Connecting Point-Level and Gridded Moments in the Analysis of Climate Data*

HANNAH DIRECTOR AND LUKE BORNN

Department of Statistics, Harvard University, Cambridge, Massachusetts

(Manuscript received 13 August 2014, in final form 31 December 2014)

ABSTRACT

The need to draw climate-related inferences from historical data makes understanding the biases and errors in these data critical. While climate data are collected at point-level monitoring sites, they are often postprocessed by averaging sites within a geographic area to align the data to a grid, easing analysis and visualization. Although this aggregation generally provides reasonable estimates of the mean, its use can be problematic for characterizing the full distribution of climate measures. Specifically, the process of averaging point-level data up to grid level can lead to inconsistencies, particularly when the grid box is heterogeneous and extremes are of interest. Point-level data are measured at individual points, while gridded data are the averaged product of many measurements within a larger spatial area. Because of this aggregation, point-level and grid-level distributions differ in many fundamental properties, such as their shape, skew, and tail behavior. This paper highlights these differences and their effects on analyses pertaining to current climatological questions. Mathematical relationships are derived to link the distributions of grid-level climate measures to the distributions of point-level climate measures using the notion of effective sample size. Then, these relationships are leveraged to propose a correction factor to use when modeling higher moments and extreme events.

1. Introduction

As the scientific community's interest in climate-related phenomena grows, the need to thoroughly understand the underlying data is critical. Historical climate data, such as instrumental temperature and precipitation records, are quite noisy as a result of changing measurement methods and inconsistent spatial and temporal coverage. To compensate for this noise, various aggregation methods have been developed, typically involving averaging measurements taken at various stations within a particular geographical area. Such averaging produces more consistent estimates of the mean, but complicates our understanding of the true distributional properties of these quantities. Therefore, measuring, analyzing, and predicting various aspects of climate requires constantly shifting between two fundamentally different data types with inherently different properties.

Point-level data are measured at a specific geographic location, while gridded data locally aggregate measurements to create areal averages. Failure to distinguish between these two data types can significantly affect the scientific validity and real-world impact of an analysis. In this paper, we discuss some of the statistical properties of distributions of gridded and point-level data, develop quantitative relationships connecting distributions of these data types, and illustrate the utility of these relationships by using them to predict climate extremes.

Many major climate data repositories follow the convention of averaging measurements within spatial areas and reporting those averages. These datasets form a core pool of information from which many other scientists begin their research. On a global scale, NASA's Goddard Institute for Space Studies, the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administation (NOAA), and a partnership between the Met Office Hadley Centre and the University of East Anglia Climatic Research Unit (CRU) all produce gridded products of various climate measurements over time (Hansen et al. 2010;

Denotes Open Access content.

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-14-00571.s1.

Corresponding author address: Luke Bornn, Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138. E-mail: bornn@stat.harvard.edu

DOI: 10.1175/JCLI-D-14-00571.1

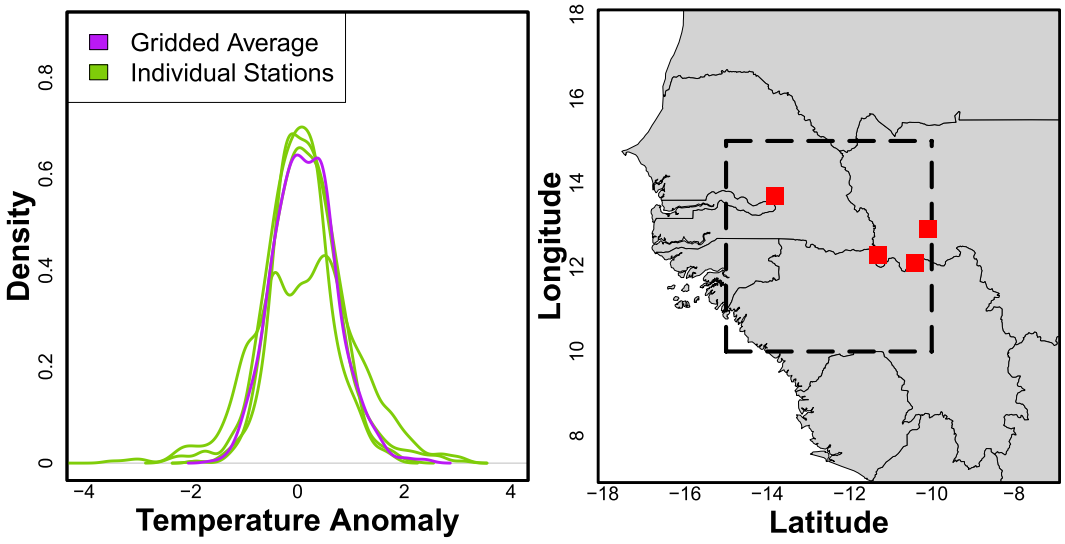

FIG. 1. (left) Kernel densities of the CRU monthly temperature anomalies from 1950 to 2010 for each station and for the gridbox average for the area spanning 10°–15°N, 10°–15°W where (right) the stations are mapped in red on the 5° × 5° square grid box. The bandwidth parameter for the kernel densities is set to 0.15. The difference in the individual stations' densities and the density of the gridded average illustrates how point-level data are generally more variable than gridded data and have heavier tails. This suggests that using different distributions to describe point-level and gridded distributions would improve accuracy.

Lawrimore et al. 2011; Peterson and Vose 1997; Jones et al. 2012). Similar datasets are produced for many regions, such as those maintained by the U.S. Historical Climatology Network (Menne et al. 2014). Climate models also only output estimates of climate measures at the level of areal regions. This means that many of the issues stemming from transferring information from the point level to the grid level that apply to historical data are also relevant for downscaling climate models (Klein Tank et al. 2009; Zwiers et al. 2013).

In developing gridded products, three main processing techniques are used, generally referred to as the anomaly method, the reference station method, and the first difference method (Peterson et al. 1998). The anomaly method requires selecting a base period of years, finding the average measurements for each station during that period, calculating the difference (anomaly) between the observed value and the base period value, and then averaging these differences (Jones et al. 1986). The reference station method extends the anomaly method to allow for cases where there are an insufficient number of measurements in the base period by allowing multiple stations to be combined to produce a full base period time series (Hansen and Lebedeff 1987). The first difference method eliminates the use of a base period entirely and instead gives the year-to-year differences in the climate measures of interest (Peterson et al. 1998). After processing the time series, the station records are aggregated over a spatial area using either a simple average or a weighting process that accounts for a station's proximity to other stations and/or to the center of the grid box. Further complicating the interpretation of these aggregates, the temporal values being averaged are of varying data types. For example, temperature measurements are usually reported as daily averages while precipitation measurements are usually reported as cumulative daily totals.

Despite the ubiquity of spatial averaging, its application can be inappropriate in some circumstances. In particular, some uses, such as predicting extreme events, are concerned with distributions at the level of individual points. Aggregation elucidates the properties of the distribution of the averages, but can fail to answer the pertinent scientific question in these cases: how to describe and predict the climate behavior of the region from which the spatial average was taken. This is essentially a case of the classical ecological fallacy in that conclusions about individual sites are incorrectly assumed to have the same properties as the average of a group of sites (Robinson 1950). Figure 1 illustrates this effect by plotting kernel densities of the average monthly temperature anomalies for a sample grid box alongside kernel densities for the monthly temperature anomaly observed at individual stations within the grid box. This highlights how the distributions for individual points tend to differ from the distributions of the average.

Failing to account for differences between individual measurements and averages can have significant effects.

For instance, changes in the frequency and intensity of extreme events are often a concern of scientists, since ostensibly they have more immediate impact than changes in averages. However, this effect is heavily dependent on how extreme events are defined. Extremes in the areal average are neither the same as extremes at individual points in a grid box nor are they the same as extremes defined by taking the most extreme value at any point within an area. All such quantities are of interest, but in different situations. For example, to assess the amount of precipitation in a catchment, areal averages should be used, but to assess extreme temperatures observed at one's home, point-level measurements are needed. The difference between these values is not trivial. For instance, in one study of the United States, the number of extremes for point-level rain gauge data and gridded reanalysis data were shown to differ by as much as 2–3 times (Mannshardt-Shamseldin et al. 2010). Further, there is interest in whether global warming is being driven solely by a change in the mean temperature or by a change in the mean temperature and a change in the variance of the temperature distribution (Schär et al. 2004; Rhines and Huybers 2013; Huntingford et al. 2013). This question depends on the entire distribution of temperature, not just the average, so an answer cannot be obtained without considering the effects of gridding on the entire distribution. Nevertheless, the analysis of data employing some form of spatial averaging is often used for analyzing climate extremes (Efthymiadis et al. 2011; Morak et al. 2013) and the variance in temperature distributions (Hansen et al. 2012; Huntingford et al. 2013; Schär et al. 2004). A central aim of this paper is to facilitate investigation into these questions, and many others like them, by clarifying the effects of spatial aggregation on the distribution of climate measures.

This paper fits into the larger class of work on change of support problems, which address inference for data observed and analyzed on different scales. These scaling problems were initially addressed within a range of field-specific contexts such as agriculture (Fairfield Smith 1938), sociology (Robinson 1950), and epidemiology (Morgenstern 1982). Tobler (1979) first introduced areal interpolation in the statistics literature when he proposed averaging intensity surfaces in such a way that volume is preserved. Flowerdew and Green (1989, 1994) suggested more informative covariates for interpolation and fitting via the expectation–maximization (EM) algorithm. Recent work has proposed combining samples of individual-level data with aggregate data to better overcome the ecological fallacy (Wakefield and Lyons 2010). Hierarchical Bayesian approaches have also been introduced (Mugglin and Carlin 1998; Banerjee et al. 2004). Thorough reviews of the change of support literature can be found in Gotway and Young (2002) and more recently Gelfand (2010).

In climate science, Osborn and Hulme (1997) first developed a quantitative relationship between the variance of the areal average and point-level data. Additional work has focused on quantifying the variance of some gridded products, particularly temperature measurements. For example, the current Hadley Centre/CRU temperature (HadCRUT) dataset provides uncertainty estimates for each grid box that are obtained by summing estimates of several individual components that have been identified as contributing to uncertainty (Brohan et al. 2006). Uncertainty attributed to incomplete spatial coverage has been explored by casting it as a sampling problem using the method postulated by Jones et al. (2001) (Brohan et al. 2006).

Our work adds to this literature by focusing on the downscaling of distributional properties. In particular, we provide a way to link the moments of grid-level distributions with the moments of point-level distributions when intrasite correlation is fixed and known or estimable. We also show how these relationships can be used to connect point-level and gridded distributions for the purpose of downscaling extremes. It is important to note that we only propose linking distributions and properties of distributions, not individual data values or time series. As has been discussed for precipitation, point-level data cannot be deterministically estimated from grid-level data, because point-level data exhibit stochastic noise that has been marginalized out of gridded data (Maraun 2013; Wong et al. 2014). Similarly, mapping a single point-level time series or data point directly to the grid level is not possible, since grid-level data need to account for all subgrid variability, information that is not fully available from a single station's time series or data point. Despite these limitations, our method provides a simple correction for linking distributions at different scales, which is valuable in situations where information is only available about a measure's distribution at the grid level or point level, but the distributions of behaviors at the other level is of greater interest.

2. Moment relationships

a. *Theoretical moments*

1) SIMPLIFIED I.I.D. MODEL

To develop an understanding of how point-level and grid-level measurements differ, we investigate how the first four statistical moments of station and grid-level distributions relate. We begin with a simplified model of gridding and assume that the measurements for each

station in any grid box are independent and identically distributed (i.i.d.). We will eventually relax this assumption to account for the stations' spatial correlation, but this assumption provides an intuitive way to understand the implications of not distinguishing between the distributions of station and point-level data.

The first four moments provide a fairly thorough characterization of a probability distribution. The first and second moments (mean and variance) are used regularly, while the third moment (skewness) and the fourth moment (kurtosis) are used less often. Skewness provides a measure of the symmetry of the distribution. A symmetric distribution, such as the normal distribution, is defined to have a skewness of zero. A distribution that has a longer left tail than right tail is referred to as skewed left and has a negative skew value whereas a distribution that has a longer right tail than left tail is referred to as skewed right and has a positive skew value. The fourth moment, kurtosis, is often referred to as a measure of "peakedness." For symmetric and unimodal distributions, a distribution with a high kurtosis has more mass in the tails and less mass in the center peak, while a distribution with lower kurtosis has more mass in the center peak and less mass in the tails. Various measures of these higher-order moments exist, but these differences have little bearing on their interpretation. See Joanes and Gill (1998) for further discussion.

Here we define skewness and kurtosis using cumulants. Letting X be a random variable, $\text{Skew}(X) = \kappa_3(X)/[\kappa_2(X)]^{3/2}$ and $\text{Kurt}(X) = \kappa_4(X)/[\kappa_2(X)]^2$. The ith cumulant, $\kappa_i(X)$, is the coefficient of the power series expansion of the cumulant generating function, $g_x(t)$, which is the logarithm of the moment generating function. Thus, $g_x(t) = \log\mathbb{E}(e^{tx}) = \sum_{i=1}^{\infty}\kappa_i t^i/i!$. Cumulants satisfy several useful properties including that $\kappa_i(cX) = c^i\kappa_i(X)$ and $\kappa_i(X + Y) = \kappa_i(X) + \kappa_i(Y)$, where X and Y are random variables and c is any constant (Hald 2000). This implies that $\sum_{j=1}^{n}\kappa_i(X_j) = n\kappa_i(X_1)$, where X_j, $j = 1, 2, \ldots, n$, is a random variable. Cumulants are also used to relate the quantiles of any distribution to the cumulative distribution function of a normal distribution via a series expansion originally developed by Thorvald Thiele in 1899 and later expanded on by Cornish and Fisher and by Hill and Davis (Hald 2000; Cornish and Fisher 1938; Hill and Davis 1968).

Using these moment definitions, properties of the distributions of point-level and gridded data can be expressed and related succinctly. Letting X_i represent i.i.d. measurement i in an arbitrary grid box, $\overline{X}$ represent the grid box average, and n represent the number of stations, or sample size, in this grid box, relationships can be derived relating each of the first four moments of X_i and each of the first four moments of $\overline{X}$ solely as a function of n. The relationships for the mean and variance of independent random variables and their averages are well known, specifically, $\mathrm{E}(\overline{X}) = \mathrm{E}(X_i)$ and $\mathrm{Var}(\overline{X}) = \mathrm{Var}(X_i)/n$. For distributions that have defined cumulants, the skewness and kurtosis relationships are less well studied, but can be derived as follows using the properties of cumulants:

TABLE 1. Mathematical definitions of the first four moments, where X_i represents a single observation and $\overline{X}$ represents the mean of a group of observations and the relationships between these individual and averaged values.

Moment	General	Cumulant	Relationship
Mean (μ)	$\mathrm{E}(X)$	κ_1	$\mathrm{E}(\overline{X}) = \mathrm{E}(X_i)$
Variance (σ^2)	$\mathrm{E}[(X-\mu)^2]$	κ_2	$\mathrm{Var}(\overline{X}) = \frac{1}{n}\mathrm{Var}(X_i)$
Skewness (γ_1)	$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$	$\frac{\kappa_3}{\kappa_2^{3/2}}$	$\mathrm{Skew}(\overline{X}) = \frac{1}{\sqrt{n}}\mathrm{Skew}(X_i)$
Kurtosis (γ_2)	$\frac{\mathrm{E}[(X-\mu)^4]}{\mathrm{E}[(X-\mu)^2]^2}$	$\frac{\kappa_4}{\kappa_2^2}$	$\mathrm{Kurt}(\overline{X}) = \frac{1}{n}\mathrm{Kurt}(X_i)$

$$\mathrm{Skew}(\overline{X}) = \frac{\kappa_3\left(\frac{1}{n}\sum X_i\right)}{\left[\kappa_2\left(\frac{1}{n}\sum X_i\right)\right]^{3/2}} = \frac{\left(\frac{1}{n}\right)^3 n\kappa_3(X_i)}{\left[\left(\frac{1}{n}\right)^2 n\kappa_2(X_i)\right]^{3/2}}$$
$$= \frac{1}{\sqrt{n}}\frac{\kappa_3(X_i)}{\kappa_2(X_i)^{3/2}} = \frac{1}{\sqrt{n}}\mathrm{Skew}(X_i) \quad \text{and} \qquad (1)$$

$$\mathrm{Kurt}(\overline{X}) = \frac{\kappa_4\left(\frac{1}{n}\sum X_i\right)}{\left[\kappa_2\left(\frac{1}{n}\sum X_i\right)\right]^2} = \frac{\left(\frac{1}{n}\right)^4 n\kappa_4(X_i)}{\left[\left(\frac{1}{n}\right)^2 n\kappa_2(X_i)\right]^2}$$
$$= \frac{1}{n}\frac{\kappa_4(X_i)}{\kappa_2(X_i)^2} = \frac{1}{n}\mathrm{Kurt}(X_i). \qquad (2)$$

Table 1 summarizes these relationships. These equations illustrate that estimating point-level moments directly from averaged data results in estimates of variance and kurtosis that are biased by a factor of $1/n$ and an estimate of skewness that is biased by a factor of $1/\sqrt{n}$. In other words, although the mean of both distributions is the same, many of their other moments differ. Because of this mischaracterization of the variance, skewness, and kurtosis, when using grid-level data in place of point-level data, predictions of what proportion of the data are above or below a particular extreme threshold will be incorrect. Similarly, using gridded data to assess whether the variance of temperature distributions is changing over time may lead to flawed conclusions, because changes in the number

of stations contributing data will cause changes in the reported gridded variance regardless of the true variance's behavior. Analogous problems exist for skewness, kurtosis, and tail behavior.

Fortunately, these differences in higher-order moments can be accounted for under certain assumptions. In fact, the relationships themselves suggest a solution. Under the assumption of i.i.d. measurements, all of the point-level moments can be directly represented as a function of the gridded moments and the number of stations within the grid box, n. So, it appears we can easily convert between these two sets of moments as long as n is known. To use a gridded moment to predict a point-level moment, it seems we need only multiply the gridded variance and kurtosis values by a factor of n and multiply the skewness value by a factor of $\sqrt{n}$. Likewise, to use a point-level moment to predict a gridded moment, it appears we need only divide the point-level moment by n for variance and kurtosis and by $\sqrt{n}$ for skewness.

2) Correlation and effective sample size

However, in drawing this conclusion, we are assuming that the point-level measurements are i.i.d., which is generally not the case. Stations within a grid box tend to be correlated, since weather events, such as heavy rainfall or lower temperatures, affect entire regions concurrently. Because of this spatial correlation, the n samples within a grid box contain less information than would be contained in n truly independent samples.

To address this issue, we instead use effective sample size n_{eff} to replace n in our calculations. Effective sample size is a common statistical measure that estimates how many i.i.d. samples a correlated sample represents. It was designed to match the standard error of the mean, so is only a reasonable, not a perfect, analog for the true amount of information contained in measurements for higher moments. However, as our later results demonstrate, it performs adequately for all moments in this context. Further methodological research could investigate effective sample size measures optimized for higher moments.

To estimate the effective sample size in a grid box, we must assume that the intrasite correlations are known or estimable. We also typically assume that the spatial correlations are constant over time, although this assumption can be relaxed. As an example, if it is known when the intrasite correlations changed, such as in the case of the recorded movement of a monitoring site, the effective sample size and corresponding adjustments can be calculated separately before and after the change, leading to different adjusted distributions for each time period. Alternatively, if the correlations are suspected to be frequently changing, particularly in an unspecified way, a sampling method can be used to identify the distribution of the correlations and corresponding adjustments over time. Applying each of these adjustments and combining the results will give a distribution on the parameters that reflects changes in spatial correlation over time. An example of a reasonable sampling method, the moving block bootstrap, will be discussed in the next section.

Given these assumptions, the following formula for spatially correlated data is appropriate where x_i is defined to be the measurement of an individual station in the grid box of interest at time i (Fortin and Dale 2005):

$$n_{\text{eff}} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cor}(x_i, x_j)}. \tag{3}$$

In our analysis, we use the empirical pairwise correlation from the point-level data to estimate each intrasite correlation. Although access to the full point-level data would be unusual in many applications, we employ this estimation technique because it requires minimal assumptions and is broadly applicable. This ensures that the moment relationships we establish in this paper are not in some way confounded or dependent on a context-specific estimation technique. In further work, the method of estimating the correlation should be determined based on the data available and the particular measurement of interest. In many cases, the empirical correlation from historical or neighboring point-level data can be used to impute the correlation of interest. For some measures, research has also been conducted to understand and quantify intrasite correlation (Hansen and Lebedeff 1987; Jones et al. 1986). More broadly, it is known that correlation among sites is affected by such factors as a grid box's location (Haylock et al. 2008), orientation and weather patterns (Hansen and Lebedeff 1987), seasonality (Osborn and Hulme 1997), site density and homogeneity, and the type of climate measure being assessed. Such information can be used to build reasonable models for the correlations when surrogate point-level data are not directly available. Ideal models will vary depending on the measure of interest and what covariate data are available, but will benefit from being robust to outliers and potential deviations from model assumptions. Since climate data tend to be noisy, methods that are too sensitive to violations of assumptions or outliers may result in inaccurate correlation estimates, which will subsequently bias estimates of the effective sample size and the corresponding adjusted moments.

3) Uncertainty quantification

Regardless of how correlation is calculated, it will be an estimated measure. This means that uncertainty in

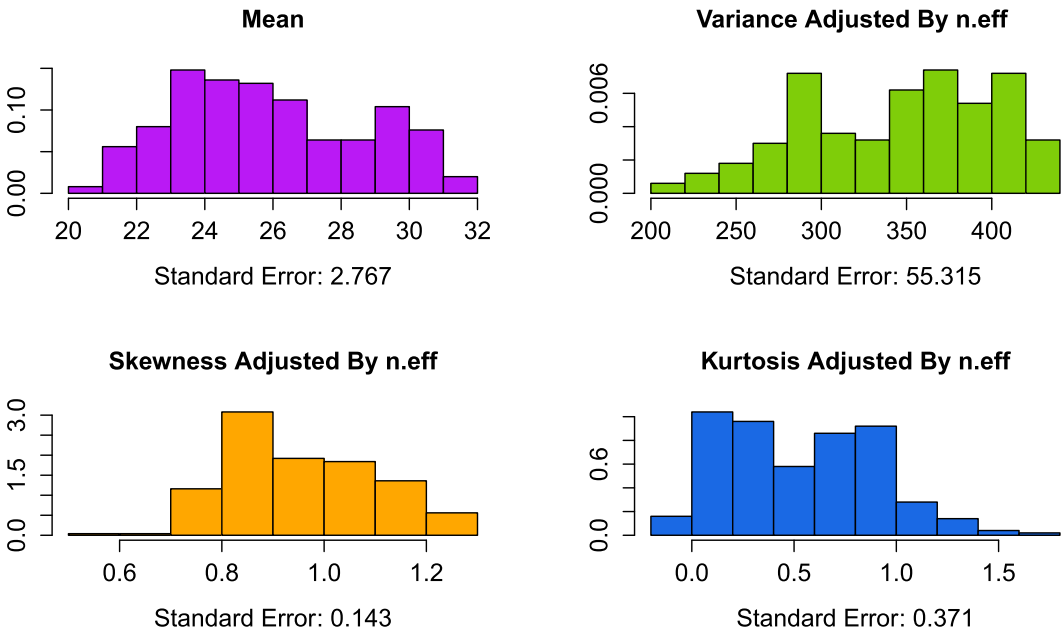

FIG. 2. Approximate sampling distributions for the point-level mean, variance, skewness, and kurtosis for monthly precipitation in the grid box spanning 30°–35°N, 105°–110°W obtained using the Global Historical Climate Network precipitation dataset from 1950 to 2010 for stations with >10% missing values omitted. Distributions were approximated via moving block bootstrap sampling with 250 ninety-six-month periods sampled with replacement. For each sampled time period, the effective sample size was calculated as described in section 2b and the first four grid-level sample moments were estimated directly from the data. Then, the grid-level variance and kurtosis were multiplied by n_{eff} and the grid-level skewness was multiplied by $\sqrt{n_{\text{eff}}}$ to obtain point-level skewness for the sampled period. The grid-level mean was left unchanged, since point-level and grid-level means are the same. Combining the results for all sampled time periods, we obtain a point-level sampling distribution from which we can estimate the uncertainty in our parameter estimates. Here, we report the standard error estimate for each parameter below its sampling distribution. Uncertainty for the mean stems only from sampling variability, while the uncertainty in the other moments is a function of both sampling variability and uncertainty in the effective sample size estimate.

the correlation estimates will result in uncertainty in the corresponding estimates of the effective sample sizes and adjusted parameter estimates. Similarly, since the locations and times where measurements were taken are samples, there is also uncertainty in the moments calculated directly at the observed scale. One way to quantify this uncertainty is to apply the well-known bootstrap method. In bootstrapping, samples are taken from the observed data with replacement to approximate the true sampling distribution. Here, because the measurements are correlated over time, we apply the moving block bootstrap, where random periods of time, or blocks, are sampled. Specifically, we set the number of blocks to N and the block length to b and let X_i, $i = 1, \ldots, T$, represent each measurement in the time series. Then N values are selected with equal probability from the time series as starting values for blocks spanning, $X_i, \ldots, X_{i+b-1}$. This procedure gives N potentially overlapping bootstrap times series. Subsequently, for each bootstrap time series the effective sample size, and sample moments are calculated independently. Combining the results for all N blocks gives an approximate sampling distribution for each measure, from which estimates of uncertainty such as the standard error can be directly calculated (see, e.g., Kunsch 1989; Efron and Tibshirani 1994). In Fig. 2, we illustrate this process by finding and plotting the sampling distribution for the moments of a sample grid box and the corresponding estimated standard errors.

b. *Empirical moments*

1) DATA

The previous discussion has been largely theoretical, leaving open the possibility that these relationships could be overwhelmed by the considerable noise in real

climate data. Therefore, we illustrate these relationships with two well-known climate datasets. Throughout the remainder of the paper, we consider the Climatic Research Unit (CRU) temperature anomaly dataset version 4 from 1950 to 2010 and the Global Historical Climate Network (GHCN) total precipitation dataset version 2 from 1950 to 2010 (Jones et al. 2012; Peterson et al. 1998). Replicate analysis on GHCN temperature anomaly data is included online as supplementary material.

The CRU dataset is the land component of the HadCRUT dataset produced by the Climatic Research Unit at the University of East Anglia in conjunction with the Hadley Centre. Temperatures are expressed as monthly anomalies from a base period of 1961–90 and each station's time series is reported along with a mean for each 5° × 5° grid box. Both the point-level and gridded data undergo extensive processing and quality checks before their publication; see Jones et al. (2012) for additional details. To ensure a relatively constant sample size over the whole time period, stations that had more than 10% missing values from 1950 to 2010 were omitted. Because of this stipulation, for a small number of grid boxes, the mean of all stations maintained does not exactly equal the grid box mean reported by CRU. Therefore, we limit our analysis to grid boxes for which the calculated mean differs by less than 0.05 from the grid box values reported by CRU, resulting in the omission of 34 of the 185 otherwise eligible grid boxes. Although perfect equality would be desirable, this is outweighed by the value of examining a publically available gridded product and the corresponding station data.

The GHCN total precipitation dataset is produced by NOAA's National Climatic Data Center. It gives the total monthly precipitation measured at stations in the Northern Hemisphere. The data were processed for quality control and identification of duplicate records. Additional information on the methods used is available on the National Climatic Data Center website (Peterson et al. 1998). Since there is no corresponding gridded dataset, we have calculated the mean of all stations in each 5° × 5° grid box to produce a spatially gridded product. We have again limited the analysis to the years 1950–2010 and have omitted any station that is missing more than 10% of measurements during this time period.

2) Observed relationships

By considering the first four sample moments for gridded and point-level data, the validity of the theoretical moment relationships can be verified empirically. Toward that end, we use the monthly time series of each grid box in the CRU data to calculate a sample mean, sample variance, sample skewness, and sample kurtosis for each grid box. Similarly, for each grid box in the GHCN precipitation dataset, a series of monthly means is obtained by averaging the measurements of each station in each grid box for all months. The sample mean, sample variance, sample skewness, and sample kurtosis are then calculated from this time series. We also find the same four sample moments for each individual station's series of monthly means. In Fig. 3, each grid box's moments are plotted against their corresponding station's moments to illustrate how the moments differ for both datasets. In Fig. 4, each of the moments for the gridded data are plotted against the corresponding mean of the stations' moments where the point-level variance and kurtosis have been multiplied by a factor of $1/n_{\text{eff}}$ and the skewness has been multiplied by a factor of $1/\sqrt{n_{\text{eff}}}$. To confirm the relationships in Table 1, these points should form a line through $y = x$, indicating their equality. For comparison, we also plot these same relationships using the true sample size rather than the effective sample size.

For both the CRU and the GHCN data, the observed results largely confirm our theoretical analysis. In Fig. 4, the points relating grid-level and point-level means for both the CRU and GHCN data form a straight line, reflecting their equivalence. For the CRU data, the points on the variance graph are generally below the line $y = x$, while the skewness and kurtosis points go roughly through the line $y = x$. On the other hand, for the GHCN data, the points for variance, skewness, and kurtosis are all below the line $y = x$. This suggests that the higher-order moments of non-normally distributed data are particularly affected by gridding.

Further, Fig. 4 shows that these differences are well captured by the relationships in Table 1. For both datasets, the grid-level and point-level data adjusted by the effective sample size form a line through $y = x$, suggesting equality. The gridded and point-level means form a perfectly straight line for the GHCN temperature data whereas for the CRU data the line is only roughly straight. This distinction is attributable to the difference in gridding method applied to the two datasets. For both the GHCN and CRU data, the point-level variance adjusted by effective sample size plotted against the point-level variance forms a line through $y = x$ while the point-level variance adjusted by the true sample size yields a set of points consistently above $y = x$. This demonstrates that the use of n as an adjustment factor overcorrects for the effects of aggregation—an issue largely corrected by the use of effective sample size. For the remaining moments, the points relating the point-level moments adjusted by the sample size and the grid box moments are scattered above the line $y = x$, while

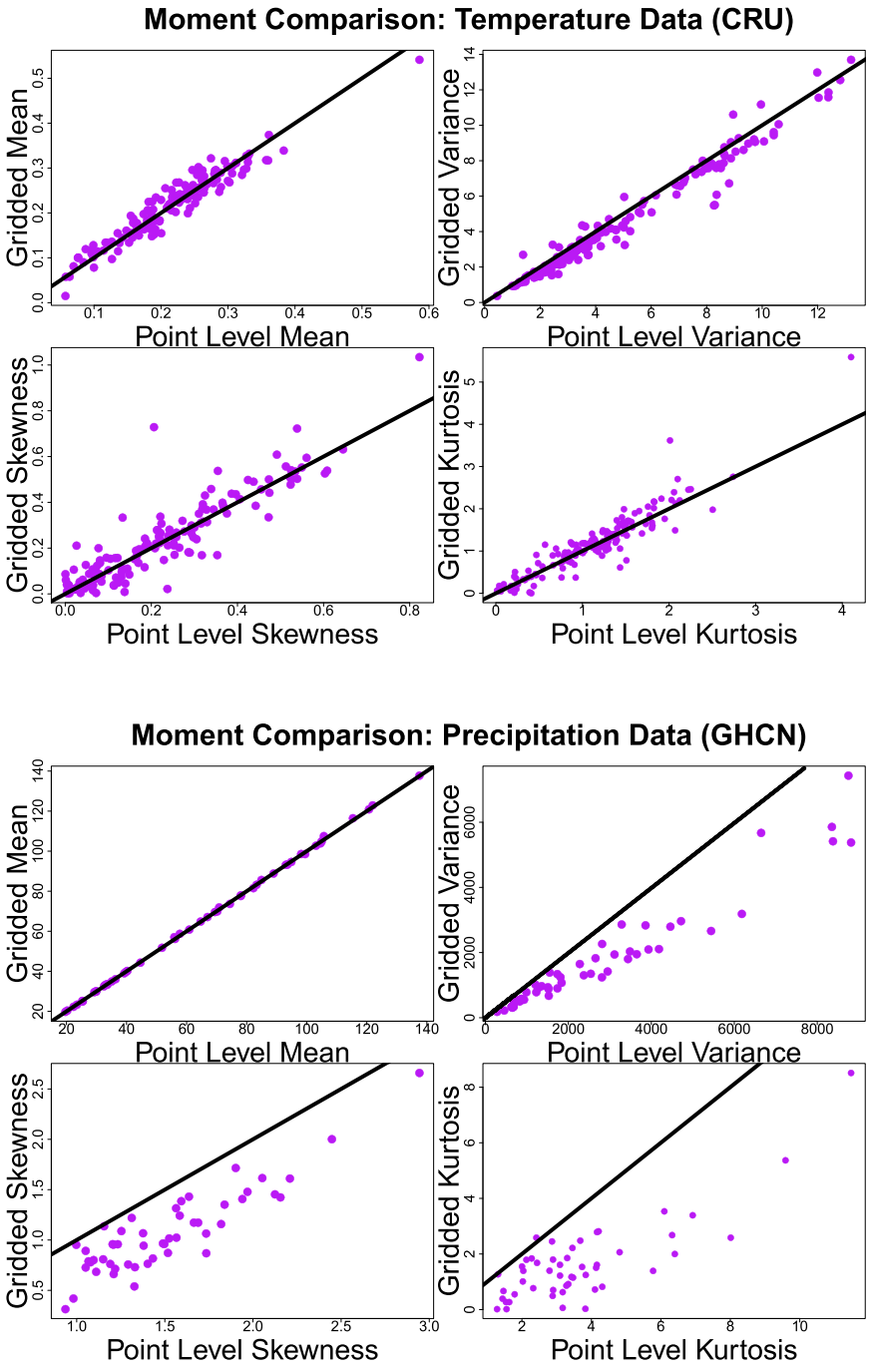

FIG. 3. Plots of the sample moments of the gridded data vs the sample moments of the point-level data for the (top) CRU temperature data and (bottom) GHCN precipitation data. Each point represents one 5° × 5° grid box and the black line is $y = x$. The points for the CRU temperature variance relationships and for the GHCN precipitation variance, skewness, and kurtosis relationships are systematically below the line $y = x$, which suggests that the gridded values for these moments differ from the point-level moments.

the points relating the point-level moments adjusted by the effective sample size and the grid box moments form a straight line, again supporting the use of Eqs. (1) and (2). There is somewhat more precision for variance than for the skewness and kurtosis. In summary, the theoretical relationships presented in Table 1, when accompanied by the notion of effective sample size, are reasonable mathematical representations of

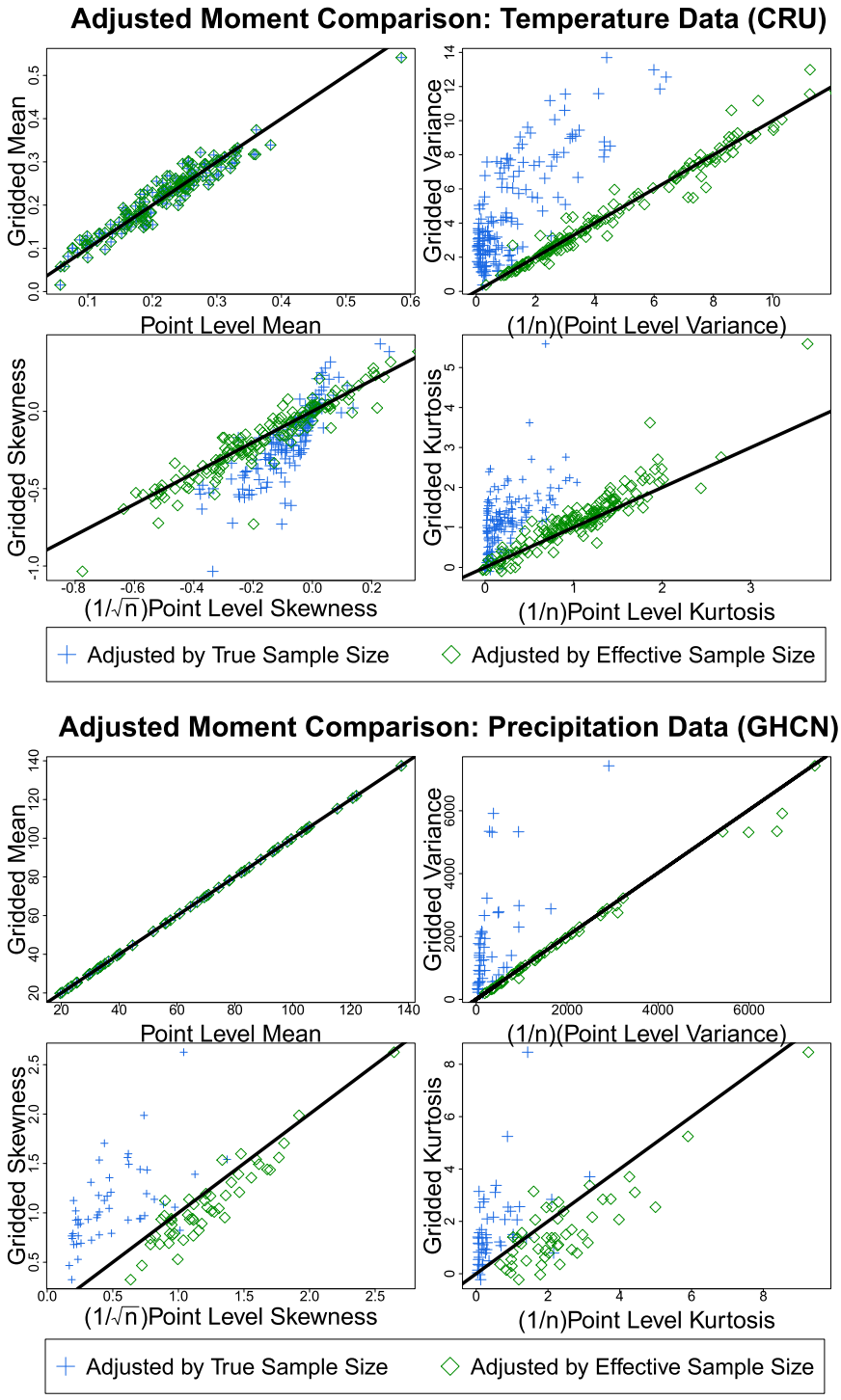

FIG. 4. As in Fig. 3, but for the point-level data adjusted by factors of the sample size n in blue and the effective sample size n_{eff} in green. Each point represents one 5° × 5° grid box and the black line is $y = x$. Each point-level moment adjusted by the effective sample size plotted against its corresponding gridded moment forms a roughly straight line, which supports the relationships in Table 1.

the observed relationships between point-level and gridded moments.

3. Extremes

a. Methods

Having established relationships between gridded and point-level moments, we now demonstrate their practical importance by examining their effects on a problem currently faced by climate scientists: how to predict extreme events. Extremes remain less understood than averages. This results primarily from the greater challenge presented in understanding extremes than in understanding averages, since extremes, by definition, rarely occur. Thus, extreme value theory often involves augmenting the limited data on rare extreme events with other information. A common approach is to estimate the distribution of extremes by treating the individual maxima from small increments of time as random samples from the true distribution of extremes. See Coles et al. (2001) for further details on classic extreme value theory.

Here, we take a similar approach and avoid looking at the extreme events alone, but instead focus on characterizing the entire distribution of the data, subsequently using that characterization to make a conservative estimate of the occurrence of extreme events at the point level. Single station time series are likely to have considerable noise and might be inaccurate, but studying extremes of the mathematical average will not accurately explain point-level behavior. To balance these concerns, we leverage the relationships between gridded and point-level moments established in Table 1. Empirical moments of the gridded data are used as initial estimates for the moments of the point-level distribution but are then multiplied by a factor of the effective sample size as given in Table 1. This gives point-level estimates that are neither biased nor adversely affected by the noise in individual station records. These point-level moments then can be used to define the point-level distribution. As with finding method of moments estimators, one only needs to find as many moments as there are parameters in the distribution, after which one may solve for the distributional parameters.

To demonstrate our method's effectiveness, we predict the percent of extremes that would be observed in each grid box. Then, we compare this prediction with the percent of extremes actually observed at the point-level. For comparison, we consider three adjustments of the variance:

1) Unadjusted: The grid-level estimates of mean and variance are used directly for point-level data.
2) Adjusted by Var/n: The grid-level variance is divided by n in order to estimate the point-level variance and the grid-level mean is used directly to estimate the point-level mean.
3) Adjusted by Var/n_{eff}: The grid-level variance is divided by n_{eff} [Eq. (3)] in order to estimate the point-level variance and the grid-level mean is used directly to estimate the point-level mean.

Once these corrections have been applied, the resulting adjusted moments are used to calculate the parameters of the distribution. For the temperature data, we elected to use a simple Gaussian model for the distribution of the temperature anomalies at both the station and grid box level. Similarly, we selected a gamma distribution as the parametric model for the precipitation data at both levels. Although more accurate distributional models could certainly be obtained, our use of simple models makes it easier to isolate the improvement in prediction directly attributable to the improvement in the moment estimates. Thus, our results give a lower bound for the method's effectiveness. Practitioners in further applications could exploit prior knowledge about the particular variable of interest or could employ more extensive model evaluation techniques to obtain greater accuracy. Since the underlying distribution for temperature anomalies may have heavier tails than can be modeled with a normal distribution, our method provides a conservative estimate of the expected number of extremes.

For each of these three methods, we calculate the percent of the fitted distribution lying above or below several thresholds. For demonstrative purposes, we have arbitrarily selected the top and bottom 2.5% and 1% of the anomaly data for all stations worldwide from 1950 to 2010 as the thresholds of interest for the CRU data and have selected the top 20%, 10%, 5%, and 2.5% of all data for the GHCN dataset. Any other thresholds of scientific interest could just as easily have been selected. We then report the difference in predicted versus observed percent of extremes for each threshold and adjustment. Results for temperature are displayed in Table 2 and results for precipitation are displayed in Table 3.

b. Results

As the values in Tables 2 and 3 indicate, for both precipitation and temperature the difference between the observed and predicted extremes is generally smallest when the moments are adjusted using effective sample size. However, the difference in the methods' performance is greater for precipitation than for temperature. This can likely be attributed to the different underlying distributions of temperature and precipitation. For data approximately following a Gaussian

TABLE 2. The average over all grid boxes of the observed percent of CRU temperature data measurements above or below various thresholds minus the percent of measurements predicted for each grid box to be above or below these thresholds. The predictions are obtained using a Gaussian distribution as described in section 3, where the mean of the distribution is estimated directly from the grid-level data. For the unadjusted prediction the grid-level estimate of variance is used directly for point-level data; for the adjusted by Var/n prediction, the grid-level variance is divided by the sample size n in order to estimate the point-level variance; and for the adjusted by Var/n_{eff} prediction, the grid-level variance is divided by the effective sample size n_{eff} in order to estimate the point-level variance.

Variance adjustment	Thresholds			
	Lowest 2.5%	Lowest 5%	Highest 2.5%	Highest 5%
Unadjusted	0.60	0.33	0.27	0.16
Adjusted by Var/n	−13.42	17.09	−14.63	−17.62
Adjusted by Var/n_{eff}	0.47	−0.01	0.10	−0.21

distribution, not adjusting the moments to the point level will only result in an inaccurate variance estimate, but for data approximately following a gamma distribution inaccuracies in the higher moments will change the shape of the distribution entirely.

In Fig. 5 for temperature and Fig. 6 for precipitation, we display a matrix of plots. Each subfigure is a series of boxplots arranged in order of increasing sample size showing the difference between the percent observed and percent predicted extremes. In general, the boxplots are centered above zero for the unadjusted moments, below zero for the moments adjusted by the true sample size, and around zero for the moments adjusted by the effective sample size. This reinforces the improvement in accuracy obtained by downscaling the moments using factors of the effective sample size. Additionally, displaying the boxplots sorted by sample size illustrates how the improvement in accuracy is not uniform across all grid boxes. Plotted in red on each set of boxplots is a line regressing the mean difference between observed and predicted extremes on the sample size n. These lines highlight that as n increases, the biasing effects of gridding increase, so that for larger sample sizes the underprediction of extremes is more significant than for smaller sample sizes. As such, the effect of gridding is noticeably affected by the number of stations used in calculating the grid-level averages. If the true sample size is used to adjust the variance, the overprediction of extremes generally increases as the sample size increases. On the other hand, using the effective sample size to adjust the variance results in the difference between the predicted and observed percent of extremes being near zero for all sample sizes. Overall, downscaling moments by their effective sample size provides an accurate way to predict extremes across grid boxes of all sample sizes when the intrasite correlations are estimable and constant. In addition, these regression lines reinforce the theoretical relationships in Table 1 by highlighting that the key factor in relating moments of grid-level distributions to moments of point-level distributions is effective sample size. We emphasize that these relationships are applicable only for connecting distributions as a whole, and should not be applied in an effort to recreate individual time series.

TABLE 3. As in Table 2, but for GHCN precipitation data measurements. The predictions are obtained using a gamma distribution as described in section 3.

Variance adjustment	Thresholds			
	Highest 20%	Highest 10%	Highest 5%	Highest 2.5%
Unadjusted	1.48	2.00	1.62	1.11
Adjusted by Var/n	3.07	−1.96	−3.83	−4.22
Adjusted by Var/n_{eff}	0.38	0.16	0.17	0.21

4. Discussion

In this paper, we have demonstrated the importance of understanding the effects of gridding when working with spatially referenced climate data. As documented in Table 1, all moments except the mean differ for point-level and gridded distributions. To draw conclusions in many circumstances, it must be recognized that averaging fundamentally changes a measurement's distribution. The distribution of means is distinct from the distribution of the individual data points that formed those means. So, for all nontrivial analysis, one must account for these differences.

In particular, this work highlights how using gridded distributions to predict climate extremes is not a reliable method for estimating point-level effects. The extent of this inaccuracy is determined by the sample size from which the grid box estimate was calculated, as well as the heterogeneity within the grid box. Analysis of extremes or higher-order moments from point-level distributions will be inaccurate unless one is eliciting trends by comparing only point-level distributions to other point-level distributions or gridded distributions to other gridded distributions. Even then, there must be a roughly consistent number of stations during all periods of comparison. Otherwise, even observed trends in higher-order moments or extremes will be indistinguishable from changes in the number of stations. Given the large amount of missing data in most climatic datasets, this will rarely be the case. Our method of downscaling extremes by downscaling distributional moments helps to alleviate

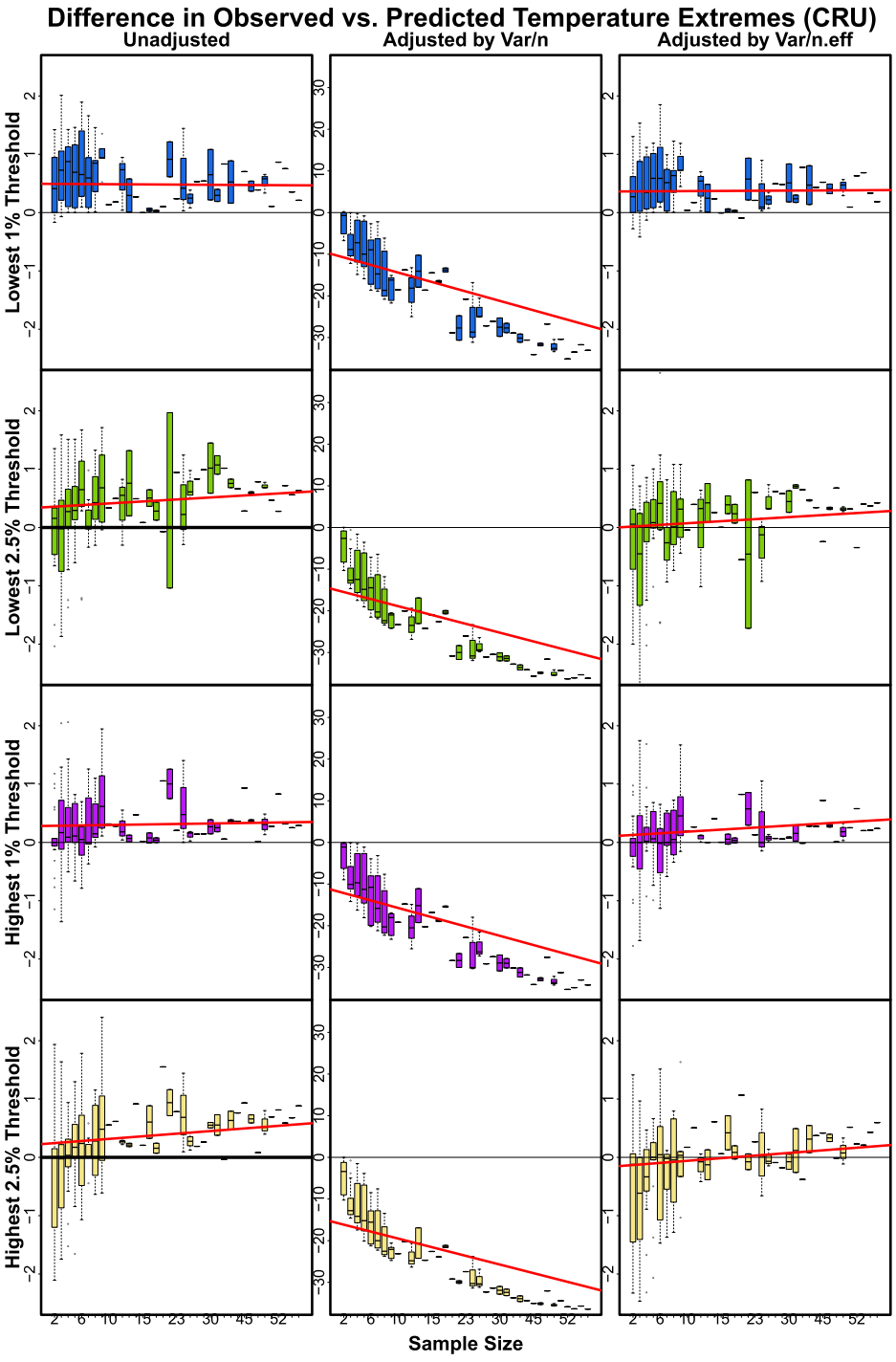

FIG. 5. Comparison of the difference in percent of predicted extremes vs percent of observed extremes for each grid box sorted by sample size for the CRU temperature data: (top)–(bottom) lowest 1% and 2.5%, and highest 1% and 2.5% thresholds. The predictions are obtained using a Gaussian distribution as described in section 3, where the variance (left) has not been adjusted, (center) has been adjusted by the sample size n, and (right) has been adjusted by the effective sample size n_{eff}. Plotted in red on each set of boxplots is the regression line between the mean difference between observed and predicted extremes and the sample size n. It shows that as sample size increases, the biasing effects of gridding increase; so that for larger sample sizes, the underprediction of extremes is more significant than for smaller sample sizes (left). Adjusting the variance by the true sample size overcorrects this result and leads to an overprediction of extremes (center). Adjusting the variance by the effective sample size gives the most accurate prediction (right).

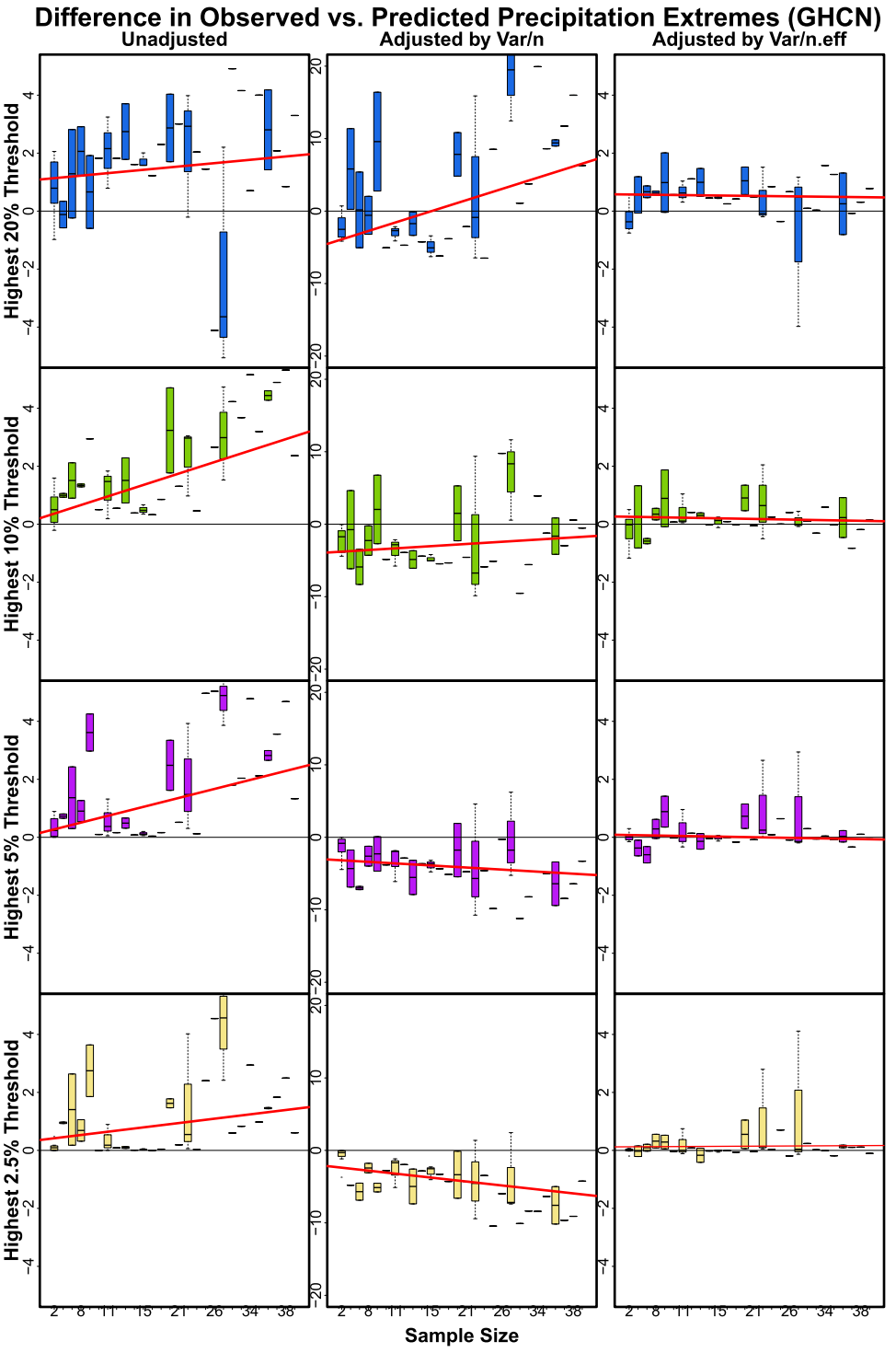

FIG. 6. As in Fig. 5, but for the GHCN precipitation data with (top)–(bottom): highest 20%, 10%, 5%, and 2.5% thresholds. These results largely mirror those observed in the CRU data except that the improvement in performance obtained by adjusting the variance by the effective sample size is greater for the GHCN precipitation data than for the CRU temperature data and the trend for analysis using true sample size is less clear.

this concern by enabling better identification of point-level variance and better predictions of extremes. Since individual time series and data values cannot be deterministically related from the point level to the grid level and vice versa, we achieve this result by relating the distributions of climate measures at these two scales. Further, this approach is completely distribution-free, making no assumptions on the distributional characteristics of the underlying data.

Another key conclusion from this work is the importance of sample size and correlation among station measurements. For future gridded products, retaining and publicizing information about the original sample size and correlation among samples would be valuable for accurate prediction. These results also point to another way the uncertainty in the output of climate models could be interpreted. Because of computational constraints, these models typically only give a single estimate for each grid box. To better understand the uncertainty in these gridded predictions, a correspondence could be developed between the information contained in climate model output and the equivalent number of measuring stations that would give the same amount of information. Doing so would make connecting what is predicted to happen in the average and what is predicted to happen at the point level much easier.

Overall, this work has extended knowledge of the statistical properties of gridded and point-level data by relating the moments of each distribution type, and has utilized these relationships to better predict climate extremes from gridded data. Cognizance of these relationships should enable better use of instrumental climate data, ultimately enabling better understanding of past and future climate trends.

Acknowledgments. The authors thank the editor and several anonymous reviewers for their valuable suggestions.

REFERENCES

Banerjee, S., A. E. Gelfand, and B. P. Carlin, 2004: *Hierarchical Modeling and Analysis for Spatial Data.* CRC Press, 472 pp.

Brohan, P., J. Kennedy, I. Harris, S. Tett, and P. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Coles, S., J. Bawa, L. Trenner, and P. Dorazio, 2001: *An Introduction to Statistical Modeling of Extreme Values.* Springer, 208 pp.

Cornish, E., and R. Fisher, 1938: Moments and cumulants in the specification of distributions. *Rev. Inst. Int. Stat.,* **5,** 307–320, doi:10.2307/1400905.

Efron, B., and R. Tibshirani, 1994: *An Introduction to the Bootstrap.* CRC Press, 456 pp.

Efthymiadis, D., C. Goodess, and P. Jones, 2011: Trends in Mediterranean gridded temperature extremes and large-scale circulation influences. *Nat. Hazards Earth Syst. Sci.,* **11,** 2199–2214, doi:10.5194/nhess-11-2199-2011.

Fairfield Smith, H., 1938: An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.,* **28,** 1–23, doi:10.1017/S0021859600050516.

Flowerdew, R., and M. Green, 1989: Statistical methods for inference between incompatible zonal systems. *Accuracy of Spatial Databases,* M. F. Goodchild, and S. Gopal, Eds., CRC Press, 239–247.

——, and ——, 1994: Areal interpolation and types of data. *Spatial Analysis and GIS,* S. Fotheringham, and P. Rogerson, Eds., Taylor & Francis, 121–145.

Fortin, M., and M. Dale, 2005: *Spatial Analysis: A Guide for Ecologists.* Cambridge University Press, 365 pp.

Gelfand, A., 2010: Misaligned spatial data: The change of support problem. *Handbook of Spatial Statistics,* A. Gelfand et al., Eds., CRC Press, 517–540.

Gotway, C., and L. Young, 2002: Combining incompatible spatial data. *J. Amer. Stat. Assoc.,* **97,** 632–648, doi:10.1198/016214502760047140.

Hald, A., 2000: The early history of the cumulants and the Gram–Charlier series. *Int. Stat. Rev.,* **68,** 137–153, doi:10.1111/j.1751-5823.2000.tb00318.x.

Hansen, J., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. Geophys. Res.,* **92,** 13 345–13 372, doi:10.1029/JD092iD11p13345.

——, R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.,* **48,** RG4004, doi:10.1029/2010RG000345.

——, M. Sato, and R. Ruedy, 2012: Perception of climate change. *Proc. Natl. Acad. Sci. USA,* **109,** E2415–E2423, doi:10.1073/pnas.1205276109.

Haylock, M., N. Hofstra, A. Klein Tank, E. Klok, P. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.,* **113,** D20119, doi:10.1029/2008JD010201.

Hill, G., and A. Davis, 1968: Generalized asymptotic expansions of Cornish–Fisher type. *Ann. Math. Stat.,* **39,** 1264–1273, doi:10.1214/aoms/1177698251.

Huntingford, C., P. Jones, V. Livina, T. Lenton, and P. Cox, 2013: No increase in global temperature variability despite changing regional patterns. *Nature,* **500,** 327–330, doi:10.1038/nature12310.

Joanes, D., and C. Gill, 1998: Comparing measures of sample skewness and kurtosis. *J. Roy. Stat. Soc.,* **47D,** 183–189, doi:10.1111/1467-9884.00122.

Jones, P., S. Raper, R. Bradley, H. Diaz, P. Kellyo, and T. Wigley, 1986: Northern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.,* **25,** 161–179, doi:10.1175/1520-0450(1986)025<0161:NHSATV>2.0.CO;2.

——, T. Osborn, K. Briffa, C. Folland, E. Horton, L. Alexander, D. Parker, and N. Rayner, 2001: Adjusting for sampling density in grid box land and ocean surface temperature time series. *J. Geophys. Res.,* **106,** 3371–3380, doi:10.1029/2000JD900564.

——, D. Lister, T. Osborn, C. Harpham, M. Salmon, and C. 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.

Klein Tank, A., F. Zwiers, and X. Zhang, 2009: Guidelines on analysis of extremes in a changing climate in support of

informed decisions for adaptation. WCDMP Rep. 72, WMO, 52 pp.

Kunsch, H., 1989: The jackknife and the bootstrap for general stationary observations. *Ann. Stat.,* **17,** 1217–1241, doi:10.1214/aos/1176347265.

Lawrimore, J., M. Menne, B. Gleason, C. Williams, D. Wuertz, R. Vose, and J. Rennie, 2011: An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.,* **116,** D19121, doi:10.1029/2011JD016187.

Mannshardt-Shamseldin, E., R. Smith, S. Sain, L. Mearns, and D. Cooley, 2010: Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *Ann. Appl. Stat.,* **4,** 484–502, doi:10.1214/09-AOAS287.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate,* **26,** 2137–2143, doi:10.1175/JCLI-D-12-00821.1.

Menne, M., C. N. Williams Jr., and R. Vose, cited 2014: Long-term daily and monthly climate records from stations across the contiguous United States. [Available online at http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html.]

Morak, S., G. Hegerl, and N. Christidis, 2013: Detectable changes in the frequency of temperature extremes. *J. Climate,* **26,** 1561–1574, doi:10.1175/JCLI-D-11-00678.1.

Morgenstern, H., 1982: Uses of ecologic analysis in epidemiologic research. *Amer. J. Public Health,* **72,** 1336–1344.

Mugglin, A. S., and B. P. Carlin, 1998: Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *J. Agric. Biol. Environ. Stat.,* **3,** 111–130, doi:10.2307/1400646.

Osborn, T., and M. Hulme, 1997: Development of a relationship between station and grid-box rainday frequencies for climate model evaluation. *J. Climate,* **10,** 1885–1908, doi:10.1175/1520-0442(1997)010<1885:DOARBS>2.0.CO;2.

Peterson, T., and R. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. *Bull. Amer. Meteor. Soc.,* **78,** 2837–2849, doi:10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

——, T. Karl, P. Jamason, R. Knight, and D. Easterling, 1998: First difference method: Maximizing station density for the calculation of long-term global temperature change. *J. Geophys. Res.,* **103,** 25 967–25 974, doi:10.1029/98JD01168.

Rhines, A., and P. Huybers, 2013: Frequent summer temperature extremes reflect changes in the mean, not the variance. *Proc. Natl. Acad. Sci. USA,* **110,** E546, doi:10.1073/pnas.1218748110.

Robinson, W., 1950: Ecological correlations and the behavior of individuals. *Amer. Sociol. Rev.,* **15,** 351–357, doi:10.2307/2087176.

Schär, C., P. Vidale, D. Lüthi, C. Frei, C. Häberli, M. Liniger, and C. Appenzeller, 2004: The role of increasing temperature variability in European summer heatwaves. *Nature,* **427,** 332–336, doi:10.1038/nature02300.

Tobler, W. R., 1979: Smooth pycnophylactic interpolation for geographical regions. *J. Amer. Stat. Assoc.,* **74,** 519–530, doi:10.1080/01621459.1979.10481647.

Wakefield, J., and H. Lyons, 2010: Spatial aggregation and the ecological fallacy. *Handbook of Spatial Statistics,* A. Gelfand et al., Eds., CRC Press, 541–558.

Wong, G., D. Maraun, M. Vrac, M. Widmann, J. Eden, and T. Kent, 2014: Stochastic model output statistics for bias correcting and downscaling precipitation including extremes. *J. Climate,* **27,** 6940–6959, doi:10.1175/JCLI-D-13-00604.1.

Zwiers, F. W., and Coauthors, 2013: Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. *Climate Science for Serving Society,* Springer, 339–389.