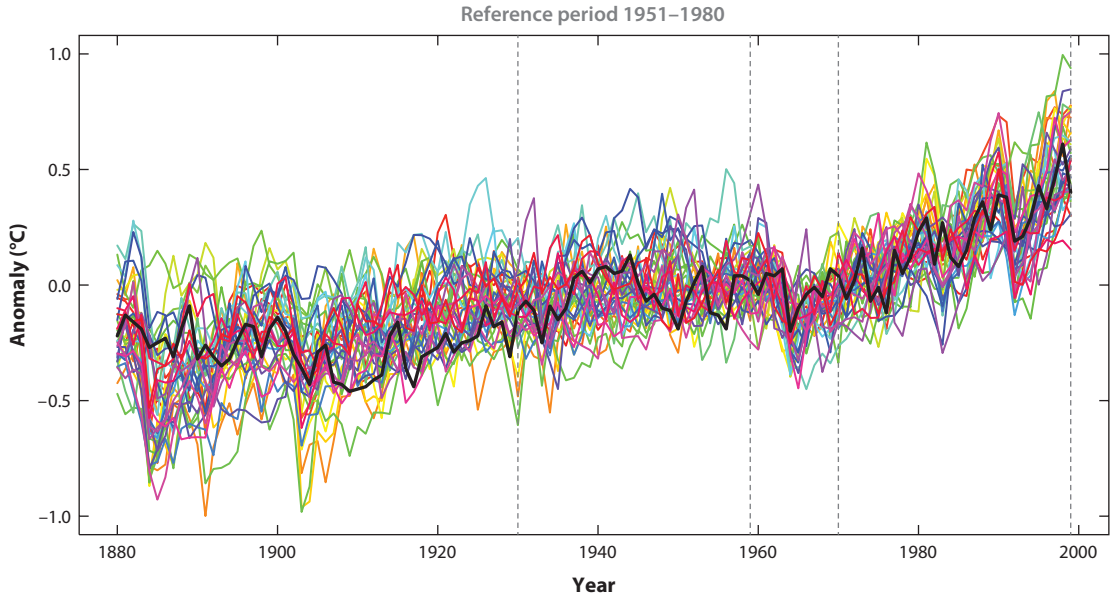


Figure 5

Global average temperature anomalies (*colored paths*) for 1880–1999 relative to the reference period 1951–1980 from 38 models in the Coupled Model Intercomparison Project Phase 5 (CMIP5). The thick black line is the Goddard Institute for Space Studies estimate of global mean temperature. The dashed vertical lines are time periods used in **Figure 6** to compare distributions of models and data.

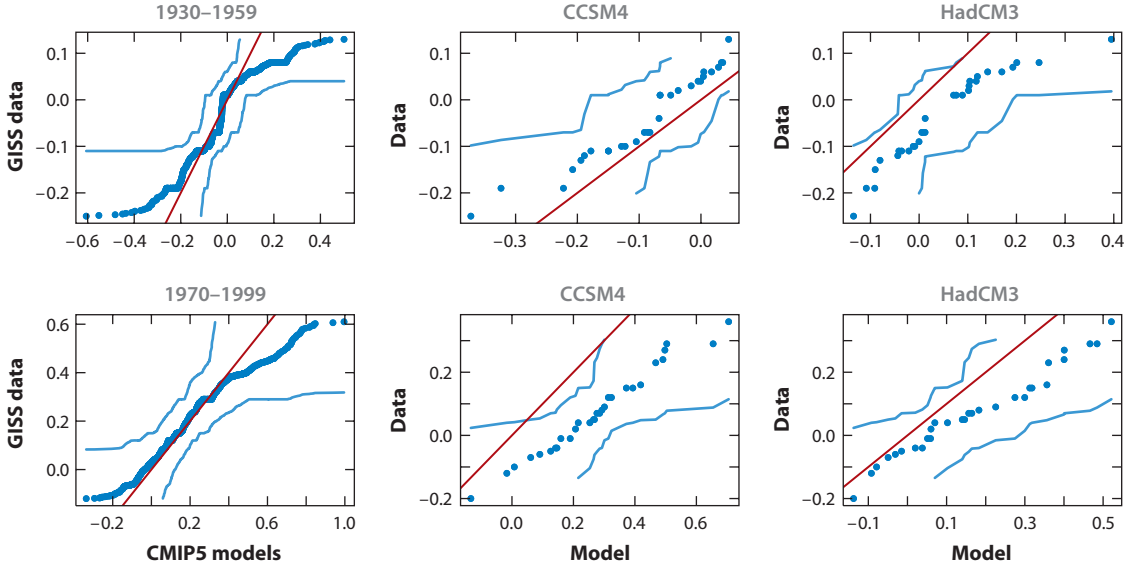


Figure 6

(*Left column*) Quantile-quantile (Q-Q) plots of Goddard Institute for Space Studies (GISS) data against the ensemble of the Coupled Model Intercomparison Project Phase 5 (CMIP5) simulations for 1930–1959 (*top row*) and 1970–1999 (*bottom row*). (*Middle* and *right columns*) Q-Q plots of the data against particular models: CCSM4 from the National Center for Atmospheric Research (NCAR) and HadCM3 from the Hadley Centre. The red lines are lines of equal distribution.

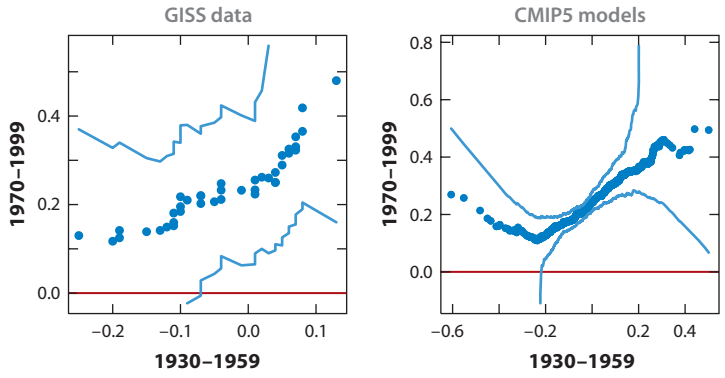


Figure 7

Shift function estimates with 95% simultaneous confidence bands under the assumption of independence in time and independence between models for data (*left*) and the ensemble of the Coupled Model Intercomparison Project Phase 5 (CMIP5) models (*right*) between the time periods 1930–1959 and 1970–1999. The horizontal red line corresponds to equal distribution for the two time periods. Abbreviation: GISS, Goddard Institute for Space Studies.

difference exists between the sets of data. A nonzero horizontal line that fits in the band indicates a location shift, whereas a nonhorizontal line that fits indicates location-scale change. Nonlinear curves correspond to more complicated location-dependent shifts. The figure clearly indicates location-scale changes between the two time periods for the data as well as for the ensemble of models. In both cases, part of the horizontal line at level 0 falls outside of the confidence band, indicating a significant change between the periods.

The confidence bands in the Q-Q and shift function plots above are asymptotic bands based on the Kolmogorov–Smirnov distribution for i.i.d. data. Because many climate series have serial dependence and are nonstationary (in fact, this defines climate change from a probabilistic point of view), developing tools that deal with empirical processes of nonstationary dependent data is necessary. Although the asymptotics for i.i.d. data lead to a Brownian bridge process, dependent data can result in other Gaussian (Bachmann & Dette 2005, Dehling et al. 2009) or non-Gaussian processes (Ould Haye & Philippe 2011), depending on whether the dependence is short- or long-term. When the data are not stationary, work going at least as far back as Shorack (1973) indicates that the appropriate quantity of interest, estimated by the empirical distribution function, is the average distribution function over time. In the context of climate data, this average distribution becomes our statistical estimate of climate. Developing simultaneous confidence bands for the average distribution function or for the marginal density of time series data entails finding computational approaches to suprema for Gaussian processes [using tools such as in Cierco-Aroylle et al. (2003) and Åberg & Guttorp (2008)] or Rosenblatt processes (Veillette & Taqqu 2010, Taqqu 2011), depending on the degree of dependence in the data. For multivariate extensions, the standard theory leads to approximations by a Kiefer process (Philipp & Pinzur 1980), and extensions of this theory to dependent situations exist (e.g., Rüschendorf 1974).

3.4. Uncertainty Propagation

Consider a city council in a coastal city, trying to determine if sea level change is going to affect its planning policies. It will need a local projection of the future sea level. What it needs to do is the following:

- Relate global mean sea level change to global mean temperature and temperature change (Vermeer & Rahmstorf 2009).
- Relate global sea level change to local sea level change using data from tide gauges, correcting for land surface changes due to, e.g., glacial rebound and tectonic activity (Tebaldi et al. 2012).
- Look at temperature projections under different scenarios (Taylor et al. 2012) to project global sea level rise.
- Apply the local relationship to the projected global sea level rise.

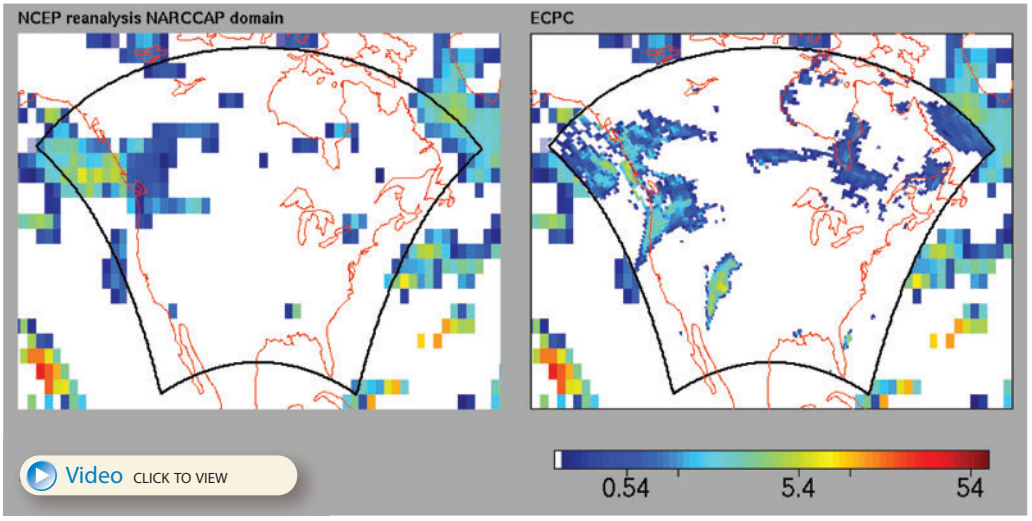
This process is sometimes called statistical downscaling. Of course, each step in this chain of calculations has uncertainties associated with it. Thus, although many regional projections (e.g., Mote et al. 2008) simply give a single number for a particular location and scenario, a statistician would prefer to carry the uncertainty all the way through the calculations, using perhaps an ensemble of different projections for each scenario combined with random draws from the two regression models (global and local) to create a likely range of possible sea level rise outcomes in the given location. That this procedure is not generally carried out can have serious consequences for local planning efforts. An example of such a propagation of uncertainty can be found at **http://courses.washington.edu/statclim/what.html**.

4. DISCUSSION

One aspect of statistics that has found surprisingly little application in climate science is the design of experiments. Most scientific simulation studies use a wasteful full factorial design. In climate science, an exception is the North American Regional Climate Change Assessment Program (NARCCAP) experiment (**http://www.narccap.ucar.edu/index.html**), intended to study differences between regional climate models (climate models run on fine resolution on a bounded region such as North America) and the global models that provide boundary conditions for the regional models. **Video 1** shows the global model outside the North American domain and the regional model inside it. The experiment uses eight regional models and four global models, employing a fractional factorial design, which reduces the necessary runs by a factor of two or doubles the number of models that can be considered. The main project runs 30 years using historical simulations and 30 years of future projections for each design combination of regional and global models. The design enables estimation of the interaction between regional and global models, which is often the quantity that carries the most information (see Sain et al. 2011 for an ANOVA-type analysis of this kind of experiment).

In an important work, Lindgren et al. (2011) suggest using a Markov random field approach to estimate global temperature. This approach would enable a proper spatial estimation of a nonstationary field on the globe, an estimation that no global temperature group has done so far. An advantage of the approach would be the ease of generating ensembles of temperature reconstructions (not only for the global average). Because of the speed of the INLA software (Rue et al. 2009) used to fit the models in this approach, carrying out experiments with various subsets of temperature stations and ocean data would also be straightforward. Inclusion of satellite data into the analysis is (at least in principle) simple as well.

Of course, many other aspects of climate science could benefit from additional statistical input. For a closing example, consider the definition of a season. First of all, not all parts of the world have four distinct seasons. The simple division of a year into groups of three months is commonly used when looking at seasonal aspects of model fit. But the method is of no use if the interest is in studying changes in seasons (Alpert et al. 2004, Trenberth 1983). Different meteorological



Video 1

A global climate model (*left*) and a regional model (*right*) using precipitation output with boundary conditions from the global model. The data are taken from the North American Regional Climate Change Assessment Program (NARCCAP) experiment. Movie created by Douglas Nychka and Stephan Sain of the National Center for Atmospheric Research (NCAR). To view the video, access this article on the Annual Reviews website at **http://www.annualreviews.org**.

services have their own definition of seasons. For example, in Sweden, the definition (translated from **http://www.smhi.se/kunskapsbanken/meteorologi/arstider-1.1082**) is the following:

> Winter is the period when the daily mean temperature permanently is below 0°C, and summer when it is permanently above 10°C. By studying means over several years one gets a smoothed temperature curve, and it is then not hard to find dates when the different temperature limits are passed.
>
> But for a particular year it may be debatable as to what constitutes "permanently." The present rule is seven days for spring, and five for other seasons. The season is said to begin at the first of these days. However, spring cannot start before February 15, and autumn must end by February 14. The earliest possible autumn is August 1, and the latest possible spring is July 31.

Thus, based on this definition, one can have years with no winter or no summer. The statistical issue here is one of constrained clustering. Seeing what answers different clustering techniques yield would be interesting.

Many issues exist in addition to those raised in this review. Statisticians should be able to contribute considerably to issues such as the visual comparison of spatial fields, nonparametric estimation of nonlinear trends in data with multiple scales of dependence, stochastic models for downscaling climate models to regional and local scales, treatment of spatiotemporal extremes in data and in climate models, studies of public health and ecological effects of a changing climate, and much more. This scientific endeavor can benefit considerably from statistical contributions because data are abundant and the potential societal impact is great.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was partially supported by the US National Science Foundation grant DMS-1106862, the Research Network for Statistical Methods for Oceanic and Atmospheric Sciences (STATMOS). I am grateful to the Swedish Meteorological and Hydrological Institute for access to the minute-resolution radiation measurements from Visby; the World Climate Research Program's Working Group on Coupled Modelling, which is responsible for CMIP, and the climate modeling groups producing and making available the CMIP5 model output used in Section 3.3; GISS for the GISTEMP series; the US NCDC for providing the CONUS annual mean temperature data; and Samuel Shen for making the data and standard errors for his paper (Shen et al. 2012) available. I have benefited from discussions with and advice from Doug Nychka and Claudia Tebaldi over many years. I also want to thank an anonymous reviewer, Peter Craigmile, and Jonathan Rougier for helpful comments on an earlier draft.

LITERATURE CITED

Åberg S, Guttorp P. 2008. Distribution of the maximum in air pollution fields. *Environmetrics* 19:183–208

Alpert P, Osetinsky I, Ziv B, Shafir H. 2004. A new seasons definition based on classified daily synoptic systems: an example for the eastern Mediterranean. *Int. J. Climatol.* 24:1013–21

Bachmann D, Dette H. 2005. A note on the Bickel-Rosenblatt test in autoregressive time series. *Stat. Probab. Lett.* 74:221–34

Cierco-Aroylle C, Croquette A, Delmas C. 2003. Computing the distribution of the maximum of Gaussian random processes. *Methodol. Comput. Appl. Probab.* 5:427–38

Craigmile PF, Guttorp P. 2011. Space-time modeling of trends in temperature data. *J. Time Ser. Anal.* 32:378–95

Dehling H, Durieu O, Volny D. 2009. New techniques for empirical processes of dependent data. *Stoch. Process. Appl.* 119:3699–718

Doksum KA. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Stat.* 2:267–77

Doksum KA, Sievers GL. 1976. Plotting with confidence: graphical comparisons of two populations. *Biometrika* 63:421–34

Field CB, Barros V, Stocker TF, Qin D, eds. 2012. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge, UK: Cambridge Univ. Press

Foster G, Rahmstorf S. 2011. Global temperature evolution 1979–2010. *Environ. Res. Lett.* 6:044022

Gent PR, Danabasoglu G, Donner LJ, Holland MM, Hunke EC, et al. 2011. The Community Climate System Model version 4. *J. Clim.* 24:4973–91

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, et al. 2000. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dyn.* 16:147–68

Guttorp P. 2011. The role of statisticians in international science policy. *Environmetrics* 22:817–25

Guttorp P, Kim TY. 2013. Uncertainty in ranking the hottest years of US surface temperatures. *J. Clim.* 26:6323–28

Hansen J, Ruedy R, Sato M, Lo K. 2010. Global surface temperature change. *Rev. Geophys.* 48:RG4004

Li B, Nychka DW, Ammann CM. 2010. The value of multi-proxy reconstruction of past climate (with discussions and rejoinder). *J. Am. Stat. Assoc.* 105:883–911

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. B* 73:423–98

Lund R, Wang XL, Lu Q, Reeves J, Gallagher C, Feng Y. 2007. Changepoint detection in periodic and autocorrelated time series. *J. Clim.* 20:5178–90

Ma Y, Guttorp P. 2013. Estimating daily mean temperature from synoptic climate observations. *Int. J. Climatol.* 33:1264–69

Mote P, Petersen A, Reeder S, Shipman H, Binder LW. 2008. *Sea level rise in the coastal waters of Washington State*. Tech. Rep., Univ. Wash. Clim. Impacts Group, Seattle, WA

Ould Haye M, Philippe A. 2011. Marginal density estimation for linear processes with cyclical long memory. *Stat. Probab. Lett.* 81:1354–64

Peixoto JP, Oort AH. 1992. *Physics of Climate*. New York: Am. Inst. Phys.

Philipp W, Pinzur L. 1980. Almost sure approximation theorems for the multivariate empirical process. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* 54:1–13

Rohde R, Muller R, Jacobsen R, Perlmutter S, Rosenfeld A, et al. 2013. Berkeley Earth temperature averaging process. *Geoinform. Geostat. Overv.* 1:2

Rougier J, Goldstein M. 2014. Climate simulators and climate projections. *Annu. Rev. Stat. Appl.* 1:103–23

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. B* 71:319–92

Rüschendorf L. 1974. On the empirical process of multivariate, dependent random variables. *J. Multivar. Anal.* 4:469–78

Sain S, Nychka D, Mearns L. 2011. Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics* 22:700–11

Sampson PD, Guttorp P. 1992. Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Stat. Assoc.* 87:108–19

Shen SSP, Lee SK, Lawrimore J. 2012. Uncertainties, trends, and hottest and coldest years of U.S. surface air temperature since 1895: an update based on the USHCN V2 TOB data. *J. Clim.* 25:4185–203

Shorack G. 1973. Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-i.i.d. case. *Ann. Stat.* 1:146–52

Smith RL. 1993. Long-range dependence and global warming. In *Statistics for the Environment*, ed. V Barnett, F Turkman, pp. 141–61. Chichester, UK: Wiley

Solomon S, Qin D, Manning M, Chen Z, Marquis M, et al., eds. 2007. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge Univ. Press

Taqqu MS. 2011. The Rosenblatt process. In *Selected Works of Murray Rosenblatt*, ed. RA Davis, K-S Lii, DN Politis, pp. 29–45. New York: Springer

Taylor KE, Stouffer RJ, Meehl GA. 2012. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* 93:485–98

Tebaldi C, Strauss BH, Zervas CE. 2012. Modelling sea level rise impacts on storm surges along US coasts. *Environ. Res. Lett.* 7:014032

Thompson WC, Schumann EL. 1987. Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. *Law Hum. Behav.* 11:167–87

Tingley M, Craigmile P, Haran M, Li B, Mannshardt E, Rajaratnam B. 2012. Piecing together the past: statistical insights into paleoclimatic reconstructions. *Q. Sci. Rev.* 35:1–22

Trenberth K. 1983. What are the seasons? *Bull. Am. Meteorol. Soc.* 64:1276–82

Trewin B. 2010. Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change* 1:490–506

Veillette M, Taqqu MS. 2010. A technique for computing the PDFs and CDFs of nonnegative infinitely divisible random variables. *J. Appl. Probab.* 48:217–37

Vermeer M, Rahmstorf S. 2009. Global sea level linked to global temperature. *Proc. Natl. Acad. Sci. USA* 106:21527–32

Wilk MB, Gnanadesikan R. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55:1–17

WMO. 2010. *Guide to Climatological Practices*. Geneva, Switz.: World Meteorol. Organ. 3rd ed.