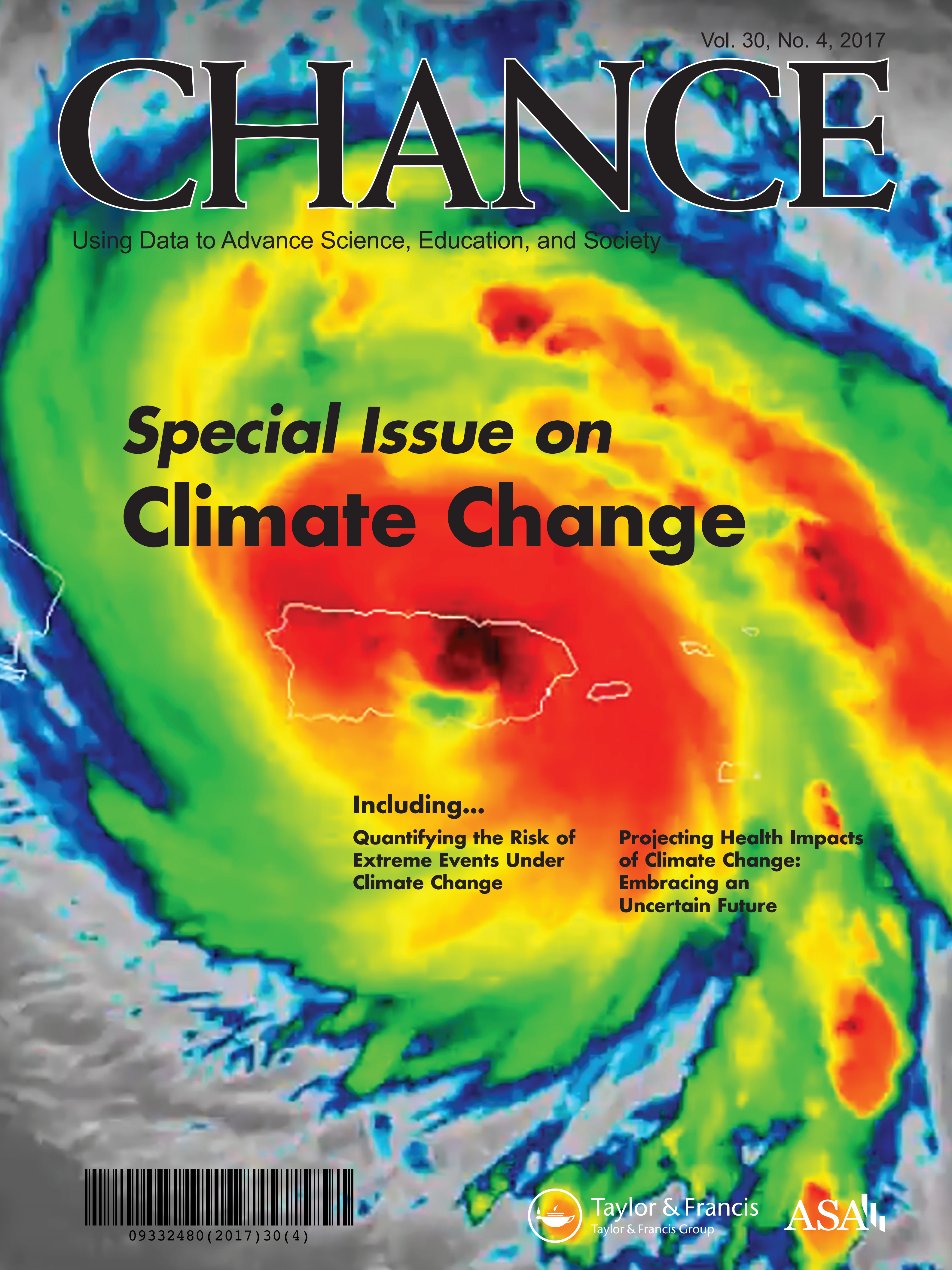


Vol. 30, No. 4, 2017

CHANCE

Using Data to Advance Science, Education, and Society

Special Issue on Climate Change

Including...

Quantifying the Risk of Extreme Events Under Climate Change

Projecting Health Impacts of Climate Change: Embracing an Uncertain Future

09332480(2017)30(4)

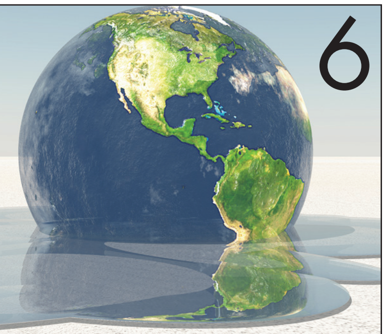
Taylor & Francis
Taylor & Francis Group

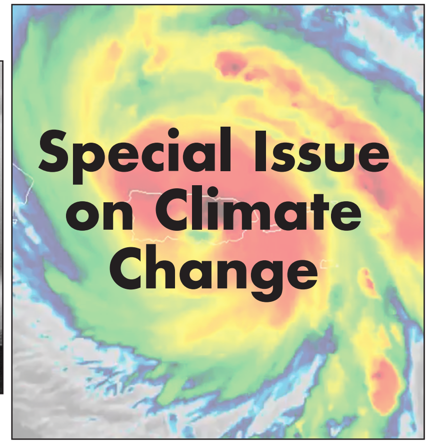
ASA

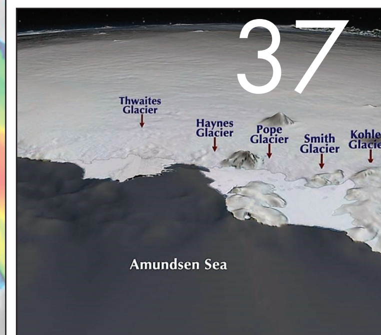
CHANCE

Using Data to Advance Science, Education, and Society

http://chance.amstat.org







ARTICLES

COLUMNS

DEPARTMENTS

Abstracted/indexed in Academic OneFile, Academic Search, ASFA, CSA/Proquest, Current Abstracts, Current Index to Statistics, Gale, Google Scholar, MathEDUC, Mathematical Reviews, OCLC, Summon by Serial Solutions, TOC Premier, Zentralblatt Math.

Cover image courtesy of NOAA. NOAA's GOES16 satellite captured this colorized-infrared imagery of Hurricane Maria over Puerto Rico on September 20, 2017.

WEBSITE

http://chance.amstat.org

AIMS AND SCOPE

CHANCE is designed for anyone who has an interest in using data to advance science, education, and society. *CHANCE* is a non-technical magazine highlighting applications that demonstrate sound statistical practice. *CHANCE* represents a cultural record of an evolving field, intended to entertain as well as inform.

SUBSCRIPTION INFORMATION

CHANCE (ISSN: 0933-2480) is co-published quarterly in February, April, September, and November for a total of four issues per year by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA, and Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

U.S. Postmaster: Please send address changes to *CHANCE*, Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

ASA MEMBER SUBSCRIPTION RATES

ASA members who wish to subscribe to *CHANCE* should go to ASA Members Only, *www.amstat.org/membersonly* and select the "My Account" tab and then "Add a Publication." ASA members' publications period will correspond with their membership cycle.

ANNUAL SUBSCRIPTION, VOLUME 30, 2017

Print ISSN - 0933-2480, Online ISSN - 1867-2280. Institutional subscribers: $128/£78/€102. Personal subscribers: $58/£35/€47.

SUBSCRIPTION OFFICES

USA/North America: Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: (215) 625-8900; Fax: (215) 207-0050. UK/Europe: Taylor & Francis Customer Service, Sheepen Place, Colchester, Essex, CO3 3LP, United Kingdom. Telephone: +44-(0)-20-7017-5544; fax: +44-(0)-20-7017-5198.

OFFICE OF PUBLICATION

American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA. Telephone: (703) 684-1221. Editorial Production: Megan Murphy, Communications Manager; Valerie Nirala, Publications Coordinator; Ruth E. Thaler-Carter, Copyeditor; Melissa Gotherman, Graphic Designer.
Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: (215) 625-8900; Fax: (215) 207-0047.

RESPONSIBLE FOR ADVERTISEMENTS

Brian Evans
Telephone: (215) 606-4311. Email: *brian.evans@taylorandfrancis.com*.
Printed in the United States on acid-free paper.

EDITOR'S LETTER

Scott Evans

Dear CHANCE *Colleagues,*

The United States is still feeling the effects of three major hurricanes this year. Hurricane Harvey brought record rainfall to Houston. Hurricane Irma ravaged Florida. Hurricane Maria hit Puerto Rico, causing more than 50 deaths, and months later, most of Puerto Rico is still without power, many people lack access to clean drinking water, and suspected cases of leptospirosis (a bacterial infection) are rising.

Climate change is the environmental challenge of our and the next generation. Although there are deniers, NASA and climatology scientists have described the compelling evidence of climate change.

Climate change is the theme of this special issue of *CHANCE*. Here are some of the reasons why. Global temperatures are rising, with 2016 the warmest year on record and the third year in a row with record-setting surface temperatures. Eight of the 12 months were the warmest on record for those respective months.

Earth's average surface temperature has risen about 2.0/1.1 degrees Fahrenheit/Celsius since the late 19th century, when records began on a global scale. This change is driven largely by increased carbon dioxide (CO^2) in our atmosphere, now at its highest point in 3 million years at 400 parts per million (2016), and other human-made emissions into the atmosphere.

Most of the warming occurred in the most recent 35 years, with 16 of the 17 warmest years on record occurring since 2001. The oceans have absorbed much of this increased heat, with water temperatures rising 0.302 degrees Fahrenheit since 1969.

Other effects of climate change are readily observable. The Greenland and Antarctic ice sheets have decreased in mass. Data from NASA's Gravity Recovery and Climate Experiment show Greenland lost 150 to 250 cubic kilometers (36 to 60 cubic miles) of ice per year between 2002 and 2006, while Antarctica lost about 152 cubic kilometers (36 cubic miles) of ice between 2002 and 2005. Glaciers are retreating almost everywhere around the world, including in the Alps, Himalayas, Andes, Rockies, Alaska, and Africa. Satellite observations reveal that the amount of spring snow cover in the Northern Hemisphere has decreased over the last 50 years and that the snow is melting earlier. Global sea levels rose 8 inches in the last century, with the rate in the last two decades nearly doubling that of the last century. The extent and thickness of Arctic sea ice has declined rapidly. Extreme weather events have increased in frequency and intensity. Since 1950, the number of record high temperature events in the United States has increased while the number of record low temperature events is decreasing. Since the beginning of the Industrial Revolution, the acidity of surface ocean waters has increased by 30 percent.

The impact of climate change is enormous and growing: 800 million people (1% of the world's population) are vulnerable to climate change impacts such as droughts, floods, heat waves, extreme weather events, and sea-level rise.

Humans play a role in creating problems but can also play a role in addressing them. The protection of nature is one key part of the solution, since 11% of global greenhouse gas emissions caused by humans can be blamed on deforestation, comparable to the emissions from all of the cars and trucks on the planet.

A worldwide effort on climate change is crucial. The Paris Agreement or Paris Climate Accord plots a new course in the global climate effort, bringing nations together for the common cause to combat climate change and adapt to its effects, with enhanced support to assist developing countries to do so. As of October 2017, 195 United Nations Framework Convention on Climate Change (UNFCCC) members have signed the agreement and 169 have become party to it, agreeing to limit global warming and adapt to climate change, in part through the use of nature-based solutions. Unfortunately, the United States has threatened to withdraw from the agreement.

Eight articles and an editorial discuss climate change and its impact. **Dr. Peter Craigmile** served as a special guest editor for this special issue. I wish to thank him and the ASA's Advisory Committee on Climate Change Policy for helping to organize this special issue.

In our columns, **Azka Javaid**, **Xiaofei Wang**, and **Nick Horton** discuss helping statistics students assess research findings using a "study of studies" in Taking a Chance in the Classroom. We also welcome the debut of a new column, Teaching Statistics in the Health Sciences with editors **Bob Oster** and **Ed Gracely**. **Aimee Schwab-McCoy** discusses use of peer consulting in applied statistics courses in the first article.

Scott Evans

The Role of Statistics in Climate Research

Peter F. Craigmile

There is much discussion today about the role that the climate system plays in many aspects of our lives. Part of this conversation revolves around humanity's interest in quantifying the state of the climate system now and in the future (*understanding the science*). The other part of the conversation involves assessing the impact of climate change, while potentially ascertaining reasons for the change.

The climate system is a complicated beast, involving the interaction of many physical processes that are changing over the air, land, sea, and ice-sheet. It has been long known that our climate system is *chaotic*—a nonlinear dynamical system—that can be hard to understand. Even simplified climate models give rise to multimodal relationships among variables, requiring careful analysis. Thus, in climate science, it can be difficult to quantify, *with certainty*, the current state of the system.

Statistical tools are required to assess, with uncertainty, what is currently known and where the system may be headed. Adding to the challenge, climate science, like many disciplines, is experiencing a data explosion. Satellites are observing the globe in increasing detail. Sophisticated climate models are producing model output on finer and finer spatial scales. Statisticians and new statistical techniques are required to use these data sources to improve our understanding of the current climate system and what it may look like in the future.

Recent and ongoing statistical research is creating new methodologies that can help to better answer open questions in climate science. Statisticians are developing methodologies to model large spatio-temporal data sets. They are building techniques to quantify and account for model discrepancy and to combine results from multiple models. New methods are being developed for modeling extremes and identifying changepoints. Hierarchical statistical modeling methods are able to blend physical and statistical models for improved statistical inference. Computer model output can be calibrated using statistical models.

This special issue of *CHANCE* draws together a wide range of articles that introduce how statistics can be used in climate science. Guttorp demonstrates, using time series analysis, how we model and assess global temperatures using observations, and compares these measurements to historical climate model simulations. Li and I extend these statistical comparisons to include spatio-temporal data from paleoclimate proxies, as well as instrumental data and climate model simulations. Jun introduces, more formally, what a global climate model is, and demonstrates how we learn about the commonalities and differences between climate model simulations using statistical tools.

The rest of the special issue starts to look at learning about climate impacts. Hammerling introduces the area of *detection and attribution*, usually carried out using counterfactual simulations of the climate system, to explain the possible factors causing climate change. Gilleland, Katz, and Naveau set forth the statistical science underlying the modeling of climate extremes. Haran, Chang, Keller, Nicholas, and Pollard communicate the science of ice sheets, and demonstrate how statisticians are able to work with scientists to predict future features of the Antarctic ice sheet. Wikle shows how statistical methods can be used to assess the synchronicity between the climate system and different ecological systems, and Chang, Sarnat, and Liu tell the story of predicting health effects under differing projections of the climate system.

A goal of this special issue is to encourage statisticians and climate scientists to work together on the hard problems that are at the interface of the two disciplines. As data sets and climate models become more complex, the challenges of carrying out climate research in small teams increase.

There is a need to build interdisciplinary research teams, bringing together varied skill sets, that can be used to solve important problems in climate science. It is necessary to understand what expertise each community can bring to working on interdisciplinary problems in climate science, but also to learn what skills each community needs to develop to be successful.

Both disciplines also could consider how each could train students and young researchers to work at the disciplines' intersection.

This special issue was organized by the ASA's Advisory Committee on Climate Change Policy, and several of the authors are either current or past members of this committee. The role of this committee is to "advise Congress on climate change issues, with special emphasis on the roles of statistics and statisticians in advancing science and informing policy." Further details about the work of this committee can be found at *ww2.amstat.org/committees/ccpac/*. C

Further Reading

Berliner, L.M. 2003. Uncertainty and climate change. *Statistical Science 18*, 430–435.

Katz, R.W., Craigmile, P.F., Guttorp, P., Haran, M., Sanso, B., and Stein, M.L. 2013. Uncertainty analysis in climate change assessments. *Nature Climate Change 3*, 769–771.

Sanso, B., Berliner, L.M., Cooley, D.S., Craigmile, P., Cressie, N.A., Haran, M., Lund, R.B., Nychka, D.W., Paciorek, C., Sain, S.R., Smith, R.L., and Stein, M.L. 2014. Statistical Science: Contributions to the Administrations Research Priority on Climate Change. A White Paper of the American Statistical Associations Advisory Committee for Climate Change Policy. *https://www.amstat.org/ASA/Science-Policy-and-Advocacy/home.aspx*.

Core Writing Team, Pachauri, R.K., and Meyer, L.A. (Eds.). Intergovernmental Panel on Climate Change. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II, and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC.

Program on Mathematical and Statistical Methods for Climate and the Earth System. 2017–18. Durham, NC: Statistical and Applied Mathematical Sciences Institute. *https://www.samsi.info/programs-and-activities/year-long-research-programs/2017-18-program-on-mathematical-and-statistical-methods-for-climate-and-the-earth-system-clim/*.

About the Author

Peter F. Craigmile is a professor of statistics at The Ohio State University in Columbus, Ohio. In addition to other research interests, he likes to develop statistical methodology for analyzing data collected over space and time, which has application to modeling climate. He has served as chair of the ASA's Advisory Committee on Climate Change Policy.

How We Know that the Earth is Warming

Peter Guttorp

There is no doubt that global temperatures are increasing, and that human greenhouse gas emissions largely are to blame, but how do we go about measuring global temperature? It is not just a matter of reading an instrument.

In Figure 1, we see a variety of curves depicting annual global mean temperature. They are not the same, although they all show a strong increase after about 1980. Different groups, using different data and different techniques, have computed the different curves. It would be hoped that the curves would all be measurements of the annual global mean temperature, but global mean temperature is not something that can be measured directly using an instrument. On the other hand, it is the quantity most commonly used to indicate global warming.

Where do the numbers come from? We will go through some issues that are associated with determining surface temperature, and illustrate some of the uses of these temperatures.

Local Daily Mean

The basic measurements that go into the calculation of global mean temperature are readings of thermometers or other instruments determining temperature. For land stations, these instruments are typically kept in some kind of box in an open, flat space covered with grass (see Figure 2). The box keeps direct sunlight from hitting the instrument but allows wind to penetrate the box.

Readings are done at different schedules in different countries. The modern instruments measure continuously, but the measurements are not always recorded. In the United States, daily maximum and minimum temperature are recorded, and their average is the daily mean temperature. In Sweden, three hourly readings throughout the day are combined with the minimum and the maximum to calculate the daily mean temperature. In Iceland, linear combinations of two readings in the morning and afternoon are used.

Modern instruments can compute the daily average automatically, but to compare to historical data, a specific averaging method has to be applied.

Local Annual Mean Temperature

Once you have a daily mean temperature, it is easy to compute an annual mean temperature: Sum all the daily means and divide by the

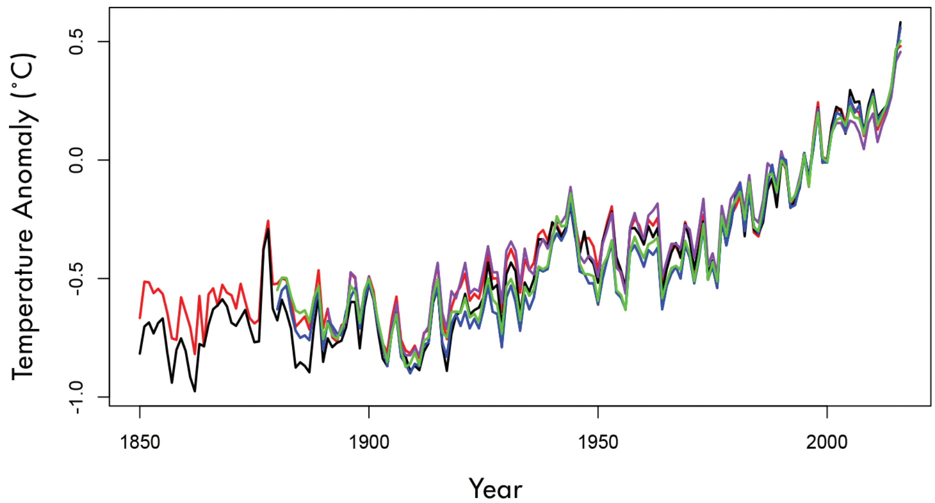


Figure 1. Five estimates of the annual global mean anomalies relative to 1981–2010: Black is from Berkeley Earth, red from the UK Met Office Hadley Center, purple from the Japanese Met Office, blue from the Goddard Institute for Space Science (GISS), and green from the National Oceanic and Atmospheric Administration (NOAA).

Figure 2. Thermometer and other instruments at Stockholm Observatory, where measurements have been made daily since 1756. The station has been moved short distances twice during this time. The box to the left is a Stephenson screen, and was used for the measurements until 2006. The pipe sticking up in the middle contains the modern measurement devise that has been used since then.
Photograph courtesy of Peter Guttorp.

number of days in the year. What is often used instead of the annual mean is something called a mean anomaly: How much did the year deviate from the average over a reference period? This makes it easier to compare sites at different altitudes, for example. A station at a higher elevation always tends to be colder than one at a lower elevation, but anomalies allow us to see if both sites are colder than usual.

The largest collection of land station data, used in the Berkeley Earth global temperature series, has some 39,000 stations and a total of 1.6 billion temperature measurements.

Sea Surface Temperature

Since more than two-thirds of the surface of our planet is water, it is not enough to take temperature measurements on land to compute a global average. Ocean-faring ships have long kept daily logbooks, with measurements of wind, air temperature, and water temperature. The water temperature used to be taken in a bucket of seawater. Later, it would be measured at the cooling water intake for the motor. Of course, ships do not travel everywhere on the oceans and, therefore, there are fairly large areas of ocean

The main groups estimating global mean temperature

- Hadley Center of the UK Met Office with the Climate Research Unit of the University of East Anglia, United Kingdom
- Goddard Institute for Space Science (part of NASA), USA
- National Centers for Environmental Information (part of NOAA), USA
- Japanese Met Office, Japan
- Berkeley Earth Project, USA

A simultaneous confidence band for n normally distributed estimates can be obtained by the Bonferroni inequality

$$P\left(\bigcup_{I=1}^{n} E_i\right) \le \sum_{i=1}^{n} P(E_i).$$

In fact, we want the complement— $P\left(\bigcap_{i=1}^{n} E_i^c\right) \ge 1 - \sum_{i=1}^{n} P(E_i^c)$.

Let E_i be the event that the true value at time i is not covered by its (pointwise) confidence set. If we let each confidence set have level $1 - \alpha / n$, we see that the probability that all parameters (in our case, the global average temperature for each year) are covered by their respective intervals is at least $1 - n(\alpha / n) = 1 - \alpha$. The confidence band then is $t_i \pm \Phi^{-1}(1 - \alpha / n)\mathrm{se}(t_i)$ where t_i is the estimated global mean temperature for year i, $\mathrm{se}(t_i)$ is the standard error of the estimate, and Φ^{-1} is the normal quantile function (inverse of the cdf).

where we have no sea surface temperature measurements from ships.

In some of these areas, there are buoys that measure the temperature. Over the last several decades, there have been satellite measurements of sea surface temperature; for over a decade, floats that measure the temperature profile of the water have been dropped all over the oceans.

The largest collection of ocean data, the ICOADS 3.0 data set, uses about 1.2 billion different records.

Combining All the Measurements

To combine the many measurements over land and oceans into an average global temperature requires estimating the temperature anomaly where there are no actual measurements, such as on a regular grid, and then averaging the estimates and measurements (if any) over the grid. Such estimation tools are derived in what is called *spatial statistics*, although atmospheric scientists have developed some methods on their own (through what they call objective analysis).

In essence, a statistician would treat the problem as one of regression, with data that are spatially dependent. The process has to take into account the fact that data are on a globe and not in the plane. The average temperature anomaly for land and ocean can be computed separately, and the global mean temperature would then be the area weighted average of the two means.

Uncertainty

There are several sources of uncertainty in the determination of global mean temperature. First of all, each measurement has error associated with it. Second, how to deal with missing areas of measurement causes uncertainty. The choice of measurement stations can also be a source of uncertainty, as can the homogenization of measurements, such as when stations are moved or measurement devices updated. There are other sources of uncertainty as well.

It is important to try to quantify the uncertainty in global mean temperature. Different groups approach this issue in different ways. Figure 3 uses the Hadley series uncertainties to compute a simultaneous Bonferroni-based 95% confidence band for global average temperature. The term simultaneous means that the confidence band covers all the true temperatures at the same time with 95% probability, as opposed to a pointwise confidence interval, which only covers the true temperature at a particular time point with 95% probability.

Ranking

In January 2017, NOAA made the claim that the global mean temperature had set a record for the third straight year. This statement

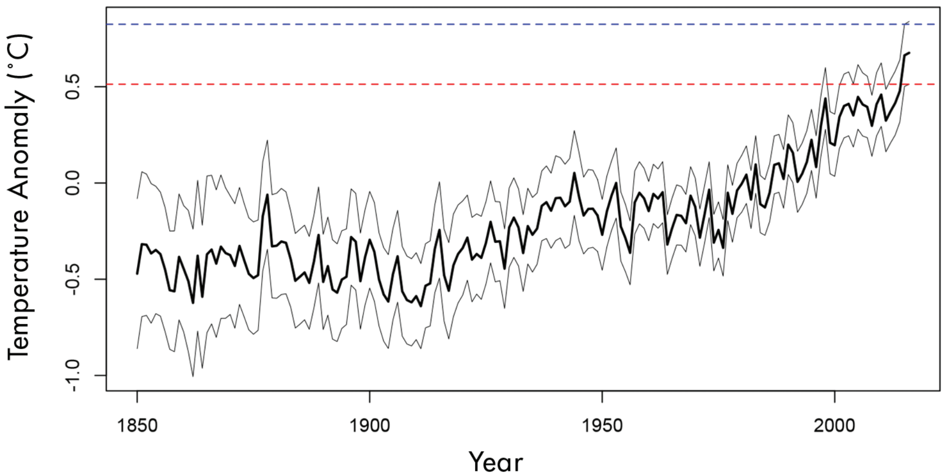


Figure 3. Hadley series with red dashed line being the lower 95% simultaneous confidence bound on the 2016 temperature and blue dashed line the upper bound on the 2015 temperature.

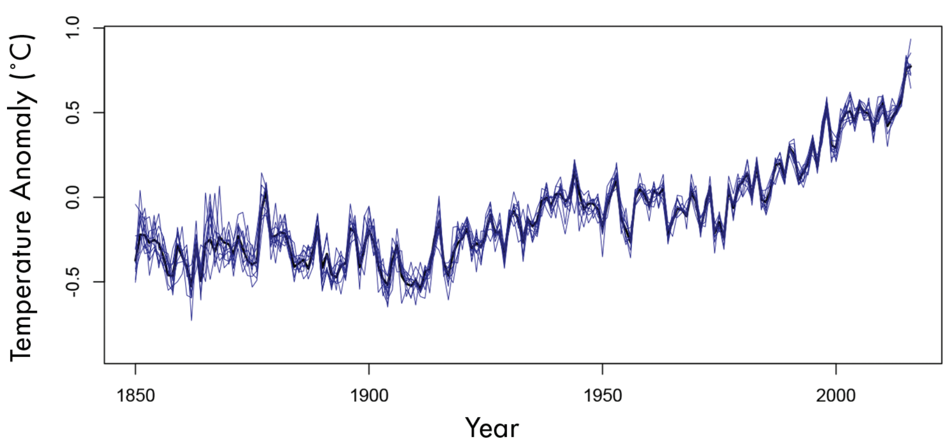


Figure 4. 10 realizations (blue) of possible Hadley temperature series and the Hadley estimate of global mean temperature (black).

is not quite accurate: For the third straight year, the *estimated* global mean temperature had set a record. In fact, four of the five series had this feature, while the Berkeley series showed 2005 as warmer than and 2010 tied with 2014. Only two of the estimates (the Hadley series and the Berkeley series) provide uncertainty estimates.

Figure 3 shows the Hadley series with associated simultaneous 95% confidence bands. If the 2015 actual temperature (which we do not know) were at the high end of its confidence band (blue dashed line), and the 2016 was at the low end of its band (red dashed line), it is quite possible that 2015 could have been substantially warmer than 2016, but that 2016 clearly was warmer than any year before 1998.

How can we say something about the uncertainty in the rankings as opposed to the estimates? One way is to simulate repeated draws from the sampling distribution of the estimates. Since we are averaging a large number of measurements, many of which are nearly uncorrelated, a central limit theorem leads us to treat the estimates as normal, with mean equal to the actual estimate and standard deviation equal to the standard error of the estimate. Figure 4 shows 10 such realizations of the Hadley temperature series.

For each of the realizations, we can calculate the rank of 2016. The distribution of that rank tells us how likely 2016 is to be the warmest year on record: It is warmest

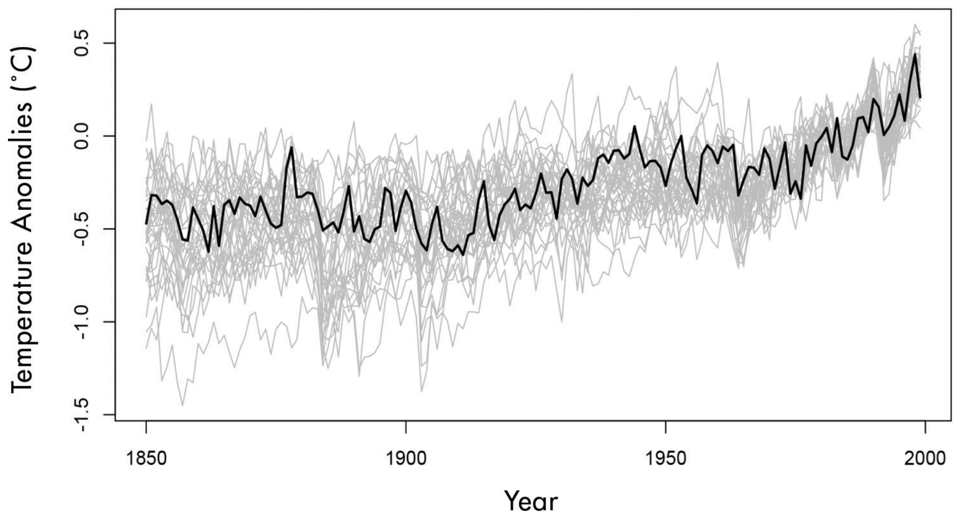


Figure 5. Global annual mean temperature anomalies from 32 CMIP5 models with historical simulations (gray), and the Hadley Center data series (black). Reference period is 1970–1999.

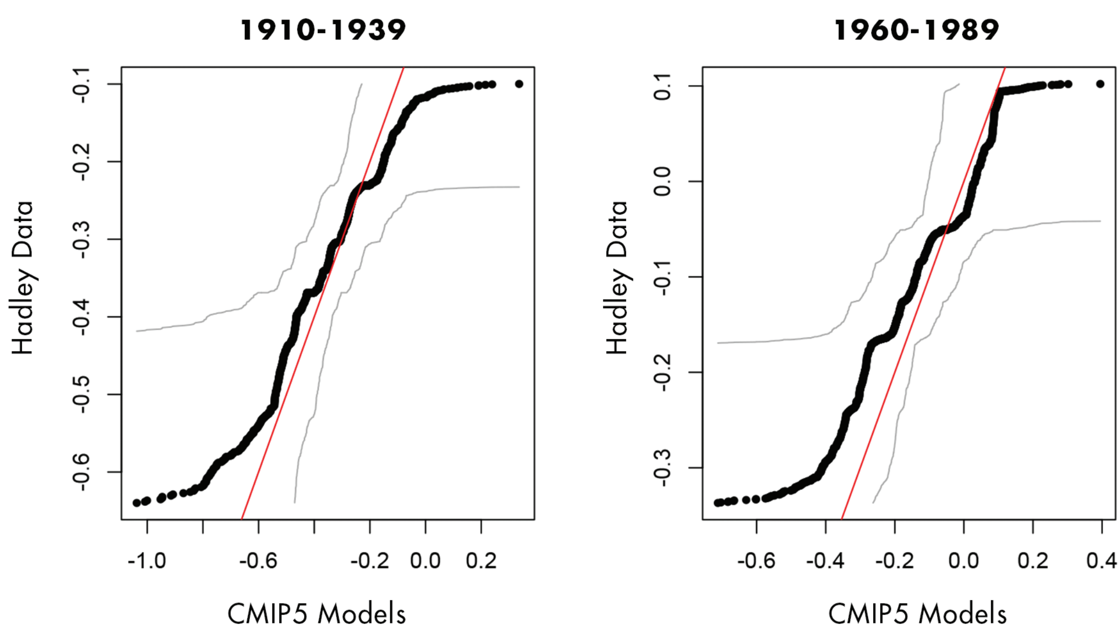


Figure 6. QQ-plots of historical climate model simulations against Hadley Center data or two 30-year periods. The gray lines are simultaneous 95% confidence bands, and the red lines are lines of equal distributions.

in 58% of the simulations, while 2015 is warmest in 42%. In 10,000 simulations, 2016 was as low as the eighth-warmest in one of them.

How about all three years—2014–2016—being record-breakers? That happened in 21% of the simulations, and in the actual Hadley estimates, of course.

Many tests have been developed to compare some aspects of distributions, such as means or medians. To compare two entire distributions, we can plot the quantiles of one against the other (called a quantile-quantile plot or QQ-plot). An advantage of this plot is that if the distributions are the same, then the plot will be a straight line. Of course, we will be estimating the quantiles from data, so there will be uncertainty. Another advantage of the QQ-plot is being able to develop simultaneous confidence bands, enabling a simple test of equal distributions: Does the line *y*=*x* fit inside the confidence band?

Models and Data

Climate, from a statistical point of view, is the distribution of weather. Climate change means that this distribution is changing over time. The World Meteorological Organization recommends using 30 years to estimate climate. This definition indicates, for example, that it does not make sense to look at shorter stretches of data to try to assess questions such as "Is global warming slowing down?"

What is a Climate Model?

A climate model is a deterministic model describing the atmosphere, sometimes the oceans, and sometimes also the biosphere. It is based on a numerical solution of coupled partial differential equations on a grid. In fact, the equations for the atmosphere are essentially the same as for weather prediction, but the latter is an initial value problem (we use today's weather to forecast tomorrow's) of a chaotic system, while the climate models has to show long-term stability. Many processes, such as hurricanes or thunder storms, are important in transferring heat between different layers in the model, but often take place at a scale that is at most similar to a grid square, and sometimes much smaller.

Different climate models deal with this subgrid variability differently and, as a consequence, the detailed outputs are different. CMIP5 is a large collection of model runs, using the same input variables (solar radiation, volcanic eruptions, greenhouse gas concentrations, etc.). These model outputs were used for the latest IPCC report in 2013. Figure 5 shows the global mean temperature anomalies (with respect to 1970–1999) with the corresponding Hadley Center series.

Comparing Distributions

It is not trivial to compare climate model output to data. Remember, the climate model represents the *distribution* of the data. The observations in Figure 5 are, therefore, not directly comparable to the model runs. Instead, we need to compare the distributions of model output and data, respectively.

Figure 6 compares these distributions using QQ-plots for two different 30-year stretches. In both cases, the distribution of the data fit the distribution of the ensemble of model outputs quite well, in that the red *y*=*x* line falls inside the simultaneous 95% confidence bands. Since we have 32 x 30 observations of the models, and only 30 of the data, the empirical tails of the model distribution are much longer than the tails of the data, but the confidence band is quite wide in the tails, meaning that we are very uncertain there. Thus, for these two time intervals and for the global mean temperature variable, the ensemble of CMIP5 models and the Hadley Center data seem to have the same distribution—they are describing the same climate. C

Further Reading

Arguez, A., Karl, T.R., Squires, M.F., and Vose, R.S. 2013. Uncertainty in annual rankings from NOAA's global temperature time series. *Geophysical Research Letters 40*:5,965–5,969.

Doksum, K. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist. 2*:267–277.

Guttorp, P. 2014. Statistics and Climate. *Annual Reviews of Statistics and its Applications 1*, 87–101.

Katz, R.W., Craigmile, P.F., Guttorp, P., Haran, M., Sanso, B., and Stein, M.L. 2013. Uncertainty analysis in climate change assessments. *Nature Climate Change 3*, 769–771.

About the Author

Peter Guttorp is a professor at the Norwegian Computing Center in Oslo and professor emeritus at the University of Washington in Seattle. He has worked on stochastic models in a variety of scientific applications, such as hydrology, climatology, and hematology. He has published six books and about 200 scientific papers.

Instruments, Proxies, and Simulations: Exploring Imperfect Measures of Climate

Peter F. Craigmile and Bo Li

Scientists working in climate research routinely use statistical models and methodologies to learn about the dynamics of climate and the interaction between different variables that configure the Earth's climate system. A rich variety of data sources is available for studying climate. In recent years, there has been a growing interest in exploring the role that different measures of climate can play, as well as the capacity of each measure in learning about the climate system.

It could be argued that the quantity of data is increasing, but while there are many more sources, data quality and coverage are not consistent. These rich data provide us an opportunity to explore and compare their characteristics.

Comparing three different measures of temperature in upper-latitude North America—from instruments, measured indirectly from paleoclimate proxies, and from climate model simulations—reveals lessons using simple statistical tools that probably can be generalized to other data sources.

Instrumental Measures

Instrumental measures provide the most-accurate measure of the Earth's climate. They are calibrated to be direct measures of specific variables such as temperature, precipitation, and pressure. However, because of, for example, measurement error, instrumental measures are not perfect. Direct measures may be influenced by urbanization, causing so-called local "heat island" effects, and will be more uncertain when monitoring stations move location over time.

Instrumental measures are often provided as a *data product*, a transformation of the original observations into a form that is more suitable for a specific analysis. For example, Brohan, et al.'s Climate Research Unit (CRU) gridded temperature product is the result of spatially averaging observations onto a 5 degree latitude by 5 degree longitude grid.

A problem with direct measures of the climate system can be the lack of spatial and temporal coverage. Figure 1 illustrates this problem for the CRU data product, restricted to the region of interest—routine measurements of temperature started around 1850, but the spatial coverage has varied substantially by year. Before 1850, there are only very few or no measurements that could represent the global climate. If we want to understand climate in the past, we need to look for other indirect measures of climate.

Paleoclimate Proxies

Paleoclimate proxies, whose growth or formulation is often susceptible to climate, are usually measures from media such as tree rings (e.g., Figure 2), ice cores, and pollen. Paleoclimate proxies provide valuable information about the Earth's past climate when no or too few instrumental measurements are available. However, this information is imperfect because either the exact relationship between climate and the measurements provided by the proxy is hard to capture or the proxy measurements integrate other, non-climatic, factors.

For example, tree growth, as measured by the density of tree rings, while positively correlated to the surrounding temperature and precipitation during the growing season, does not allow us to infer perfectly the exact climatic conditions in that year.

Proxies are often preprocessed to remove effects not due to climate (e.g., age effects are often removed from tree ring proxy data) and, as with instrumental measures, may be available as gridded data products.

Briffa, et al.'s maximum late wood tree ring density data set is an example of such a paleoclimate data product. Figure 1 indicates that tree ring densities are observed more sparsely in space than direct measures of temperature after 1920, but they still maintain reasonable spatial coverage for a certain period before 1920 and, thus, can be useful for deducing climate in the past. Therefore, even with all the uncertainties in the imperfect

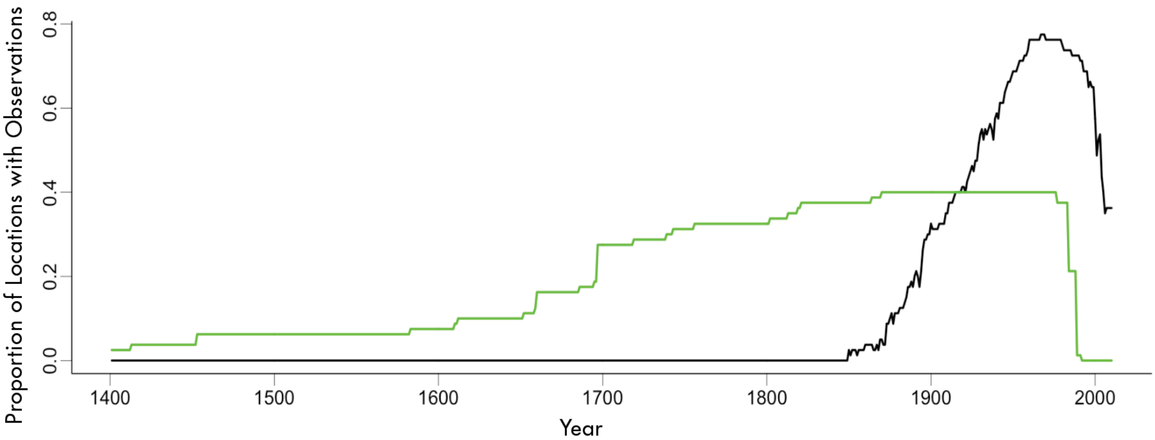


Figure 1. For two gridded climate data sets over upper-latitude North America, the proportion of grid boxes with observations, by year: temperatures in black; tree ring densities in green.

Figure 2. A popular paleoclimate proxy is tree growth, as measured by the density of tree rings.

proxy measurements, their longer temporal coverage, compared with instrumental measures, still makes them compelling for the purpose of reconstructing past climate.

Climate Model Simulations

The *output* or *simulations from computer climate models* cannot be considered as real measurements of the Earth's climate. Instead, they are synthetic climate conditions generated by mathematically quantifying various feedbacks and interactions between drivers of climate, also known as *forcings*, such as greenhouse gases, and climate variables such as temperature or precipitation.

Climate model simulations provide useful information about the "distribution" of such climate variables. ("Distribution" means the possible values that could be observed over space, time, under varied—initial—conditions of the computer model.) While climate models are "tuned" to reproduce global mean values of climate variables, such as temperature, it is still interesting to study their spatially and temporally varying distribution.

As an example, Phipps, et al.'s CSIRO Mark 3L climate model last-millennium simulations

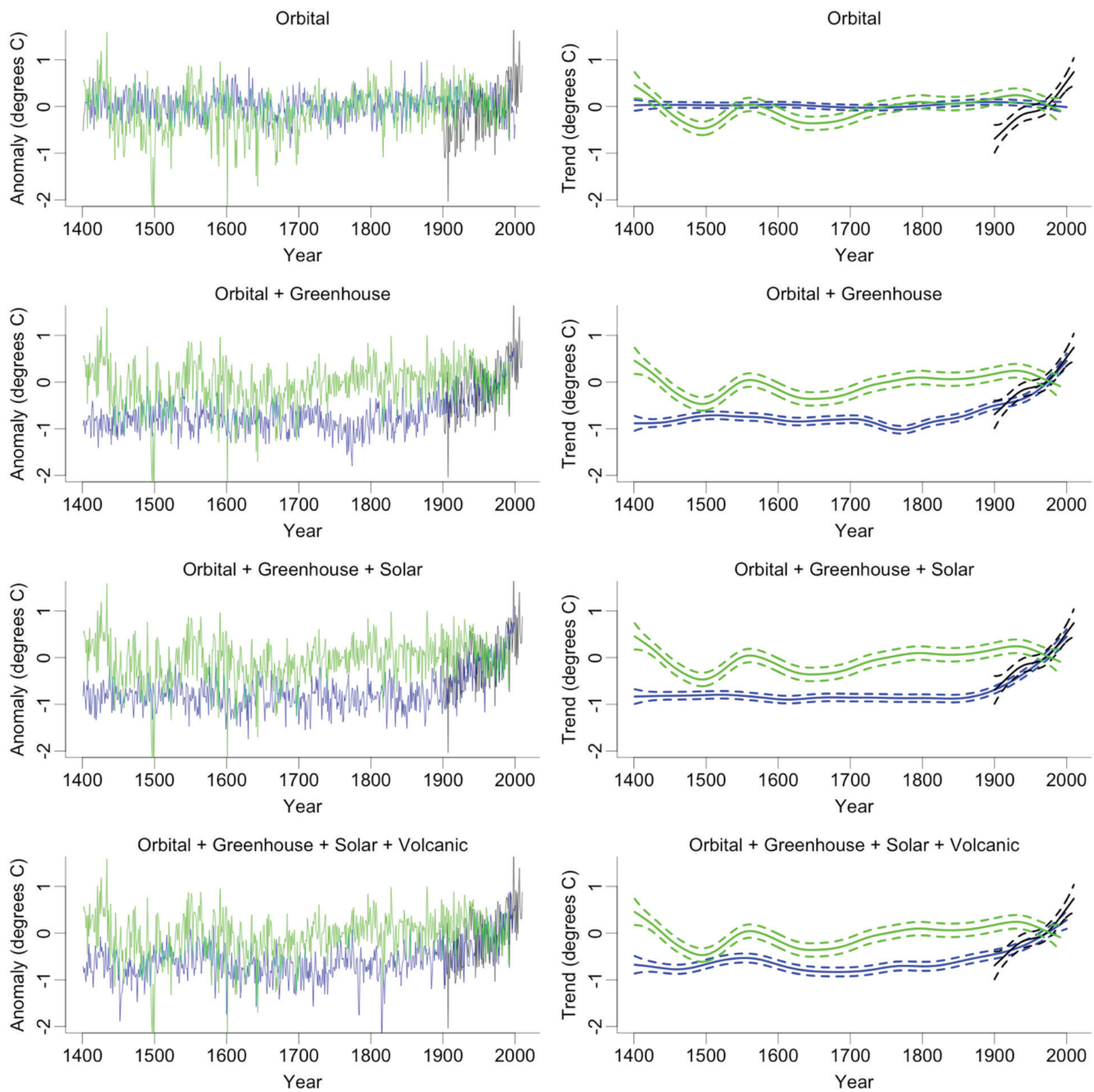


Figure 3. Averaging values over the grid boxes in upper-latitude North America, plots of the temperature data product (in black), tree-ring density values (in green), and an average climate model run (in blue) by year (left panels). Each row displays a different forcing scenario for the climate model runs. Values are shown as anomalies (relative to the mean value from 1961–1990). Estimated decadal trends for each global series (solid lines), with associated 95% pointwise confidence intervals (dashed lines) (right panels).

produce synthetic climate under a number of forcing scenarios. These include greenhouse gas concentrations, solar irradiance, orbital features (such as induced by the tilt and orbit of the Earth), and volcanic emissions. There is a growing interest in running climate models as experimental designs. By varying the factors that go into the climate model (e.g., the forcing scenarios or the initial conditions), we can see how the distribution of climate changes. Using weighted averages, the climate simulations were converted to the same spatial resolution as the instrument and paleoclimate measures.

Table 1—Correlations between Indirect Measures of Temperature

	With Instrumental		With Paleoclimate	
Indirect Measure	**Correlation**	**95% CI**	**Correlation**	**95% CI**
Paleoclimate	0.41	(0.25, 0.67)		
Sim: orbital	–0.22	(–0.38, –0.02)	–0.05	(–0.14, 0.04)
Sim: orbital + greenhouse	0.41	(0.13, 0.56)	0.08	(0.01, 0.15)
Sim: orbital + greenhouse + solar	0.45	(0.19, 0.60)	0.09	(0.02, 0.17)
Sim: orbital + greenhouse + solar + volcanic	0.31	(0.06, 0.46)	0.16	(0.08, 0.25)

Averaging over the grid boxes in upper-latitude North America, the correlation between each indirect measure of temperature (paleoclimate and climate simulations) and the instrumental data. The 95% confidence intervals for the correlations take into account the temporal correlation within each series.

With these three data products, we now demonstrate, using a simple statistical analysis, how they relate to each other. In particular, we assess when and to what degree the different measures agree, and when and to what degree they diverge as different variations of real climate. Throughout, we take the opinion that while the instrumental temperatures are measured with error, when available, they are the most reliable measures of temperature.

Global Comparisons

We first examine how consistently each data product represents the global climate. The left-hand panel of Figure 3 compares averages, over upper-latitude North America, of the instrumental temperature data product (in black), tree ring density values (in green), and the average of three climate model runs from the CSIRO model (in blue), by year. Values are presented as anomalies relative to the mean value from 1961–1990.

In each row, we vary the forcing scenario, starting with only the orbital forcing and adding one forcing at a time until orbital, greenhouse, solar, and volcanic forcings are present in the climate model. This figure not only displays the differences in temporal coverage for the three data sets, but also gives a clear visualization of the discrepancies in their values.

The right-hand panels of Figure 3 compare estimated decadal trends for each global series anomaly as we vary the forcing scenario in different rows. (The trends were estimated via smoothing splines, using a statistical model that allows the errors to be correlated through time.) The dashed line around each trend estimate denotes pointwise 95% confidence intervals for each trend.

Table 1 summarizes the correlations between the indirect and direct measures of climate, calculated over the periods for which a given pair of series are both available. For uncertainty quantification, we produce 95% confidence intervals for each correlation coefficient. (The confidence interval calculation uses a statistical method, called the *moving block bootstrap*, that allows each series to be correlated through time.)

The linear association between paleoclimate (tree ring densities) and instrumental temperatures is moderately strong at this global spatial scale, and is highly statistically significant. The significant association between paleoclimate proxies and temperatures at global scale has been observed and is used widely to learn about past temperatures. Comparing the decadal trends between paleoclimate and instrumental reveals that observed increasing temperatures in recent decades are not present in the tree ring records (Rosanne D'Arrigo and colleagues call this the "divergence problem").

After detrending, there is a stronger association between the global paleoclimate and instrumental values (a correlation of 0.64, with a 95% confidence interval of between 0.52 and 0.77).

Comparing the instrumental temperatures to different climate model simulations shows that the climate models lacking in forcings tend to be poorly correlated with the instrumental temperatures. Indeed, with a climate model containing only orbital forcings, we obtain no statistically significant association between temperatures and the climate model simulations. The most complicated forcing scenario (involving orbital, greenhouse, solar, and volcanic forcings) is less linearly associated with the direct measure of temperature than the climate model without volcanic forcings.

A compelling reason for this can be the difficulty in absorbing and digesting the unevenly spread and rugged volcanic forcings in climate models. Most of the agreement between the instrumental temperatures and climate model simulations are in the decadal trends (after detrending, there are no significant correlations between the residual instrumental values and the climate model simulations).

Despite the associations between paleoclimate and instrumental, and between simulation and instrumental, the correlation between paleoclimate and the climate model simulations averaged over upper-latitude North America is weak, although significant.

This conclusion, nevertheless, is only based on the linear association between different data sets. Whether there is a non-linear association between proxies and model simulations requires further investigation. It is interesting that unlike with instrumental temperatures, there is no decrease in the correlations at the most-complicated forcing scenario. As more forcings are added to the climate model, the correlation between the climate model simulation and tree ring densities increases.

Both climate models and these paleoclimate records are informative about climate on the global scale. However, the climate information in proxy records and model simulation may come from different perspectives and, thus, integrating the two data sources may provide more climate information than each individual source.

Local Comparisons

The next step is to move to a local scale and see if the same conclusion in the global comparisons still apply. Figure 4 demonstrates the various correlations between the indirect (paleoclimate and climate model runs) and the direct measure of temperature. This case is restricted to comparing with the climate model that contains all four forcings and does not compare decadal trends at each location. The blue values indicate those correlations that are significantly positive on the basis of a hypothesis test for the correlation parameter, again accounting for possible dependence in each time series.

The correlations between the instrumental temperatures and tree ring densities are the most significant and consistent throughout the different grid boxes across upper-latitude North America, although there is some spatial variation in the strength of the relationship. It is also worth noting that these grid-specific correlations in general appear to be stronger than the correlation at the global scale, which is 0.41. This may indicate that with proxy data we may first make reconstructions at the grid level and then average them to obtain the global reconstruction.

In terms of relating climate simulations with the instrumental and the paleoclimate measures, only remote regions of statistically significant linear association (in the east for instrumental and climate model simulations; in the south and west for paleoclimate and the climate model simulations). It is surprising to find that the correlation between simulations and instrumentals largely downgrade at the grid level, as opposed to a fairly strong correlation at the global scale. The proxies and simulations remain weakly associated at the grid level.

The big patch of weak and nonsignificant correlations in the middle and lower panels of Figure 4 implies that proxies and simulations can give completely different estimates of climate, where estimates from proxies perhaps have more credibility due to the relatively stronger correlation shown in the upper panel of Figure 4.

These findings seem to suggest that the optimal way to use model simulations in a spatial reconstruction of climate is to have the simulations regulating the climate at the global scale rather than at the local scale. We might be able to use the local information from climate model simulations if nonlinear associations between the model simulations and instrumental measures at a local scale were identified. This raises a question of how much capacity we have in recovering the local climate of the past based on the available data products.

Conclusions

Global average temperatures in the upper-latitudes of North America show agreements between all three data products and the impact of climate forcings. The signal from solar forcings seems weak, and the signal from orbital and volcanism can be hard to interpret. For local temperatures, the climate footprint from climate forcings is weaker over space. The real climate, especially expressed at local scales, is very volatile over space, as is seen these three data products (instrumental, paleoclimate, and climate model simulations) express local climate signals quite differently.

Further research, including more data collection of direct and indirect measures, is required to improve climate reconstruction.

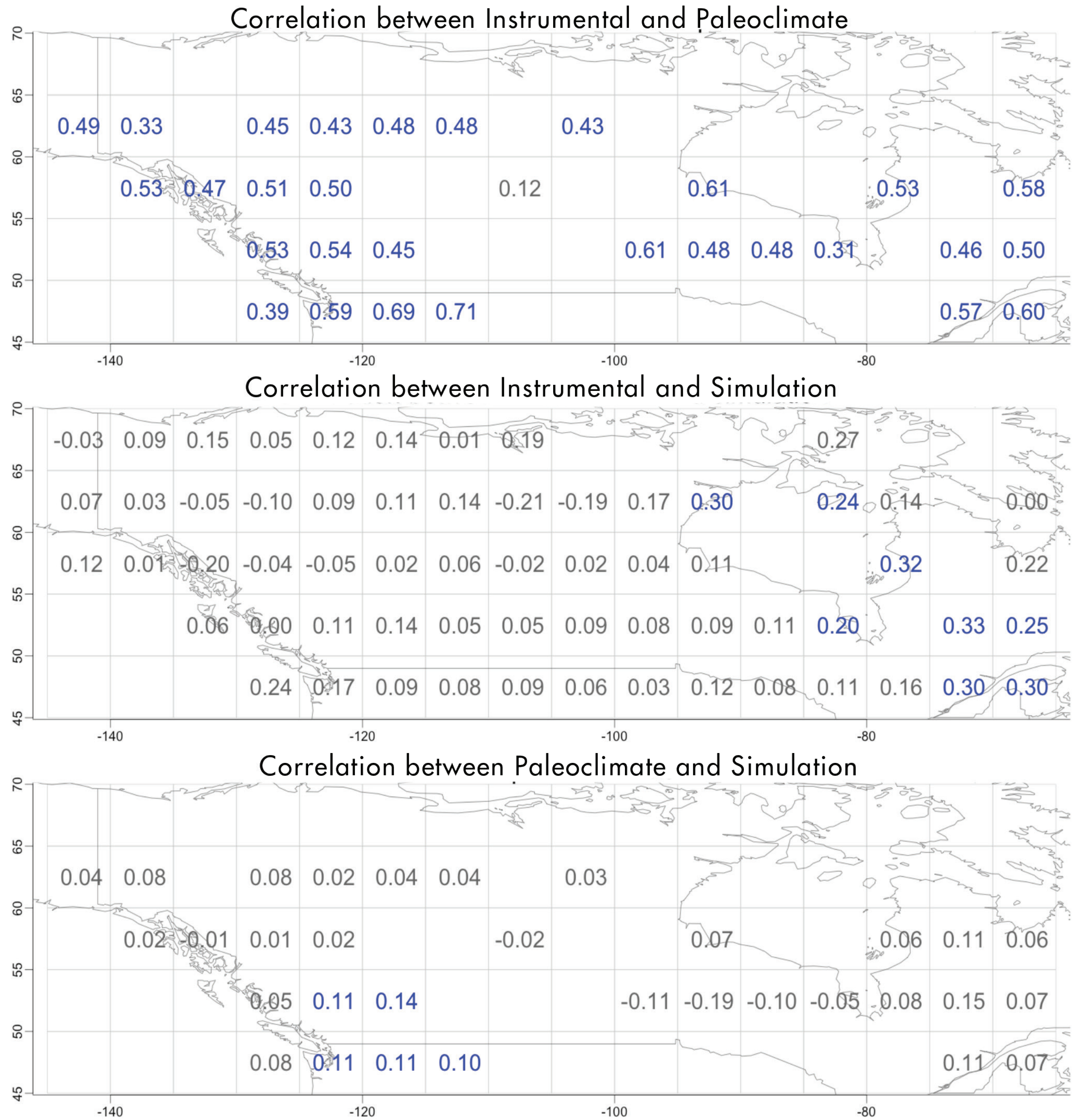


Figure 4. For each grid box, correlations between each indirect measure of temperature (paleoclimate and climate simulations) and the instrumental data. The blue numbers denote significantly positive correlations at a marginal significance level of 0.05.

To explore the relationship between different data sets based on calculating estimates of decadal trends and a simple correlation coefficient, we calculated each correlation separately for each grid box for the local comparisons.

By no means do we indicate that the correlation coefficient perfectly measures how well we can carry out climate reconstructions. It is known that there is a nonlinear relationship between the real climate and indirect measures of climate (e.g., tree growth is not linearly related perfectly to temperature and precipitation). Also, the correction of instrumental measures to make them more reliable is a progressing research area and will lead to higher-quality instrumental climate products.

Finally, the spatial reconstruction of climate using more-sophisticated statistical models often allows for the reconstruction at one single point to borrow information from their neighboring points, by taking account of the spatial correlation. This strengthens the ability to reconstruct the underlying climate processes over time and space using direct and indirect measures.

This simple analysis can be considered as a preliminary exploration of different climate data sets. More rigorous statistical comparisons to understand their common characteristics as well as their own individual features are called for. This study demonstrates that it is more appropriate to use statistical models that account for the uncertainty inherent in each data source to allow for a more precise uncertainty quantification in reconstructing, and more generally modeling, climate. C

About the Authors

Peter F. Craigmile is a professor of statistics at The Ohio State University in Columbus, Ohio. In addition to other research interests, he likes to develop statistical methodology for analyzing data collected over space and time, which has application to modeling climate. He has served as chair of the ASA's Advisory Committee on Climate Change Policy.

Bo Li is an associate professor of statistics at the University of Illinois at Urbana-Champaign, Illinois. She enjoys modeling spatial and spatio-temporal data; in particular, the data regarding the climate, environment, and health. She is currently serving as chair of the ISBA Environmental Science Section and on the ASA Advisory Committee on Climate Change Policy.

Further Reading

Bradley, R.S. 2014. *Paleoclimatology* (third edition). Boston, MA: Academic Press.

Briffa, K.R., Osborn, T.J., Schweingruber, F.H., Jones, P.D., Shiyatov, S.G., and Vaganov, E.A. 2002. Tree-ring width and density data around the Northern Hemisphere (Parts 1 and 2). *The Holocene* 6:737–789.

Brohan, P., Kennedy, J.J., Harris, I., Tett, S.F.B., and Jones, P.D. 2006. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research* 2:99–113.

D'Arrigo, R., Wilson, R., Liepert, B., and Cherubini, P. 2008. On the "divergence problem" in northern forests: A review of the tree-ring evidence and possible causes. *Global and Planetary Change* 60:289–305.

Li, B., Nychka, W.D., and Ammann, C.M. 2010. The value of multi-proxy reconstruction of past climate (with discussions and rejoinder). *Journal of the American Statistical Association* *105*:883-911.

Smerdon, J.E. 2017. What was Earth's climate like before we were measuring it? *Significance* *14*:24–29.

Tingley, M., Craigmile, P.F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B. 2012. Piecing together the past: Statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews 35*:1-22.

Tingley, M., Craigmile, P.F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B. 2015. On discriminating between GCM forcing configurations using Bayesian reconstructions of Late-Holocene temperatures. *Journal of Climate 28*:8, 264–8, 281.

Washington, W.W., and Parkinson, C.L. 2005. *An Introduction to Three-Dimensional Climate Modeling* (second edition). Sausalito, CA: University Science Books.

Climate Model Intercomparison

Mikyoung Jun

Despite some climate "skeptics" over global warming and climate change due to human influence, it is safe to say that scientific consensus has been achieved and the majority of the public believes that it is real. It can be observed that global surface temperature keeps rising and that arctic sea ice is shrinking rapidly. Numerous scientific research studies report that the increase in global surface temperatures is due to human-induced emission of greenhouse gases.

Since we can only observe the past and present climate, though, we need tools and methods for accurate prediction of future climate change. Furthermore, temperature or precipitation are not the only causes, but rather, numerous climate variables vary and interact together over space and time. Thus, we need sophisticated tools and methods for understanding the current climate and forecasting the future climate as accurately as possible.

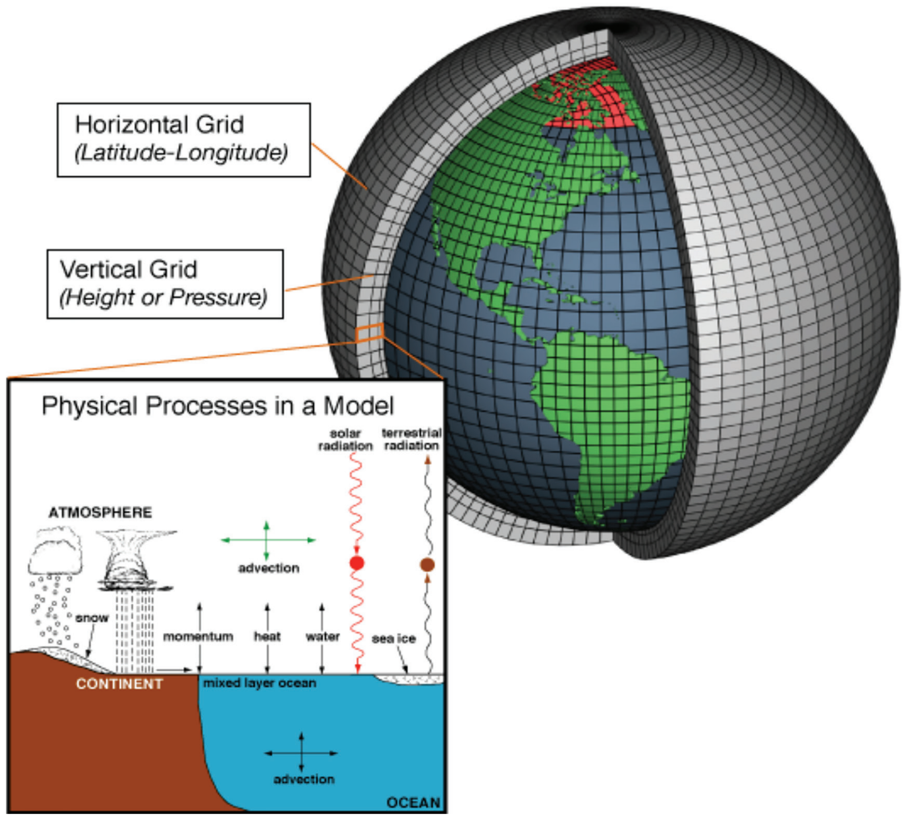


Figure 1: Schematic view of a global atmospheric and ocean model. Source: NOAA.

What are Global Climate Models?

Coupled Atmospheric and Ocean General Circulation Models (AOGCMs, commonly called *climate models* or *global climate models*) are complex systems of partial differential equations intended to simulate the climate in a four-dimensional domain: surface of the Earth (2-dimensional) × vertical height (or pressure) × time. The models are called "coupled" since the atmospheric and ocean components interact in a complex manner (for instance, sea surface temperature information is fed into the atmospheric component, and atmospheric wind stress information is fed into the ocean component).

The outputs of these climate models are the values of multiple climate variables (temperature, precipitation, radiation, relative humidity, wind pressure, etc.) on a gridded domain (latitude × longitude × vertical height or pressure; see Figure 1) across time. They also involve several submodels, such as a sea ice model. Many physical processes are parametrized; there are thousands of parameters for physical processes such as solar radiation, convection, and cloud cover, and the inset in Figure 1 shows only a few examples of the physical processes that go into a global climate model.

These coupled AOGCMs are the main sources of information for understanding the climate in general and especially for predicting future climate change (since we do not have data from the future).

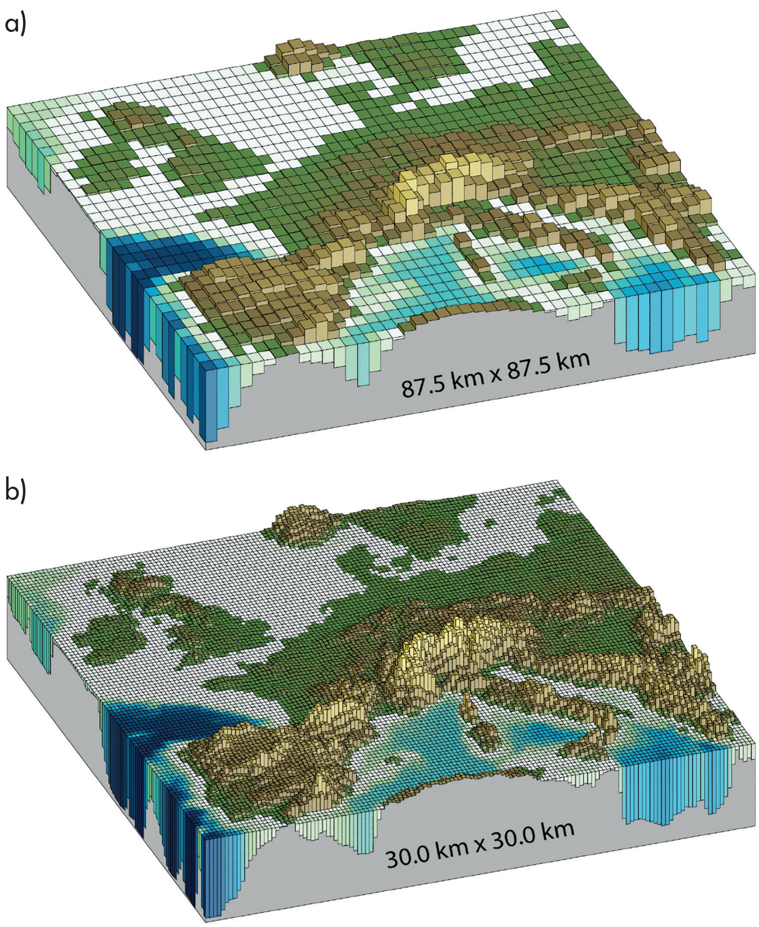


Figure 2. Comparison of grids for (a) a global climate model and (b) a regional climate model, with (b) showing the increased grid in both horizontal and vertical dimensions.
Source: IPCC AR5.

Another use of these models is to simulate the effect of greenhouse gases on the future climate. For example, varying emission scenarios (how much greenhouse gas gets emitted into the air), can demonstrate how much surface temperature will increase (or decrease) given an emission scenario for the future.

Along with the great achievements made in developing global climate models, serious effort has been devoted to developing Regional Climate Models (RCMs). RCMs are similar to global climate models, but they are forced by specified lateral and ocean conditions from a global climate model or observation-based data set. They account for high-resolution topographical data, land–sea contrasts, surface characteristics, etc. They only cover a limited domain, and the values at their boundaries are specified. Figure 2 provides an illustration showing the differences in horizontal and vertical grid resolutions between a global climate model and an RCM over a spatial region.

Are Climate Models Accurate?

Climate models are essential tools in climate study. They are used for weather forecasting and understanding the past and present climate, and are irreplaceable tools for understanding climate change. They are huge systems and cost billions of dollars to be developed and to run on supercomputers.

Climate models improve over time, owing to tremendous efforts on the part of scientists and modelers around the world. However, they are still far from perfect. Although numerous state-of-the-art models have been developed by top-notch scientists and climate modelers, they often give very different answers about certain aspects of the climate.

Some of these differences may be due to the climate system's internal variability (due to natural internal processes within the climate system). There is also the climate system's external variability (due to natural and anthropogenic external forcing). Forcing refers to a change in the net energy exchange between the climate system and the environment, such as solar intensity cycles, anthropogenic emissions of greenhouse gases, and volcanic eruptions.

The volcanic and solar forcing reconstruction by climate models may differ. Climate models also may get aerosol effects wrong. Aerosols can affect climate in several ways. For instance, they can reflect incoming sunlight back to outer space when the sky is clear, which can trap solar energy within the atmosphere. Climate models incorporate the effects of aerosols on clouds through parametrizations, and problems in these parametrizations could result in incorrect projections of climate.

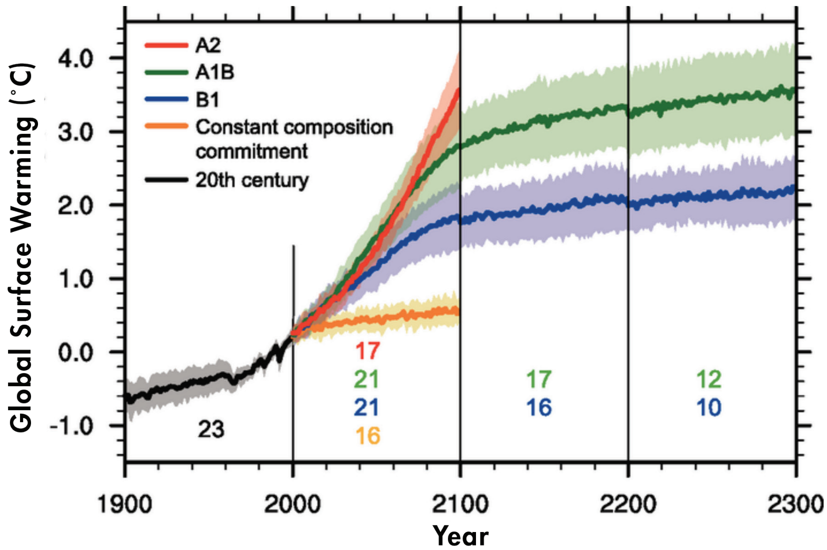


Figure 3. Multi-model means of surface warming (relative to 1980–1999) for the emission scenarios A1B, A2, and B1, shown as continuations of the 20th-century simulation. This figure is based on CMIP3 runs. Six emission scenarios were considered, including the three in the figure. Scenario A1 represents an integrated world (rapid economic growth, globalization), and A1B is a subset of the A1 family that represents a balanced emphasis on all energy sources. Scenario A2 represents a more-divided world with regionalization, based on economic focus. Scenario B1 represents a more-integrated and ecologically friendly world. More details on all scenarios can be found on the IPCC emissions scenarios website.

Source: IPCC AR4; *www.ipcc.ch/ipccreports/sres/emission/index.php?idp=3*.

It is widely acknowledged that current versions of climate models have great room for improvement. There are areas such as sea ice regions where climate models suffer more than in other areas, and there are climate variables such as precipitation for which climate models have a much harder time, compared to other climate variables such as surface temperature. Therefore, a thorough understanding and validation of these climate models is imperative.

CMIP and Climate Model Intercomparison

The Intergovernmental Panel on Climate Change (IPCC, *http://www.ipcc.ch*), established by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO) in 1988, publishes Assessment Reports (ARs) on climate change based on results from various AOGCMs developed by multiple organizations worldwide.

The Coupled Model Intercomparison Project (CMIP) is an experimental protocol established in 1995 by the World Climate Research Programme (WCRP), under the United Nations, for climate model diagnosis, validation, intercomparison, documentation, and data access. A large number of members of the international climate modeling community is participating, and much of the CMIP data are archived and publicly available from the Program for Climate Model Diagnosis and Intercomparison.

Phase three of the CMIP (CMIP3), which provided materials for the IPCC's fourth Assessment Report (AR4), included about 20 climate models and completed the data for the archives, mostly in 2005 and 2006. The number of climate models participating in the next phase, CMIP5, was much greater, and the performance of each of these climate models was shown to be generally better than those of CMIP3. The archive for CMIP6 is currently underway.

Scientific publications have noted that there are often significant disagreements between the climate models participating in CMIP3, as well as its successors. For example, the projected global temperature increase in 2100 and thereafter can vary by more than 1 degree (see Figure 3). The "accuracy" of climate model outputs may mean different things; it might be about the "mean" (i.e., the center) of the prediction to some people and about the "uncertainty" (i.e., the spread) to others. At any rate, one might expect to see smaller uncertainty from one CMIP generation to the next in terms of how similar or different the climate models are in that generation.

However, according to Knutti and Sedláček (2012), this is not the case. The authors show that the uncertainty (or spread) in CMIP5 models is comparable to the uncertainty in CMIP3 models. This rather disappointing "improvement" in CMIP5 models over those in CMIP3 does not necessarily mean that the models did not get better. It may well be due to the fact that the climate models became more complicated (for example, more forcing terms and aerosols were added to the models in CMIP5). Knutti and Sedláček also note that direct comparison of the models in CMIP3 and the models in CMIP5 is not possible, since the two generations use different emission scenarios.

They further argue that there are many other factors that contribute to model disagreements, such as (i) the inherent limitation of computational resources; (ii) lack of a complete scientific understanding of climate; (iii) lack of long-term observations (our observation records for the climate do not go back further than the 1800s, and satellite data have only became available since around the 1970s; and (iv) lack of consensus on the metric to use to determine how similar or dissimilar model outputs are.

Despite the complexity of climate models and their vast amounts of output, common practices for validating model outputs and comparing across different climate models rely on simple spatial or temporal means and variances. With the advancement of spatial and spatio-temporal methods in statistics, statisticians can go beyond the current common practice to better understand the spatial and temporal variation structures of climate model outputs.

In particular, statistical techniques for multivariate spatial and spatio-temporal distributions allow statisticians to validate climate model outputs, considering multiple climate variables jointly at various spatial and temporal scales (i.e., how one climate variable covaries in space and/or time with other climate variables and how they interact in space and/or time).

One example is the work of Philbin and Jun (2015), who consider surface temperature and precipitation jointly and focus on validating climate models in terms of how well they produce the joint distributional structure of the two climate variables. They treat each climate variable (30-year seasonal averages from 1981 to 2010) as spatial data and estimate key (statistical) parameters that jointly describe the spatial variations of the two climate variables. In addition to comparing those (statistical) parameter estimates across different climate models, they also compare them with the (statistical) parameter estimates of observations.

The idea behind this is that, if each climate model represents something more or less similar to the "true" climate (in this case, the observations are considered to be the truth, subject to measurement error), then those (statistical) parameter estimates from the climate models, as well as those from observations cannot be drastically different.

Figure 4 shows an example of the results in Philbin and Jun. The figure displays estimates for the so-called *smoothness* parameter (how smoothly each climate variable varies over space) and the cross-correlation (correlation between the temperature and precipitation value at a given location) for the climate models and observations considered.

While some models agree with each other (and with the observations), some models are quite distinct from the others. Philbin and Jun report that statistical models predict surface temperature and precipitation data with less smoothness than the observational data for the tropics. The estimated spatial cross-correlations of these two climate variables are quite different for most climate models in the mid-latitudes.

Climate Model Interdependence

Another interesting but challenging issue for climate model intercomparison is the models' *interdependence*. As Knutti, et al. (2013) note, most models are strongly tied to their predecessors, and some also exchange ideas and codes with other models. As a result, climate models are correlated with not only each other, but also with the earlier generation. Here, we do not simply mean how similar the climate model outputs are. Rather, we are concerned with how similarly *wrong* the climate models are.

Nevertheless, the current practice of using these "ensembles" of climate models to quantify the uncertainty of projected climate change (as in Figure 3) often assumes that these multiple climate models provide "independent" information about the future climate. This assumption of independence within climate model ensembles leads to incorrectly assessing the levels of uncertainty (i.e., overconfidence) of future climate projections.

For instance, even if 20 climate models are considered in a study, you will not get 20 sets

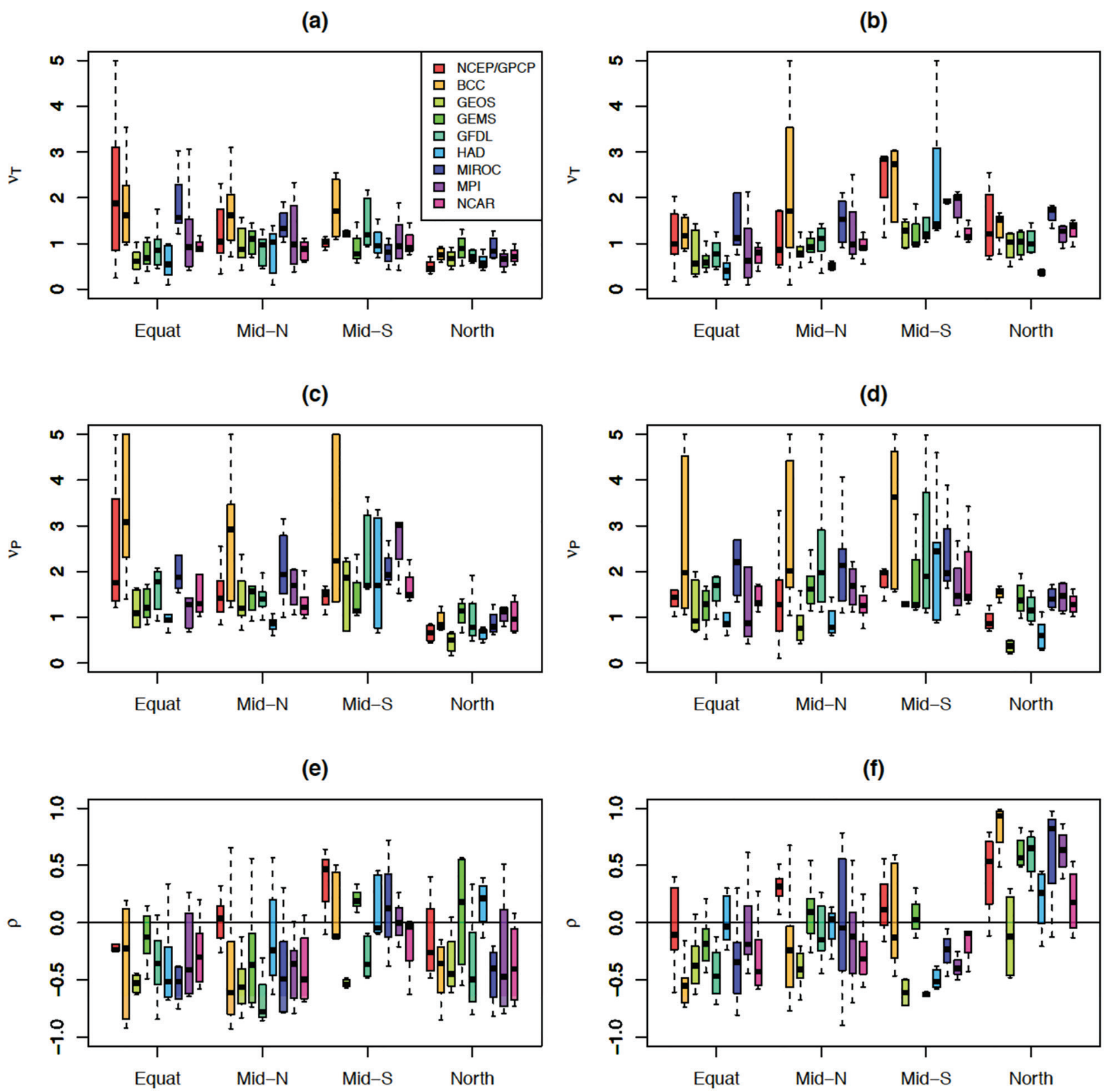


Figure 4. Climate model intercomparison results from Philbin and Jun (2015). The top row compares the smoothness of the temperature, the middle row the smoothness of precipitation, and the bottom row the cross-correlations for different regions of the Earth. The left column shows boreal summer and the right, boreal winter. Colors represent modeling groups as shown in the legend in 4(a).

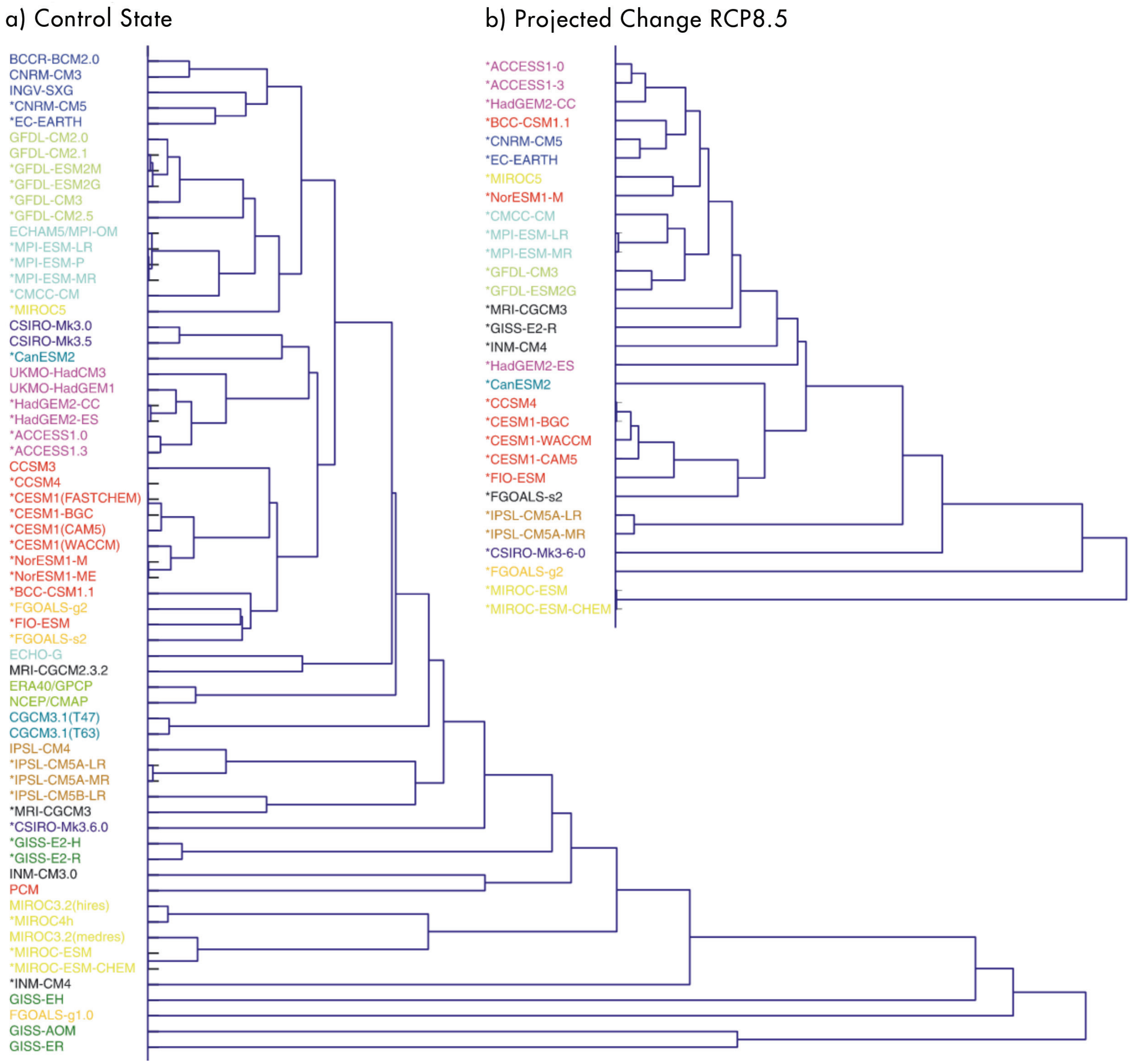


Figure 5. (a) The model "family tree" from CMIP3 and CMIP5 (the latter marked with asterisks). A "distance matrix" for the temperature and precipitation fields was used to determine the dendrogram. (b) Same as (a), except that distance is calculated based on the predicted change in temperature and precipitation for the end of the 21st century. Here, the control state refers to the emission scenario with no external forcing, and rcp8.5 refers to the emission scenario with high greenhouse gas emissions. Each climate model is represented by its official CMIP model name, a combination of the institutional id and its model name (e.g., IPSL—Institut Pierre-Simon Laplace-CM4 [one of IPSL's models]).

Source: Knutti, et al., 2013.

of independent information, but rather, a smaller amount of information due to their dependence—that is, their similarities.

However, it is not straightforward to define and quantify model interdependence for a given pair of climate models. For any given pair of climate models, how dependent (or similar) they are may be different for each climate variable, spatial region, and/or time period considered.

One notable work for understanding model interdependence is the "family tree" approach to climate models in Knutti, et al. (2013; see Figure 5). The concept of a family tree is easy to understand, although, as seen in Figure 5, the tree structure is quite complex in this case. As might be expected, most climate models developed by the same institution are related at low levels of the family tree.

In a more statistical sense, Jun, et al. (2008) quantified the dependence of a given pair of climate model errors using CMIP3 data. They considered 19 climate models from the CMIP3 archive and found that there are only about five *degrees of freedom* in those 19 climate models (in other words, the information provided by the 19 climate models is ultimately equivalent to the amount of information that could be provided by five independent climate models). Their study considered only the surface temperature variable as seasonal averages over three decades. Their conclusion might change if another climate variable or a different time scale (e.g., yearly average or monthly average) is considered.

Even if we can identify the actual degrees of freedom in an ensemble of climate models, it is not clear how we can incorporate this information into our study on climate change. We cannot simply identify five independent climate models among all the climate models considered. Instead, it is more likely that there are five linear combinations of all the climate models that are somehow independent of each other.

Furthermore, if a particular climate model is seriously wrong and has completely different outputs from the rest of the climate models, it could be mistakenly viewed as a valid, independent model (thus adding one degree of freedom to the entire pool of climate models). However, it most likely would not be useful (but, in fact, actually harmful) for the study of climate change.

Conclusions

Now that CMIP6 is under way (see Eyring, et al., 2016), scientists and statisticians are excited about seeing greater improvement in climate models over previous ones and, at the same time, face challenges that the new model archive will bring regarding model intercomparison.

New models will be much more complex, with more physical processes and parametrizations. They will use much finer grid resolutions (thus much bigger data). There also will simply be many more climate models. The combination of excitement and challenges that must be confronted will never end, but it will surely bring us a better understanding of the future climate and the human influence on climate change. C

Further Reading

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., and Taylor, K.E. 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. *Model Development* *9*:1,937–1,958.

Jun, M., Knutti, R., and Nychka, D.W. 2008. Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There? *Journal of the American Statistical Association 103*:934–947.

Knutti, R., Masson, D., and Gettelman, A. 2013. Climate model genealogy: Generation CMIP5 and how we got there, *Geophysical Research Letters 40*: 1,194–1,199.

Knutti, R., and Sedláček, J. 2012. Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Climate Change*. doi: 10.1038/NCLIMATE1716.

Philbin, R., and Jun, M. 2015. Bivariate spatial analysis of temperature and precipitation from general circulation models and observations, *Advances in Statistical Climatology, Meteorology, and Oceanography 1*:29–44.

About the Author

Mikyoung Jun is an associate professor in the Department of Statistics at Texas A&M University. She received her PhD in statistics from University of Chicago in 2005. Her main research interests are development of statistical models for physical processes on a global scale and their applications to various climate applications.

Climate Change Detection and Attribution: Letting Go of the Null?

Dorit Hammerling

Figure 1. Are events such as Hurricane Sandy linked to climate change, and will they become more common or intense in the future?

When Hurricane Sandy hit the New Jersey shore at the end of October 2012, news headlines abounded that questioned the connection to climate change. Sandy was the second-costliest hurricane in U.S. history, topped only by Hurricane Katrina (more-recent events, such as Hurricanes Harvey and Irma, may change that line-up). When such events occur, people understandably wonder: What is the connection to climate change? Has climate change made the occurrence of events such as Hurricane Sandy more likely, and will such events become even more frequent or more severe in the future?

Climate change can be loosely defined as a change in the average weather conditions that could influence the occurrence and intensity of storms, but scientists have mainly shied away from making such connections. The argument is along the lines that there is too much natural variability in the weather and climate system to attribute individual events, such as Hurricane Sandy, to climate change. There is, however, a whole line of research investigating the linkage between what can be observed over longer time periods and climate change, referred to as *climate change detection and attribution* or *optimal fingerprinting*. For some types of observations, this linkage can be clearly made using statistical methods.

Climate change detection and attribution is an established

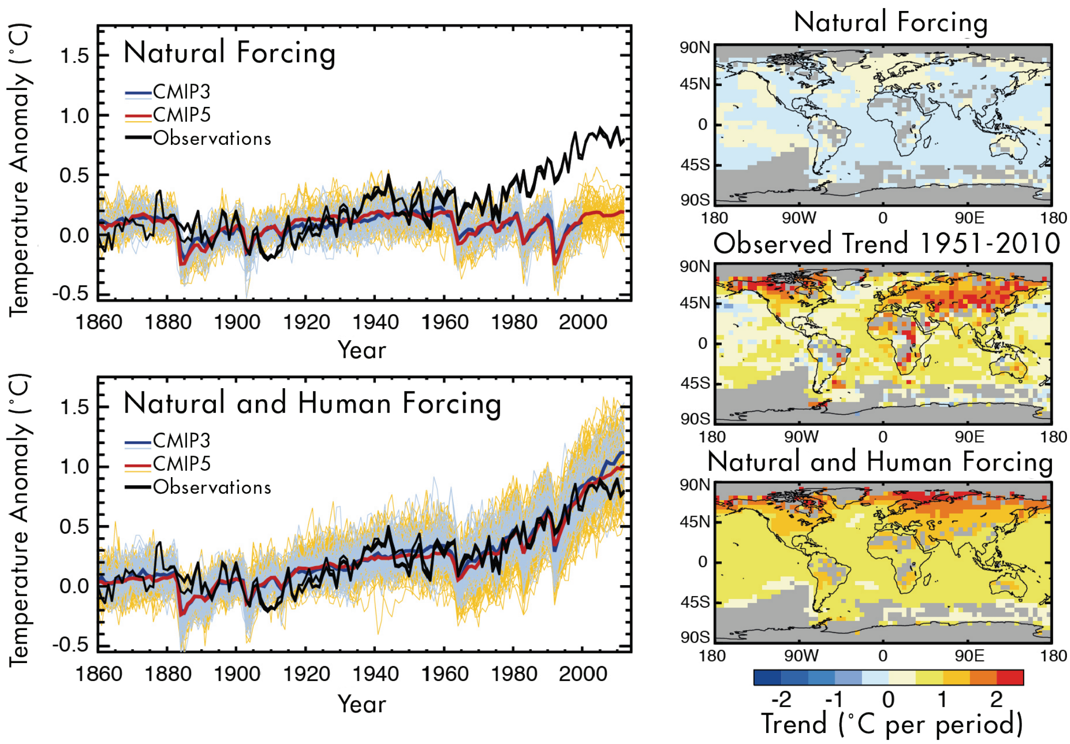


Figure 2. Example of a detection and attribution study, reproduced from FAQ 10.1, Figure 1, IPCC 2013: The Physical Science Basis. Time series of global and annual-averaged surface temperature change from 1860 to 2010. The top-left panel shows results from two ensembles of climate models driven only with natural forcings, shown as thin blue and yellow lines; ensemble average temperature changes are thick blue and red lines. Three different observed estimates are shown as black lines. The lower-left panel shows simulations by the same models, but driven with both natural forcing and human-induced changes in greenhouse gases and aerosols. Spatial patterns of local surface temperature trends from 1951 to 2010 (right). The upper panel shows the pattern of trends from a large ensemble of Coupled Model Intercomparison Project Phase 5 (CMIP5) simulations driven with just natural forcings. The bottom panel shows trends from a corresponding ensemble of simulations driven with natural + human forcings. The middle panel shows the pattern of observed trends from the Hadley Centre/Climatic Research Unit gridded surface temperature data set 4 (HadCRUT4) during this period.

methodology in the climate science community and, more recently, an area of research within the statistics community. The framework of detection and attribution is the tool set that lets the Intergovernmental Panel on Climate Change (IPCC) make statements such as, “It is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century.”

Figure 2 shows a typical example how climate change detection and attribution is used to establish a link between human—also referred to as anthropogenic—actions and observations. The main questions addressed are: Could our observations have happened by chance, or is there really a change? If there is indeed a change, can we attribute it to human actions?

From a data point of view, three main ingredients are required to answer this question: (1) Observational data sets and (2) results from climate models driven only by natural phenomena, and (3) results from climate models that also incorporate human-driven phenomena such as increased greenhouse gases. Getting at these three ingredients is not easy and requires worldwide resources and collaboration. Observations are taken through a variety of means, ranging from ocean buoys to weather stations to satellites, and are then exchanged and archived at specialized institutions throughout the world.

Climate models are complex numerical models based on physics that amount to hundreds of thousands, if not millions, lines of computer code to model the Earth for the past, present, and future. They are developed at scientific institutions worldwide and are constantly improved upon to reflect the Earth's physical systems in ever-increasing detail. Climate models are so complex and computationally intensive that they can usually only run on supercomputers. They incorporate a multitude of processes, ranging from oceans currents to river runoff, land use, storms, clouds, and sea ice.

The domain of the models is vast: from the bottom of the ocean to the outer layers of the atmosphere. As such, they allow for an endless number of studies to investigate what the Earth looked like in the past—for example, during the Ice Age—and what our future will look like as a function of different choices made in moving forward.

All three ingredients—observational data, results from climate models using only natural forcing, and results from climate models that also incorporate human forcing—are visualized in Figure 2. Looking at the black lines in the left-hand panels of Figure 2 shows that there are some differences between observational data sets coming from different institutions. Those differences in observational products, however, are small compared to the differences between runs from different climate models. These are shown by the yellow and light-blue lines in the left panels in Figure 2.

These differences stem from, among other things, different ways of implementing the physical equations and parameterizing processes, different spatial and temporal resolutions, and numerical schemes between the models. Even when only a single climate model is used, the climate model results can exhibit considerable dispersion, due to a fascinating phenomenon popularly known as the butterfly effect, where minuscule changes in the starting conditions can lead to quite different results in the outcomes.

Even though there is a wide spread among them, all the lines follow a common larger trend, and it is clear simply from visual inspection that models that are only forced by natural events cannot explain the observations, while the model runs also driven by human-induced changes in greenhouse gases and aerosols can.

The tool to assess such questions in a more statistically rigorous way is multivariable regression analysis, where multivariable refers to the fact that more than one predictor variable is investigated to explain the response variable. In most cases, the setup is also multivariate, which refers to the fact that the response variable is of a dimension larger than one. Typically, thousands of locations in space and time are considered. This model can be written as

$$y = x_1 * \beta_1 + x_2 * \beta_2 + \varepsilon,$$

where in our example, the response variable, y, is the observed temperature and the predictor variables are the climate model runs driven by natural forcings, x_1, and the climate models driven by human forcings, x_2, with their respective scalar coefficients, β_1 and β_2. The goal of the analysis is to determine whether $\beta_2 \neq 0$, which implies that climate change due to human activity has been detected.

A step further is to test for all the βs to be equal to unity, which then implies attribution to the specific set of forcing scenarios being investigated. The last term in the equation is the error term ε. This term is normally distributed with a mean of zero and covariance matrix of the same dimension as y. This covariance matrix describes the dependence structure of the errors.

While the model is in principle simple, practical issues often arise due to the high dimensionality of the problem. If a climate model runs at a spatial resolution of 1 degree in latitude and longitude, that results in a total 180 x 360 = 64,800 locations. The most-challenging technical detail is to estimate a covariance matrix corresponding to all these locations. The covariance matrix describes the relationship of the errors between different locations, which in a system like the Earth, with many interconnected processes, is critical.

What is done in practice is to aggregate the data to a coarser resolution. Resolutions typically used are 5 degrees by 5 degrees or 10 degrees by 10 degrees, but even using these aggregations, there is still a large covariance matrix to be estimated. This covariance matrix is estimated from so-called *control runs*. Control runs are climate model runs where no external forcing has been applied and only internal variability is present, but estimating large covariance matrices from noisy data is notoriously difficult. Just estimating this covariance matrix in an ad hoc fashion and plugging it in the model would neglect the uncertainty associated with its estimation.

This has, in turn, led to the recent methodological developments using so-called *Bayesian hierarchical models* to account for these uncertainties more comprehensively and translate them to the final estimates. Bayesian hierarchical models are a flexible class of models, where the model parameters themselves are considered random and uncertainties can easily be propagated due to the hierarchical setup of the model.

Apart from technical difficulties related to the covariance estimation, a more-fundamental problem with the current methodological setup is that changes in events related to atmospheric circulation, such as storms, cannot be characterized robustly due to their underlying chaotic nature.

This is in contrast to changes in thermodynamic state variables such as global temperature, which can be relatively well-characterized. The climate change signal in atmospheric circulation events, if it exists, is very small and easily drowned out by natural variability. However, it is still of societal importance to assess how the impact of events driven by atmospheric circulation might be altered by climate change present in the form of a new thermodynamic state.

Motivated by this conundrum, a group of researchers recently proposed a new way of viewing and formulating the problem. Rather than trying to assess the probability of the event occurring, they suggest viewing the event as given, and assessing to which degree changes in the thermodynamic state, which we know has been influenced by climate change, altered the severity of the impact of the event. This is de facto moving away from the current null hypothesis that there is no climate change.

Under the current null hypothesis, it takes strong evidence to dispute or overturn the hypothesis. One issue argued by the advocates of this new direction is that the current setup leads to many cases where we fail to reject the null hypothesis even though we should. In technical terms, such a scenario is referred to as a *type II error* or *false negative*. Given that climate change has been established as real and already has created a "new normal," requiring to prove this fact again every time is pointless, and worse, eliminates the power to answer the kinds of questions that are truly relevant to society.

In the case of Hurricane Sandy, this alternative framework has already been explored through an experiment by the European Centre for Medium-Range Weather Forecasts. In addition to the actual forecast, their researchers simulated the superstorm under otherwise-equal conditions by exchanging the unusually high sea surface temperatures, arguably caused by climate change, with the climatological values along the coastline of the eastern U.S. Although the storm track remained almost the same, the main features characterizing the storm diminished in their intensity when using the climatological values. Most notably, precipitation was 35% higher under the actual condition than the climatological average.

Given that most of the damage resulted from the flooding after the storm surge, it can be argued that the combination of higher sea levels due to climate change—and the unusually high sea surface temperatures also connected to climate change—were responsible for at least a portion of the estimated $71 billion in damages.

While a subsequent quantitative study linking a decreased storm surge to a reduction in damages hasn't been made as part of this experiment, it illustrates and provides the foundation of how it could be done. Thus, this new way of viewing the problem could be a game changer in the attribution of extreme events by providing a framework to quantify the portion of the damage that can be attributed to climate change, even for events that themselves cannot be directly attributed to climate change using traditional methods. C

Further Reading

Hegerl, G.C., Zwiers, F.W., Braconnot, P., Gillett, N.P., Luo, Y., Marengo Orsini, J.A., Nicholls, N., Penner, J.E., and Stott, P.A. 2007. Understanding and attributing climate change in Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.

Bindoff, N.L., Stott, P.A., Achuta-Rao, K.M., Allen, M.R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I.I., Overland, J., Perlwitz, J., Sebbari, R., and Zhang, X. 2013. Detection and Attribution of Climate Change: from Global to Regional in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Trenberth, K.E., Fasullo, J.T., and Shepherd, T.G. 2015. Attribution of climate extreme events, *Nature Climate Change 5*: 725–730.

Washington, W.M., and Parkinson, C.L. 2005. *An Introduction to Three-dimensional Climate Modeling*. Herndon, VA: University Science Books.

Katzfuss, M., Hammerling, D., and Smith, R.L. 2017. A Bayesian hierarchical model for climate change detection and attribution. *Geophysical Research Letters 44*:5,720–5,728. doi:10.1002/2017GL073688.

About the Author

Dorit Hammerling, PhD is the section leader for statistics and data science at the Institute for Mathematics Applied to Geosciences of the National Center for Atmospheric Research in Boulder, Colorado.

Quantifying the Risk of Extreme Events under Climate Change

Eric Gilleland, Richard W. Katz, and Philippe Naveau

We are all familiar with the proverbial "100-year flood," a concept instrumental in engineering design for water resources management. This concept, generally termed a "return level" corresponding to a specified "return period," is based on the assumption of an unchanging (or "stationary") climate. In particular, Emil Gumbel, a pioneer in the application of the statistics of extremes, cautioned as long ago as 1941:

> … to apply any theory we have to suppose that the data are homogeneous, i.e. that no systematical change of climate and no important change in the basin have occurred within the observation period and that no such changes will take place in the period for which extrapolations are made.

Today, with the rapid increase in greenhouse gases and its consequences on the climate system, including the possibility of increases in the frequency and intensity of extremes, the assumptions spelled out by Gumbel are no longer necessarily tenable. An open question is how best to convey the risk of extreme events under a changing (or "non-stationary") climate.

A further, related challenge concerns the quantification of the risk of the simultaneous occurrence of two or more extreme events (e.g., both extreme wave height and sea level in the case of coastal flooding), sometimes termed "compound" events. In the engineering design community, there has been some resistance to taking climate change into account in flood plain management, as well as reluctance to reconsider the concept of a 100-year flood under non-stationarity.

Quantifying Risk Assuming Stationarity

Even though our climate is changing, the assumption of a stationary or unchanging climate is a convenient starting point. Under stationarity, the 100-year flood corresponds to the 0.99 quantile of the probability distribution of the annual peak flow (or some other measure of flooding) of a river at a given location.

More generally, let p denote the probability of a flood in a given year. In that case, a T-year flood (i.e., the return level with a return period of T years) would correspond to the $1 - p$ quantile, where $p = 1/T$ (see Figure 1). The concept of a 100-year flood has several interpretations. Over any 100-year time period, there should be one 100-year flood on average. If we make the additional, and not too unrealistic, assumption that flood events in different years are probabilistically independent, then the waiting time on average until the next 100-year flood should also be 100 years.

Nevertheless, even under stationarity, the concept of a 100-year flood is commonly misunderstood by the public. On the one hand, 100 years could pass with no 100-year floods or, on the other hand, 100-year floods could even occur in consecutive years. To minimize confusion, the U.S. Geological Survey has advocated replacing the terminology "100-year flood" with "a flood with an average recurrence interval of 100 years."

To better convey the risk of flooding (e.g., to convince homeowners of the necessity to purchase flood insurance), the risk of a 100-year flood can be expressed in terms of the probability of one or more 100-year floods over some design lifetime L (e.g., the average life span of a house)—say, 30 years. Continuing to assume temporal independence in addition to stationarity, this "risk of failure" can be expressed from elementary probabilistic reasoning as

$$1 - (1 - p)^L = 1 - (1 - 1/T)^L = 1 - (1 - 1/100)^{30} \approx 0.26$$

In other words, there is roughly a 26% chance that a 100-year flood would be experienced at least once within a 30-year time period, or quite a bit more likely than many people would naively guess.

Of course, there is much uncertainty in estimated return levels for long return periods (such as 100 years) based on a limited time series of observations, especially only with extremes such as annual peak flow

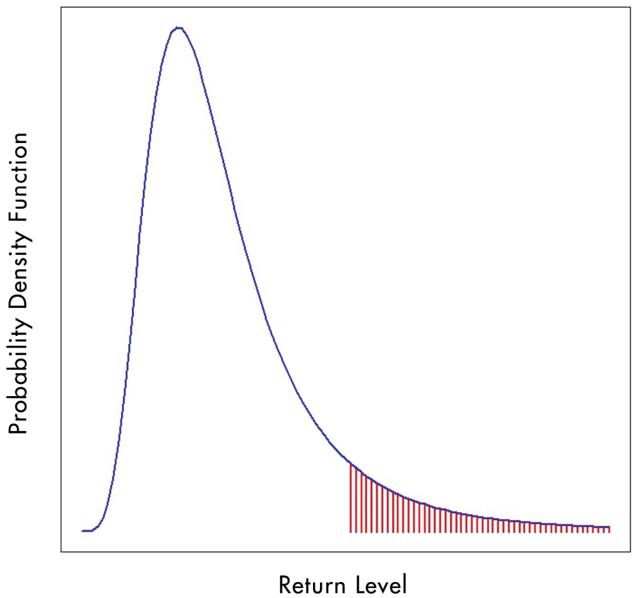


Figure 1. Probability density function for annual peak flow, with area (highlighted in red) under curve above return level equaling probability *p* of a flood in a given year.

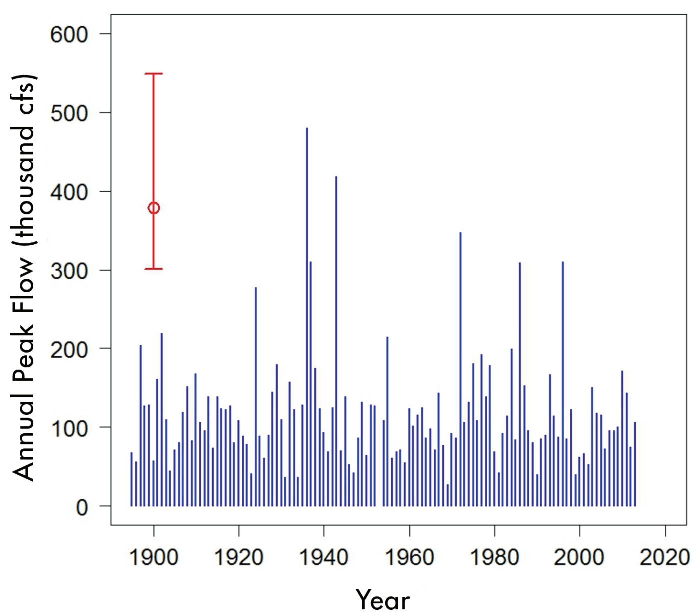
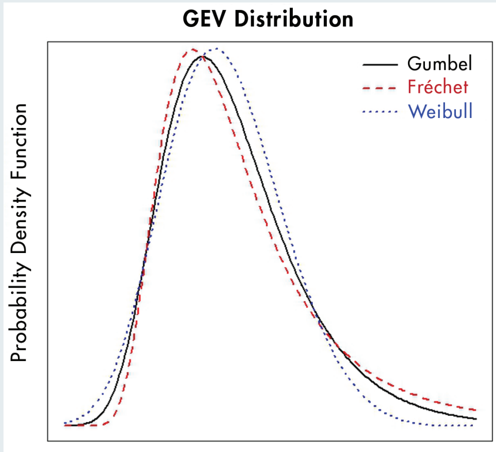


Figure 2. Time series of annual peak flow (1,000 cfs, for water year October–September) of Potomac River at Point of Rocks, MD, for period 1895–2013, along with point estimate (circle) and 95% confidence interval for 100-year flood (in red).

being extracted from the record (that is, only one value per year). Nevertheless, such uncertainty can be readily quantified using advanced methods based on the statistics of extremes.

As an example, Figure 2 shows the time series of annual peak flow (water year October – September) for the Potomac River at Point of Rocks, MD, during 1895–2013. To illustrate the stationary case, this example was chosen because of a lack of any obvious trend. Also superimposed on the figure is the estimated 100-year flood, based on fitting a generalized extreme value distribution (see sidebar), along with

GENERALIZED EXTREME VALUE DISTRIBUTION



The generalized extreme value (GEV) distribution arises as the approximate distribution for the maximum value of a long sequence of observations. It has three parameters: Location parameter governs the center of the distribution; scale parameter governs the spread of the distribution; and shape parameter, determined by the upper tail of the distribution of the individual observations. Depending on the sign of the shape parameter, the GEV distribution has three types: the Fréchet type with a "heavy" tail (i.e., slowly decaying in the form of a power law); the Weibull type with a finite upper tail; and the Gumbel type, intermediate between the Fréchet and Weibull types.

a 95% confidence interval (these calculations were performed using the extRemes package in the open source statistical programming language R). Besides being relatively wide, a disappointing (but not surprising) feature of this interval for practitioners concerns the need to hedge more on the upper end, or precisely the values that matter the most for engineering design.

Quantifying Risk under Climate Change

Evidence for shifts in the frequency and intensity of extremes (e.g., for precipitation, the major driver of floods) as part of climate change is already strong. Some of this evidence is empirical in the form of statistically significant trends, while other evidence is theoretical in the form of projections of climate change based on numerical models of the climate system. Yet, there has been heated debate in the hydrology community about whether stationarity is "dead."

On the one hand, some argue that water resources management based on the stationarity assumption is obsolete and should be completely abandoned. On the other hand, others argue that, even given statistically significant trends in extremes, the uncertainty is too great to justify modifying engineering design to take non-stationarity into account. Some of this resistance is based on legitimate concerns about the difficulty of extrapolating any trends into the future. Still other opposition seems recalcitrant, just using uncertainty as an excuse to avoid taking any action.

In some situations (such as when there are too many parameters to estimate relative to the number of observations), fitting simpler statistical models may be preferable based on the principle of parsimony (or Occam's razor). Yet, in the climate change context, ignoring apparent trends in extremes does not appear to be a defensible strategy.

As an example of a real shift in floods over time, Figure 3 shows the time series of annual peak flow (water year October–September) at Mercer Creek, WA, during 1956–2010, a small drainage basin that underwent rapid development in the 1970s. As could have been anticipated from basic hydrologic principles, a rapid increase in the annual flood magnitude occurred during the 1970s and early 1980s, even though the average flow showed no evidence of change. Also superimposed on the figure are the estimated 20-year floods, based on fitting the generalized extreme value distribution separately to the two time periods 1956–1977 and 1978–2010, along with 95% confidence intervals (20-year floods are shown instead of 100-year floods because the two time periods are so short).

The question could be raised of how best to quantify the flood risk for Mercer Creek starting in the late 1970s. First of all, it would not be feasible to modify the flood plain gradually on an annual basis. Further, adopting a risk measure that extends or generalizes either the expected frequency or expected waiting time interpretations of a 100-year flood to non-stationarity would not necessarily be straightforward. Instead, the same concept risk of failure, as discussed under stationarity, could provide a more-meaningful measure of flood risk over a specified future design life time, but now the probability of an

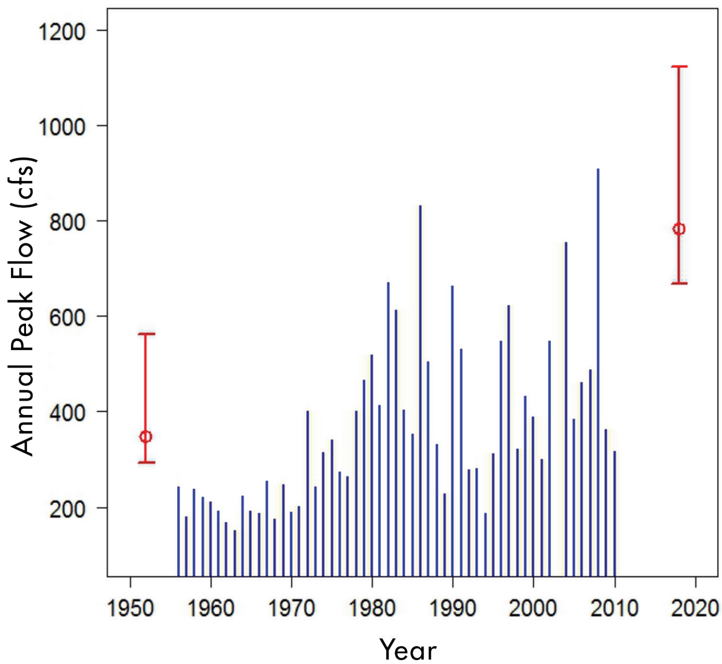


Figure 3. Time series of annual peak flow (cfs, for water year October–September) at Mercer Creek, WA, during 1956–2010 (*Note*: Peak flow is missing for year 2003), along with point estimates (circles) and 95% confidence intervals for 20-year floods (in red) for 1956–1977 (left) and 1978–2010 (right).

extreme event p would vary depending on the year.

Return Periods in a Multivariate Context

A return level of interest often concerns a more-complicated event than simply being above a high value (or below a low value). For example, the severity of a drought depends on multiple factors, such as the location, length of time without precipitation or other source of water, temperature, and possibly other conditions. Even when such an event can be defined reasonably, it will often depend on more than one variable.

To illustrate, Figure 4 shows four of the most classical cases of bivariate extreme events. Each panel leads to a different return period despite the same high threshold being involved. Practitioners should be aware that such return periods are relative quantities and cannot be compared whenever the event of interest differs. Deciding and clearly stating which event is the object of interest may be the most important step of a risk analysis, but it is often overlooked. Bypassing this step often leads to misguided conclusions.

It is very tempting to compare the return period of a single event, such as a drought defined in terms of extreme low precipitation, with the return period of a joint event, such as both extreme high temperature and extreme low precipitation. However, as recently reported by Francesco Serinaldi, a hydrologist at Newcastle University, univariate and multivariate return periods are not comparable. Nevertheless, research by Amir AghaKouchak, a hydrologist at the University of California Irvine, and colleagues involved precisely this type of analysis and claimed that a univariate return-period analysis substantially underestimates the occurrence probability of the 2014 California drought when ignoring the effect of temperature.

One simple reason for this potentially misleading statement is that the geometrical structure of any type of joint event makes it fundamentally different from a univariate one. For example, compare the green area in the upper left panel of Figure 4 to the colored areas in the other panels, and note that it is always different from them and, therefore, not comparable in terms of return periods.

In this context, it has been argued that the definition of return period is meaningless because it can create confusion, especially in a multivariate setting. While it is true that "apples" should not be compared to "oranges," simply dismissing the entire concept of a return period is tantamount to throwing the baby out with the bathwater. For planning purposes,

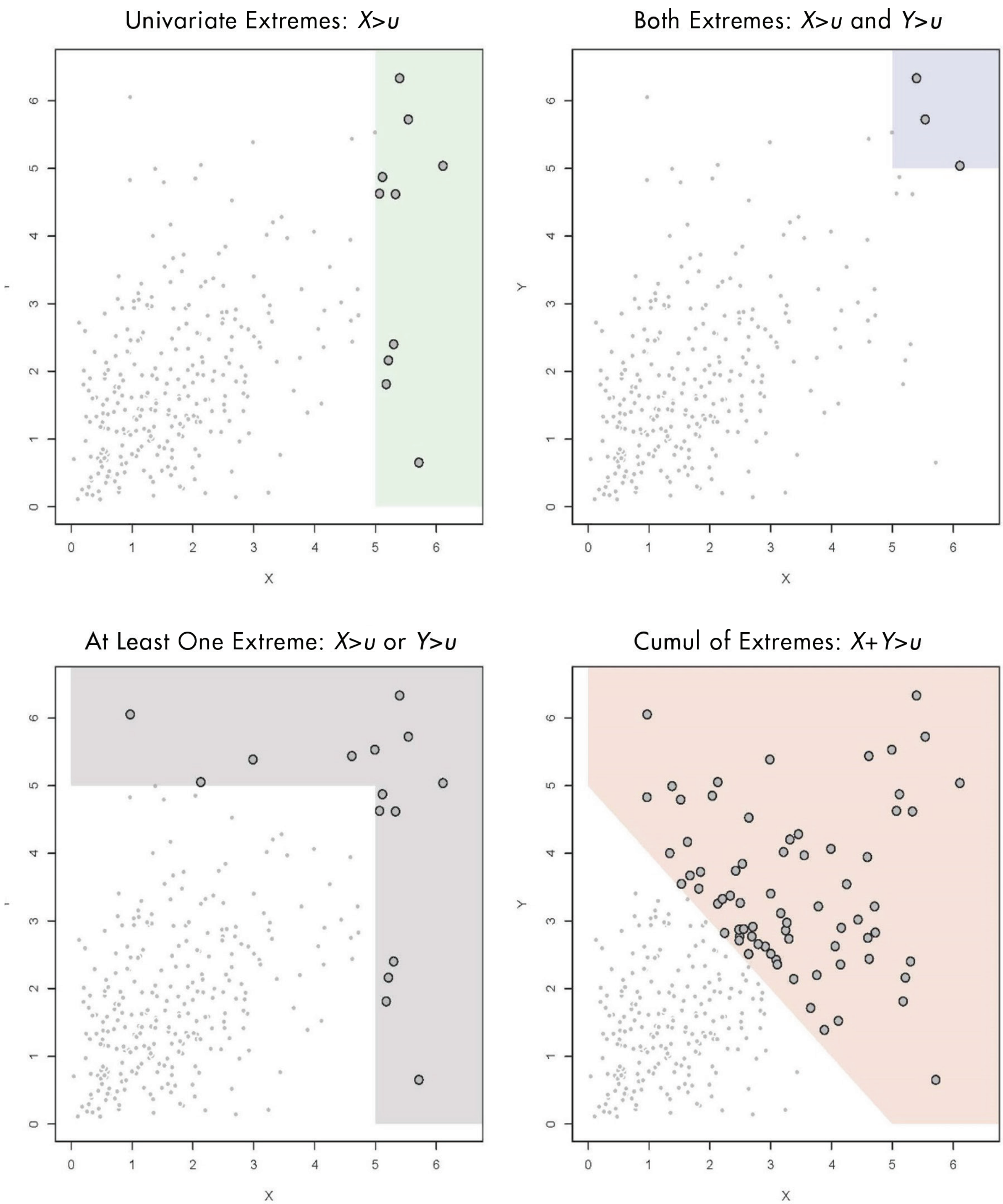


Figure 4. Four classical types of events in the same bivariate setting, given the same threshold $u = 5$. Upper left panel ($X > u$), upper right panel ($X > u$ and $Y > u$), lower left panel ($X > u$ or $Y > u$), and lower right panel ($X + Y > u$). For simplicity, it is assumed that the two random variables, X and Y, have the same marginal distribution so the same threshold u can be used.

some form of risk measure remains essential.

The key point to reduce incorrect claims is to simply remind risk managers that a return period should never be interpreted alone—that is, without specifying the associated event of interest. Despite the challenges, the exercise can even be carried out in a multivariate framework.

For example, an event can be defined as the joint probability that all variables of interest, say two random variables X and Y, fall above a specified amount, say a high threshold u (blue points in upper right panel of Figure 4). In that case, there is no quandary in setting this probability p to $1/T$ and solving for the value of u that satisfies the equation. Subsequently, T can be interpreted as the return period for this specific event.

The event can easily be changed by re-defining the probability. For example, the event could be considered to be that the sum of all variates is greater than the threshold u (brown points in lower right panel of Figure 4). Then the return period simply need be interpreted within the context of this sum.

One delicate statistical point is to assess how the joint probabilities of bivariate extremes (the event that both X and Y exceed the threshold u; see blue area in the lower panel of Figure 5) can change, even when the marginal distributions are identical (here, standard Gaussian). The panels of Figure 5 display simulated bivariate realizations from two different dependence types (Gaussian and one based on multivariate extreme value theory, MEVT). As the threshold u increases, there is asymptotic independence or lack of "clustering" at high levels for the Gaussian case, whereas there is asymptotic dependence or "clustering" at high levels for the MEVT.

In other words, any Gaussian bivariate vector is unable to

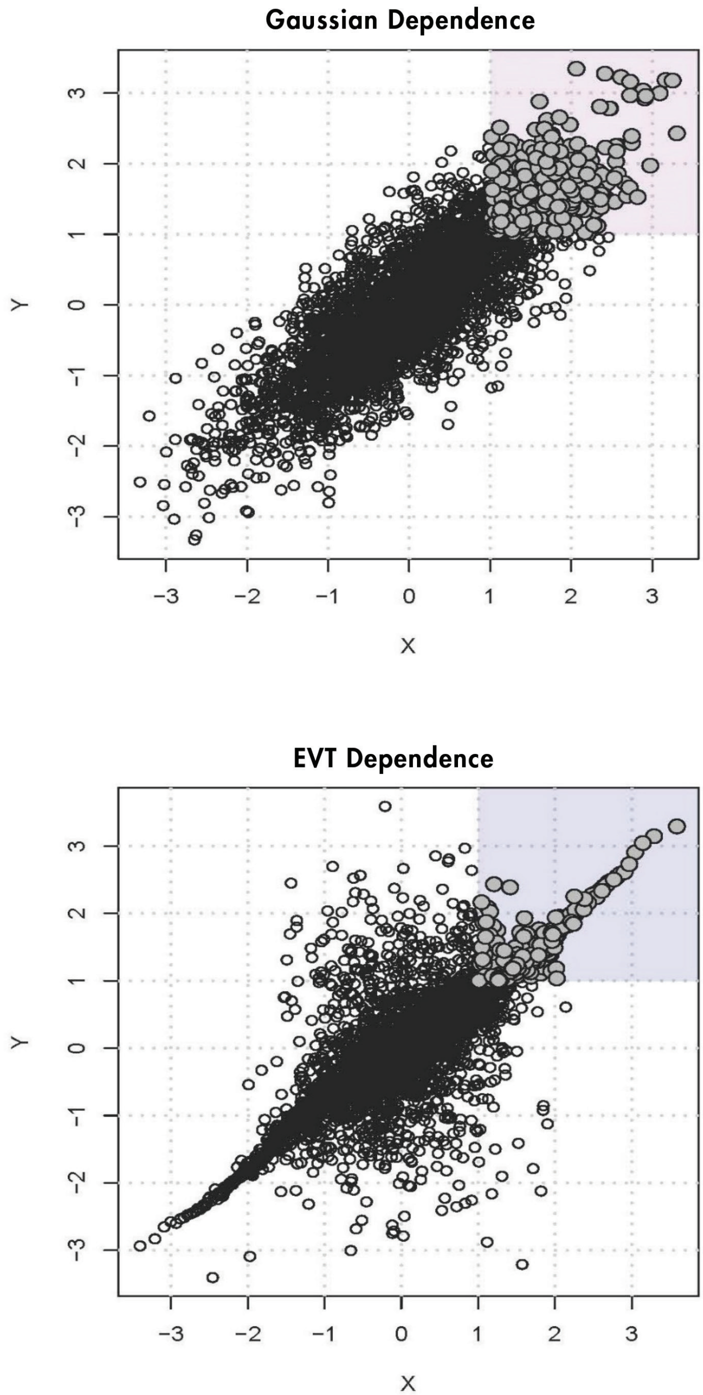


Figure 5. Comparison between asymptotic independence and dependence for bivariate extremes (30,000 simulated realizations). The upper panel corresponds to a linear relationship based on Gaussian distributions with a correlation coefficient of 0.8. The lower panel represents the same form of linear relationship, but instead based on Cauchy distributions.

model strong dependence among extremes, so other statistical models, especially those based on MEVT, would provide a potentially better way to represent extreme dependence. There is much evidence of such dependence in climate extremes (e.g., between extreme high temperatures on consecutive days at the same location or on the same day at nearby locations).

Outside the context of climate, a compelling example of the danger of relying on Gaussian dependence for extremes concerns the so-called "formula that killed Wall Street." Financial engineers were unaware that a formula based on the correlation coefficient would substantially underestimate the simultaneous risk of default for two assets. This misconception contributed to the worldwide financial collapse in 2008.

One added complication is that severe climate events can occur when some (or even all) variables involved assume states not necessarily considered extreme. For instance, severe flooding can occur when not-so-intense rain falls over frozen ground or snow pack, as has recently occurred at Yosemite National Park in California.

Comparing Return Periods in Climate Studies

To deal with non-stationarity issues in climatology, ensemble runs from numerical climate models offer a way to study independent draws from different types of climate worlds. In this context, an emerging topic in the field of *Detection and Attribution* (D&A) is the so-called event attribution paradigm. The archetypical example is the 2003 heat wave over Europe, where the question of interest concerns whether or not the return period of such a hot summer would have been different under a stationary climate; that is, without the influence from anthropogenic forcings or increases in greenhouse gases.

To address this question, climatologists run experiments with numerical models of the climate system and estimate a measure known as the *Fraction of Attributable Risk* (FAR). The FAR is defined as the relative ratio of two probabilities of the same event under two different alternate realities: the world that is and that which might have been. The common event is defined by exceeding a fixed threshold u, so the FAR is simply a relative ratio of return periods for the same event. For the European heat wave example, the FAR was estimated to be greater than 0.5, or more than double the risk of such an event being attributable to anthropogenic factors.

Hydrologists often play a similar game, but they fix the return period and estimate the associated return level. A risk manager may subsequently need to revise the return period under a warming climate.

In any case, one common challenge for all these communities is to move away from a single-event analysis based on one variable and unrealistic stationary assumption toward more-complex events (e.g., concurrent extremes) in a non-stationary world. Such a revision will lead to rich statistical questions; for example, how to invent novel statistical summaries that are easily interpretable, inferable, and comparable, and can generalize the old concept of return levels. C

About the Authors

Eric Gilleland is a project scientist in the Research Applications Laboratory at the National Center for Atmospheric Research. He has interests in spatial and extreme-value statistical analyses, as well as forecast verification methodology.

Richard W. Katz is senior scientist emeritus at the Institute for Mathematics Applied to Geosciences at the National Center for Atmospheric Research. He is interested in the application of statistics to the environmental and geophysical sciences, especially to weather and climate extremes. He has served as the chair of the ASA Advisory Committee on Climate Change Policy.

Philippe Naveau is a senior scientist at the Laboratoire des Sciences du Climat et de l'Environnement (LSCE) at the Centre National de la Recherche Scientifique (France). His main research interests focus on environmental statistics, especially the modeling of climate and weather extremes.

Further Reading

AghaKouchak, A., Cheng, L., Mazdiyasni, O., and Farahmand, A. 2014. Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 California drought. *Geophysical Research Letters 41*:8847–8852, doi: 10.1002/2014GL062308.

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., and Stouffer, R.J. 2008. Stationarity is dead: Whither water management? *Science 319*:573–574.

Rootzén, H., and Katz, R.W. 2013. Design Life Level: Quantifying risk in a changing climate. *Water Resources Research 49*. doi: 10.1002/wrcr.20425.

Serinaldi, F. 2016. Can we tell more than we can know? The limits of bivariate drought analysis in the United States. *Stochastic Environmental Research and Risk Assessment 30*:1691–1704.

Serinaldi, F., and Kilsby, C.G. 2015. Stationarity is undead: Uncertainty dominates the distribution of extremes. *Advances in Water Resources 77*:17–36.

Stott, P.A., Stone, D.A., and Allen, M.R. 2004. Human contribution to the European heatwave of 2003. *Nature 432*:610–614.

Statistics and the Future of the Antarctic Ice Sheet

Murali Haran, Won Chang, Klaus Keller, Robert Nicholas, and David Pollard

One of the enduring symbols of the impact of climate change is that of a polar bear drifting in the sea, alone on its own piece of ice. For those who are left untouched by the loneliness of drifting polar bears, images of partially submerged lands and the devastation wrought by storm surges showcase some potentially frightening impacts of sea level rise on human life. The threat of sea level rise, in turn, is linked to the melting of ice sheets. Ice sheets are, therefore, important to understanding our planet, as well as learning about how our future may be affected by climate change. A promising approach to improving our understanding of ice sheets and derive sound projections of their future is to combine ice sheet physics, statistical modeling, and computing.

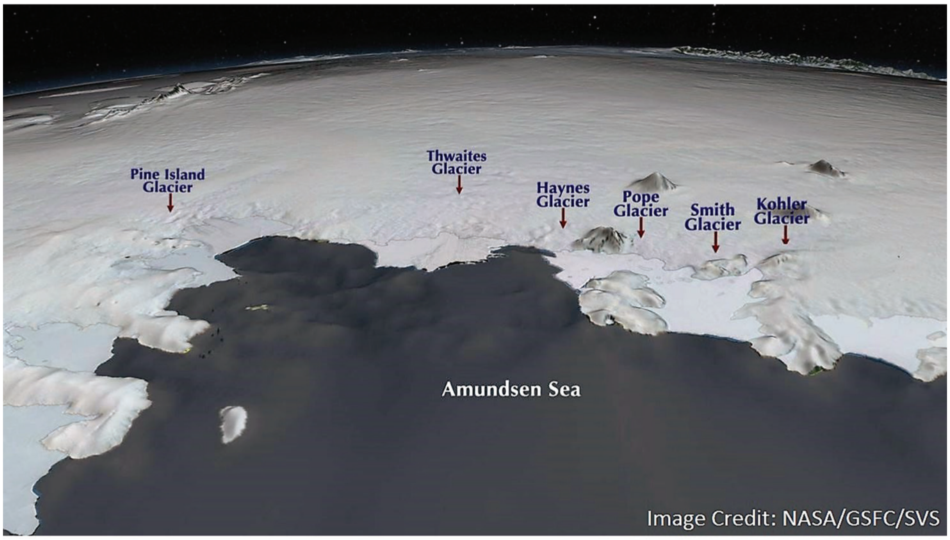


Figure 1. The West Antarctic Ice Sheet, viewed from the Amundsen Sea, with important glaciers highlighted.

First, what exactly is an ice sheet? It is an enormous mass of glacial land ice, more than 50,000 square kilometers in extent. The Antarctic ice sheet extends over 14 million square kilometers while the Greenland ice sheet extends over 1.7 million square kilometers.

To put this in perspective, the area covered by the Antarctic ice sheet is comparable to the continental United States and Mexico combined. In fact, the Greenland and Antarctic ice sheets contain more than 99% of the freshwater ice in the world. Roughly speaking, melting the entire Greenland ice sheet would result in sea level rise of around 7 meters (23 feet) while if the entire Antarctic ice sheet melted, it would result in sea level rise of around 57 meters (187 feet).

It is easy to imagine how even partial melting of these giant ice sheets can potentially lead to large sea level rise, making, for instance, low-lying coastal regions more vulnerable to future storm surges. Hence, a number of high-profile articles and documentaries have placed melting ice in polar regions squarely at the center of the discussion of the impacts of climate change. Knowing something about how ice sheets are changing has very practical consequences, such as when making decisions about how and where to build infrastructure on the coasts, and how to assess risk to property due to climate change in the future.

Risk is defined in terms of probabilities (risk of an event = expected loss under a probability distribution on that event), carefully describing the risks associated with a climatic event requires estimating future probabilities. That makes studying the future of ice sheets in a statistically sound fashion of interest from a both scientific and a policy and decision-making perspective.

How do scientists study the future of ice sheets? What role (if any) does statistical thinking play in studying ice sheets?

A careful study of ice sheets involves four main pieces: (1) physics for modeling the behavior of the ice sheets over time; (2) computing and applied mathematics (mostly solving differential equations) for translating the physical model into computer simulation code; (3) data sets that are informative about the past and current state of the ice sheet and related climate variables, such as ocean temperatures and

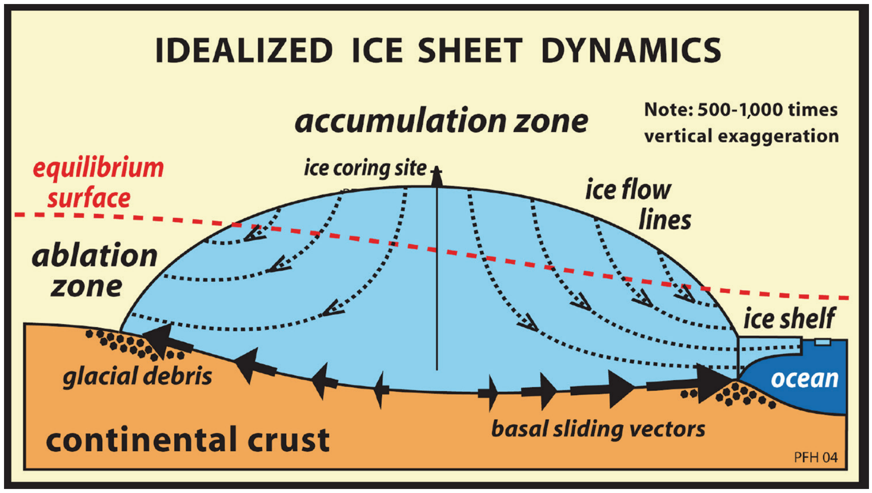


Figure 2. The arrows in the figure illustrate the direction of ice flow. Ice accumulates on top through precipitation (snowfall), and flows downward due to gravity. Ablation means the melting or evaporation of ice. This cartoon illustrates how the ice sheet rests (and slides) on the continental crust, and points out its important interaction with the ocean waters. Parameters (inputs) that determine how the ice sheet slides, and how it interacts with the surrounding ocean waters, are key to future projections of the ice sheet.
From *www.snowballearth.org*, courtesy Paul Hoffman.

snowfall accumulation; and (4) statistical methods that combine information from the physical model with observations of the ice sheet.

This is an interesting and challenging area of research because sound scientific research requires an interdisciplinary collaboration between ice sheet modelers and statisticians. The statistical challenges involve combining information from disparate sources such as the physical model and observational data sets. The size and complexity of the data and models pose serious computational challenges.

This article provides a taste of some of the interesting scientific questions about ice sheets, the resulting statistical problems, and an outline of a statistical method that can be used to solve this problem. The discussion is broadly targeted at ice sheets, but we focus predominantly on the West Antarctic Ice Sheet (WAIS) and PSU3D-ICE, Pollard and DeConto's ice sheet model.

These four aspects involved in the study of ice sheets are common to many other research areas in climate science where models, uncertain parameters, and multiple sources of observations have to be brought together to understand the past, present, and future state of the climate.

In fact, similar statistical problems arise often in other scientific disciplines where complex dynamical models are used, and the applications of the statistical methods of emulation and calibration described here even extend to many manufacturing and engineering processes.

The Physics of Ice Sheets

Ice sheets are created by long-term snowfall accumulation. When snowfall exceeds snow melt each year, it builds layer upon layer of snow, the weight of which compresses the underlying snow to form ice. Over thousands of years, this has resulted in massive ice sheets that can be thousands of feet thick.

The flow of the ice sheet is due to the height of this thick ice and snow. Ice sheet experts have worked extensively on building physical models that describe how ice sheets flow and evolve over time. Figure 2 provides a simplified view of the physics involved. It shows, for instance, that the ice flows downslope from the highest central regions toward the edges of the ice sheet.

Figure 2 provides a sense of how the ice sheet rests on the continental crust and how the ocean interacts with the ice sheet. The multiple parallel curves represent different ice flow lines, corresponding to different heights of ice (central regions of the West Antarctic Ice Sheet [WAIS] are more than 2,000 meters high). Gravity is a fundamental driver of the flow, causing stresses and deformation that tend to flatten the sheet surface over tens of thousands of years.

The basic physical principle underlying the ice sheet is the conservation of mass, which ensures that the local thickening or thinning of ice is balanced by ice added or removed. Ice can be added by snowfall, and is removed by *ablation*, the process by which snow or ice melts and flows away in stream-like channels and crevasses, evaporates, or is blown away by the wind.

The underlying topography, as well as the slipperiness of the bedrock surface, also influences the behavior of the ice sheet. Overall, the ice sheet surface tends to be a smooth dome, but high peaks may protrude though the ice, exposing the land in places.

The edge of the ice sheet is particularly vulnerable as it interacts with the ocean. Where the ice sheet meets, or *abuts*, the ocean, it can form a vertical cliff or can continue as an ice shelf—a floating table of ice hundreds of meters thick, flowing out toward the open ocean. Sometimes part of the ice shelf breaks off (known as *calving*) to create floating icebergs. Individual calving events can be dramatic and spectacular, especially if the ice cliff abutting the ocean is relatively tall.

This brief outline of ice sheet dynamics shows that there are many inputs or parameters of the model that, when changed, can have a considerable influence on ice sheet behavior. For instance, the slipperiness of the bedrock surface—the amount of friction between the ice sheet and the bedrock surface—affects how fast the ice sheet slides over it.

The ocean melt coefficient is a parameter that describes the sensitivity of the ice sheet to temperature changes in the surrounding ocean. Changes to this parameter will cause the ice sheet to react very differently to the changes in the surrounding ocean temperatures. Different parameter values will result in very different projections of the future of the ice sheet. Figuring out reasonable parameter values to use is, therefore, a very important research problem, and it makes sense to find parameter values that allow the ice sheet model to credibly reproduce both the past and current behavior of the ice sheet.

In fact, parameter inference is precisely the problem focused on here. Careful science requires not only providing "best" values of the parameters (point estimates) but also providing uncertainties about the parameter values.

Computer Models for Studying Ice Sheets

To study how the ice sheet behaves under various parameter settings and the impact of external climate variables or external *forcings* (physics external to the system that affect the ice sheet) on the ice sheets, scientists create computer programs that incorporate the physics of the ice sheet, as well as the various forces acting on it. These days, using computer simulations to learn about the behavior of an ice sheet in response to internal and external conditions is common in the earth and atmospheric sciences, and is often used in many science and engineering problems.

In our work on the West Antarctic Ice Sheet (WAIS), we use the PSU3D-ICE model, which strikes a balance between detailed physical modeling and computational efficiency. This balance allows it to produce realistic long-term behavior of the ice sheet without attempting to incorporate very-high-resolution physical modeling. This allows the long runs to be accomplished with a reasonable amount of computational effort.

Many decisions have to be made about how to run the ice sheet model. For instance, an important choice is to determine how far back to start the ice sheet model to "spin it up" to the present time (we start it 40,000 years before present). The spin-up phase of the model involves running it until it reaches a "steady state" that does not, ideally, depend too much on the initial values chosen to run the model.

Another choice is the kind of external forcings (physics external to the system that affect the ice sheet) to use on the ice sheet dynamics; we use well-established data sets and models to provide the atmospheric and oceanic external forcings. The computer model output is in the form of a spatial grid. Therefore, we also must determine the resolution at which we want model output, with a higher resolution typically taking more computational time. Here, we simply obtain information that is close to the same scale at which the observational data sets (described below) are available.

Finally, and crucially, we must determine a study design that suggests which parameter values to use when running the model, since we are constrained by computational considerations.

Ice Sheet Data

Detailed modern observations of WAIS are constructed combining many different types of observations, including satellite altimetry, airborne and ground data surveys, and ground radar surveys. These data are useful for learning about (referred to as *constraining* in the geosciences literature) important parameters of the model. However, to obtain better projections of WAIS on the scale of hundreds to thousands of years in the future, it is important to also use the long-term behavior of the ice sheet to learn about the parameters. The parameters inferred must be capable of producing realistic behavior of the

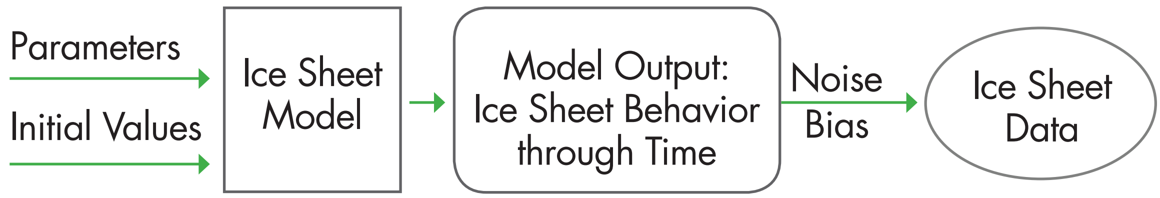


Figure 3. Parameters and initial values drive the ice sheet model. Its output describes the behavior of the ice sheet through time. Because this is an imperfect model, we account for noise (measurement error) and biases (missing processes in the model) to develop a model for the ice sheet observations. The ice sheet model is a "black box"—we only see model output for any given set of parameters. Example parameters include those that determine the basal sliding of the ice sheet and the sensitivity of the ice sheet to the surrounding ocean water temperatures. Emulation approximates via a Gaussian process of how this ice sheet model maps parameters into model output. This approximate model, combined with a model for error and bias, is used as a statistical model for the ice sheet data (observations) on the far right.

ice sheet over much longer periods of time.

Data from the distant past, going back hundreds of thousands of years or more, are reconstructions of the ice sheet's past. These are based on recent measurements, such as sonar data about ocean floor features as well as shallow sediment cores, which have been used by researchers to provide maps of approximate grounding lines—the location where the ice sheet transitions from lying on bedrock to hanging over the ocean—at 5,000-year intervals, from 25,000 years ago to the present. Hence, these resulting data are in the form of time series.

What Makes This a Challenging Statistics Problem?

It should already be clear that this is a statistical problem. Multiple data sets are involved, after all, and there is an interest in inferring parameter values and making predictions. Let us consider two important scientific questions: Given the recent satellite observations of the ice sheet and the paleo-reconstructed data about the ice sheet in the distant past, what are the likely values of the ice sheet model parameters? What can we say about the future of the ice sheet based on what we know about the past?

We can translate the first question into the language of statistics and probability: Given the two data sets and what we know about the ice sheet model by running it at various parameter settings, what is our estimate of the probability distribution of the model parameters? The probability distribution captures our knowledge about the parameters given what we knew about the parameters (prior scientific information) and what information the model runs and the observations provide about the parameters.

This fits naturally into the language of Bayesian inference which allows us to combine prior information with information from the data to obtain a posterior distribution of the parameters.

An advantage of this approach is that once we have an estimate of the posterior probability distribution of the model parameters, it can be used to answer the second question. Essentially, we need only see what the ice sheet model projections look like at various parameter settings, and weight the probability of these projections according to the posterior distribution of the parameters.

To summarize, we have: (1) a deterministic computer model that describes the ice sheet behavior as a function of parameters, but only simulations of this model at a limited number of parameter settings, and (2) observations of the ice sheet, both modern satellite ones and paleo-reconstructions of the ice sheet from the distant past. We need to formulate a *statistical* model that combines all of this information, while allowing for measurement errors and imperfections in the computer model.

The formulation of the statistical problem may seem pretty standard, except for one important twist: For Bayesian inference, we need both a prior distribution of the parameters and a probability model that connects the observations with the parameters. More specifically, the probability model provides a distribution for the observations—in this case, the satellite data and the paleo-reconstructed data—at each parameter value. This probability model is used to obtain a likelihood function, and then the rest of it is, modulo computational challenges, routine Bayesian inference.

Here, the only connection we have between the parameters and the observations is via the ice sheet model. This poses some challenges: (1) The model is deterministic, not probabilistic, so it does not provide a probability model on its own; (2) We only see the model output

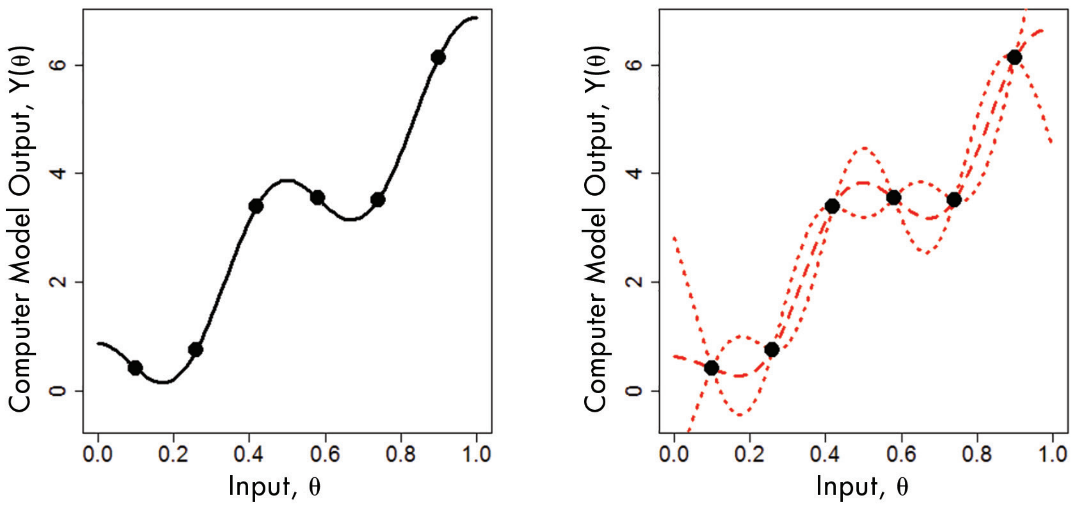


Figure 4. Emulation for a toy example: Black dots correspond to input parameters for which the computer model was run (left); red dashed lines are interpolations by a Gaussian process—they provide approximate computer model output at every parameter value (right). The dotted red curves correspond to uncertainties; there is greater uncertainty as we get further from places where we have data.

at a few (relatively small number of) parameter values; and (3) We know that the ice sheet model is an imperfect representation of the observations. The problem and an outline of how we can think about solving it are summarized in Figure 3.

What makes the problem challenging is the fact that the output from the model is high-dimensional and in the form of spatial or temporal data. These data are not always not easily modeled using Gaussian models. New, computationally efficient statistical methodology is, therefore, necessary for addressing these issues.

Ice Sheet Model Emulation and Calibration

How do we solve this problem? We can do this in two main stages.

We first approximate the ice sheet model with a statistical model; that is, we develop a statistical model that can predict how the ice sheet model will behave at new parameter values. Think of this problem as needing a flexible regression-type approach: Given many predictors (various parameter settings) and corresponding model output (responses), the model output at new predictors (any new parameter setting) can be predicted, along with some uncertainty about the model output. This kind of uncertainty may be referred to as epistemic, meaning that the uncertainty arises from our lack of knowledge (*episteme* is Greek for "knowledge") about what the model will do, *not* the fact that there is anything random associated with the model—it is deterministic.

This process of approximating the model is called *emulation*. Emulation results in a probability model that links the parameters of the ice sheet model to the output of the ice sheet model. The statistical model used for emulation is a *Gaussian process*, a popular model in spatial statistics, which is well-suited to interpolating functions.

Consider the simple example in Figure 4, where we consider a collection of random variables that are a function of a single parameter. There are, of course, infinite such random variables on any given range of parameter values—say, between 0 and 1. A Gaussian process model states that any finite collection of random variables—for example, the six function values between 0 and 1 (black dots on Figure 4)—has a joint normal distribution. Crucially, the dependence among the random variables decreases as a function of the distance between them, making two random variables that are close to each other (in parameter value) more dependent and, thus, more alike.

This suggests how Gaussian processes provide a useful approach for interpolation: The predicted value for a random variable at any parameter value, such as a function value in between the six black dots, is more like—more dependent on—values that are close to it, and depends less on values that are far from it. The precise dependence between the random variables at various parameter

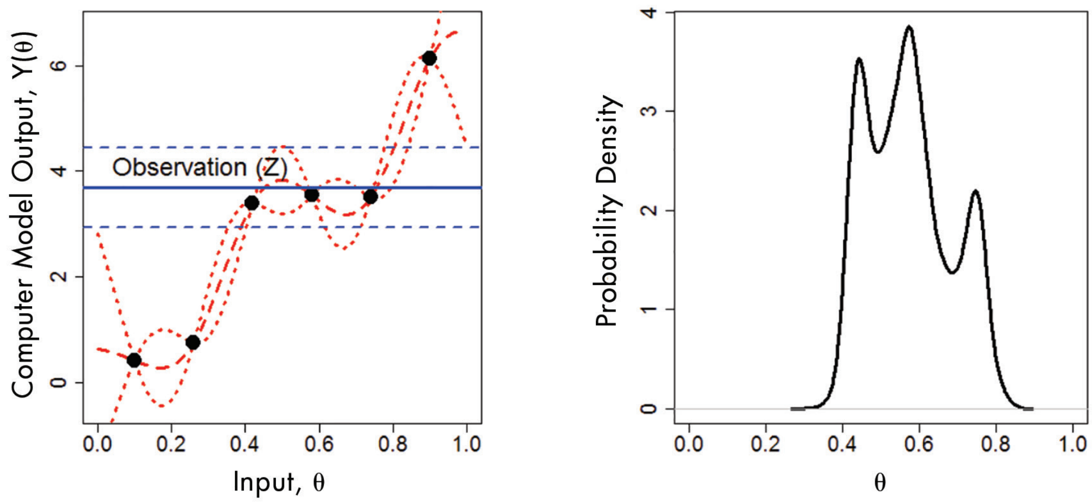


Figure 5. Calibration for a toy example. The blue horizontal line (left) represents a single data point. Calibration attempts to find parameter values that are compatible with that observation while taking into account uncertainties due to variability represented by the blue dotted lines). Bayesian inference provides the (posterior) distribution on the right, which summarizes what we know about the parameter. There are three peaks in this density, corresponding to three black dots (left figure) closest to the observations.

values is controlled by a covariance function that describes how covariances change as a function of distance. Hence, Gaussian processes provide a simple and effective way to interpolate a function, using dependence, without having to detemine a specific form for the function.

This idea extends in principle to functions of multiple parameters as well. Figure 4 (right) shows what a Gaussian process interpolator produces for a toy example with only one input parameter.

We need the model for the observations to allow for the fact that the ice sheet model is not a perfect representation of the observations of the ice sheet. For this, we add a component to the model for errors (variability in the observations) and sources of systematic biases, called a model-data "discrepancy" term. Once we put these pieces together, we have a model that is potentially useful for observations of the ice sheet that serves as the probability model for the observations given the parameters.

Figure 5 shows how calibration works for a toy example where the model output is just a scalar and the observation consists of only a scalar as well.

We can summarize the entire approach as follows.

(1) **Generate an ensemble of model runs:** Run the ice sheet model at various parameter settings. This provides pairs of parameters and model output, just as in a regression problem.

(2) **Emulate the ice sheet model:** Use a statistical model to approximate the relationship between the parameters and the model output. This is similar to fitting a flexible regression model, except the response is multivariate, spatial (satellite data), and temporal (paleo-reconstructed data).

(3) **Construct a model for the observations:** This is the fitted Gaussian process model + a model for errors and biases. We only specify the form of the errors and biases; their parameters still need to be inferred from the data (from Step 4, below).

(4) **Calibration:** Fit the model to the observations. This gives us a distribution on the parameters, while providing some information about the errors and biases.

(5) **Project the future of the ice sheet:** Use the posterior distribution on the parameters to run the model forward and provide the future of the ice sheet in the form of a ("posterior predictive") distribution.

Of course, here the model output is quite a bit more complicated than a standard regression response because the model output is a map of the current ice sheet (a spatial data set), along with information about the ice sheet's past over time (a time series data set). The relationship between the parameters and the model output also can be quite complicated.

There are additional complications because the data tend not to be Gaussian. For example, the ice sheet data are modeled as presence-absence. A spatial generalized linear model version of a Gaussian process lets us approximate

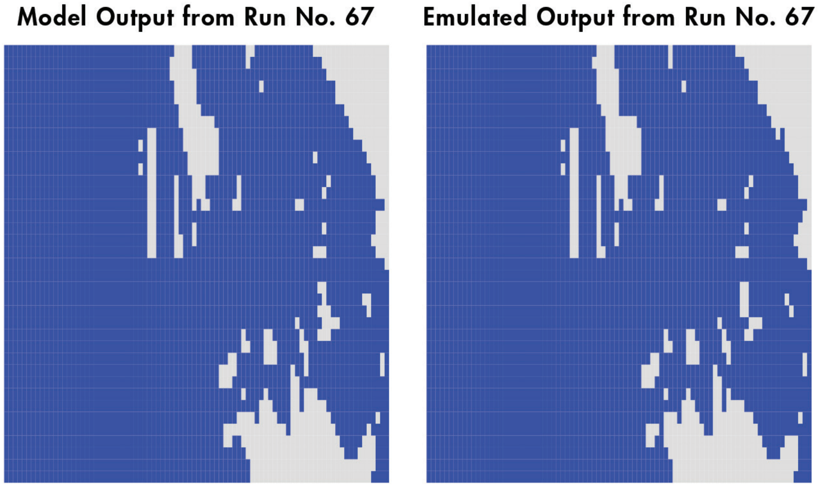


Figure 6. Comparison of actual model output (left) with emulated model output (right). Blue corresponds to "no ice sheet" and white corresponds to "ice sheet presence." The emulator is able to mimic the model run quite well.

the deterministic model with non-Gaussian output by a probabilistic model. The high-dimensionality of the data also necessitates some dimension-reduction approaches. We use a principle components analysis-based approach.

The Gaussian process methodology is remarkably flexible, allowing us to emulate the ice sheet model quite well. How well can be studied by using cross-validation, such as by leaving out 10 percent of the model runs (parameter settings) when fitting the Gaussian process model to the ice sheet model runs, then looking at what the Gaussian process model predicts for the parameter settings that were left out. If it resembles what the model actually outputs at those parameter settings, it suggests that emulation is working well. Figure 6 illustrates this.

Results

Using emulation and calibration methods to these data and models makes it possible to obtain parameter estimates and resulting probability distributions for future projections. This is summarized in Figure 7, which shows the distribution of potential sea level rise due to the melting of WAIS in 500 years. Calibration with both the modern and paleo data results in different sea level rise projections (red curve, "current approach") when compared to projections with calibration using only the modern data (dash-dot blue curve, "modern obs only").

In particular, using both sources of data eliminates any possibility of there being no sea level rise; that is, the value 0 is included in the distribution for the modern data, while it is essentially excluded when both data are used.

Our research shows that sea level rise is inevitable, although our results are relatively conservative in stating that it is most likely to be around 2 meters. Even 2 meters of sea level rise will leave many low-lying regions in the world completely submerged, and would put many more areas—such as the Netherlands and the Maldives—at high risk of potentially devastating storm surge damage. Future storm surges are likely to cause much greater devastation through flooding. Recently developed models that incorporate a few additional features of the ice sheet dynamics suggest that sea level rise may be even more dramatic.

Caveats

With all the complicated sources of information that have gone into this research, we have to be cautious about our conclusions. The ice sheet model does not include all the processes that affect the ice sheet. Uncertainties in the paleo data have not been accounted for. The ice sheet model will behave differently for different initial values; ideally, we would incorporate uncertainties due to this variation, too.

Similarly, external forcings—climate variables that are external to the ice sheet—may change over time in a number of different ways. These also affect how the ice sheet behaves.

Incorporating all these uncertainties is daunting largely because of the computational challenges involved. Hence, whatever we say about the behavior of the ice sheets in the future is necessarily imperfect. However, the information summarized here incorporates

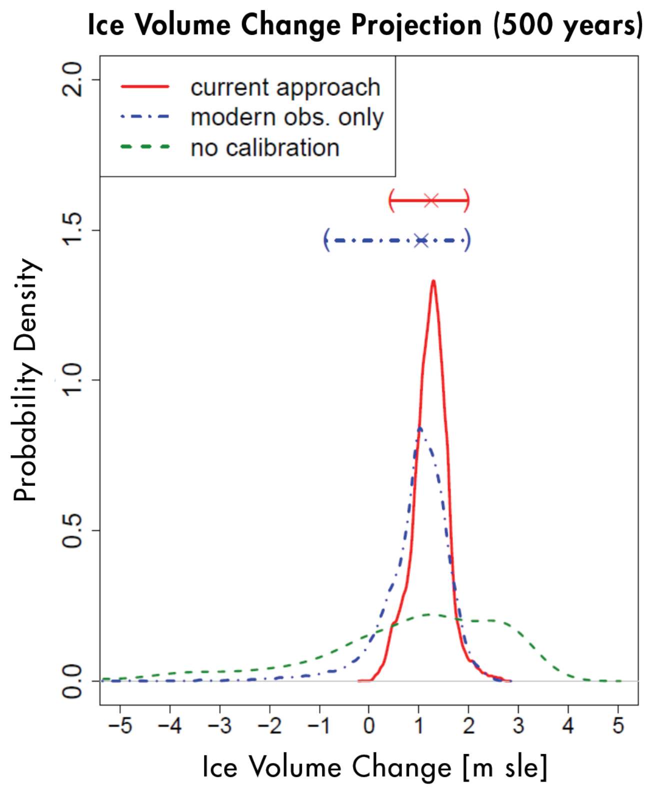


Figure 7. Posterior predictive distribution of projected ice sheet contribution to sea level rise. Adding the paleoclimate data results in a much-sharper projection (red curve) than when only modern satellite data are used (blue curve). In particular, the possibility of no (zero) sea level rise due to ice volume change is virtually eliminated in the red curve. Reproduced from Chang, et al.

cutting-edge physics and multiple observation data sets, and pieces the information together in a principled manner, so in spite of all these caveats, we have made progress. To quote Einstein, "…all our science, measured against reality, is primitive and childlike—and yet it is the most precious thing we have."

About the Authors

Murali Haran is a professor in the Department of Statistics at Penn State University.

Won Chang is an assistant professor in the Department of Mathematical Sciences at the University of Cincinnati.

Klaus Keller is a professor in the Department of Geosciences and director of SCRiM (Sustainable Climate Risk Management) at Penn State University, and an adjunct professor in the Department of Engineering and Public Policy at Carnegie Mellon University.

Rob Nicholas is a research associate at the Earth and Environmental Systems Institute and managing director of SCRiM at Penn State University.

David Pollard is a senior scientist at the Earth and Environmental Systems Institute at Penn State University.

Further Reading

http://www.antarcticglaciers.org/modern-glaciers/introduction-glacier-mass-balance/

Bakker, A., Louchard, D., and Keller, K. 2017. Deep uncertainties surrounding sea-level projections: Sources and Implications. *Climatic Change Letters 140*:339–347

Chang, W., Haran, M., Applegate, P., and Pollard, D. 2016. Improving Ice Sheet Model Calibration Using Paleoclimate and Modern Data. *Annals of Applied Statistics 10*:2,274–2,302.

Diaz, D., and Keller, K. 2016. A potential disintegration of the West Antarctic Ice Sheet: Implications for economic analyses of climate policy. *American Economic Review 106*:607–611.

National Snow and Ice Data Center. State of the Cryosphere. *https://nsidc.org/cryosphere/sotc/ice_sheets.html*.

Wong, T.E., Bakker, A., and Keller, K. 2017. Impacts of Antarctic Fast Dynamics on Sea-Level Projections and Coastal Flood Defense. *Climate Change 144*(2):347–364.

Ecological Impacts of Climate Change: the Importance of Temporal and Spatial Synchrony

Christopher K. Wikle

Consensus has been building in recent decades that human activities are contributing to substantial modification of the Earth's climate system, leading to growing interest in the detection and assessment of potential biological impacts associated with this changing climate. Multiple approaches have attempted to link ecological impacts with climate change, but all have been limited, more or less, by the difficulty of adequately representing ecosystem complexities in these analyses.

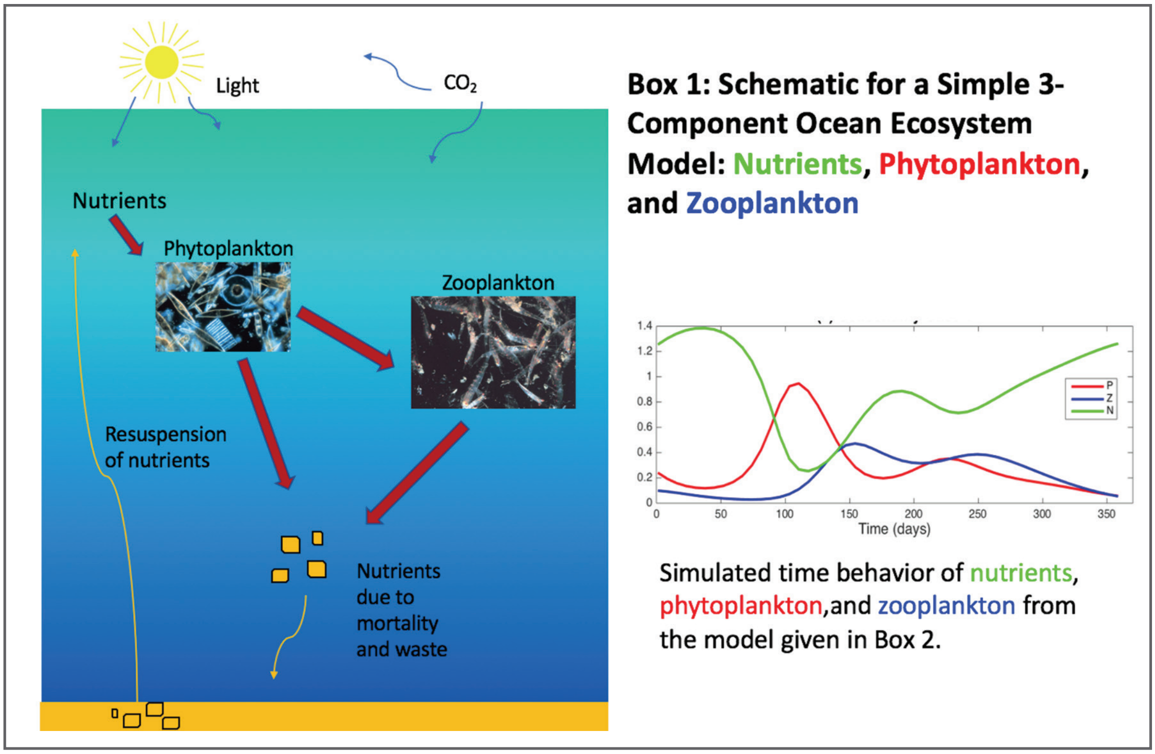
That is, to understand how an ecosystem will respond to climate changes requires accounting for the myriad interactions between biological and physical processes across spatial scales ranging from very local to global, and temporal scales ranging from within a day to decades.

This problem is particularly visible in ecological processes that exhibit synchrony, which can be defined to occur when critical phases of species' life cycles are closely linked to temporal cycles in environmental variables. For example, the onset of budding in a plant is tied to light availability and temperature, and the return of migratory waterfowl in the spring could be affected by weather conditions at various geographical locations.

Such synchronous behavior can be exhibited across space as well, and the synchrony of one species can affect those species that co-occur in a geographical area. For example, spawning groups on coral reefs could be affected by ocean conditions, and such conditions might alter the competitive balances of the various populations on the reef.

As the climate changes, there is growing evidence that many of the cyclical environmental variables involved in these types of synchronous behaviors will also change. For instance, temperature cycles are driven by the solar cycle, which is very predictable, but temperature cycles are also affected by shifts in global weather patterns, local land use changes, and contrasts between heating on land versus the sea. Similarly, precipitation often occurs in cycles, such as with the monsoon wet seasons, but the timing of the onsets of these wet periods is again a function of global weather patterns, which can change year-to-year.

It is not clear whether our current ability to project climate into the future has the sufficient resolution in time and space to



Box 1. Cartoon depiction of the simple 3-component ocean ecosystem with nutrients, phytoplankton, and zooplankton (left). The time series plot (right) shows a simulation of this system across a year. Nutrients are brought to the surface of the ocean through upwelling (see Figure 1). With the availability of sunlight, phytoplankton abundance increases due to photosynthesis, providing a food source for zooplankton, which deplete the phytoplankton population and eventually die off due to lack of a food source. The dead zooplankton and waste drop to the ocean bottom as detritus, which provides a source of nutrients for the cycle to continue.

Author-created figure with two Wikimedia images of phytoplankton and zooplankton: *https://commons.wikimedia.org/wiki/File:Diatoms_through_the_microscope.jpg* and *https://en.wikipedia.org/wiki/Swarm_behaviour#/media/File:Krill_swarm.jpg*.

adequately characterize the impact of these changing cyclical variations on ecosystems.

In the concept of temporal synchrony, time cycles in an environmental variable influence an ecological process. In particular, this is illustrated with an important ecosystem in the ocean corresponding to the life cycle of phytoplankton and zooplankton, both of which are very important components at the lower end of the ocean food chain. Phytoplankton are a type of algae that, like terrestrial plants, need sunlight and inorganic nutrients to grow by photosynthesis. Zooplankton are animals that rely on phytoplankton as a food source, and are themselves an important food source for animals higher in the food chain, such as fish and marine mammals.

Seemingly small changes in the winds and ocean currents near the shore can have a significant impact on this ecosystem. Changes in temporal synchrony for this system can change with space because the ocean currents vary significantly across space. The interactions between the ocean and this ecosystem also might change in the future, and it is difficult to obtain realistic projections of future climate conditions at fine-scale geographical resolutions.

Overall, our ability to manage ecosystems that show strong sensitivity to weather and climate variables depends on our ability to obtain realistic future projections of these variables. It is especially important that these projections be realistic in their ability to capture small geographic changes in seasonal cycles.

An Example of Temporal Synchrony

Phenology is the study of the seasonality or cyclic behavior in plant and animal life. Perhaps not surprisingly, the phenology of many plant and animal species is tied to the solar cycle, but it is also tied to temperature, precipitation, and habitat conditions, which have

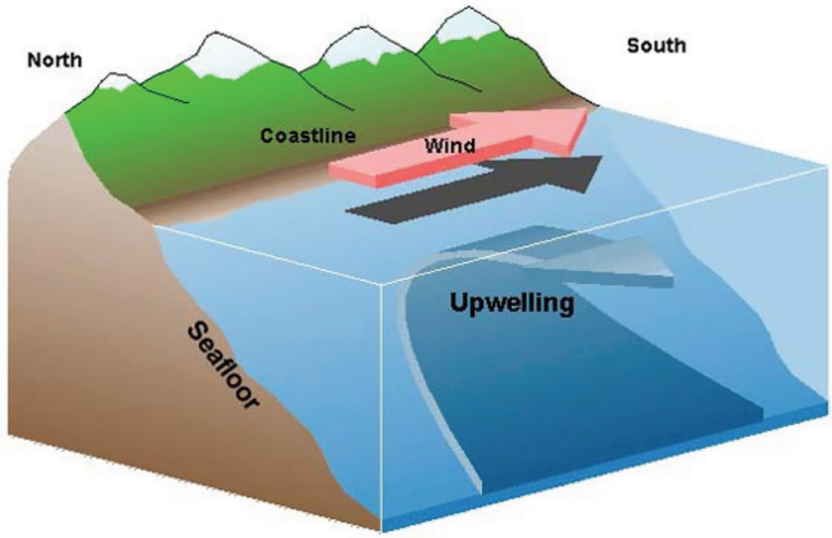


Figure 1. Cartoon depiction of coastal upwelling. Coriolis effects from southerly winds blowing along the coast push near-coast surface waters out to sea, and that water is replaced by colder water from below, which also brings nutrients up from the bottom.
Images source: *https://commons.wikimedia.org/wiki/File:Upwelling2.jpg.*

less-regular cycles. For example, phytoplankton blooms in the ocean require photosynthesis, and thus, are dependent on the solar cycle for light availability. Phytoplankton also require nutrients to thrive, so for there to be a substantial increase in phytoplankton abundance, there must be nutrients available at the same time that light is available. Nutrient availability is related to surface winds and deeper ocean currents.

Other examples of the connection between environmental cycles and ecological cycles include familiar processes such as when budding occurs in the spring and insect larvae transition to adulthood. Often, the interaction between species in an ecosystem, known as "symbiosis," also depends on these cycles. Since there is a predator-prey relationship between phytoplankton and zooplankton in the ocean, for instance, if the phytoplankton are limited by either the available light or nutrients, this will affect the zooplankton population because they feed upon the phytoplankton.

There is mounting evidence that climate change modification of seasonal cycles can have an impact on this cross-species symbiosis. A well-studied example of this is the relative explosion of bark beetles that have been decimating western North American forests in recent years. Historically, very cold winter temperatures kept the beetle population in check. In recent winters, the temperatures have not been consistently cold enough to limit the growth of the beetle population.

As another example, consider a general predator-prey ecosystem setting, where habitat conditions vary as a function of an environmental time series covariate that changes according to an annual cycle. Assume that prey abundance is a function of these habitat conditions, but the ability of the prey to use this habitat is dependent on a different annual cycle. Since the predator is dependent on the availability of the prey, seasonality is indirectly driving the population dynamics of the predator. If this seasonality is modified by climate change, it could have significant effects on the abundance of both the prey and the predators.

There are a number of systems for which this could serve as an analogy, but we focus our attention on the ocean ecosystem for illustration.

In considering the lowest levels of the ocean food chain near the shore, a simple model for this ecosystem consists of phytoplankton (P), zooplankton (Z), and nutrients (N). In this simple ecosystem representation, the "habitat" corresponds to the available nutrients and sunlight, phytoplankton are the prey and zooplankton correspond to the predators. More specifically, as outlined in Box 1, nutrients are taken up by phytoplankton,

Box 2: Simple Nutrient, Phytoplankton, Zooplankton (NPZ) Ecosystem Model

(after Franks, et al., 1986; Miller, 2004)

$$\frac{dP}{dt} = \frac{V_m s(t) N P}{K_s + N} - m_p P - Z R_m (1 - e^{-\lambda P}) - \alpha w(t) P$$

Phytoplankton rate of change = nutrient uptake – mortality – grazing from zooplankton – loss from mixing

$$\frac{dZ}{dt} = (1-\gamma) Z R_m (1 - e^{-\lambda P}) - m_z Z$$

Zooplankton rate of change = growing efficiency * grazing – mortality

$$\frac{dN}{dt} = -\frac{V_m s(t) N P}{K_s + N} + m_p P + m_z Z - (1 - m_s) \gamma Z R_m (1 - e^{-\lambda P}) + \alpha w(t)(N - N_0)$$

Nutrient rate of change = – nutrient uptake + (1-growth efficiency)*grazing + mortality + gain from mixing

Variables and Parameters	Values
$P =$ phytoplankton stock size	start at 0.3 $\mu g L^{-1}$
$V_m =$ max phytoplankton growth rate	2 day^{-1}
$s(t) =$ seasonally varying light availability	sinusoid peaking at V_m on June 1
$N =$ nutrient concentration	start at 1.2 $\mu g L^{-1}$
$K_s =$ half saturation constant	$1.0 mugL^{-1}$
$m_p =$ phytoplankton mortality rate	0.1 day^{-1}
$Z =$ zooplankton stock size	start at 0.1 $\mu g L^{-1}$
$R_m =$ max zooplankton grazing rate	1.7 day^{-1}
$\lambda =$ Ivlev constant	1.0 $\mu g L^{-1}$
$\alpha =$ baseline turbulent mixing fraction	0.2 day^{-1}
$w(t) =$ upwelling mixing fraction	two-harmonic series
$\gamma =$ zooplankton growth efficiency	0.3
$m_z =$ zooplankton mortality rate	0.3 day^{-1}
$m_s =$ fraction of fecal loss of N	0.5
$N_0 =$ initial N value	1.3 $\mu g L^{-1}$

Box 2: A set of differential equations (top) describing the simple 3-component ecosystem model shown in Box 1. The variables and parameters in the model are defined in the list (bottom).

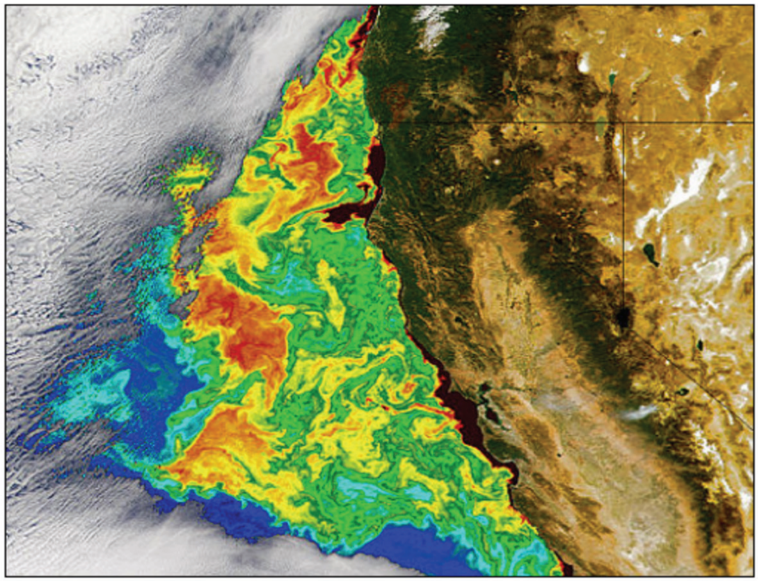


Figure 2. Image from NASA's SeaWiFS instrument that measures ocean color. Ocean color is a good proxy for phytoplankton abundance near the surface of the ocean. This image shows the ocean color on October 6, 2002, along the northern coast of California. The relative abundance of phytoplankton is given by the color, with blue the lowest and red the highest abundance.
Image Source: NASA SeaWiFS Ocean Color.

phytoplankton are eaten by zooplankton, zooplankton losses due to metabolic processes are then available as nutrients, and the cycle continues.

Although this simple model illustrates predator-prey and nutrient dependence in the ecosystem, it requires a mechanism to transport nutrients so they are available to be consumed by phytoplankton. For example, there is often a vertical transport of water toward the surface in coastal environments known as "upwelling" (see Figure 1).

Upwelling transports nutrients toward the surface, where they are available for phytoplankton photosynthesis. Upwelling is a physical process in the ocean driven by near surface winds, so changes in wind conditions on short-term, seasonal, inter-annual, and climatological time scales will affect it. This is accounted for in the very simple model of the ocean ecosystem outlined in Box 2 through a term that considers seasonal changes of the ocean "mixed layer."

The mixed layer of the ocean is a shallow layer in which turbulence caused by surface winds and waves mix the water column so the density of the water is about the same throughout the layer. The mixing leads to the loss of a portion of the available phytoplankton near the surface, but brings much-needed additional nutrients from below for phytoplankton growth, assuming light availability for photosynthesis (this is a highly idealized representation of the real-world ecosystem).

Because of seasonal changes in near-surface winds and larger-scale ocean currents, the mixed layer changes throughout a season. In addition, these physical processes vary quite substantially over small geographical areas.

For example, Figure 2 shows an image along the northern coast of California obtained from the NASA SeaWiFS instrument that measures ocean color. Ocean color is a good proxy for phytoplankton abundance in the near surface levels of the ocean. It is clear from this image that there is substantial variation in phytoplankton in this region.

Consider the effects on this ocean ecosystem from the two

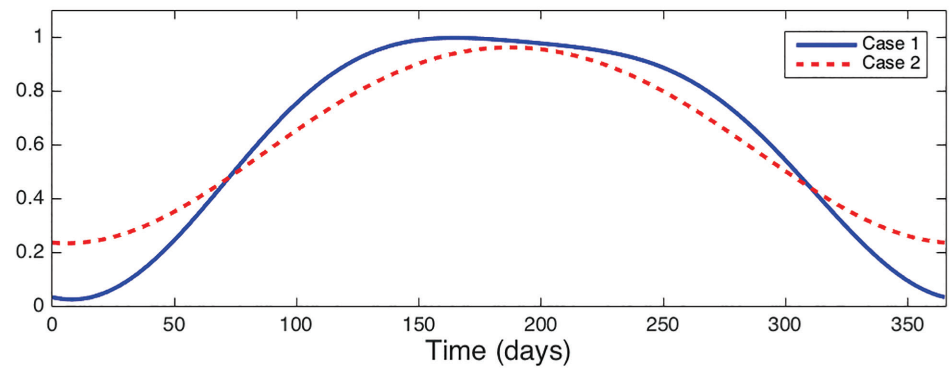


Figure 3. Time series showing seasonal ocean upwelling for the two simulations in Figure 4. These curves were derived from a regression analysis on the daily upwelling index at (60N, 146W), as available from NOAA's Pacific Fisheries Environmental Laboratory (*www.pfeg.noaa.gov/products/PFEL/modeled/indices/upwelling/NA/data_download.html*). The Case (1) (blue solid) curve corresponds to the period from January 1, 1967–December 31, 1991, and the Case (2) (red dashed) curve corresponds to the period from January 1, 1992–December 31, 2016. Although these time series seem fairly similar, their impact on the simple ecological system described in Box 2 can be substantial (see Figure 4).

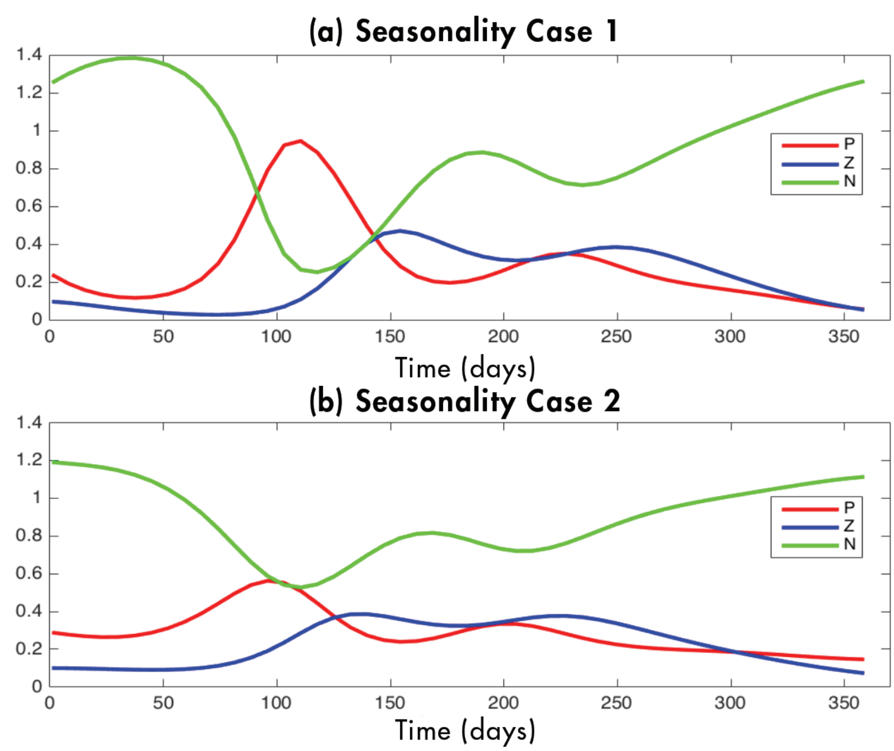


Figure 4. Panel (a) Simulation of NPZ model in Box 2 in which the seasonal upwelling is given by the blue solid line in Figure 3. Panel (b) shows the same simulation except with seasonal upwelling, given by the red dashed line in Figure 3. In both cases, phytoplankton is indicated by the red curve, zooplankton by the blue curve, and nutrients by the green curve. Both the intensity and the timing of the peaks in phytoplankton and zooplankton are affected by the relatively small differences in the seasonal upwelling.

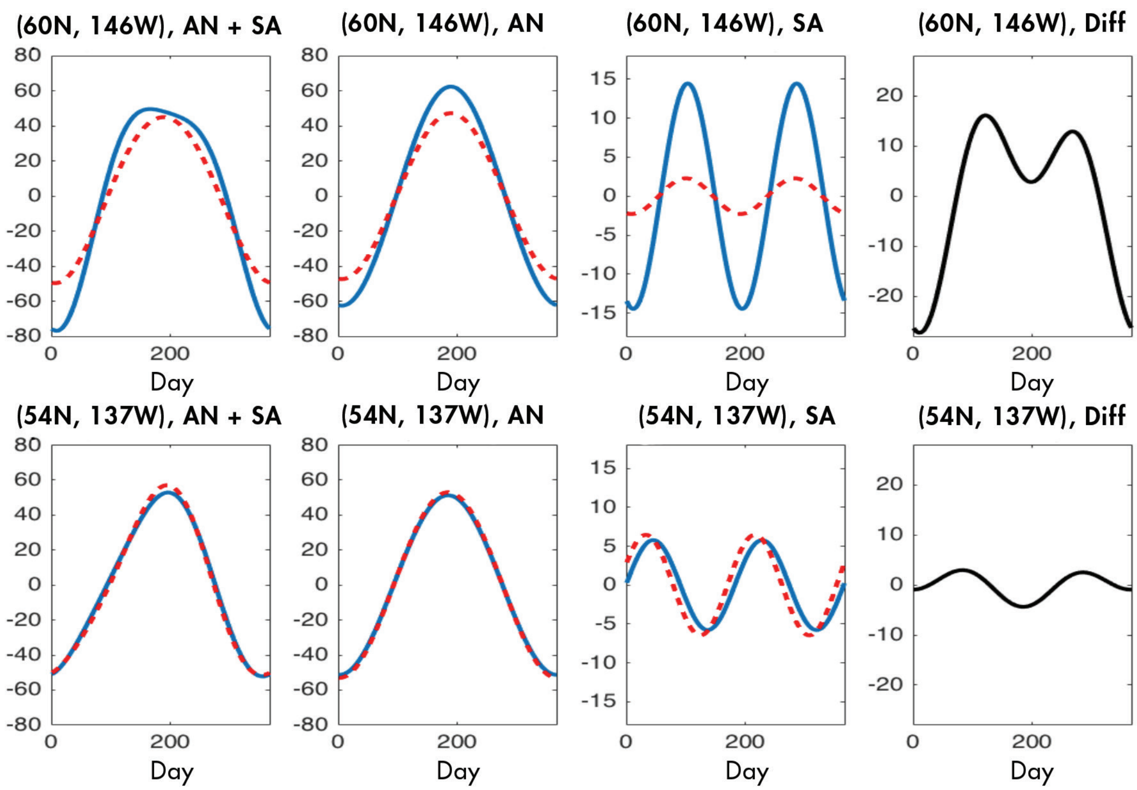


Figure 5. The first three top panels show (left to right) the combined annual and semiannual components, annual components, and semiannual components, for January 1, 1967–December 31, 1991 (blue), and the period from January 1, 1992–December 31, 2016 (dashed red) for the ocean upwelling index at (60N, 146W). The bottom panels show the same plots, but for a station at (54N, 137W). The solid black lines in the right panels show the differences between the total seasonality (annual + semiannual cycles) for the two time periods (first period–second period); that is, the difference between the solid blue and dashed red lines in the left-most panels. This shows that the main components of the seasonal cycle can change with time and vary substantially at nearby spatial locations.

different seasonal time series of ocean upwelling, shown in Figure 3. Figure 4a shows the nutrient, phytoplankton, and zooplankton time series over a year from the integration of the model in Box 2, where the seasonal upwelling is specified by the blue solid-line time series in Figure 3. Figure 4b shows the same simulation, but the seasonal upwelling is given by the red-dashed seasonal time series given in Figure 3.

There are clear differences between these two cases, with a later and more-intense phytoplankton peak in case 1. The difference between these two scenarios is also evident for the zooplankton population shown in Figure 4, with later and a more-intense initial peak in case 1. These marked differences are perhaps somewhat surprising, given that the seasonal variation between the upwelling components in these two cases does not seem that substantial. Yet, differences such as these in zooplankton response could have important effects at the higher levels of the ocean food chain, because zooplankton are a primary food source for many fish and marine mammals.

A reasonable question in response to the simple simulation in Figure 4 is whether such changes actually occur in real-world time series. Indeed, they do. The plots in Figure 3 correspond to a scaled version of the first two harmonics from the daily upwelling index for a location in the eastern Pacific Ocean at (60N, 146W) for the periods January 1, 1967–December 31, 1991, and January 1, 1992–December 31, 2016, respectively. This plot shows that there is a small, but fairly distinct, difference in the seasonal variation between the two time-periods.

This can be seen more clearly in the first three panels in the top row of Figure 5, which shows the first two seasonal harmonic components for these ocean upwelling index series. An annual seasonal harmonic corresponds to fitting a sinusoid with a period of one year to the time series, estimating the phase and amplitude of the sinusoid. The second (semi-annual) harmonic corresponds to fitting a sinusoid with a period of one-half year to the time series, again estimating its amplitude and phase.

It is clear that, in this case, there is no substantial change in the phase of the annual and semi-annual cycles for each time series, but there is a large difference in the respective amplitudes. This difference in amplitudes lead to the noticeable differences between the seasonal cycles of the time series, as shown in the last panel in the top row of Figure 5. In this case, the seasonal cycle is just the sum of the fitted annual and semi-annual harmonics.

In addition, note that this seasonal cycle in the upwelling time series changes as a function of geographic location. For example, the seasonal cycles for a "nearby" location (54N, 137W), given in the bottom row of Figure 5, show relatively little difference between the early and later time periods. This leads us to believe that the spatial variation in the change in seasonality could be important ecologically.

As we have shown, fairly small differences in the seasonal abundance in one species can have a somewhat dramatic impact on interacting species. This suggests that regional differences in temporal synchrony could lead to large differences in the spatial distribution of the relevant ecological variables.

This notion of geographical differences in temporal synchrony is related to the concept of "spatial synchrony" in the literature.

Spatial Synchrony

It has long been known that ecological populations that are geographically separated can change together, which is called "spatial synchrony." Such synchronies can occur due to dispersal among meta-populations; dependence on a common covariate, such as a weather variable; or interactions with other nearby populations. This suggests that if cycles in the environmental covariates mentioned in the previous section vary non-homogeneously across a geographic landscape, it could affect spatial synchrony.

There is evidence of such climate change induced variations in spatial synchrony in the literature. For example, bark beetle infestations in the western U.S., the distribution of migratory breeding birds, the gypsy moth invasion in the eastern U.S., and the ecosystem dynamics described above have all been shown to exhibit this effect.

Figure 5 shows the different time patterns in the first two harmonics between a location at (60N, 146W) and one at (54N, 137W). More generally, it is known that spatial differences of the first and second harmonic phase and amplitude in atmospheric temperature and precipitation fields exhibit distinct large-scale geographic differences, due to the difference in heating rates between land and sea. It is expected that these patterns are also present in the ocean and that they may be altered in a human-induced climate change environment.

Consider the satellite image off the coast of California in Figure 2. The abrupt transition in ocean color across small spatial regions suggests that trying to understand ecological impacts due to spatial synchrony in a changing environment will require climate projections that capture the seasonal cycle in reasonable temporal detail and are at a high spatial resolution. Indeed, one of the mechanisms of spatial synchrony is the movement of a population from one region to another.

For example, it is well known that the horizontal transport of phytoplankton, zooplankton, and nutrients in the ocean ecosystem is important to maintaining healthy populations. Thus, it is essential to capture the time and spatial scales of this process realistically to obtain future projections of the ocean ecosystem that are useful to managers and policy makers. Currently, this is done through projecting output from fairly low-resolution global climate simulation models to smaller scales, a process known as "downscaling."

Challenges of Downscaling

The current generation of global climate models (GCMs) used to project future climate under different greenhouse gas emission scenarios cannot be evaluated at resolutions that can accommodate the spatial and temporal scales adequately to project potential ecological impacts discussed above. That is, the ecological response to large-scale climate change must take into account local scale and site-specific features.

To obtain fine-scale information, the GCM output is typically downscaled either dynamically or statistically. Dynamic downscaling is performed by linking a higher-resolution numerical simulation model, typically applicable at some regional scale, with the global-scale GCM. In contrast, statistical downscaling calibrates a GCM run over a historical period with historical small spatial-scale observations, such as those that come from historical weather stations. Typically, this calibration is a type of regression model. This fitted model relationship is then assumed to hold for a GCM run under different conditions.

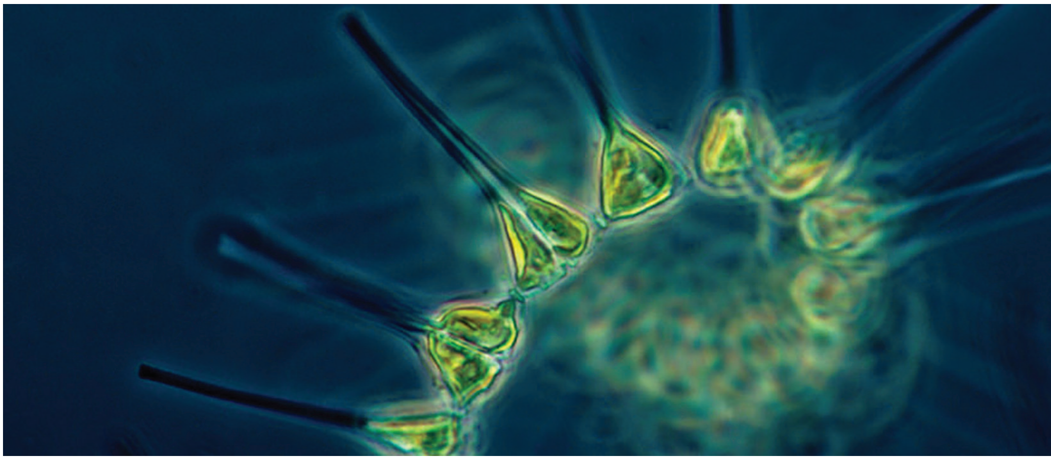
In both cases, the downscaling relationship is used to obtain higher-resolution projections from GCMs run in future climate change scenarios. The fundamental assumption in downscaling is that the inherent relationships and biases that drive GCM models and downscaling procedures are stationary—that is, the bias corrections and calibrations derived under historical conditions are appropriate for future scenarios. It is not always clear that this assumption is valid.

Only recently have comprehensive studies attempted to quantify the uncertainty associated with downscaling methods. It is widely accepted that these procedures are likely to be better at projecting temperature than precipitation and are better at capturing climatological averages than extremes.

Many studies have demonstrated spatial biases among various regional-scale climate model outputs. Recent studies in the ecological literature have been somewhat conflicted about the ability of these models to accommodate ecologically realistic spatial structure at fine geographic resolutions. In particular, the ability of these models to project realistic ecological behavior varies by region, season, type of method, and source of validation data.

As stated in Bucklin, et al. (2013), the "uncertainty associated with alternative downscaling methods may rival that of other, more widely appreciated sources of variation, such as the general circulation model or emission scenario with which future climate projections are created."

This is disconcerting, since it suggests that if these models cannot capture the realistic spatial distribution of temporal environmental variables such as temperature, precipitation, and upwelling adequately, then the projected ecological impacts on processes subject to temporal and spatial synchrony are likely to be too uncertain to be useful for managers interested in mitigating the effects of climate change.



Phytoplankton is the base of several aquatic food webs. In a balanced ecosystem, they provide food for a wide range of sea creatures including whales, shrimp, snails, and jellyfish.
Photo courtesy of NOAA.

Discussion

The yearly time course of many ecological processes is shown to be tied closely to cycles in environmental variables and that the cycles in one component of a complex ecosystem can directly affect another component, such as with a predator-prey system. If the cycles in the environmental variables are altered due to climate change—and there is evidence to suggest that these changes are indeed happening—then it can have serious effects on the ecosystem.

In some cases, cycles in environmental and ecological variables change quite substantially over fairly small geographical scales, so, to project potential impacts of climate change, climate simulation models must represent seasonality accurately over very small scales—which, unfortunately, is beyond the capacity of the current generation of global climate models. Attempts to mitigate this through the use of downscaling techniques probably cannot yet give realistic fine-scale spatial representations of seasonal variability.

Many authors caution that we must be aware that the myriad interactions across scales and species, and between physical, chemical, and biological systems, makes the study of ecological impacts due to climate change a challenging problem. Adequately quantifying uncertainty across all of these factors is still a moving target, and will continue to be the subject of research in the future. The more-specific problem of capturing spatial and temporal variation in temporal and spatial synchrony of ecological processes has been given relatively little attention, and is a fruitful area of research for statisticians.

Hierarchical statistical models have been shown to be effective for similar problems because of their ability to easily account for uncertainties in different sources of data, accommodate different data resolutions, and build complex multivariate spatio-temporal dependencies. These models are seeing increasing use at the interface of climate and statistics and at the interface of ecology and statistics. It is only natural that they will provide a framework to consider the complex interaction necessary to model climate changes to spatial and temporal synchrony for complex ecosystems.

Indeed, understanding synchrony of ecological processes under historical climate conditions can provide insight into potential impacts of future climate change. In this regard, new developments in downscaling and higher resolution GCMs that resolve cloud dynamics will surely improve the ability of the models to capture realistic local behavior. The output of such studies will provide essential information for mangers and policy makers who must attempt to mitigate the unwanted effects of climate change. C

Further Reading

Bucklin, D.N., Watling, J.I., Speroterra, C., Brandt, L.A., Mazzotti, F.J., and Romañach, S.S. 2013. Climate downscaling effects on predictive ecological models: a case study for threatened and endangered vertebrates in the southeastern United States. *Regional Environmental Change 13*(1):57-68.

Defriez, E.J., Sheppard, L.W., Reid, P.C., and Reuman, D.C. 2016. Climate change-related regime shifts have altered spatial synchrony of plankton dynamics in the North Sea. *Global Change Biology 22*(2):069–2,080.

Ekström, M., Grose, M.R., and Whetton, P.H. 2015. An appraisal of downscaling methods used in climate change research. *Wiley Interdisciplinary Reviews: Climate Change 6*(3):301–319.

Franks, P.J.S., Wroblewski, J.S., and Flierl, G.R. 1986. Behavior of a simple plankton model with food-level acclimation by herbivores. *Marine Biology 91*, 121–129.

Liebhold, A., Koenig, W.D., and Bjørnstad, O.N. 2004. Spatial synchrony in population dynamics. *Annual Review of Ecology Evolution and Systematics 35*:467–490.

Miller, C. 2004. *Biological Oceanography*. Malden, MA: Blackwell Science Ltd.

Molinos, J.G., and Donohue, I. 2014. Downscaling the nonstationary effect of climate forcing on local-scale dynamics: the importance of environmental filters. *Climatic Change 124*(1–2):333–346.

Parmesan, C., Burrows, M.T., Duarte, C.M., Poloczanska, E.S., Richardson, A.J., Schoeman, D.S., and Singer, M.C. 2013. Beyond climate change attribution in conservation and ecological research. *Ecology Letters 16*(s1):58–71.

Scholes, R.J. 2017. Taking the mumbo out of the jumbo: progress toward a robust basis for ecological scaling. *Ecosystems*, 1–10.

Cressie, N., and Wikle, C.K. 2011. *Statistics for Spatio-Temporal Data*. Hoboken, NJ: John Wiley and Sons.

About the Author

Christopher K. Wikle is Curators' Distinguished Professor of Statistics at the University of Missouri, with additional appointments in soil, environmental, and atmospheric sciences and the Truman School of Public Affairs. His primary research interests are in spatio-temporal statistics applied to environmental and ecological processes, with particular interest in dynamics. He is a fellow of the ASA and has received the ASA ENVR Section Distinguished Achievement Award.

Projecting Health Impacts of Climate Change: Embracing an Uncertain Future

Howard H. Chang, Stefanie Ebelt Sarnat, and Yang Liu

Global climate change affects human health most notably by increasing the frequency and intensity of dangerous heat waves, wildfires, and hurricanes. In addition to extreme weather events, climate change can lead to a myriad of persistent environmental changes that affect public health. One important area of climate science research is to understand health risks associated with the changing environment.

Changes in local weather patterns can create favorable conditions for higher concentration of ground-level ozone, an air pollutant linked to asthma exacerbation. After heavy rainfalls, drinking water quality can decrease due to contaminated runoffs, especially in rural settings. Ecological changes in vector habitats can encourage the transmission of infectious diseases such as dengue and malaria.

Health impact assessment refers to the analytic framework for evaluating how a policy or program affects population health. It is frequently applied in climate and public health research to quantify *future* health and economic burdens attributable to various consequences of climate change. For example, the World Health Organization estimated that climate change is expected to cause approximately 250,000 additional deaths globally annually from 2030 to 2050. This estimate includes deaths due to heat, flooding, diarrheal diseases, malnutrition, and infectious diseases that can vary across age groups and world regions.

Quantitative measures of future health impacts can play an important role in communicating the significance of climate change. Health impact assessment can also be conducted at the national or local level to provide crucial information for decision-makers who are developing long-term strategies for mitigating environmental risks and improving disaster preparedness.

Performing health impact assessment entails the integration of various data. For projecting future climate-related health impacts, analyses require three sources of information: (1) health effects of environmental exposures, (2) projections of future exposures, and (3) distributions of exposures and effects in the future population. However, each information source is subject to uncertainty because of data availability and assumptions made for the future.

The use of the word *projection*, instead of prediction or forecast, explicitly reflects the high degree of uncertainty in these estimates. Because uncertainty in projection is unavoidable, one important aspect for health impact analysis is the need to consider, quantify, and incorporate various sources of uncertainty in the final estimate.

This case study projects the number of emergency department (ED) visits due to future heat waves in Atlanta, Georgia. The analysis consists of three steps: estimating the association between heat waves and ED visits for all internal causes using historical data from 1992 to 2012; estimating future heat wave occurrences for 2055–2059 using climate model outputs; and estimating future ED visits attributable to heat waves. The case study describes and illustrates how statistical modeling contributes to various components of performing health impact projections.

Estimating Health Effects of Heat Waves

Estimates of large-scale health effects of environmental exposures are typically obtained from population-based epidemiologic studies. These studies link databases for adverse health outcomes (such as records of death certificates or medical records) to measurements of environmental risks (temperature, water quality, or air pollution). Environmental health studies are predominantly observational because exposures are seldom randomly assigned to the population.

Moreover, risks between common adverse health outcomes and exposures are often small, despite having considerable public health consequences thanks to their ubiquitous exposures. Various study designs and statistical methods have been developed specifically for population-based environmental epidemiology.

To estimate the association between heat waves and ED visits, we first obtained the following health and exposure data sets: daily counts of ED visits for all internal causes from local hospitals and the Georgia Hospital Association, and daily meteorology data from the National Climatic Data Center collected at Hartsfield International Airport.

A heat wave is typically characterized by a period of sustained hotter temperature compared to historical records. To reflect this prolonged exposure to extreme heat, we defined heat waves as periods of ≥2 consecutive days with daily maximum temperature beyond the 98th percentiles, calculated using records from 1945 to 2012. Heat wave days were indicated as a 1 and non-heat wave days were given a 0; the first day of a heat wave period was classified as a non-heat wave day so the research would capture only the health effect of *sustained* heat exposure. Overall, the data set contained about 19 million ED visits and 91 heat wave days.

Figure 1 shows time series plots of daily maximum temperature and ED visits in the 20-county Atlanta metropolitan area for 2001 to 2010. Note that daily ED visit shows an overall increasing trend due to population growth and a seasonal trend, with higher ED visits during the winter. This is probably due to increased respiratory infections during colder months. Particularly, the two spikes observed in the winter of 2003–2004 and the winter of 2009 may correspond to, respectively, the emergence of the Fujian H3N2 subtype of the influenza A virus and the 2009 H1N1 influenza pandemic. However, heat waves occur only during the summer.

Because of this seasonal difference in ED visits, a simple regression model that treats the daily ED visit count as the outcome variable and our exposure of interest (heat wave vs. non-heat wave days) as the predictor results in an overall protective effect of heat waves (about 28% fewer ED visits on heat wave days).

To estimate daily changes in ED visits that are attributable *only* to heat waves (i.e., not just due to the summer-versus-winter difference), the model has to include variables, known as confounders, that are causally related to ED visits and also independently associated with heat waves. Figure 2 illustrates relationships between outcome (Y), exposure (X), and confounder (C).

For a variable to be a confounder, it has to be a risk factor for the outcome, be associated with the exposure, and not be in the causal pathway between the exposure and the outcome. Therefore, a typical health effect model for heat waves will include variables such as meteorology (continuous temperature and humidity) and time trends (long-term and seasonal).

After adjusting for confounder variables flexibly in the model to account for their non-linear associations with ED visits, there is a relative risk of 1.020 with a 95% confidence interval of (1.013, 1.028). Relative risk is a commonly used measure in epidemiology to describe how a binary disease outcome varies by a risk factor. It is defined as the ratio of disease rates between the exposed and the unexposed group, where a relative risk greater (less) than 1 indicates the risk factor positively (negatively) affects disease occurrence.

In this analysis, a relative risk of 1.020 implies that there is an estimated 2% increase in ED visits during heat wave days compared to non-heat wave days with the same temperature, humidity, and season.

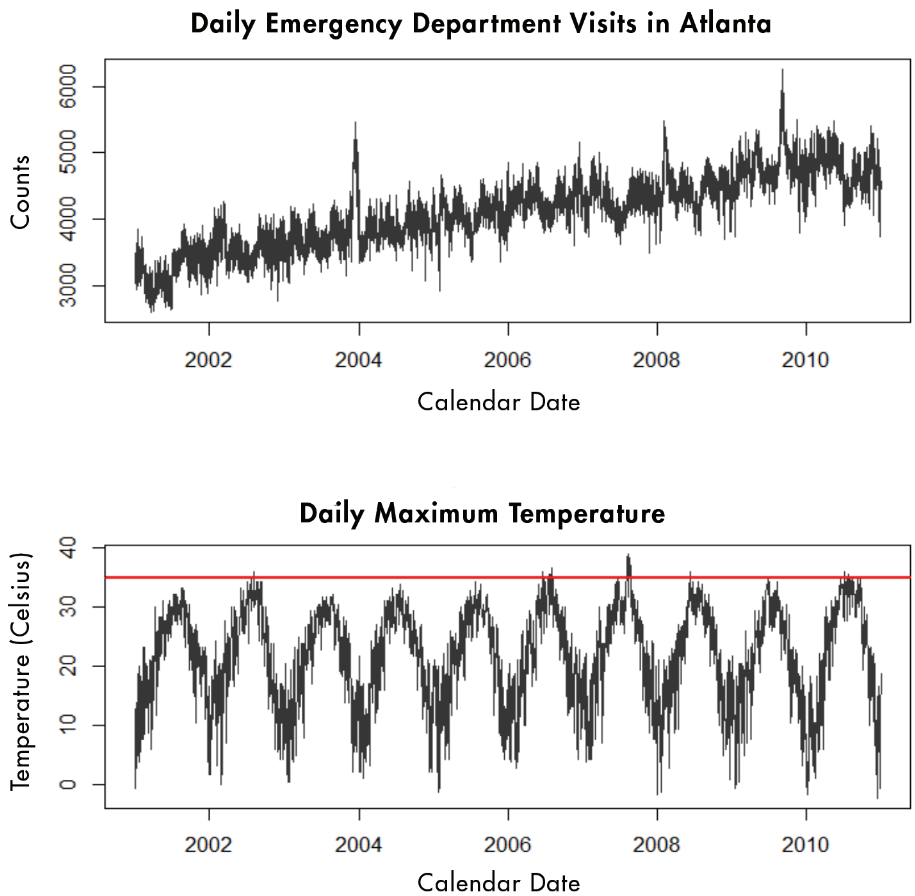


Figure 1. Time series plots of daily emergency department visits for all internal causes in 20-county Atlanta metropolitan area and daily maximum temperature, 2001–2010. The 98th percentile of daily maximum temperature threshold for the heat wave definition is indicated by the red line.

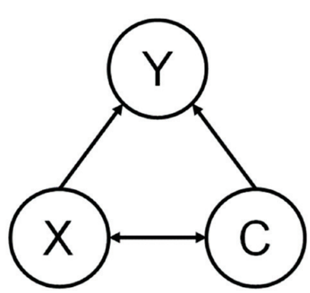


Figure 2. Illustration of relations between outcome (Y), exposure (X), and confounder (C).

The modeling decisions made in this case study include first defining extreme temperature as the 98th percentile of historical records. Using the 99th percentile is another possibility, but will result in a reduction of statistical power due to a reduced number of heat waves identified. Heat waves use daily maximum temperature; other temperature metrics—such as minimum temperature, which reflects night-time temperature, and apparent temperature, which reflects human discomfort—may create different risks. Heat waves may also have a delayed effect, driven by physiological responses to heat or healthcare-seeking behaviors.

This case study found a weaker increase in ED visits (relative risk of

1.012) when exposure was defined as the previous day being a heat wave day, compared to the *previous* day not being a heat wave day.

Finally, the health outcome represents an aggregated measure of ED visit morbidity (e.g., across all ages and outcomes). Audiences for health impact assessments are often interested in identifying subpopulations that are particularly vulnerable to the negative consequences of climate change. For example, children with asthma may be more sensitive to increasing air pollution concentration, and individuals who spend more time outdoor may be more at risk for heat stroke.

Projecting Future Heat Waves

Projections of future meteorology are accomplished mainly through computationally expensive regional or global climate models. Climate models are three-dimensional mathematical representations of the Earth's climate and reflect state-of-the-art scientific knowledge of the physical and chemical processes. Climate projections have to account for both the current climate and factors that affect the climate system. The International Panel on Climate Change (IPCC) has developed various scenarios to represent a range of future greenhouse gas emissions. The most-recent scenarios are known as Representative Concentration Pathways (RCPs).

These scenarios are often tied to additional assumptions about future technology progress, regulatory policies, and coordination across countries. Because of the complexity in climate modeling, a global community of scientists has developed different modeling strategies with regard to the Earth's atmosphere, ocean, land surface, and ice, and their interactions. The climate scenarios and the choice of climate model constitute an important source of variability in health projection.

Climate model simulations typically include both historical (hindcast) and future periods. Because climate models are deterministic computer models, their outputs can exhibit complex biases over space and time when compared to historical measurements. Biases can arise from insufficient characterization of the climate system with mathematical equations and discretization of the continuous environmental field in space and time.

For health impact projects, two common approaches are used to address bias in climate model outputs. The first approach evaluates health risks using both the hindcast and the future periods. Health impacts attributed to climate change are then calculated by taking the difference between these two periods, assuming the bias cancels out. However, this approach forces the hindcast period to serve as the reference period.

More recently, methods have been developed to *bias-correct* climate model outputs. Bias-correction is accomplished by first modeling the discrepancy between observations and hindcast simulations, and then assuming this bias can be extrapolated to future periods. While bias correction offers a more-flexible framework for evaluating future climates, it makes the assumption that the bias observed during the historical period will remain the same in the future period. This is a particular concern if bias correction is done one variable at a time, ignoring complex interactions between variables in a dynamic climate system.

Climate models perform simulations over a three-dimensional grid over the Earth's surface. This heat wave and ED visit case study used high-resolution regional climate model outputs from the Weather Research and Forecasting model (WRF) 3.2.1. It involved running the WRF model for a historical period 2001–2004 and a future period of 2055–2059 at a spatial grid resolution of 12 km over the continental U.S. We examined a low greenhouse gas emission scenario (RCP 4.5) and a high-emission scenario (RCP 8.5). The regional WRF outputs were based on inputs from the Community Earth System Model version 1.0 (CESM 1.0) climate model, which was run globally with a cruder spatial resolution.

We first extracted the single grid cell from WRF that includes the weather station in Atlanta used to conduct the previous health effect analysis. Figure 3 shows the quantile functions of daily maximum temperature (March–October) during the historical period 2001–2004 for observations and climate model simulations. Even though the day-to-day correlation is only moderate (Pearson's correlation of 0.63), the climate model is able to capture the overall *distribution* of daily maximum temperature quite well, with a small positive bias at the extremes.

One bias-correction method—quantile-mapping—tries to resolve the difference in quantile functions. Working with quantile function has the advantage of not having to assume a distribution for maximum temperature (e.g., normal or log-normal). We first estimated the bias across quantile levels between daily maximum temperature observations and model simulations during 2001–2004, and then calculated the number of heat wave days for two RCP scenarios using bias-corrected future projections for 2055–2059 (Figure 4).

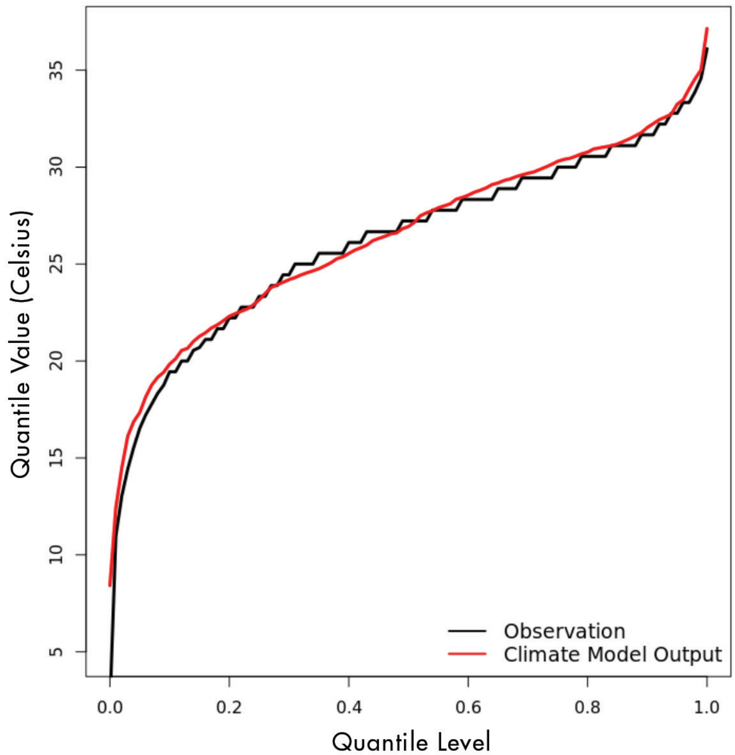


Figure 3. Quantile functions of daily maximum temperature (March–October) in Atlanta, Georgia, from 2001 and 2004.

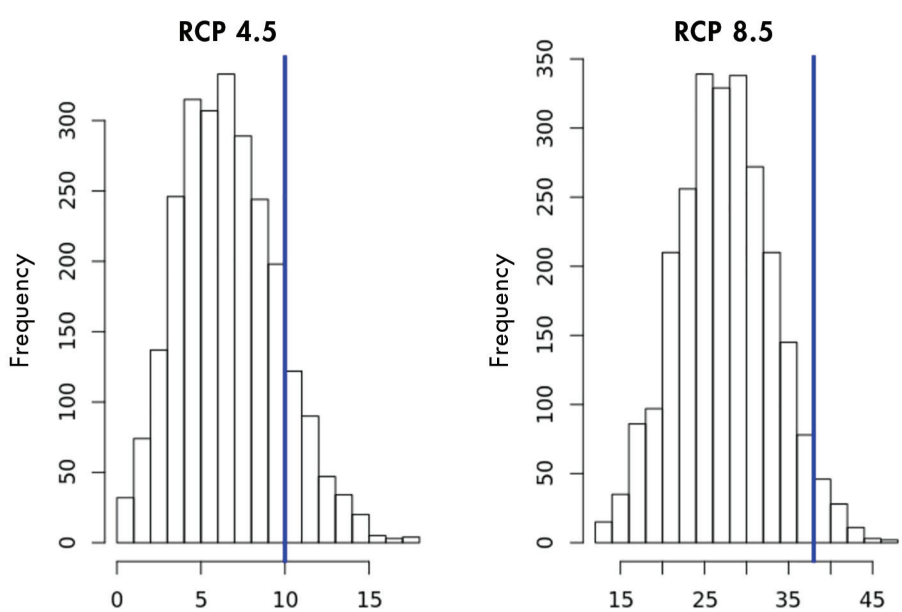


Figure 4. Projected number of heat wave days in Atlanta, Georgia, for the period 2055 to 2059 under two Representative Concentration Pathways (RCP). The histogram represents uncertainty in the projection and the blue line indicates projection by climate model outputs without bias correction.

Table 1—Projection Intervals of Annual Emergency Department Visits for All Internal Causes Attributable to Heat Waves in Atlanta from 2055–2059

SSP Narrative	% Population Change	Emission Pathway RCP 4.5	Emission Pathway RCP 8.5
SSP 1: Sustainability	24.8	34 – 226	281 – 668
SSP 2: Middle of the Road	23.0	33 – 222	277 – 653
SSP 3: Regional Rivalry	9.1	29 – 201	247 – 584
SSP 4: Inequality – Road Divided	18.5	32 – 218	268 – 636
SSP 5: Fossil-fuel Development	37.4	37 – 257	310 – 740

Projections were conducted under five different Shared Socioeconomic Pathways (SSPs) and two Representative Concentration Pathways (RCPs).

Here, the histogram reflects statistical uncertainty about the projection and the blue line indicates projection by climate model outputs without bias correction. Deterministic projections are higher than the mean of bias-corrected projections, probably because model outputs show positive bias at higher temperatures during the hindcast period.

Projecting Future Health Impacts

Several important characteristics of the future population will determine the projected number of future ED visits due to heat waves. Because the estimated risk association between ED visits and heat waves is a relative rate, the health burden depends on the baseline number of ED visits. This baseline number is determined by both the at-risk population size and the baseline risk.

Given its historical trend, the Atlanta population is likely to be larger in the 2050s compared to the current day. Population change is driven by migration, life expectancy, and fertility; these variables are interrelated to economic development, urbanization, and anthropogenic emission.

The future baseline ED visit rate also may change in response to the overall health of the population, as well as changes in healthcare access. Similar to IPCC's emission scenarios, a set of five scenarios for future populations known as the Shared Socioeconomic Pathways (SSP) has been developed. Again, a community of demography and economy modelers has been involved to provide projections of future country-specific population size, age and sex composition, and gross domestic product.

Another consideration for projecting health impacts is whether the estimated heat wave and ED visit association will change in the future. For instance, adaptation describes how we can reduce our susceptibility to environmental changes and, thus, reduce future health risks.

Adaptation can be accomplished via various routes. In response to heat waves, local authorities may develop more-effective warning systems or provide more cooling centers. Increasing prevalence of air conditioning may also reduce the risk of exposure.

Finally, there is recent epidemiologic evidence that over the last few decades, associations between high temperature and mortality have decreased.

We obtained population projections for the U.S. for five SSP scenarios from the National Center for Atmospheric Research using the Community Demographic Model and assumed Atlanta's population will increase proportional as the U.S. We assumed the baseline ED visit rates to be the same as the most recent five-year period (2,630 visits per day), and the relative risk for heat wave to be unchanged (i.e. no adaptation).

To incorporate uncertainties from both health effect estimates and projected future heat waves, we conducted a Monte Carlo simulation experiment by repeatedly simulating realizations using the point estimates of the relative risk and its standard error and posterior samples of the number of heat waves during 2055–2059. Table 1 gives the projected number of future ED visits per year attributable to heat waves under different combinations of population and emission scenarios.

The projection interval represents the 2.5% and 97.5% quantiles of the simulation. Each interval reflects uncertainty in the health effects and in the number of projected future heat wave days, while uncertainty in emission and population changes is described by the combination of different SSP and RCP scenarios. It should be noted that sometimes not all combinations of RCP and SSP are realistic scenarios for a given population.

In this case study, differences between the two future RCP emission scenarios appear to have a larger impact on projection uncertainty than future population changes under different SSPs. Specifically, projection intervals are similar across different SSP narratives under either RCP 4.5 or RCP 8.5. In contrast, the lower interval bound under RCP 8.5 within each SSP narrative is consistently higher than the upper interval bound under RCP 4.5.

Conclusions

Climate change research is highly interdisciplinary, bringing together tremendous amounts of data, theory, and modeling efforts to provide timely knowledge for one of the most-pressing issues of our time. Uncertainty in projecting the future climate and its consequences is well-recognized and will remain an integral part of the scientific endeavor.

Uncertainty arises from future storylines (e.g., emission and socioeconomic scenarios, adaption) and data availability and quality (e.g., heterogeneous health risks, standard error associated with health risk estimates, bias in climate model outputs). Uncertainty analysis should be viewed as an integral component of projection analysis, where statistical modeling techniques and probabilistic reasoning can help ensure these findings are informative, accurate, and reproducible. **C**

Further Reading

Chen, T., et al. 2017. Time-series analysis of heat waves and emergency department visits in Atlanta, 1993 to 2012. *Environmental Health Perspectives*. In press.

Riahi, et al. 2016. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Climate Change*. *http://dx.doi.org/10.1016/j.gloenvcha.2016.05.009*.

U.S. Global Change Research Program. 2016. The impacts of climate change on human health in the United States: A Scientific Assessment. *http://dx.doi.org/10.7930/J0R49NQX*.

Van Vuuren, et al. 2011. The representative concentration pathways: an overview. *Climate Change 109*:5.

World Health Organization. 2014. Quantitative risk assessment of the effects of climate change on selected causes of death, 2030s and 2050s. Geneva: World Health Organization.

About the Authors

Howard H. Chang is an associate professor in the Department of Biostatistics and Bioinformatics at Emory University. His research focuses on the development and application of statistical methods for analyzing complex spatial-temporal environmental and health data. He collaborates extensively in air quality modeling, environmental epidemiology, infectious disease, and climate science.

Stefanie Ebelt Sarnat is an associate professor in the Department of Environmental Health at Emory University. Her research focuses on assessing health impacts of air pollution, meteorological conditions, and weather extremes using population- and panel-based approaches. Her work is grounded in exposure science and environmental epidemiology, with strong collaborations in biostatistics, atmospheric and climate sciences, and environmental engineering.

Yang Liu is an associate professor in the Department of Environmental Health at Emory University. His research interests include satellite aerosol retrieval and product design, application of satellite remote sensing in air pollution exposure modeling and health effects research, potential impacts of global climate change on public health, GIS, and spatial statistics. Over the past 10 years, he has led or participated in many projects to study the characteristics of air pollution and the impact of climate change on air quality and human health using remote sensing and model simulations.

Using a "Study of Studies" to Help Statistics Students Assess Research Findings

Azka Javaid, Xiaofei Wang, and Nicholas J. Horton

The American Statistical Association (ASA) Undergraduate Guidelines Workgroup in the Curriculum Guidelines for Undergraduate Programs in Statistical Science emphasizes the importance of study design as an essential skill for undergraduate programs in statistics (ASA, 2014). The guidelines also stress communication skills, along with teamwork and collaboration, as vital elements for statistical practice. In addition, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report notes the importance of concepts like bias and causal inference in study design (ASA, 2016). To achieve these goals, the GAISE College Report includes several recommendations, one of which reiterates the importance of fostering active learning through discussions.

This article describes an activity that is appropriate for students in introductory and intermediate statistics courses to practice interpreting research results and scrutinizing the design and analysis of studies. The activity includes a component of group work to improve students' communication and collaboration skills.

Traditional textbooks used in the introductory statistics curriculum stress the importance of survey design. As an example, the fourth edition of *Intro Stats (IS)* reinforces concepts relevant to experimental design, including bias, randomization and sample size in Chapter 12 ("Sample Surveys") (De Veaux, Velleman, and Bock, 2013).

Additional sample design concepts, such as observational studies, control groups, statistical significance, and confounding variables, are presented in Chapter 13 ("Experiments and Observational Studies") of *Intro Stats*. Other textbooks (e.g., *OpenIntro Statistics*, 2015) follow a similar approach.

How can textbook readings be reinforced in a class? This activity can help students explore aspects of design, assess research findings in published papers, and critique representations and interpretations of original research.

The Activity

This activity is based on a "Study of Studies" column in the *Atlantic* magazine. Each "Study of Studies" analyzes a different topic using published research articles. Table 1 lists the name, author, and date for all the past published "Studies." For this activity, we used one of the "Study of Studies" articles —"Diner Beware: How restaurants trick you into eating less and spending more" (*http://theatln.tc/2yBKLGk)* (Lam, 2015). Author Bourree Lam analyzes how restaurants manipulate seating arrangement, server posture, plate color and size, and music to attract more customers and revenue.

Implementation

The students received copies of the one-page "Diner Beware" column, which was read aloud by the class. Next, they were split into groups of two to four students and each received a copy of one of the 12 research articles cited in "Diner Beware." The research articles ranged in length from 4 to 28 pages, with an average of 10 pages.

The students were asked to skim the research article and, as a group, summarize the original research study design (i.e., describe the study's sample design, determine whether the study was randomized or

Table 1—Past "Study of Studies" Published in the *Atlantic*

Title	Author	Publication Date
You Can Be Too Beautiful	James Hamblin	March 2013
The Queen Bee's Guide to Parenting	Lindsey Abrams	April 2013
Various Ways You Might Accidentally Get Drunk	James Hamblin	May 2013
The Unexpected Ways a Fetus is Shaped by a Mother's Environment	Lindsey Abrams	June 2013
The Worst Time to Have Surgery	James Hamblin	July/August 2013
Is There Really Such a Thing as a 'Workaholic'?	Jordan Weissmann	September 2013
Violence Is Contagious	Rebecca J. Rosen	October 2013
Why You Look Like Your Dog	Sarah Yager	November 2013
How Women Change Men	Sarah Yager	December 2013
Who Cheats—and Why	Julie Beck	January/February 2014
Why You Can't Keep a Secret	Sarah Yager	March 2014
The Optimal Office	Julie Beck	April 2014
Our Gullible Brains	Sarah Yager	May 2014
Funny or Die	Julie Beck	June 2014
What is Art?	Matthew Hutson	July/August 2014
How to Look Smart	Julie Beck	September 2014
Status Anxiety	Matthew Hutson	October 2014
Keeping the Faith	Emma Green	November 2014
Faking It	Julie Beck	December 2014
You Are Just Like Me!	Matthew Hutson	January/February 2015
The Secret of Superstition	Matthew Hutson	March 2015
Diner Beware	Bourree Lam	April 2015
When Emotional Intelligence Goes Wrong	Andrew Giambrone	May 2015
The Hypocrisy of Professional Ethicists	Emma Green	June 2015
Palm Reading Is Real?	Eleanor Smith	July/August 2015
A Scientific Look at Bad Science	Bourree Lam	September 2015
Why We Compete	Matthew Hutson	October 2015
The Strange Origins of Urban Legends	Matthew Hutson	November 2015
Why You Bought That Ugly Sweater	Eleanor Smith	December 2015
A Strategic Guide to Swearing	Stephanie Hayes	January/February 2016
People Are Pretty Bad At Reading Faces	Naomi Sharp	March 2016
CEOs Behaving Badly	Alyza Sebenius	April 2016
How to Boast on the Sly	Matthew Hutson	May 2016
Life Isn't Fair	Matthew Hutson	June 2016
The Science of Beer Goggles	Stephanie Hayes	July/August 2016
The Charisma Effect	Matthew Hutson	September 2016
Do People Need Small Talk to Be Happy?	Stephanie Hayes	October 2016
How Voters Respond to Electoral Defeat	Ben Rowen	November 2016
Why Kids Need Recess	Alia Wong	December 2016
Awesomeness Is Everything	Matthew Hutson	January/February 2017
Unsafe at Any Speed	Jake Pelini	March 2017
How to Buy Happiness	Isabella Kwai	April 2017
Puppy Love	Katherine Riley	May 2017
Boredom Is Good for You	Jude Stewart	June 2017
How Vanity Could Save the Planet	Matthew Hutson	July/August 2017
How to Cut in Line	Jude Stewart	September 2017
When the Mind Wanders	Jake Pelini	October 2017

Table 2—12 papers in Bourree Lam's "Study of Studies" on Restaurants and Dining, the *Atlantic* (April 2015)

Title and Author	Publication
Odors and Consumer Behavior in a Restaurant (Guéguen and Petr, 2006)	*International Journal of Hospitality Management*
Plate Size and Color Suggestibility (Ittersum and Wansink, 2012)	*Journal of Consumer Research*
Assessing the Influence of the Color of the Plate on the Perception of a Complex Food in a Restaurant Setting (Fiszman, Giboreau, and Spence, 2013)	*Flavour*
Dining in the Dark (Scheibehenne, Todd, and Wansink, 2010)	*Appetite*
The Effect of Musical Style on Restaurant Customers' Spending (North, Shilcock, and Hargreaves, 2003)	*Environment and Behavior*
The Influence of Background Music on the Behavior of Restaurant Patrons (Milliman, 1986)	*Journal of Consumer Research*
The Impact of Restaurant Table Characteristics on Meal Duration and Spending (Kimes and Robson, 2004)	*Cornell Hotel and Restaurant Administration Quarterly*
How a Crowded Restaurant Affects Consumers' Attribution Behavior (Tse, Sin, and Yim, 2002)	*International Journal of Hospitality Management*
Lower Buffet Prices Lead to Less Taste Satisfaction (Just, Sigirci and Wansink, 2014)	*Journal of Sensory Studies*
Determinants and Consequences of Female Attractiveness and Sexiness (Lynn, 2009)	*Archives of Sexual Behavior*
Effect of Server Posture on Restaurant Tipping (Lynn and Mynier, 1993)	*Journal of Applied Social Psychology*
Effect on Restaurant Tipping of Male and Female Servers Drawing a Happy, Smiling Face on the Backs of Customers' Checks (Rind and Bordia, 1996)	*Journal of Applied Social Psychology*

observational), and assess the validity of the claims presented in their "Diner Beware" article.

Students made a brief set of slides summarizing their original article using RMarkdown (Baumer, Çetinkaya Rundel, Bray, Loi, and Horton, 2014). The RMarkdown slides were then shared with the class via RPubs, a platform for web publishing from RStudio (the slides could be submitted in other ways, such as by e-mailing the instructor). Lastly, students were given 5 to 10 minutes to present their findings. The student presentations were intended to improve communication skills and give students experience with technological innovations like RPubs.

An example of this process can be presented with the research article titled "Odors and consumer behavior in a restaurant" (Guéguen and Petr, 2006). Guéguen and Petr's work analyzed the effect of lemon and lavender scents on the duration of time and the amount of money spent by customers in a restaurant. They carried out their study from 8 p.m.–11 p.m. on three Saturdays in May with 88 patrons, and hypothesized that lavender is considered a relaxing odor while lemon is a stimulating odor.

Another example is presented by the research article titled "The Impact of Restaurant Table Characteristics on Meal Duration and

Spending" (Kimes and Robson, 2004). The authors assessed how table type and table location can affect average spending per minute (SPM) of a customer.

Lam summarizes Guéguen and Petr's research article with only this: "… particular scents also have an effect: diners who got a whiff of lavender stayed longer and spent more than those who smelled lemon, or no scent" (Lam, 2015). Similarly, Lam provides a terse summary for Kimes and Robson's article: "Diners at banquettes stayed the longest … Diners at bad tables—next to the kitchen door, say—spent nearly as much as others but soon fled." The students were asked to reconcile these statements with the conclusions in the original research articles. Lam's summaries are intentionally terse: The format provides an opportunity to show students the pitfalls of taking short news articles at face value.

Results

The activity was conducted with introductory and intermediate statistics students at Amherst College during the fall 2015 and spring 2016 academic semesters. The Amherst College Institutional Review Board (IRB) approved this study. On average, 20–25 students in each class engaged in the activity. Approximately 80 minutes were allotted to the activity.

In summary, students correctly identified basic conceptual elements in the designs of the original studies. These elements include sample size, the research question, conclusion, and classification of the study as observational or randomized. Many students were skeptical of the brief claims about the original studies given in the "Study of Studies."

For example, student work correctly identified Guéguen and Petr's sample size of patrons from a small pizzeria in Brittany, France. The students also describe how "lavender, but not lemon, increased the length of stay of customers and the amount of purchasing," which indicates that the students' picked up on Guéguen and Petr's hypothesis and research conclusions.

The students criticized the way the conclusions were portrayed in the "Diner Beware" article, pointing out that the "Diner Beware" summary does not account for the possibility of "cultural bias/geographical bias." Geographical bias stems from the fact that the study was only conducted in a small town in France, so the conclusions regarding scent and customer spending behavior may not generalize well to people of non-French heritage or individuals from urban areas.

In their article, Guéguen and Petr acknowledge that a small sample size and the use of only one restaurant are limitations of their study; students picked up on these caveats. Students also recognized that "limiting the study to three Saturdays in May between 8 p.m.–11 p.m. further creates sampling bias (targets a specific population)." Daytime and weekday visitors are evidently not represented. Moreover, since there was no replication, it is highly possible that another factor may have confounded the results.

Student analysis of Kimes and Robson's article also revealed comprehension of the research's design. In their analysis, the students correctly identified the sample size of 1,413 and the single-blinded nature of the study, since in students' words, "the participants did not know the true nature of the experiment." The students expressed skepticism regarding the causal statements made in Lam's article regarding Kimes and Robson's study, considering the observational nature of the study and the fact that Kimes and Robson "excluded some information, like the bar and patio seating" and that they "only took data from busy times."

Kimes and Robson's limitations stem from the fact that they only used one restaurant to draw conclusions, a shortcoming that relates to the limitation students picked up in regard to the limited focus of the study (i.e., inattention paid to less-busy hours).

Another student group summarized the limitation of Lam's synthesis as the inability to generalize the original research's findings, since the "conclusion for this specific restaurant may not apply to all restaurants."

Students' propensity for critique allows for challenging conventions, which produces a skepticism- and curiosity-driven outlook. This outlook, though, may have to be calibrated since student skepticism may be excessive.

Discussion

In an activity that linked summaries of research studies with published scientific papers, students generally accurately reported the original research's study design—in particular, the study's sample size; whether it was observational or experimental; and the general hypothesis, as well as the overarching conclusions. Students were often critical of the extremely terse representations of the original research in Lam's "Diner Beware" article in the *Atlantic*'s "Study of Studies" column. This is not surprising, given that the goal of the "Study of Studies" is to introduce provocative or idiosyncratic research findings, not to review or assess them comprehensively.

If time permits, the instructor might spend some time on debunking misplaced criticism, ensuring that students have a thorough understanding of the original research, can acknowledge credible published findings, and do not develop "knee-jerk" skepticism.

Overall, this activity was implemented successfully. It raised awareness about study design and secondary representations of original research. The activity can be undertaken with introductory and intermediate statistics students in a single class period, and may help improve communication skills by fostering discussion about experimental design.

We recommend that the study be undertaken after one lecture or more in study design. Conducting the study after few lectures would provide an informal student assessment and in the process, help reinforce previously learned study design concepts.

Numerous other articles published in the "Study of Studies" column could be used in the same way as the "Diner Beware" article (see Table 1 for a comprehensive list of candidate articles).

Acknowledgments

The authors would like to acknowledge support from the Gregory Call Fund of the Amherst College Dean of the Faculty. The authors thank Tasheena Narraidoo for her assistance. ◘

Further Reading

ASA. 2014. Curriculum guidelines for undergraduate programs in statistical science. *www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx*.

ASA. 2014. Revised guidelines for assessment and instruction in statistics education (GAISE) college report. *www.amstat.org/education/gaise*.

Baumer, B., Çetinkaya Rundel, M., Bray, A., Loi, L., and Horton, N.J. 2014. R markdown: Integrating a reproducible analysis tool into introductory statistics. *Technology Innovations in Statistics Education 8*:1–29.

De Veaux, D., Velleman, P.F., and Bock, D.E. 2013. *Intro Stats* (4th ed.). Pearson.

Diez, D.M., Barr, C. D., and Çetinkaya Rundel, M. 2015. *OpenIntro Statistics* (3rd ed.).

Guéguen, N., and Petr, C. 2006. Odors and consumer behavior in a restaurant. *International Journal of Hospitality Management 25*:335–339.

Kimes, S.E., and Robson, S.K.A. 2004. The impact of restaurant table characteristics on meal duration and spending. *Cornell Hotel and Restaurant Administration Quarterly 45*:333–346.

Lam, B. 2015. Diner beware: How restaurants trick you into eating less and spending more. *The Atlantic 4*:29.

About the Authors

Azka Javaid is a data scientist at IBM Watson Platform for Health Analytics. She received her bachelor's degree in computer science and statistics from Amherst College in 2017. Her interests include R software development and data engineering. In her free time, she enjoys data wrangling and building Shiny applications. She has published work on assessing race and socioeconomic disparities in critical care settings.

Xiaofei Wang is a lecturer/research scholar in the Department of Statistics and Data Science at Yale University. She obtained her doctorate in statistics from Yale University in 2014. Her current interests span statistical consulting, Bayesian statistics, statistical computing, statistics education, and machine learning.

Nicholas Horton is a professor of statistics at Amherst College, with interests in longitudinal regression, missing data methods, statistical computing, and statistical and data science education. He co-authored a series of books about statistical computing and data science. He is an accredited statistician, a fellow of the American Statistical Association, former member of the ASA board of directors, and past chair of the ASA's section on Statistical Education.

Bob Oster and Ed Gracely
Column Editors

Welcome to the newest column in *CHANCE*, devoted to teaching statistics, especially in the health sciences.

We won't deal with statistics classes in grade school (not around when many of us were young), nor with the education of future statisticians. Our interest is statistical pedagogy, mainly for those who will need to *understand* data analyses (such as for reading journals in their fields) and/or *use* statistics as a part of their own research, but are not going to be statisticians.

The ASA Section on Teaching of Statistics in the Health Sciences (TSHS) focuses its educational efforts on health professionals. From medical students to physicians, physical therapists to music therapists, biomedical bench scientists to epidemiologists, and more, they all read extensively in their respective literatures, and many of them perform research in their fields as well. We teach them all!

Don't stop reading this section if you work in a different field, however. Much of what we do in TSHS is quite general and overlaps with what other statistical educators do. From best practices and how to explain *p*-values to novices, to innovative and modern techniques like flipped classrooms, clickers, and problem-based learning, we touch on a wide variety of useful methods. If learning new techniques and ideas interests you, whatever you teach, then check out this column. Skilled teachers will share their experiences and ideas, and a variety of other content likely to be of interest to statistics educators.

We start our new column with this article about peer consulting.

Life on an Island: Using Peer Consulting in Applied Statistics Courses

Aimee Schwab-McCoy

Picture the first class session after a long holiday break. Ask your students, "Did you work on your semester projects over the break?" How would they react? Maybe confusion: "Were we supposed to be working?" Maybe confidence: "Of *course* we were working, professor!" Now, picture how they might react when you say this: "Well, I hope you weren't working on them, because as of today, they're *someone else's projects*! Your new task for the rest of the semester is to serve as *statistical consultants* for your fellow students."

This is exactly what the students at the University of Nebraska-Lincoln (UNL) heard in a second statistics course, as an exercise in peer consulting.

Clearly, statistical consulting is an important piece of professional statistical practice, but it is an experience that few students have in their undergraduate statistics curriculum. While the Center for Applied Statistics and Evaluation (CASE) at Truman State University and the Statistical Consulting Center at Winona State University provide excellent examples of thriving undergraduate consulting centers, many statistics programs may not have the time or resources to dedicate to an undergraduate consulting experience. An alternative option is to introduce peer consulting into the curriculum through in-class activities or semester-long projects.

What is "peer consulting" and why is it useful? Simply put, in a peer consulting experience, students play out the roles of consultants and clients in a research setting. This column describes one model for implementing peer consulting in an undergraduate statistics course, and offers tips

Team A	Team B	Team C	Team D	Team E
• Client: Team B • Consultant: Team E	• Client: Team C • Consultant: Team A	• Client: Team D • Consultant: Team B	• Client: Team E • Consultant: Team C	• Client: Team A • Consultant: Team D

Figure 1. How a hypothetical class with five student groups might have client-consultant pairs assigned. Each student group is now working with two others: as the client of one group and the consultant to another.

and guidelines for use in your own classrooms.

Giving students realistic peer consulting experiences was the motivation for a semester-long project in "STAT 318: Introduction to Statistics II" at the University of Nebraska-Lincoln. The course is a follow-up to a traditional algebra-based introductory course, covering topics such as analysis of variance, experimental design, non-parametrics and randomization tests, and simple/multiple linear regression. The course emphasizes professional statistical skills and communication, and usually assesses that with at least one significant writing component. Enrollment typically ranges from 15–25 students, enabling active in-class discussion and team exercises.

In the spring 2014 semester, 22 students were enrolled, most of whom were juniors and seniors. In a "getting to know you" informal survey administered during the first week of class, most of the students expressed interest in pursuing either graduate school in statistics or employment as a statistician or data scientist. Many, but not all, of the students were mathematics majors and in general, students felt confident in their mathematical skills. However, students expressed significantly lower levels of confidence in their hands-on data analysis skills.

The class project began like many statistics projects typically do. During Phase One, students worked in teams of two or three to design an experimental or observational study to carry out on a virtual population of subjects living on the Islands (*http://islands.smp.uq.edu.au*). Each group collected four to six variables during their experiment, in addition to basic demographic information such as name, age, and hometown.

The students submitted written project proposals at the end of Week 6 in a 16-week semester that included the research objective and study motivation, a literature review, and a complete outline of the study procedure. Each group received feedback on their project proposals before being approved to begin data collection. Finalized data sets were due at the end of Week 10, which gave the student groups about four weeks to complete their data collection process. At UNL, Week 10 is usually immediately before spring break.

Date	Material Due
Week 3	Tentative research objective, project groups
Week 6	Project proposals
Week 10	Data collection completed
Spring break	Client-consultant pairs assigned
Week 11	Initial in-class consultation
Week 14	Second in-class consultation
Week 15	Presentations

After the students returned from spring break, they were in for a surprise: During the first class meeting, the *real* objective of the project was explained—each student group would now work as statistical consultants on a project other than their own. The assignments went in a "round robin" fashion. For example, Team B's data would be given to Team C, who would serve as their statistical consultants and continue the analysis. Team B then receives Team A's data, and in addition to being the *clients* for Team C, they would now also be the *consultants* for Team A (see Figure 1).

Assigning student groups to client-consultant pairs had several advantages. In terms of classroom dynamics, teams had been working well together on their projects up to this point, but tended to use the same groupings for in-class activities and out-of-class study. With the client-consultant pairs, students were now working with others outside their majors or established peer groups.

More importantly, this gave the student groups experience in the role of a real statistical consultant. Very few of these students had any prior research or data analysis experiences beyond the classroom, and none had ever worked with a client. This allowed students to practice the "soft" statistical skills of written and verbal communication, as well as listening to a client's goals and expertise on a topic.

The student groups in this course also experienced a side of statistical consulting that very few statisticians themselves ever see: the client's perspective. Students were engaged in their projects, and had developed a genuine curiosity about what the results would show. By allowing other groups to analyze their data, they experienced what it is like to be the one asking the questions about an analysis.

Client-consultant pairs met twice during class to discuss the project. At the end of Week 11, each group met to ask their "clients" questions about the research project they would be picking up. Each team participated in an initial consultation twice: once as the consultant, and once as the client. The initial consultations were "client-led": Each client group was expected to describe the research objective for their consultants, and explain any necessary background about the research objective or study design. Consultants took notes and asked questions where necessary.

Three weeks later, near the end of the semester, client-consultant pairs met again to discuss their preliminary analyses. This time, the "consultants" took the lead during the meetings to present preliminary data analysis results, justify selection of statistical procedures or models, and provide an initial answer to the research objective. During this meeting, the "clients" asked questions, took notes, and suggested additional parts of the study design that might be relevant to the final analysis.

For each client-consultant meeting, the clients and consultants received a list of possible objectives or points of discussion to help them get started, but the discussions were largely their own.

To aid with assessment, all client-consultant meetings that took place during class were recorded. This allowed the instructor to check in with other groups as the meetings were happening in one area of the classroom, and encouraged the students to take charge of the discussions since the instructor was not there to guide them. Grades were assigned for each client-consultant meeting after class based on the recordings; these were mostly participation-based.

Consultants presented two final products: a 5- to 10-minute in-class presentation summarizing their results and statistical methods used, and a research poster detailing the analysis in more depth. These posters were similar to what would be presented at an undergraduate research fair. Each student also submitted a "self-assessment" at the end of the course. In this self-assessment, students evaluated themselves in three roles: as clients, as consultants, and in their partner groups.

This peer consulting experience was designed to teach a variety of professional statistical skills not typically found in the undergraduate curriculum. Students learned statistical *thinking* skills by working as a team to answer a scientific research question by designing a study and collecting data. Students also learned first-hand what it is like to join a statistical investigation after data collection, and obtained practice in making decisions about the most-effective or appropriate analysis for a given situation.

All studies contained at least four to six variables, in addition to demographic information about each Islander, giving students a rich, realistic, multivariate data set to explore.

In most consulting situations, the client is interested in the contextual implications, conceptual understanding, and bigger picture of the analysis, not necessarily the fine-tuned details. In this project, students learned about the nuances of statistical writing for two different professional audiences: the fellow statistician and the client. The data collected in this project is, of course, not real, since the Islanders are a computer-simulated population. However, it is *realistic*.

On the Island, data are not recorded automatically. Students must record the data themselves, much like most researchers in the social sciences and STEM fields. This may seem like a hassle to students, but it mimics the actual data collection process in scientific research. Students get a firsthand sense for the costs associated with collecting data, both in terms of financial resources with real subjects and the researcher's time commitment.

From the instructor's perspective, there were several benefits and challenges to this peer consulting project. Some of the benefits included:

- Increased communication between student groups—by assigning student teams to work with other student teams, there were more pathways for students to work together.
- More accountability in terms of group performance—if a group or team member didn't meet expectations, they weren't just letting down their instructor. They also had a client with expectations for their work.
- More realistic motivation to learn and practice—group work was not just to keep students busy; the project was designed to be as close to real life as possible in terms of project deliverables (report for client and scientific poster), timeline (6 weeks overall to do the consulting), and design and data collection.

- Increased curiosity—students working with the Island tend to grow attached to their virtual population. Since this is a designed population, students and instructors don't know which parts of the simulation are "true" to real life and which ones aren't, further stimulating student curiosity.

- Exposure to multiple studies—students designed their own studies and collected data, but then analyzed completely different scenarios. To do this, students had to think about how experimental design and statistical analysis interact. They may have designed a study thinking it was a "regression example," but their final consulting product may have ended up being drastically different.

- A view into the life of a consulting statistician—to the statistician's lament, we are often brought in to a study or experiment *after* data are collected, not before. By the time the students viewed their clients' data, their statistical skills had become more sophisticated, and many groups were able to identify things they would have done differently in both their clients' data and their own data. At the end of the semester, a student wrote: "one big benefit of presenting another group's project is when we found issues with the data they'd gathered, we couldn't go back and 'fix' it by gathering more data. … [it] much more closely mirrored real life."

Some of the challenges unique to peer consulting projects are:

- Increased time commitment for the instructor—not only is the instructor supervising and monitoring student teams, but now there is now the added dimension of how student teams interact with each other. Making expectations clear can help. Consider building a common communication space in your school's learning management system (LMS) for shared documents and to monitor student participation.

- Less-than-favorable student reactions—some students will be very attached to their research question or objective, and may not see the point of peer consulting immediately. Sell it! Explain your rationale clearly, use examples of your own consulting work, or perhaps ask a professional consultant to speak to the class about the importance of consulting in statistical practice. Some students may continue to see it as added busywork, so the instructor must continue to emphasize the importance of learning good professional communication skills.

- Added time in class for client-consultant meetings—another option may be to schedule meetings outside of class during office hours—this is an even greater instructor time commitment. Formal meetings should be assessed in some way, not necessarily by the instructor. Consider assigning students reflective writing prompts after each meeting, to assess their experiences as clients and as consultants.

- Timing and execution—timing is critical to introducing peer consulting exercises. Wait too long and students will begin analyzing their original data set. In this course, the consulting element was introduced as a surprise, but for a variety of reasons (scheduling, rumors from former students), this might not always work.

What is most important about the peer consulting exercise isn't the surprise element—although that part is fun—but the *investment*. Students should be invested in the project on both sides, as a client and as a consultant. If students know that their projects will be switched in the second half of the semester, they may be tempted to put less effort into the experimental design or data collection. This could be avoided by grading student research proposals, or even giving students a choice about which project they'd like to peer consult for.

No matter whether the switch to a peer consulting model is known ahead of time or a surprise, consider designing student projects with checkpoints to prevent "working ahead" of the switch.

This peer consulting project was used successfully in a second statistics course, but there are ways to implement peer consulting in courses at other levels. Instructors might consider using consulting projects across the curriculum, where students in an upper-level statistics course could serve as peer consultants for a lower-level course. Another possible implementation might be collaborating across sections, schools, or departments. The added separation may make the consulting experience feel more realistic, and this gives the instructor the chance to collaborate and share the added workload of peer consulting projects with colleagues in other departments or institutions.

A major component of this project was the client-consultant meetings. Where possible, client-consultant meetings should include both an in-person component to help students develop interpersonal communication

WELCOME TO THE ISLANDS

What is "The Islands"? The Islands is a simulated population of human subjects explicitly designed to support learning about experimental design, data collection, and statistical inference and modeling for statistics courses. The Islands was initially developed by Michael Bulmer of the University of Australia, Queensland, but is now used as a teaching tool in courses worldwide. Think of it like playing the Sims in your statistics courses: Students can design any research study they like, and carry it out on "human" subjects without the various hurdles to actually collecting data on real human subjects.

The background story of the Islands goes like this: The Islands comprise a group of three virtual island nations—Ironbard, Providence, and Bonne Santé—that were settled by a small band of shipwreck survivors more than two centuries ago. As time has passed, all three Island nations have flourished into a rich society with more than 40,000 residents. When students first "arrive" at the Islands, they'll be greeted with a map of all three island nations, with important settlements and geographic landmarks identified. From there, students can head to any of the 27 towns and villages to find their subjects, and start experimenting.

Once virtual Islander have "granted consent" to participate in a research study, they can be asked a series of survey questions or complete any one of an extensive list of tasks. Their life histories are also available for students to view, including their friendships with other Islanders, disease histories, academic transcripts, and net wealth. Students can gather and record the data from each subject, then move onto the next participant.

Students using the Islands in the classroom have a wealth of resources to get started. The first stop for students should be the visitor center near Arcadia, where they can gain a brief orientation to the Islands and read some frequently asked questions from previous scientific expeditions (statistics classes). Students can also choose to study towns and villages as a whole by visiting the local bureau to peruse employment records, or perhaps the town hall to study causes of death.

In some towns, the local clinic can provide a list of recent patients, along with their symptoms, diagnoses, and outcomes. Those interested in Island history and folklore may choose to visit the Museum to read logbooks from the initial settlers, and maybe even contact their ghosts. The climate stations on each island provide some insight into local weather patterns, and experimental field stations give students the option to study local agriculture.

Finally, a student in need of some inspiration might visit the Academy to browse academic journals such as *Proceedings of the Islands Academy* or the *Journal of Island Studies* (both of which contain examples of previously submitted student projects).

There are many advantages to using the Islands in courses. Giving students access to a rich virtual population encourages them to investigate interesting research questions in a more-realistic way. Students experience the practical concerns of data collection, such as choosing a suitable sample, and learn that collecting real data takes time and effort. Interested in teaching with the Islands? Check out the resources below to get started.

and "conversational statistical literacy," and a written component to encourage clear and concise statistical writing. There should be an assessment of the client-consultant meetings, to help hold students accountable for reaching objectives and being actively involved in their roles. One suggestion is to have students watch recordings of their own consulting meetings and reflect on their discussions.

Finally, a good way to get started with client-consultant meetings is to model a statistical consulting meeting for students. This gives students a clear idea of the expectations associated with client-consultant meetings and a sense of the importance of statistical consulting to practice.

In their course evaluations and final self-assessments, many students reported that peer consulting was an enriching experience. It was also a rewarding challenge to take on as the instructor. If you're interested in designing peer consulting projects or activities for your own courses, make sure to:

1. Consider timing carefully. Switching groups after a break was ideal, because generally students don't do academic work over a break. This avoids students doing work on a project analysis before switching groups.

2. Sell it! Explain the purpose of the switch and the importance of consulting to practice. Emphasize the importance of building communication skills—they aren't expected to be perfect consultants immediately. This is a chance for them to build and develop skills in a collaborative environment.

3. Make expectations clear from the beginning. For most, if not all, this will be a completely new experience. Let them know what your expectations are, and what their expectations should be of their clients and consultants.

About the Author

Aimee Schwab-McCoy is an assistant professor of statistics in the Mathematics Department at Creighton University. She received her PhD in statistics from the University of Nebraska-Lincoln. Her current research projects revolve around curriculum development and assessment in statistics, data science, and quantitative literacy. She is also an active statistical consultant with partnerships in the health and social sciences.

Further Reading

Baglin, J., Bedford, A., and Bulmer, M. 2013. Students' experiences and perceptions of using a virtual environment for project-based assessment in an online introductory statistics course. *Technology Innovations in Statistics Education* 7(2):1–15.

Bulmer, M. 2010. Technologies for enhancing project assessment in large classes. In C. Reading (ed.), *Proceedings of the Eighth International Conference on Teaching Statistic*s. Ljubljana, Slovenia.

Carver, R.H., Everson, M.G., Gabrosek, J., Holmes Rowell, G., Horton, N.J., Lock, R., Mocko, M., Rossman, A., Holmes Rowell, G., Velleman, P., Witmer, J., and Wood, B. 2016. *Guidelines for Assessment and Instruction in Statistics Education: College Report*. Washington, DC: American Statistical Association.

The Islands in Schools Project: *www.islandsinschools.com.au/*.