Rick Katz and Eric Gilleland

Extreme Events Tutorial

Climate impacts society in large part through extreme events (e.g., deaths associated with an intense heat wave). We review the statistical theory of extreme values, emphasizing on applications to weather and climate extremes and their impacts. Extreme value theory does not at all resemble the corresponding statistical theory for averages, in which the normal distribution arises through the central limit theorem.

The natural stochastic model for the occurrence of rare events is a Poisson process, with the frequency of rare events having an approximate Poisson distribution. At least in part for this reason, the Poisson distribution is commonly used to model mortality or morbidity rates (i.e., count data) in epidemiological studies (e.g., relating the number of deaths to environmental variables).

The fundamental theorem for extremes establishes that the distribution of the maximum of a sequence of observations (independent and identically distributed) must be approximately in the form of the generalized extreme value (GEV) distribution. In practice, the GEV is fitted to "block maxima" (e.g., the highest temperature over a year). Through a simple transformation, the theory applies equally well to minima (e.g., lowest temperature over a year). The GEV distribution includes three types: the Type (i) or Gumbel distribution, the Type (ii) or Fréchet distribution, and the Type (iii) or Weibull distribution. None of these distributions closely resembles the normal. The Weibull type typically arises as an approximation for annual maxima or minima of temperature, the Fréchet for annual maxima of precipitation amount.

The "peaks over threshold" is an alternative approach, more flexible and potentially more powerful. Originally developed in hydrology, the basic idea is to exploit more of the available information about the upper tail of the distribution (e.g., the two highest values in the entire record might have occurred in the same year). The approach involves combining the Poisson process for the occurrence of an exceedance of a high threshold with a generalized Pareto (GP) distribution for the excess over the threshold. Because the maximum of a sequence falls below a threshold if and only if no exceedances of the threshold occur, the peaks over threshold approach can be used to indirectly fit the GEV distribution for block maxima.

The GP approximation for the upper tail of a distribution is the analogue of the GEV approximation for block maxima. It has three types: the type (i) or exponential distribution, the type (ii) or Pareto distribution, and the type (iii) or beta distribution, with virtually all of the distributions used in climate research (e.g., normal, gamma, lognormal) having an approximately exponential upper tail. The beta type typically arises as an approximation for the upper or lower tail of the distribution of daily temperature, the Pareto for the upper tail of daily precipitation amount.

One advantage of extreme value methodology is that covariates can be readily incorporated into the models in a regression-like approach, but still consistent with the known statistical properties of extremes. For example, the GEV distribution can be fitted to annual maxima with a trend in one or more of its parameters (in this way, the likelihood of extreme events can shift with climate change). With the peaks over threshold approach, annual cycles in daily temperature or precipitation extremes (e.g., by using a sine wave as a covariate) can be permitted. Besides time, covariates can be genuine geophysical variables (e.g., the state of the El Niño phenomenon).

It is common to communicate the risk of extreme events in terms of return periods and return levels. A "return period" is the inverse of the probability of an event (e.g., the so-called "100-yr flood" has a 1% chance of occurring in a given year). Associated with the specified return period is a "return level" (e.g., the magnitude of flood corresponding to a 100-yr return period). An advantage of the statistical theory of extremes is that it permits confidence intervals to be attached to estimated return levels, taking into account sampling error.

In the peaks over threshold approach, one issue that arises is the clustering of extremes because of the temporal dependence of climate variables. Such clustering is evident for temperature extremes, but not necessarily for precipitation extremes. In practice, clusters are identified and the data are "de-clustered" by only fitting the GP distribution to cluster maxima. One way to characterize heat waves, of particular interest because of their health impacts, would be in terms of statistics of such clusters (length, magnitude, etc.).

In epidemiological studies, Poisson regression (i.e., a Poisson distribution for mortality or morbidity rates being modeled as a function of covariates such as environmental variables) is the methodology of choice. Not many direct applications of the statistical theory of extreme values in epidemiology exist. In research on longevity, the GEV distribution has been fitted to maximum life span. So it would be a natural generalization to use extremal distributions as a model for the longest possible survival times after the onset of a particular disease.

Until recent years, a major obstacle to the application of the statistics of extremes was the lack of software. Now software exists to fit extremal models, including both the block maxima and peaks over threshold approaches and permitting the inclusion of covariates. A tutorial is provided using the software package ("extRemes"), consisting of a graphical user interface with functions being written in the open-source statistical programming language R.

Recommended reading

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Katz, R.W., M.B. Parlange, and P. Naveau, 2002: Statistics of extremes in hydrology. *Advances in Water Resources*, **25**, 1287-1304. (available at: www.esig.ucar.edu/HP_rick/awr.pdf)

Web resources

Extremes Toolkit: www.esig.ucar.edu/extremevalues/evtk.html

Extremes Tutorial: www.esig.ucar.edu/extremevalues/tutorial/ (html version), www.esig.ucar.edu/extremevalues/tutorial.pdf (pdf file)

R statistical programming language: www.r-project.org

Statistics of Weather and Climate Extremes: www.esig.ucar.edu/extremevalues/extreme.html

\

Background

Rick Katz is Senior Scientist in the Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, CO. He received a Ph.D. in statistics from Pennsylvania State University. He is one of the founders of the NCAR Geophysical Statistics Project, an activity which promotes collaboration between geophysical scientists and statisticians. His research interests focus on the application of statistics to weather and climate and their societal impacts. Specific areas of application include assessing the economic value of weather and climate forecasts, stochastic modeling of climate time series, statistical downscaling, and statistics of extreme weather and climate events and their impacts.