Uncertainty in Ranking the Hottest Years of U.S. Surface Temperatures

PETER GUTTORP

University of Washington, Seattle, Washington, and Norwegian Computing Center, Oslo, Norway

TAE YEN KIM

University of Washington, Seattle, Washington

(Manuscript received 26 October 2012, in final form 19 February 2013)

ABSTRACT

Ranking years based on statistical estimates of regional and temporal averages is subject to uncertainty. This uncertainty can in fact be quite substantial and can be described by the rank distribution of an ensemble of such averages. The authors develop a method for estimating it using simulation. The effect of temporal correlation is quite limited in the case studied in this paper: the contiguous United States' annual-mean temperature. The method also allows assessment of derived quantities such as the probability of a given year being one of the 10 warmest in the historical record.

1. Introduction

A popular indicator of global warming is the ranking of recent years among the historical record. This can be done for individual stations, for countries, for continents, or globally. For example, the recent paper by Shen et al. (2012) produced lists of the 10 hottest and 10 coldest years for the contiguous United States. This ranking was based on their statistical estimate of U.S. temperature anomalies using the U.S. Historical Climatology Network dataset, version 2 (Menne et al. 2009), corrected for time of observation differences. In this note we will use the same data, focusing only on the annual T_{mean}.

The ranking in Shen et al. (2012) was based on their statistical estimate of U.S. temperature. However, the ranking of years must take into account the fact that U.S. land temperature is a statistical estimate, not a direct measurement, and as such has a standard error, with components from measurement error at individual stations, spatial dependence, orographic effects, etc. The paper by Shen et al. is mainly concerned with the statistical estimation of this standard error. The data used by Shen et al. are not the same as those used by the National Oceanic and Atmospheric Administration (NOAA) to estimate U.S. contiguous states' average temperature, as the latter have been subject to data homogenization while the former have not.

How does the uncertainty in the statistical estimate of U.S. annual-average land temperature affect rankings? For example, from Fig. 1 (adapted from Fig. 6 in Shen et al. 2012), we see that the year 1921 (green horizontal line), ranked fourth in their Table 4, may actually have been the hottest year on record, because the higher green horizontal dashed line, corresponding to the upper end of a 95% confidence interval for the actual average yearly U.S. temperature for 1921, crosses the intervals for all three years that are ranked hotter than 1921. On the other hand, the same year can be ranked as low as fourteenth, because the lower green horizontal dashed line (the lower end of the confidence interval) crosses or falls below 13 of the confidence intervals. Now, these intervals are not simultaneous (i.e., the joint confidence level of all the intervals taken together is less than 95%). Thus, a more precise assessment of the uncertainty in rankings is needed. That is the purpose of this note.

In section 2, we describe a simple approach to simulating the distribution of ranks for each year from an ensemble of time series with the same mean and variance as the Shen et al. (2012) series. The spread of this rank distribution describes the uncertainty in the rank for that year. The method in section 2 does not take into account the temporal dependence of the time series of

Corresponding author address: Peter Guttorp, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
E-mail: peter@stat.washington.edu

DOI: 10.1175/JCLI-D-12-00760.1

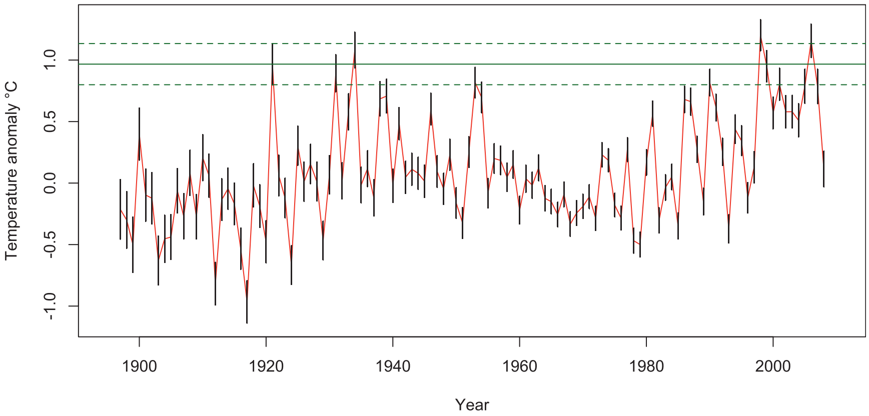

FIG. 1. Temperature anomalies for contiguous U.S. annual-average temperature in red. The black vertical lines are 95% confidence limits for the corresponding values (Shen et al. 2012). The solid green horizontal line corresponds to the average temperature anomaly for 1921, and the dashed green horizontal lines are the upper and lower 95% confidence limits for that value.

land temperature averages, and section 3 describes how a more realistic simulation is done. We close the paper with a discussion in section 4.

2. Simulating the distribution of ranks

If the U.S. temperatures at different years can be treated as independent, it is fairly easy to simulate a temperature series with the same stochastic structure (i.e., mean and standard deviation) as the observations. Because the averaging is over a fairly large number of stations, many of which are only weakly spatially dependent, a normal assumption for the yearly estimates is reasonable, using a central limit theorem such as that of Pinkse et al. (2007). Thus, for each year [using the function rnorm in the statistical software R (R Development Core Team 2012); the code used for all calculations in the paper and the complete versions of Figs. 3 and 5 (described in greater detail below) are freely available at http://www.statmos.washington.edu/wp/?p=696] we simply draw a sequence of random normal numbers with mean equal to the yearly U.S. average temperature and a standard deviation equal to its standard error. Figure 2 shows 10 such simulated paths together with the original time series (in blue) from Fig. 1.

We then calculate the rank of each year in the simulated series (using the R function rank). Repeating a large number of times (100 000 in order to be near certain of the accuracy of two decimal places in proportions), we obtain a distribution of ranks for each year. Figure 3 shows the rank distribution for the years 1997–2008. The number in the header is the simulated probability that the given year is the warmest in the record from 1897 to 2008.

We see from Fig. 3 that years such as 1997 and 2008 have a very poorly defined rank, that is, that the exact ascending rank [70 and 67 out of 112 in the series by Shen et al. (2012)] is very uncertain, while 1998 and 2006 are clearly among the warmest years in the record, having low rank uncertainty; 1998 has a probability of 0.64 of being the warmest year, and 2006 has a probability of 0.28. How about 1921? It turns out to have a probability

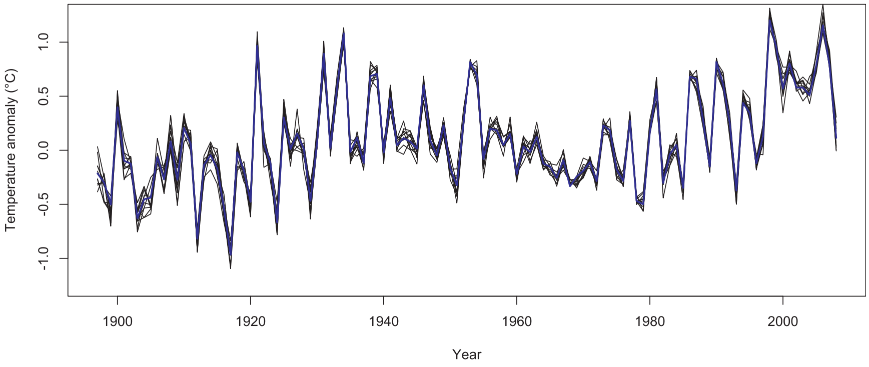

FIG. 2. Ten simulated paths (in black) and the original series (in blue).

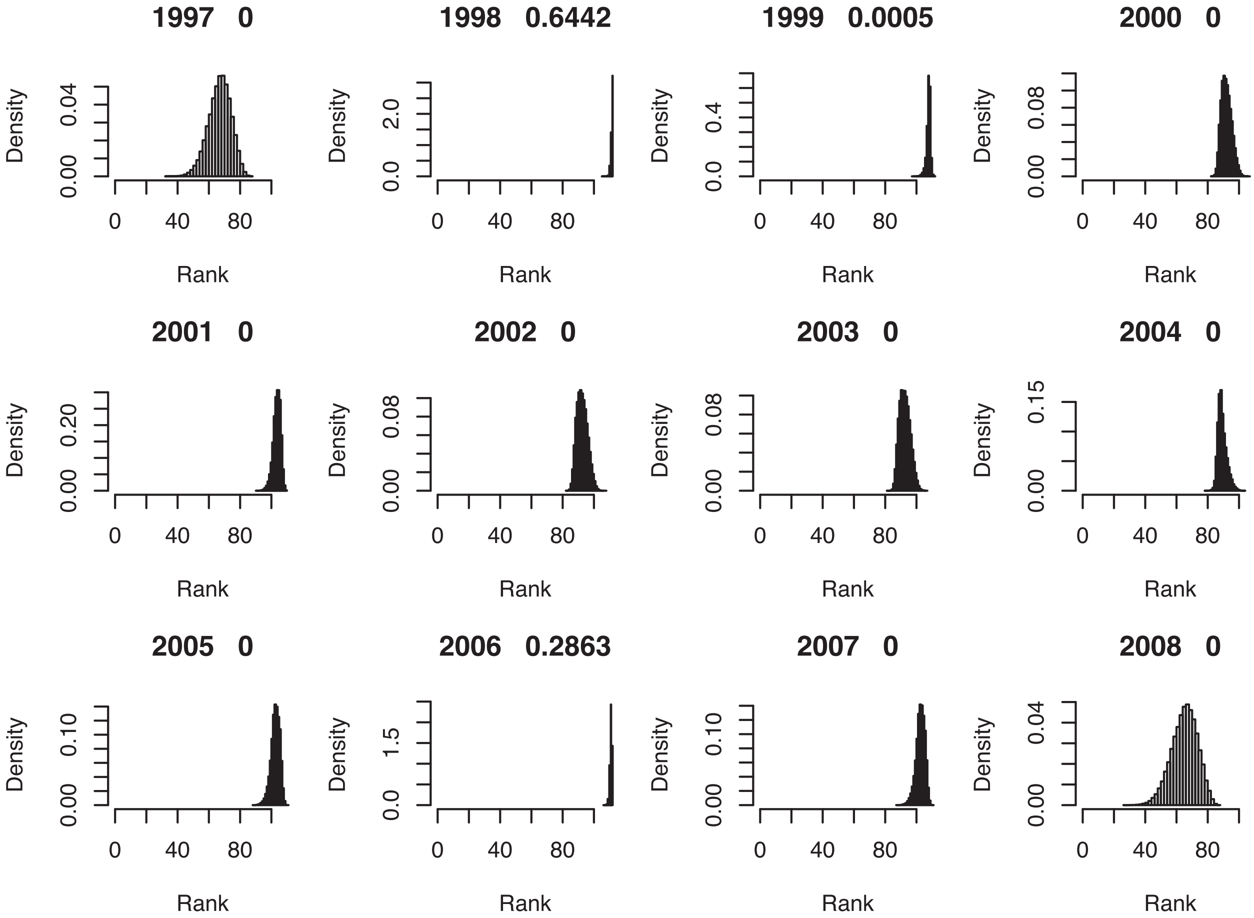

FIG. 3. Simulated rank distributions in ascending order for the years 1997–2008, assuming normality and independence between years and using 100 000 simulated paths. The right-hand number in each figure panel heading is the estimated probability of that year being the warmest among the 112 years (1897–2008) considered.

of 0.01 of being the warmest year on record under the assumptions in this section.

3. Taking into account temporal dependence

The simulation in section 2 assumed that different years are independent, and that the nonstationarity in the data only results from the mean and variance changing with time. However, because of large-scale circulation patterns and atmosphere–ocean interactions, we would expect successive years to be dependent random variables.

To estimate the dependence pattern, we first remove a linear trend using ordinary least squares. Studying the dependence structure of the residuals from this trend using the autocorrelation structure and the Akaike information criterion (AIC; R function auto.arima) suggests an autoregressive-moving-average (ARMA) model of orders 3 and 1, respectively. We fit this model using generalized least squares (so that we take proper account of the dependence as well as the varying variance; R function gls). The normalized residuals from this fit are uncorrelated [as checked by spectral analysis, autocorrelation function, time series order estimation using AIC, and a white noise test by Lobato and Velasco (2004) that yields a p value of 0.977] and Gaussian (as tested by a Q–Q plot with simultaneous confidence band). Using the Bayesian information criterion (BIC) instead of AIC to determine time series order is often recommended because of the tendency of BIC to select more parsimonious models. Indeed, in this case BIC selects a moving-average model of order 1, but there is considerable time series structure left in the residuals, as judged by the spectrum, autocorrelation function, and white noise test. We therefore will use the AIC-selected model.

We create new time series innovations by permuting the normalized residuals. Using the estimated time series model and the permuted innovations, we then compute a new path using the R function arima.sim. We multiply this path with the standard errors and add back

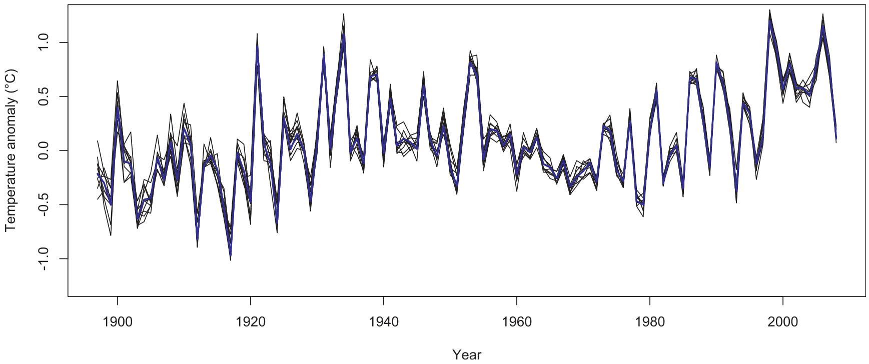

FIG. 4. Ten simulated dependent paths (in black) and the original series (in blue).

the U.S. yearly mean. This method is essentially the time series bootstrapping method given in Efron and Tibshirani (1993, section 8.5), with the slight modification that we are using a permutation bootstrap (instead of resampling from the residuals). The difference from the method in section 2 is that we use dependent random variables that closely mimic the estimated dependence structure in the original time series instead of the independent normal random variables used in section 2. Similarly to Fig. 2, Fig. 4 shows 10 simulated paths in black with the original monthly series in blue.

The counterpart of Fig. 3 for dependent data is Fig. 5. We see that the uncertainty about the ranks is somewhat larger when taking into account temporal dependence, but that the rank distributions look very similar. One way to describe this uncertainty is by computing the

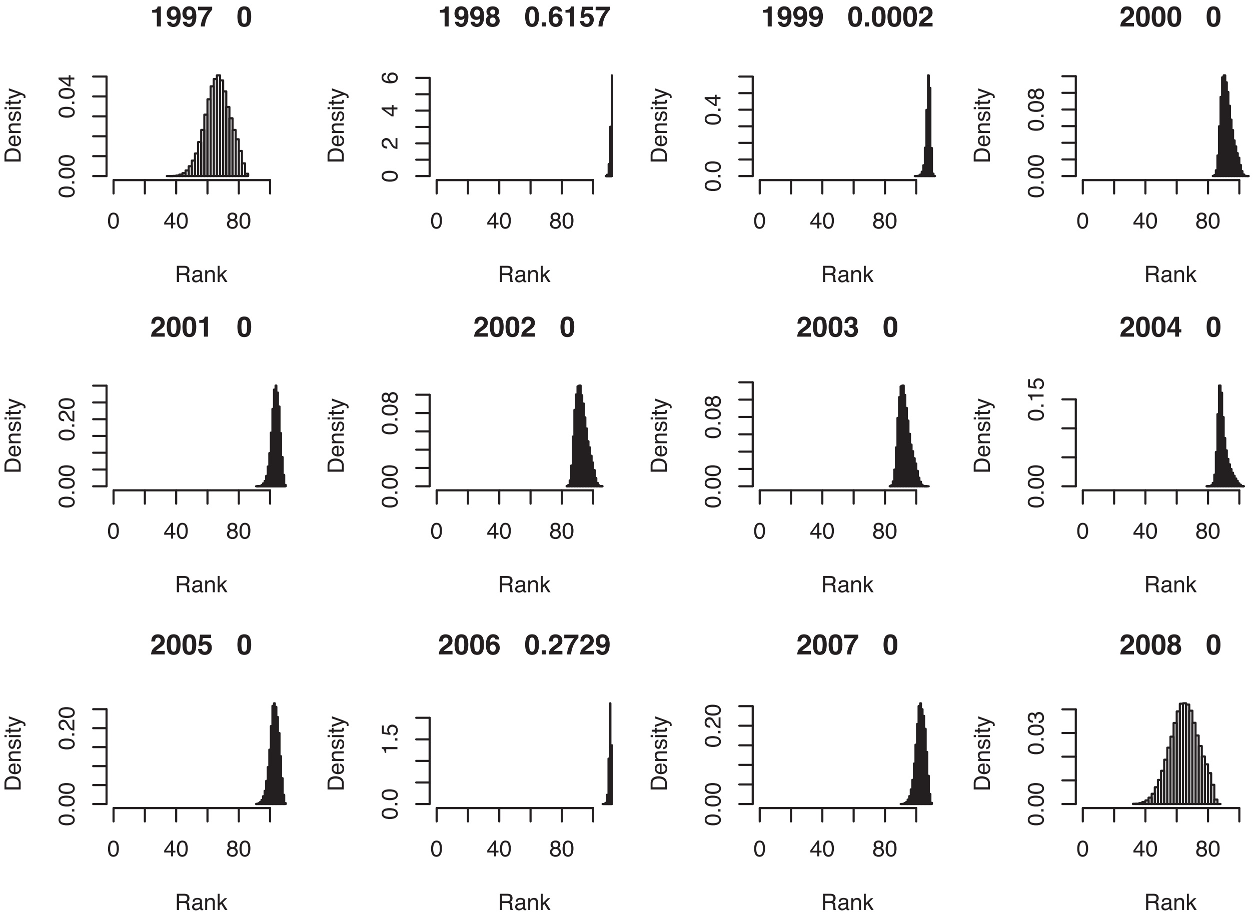

FIG. 5. As in Fig. 3, but for dependent data.

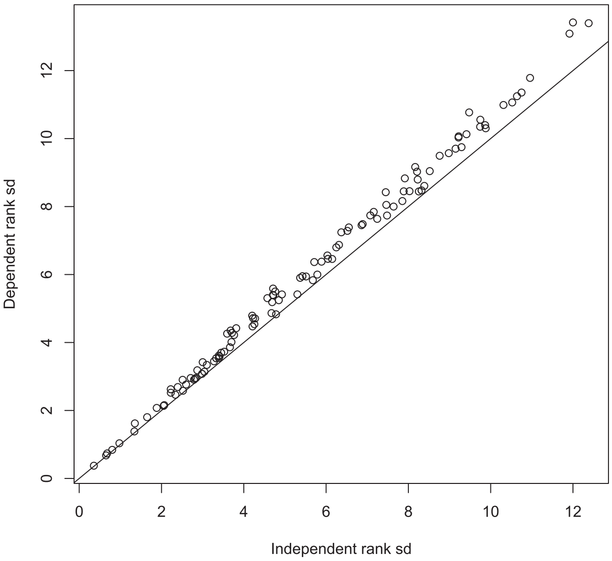

FIG. 6. Std dev of the rank distribution from the ensemble simulated in section 2, plotted against the std dev of the rank distribution from the ensemble simulated in section 3. The solid line has slope 1.

TABLE 1. Probability of a given year being among the 10 warmest on record.

Year	Independent model	Dependent model	Rank in Shen et al. (2012)
1921	0.99	0.99	4
1931	0.95	0.94	6
1934	1.00	1.00	3
1938	0.12	0.14	—
1939	0.18	0.21	—
1953	0.78	0.74	8
1954	0.12	0.15	—
1990	0.82	0.79	7
1998	1.00	1.00	1
1999	1.00	1.00	5
2001	0.71	0.69	9
2005	0.61	0.58	10
2006	1.00	1.00	2
2007	0.61	0.58	—

standard deviation by year of the rank distribution. Figure 6 shows the standard deviations of ranks in the independent ensemble plotted against those from the dependent ensemble, confirming the impression that the dependent rank distributions are slightly more spread out than the independent ones.

4. Discussion

The description of uncertainty in ranks using a simulated rank distribution allows easy calculation of derived quantities, such as the probability that each of the years 1897–2008 are among the 10 warmest. Because we have 100 000 rankings for each year, we just need to count the proportion of these ranks that are in the top 10 (i.e., ranked 103 or higher). Table 1 gives the results (we only present those years whose probability of being in the top 10 is at least 0.1), and a comparison with the ranking in Shen et al. (2012). Our list has four years that are not ranked top 10 by Shen et al.; all of their ranked years have top 10 probabilities of more than 0.5. The year 2007 (not ranked by Shen et al.) comes out with the same probability of being in the top 10 as the year 2005 (ranked 10). The temperature anomalies for 2005 and 2007 differ by 1.5×10^{-4}°C. Ranking the years using the average ensemble rank yields the same top 10 years as in Shen et al. (2012) for the independent ensemble, while in the dependent ensemble year 2007 replaces year 2005.

The Intergovernmental Panel on Climate Change fourth assessment synthesis report (Pachauri and Reisinger 2007) said that ''[e]leven of the last twelve years (1995–2006) rank among the twelve warmest years in the instrumental record of global surface temperature (since 1850).'' Using the Hadley Centre/Climatic Research Unit, version 4 (HadCRUT4), global temperature data (Jones et al. 2012), based on both land and marine observations, we can assess the uncertainty in the statement that the years 1995–2006 have 11 years in the top 12 since 1850 (this statement is true for the HadCRUT4 global series). Applying our method, again using independent simulations and the standard errors given by the Hadley Centre, we see that 11 years have a probability of 0.54. The probability that 10 years are in the top 12 is similar (0.42) and 9 years have a probability of 0.03, while the remaining probability is split between 8 or 12 years.

As a by-product to our method, we have estimated the linear trend in U.S. average annual temperature. Using ordinary least squares, this is highly significantly different from zero (0.554°C century^{-1} with a standard error of 0.117 and a p value of 6.1×10^{-6}). On the other hand, taking into account the changing variability and the autocorrelation structure, our generalized least squares estimate is 0.595°C century^{-1}, with a standard error of 0.320 and a p value of 0.066, similar in value but no longer significantly different from zero.

Since the original submission of this paper, the contiguous U.S. annual-mean temperature series has been updated through 2012, with the last year being the hottest on record. While we do not have standard error calculations for the NOAA series used in this ranking, we estimate them by extrapolating those of Shen et al. (2012) and transform the NOAA values to the scale used by Shen et al. The 2012 annual-mean temperature is apparently seven standard errors larger than the previously largest value (1998). This is highly unlikely under a stationary climate, and all 100 000 paths in our ensemble come out showing 2012 as the hottest year.

Acknowledgments. We are very grateful to Sam Shen for providing the U.S. average temperatures and their standard errors from his analysis. The comments of three referees have improved the presentation. Thanks to Anthony Arguez of NOAA for checking our code (and finding an error). This research was partially supported by the U.S. National Science Foundation Grant DMS-1106862, the STATMOS research network on Statistical Methods in Oceanic and Atmospheric Sciences.

REFERENCES

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman & Hall, 436 pp.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.,* **117,** D05127, doi:10.1029/2011JD017139.

Lobato, I., and C. Velasco, 2004: A simple and general test for white noise. Preprints, *Latin American Meeting 2004,* Santiago, Chile, Econometric Society, 112.

Menne, M. J., C. N. Williams Jr., and R. S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.,* **90,** 993–1007.

Pachauri, R. K., and A. Reisinger, Eds., 2007: *Climate Change 2007: Synthesis Report.* Intergovernmental Panel on Climate Change, 104 pp.

Pinkse, J., L. Shen, and M. Slade, 2007: A central limit theorem for endogenous locations and complex spatial interactions. *J. Econometrics,* **140,** 215–225.

R Development Core Team, 2012: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 409 pp.

Shen, S. S. P., S. K. Lee, and J. Lawrimore, 2012: Uncertainties, trends, and hottest and coldest years of U.S. surface air temperature since 1895: An update based on the USHCN V2 TOB data. *J. Climate,* **25,** 4185–4203.