Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression

Matz A. Haugen*, Michael L. Stein and Elisabeth J. Moyer

University of Chicago, Chicago, USA

Ryan L. Sriver

University of Illinois at Urbana-Champain, Urbana, USA

**Corresponding author address:* Matz A. Haugen, 5734 S. Ellis Ave, 60637, Chicago, USA.

E-mail: mahaugen@uchicago.edu

ABSTRACT

Understanding future changes in extreme temperature events is critically important because of their potential societal cost, but inherently challenging in a transient climate. A single model simulation is generally insufficient to empirically characterize the statistical properties of underlying processes. Ensembles of repeated simulations (with different initial conditions providing independence) not only alleviate this data insufficiency but also allow new approaches for characterizing changes in extremes. We present here a new method of using ensembles that allows characterizing changes in temperature distributions using a continuous representation of seasonality rather than breaking the dataset into seasonal blocks. That is, we assume that temperature distributions evolve smoothly both day-to-day over an annual cycle and year-to-year over longer secular trends. To demonstrate our method's utility, we analyze an ensemble of 50 simulations of the Community Earth System Model (CESM) under a scenario of increasing radiative forcing to 2100, focusing on North America. The results both confirm aspects of climate system behavior known from previous studies and also elucidate new features.Confirming results include that daily temperature bulk variability generally decreases in wintertime in the continental mid- and high-latitudes ($> 40°$). New results include that in these same wintertime distributions, the low tails "stick", i.e. experience a lesser reduction in variability, producing a more negative skew. Although the examples above concern temperature only, the technique is sufficiently general that it can be used to generate precise estimates of distribution changes in a broad range of climate variables by exploiting the power of ensembles.

1. Introduction

Climate timeseries have generally been assumed to be separable into two components: randomness inherent in the underlying physical processes, which we call natural variability, and forced secular trends that follow from increasing concentrations of greenhouse gases. Recently, the degree to which that natural variability may itself be changing has received significant scientific interest (e.g. Trenberth 2011; Donat and Alexander 2012; Deser et al. 2012a; Thompson et al. 2015; Kay et al. 2015). Potential changes in climate extremes, because of their heightened societal impacts, are of special concern (e.g. Davison and Smith 1990; Stott et al. 2004; Chavez-Demoulin and Davison 2005; Eastoe and Tawn 2009; Otto et al. 2012; Swain et al. 2014; Singh et al. 2014; Trenberth et al. 2015; Diffenbaugh et al. 2015; Huang et al. 2015a; Jalbert et al. 2017). However, detecting and understanding such changes on timescales relevant to human activities is inherently challenging and generally data-limited. The long equilibration time of the climate system means that on the timescales of interest to human society, the climate state will be evolving (or 'transient'), so that its statistical properties are not stationary. Fully characterizing this evolving natural variability of rare events is intrinsically challenging due to the limited amount of data in either observations or simulations generally available. Studies of future climate extremes often employ statistical extreme value theory to make inferences about rare events with modest amounts of data.

n this work, we study the entire distribution of temperatures in a transient climate, including rare events, by employing quantile regression on an ensemble of simulations of an identical forcing scenario from a single climate model. Given sufficient care in choosing different initial conditions, such a large ensemble will reflect the natural variability of the system, since each simulation will be be statistically independent in terms of its natural variability. The increased data provided by multiple simulations then admits more confident statements about changes in the statistical behavior of the system than can be made with a single simulation. While the use of ensembles is a relatively recent development, it is growing rapidly (e.g. Deser et al. 2012b,a, 2014; Fischer and Knutti 2014; Kay et al. 2015; Sriver et al. 2015; Rodgers et al. 2015; Hagos et al. 2016). Deser et al. (2012b), Deser et al. (2012a) and Fischer and Knutti (2014) in particular discuss how ensembles help in distinguishing internal climate variability from anthropogenic effects and allow more accurate estimates for the forced model response. There remains however considerable room for additional development of statistical methods that exploit the benefits of ensembles.

Ensembles of multiple simulations offer at least three advantages that are currently under-exploited. The most important such advantage is that the increased data volume allows examining the entire distribution of a climate variable. Studies of climate variability to date are generally divided between those that address the center of the distribution (e.g. Semenov and Bengtsson 2002; Räisänen 2002; Kitoh and Mukano 2009; Screen 2014; Schneider et al. 2015), and those that address its tails (e.g. Katz and Brown 1992; Meehl et al. 2009; Northrop and Jonathan 2011; Davison et al. 2012; Huser and Davison 2014; Trenberth et al. 2015; Huang et al. 2015b; Jalbert et al. 2017), generally via extreme value theory. A more limited body of studies address overall distributional changes in climate variables, but these generally focus on observations or observation-based data products, which are necessarily limited in terms of data amount and therefore require spatial or temporal aggregation (Donat and Alexander 2012; Stainforth et al. 2013; Chapman et al. 2013; Huybers et al. 2014; McKinnon et al. 2016; Rhines et al. 2017). When studying model projections using ensembles, the large amount of data at each location allows us to accurately estimate changes in the distribution of climate variables (e.g. temperature) without spatial aggregation.

A second potential advantage provided by data-rich ensembles is that trends in both means and variability need not be modeled as linear in time (Franzke 2015; Gao and Franzke 2017). Typically, analyses assume linear trends, but in realistic scenarios, forcings are not linear over centennial timescales, and a linear approximation can be misleading (see for example Poppick et al. 2017). The increased data provided by ensembles means that we can consider more flexible statistical models to better represent complex climate responses. As we will show, distributions of daily temperature evolve nonlinearly, and follow different trajectories even as a function of quantiles (i.e. different parts of the distribution). Analysis methods should therefore be able to take into account nonlinearities both in time and across quantiles.

Finally, a third advantage of ensembles is that they allow a more natural treatment of seasonal variation in climate variables. In situations of limited data, it is standard practice to treat seasons separately, assuming that each season has a temporally constant average and stationary statistical properties discontinuous from neighboring seasons. With ensembles of simulations, we can allow for a smooth change in the underlying trend from day to day, using a parsimonious set of parameters. By modeling the entire year on a continuum, we can explore how each season transitions to the next and how seasonal patterns change over time, features that may be highly dependent on both geographic location and quantile.

We describe here a methodology for exploiting ensembles to study changing climate variability that captures these advantages: we model the complete distribution of daily temperatures as a continuous function of both seasonality and secular climate change over time. Although the methodology is applied to temperature here, it is general and can be applied to other climate variables of interest. We also show how such an ensemble-based approach is well-positioned for the purposes of uncertainty quantification. Because each simulation is treated as an independent sample drawn from the ensemble of simulations, we circumvent the issue of dependency within each simulation. We can therefore obtain uncertainty quantifications for all estimates by resampling complete simulations from the ensemble.

In the sections that follow, we describe estimated changes in both bulk and tail variability as differences in two quantiles; a large quantile difference implies more variability in a given region of the distribution. When those quantiles lie in the high or low tails, the quantile difference is a measure of the spread or thickness of the tail. Figure 1 gives a pictorial explanation of how quantile differences reflect bulk and tail variability. Although the estimated model is seasonally continuous, we also present results assuming seasonally constant conditions, and show that the seasonal effect on temperature can indeed be explained with a reasonably smooth function. When applied to model runs of a realistic future climate scenario, results reproduce some well-understood changes (e.g. strong reduction in wintertime variability at continental mid-latitudes) and produce some new insights (e.g. strong changes in skewness driven by low tail behavior).

2. Data

We apply our algorithm to an ensemble of 50 historical/future simulations of the Community Earth System Model (CESM) (Sriver et al. 2015). The atmospheric component is the low-resolution Community Atmosphere Model version 4, with T31 spectral resolution ($\sim 3.75^\circ \times 3.75^\circ$) and 26 vertical levels. The model ocean component is the low-resolution version of the Parallel Ocean Program version 2 (Smith et al. 2010) with a nominal horizontal grid resolution of 3°, augmented to approximately 1° at the equator. The ocean model contains 60 vertical levels, down to a maximum depth of 5,500 m.

The ensemble is especially appropriate for the purpose of studying variability because it is based on a $\sim 10,000$ year pre-industrial control simulation. After a $\sim$4000 year spin-up using constant preindustrial conditions, we initialize 50 historical hindcasts (1850-2005) from snapshots of the coupled model state taken every 100 years, so that the last hindcast is initialized after approximately 9000 years of the control simulation. We then extend each hindcast to 2100 using the

Representative Concentration Pathway (RCP) 8.5 scenario. The 100-year gap between each new initialization ensures nearly independent ensemble members that fully capture internal variability within the coupled system. RCP8.5 corresponds to anthropogenic radiative forcing of roughly 8.5 W m^{-2} by 2100 (Moss et al. 2010). More information about the model and ensemble design can be found in Sriver et al. (2015).

CESM does show some known biases that affect primarily temperature means (and possibly trends in means), but also to some extent the higher-order moments of the temperature distribution, e.g. variance and skewness. Known model biases include reduced ocean heat transport, low north Atlantic sea surface temperature, and excessive northern hemisphere sea ice (Shields et al. 2012). The model generally underestimates both temperature and precipitation extremes compared with observations, i.e. the mean of the extreme value distributions is biased, but the scale and shape are consistent with observations for the continental United States (Sriver et al. 2015).

To evaluate whether the CESM simulations provide sufficiently realistic temperature distributions for the purpose of this analysis, we compare CESM temperatures with those from the ERA-Interim (European Reanalysis) data product (Dee et al. 2011). Over the region of analysis, the model adequately reproduces overall geospatial patterns in temperature mean, standard deviation, and skewness, though with some quantitative discrepancies. Figure 2 shows the model/reanalysis comparison for winter; for summer see Supplementary Online Material Figure S1. The model underestimates variability somewhat, and produces excessively cold winter temperatures in the Arctic. The resulting exaggerated temperature gradients contribute to excess variability and excessively negative skew in the northern mid-latitudes. (Skewness is proportional to the cube of temperature after subtracting off the average seasonal temperature; see Appendix A1.) To build intuition around our data, we will show throughout this work results for the three locations **a**, **b**, and **c** shown in Figure 2, which have representative but different temperature distributions. See

Supplementary Online Material Figure S2 for comparison of model and reanalysis temperature distributions in both summer and winter for these locations.

3. Methods

In the methodology presented here, we model temperature at each location as a function of both seasonality and long-term change of the annual mean climate. We use two independent variables, with seasonality represented by a variable d, the day of the year (spanning values 1 to 365), and change in annual mean temperature represented by a variable t, years elapsed since 1850 (spanning 0 to 250 for these scenarios). We thus assume that each temperature quantile can be described by two sets of basis functions that represent the two variables' independent relationships with temperature (called here $\{f_i(d)\}$ and $\{g_j(t)\}$), and interaction terms $h_i(d)s_j(t)$, where f_i, g_j, h_i, and s_j are all smooth functions of the appropriate variable. The interaction terms are required to capture effects in which long-term temperature evolution differs between seasons, e.g. the robust projection that winter temperatures warm more than summer temperatures. To impose our smoothness condition, we assume that f_i, g_j, h_i, and s_j are piecewise cubic polynomials with a continuous second derivative, also called splines. (For a review of cubic polynomial basis functions, see Hastie et al. 2009, Chapter 5.) Because the seasonality variable d is periodic, its basis functions are also assumed periodic. For more details, see Appendix A2a.

We choose the number of basis functions by evaluating a metric representing model sufficiency. Our model sufficiency criterion is aimed at capturing the long term underlying signal. We do not require estimated quantile functions to capture transient events during the historical period like volcanic eruptions. Details on how we select the number of basis functions is given in Appendix A2b. In our climate simulation output, the intra-seasonal effect requires more detailed modeling than the inter-seasonal effect. In the results shown here, we fit the model with 15 terms (that is, basis functions) for the main seasonal effect $\{f_i\}$, but the interaction terms require less seasonal complexity, so we use only 3 terms for $\{h_i\}$. We use 4 terms for both the temporal change $\{g_j\}$ and the interaction terms $\{s_j\}$. That is, modeling long-term change generally requires fewer terms than modeling seasonality. In summary, we use 32 basis functions in total including an intercept term. We then fit each q^{th} quantile of temperature

$$T_q(d,t) = \alpha + \sum_i a_i f_i(d) + \sum_j b_j g_j(t) + \sum_{i,j} c_{i,j} h_i(d) s_j(t), \tag{1}$$

where all of the coefficients depend on q but we suppress the dependence for convenience. This fit determines coefficients a_i, b_j, $c_{i,j}$ for each quantile at each location.

To simplify notation, we construct a matrix X where each column contains a basis function and each row refers to a unique value of d and t. Using this matrix, X, we construct our temperature model in vectorized form,

$$T_q = X\beta_q, \tag{2}$$

where β_q contains the basis coefficients $a_i, b_j, c_{i,j}$. If the total number of basis functions is p and the total number of observations is n, then the predictor matrix X will have dimensions $n \times p$ and β_q will be a p-length vector. To get a confidence interval around T_q, we re-estimate the coefficients, β_q, using a resampled data set. Because we have 50 simulations we resample the data by drawing whole simulations from our ensemble of 50 simulations. Each simulation is treated as a contiguous block of data, and the dependency structure within these blocks is maintained when resampling whole simulations at a time. By repeating this resampling and re-estimation procedure 100 times we obtain pointwise confidence bands around each estimated T_q. Appendix A2c provides further details about uncertainty quantification.

As an example of a typical model fit, we show in Figure 3 the seasonal cycle in CESM daily temperatures for three locations, along with estimates of low, median and high quantiles. We show here data from 1850 to demonstrate the seasonal fit rather than that of the long-term trend. All locations show strong seasonal differences in variance that are well-represented by our smooth estimates. Relevant features that are captured include an asymmetrical seasonal cycle in all locations; a clear left skewness in wintertime in all three locations (although most pronounced in the higher-latitude **a** and **b**); and a distinct springtime shoulder in the higher-latitude locations. These characteristics show the benefit of explicitly modeling seasonal variations, since analyzing seasonal averages, and perhaps even monthly averages, would throw out meaningful details.

4. Results

To facilitate comparison to previous studies, we first perform a preliminary analysis where we replicate more standard methods. That is, we examine changes in the aggregate distribution of temperatures over multi-week and multi-month intervals, before we show results from our new approach that calculates responses for individual days. Even the standard analysis readily shows that temperature distributions in the CESM ensemble change markedly over the RCP 8.5 scenario (Figures 4, 5, and 6, which compare the initial and final time windows 1850-1864 and 2086-2100). Means uniformly shift warmer, but the shapes of the distributions also change in complex ways. Figure 4 shows initial and final distributions in our example locations for aggregated 15-day periods in winter and summer. In at least two of the three depicted locations, it is clear that the distributions are becoming narrower, although quantifying exactly how the tails are changing requires more sophisticated techniques.

The changes in distributions also exhibit complex spatial patterns. Figures 5 and 6 show patterns of change in mean, standard deviation, and skewness using 3-month seasonal blocks (DJF and JJA). We see the expected strong decrease in variance in winter over land, especially in the northern mid-latitudes. By contrast, summer variance changes are much smaller and differ in sign in different locations. Temperature skewness, i.e. the asymmetry of the distribution, shows strong changes in winter over land in a dipole pattern. Winter temperature distributions are in all time periods negatively skewed throughout most of the domain, but in the north (including locations **a** and **b**), they become more negatively skewed in the future, while in the south (including location **c**), they become more symmetric. Summer skewness changes are again smaller and with less spatial coherence, other than the strong transitions in the Southern Great Plains and in Mexico/Central America, where skewness in temperature distributions actually changes sign.

Our methodology for quantile estimation provides additional information that helps to quantify how temperature distributions are changing and to estimate the uncertainty associated with each change. We can evaluate not only bulk variability – the interquartile range (IQR), the difference between the 0.25th and 0.75th quantiles – but differences between any two quantiles. We therefore evaluate the difference between two low or high quantiles, denoted Δq_{low} and Δq_{high}, which measure tail variability in the same way that interquartile range measures the variability of the bulk distribution. If the skewness of a distribution changes over time, then future distributions are not simply scaled versions of present distributions. That is, their tail variabilities must change differently than does the IQR. In the case of the northern mid-latitudes winter temperatures shown in Figure 5, where distributions become more negatively skewed as bulk variability decreases in the future, the effect could result from either/both a low tail contracting less than the bulk (or actually increasing), or a high tail contracting more than the bulk. Our methodoogy allows readily differentiating these cases.

To assess whether the high tail and/or the low tail is driving changes in skewness, we consider the fractional changes in low, high, and bulk variability. If we denote the initial and final quantile difference as $\Delta q_{p,i}$ and $\Delta q_{p,f}$ at the p^{th} percentile, the temporal change in quantile differences relative to the initial year is then

$$\rho = \frac{\Delta q_{p,f} - \Delta q_{p,i}}{\Delta q_{p,i}}. \quad (3)$$

Because our method reveals the complete temperature distribution for each day of the year for all years, we choose a representative day to understand winter and summer changes (Jan 1 and July 5, respectively), and consider the difference between the beginning and end of the scenarios, the years 1850 and 2000. For these representative days, we show in Figure 7 the fractional variability changes of ρ for low and high tails as well as the IQR.

Results show that tail changes can indeed differ strongly from changes in the bulk of the distribution. In wintertime (Figure 7, top row), in much of the northrn mid-latitudes (including locations **a** and **b**), low tails change in a way that contributes to more a more negative skew. Low tail variability contracts less than does the IQR, while high tail variability contracts more strongly. (High tails would contribute to more negative winter skew predomimantly in the Hudson's Bay region, where the model shows distinct bias.) In summertime (Figure 7, bottom row), the high tail dominates the transition to positive skew in the Southern Great Plains region (including location **c**). See Supplementary Online Material for additional related figures that provide context. Figure S3 shows the same analysis over a larger region, and Figures S4 and S5 repeat Figure 7 but for daily maximum and minimum temperatures, respectively, rather than for daily means.

To clarify the relative contributions of high and low tails to skewness changes, we also examine evolving temperature variability in the bulk and tails as a function of seasonality as well as long term change. Figure 8 shows absolute variability changes for the three example locations **a**, **b**, and **c** estimated using our quantile model, and for fractional changes see Supplementary Online Material Figure S6. Colored lines show different years in the 1850-2100 simulations, and enveloping dashed lines represent 90% confidence intervals. See Appendix A2c for details about uncertainty quantification. Horizontal variations are the seasonal evolution of variability; vertical offsets between lines are long-term changes. In all locations, wintertime skewness changes are driven by the relative changes in IQR and low tails. In the higher-latitude locations **a** and **b**, more negative winter skew results because the IQR contracts even more strongly than does the low tail variability. In other words, the low tails "stick". In the lower-latitude location **c**, more positive winter skew results because the IQR barely changes while the low tail variability contracts strongly.

The complexity of the relationships in Figure 8 also shows how misleading it may be to use a three-month block to represent a season. With the detailed information that results from our approach, we can see just how complex even the present day variability profiles are. While all three locations show higher bulk variability (IQR) in winter than summer, the transition from winter to summer happens more quickly in some locations than at others, more abruptly in the northernmost location **a** and more gradually in **c**. Low-tail variability seasonal transitions are even sharper than those of IQR in **a** and **b**, but more gradual in **c**. In contrast, high-tail variability is more seasonally constant overall than low-tail variability. Through these examples, we see how our method offers detailed information about changes in variability across seasons and annual change, usually unavailable when analyzing each season separately.

While we show only three locations in the text here, an online interactive application allows similar in-depth examination of changes in model temperature distributions at all locations within North America, available at `https://matzhaugen.com/links.html`. The application allows the user to browse through any desired location to see how the variability changes as a function of season, year and quantile difference. We include temperature histograms of the first and last simulation year for the designated location, as well as maps that show the variability change spatially.

5. Conclusions

We present a method to quantify changes in tail variability of temperature with high precision in a transient climate model. Using data from the whole year and the whole span from 1850-2100 we estimate temperature as a function of seasonality and long term change. Analyzing the whole year simultaneously as opposed to analyzing each season separately allows for more flexible modeling of seasonality. The large ensemble makes it possible to fit such models stably.

By resampling entire simulations from the ensemble of climate simulations and recalculating the quantiles, we obtain confidence bands that do not require any assumptions of independence within any one simulation. We show that the smooth quantile estimates are accurate even across small intervals of the domain of the predictors. The fidelity of these intervals serves as a criterion to determine the required complexity in the statistical model.

The techniques presented in this study are validated in part by the fact that they replicate several prior conclusions made in the literature, e.g. the well-known projected decrease in winter variability in the northern mid-latitudes (e.g. Schneider et al. 2015) most likely due to amplified warming in the arctic (Screen 2014). Our approach furthermore allows us to quantify tail variability and give corresponding confidence intervals around our estimates. In the case study of CESM runs analyzed here, we relate the changes in tail variability to changes in skewness of the temperature distributions, and find that in most of the domain analyzed, wintertime skewness changes are driven largely by the relative behavior of IQR and low tails. For example, in much of the continental northern U.S. and Canada, the low tail of temperature contracts substantially less than does the overall temperature variability.

These results may inform physical explanations for the projection that skewness in winter temperature changes in a dipole pattern across North America. It is possible that the skewness change is a result of a change in the mean location and variability of the mid-latitude jet stream (e.g. Barnes and Polvani 2013); this possibility may warrant further study.

The abundance of data available in ensemble simulations relative to single simulations allows using quantile regression to accurately estimate high quantiles, avoiding some of the limitations of extreme value theory. Unlike quantile regression, methods using extreme value theory require making assumptions about the shape of the tail of the distribution. By parameterizing the seasonally time-varying distribution of temperature through smooth functions using the whole year as our domain, we also reveal previously unavailable details about seasonal transitions. For example, we show here that springtime variability decreases occur later in the year at lower latitudes, and that seasonal transitions in tail variability differ from those in IQR. While we analyze only temperature here, our method is intended to be general enough to be applied to other climate variables such as precipitation or humidity. These detailed insights into climate variable distributions may be valuable for risk assessment studies that emphasize extreme events.

Acknowledgments

This work was supported in part by STATMOS, the Research Network for Statistical Methods for Atmospheric and Oceanic Sciences (NSF-DMS awards 1106862, 1106974 and 1107046), and RDCEP, the University of Chicago Center for Robust Decision-making in Climate and Energy Policy (NSF grant SES-0951576). We acknowledge the University of Chicago Research Computing Center, whose resources were used in the completion of this work. Ryan L. Sriver acknowledges support from the Department of Energy sponsored Program on Integrated Assessment Model Development, Diagnostics and Inter-Model Comparisons (PIAMDDI), and the Program on Coupled Human Earth Systems (PCHES).

APPENDIX

A1. Model and reanalysis comparisons

Following the discussion on the paper, we define sample mean, variance and skewness as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$\gamma = \frac{1}{n}\sum \left(\frac{x_i - \bar{x}}{s}\right)^3 . \tag{A1}$$

These definitions are used in Figures 2, 5, and 6 in the main text and in Supplementary Online Material Figures S1 and S2. We plot the standard deviation s rather than the variance s^2.

A2. Model Details

In the following, we first give details regarding the regression of temperature quantiles on a fixed set of basis functions. We then discuss how to select the number of basis functions, through a "sufficiency criterion". Lastly, we describe how we quantify uncertainty in the quantile estimates.

a. Model estimation

Given the number of basis functions in our model, represented by the columns in a matrix X with number of rows equal to the number of observations in the data set, we construct our temperature quantile estimate, $\hat{T}_q$, and corresponding coefficients, $\hat{\beta}_q$, viz.

$$\hat{T}_q = X\hat{\beta}_q \tag{A2}$$

such that the q^{th} fraction of residuals between the observations T at a particular location and their estimates, $T - \hat{T}_q$, are greater than zero and a fraction $1-q$ are less than zero. With the temperature model in Equation 2, our coefficient vector estimate, $\hat{\beta}$, contains the estimates of $a_i, b_j, c_{i,j}$. Note that the seasonal interaction terms corresponding to the coefficients $c_{i,j}$ are not necessarily the same as the main seasonal terms corresponding to a_i. In fact, we find that fewer seasonal interaction terms are needed to describe the interaction behavior.

Computationally, obtaining the above quantile is equivalent to solving the following optimization problem (Koenker and Bassett Jr 1978),

$$\min_{\beta}\left\{\sum_{d,t:T(d,t)\geq X(d,t)\beta} q|T(d,t)-X^T(d,t)\beta| \quad + \right.$$
$$\left. \sum_{d,t:T(d,t)<X(d,t)\beta} (1-q)|T(d,t)-X^T(d,t)\beta|\right\}, \quad \text{(A3)}$$

and can be implemented in either R or MATLAB using existing libraries[1]. Because we have access to 50 simulations, each location provides us with $365 \times 250 \times 50$ or approximately 4.5 million observations. Consequently, even fairly high quantiles can be accurately estimated without borrowing data from neighboring locations through a spatial model as done by e.g. Reich et al. (2011). However, making inferences about more extreme quantiles, such as the quantiles .001 or .999, cannot be guaranteed to work as well with our methods.

We do not experience issues with quantile estimates crossing in our study area even though the optimization framework above does not explicitly enforce monotonicity with increasing percentile estimates. The absence of crossing quantiles is likely also due to the large sample size. For strict enforcement of monotonicity in the quantile curves see e.g. Bondell et al. (2010).

[1]We use the R library rq and the function `rq.fit.pfn`, developed by Portnoy and Koenker (1997). Basis functions are created using `pbs` for periodic spline basis functions and `ns` for non-periodic splines. The non-periodic splines are constrained to be linear beyond the domain, 1850-2100, and are called *natural splines*.

b. Model selection

We describe our approach to selecting a modest set of basis functions that can accurately represent the temperature data. If the model chosen has too many basis functions we run the risk of overfitting out-of-sample observations. To make sure this does not happen we need a metric to quantify the goodness-of-fit of the model.

Any reasonable temperature model we fit to the data will by definition contain the desired amount of positive and negative residuals *globally* according to the desired quantile q. A more stringent requirement would be that the smooth temperature estimate contains approximately an appropriate fraction of positive and negative residuals on a *daily* basis: for each d and t,

$$S(d,t) = \frac{1}{n}\sum_{i=1}^{n} I\left[\hat{T}_i(d,t) - T_i(d,t) > 0\right] \approx q, \tag{A4}$$

where I is the indicator function and n is the total number of samples (i.e. 50 for our CESM ensemble data set). If $S(d,t)$ is close to the value q for each d and t, the model would accurately describe the data and the number of basis functions is sufficient. In reality, we are looking basis functions that obey A4 with d averaged over blocks of days to increase the sample size, e.g. 10 days blocks. It is also not the goal to capture the quantile at too short a timescale as events like volcanic eruptions would interfere with the estimate.

In order to estimate the appropriate number of basis functions, we hold out 5 simulations from the fitting process and use these to calculate our exceedences, which we call $S_{test}(d,t)$. We repeat this 10 times so that all the simulations are eventually held out, giving 10 samples of $S_{test}(d,t)$. As we increase model complexity through degrees of freedom in the basis functions, the variability of S_{test} should reach a minimum when the necessary number of basis functions is reached and the quantile estimate is the same for each time point. If the number of basis functions is increased beyond this point, we start to overfit the data and the out-of-sample variability of S_{test} will increase.

To estimate S_{test}, we block the variables in two ways, one for each variable. First, we divide each year in 10-day bins and calculate the average exceedence estimate, $\hat{S}_{test}$, in each bin. We sum over the whole domain of long term change, t, and a subset of the seasonality variable, d. Specifically, let A be a set of non-overlapping contiguous blocks of days that together cover the whole year, where $a_j, \quad j = 1, ..., m$ are the elements of the set. Also let T be the index set for long term change, $T = [1850, 2100]$, measured in years. Then, for all $a_j \in A$,

$$\hat{S}_{test}(a_j) = \frac{1}{n} \sum_{i=[1,n], d \in a_j, t \in T} I\left[\hat{T}_i(d,t) - T_i(d,t) > 0\right]. \tag{A5}$$

To get an equal number of days in each bin we use the first 360 days of the year only.

Second, we divide the long term change variable, t, in bins and repeat the process by flipping the role of the variables in Equation A5 to get a set of $\hat{S}_{test}(b_j)$ with $b_j \in B$, a set of non-overlapping contiguous blocks of long-term change indices in T. An example of the blocked exceedence estimate is shown in Figure 9. Note that the pointwise quantile estimate is contained between the error bars, suggesting that the model is sufficiently complex. The standard deviation of these estimates of $\hat{S}_{test}$ is our measure of exceedence variability.

We seek the simplest model that gives good calibration of the quantile estimates (so close to 0.05 in Figure 9). At the same time we have to watch out to not overfit the data so we also want to minimize out-of-sample variability. We find that a model with 15 seasonal, 3 seasonal-interaction and 4 temporal degrees of freedom minimizes the variability of exceedences $\hat{S}_{test}$, shown in Figure 10, where seasonality has been binned. The out-of-sample fit when binning long-term change is shown in Figure S7 in the supplement. Here, models 4-6 have approximately equal test error, so since binning seasonality suggests the complexity of model 6, we chose model 6 as the overall model. Including the possible interaction terms, the full model has 32 free parameters to be fitted, or $\hat{\beta} \in \mathbb{R}^{32}$. All model candidates are shown in Table 1. We reach the same conclusion when blocking the long term change, t, and when analyzing different spatial locations (see Figure S7).

c. Uncertainty Estimation

With a reasonable model chosen through cross-validation, we present a way to quantify its uncertainty. Because we are using multiple simulations that are assumed independent, we resample entire simulations from the set of 50 simulations. Resampling 50 new simulations with replacement from the original set of simulations yields a new dataset. From the new data set we obtain another temperature estimate with the same model basis functions but different coefficients, β^*. After repeating this resampling and re-estimating procedure 100 times we generate pointwise confidence intervals for temperature quantiles. For example, in Figure 8 we show the 90% confidence interval by selecting the pointwise 5^{th} and 95^{th} percentiles of temperature variability estimates. Because the confidence intervals are quite tight we deem the 100 new estimates (or bootstraps) sufficient to indicate that the results we describe in section 4 are not due to random variation. Larger number of bootstrap replicates might give slightly more accurate intervals but would not change our conclusions. One might also consider fewer simulations as a compromise between computation time and quality of the estimates. Assuming normally distributed confidence intervals, we would expect the standard error to scale as $1/\sqrt{n}$. Thus, if one is willing to widen the confidence intervals by a factor of 2 (approximately) only 10 simulations would suffice. However, one could compensate for this greater variability by using fewer basis functions at a cost, of course, of obtaining less resolved estimates of seasonal patterns and long-term trends in the quantiles.

References

Barnes, E. A., and L. Polvani, 2013: Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *Journal of Climate*, **26 (18)**, 7117–7135.

Bondell, H. D., B. J. Reich, and H. Wang, 2010: Noncrossing quantile regression curve estimation. *Biometrika*, **97 (4)**, 825–838.

Chapman, S. C., D. A. Stainforth, and N. W. Watkins, 2013: On estimating local long-term climate trends. *Phil. Trans. R. Soc. A*, **371 (1991)**, 20120 287.

Chavez-Demoulin, V., and A. C. Davison, 2005: Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54 (1)**, 207–222.

Davison, A. C., S. A. Padoan, and M. Ribatet, 2012: Statistical modeling of spatial extremes. *Statistical science*, 161–186.

Davison, A. C., and R. L. Smith, 1990: Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 393–442.

Dee, D. P., and Coauthors, 2011: The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, **137 (656)**, 553–597.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012a: Communication of the role of natural variability in future North American climate. *Nature Climate Change*, **2 (11)**, 775–779.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012b: Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, **38 (3-4)**, 527–546.

Deser, C., A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *Journal of Climate*, **27 (6)**, 2271–2296.

Diffenbaugh, N. S., D. L. Swain, and D. Touma, 2015: Anthropogenic warming has increased drought risk in California. *Proceedings of the National Academy of Sciences*, **112 (13)**, 3931–3936.

Donat, M. G., and L. V. Alexander, 2012: The shifting probability distribution of global daytime and night-time temperatures. *Geophysical Research Letters*, **39 (14)**.

Eastoe, E. F., and J. A. Tawn, 2009: Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58 (1)**, 25–45.

Fischer, E. M., and R. Knutti, 2014: Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, **41 (2)**, 547–554.

Franzke, C. L., 2015: Local trend disparities of european minimum and maximum temperature extremes. *Geophysical Research Letters*, **42 (15)**, 6479–6484.

Gao, M., and C. L. Franzke, 2017: Quantile regression–based spatiotemporal analysis of extreme temperature change in china. *Journal of Climate*, **30 (24)**, 9897–9914.

Hagos, S. M., L. R. Leung, J.-H. Yoon, J. Lu, and Y. Gao, 2016: A projection of changes in landfalling atmospheric river frequency and extreme precipitation over western North America from the Large Ensemble CESM simulations. *Geophysical Research Letters*.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *Elements of Statistical Learning*. 2nd ed., Springer.

Huang, W. K., M. L. Stein, D. J. McInerney, S. Sun, and E. J. Moyer, 2015a: Changes in US temperature extremes under increased CO2 in millennial-scale climate simulations. *arXiv preprint arXiv:1512.08775*.

Huang, W. K., M. L. Stein, D. J. McInerney, S. Sun, and E. J. Moyer, 2015b: Estimating changes in temperature extremes from millennial scale climate simulations using generalized extreme value (GEV) distributions. *arXiv preprint arXiv:1512.08775*.

Huser, R., and A. Davison, 2014: Space–time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76 (2)**, 439–461.

Huybers, P., K. A. McKinnon, A. Rhines, and M. Tingley, 2014: US daily temperatures: The meaning of extremes in the context of nonnormality. *Journal of Climate*, **27 (19)**, 7368–7384.

Jalbert, J., A.-C. Favre, C. Bélisle, and J.-F. Angers, 2017: A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Katz, R. W., and B. G. Brown, 1992: Extreme events in a changing climate: variability is more important than averages. *Climatic change*, **21 (3)**, 289–302.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, **96 (8)**, 1333–1349.

Kitoh, A., and T. Mukano, 2009: Changes in daily and monthly surface air temperature variability by multi-model global warming experiments. *Journal of the Meteorological Society of Japan. Ser. II*, **87 (3)**, 513–524.

Koenker, R., and G. Bassett Jr, 1978: Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

McKinnon, K. A., A. Rhines, M. P. Tingley, and P. Huybers, 2016: The changing shape of northern hemisphere summer temperature distributions. *Journal of Geophysical Research: Atmospheres*, **121 (15)**, 8849–8868.

Meehl, G. A., C. Tebaldi, G. Walton, D. Easterling, and L. McDaniel, 2009: Relative increase of record high maximum temperatures compared to record low minimum temperatures in the US. *Geophysical Research Letters*, **36 (23)**.

Moss, R. H., and Coauthors, 2010: The next generation of scenarios for climate change research and assessment. *Nature*, **463 (7282)**, 747–756.

Northrop, P. J., and P. Jonathan, 2011: Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, **22 (7)**, 799–809.

Otto, F. E. L., N. Massey, G. J. Oldenborgh, R. G. Jones, and M. R. Allen, 2012: Reconciling two approaches to attribution of the 2010 Russian heat wave. *Geophysical Research Letters*, **39 (4)**.

Poppick, A., E. J. Moyer, and M. L. Stein, 2017: Estimating trends in the global mean temperature record. *Advances in Statistical Climatology, Meteorology and Oceanography*, **3 (1)**, 33–53, doi: 10.5194/ascmo-3-33-2017, URL http://www.adv-stat-clim-meteorol-oceanogr.net/3/33/2017/.

Portnoy, S., and R. Koenker, 1997: The Gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, **12 (4)**, 279–300.

Räisänen, J., 2002: CO2-induced changes in interannual temperature and precipitation variability in 19 CMIP2 experiments. *Journal of Climate*, **15 (17)**, 2395–2411.

Reich, B. J., M. Fuentes, and D. B. Dunson, 2011: Bayesian spatial quantile regression. *Journal of the American Statistical Association*, **106 (493)**, 6–20.

Rhines, A., K. A. McKinnon, M. P. Tingley, and P. Huybers, 2017: Seasonally resolved distributional trends of North American temperatures show contraction of winter variability. *Journal of Climate*, **30 (3)**, 1139–1157.

Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth System Model. *Biogeosciences*, **12 (11)**, 3301.

Schneider, T., T. Bischoff, and H. Płotka, 2015: Physics of changes in synoptic midlatitude temperature variability. *Journal of Climate*, **28 (6)**, 2312–2331.

Screen, J. A., 2014: Arctic amplification decreases temperature variance in northern mid-to high-latitudes. *Nature Climate Change*, **4 (7)**, 577–582.

Semenov, V., and L. Bengtsson, 2002: Secular trends in daily precipitation characteristics: greenhouse gas simulation with a coupled AOGCM. *Climate Dynamics*, **19 (2)**, 123–140.

Shields, C. A., D. A. Bailey, G. Danabasoglu, M. Jochum, J. T. Kiehl, S. Levis, and S. Park, 2012: The low-resolution CCSM4. *Journal of Climate*, **25 (12)**, 3993–4014.

Singh, D., and Coauthors, 2014: Severe precipitation in Northern India in june 2013: causes, historical context, and changes in probability. *Bulletin of the American Meteorological Society*, **95 (9)**, S58.

Smith, R., and Coauthors, 2010: The parallel ocean program (POP) reference manual ocean component of the Community Climate system Model (ccsm) and Community Earth System Model (CESM). *Rep. LAUR-01853*, **141**.

Sriver, R. L., C. E. Forest, and K. Keller, 2015: Effects of initial conditions uncertainty on regional climate variability: An analysis using a low-resolution CESM ensemble. *Geophysical Research Letters*, **42 (13)**, 5468–5476.

Stainforth, D. A., S. C. Chapman, and N. W. Watkins, 2013: Mapping climate change in european temperature distributions. *Environmental Research Letters*, **8 (3)**, 034 031.

Stott, P. A., D. A. Stone, and M. R. Allen, 2004: Human contribution to the European heatwave of 2003. *Nature*, **432 (7017)**, 610–614.

Swain, D. L., M. Tsiang, M. Haugen, D. Singh, A. Charland, B. Rajaratnam, and N. S. Diffenbaugh, 2014: The extraordinary California drought of 2013/2014: Character, context, and the role of climate change. *Bulletin of the American Meteorological Society*, **95 (9)**, S3.

Thompson, D. W. J., E. A. Barnes, C. Deser, W. E. Foust, and A. S. Phillips, 2015: Quantifying the role of internal climate variability in future climate trends. *Journal of Climate*, **28 (16)**, 6443–6456.

Trenberth, K. E., 2011: Attribution of climate variations and trends to human influences and natural variability. *Wiley Interdisciplinary Reviews: Climate Change*, **2 (6)**, 925–930.

Trenberth, K. E., J. T. Fasullo, and T. G. Shepherd, 2015: Attribution of climate extreme events. *Nature Climate Change*, **5 (8)**, 725–730.

LIST OF TABLES

	Seasonal	Seasonal-Int.	Temporal
1	5	3	3
2	7	3	3
3	10	3	3
4	10	3	4
5	12	3	4
6	15	3	4
7	15	3	5
8	15	5	5
9	18	5	5

TABLE 1. Degrees of freedom in the spline basis for each independent variable, with the interaction terms including the reduced set of seasonal polynomials with degrees of freedom listed in the middle column. The temporal polynomials are the same in both the main and interaction terms.

LIST OF FIGURES

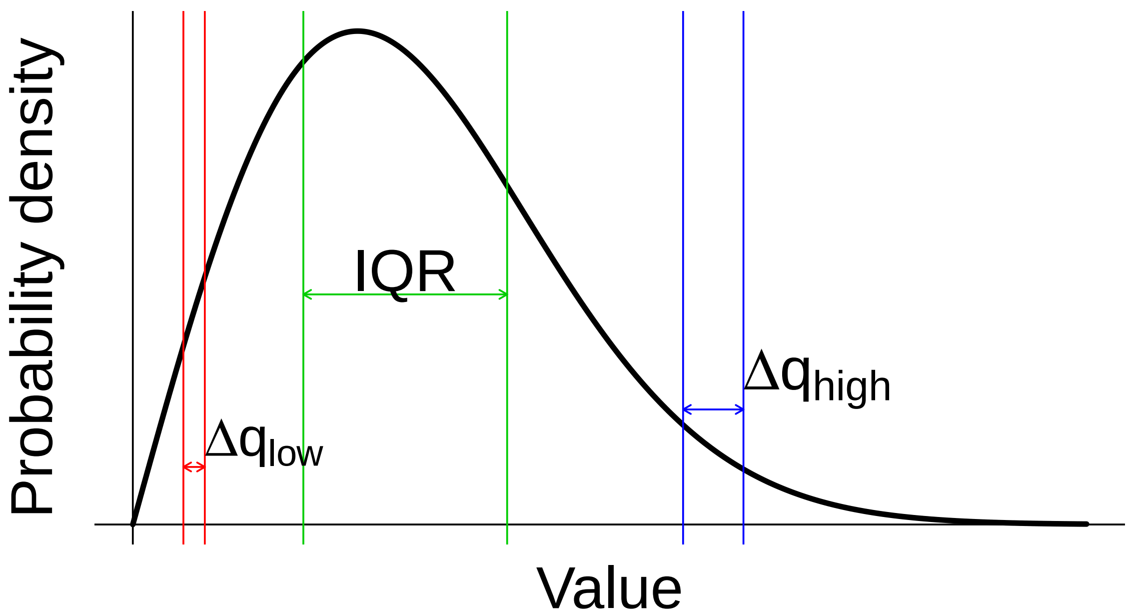


FIG. 1. An illustration of concepts and values related to distributions used in this paper. The cartoon shows a positively skewed (or "right-skewed") probability distribution and the three quantile differences discussed in this paper, in the low tail, high tail, and middle of the distribution. The p^{th} quantile in a distribution is the value such that the probability of being below it has probability p. Here Δq_{low} is the difference between the 0.025^{th} and 0.05^{th} quantiles; the IQR or Interquartile Range that between the 0.25^{th} and 0.75^{th} quantiles; and Δq_{high} that between the 0.95^{th} and 0.975^{th} quantiles. The values Δq_{low} and Δq_{high} quantify variability in extreme values while IQR quantifies variability in the bulk of the distribution.

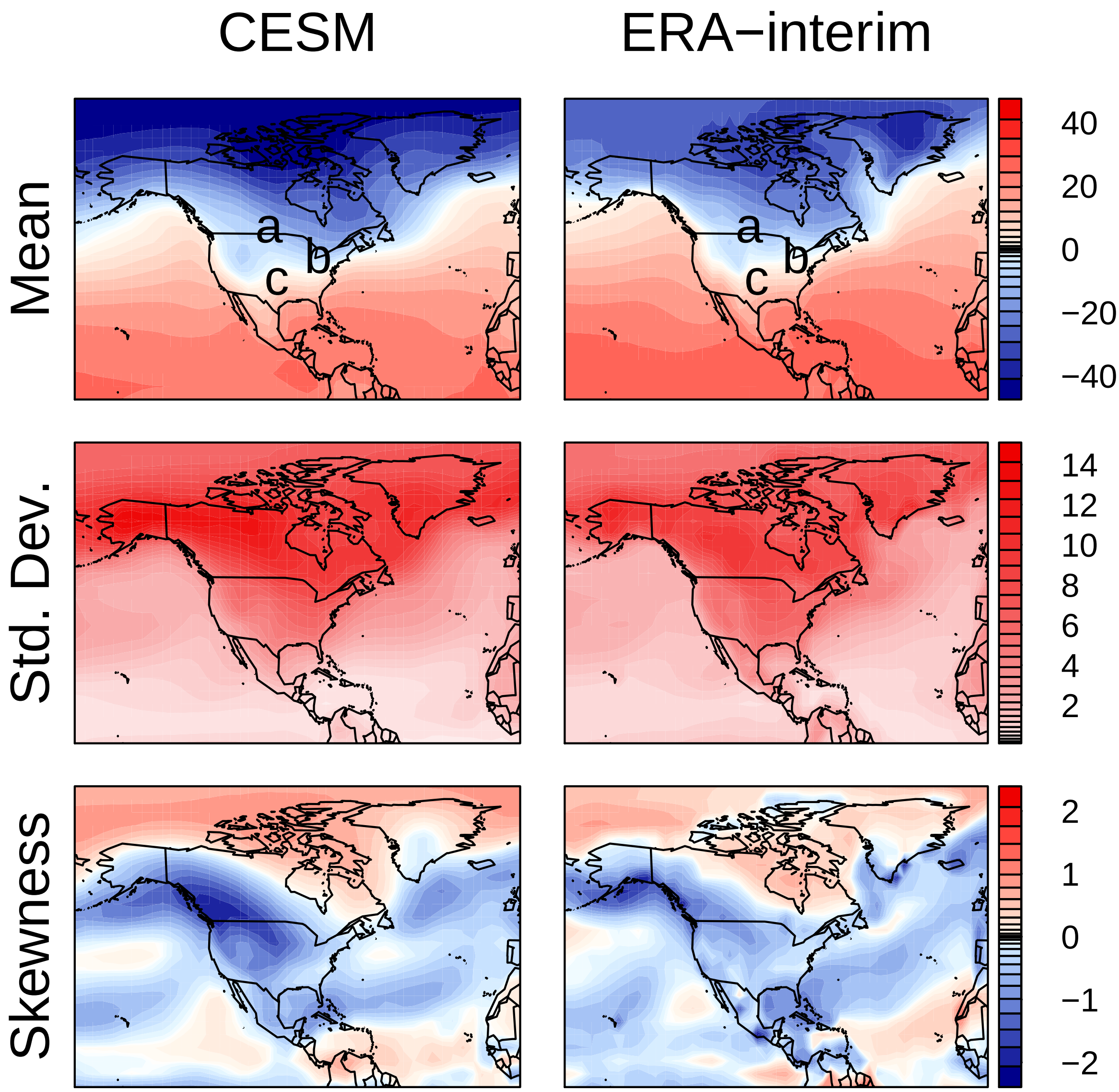


FIG. 2. Comparison of daily temperature distribution properties (mean, standard deviation, and skewness) between the CESM ensemble and ERA-Interim, for winter (DJF; aggregating all daily temperatures without deseasonalizing). We compare the years 1979-1994, the first available 15 years of the ERA-Interim dataset, and upscale the reanalysis from 0.75° to 3.75° resolution to match CESM. Units on top two rows are degrees Celsius; bottom row showing skewness is dimensionless. Winter skewness over the continental U.S. is negative in both model and reanalysis, implying a thicker lower tail; see Figure 1 and Appendix A1 for example and definitions. Overall, large-scale geospatial patterns are similar in both data sets, though some discrepancies are present (e.g. abnormally cold model Arctic winters). Letters **a-c** mark locations that will be used in examples throughout the paper; these are ordered from north to south, with latitudes and longitudes **a** $(50.1, -101)$, **b** $(42.7, -82)$, **c** $(35.3, -98)$.

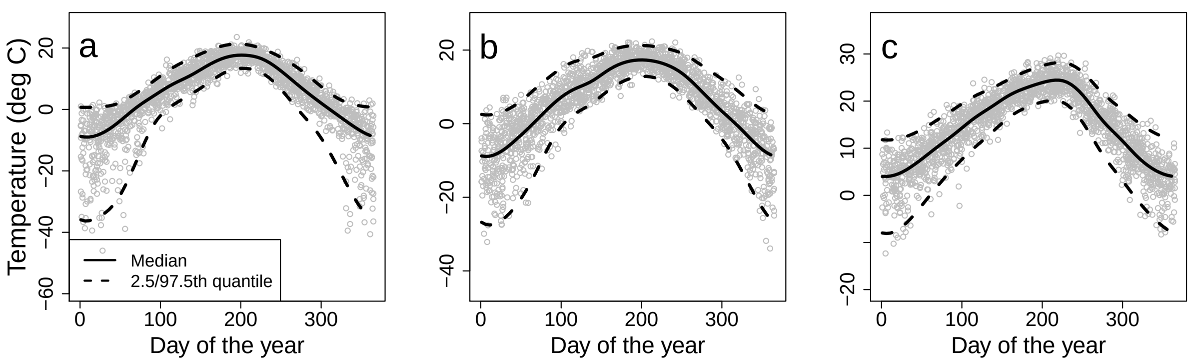


FIG. 3. Illustration of results of our quantile estimation procedure using the 50-member CESM ensemble. The figure shows ensemble daily mean temperatures for the year 1850 for the three representative locations **a**, **b**, and **c** defined in Figure 2. The ensemble provides 50 points per day but for clarity we show only 10% of the data. Solid lines show the median daily temperature and dashed lines the 2.5^{th} and 97.5^{th} quantiles estimated by our procedure. (Note the higher variabilities in winter.) Note that the location of the points exceeding the smooth median estimate are approximately uniform across time (notwithstanding the amplitudes of residuals), suggesting that the quantile estimate is accurate for each day. XX put in Michael's text here At all sites, the estimated quantile curves capture the seasonally changing patterns in the distributions reasonably well.

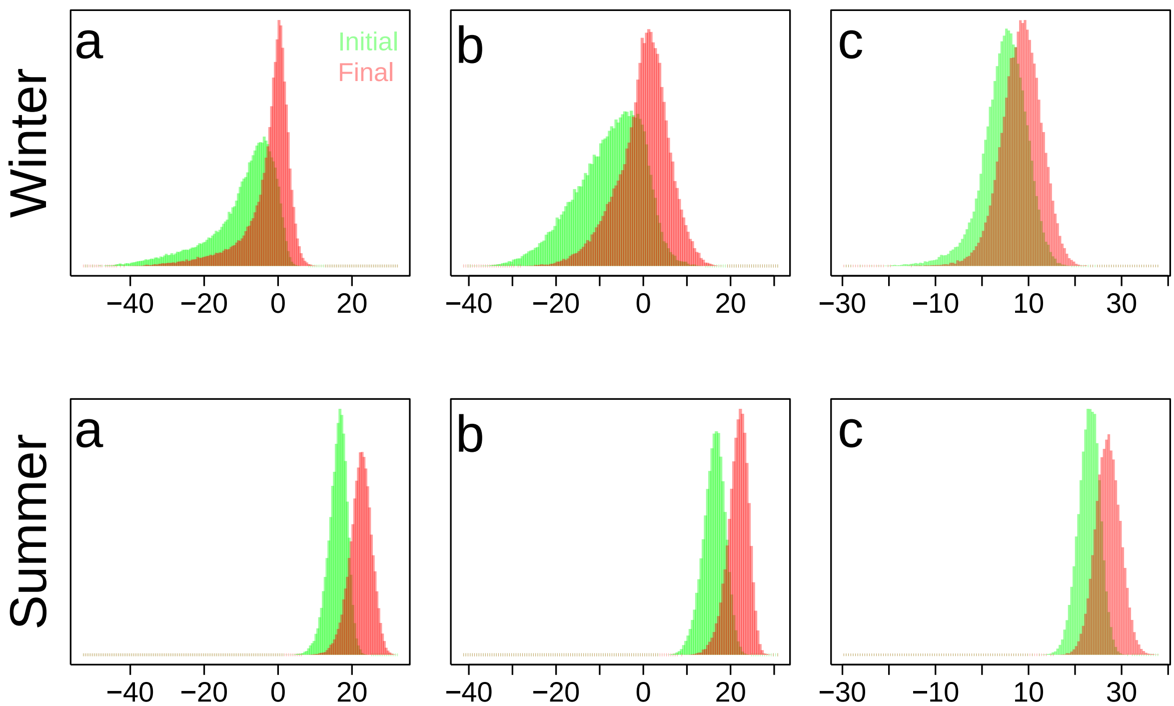


FIG. 4. Evolving distributions of daily mean temperature in the CESM ensemble RCP8.5 model runs at the locations **a, b, c** defined in Figure 2. Each distribution includes temperatures from a 15-day period over 15 model years for a total of 11,250 observations (15 days × 15 years × 50 ensemble members). Winter distributions are taken from Jan 1-15 and summer July 5-19; "initial" distributions include years 1850-1864 and "final" years 2096-2100. Changes in distributions are readily apparent, especially in winter at higher latitudes (locations **a** and **b**), but detailed quantification, especially of tail changes, requires more sophisticated techniques.

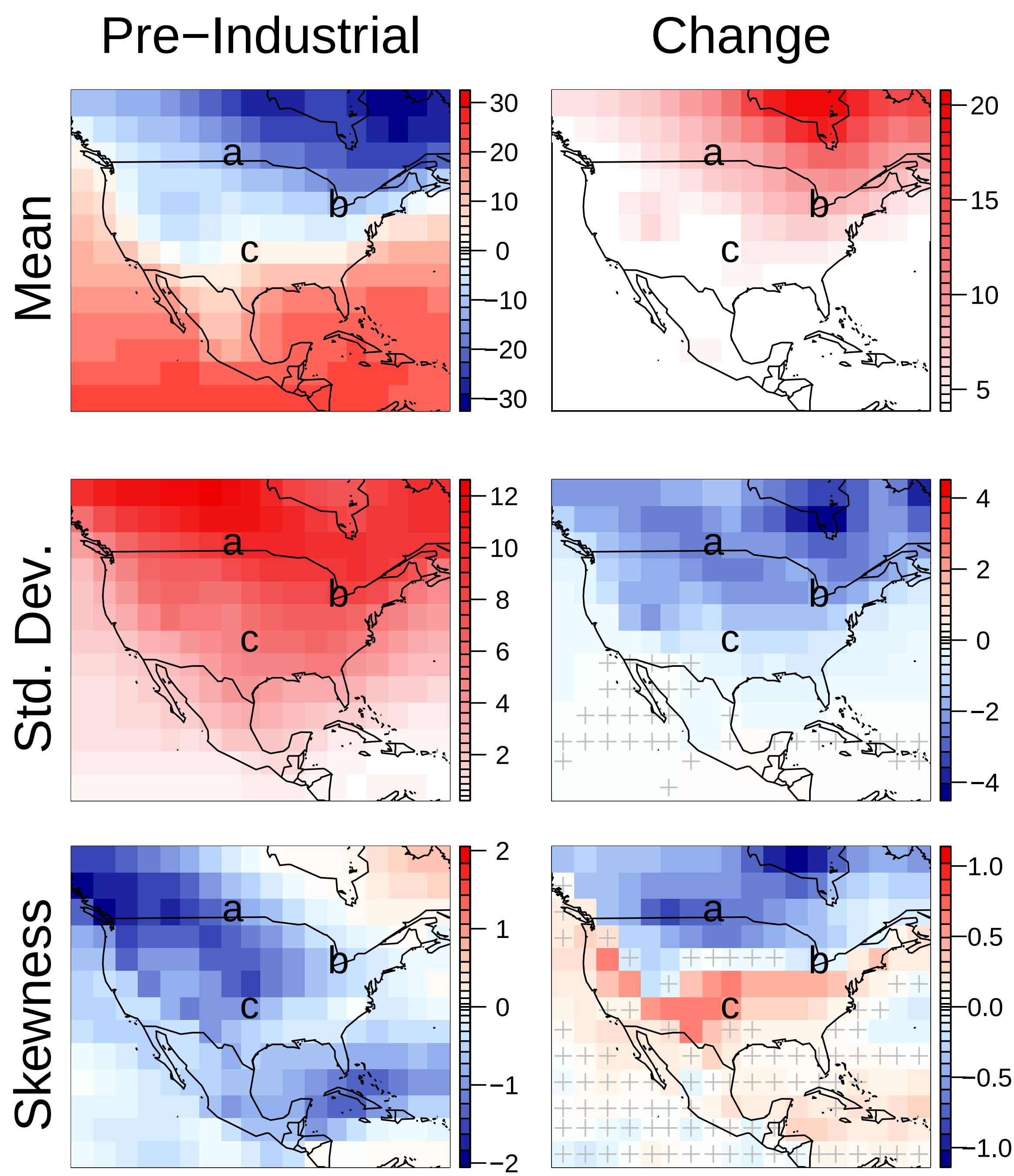
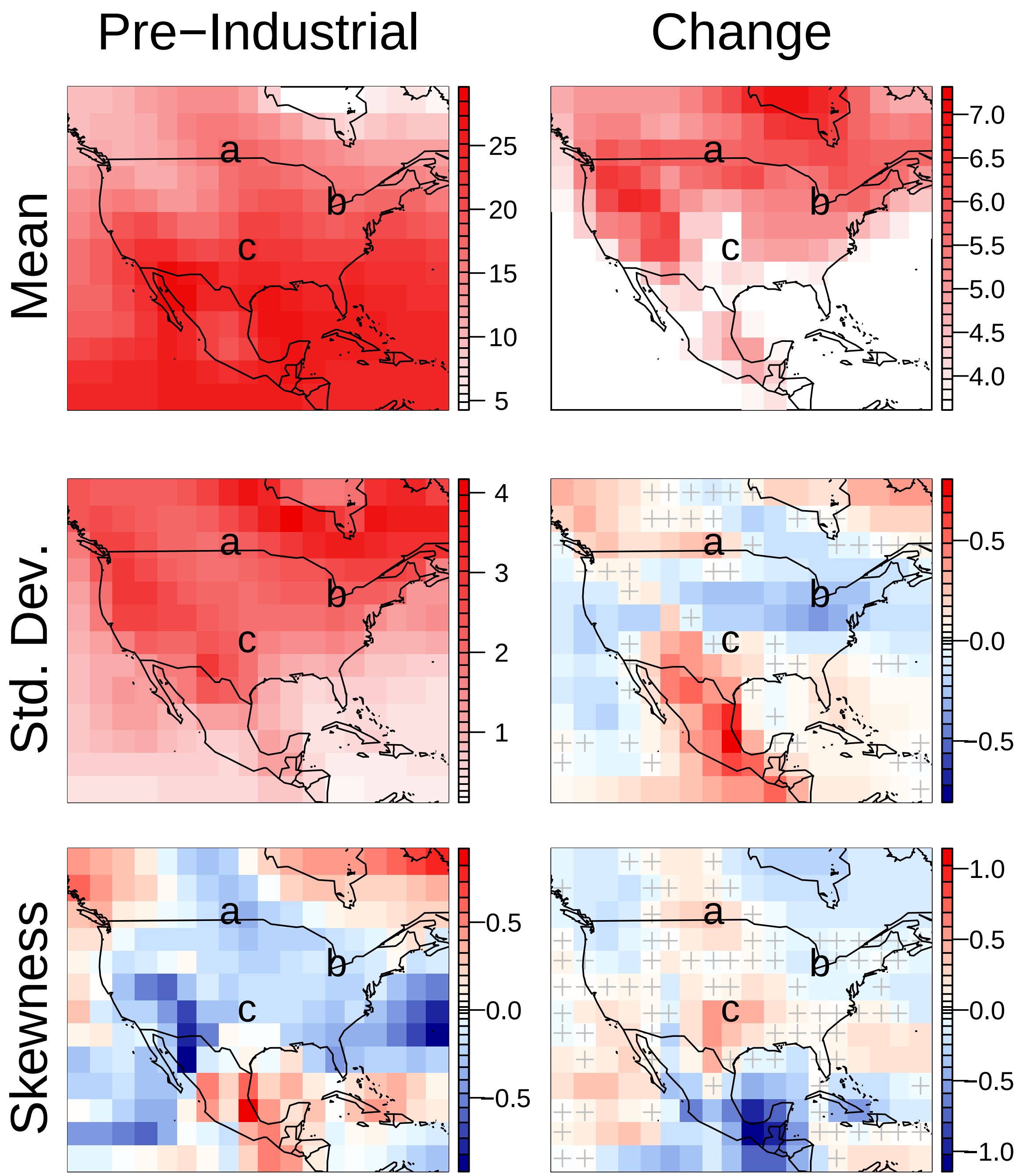


FIG. 5. Initial temperature distribution properties (left) and their changes over time (right) in the CESM ensemble RCP8.5 model runs, for aggregate wintertime (DJF) daily temperature. Initial ("pre-industrial") and final periods are defined as in Figure 4, as 15-year periods 1850–1864 and 2086–2100. Distributional moments (mean, standard deviation, and skewness) are defined as in Figure 2. Units on the top two rows are degrees Celsius, while the bottom row showing skewness is dimensionless. Gray crosses mark locations where the changes are not significant at the 0.05 level, obtained by resampling the set of 50 simulations (with replacement) and recalculating the sample moments. *Top right*: Mean temperature universally increases; note that color scale begins at XX degrees. Extreme warming in the Hudson's Bay region occurs where the model is biased low in present-day simulations. *Middle right*: As expected, standard deviation decreases strongly at higher latitudes. *Bottom right*: Changes in winter skewness show a dipole pattern, which enhances negative skew above $\sim 40°$ but reduces it at lower latitudes.



668 FIG. 6. As in Figure 5 but for aggregate summer (JJA) temperatures, and note that scales differ from those in Figure 5. Except
669 in the desert Southwest and Mexico, changes in standard deviation *(middle right)* and skewness *(bottom right)* are generally smaller
670 in summer than in winter and often not significant at the 0.05 level.

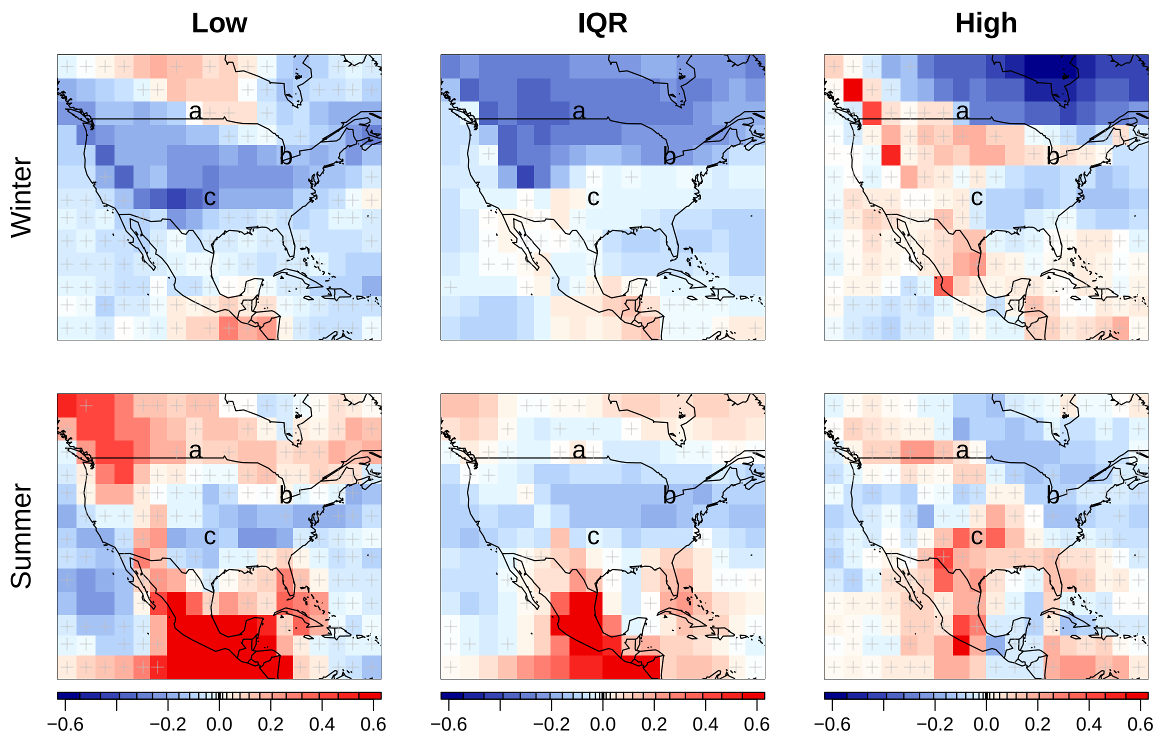


FIG. 7. Changes in daily temperature variability (quantile differences) over time in CESM ensemble RCP8.5 runs estimated using our statistical approach. Because our approach removes the need to aggregate over time when presenting changes, we show here differences in distributions for a single day and year: Jan 1 for winter (*top*) and July 5 for summer (*bottom*), with differences evaluated between the years 1850 and 2100. Changes are expressed as fractions of initial variability, so that the value 0 indicates no change with respect to the initial year. *Left*, *middle*, and *right* columns show, respectively, changes in low tail variability, IQR, and high tail variability, as previously defined. Gray crosses mark grid points where the change is less than 3 standard deviations from the original estimate. As expected, estimated changes in IQR (*middle*) are similar to changes in standard deviation seen in Figures 5 and 6. Changes in tail variability are clearly different from those in IQR, meaning that future distributions are not simply a rescaling of the present-day distributions.

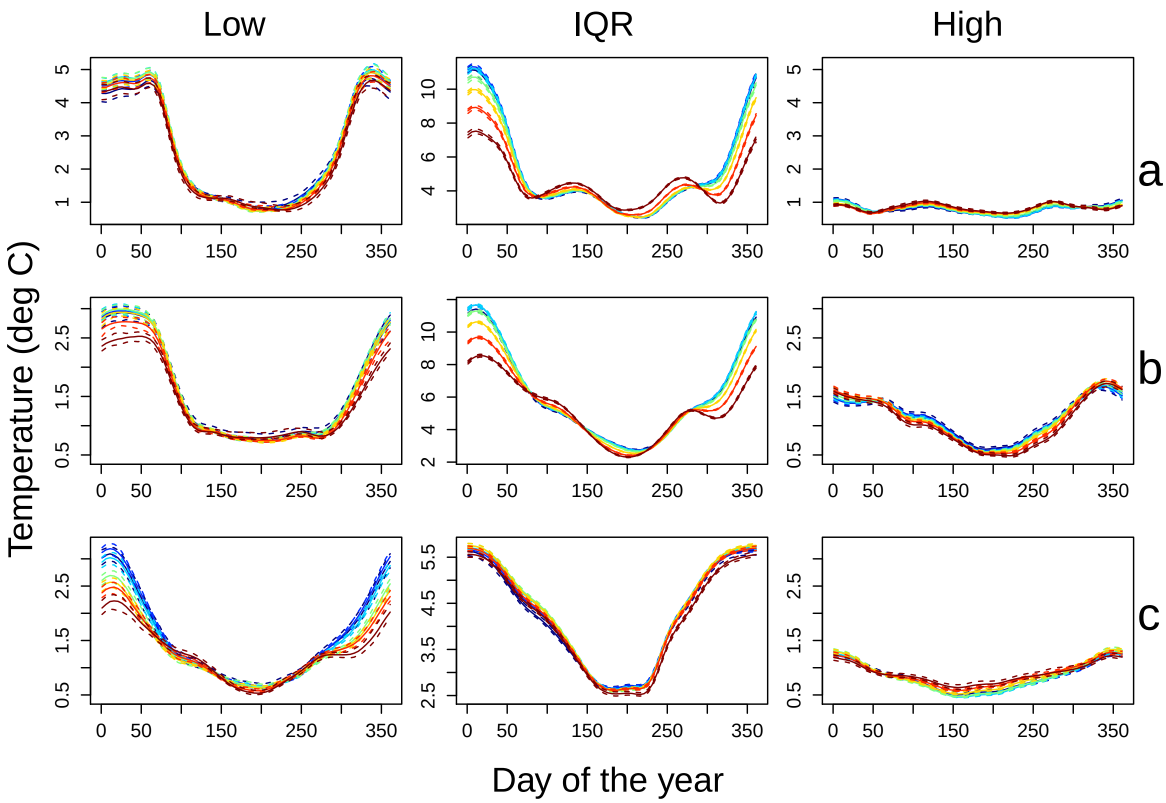
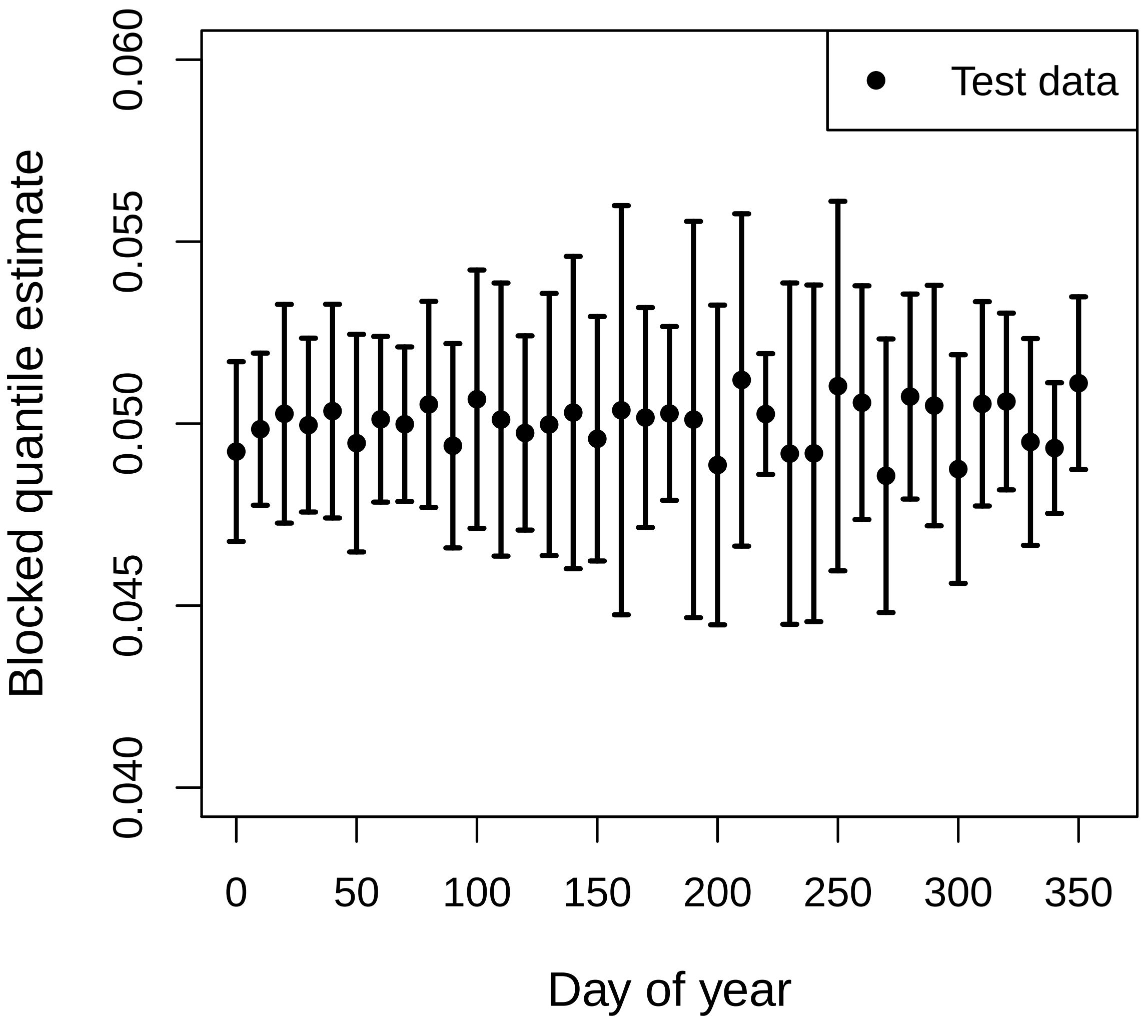


FIG. 8. Evolving daily temperature variability (quantile differences) over time in CESM ensemble RCP8.5 runs estimated using our statistical approach, for locations **a**, **b**, and **c**. Using the analysis described in Figure 7, we show absolute IQR and tail variability as a function of seasonality, with different years (at 40 year intervals) shown as different colored lines, from 1850 (dark blue) to 2090 (dark red). Dashed lines represent pointwise 90% confidence intervals. Note the complexity of seasonal cycles in variability at different locations. These results show that the dipole pattern of changes in wintertime skewness changes seen in Figure 5 is driven by low rather than high tail behavior. In wintertime, in the more northern locations **a** and **b**, IQR reduces more strongly than does low tail variability, making skew more negative. In the more southern location **c**, IQR change is negligible while low tail variability reduces strongly, making skew more positive. In all locations, absolute changes in wintertime low tail variability are larger than changes in high tails. For fractional changes, see Supplementary Online Material Figure S6.



689 FIG. 9. Exceedence probability of temperature events above the 95^{th} quantile estimate. The density is obtained by making 10-day bins and counting the number of observations that are above the quantile estimate and normalizing by the total number of exceedences aggregated across all model runs. Each bin is represented by the bin start day, i.e. an x-axis value of 0 includes the interval $(0, 10]$. We hold out 10 different sets of simulations to obtain 10 different estimates for each block of time, from which we calculate their mean shown as points and standard deviation shown as error bars around $\hat{S}_{test}$.

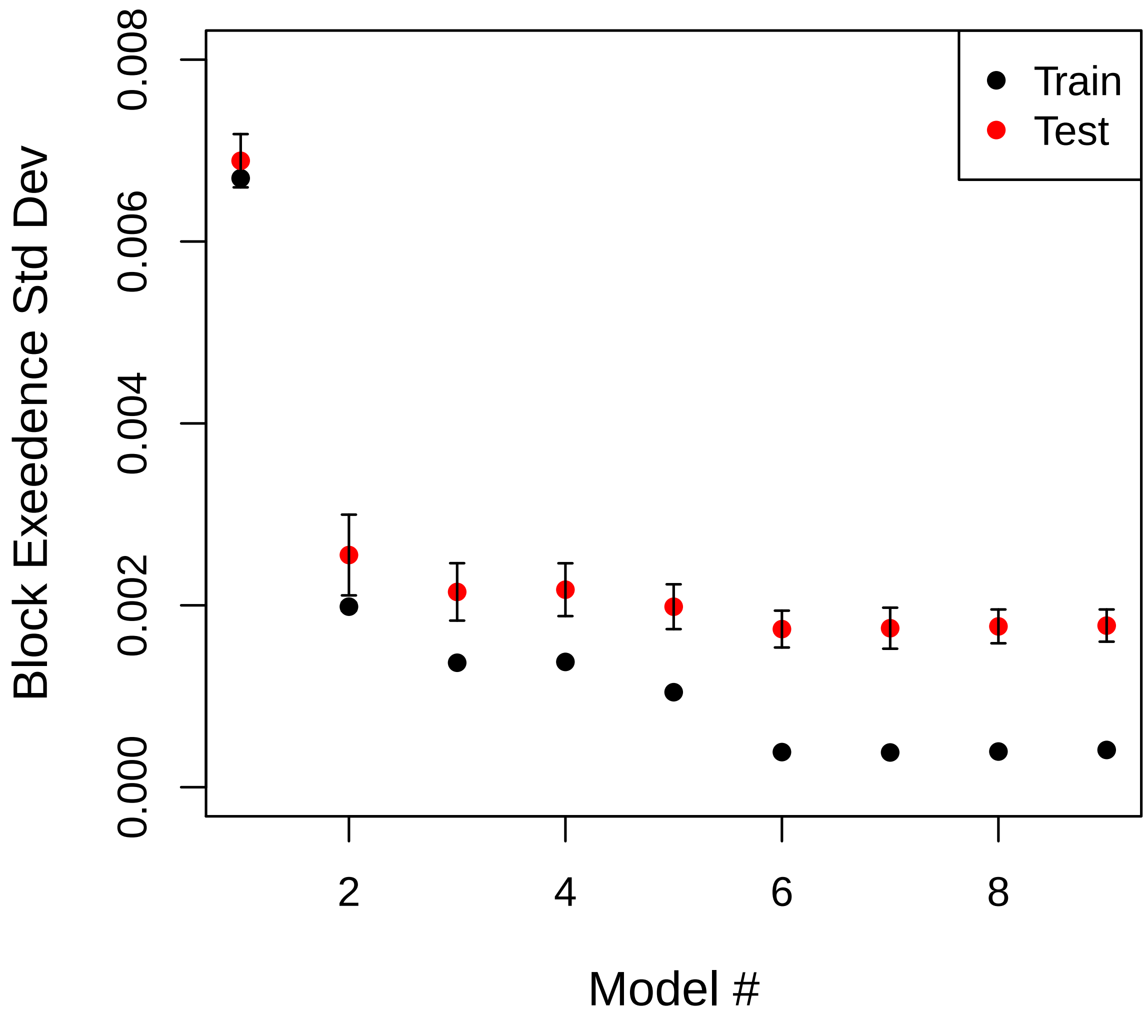


FIG. 10. Training and test exceedence standard deviation as a function of model number, where increasing model number signifies increasing degrees of freedom in the spline basis functions. The data were extracted from the gridbox located at (lat, lon) $= (31.5, -93.8)$. The exceedence is calculated by binning seasonality in 10-day blocks and summing over the long term change.