An evaluation of block-maximum-based estimation of very long return period precipitation extremes with a large ensemble climate simulation

Mohamed Ali Ben Alaya, Francis W. Zwiers, & Xuebin Zhang

2020

Original citation:

Ben Alaya, M. A., Zwiers, F. W., & Zhang, X. (2020). An evaluation of block-maximum-based estimation of very long return period precipitation extremes with a large ensemble climate simulation. *Journal of Climate*, *33*(16), 6957–6970. https://doi.org/10.1175/JCLI-D-19-0011.1

An Evaluation of Block-Maximum-Based Estimation of Very Long Return Period Precipitation Extremes with a Large Ensemble Climate Simulation

M. A. BEN ALAYA

Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada

F. ZWIERS

Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada, and Nanjing University of Information Science and Technology, Nanjing, China

X. ZHANG

Climate Research Division, Environment Canada, Toronto, Ontario, Canada

(Manuscript received 4 January 2019, in final form 4 June 2020)

ABSTRACT

The recurring devastation caused by extreme events underscores the need for reliable estimates of their intensity and frequency. Operational frequency and intensity estimates are very often obtained from generalized extreme value (GEV) distributions fitted to samples of annual maxima. GEV distributed random variables are ''max-stable,'' meaning that the maximum of a sample of several values drawn from a given GEV distribution is again GEV distributed with the same shape parameter. Long-period return value estimation relies on this property of the distribution. The data to which the models are fitted may not, however, be max-stable. Observational records are generally too short to assess whether max-stability holds in the upper tail of the observations. Large ensemble climate simulations, from which we can obtain very large samples of annual extremes, provide an opportunity to assess whether max-stability holds in a model-simulated climate and to quantify the impact of the lack of max-stability on very long period return-level estimates. We use a recent large ensemble simulation of the North American climate for this purpose. We find that the annual maxima of short-duration precipitation extremes tend not to be max-stable in the simulated climate, as indicated by systematic variation in the estimated shape parameter as block length is increased from 1 to 20 years. We explore how the lack of max-stability affects the estimation of very long period return levels and discuss reasons why short-duration precipitation extremes may not be max-stable.

1. Introduction

Even though we have an extensive understanding of key facets of climate and hydrologic systems from both dynamic and thermodynamic perspectives, for practical purposes, we do not yet have the ability to analyze and describe the intensity of many types of rare extremes based on physical reasoning. In the case of extreme precipitation, current knowledge of storm mechanisms remains insufficient to allow precise evaluation of the intensity and frequency of rare events that occur once in 100 or 1000 years such as are needed for engineering practice for water resources design and management. Probabilistic approaches using statistical frequency analysis are therefore widely used to estimate extremes for a given return period. This approach treats hydrometeorological variables as random variables governed by distribution laws where the upper tail of the distribution describes both the magnitude and frequency of extreme events. In practice, the distributional form is unknown and thus asymptotically motivated extreme value distributions are often used. This involves fitting an observed sample of extremes to such a distribution and using the fitted distribution to estimate the extreme quantiles of interest, often with extrapolation beyond

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-19-0011.s1.

Corresponding author: M. A. Ben Alaya, mohamedalibenalaya@uvic.ca

DOI: 10.1175/JCLI-D-19-0011.1

the observed data—for example, in cases in which estimates of 100-yr return level (RL) are required but station data are only available for a shorter period that is often only a few decades in length.

The probabilistic description of extremes through the extreme value theory (EVT) is usually obtained via either the block maxima approach or the peaks-over-threshold (POT) approach. The first uses the generalized extreme value distribution (GEV) to describe the probability distribution of the intensity of maxima of blocks of data (typically annual maxima), whereas the second uses the generalized Pareto distribution to model exceedances over a threshold [e.g., see Coles (2001), Smith (2003), and Kotz and Nadarajah (2000), among others]. Both approaches have been applied extensively in climate research (Caires and Sterl 2005; Kharin and Zwiers 2005). In both cases, the rate of decay of the upper tail of the distribution is determined by a shape parameter, which under suitable conditions is common between the two approaches. This rate of decay tells us (i) how quickly the largest values of the distribution increase as the probability of exceedance decays to zero and (ii) whether they can become infinitely large (nonnegative shape), or if there is an upper bound on how large they can become (negative shape).

The GEV distribution, through the block maximum approach, has been used more often than the POT approach, even though the latter may be able to make more efficient use of the information available in the observed data series (Lang et al. 1999). This is mainly because the implementation of the POT approach usually involves some subjective choices, mainly regarding the specification of a sufficiently high threshold, declustering of threshold exceedances and the treatment of the annual cycle. The lack of objective automatic procedures to address these issues hampers the wide application of POT approach in climate research for which analyses at a large number of locations are usually required. Furthermore, available data on extremes may already have been processed into block maxima, making it unsuitable for analysis with the POT approach. As we will see, the block maximum approach is also not entirely free of subjective choices, notably concerning the block length.

Past criticisms of traditional frequency analysis in water resource engineering include a series of papers by V. Klemeš (Klemeš 1986, 1987, 2000) that pointed to the limited information content of the historical data. Indeed, in statistical EVT, extrapolation to very low exceedance probabilities is performed without including any additional information from knowledge of the physical processes that generate extreme values. Generally, a model that allows extrapolation should contain more information than the data themselves, either explicit or implied. In the case of extreme value analysis via the block maximum approach, a key question is therefore whether the underlying physical processes produce block maxima that are max-stable, meaning that the maximum of a sample of several values drawn from a given GEV distribution is again GEV distributed with the same shape parameter. If there is no evidence that the data to which the distribution is fitted are indeed max-stable, there is then a question as to whether the fitted distribution can be used to extrapolate beyond available samples of extremes. This paper uses a large ensemble of historical simulations from the Canadian Regional Climate Model (CanRCM4) over North America to assess whether simulated extreme daily and subdaily precipitation amounts simulated by that model can be well described by a max-stable distribution (i.e., the GEV distribution), and to explore the implications of a lack of max-stability for long period RL estimates derived from the fitted distribution. We also consider very briefly some of the physical origins of a lack of max-stability. A modeling framework that integrates physically based information in an attempt to mitigate these problems will be described in a future paper.

2. Problem definition

The application of the extreme value theory in the real world can be viewed as a solution to a curve extrapolation problem, where the curve to be extrapolated is the upper tail of the parent distribution function of a variable of interest. Usually an estimate of part of the parent curve can be obtained from empirical data in the region where observations provide some information, while extrapolation is required outside the range of observed data. Extrapolation is constrained by the requirement that the distribution function should monotonically approach unity from below. In the case of the block maximum approach, the extremal types theorem (Fisher and Tippett 1928; Leadbetter et al. 1983) suggests the GEV distribution as a possible choice for describing a sample of block maxima, in which case extrapolation relies on max-stability property of GEV distributed random variables. In real world applications, the time series of block maxima to which the GEV is fitted may not be max-stable, possibly resulting in biased extrapolation.

The accuracy of the extrapolation to very long return period using extreme value analysis depends on two main aspects:

1) The accuracy of the estimated parameters of the extreme value distribution from the available data. This depends fundamentally on the size of the sample that is available for parameter estimation.

2) The validity of the assumption that the data to which the extreme value are max-stable.

From a statistical perspective, the latter assumption represents the main source of additional information that is required to extrapolate beyond the information contained in the data. Such an assumption cannot, however, be tested with limited observational records. To the best of our knowledge the only attempt in the current climate literature that involves a brief discussion about max-stability assumption is the work of Huang et al. (2016) that addresses temperature extremes using large datasets from climate simulations.

3. Data and methods

a. Data

We use hourly precipitation accumulations from 35 members of a 50-member large ensemble of CanRCM4 simulations covering North America at 0.44° spatial horizontal resolution (~50 km). Scinocca et al. (2016) provides a detailed description CanRCM4, which is a participant in the Coordinated Regional Climate Downscaling Experiment (CORDEX) (Giorgi et al. 2009). The CanRCM4 simulations were driven by a corresponding 50-member large ensemble simulation produced with the second generation of Canadian Earth System Model (CanESM2) that use the historical "all" forcing prescription including, solar and volcanic forcing, greenhouse gases, aerosols, ozone, and land use for the period 1951–2005, and RCP8.5 forcing for the period 2006–2100. The differences among ensemble members are due to internal variability. We use only output from the simulations for the period 1951–2000. Also, we use only the 35 simulations for which hourly precipitation was archived. Each of these simulations can be considered as a plausible realization of the real world (Deser et al. 2012) for the historical period 1951–2000. Hence, at a given location, the ensemble provides 35 times as much data as an observational record for the same period.

In our analysis, we have assumed stationarity over the 1951–2000 period that we consider as a working hypothesis. We make this assumption despite evidence of nonstationarity that is associated with the warming of the climate system from observational studies (e.g., Sun et al. 2020, manuscript submitted to *J. Climate*; Westra et al. 2013) and detection and attribution studies that compare observations with models (e.g., Min et al. 2011; Zhang et al. 2013). Such evidence emerges statistically when extremes are considered over broad continental to global scale areas, but it is not evident that attempting to model nonstationarity at local scales would reduce the root-mean-square errors of estimates of the magnitude of very rare extreme precipitation events, even when very large amounts of data are available (e.g., Li et al. 2019).

b. Methods

Hourly precipitation for individual grid cells was aggregated into 6-, 12-, and 24-h accumulations using sliding time windows; annual maxima of 1-, 6-, 12-, and 24-h accumulations were retained for analysis. Data from the 35 simulations were pooled to obtain samples of $50 \times 35 = 1750$ annual maxima for each duration. For each CanRCM4 grid cell and each duration we fitted a GEV distribution to the 1750 annual maxima using the maximum likelihood (ML) method. The large samples of 1750 annual maxima result in GEV distribution parameter estimates with very low sampling uncertainty, including the shape parameter estimates on which we will focus. The very large sample also allows us to fit versions of the GEV to block maxima for blocks that are longer than 1 year.

4. Results and discussion

Our analysis strategy proceeds as follows. In section 4a we first fit GEV distributions to block maxima at individual grid boxes and describe some of the basic characteristics of the fitted distributions and how they change as block length is increased. We next assess in section 4b the goodness of fit (GOF) of these distributions and consider differences in GOF when using annual and 10-yr block maxima. This is followed in section 4c by a detailed examination of the upper tail behavior of extreme precipitation at a selected set of representative locations. The results of these three subsections suggest that there are many locations where the samples of annual maxima from the ensemble of CanRCM4 simulations do not exhibit max-stability, which is inconsistent with the max-stability property that is inherent in the GEV distribution, We therefore examine the implications for the estimation of long period RLs in section 4d, and briefly consider in section 4e how the mixture of physical processes can potentially produce extreme precipitation that might affect tail stability.

a. Basic characteristics of the fitted GEV distributions

Figure 1 shows maps of the estimated shape parameter of the GEV distribution for CanRCM4 simulated daily and subdaily extreme precipitation for the historical period 1951–2000 over North America. As we can see, maps obtained using a single 50-yr CanRCM4 simulation are noisy, due to high uncertainty in shape parameter estimates. The second column of maps shows

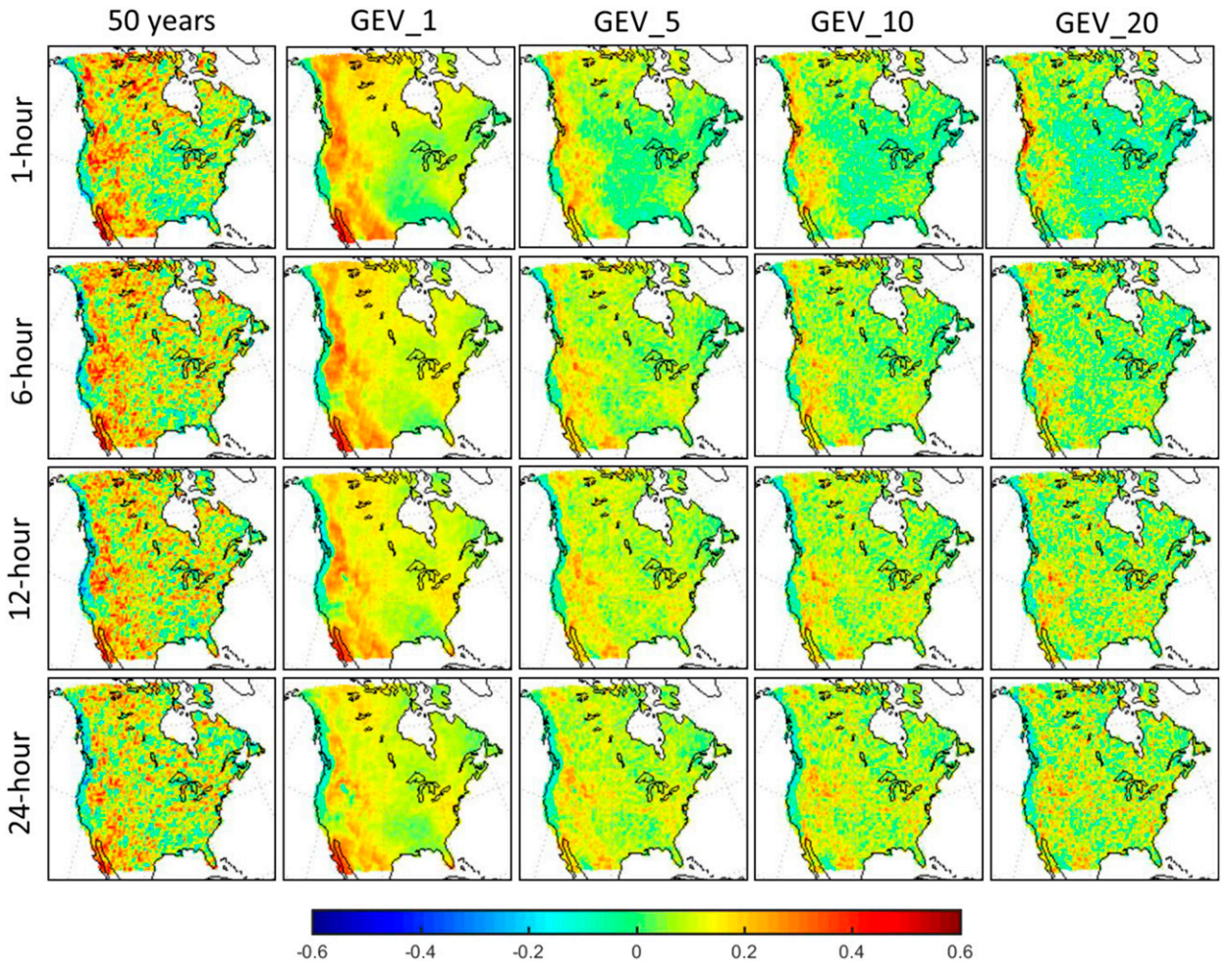

FIG. 1. Estimated shape parameters of the GEV distributions fitted to CanRCM4 simulated (top) 1-, (top middle) 6-, (bottom middle) 12-, and (bottom) 24-h precipitation accumulations for the historical period 1951–2000 over North America. Shown are (left) the shape parameter estimates when the GEV distribution is fitted to a sample of 50 annual maximum values from only one of the 35 CanRCM4 simulations and shape parameter estimates using block maxima from the ensemble of 35 CanRCM4 simulations for (left center) 1-yr (GEV-1; 1750 blocks), (center) 5-yr (GEV-5; 350 blocks), (right center) 10-yr (GEV-10; 175 blocks), and (right) 20-yr blocks (GEV_20; 87 blocks).

that the use of 1750 annual maxima pooled from the 35 ensemble members leads into much smoother maps, reflecting a substantial reduction of sampling uncertainty. Note that there is greater spatial variation in the shape parameter for the annual maxima of subdaily accumulations than for daily accumulations. Subsequent columns illustrate shape parameter estimates obtained for blocks of lengths of 5, 10, and 20 years. Spatial noise that likely originates from sampling variability again becomes apparent as block length increases and the number of blocks correspondingly decreases. Nevertheless, the magnitudes of the estimated shape parameter appear to decrease, on average, over North America with increasing block length.

Figure 2 shows how the shape parameter estimates, averaged over the continent, vary as a function of block length. These are compared with the median shape parameter estimate that is obtained using annual maxima from a single CanRCM4 simulation. The 80% uncertainty intervals that are displayed reflect the total variation in the shape parameter estimates, combining spatial variations with uncertainty in parameter estimates due to sampling variation. As expected, the uncertainty in the shape parameter estimates for the distribution of annual maxima is much larger when using only a single climate simulation (i.e., a sample of 50 annual maxima) than when using the 35 ensemble simulations (1750 annual maxima). Similarly, the variation in parameter estimates increases with increasing block length due to the decline in the number of available blocks. Systematic variations in the estimated shape parameter with block length can also be seen (Fig. S1 in

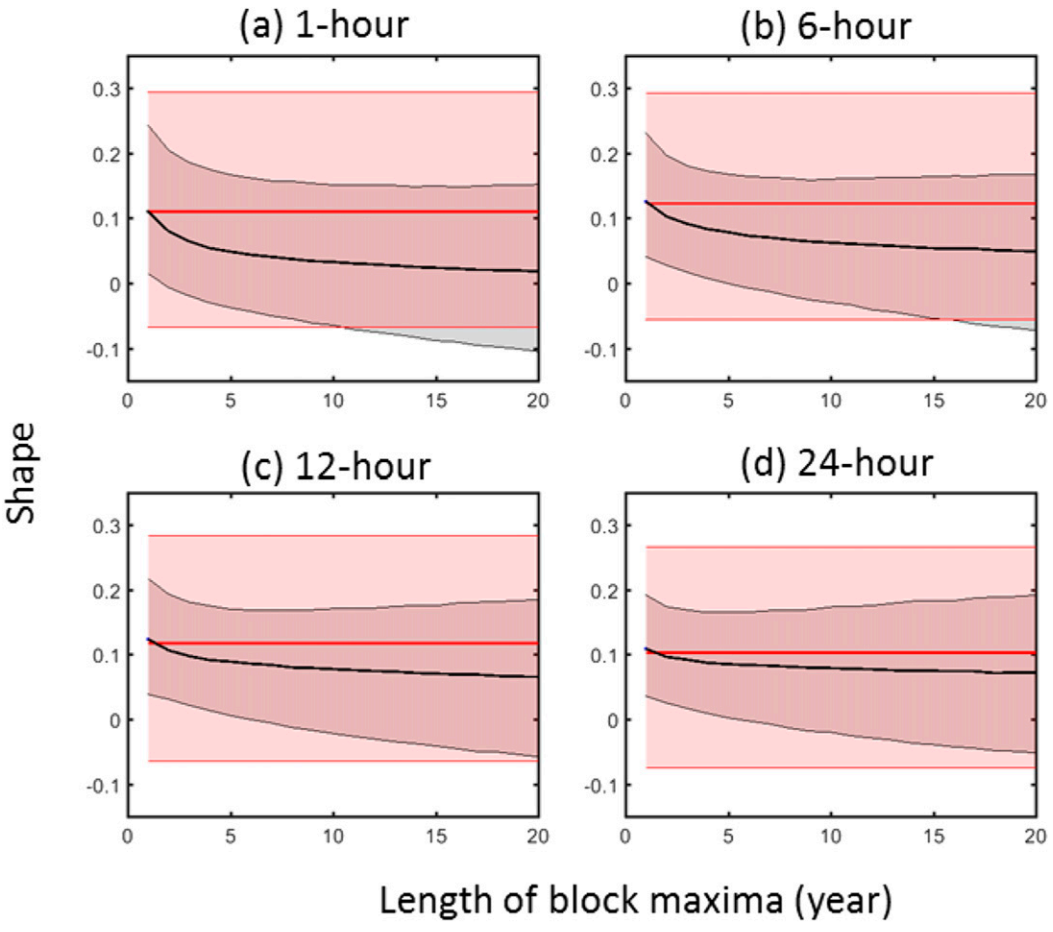

FIG. 2. Spatial averages over North America of estimated shape parameters of the GEV distributions fitted to CanRCM4 simulated (a) 1-, (b) 6-, (c) 12-, and (d) 24-h precipitation accumulations for the historical period 1951–2000 at individual grid boxes as a function of block length using 35 CanRCM4 simulations (black solid line) together with 80% uncertainty intervals (gray shading). The uncertainty intervals are obtained by calculating the 10th and 90th percentiles of shape parameter estimates across grid points over North America. For reference, the solid red line and red shading show corresponding results for GEV distributions fitted to annual maxima only from a single 50-yr CanRCM4 simulation.

the online supplemental material) in individual Bukovsky regions (Bukovsky 2012) (supplemental Fig. S2), with different types of variation evident in different regions.

b. Goodness of fit

Figures 1 and 2 and Fig. S1 suggest that annual maxima simulated by CanRCM4 may not have the max-stability properties that are implicit in GEV distribution. Given the large sample of annual maxima that is available, there should therefore be an indication that the GEV does not fit this sample well, which is indeed the case. Figure 3a shows that the null hypothesis that the samples of 1750 annual maxima follow the fitted GEV distribution was rejected using a Pearson's chi-square GOF test (see section S1 in the online supplemental material for details) for a large fraction of grid boxes, suggesting that max-stability is not valid for annual maxima over much of North America in the climate simulated by CanRCM4. While the evidence based on a single sample of 50 maxima is less clear, we note that rejection nevertheless occurs at 13.4% of grid boxes, which is larger than would be expected by random chance for a test operating at the 5% significance level given the scale of spatial dependence of extreme precipitation. This suggests that, even with this length of the record, there may be evidence that is relevant to the question of whether max-stability holds, if information is aggregated across space.

Note that since we will make several more comparisons between results that can be obtained from a single 50-yr CanRCM4 simulation and the 1750-yr sample from the CanRCM4 large ensemble, we also briefly

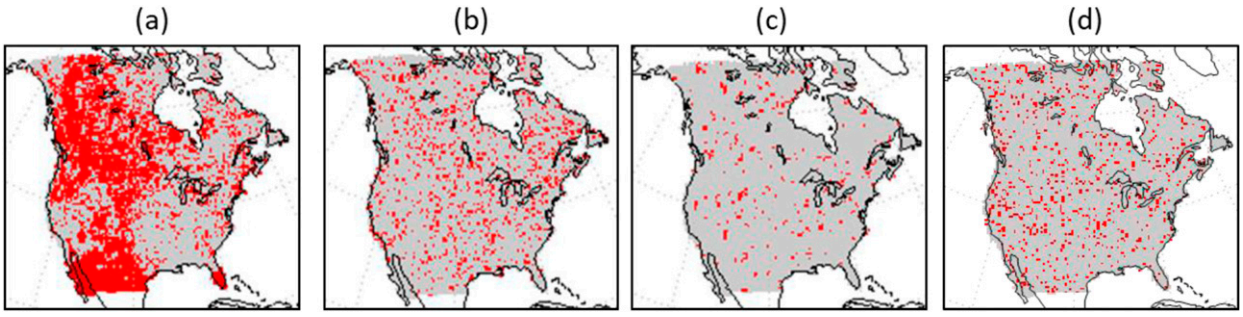

FIG. 3. Goodness-of-fit test results for extreme hourly precipitation for each grid box over North America. (a) Results of chi-square tests of the null hypothesis that the pooled sample of 1750 annual maxima from all 35 CanRCM4 simulations follows the GEV distribution. (b) As in (a), except testing that the null hypothesis that the 50 annual maxima from one CanRCM4 simulation follow the GEV distribution. (c) Results of a two-sample Kolmogorov–Smirnov test of the null hypothesis that the 50 annual maxima from one CanRCM4 simulation have the same distribution as the 1700 annual maxima from the remaining 34 CanRCM4 simulations. (d) As in (a), except testing the null hypothesis that the pooled sample 175 ten-year maxima from all 35 CanRCM4 simulations follows the GEV distributions. Red points show locations where rejections occur.

compare the maxima from a single 50-yr simulation with the remaining 1700 maxima (Fig. 3c). In this case we find no evidence to suggest that the simulation that provided the 50 maxima is statistically distinct from the other 34 CanRCM4 simulations. Indeed, the null hypothesis is rejected at about 4.2% locations, which seems entirely consistent with the specified significance level.

c. *Assessment of upper tail behavior at specific locations*

While some information indicating a lack of fit of the GEV distribution to annual maxima may be available from individual 50-yr records, such information is not reliably available (e.g., locations where rejection occurs in Fig. 3b are randomly scattered), suggesting that the GEV distribution may often be unknowingly inappropriate for fitting to annual maxima. It is therefore desirable to examine how these results regarding the stability of the shape parameter could affect the estimation of high RLs. Figure 4 shows plots of estimated RLs as a function of return period at four different locations A, B, C, and D based on GEV distributions fitted to a sample of 50 annual maxima (using a single CanRCM4 simulation) and also using the 1750 annual maxima (from 35 simulations). Shading indicates 80% confidence intervals obtained by bootstrapping. The geographical positions of the four locations are shown in Fig. 4a. The locations are chosen to illustrate the variation in tail behavior across North America that can be seen in the 1750 years of CanRCM4 output.

For location A in the Pacific Northwest region, the GEV fitted to a sample of 50 annual maxima seems to underestimate high RLs (see Fig. 4b). Underestimation apparently persists when increasing the sample size to 1750 annual maxima, even though the much larger sample substantially reduces the uncertainty of the GEV parameter estimates and high RL estimates. A reasonable hypothesis is that this underestimation may be related to the increase of the shape parameter beyond annual maxima (Fig. 4c). In contrast, the GEV distributions fitted to annual maxima at location B appear to overestimate high RLs (Fig. 4d), consistent with the decrease of the shape parameter with increasing block length (Fig. 4e). The third location, C, shows an example where the stability of the shape parameter seems to be valid beyond annual maxima (Fig. 4g), and thus the GEV distribution fitted to the full sample of 1750 annual maxima characterizes 1-h duration extreme precipitation relatively well (Fig. 4f). Location D shows a midcontinental point where substantial overestimation of extreme quantiles using the GEV fitted to a sample of 50 annual maxima appears to be principally due to overestimation of the shape parameter. At this location, the estimated shape parameter appears to stabilize for blocks longer than about 3 years. Results for an additional nine locations representative of Bukovsky regions are displayed in Figs. S4 and S6 in the online supplemental material.

We now focus on the estimation of a quantile deep in the upper tail of the distributions of extreme precipitation, namely the 1000-yr RL. Figure 4 suggests that estimates based on GEV distributions fitted to annual maxima are biased relative to empirical quantile estimates at three of the four locations considered, even when using the large samples of 1750 annual maxima to fit the distributions. Apparent biases are also seen at about one-half of the locations representative of Bukovsky regions (Fig. S6). Several sources of uncertainty need to be considered to assess the hypothesis that the biases that are apparently seen in 1000-yr RL

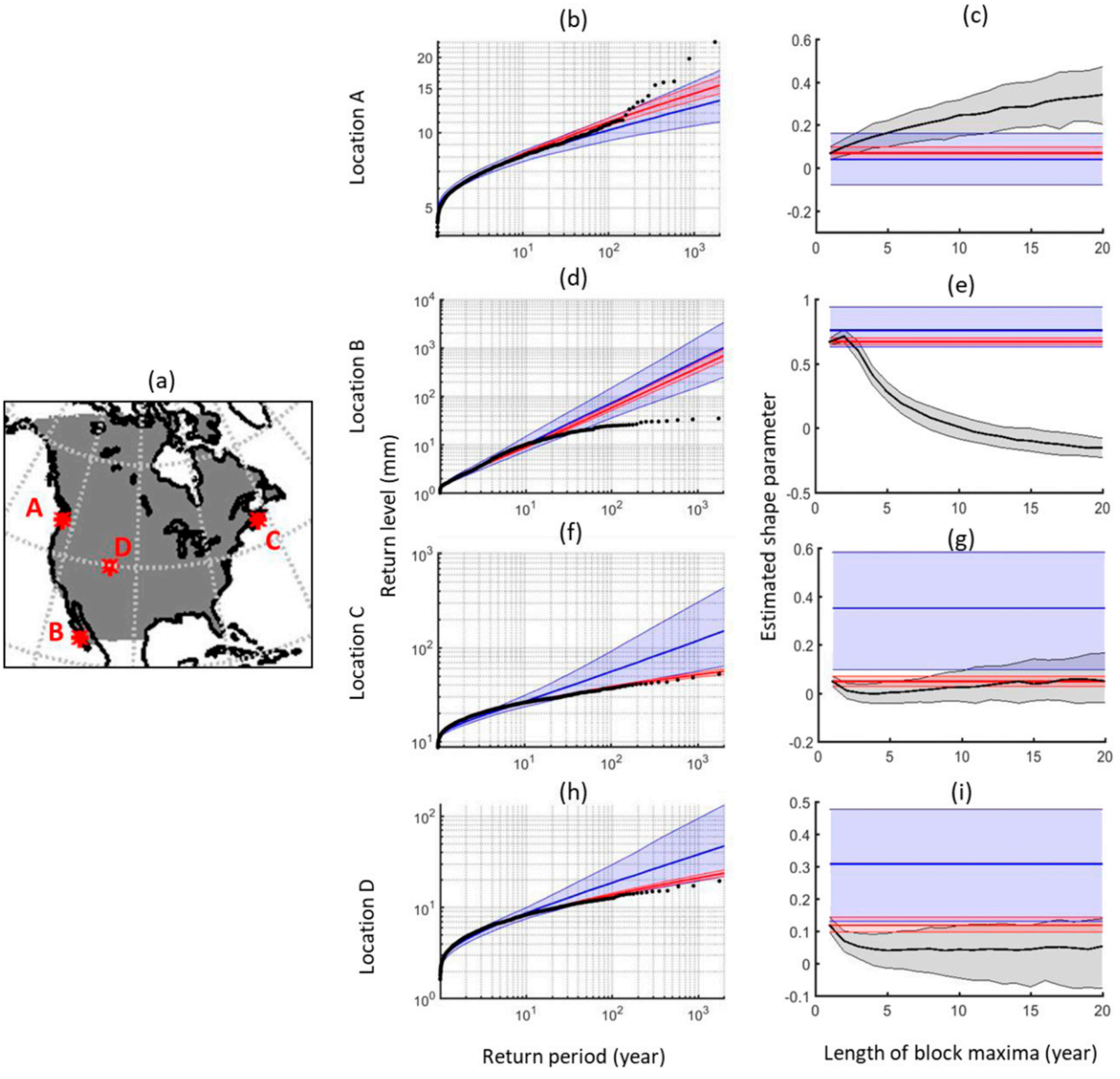

FIG. 4. (a) Geographical positions of the four locations A–D. Return-level estimates based on fitting the GEV distribution to annual maxima (using the ML method) at the four different locations (b) A, (d) B, (f) C, and (h) D using one CanRCM4 simulation of 1951–2000 (50 annual maxima, in blue) and the 35 simulations (1750 annual maxima, in red). Black dots show empirical quantile estimates obtained using the 1750 annual maxima. Also shown are estimates of the shape parameter vs block length based on 1750 years of CanRCM4 simulations are shown by the black line for the four locations (c) A, (e) B, (g) C, and (i) D. These panels also show estimated shape parameters based on annual maxima from a single CanRCM4 simulation (in blue) and the 35 ensemble members (in red), with the extension to longer blocks reflecting the max-stability assumption. Shading indicates 80% confidence intervals obtained by bootstrapping.

estimates based on fitting the GEV distribution to annual maxima are related to instability of the shape parameter. This includes (i) sampling variability, which affects both the empirical and GEV derived 1000-yr RL estimates; (ii) the possibility that the procedure used to produce the empirical RL estimate may induce bias that could cloud the comparison between the empirical and GEV based RL estimates; and (iii) the possibility that the method used to fit the GEV distribution may induce biases that cloud comparisons.

We deal with the impact of sampling variability first. Figure 5 displays, for each of the four locations, a box-and-whisker plot of 1000 paired differences between GEV-based 1000-yr RL estimates and the corresponding empirical 1000-yr RL estimates based on 1000 bootstrap samples of the 1750 annual maxima of CanRCM4

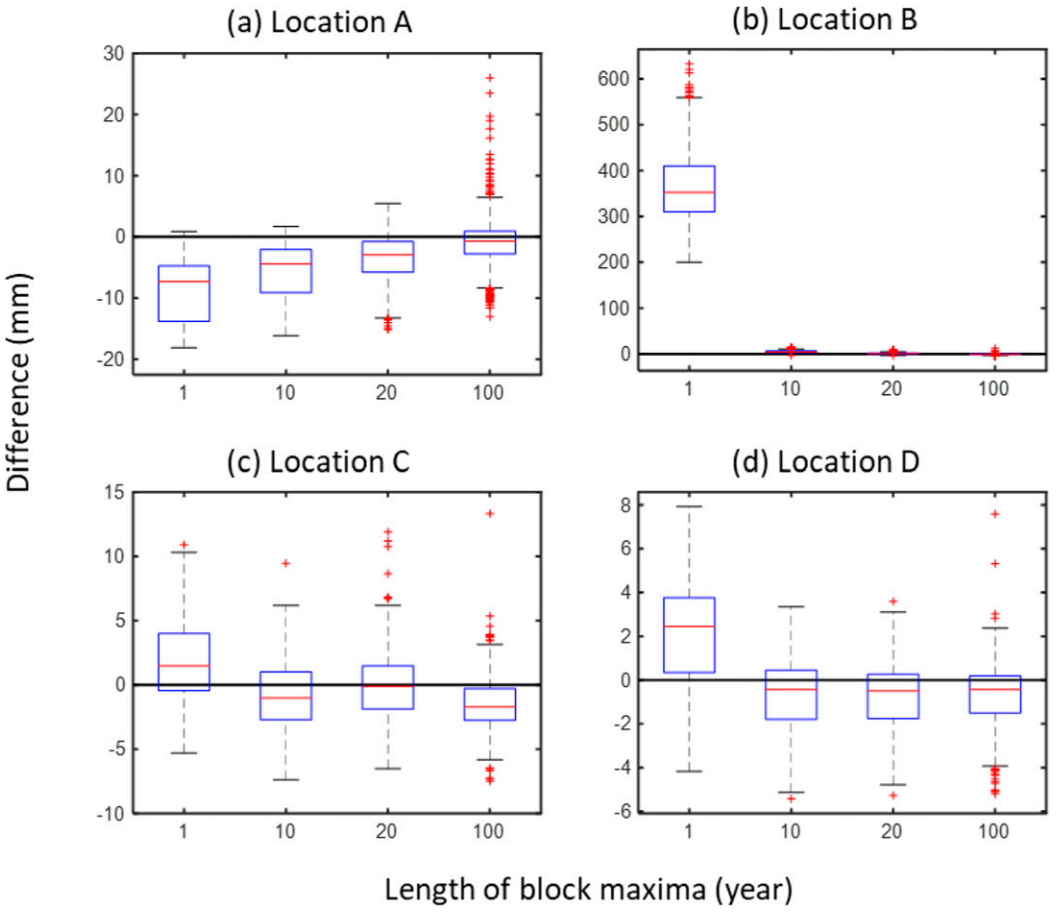

FIG. 5. Difference between GEV-based 1000-yr RL estimates for hourly precipitation and empirical 1000-yr RL estimates. Empirical estimates are derived from the pooled sample of 1750 annual maxima of CanRCM4 simulated hourly precipitation. GEV-based estimates are obtained from GEV distributions fitted via maximum likelihood to samples of block maxima of blocks of varying length, ranging from 1 yr (1750 blocks) to 100 yr (17 blocks). The differences were calculated 1000 times for each of 1000 bootstrap samples of 1750 annual maxima. Box-and-whisker plots display the distribution of the resulting bootstrap samples of differences for different block lengths at the four different locations (a) A, (b) B, (c) C, and (d) D (presented in Fig. 4a).

simulated hourly precipitation amounts using blocks of different lengths. For location A, the GEV-based RL estimates are consistently smaller than the empirical estimates when using 1- and 10-yr block lengths, and very frequently so when using 20-yr blocks, providing clear evidence that the GEV-based estimates are negatively biased relative to empirical estimates. Somewhat similar results are obtained at location B, except that in this case, the bias is in the opposite direction, and particularly marked when considering 1-yr blocks. In both cases, the sign of the bias and its evolution with block length is consistent with the evolution of the estimated GEV shape parameter seen in Figs. 4c and 4e respectively. At locations C and D we see that the bootstrap sampling distributions of the differences between the GEV-based and empirical RL estimates more consistently cover zero, indicating much weaker or nonexistent bias in the GEV-based estimate relative to the empirical RL estimates, which is consistent with the substantially weaker variation in the estimated GEV shape parameter with block length seen in Figs. 4g and 4i. Nevertheless, there is a relatively strong indication of bias at location D when considering 1-yr block maxima (Fig. 5d). Similar observations distinguishing between the behavior of apparent biases in places where there is a strong evolution of the estimated GEV shape parameter with block length, and places where that evolution is weaker, can be made for locations representative of all nine Bukovsky regions (see online supplemental Fig. S2 for locations, Figs. S4 and S6 for plots corresponding to Fig. 4, and Fig. S9 for bootstrap sampling distributions of the difference between GEV-based and empirical 1000-yr RL estimates).

A second concern flagged above is whether the empirical estimate of the 1000-yr RL is itself biased since

this estimate must be obtained at every location by interpolating between the largest two order statistics in the available sample of 1750 annual maxima. This bias is evaluated in Fig. 6 by first fitting a GEV distribution to the available sample of 1750 maxima, then repeatedly generating samples of 1750 values from the fitted distribution, next calculating an empirical 1000-yr RL estimate from each of the generated samples, and finally calculating the relative differences between those empirical estimates and the 1000-yr RL of the fitted distribution that was sampled. Figure 6 shows box-and-whisker plots of these relative differences assuming the fitted distributions at the four locations A, B, C, and D represent the ''truth.'' The median bias is negligible at all four locations. The errors in the empirical RL estimates are somewhat skewed and have particularly large spread at location B. A further analysis (Fig. S12 in the online supplemental material) considering representative locations in each of the nine Bukovsky regions (Fig. S2) also demonstrates that the expected bias in the empirical 1000-yr RL estimates is negligible, and that empirical quantile uncertainty can be particularly large in the desert region, where precipitation occurs relatively infrequently.

A third question is whether the particular method used to fit the GEV distribution induces a bias that could affect the comparison between the empirical and GEV-based RL estimates. We therefore recalculated Figs. 4–6 and supplemental Figs. S4, S6, S9, and S12 using the method of probability weighted moments instead of the ML method. The corresponding figures in the supporting information (Figs. S3, S8, S11, S5, S7, S10, and S13 in the online supplemental material) show that our findings are not sensitive to the choice of GEV fitting method.

Overall, it seems evident that reliable extrapolation to the far tail using a GEV distribution depends on both the accuracy of the estimated parameters and the validity of the stability assumption. Note, for example, that at locations A, B, and D, and also the majority of locations shown in supplemental Figs. S3–S13, the GEV distribution fitted via maximum likelihood using the 1750 annual maxima does not provide reliable estimates of extreme quantiles when the nonstability assumption seems to be not valid, even though in this case the large sample ensures small sampling uncertainty in the estimated GEV parameters. This includes the estimates of high quantiles that lie within the support of the large CanRCM4 simulated samples of annual maxima. This indicates that the behavior of extreme precipitation beyond annual maxima is very complex and cannot be simply summarized using a single extreme value distribution characterized by only three parameters.

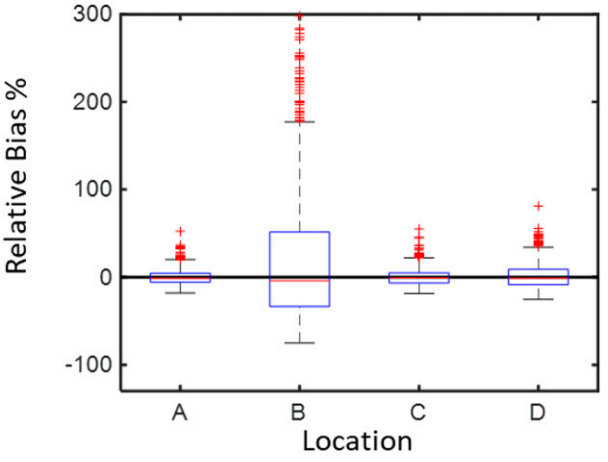

FIG. 6. Box-and-whisker plots illustrating the median, the interquartile range, and lower and largest values of relative biases of empirical 1000-yr RL estimates that are based on 1750 values sampled from four different known GEV distributions. The assumed known distributions are obtained by separately fitting GEV distributions via maximum likelihood to the 1750 annual maxima of hourly precipitation at the four locations A, B, C, and D (presented in Fig. 4a).

d. Further implications for RL estimates

We next compare maps of GEV-based 100- and 1000-yr RLs obtained using 1-yr and 10-yr blocks with empirical estimates of these RLs, in both cases using data from the 35 CanRCM4 simulations. As we can see, while the GEV distribution fitted to annual maxima provides estimates of the 100-yr RL that are comparable in magnitude to the empirical estimates over much of continent (Fig. 7a), the apparent bias in estimating the 1000-yr RL (Fig. 7b) cannot be neglected. These relative differences are spatially organized, suggesting physical origins, with substantial underestimation relative to the empirical estimates along much of the west coast of North America, and overestimation over most of the rest of the continent except in an area that stretches northward from the Gulf Coast into the central United States. These relative differences appear to be related to spatial variation in the shape parameter estimates that is seen in Fig. 1.

Figures 7c and 7d show that the bias of the GEV-based estimates relative to the empirical estimates is much reduced when the GEV distribution is fitted to 10-yr maxima. Figure 7c shows that the bias becomes negligible in the case of the 100-yr RL estimates. Its magnitude is also much reduced in the case of the 1000-yr RL estimates (Fig. 7d), for which the relative differences are noisy and without strongly apparent spatial organization, suggesting sampling variability as a dominant cause. Consistent with this, a GOF test assessing the fit

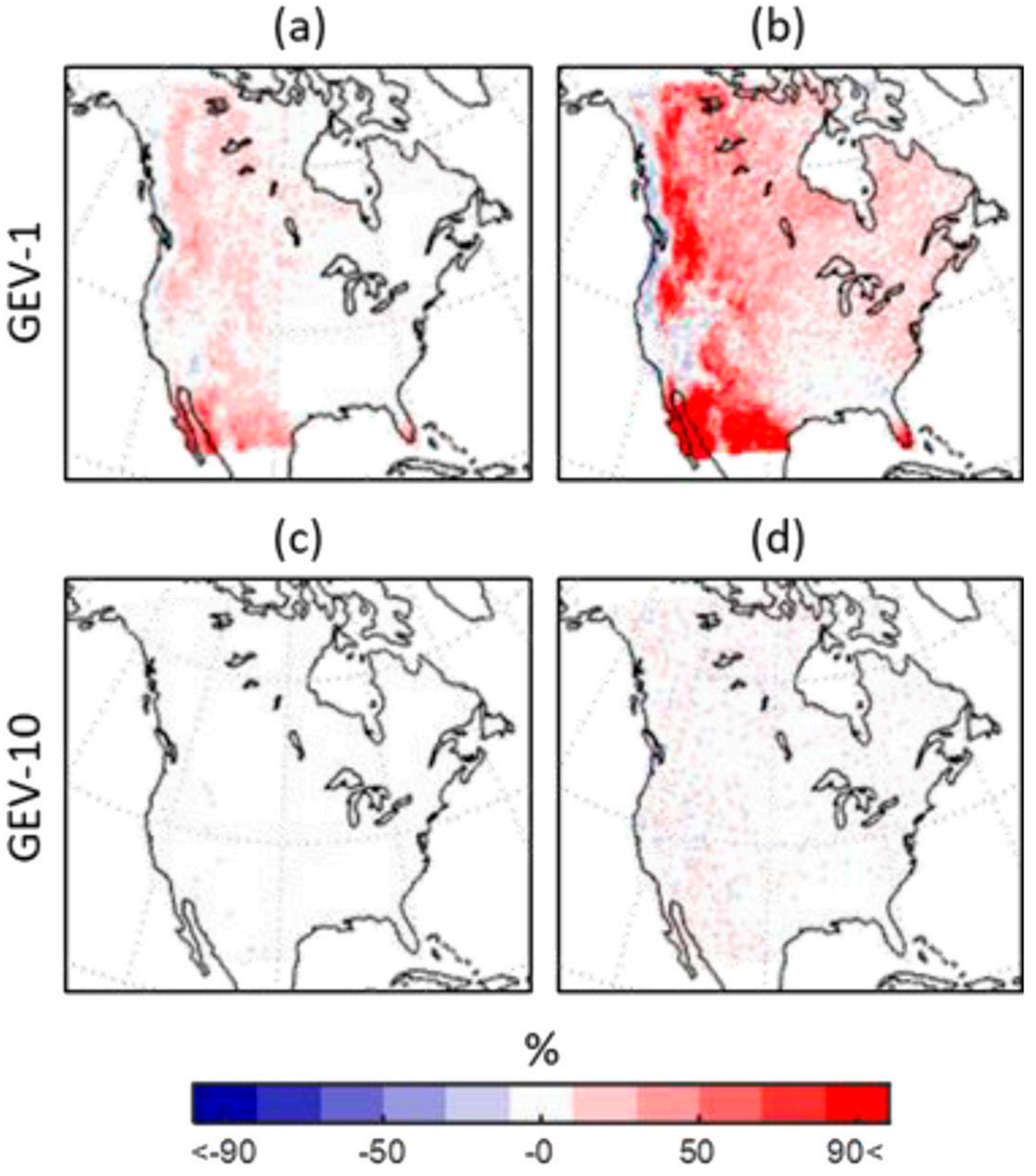

FIG. 7. Maps of the relative difference between GEV-based and empirical RL estimates for (left) 100-yr and (right) 1000-yr events. GEV-based estimates are obtained from GEV distributions fitted to (a),(b) annual maximum (GEV-1) and (c),(d) 10-yr maximum (GEV-10) hourly precipitation from the 35 CanRCM4 historical simulations of the period 1951–2000. Differences are expressed in percent relative to the corresponding empirical estimates.

of the GEV distribution to the available sample of 175 ten-year maxima (Fig. 3d) shows substantially reduced evidence of lack of fit, with rejection of the null hypothesis occurring at 9.1% of locations, which is nevertheless somewhat higher than the specified significance level of 5%.

Further evidence that the relative differences between GEV-based and empirical RL estimates may be related to changes in GEV shape parameter estimates that occur with increasing block length is shown in Fig. 8. It shows the differences between the shape parameter estimates obtained when fitting GEV distributions to annual and 10-yr block maxima of hourly precipitation. While noisy, the pattern of shape parameter differences is similar to that of the relative differences in 1000-yr RLs seen in Fig. 7b. Figure 9 links these two assessments of the effect of misfit when using annual maxima to fit the GEV distribution by plotting the relative bias in the

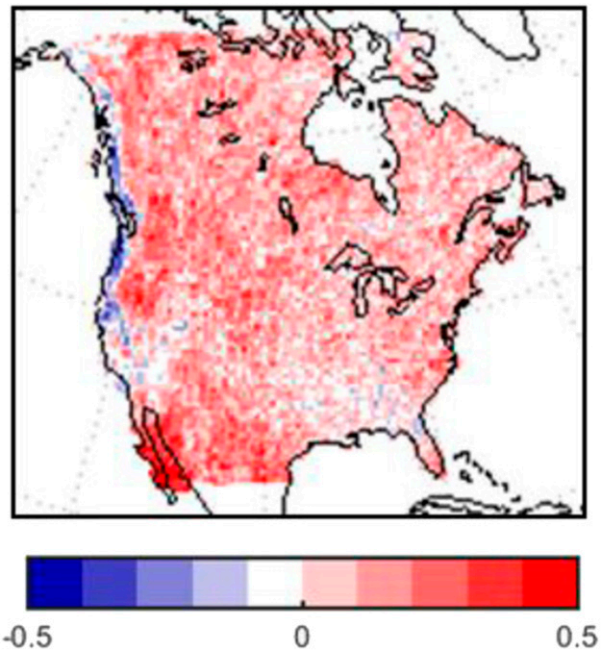

FIG. 8. Estimated difference between the shape parameters of GEV distributions fitted to annual maximum hourly precipitation and those of GEV distributions fitted to 10-yr maximum hourly precipitation for the 35 CanRCM4 historical simulations of the period 1951–2000.

100- and 1000-yr RL estimates as a function of D for all grid boxes over North America. As one penetrates deep into the upper tail of the extreme precipitation distribution (e.g., Fig. 9b), the magnitude of the change in estimated shape parameter becomes strongly predictive of the relative difference between GEV-based and empirical RL estimates.

e. On the role of physical processes

We alluded above to the possibility that the pattern of the relative differences between GEV-based and empirical RL estimates seen in Fig. 7b may have a physical basis. A disadvantage of univariate extreme value theory is that the behavior of extreme precipitation is depicted without knowledge of the physical processes that produce extremes. Instead, the additional information that allows extrapolation derives from mathematical postulates and assumptions. A key assumption is that data from the parent process must be independent and identically distributed (''iid''). While the independence assumption can be relaxed under certain conditions (e.g., Leadbetter et al. 1983), the identically distributed (''id'') assumption is likely problematic. Extreme precipitation events may be produced by a variety of physical components, with different processes producing extremes of different intensities at different frequencies. Statistically, one might therefore consider the upper tail of the unknown distribution of precipitation extremes as reflecting a mixture of distributions that are produced by different types of physical processes. If the element of the mixture producing the most extreme events occurs only rarely, it would likely be necessary to use blocks that are long enough to consistently sample events from that rarely occurring process to ensure that the sample data exhibit max-stability. This indeed appears to be the situation at location A, for example (Figs. 4b,c), where there is evidence that the largest extremes are associated with rare, very intense, atmospheric rivers (Fig. S14 in the online supplemental material), which are well simulated by the global model CanESM2 that drives CanRCM4 (Tan et al. 2019) and also well represented in CanRCM4 (Whan and Zwiers 2016).

Consideration of mixture distributions in the hydrological literature dates back at least to Waylen and Woo (1982), who used mixtures to model floods arising from different processes; to Rossi et al. (1984), who used a two-component extreme value distribution; and to more recent works such as that of Barth et al. (2019). To illustrate that the variations in shape parameter with block length observed at locations A–D shown in Fig. 4a might be consistent with the notion that extremes result from a mixture of processes, we fitted a mixture of two GEV distributions of the form $F(x) = \omega \times F_1(x; \mu_1, \sigma_1, \xi_1) + (1 - \omega) \times F_2(x; \mu_2, \sigma_2, \xi_2)$ to the 1750 annual maxima of hourly precipitation accumulations at each of those four locations, where for $i = 1, 2$, $F_1(x; \mu_i, \sigma_i, \xi_i)$ is a GEV distribution with location, scale, and shape parameters μ_i, σ_i and ξ_i, respectively, and ω is a weight parameter. The mixture distribution parameters were estimated via the ML method to obtain the following four mixture distributions $F_A(x) = 0.17 \times F_1(x; 5.6, 0.91, 0.4) + 0.83 \times F_2(x; 6.01, 0.92, -0.05)$, $F_B(x) = 0.21 \times F_1(x; 7.12, 3.96, 0.15) + 0.79 \times F_2(x; 1.93, 0.66, 0.57)$, $F_C(x) = 0.9 \times F_1(x; 15.47, 4.01, 0.08) + 0.1 \times F_2(x; 22.8, 4, -0.5)$, $F_D(x) = 0.56 \times F_1(x; 5.33, 1.69, 0.06) + 0.44 \times F_2(x; 3.28, 0.89, -0.02)$ at locations A, B, C, and D respectively. At each location we then simulated 1000 series of 1750 values from the fitted mixture distribution and fitted simple GEV distributions via maximum likelihood to the simulated samples and to maxima over longer blocks. Figure 10 illustrates the evolution of the resulting sample of shape parameters with block length. As we can see, fluctuation of the shape parameter beyond annual maxima for samples from the mixture distribution is reminiscent of that seen in Figs. 4c, 4e, 4g, and 4i.

Note that our intention is not to demonstrate that a mixture distribution is better than a single GEV for extrapolation, but merely to illustrate one possibility for

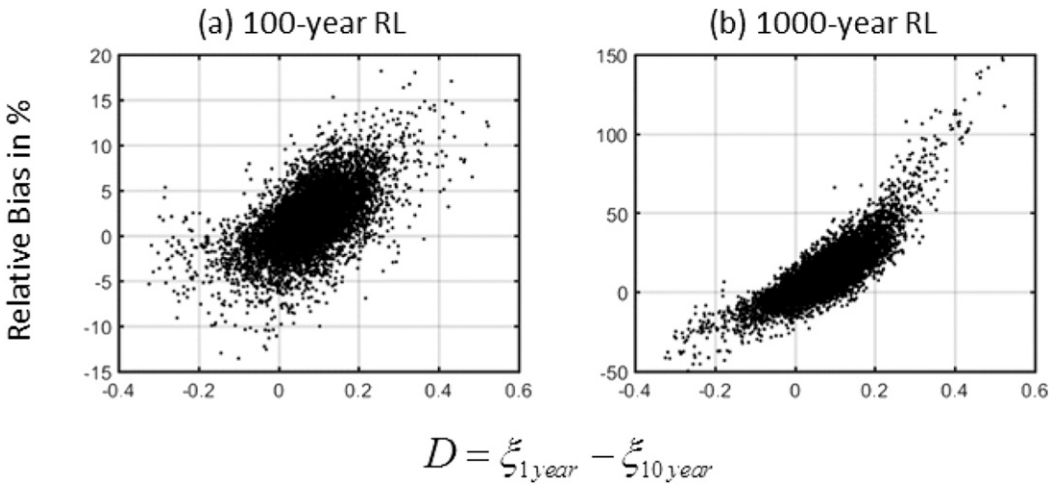

FIG. 9. Biases in GEV-based (a) 100- and (b) 1000-yr RLs estimates relative to empirical RL estimates based on samples of 1750 annual maxima of CanRCM4 simulated hourly precipitation shown as a function of the change D in the estimated GEV shape parameter when the distribution is fitted to 1- and 10-yr block maxima.

how the apparent lack of max-stability of annual maxima could arise.

5. Conclusions

A key assumption in using extreme value theory to estimate the intensity and frequency of rare events not observed in the instrumental record is that the observed process produces extremes with a distribution with an upper right-hand tail that has a stable rate of decay. This central assumption has been flagged in previous literature as being of concern, but is often not considered in applications of EVT, in part because its validity is difficult to assess with instrumental records that are generally no more than a few decades long.

We have examined the validity of the stability assumption in the climate simulated by a modern regional climate model of North America, CanRCM4, for which a large ensemble of simulations from which hourly precipitation has been archived is available. This ensemble provides 1750 years of simulated climate data that are consistent with forcing conditions that prevailed during the latter half of the twentieth century, and thus provides a realistic test bed for assessing the stability assumption. By considering block maxima of extreme precipitation for durations of 1, 6, 12, and 24 h and blocks of length varying from 1 to 20 years, we show that the model-simulated extremes exhibit a lack of max-stability across large parts of North America. Goodness-of-fit tests show that the GEV distribution does not fit the large sample of annual maxima of hourly precipitation extremes well across much of the continent, presumably because a distribution that has max-stability as an inherent property is being fitted to data that are not max-stable. The fit is seen to improve in many locations when the GEV distribution is fitted to 10-yr maxima rather than annual maxima. Lack of max-stability is also evident from the changes in the estimated GEV shape parameter that are noted as block length increases. Shape parameter estimates tend to stabilize as block length increases, which is suggestive of the possibility that block maxima for sufficiently long blocks may be max-stable in the climate of CanRCM4. Evidence of a lack of max-stability is weaker for events of longer duration, but nevertheless it cannot be ignored.

An implication of this apparent lack of max-stability is that long-period RL estimates based on GEV distributions fitted to annual maxima may be seriously biased, even when very large samples of annual maxima are available. We showed that bias in 100- and 1000-yr RLs is substantially reduced when block length is increased, suggesting that sampling events deep enough in the upper tail could ultimately, identify a point above which stable decay does occur, at least in the climate simulated by CanRCM4. We argue, as others have also done, that the lack of stability has physical origins, reflecting the fact that extreme precipitation at any given location may be produced by a number of different physical processes with different relative frequencies of occurrence. While it is not practical in operational analyses to increase block length so as to sample information from deeper in the upper tail, data pooling within an assumed homogenous region through an appropriate statistical model of spatial extremes may be helpful (Davison et al. 2012). Another approach would be to use additional meteorological data to decompose precipitation into physically

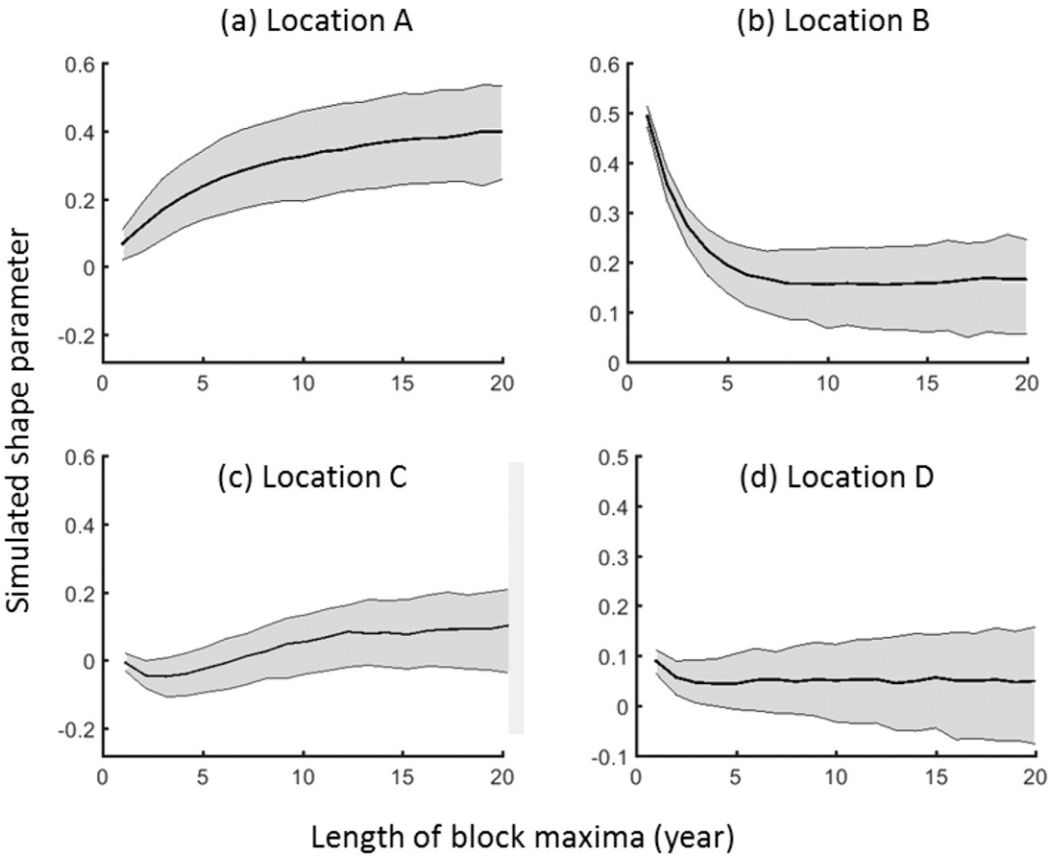

FIG. 10. Estimated shape parameters as function of block length for simple GEV distributions fitted to samples simulated from a mixture of two GEV distributions fitted to the 1750 annual maxima of CanRCM4 simulated hourly precipitation at the four locations (a) A, (b) B, (c) C, and (d) D presented in Fig. 4a. Shading indicates 80% uncertainty intervals on the basis of 1000 samples of 1750 values simulated from the fitted mixture distribution.

interpretable constituent components so as to model how the variation in the upper tail of the precipitation distribution varies with one or more of those components. A study suggesting such an approach will be published separately.

It should be borne in mind that numerous caveats apply to our findings. First, given its approximately 50-km spatial resolution, the regional climate model used in this study is not expected to be able to simulate all of the phenomena that, in nature, are responsible for extreme precipitation. For example, the model uses parameterizations that describe the impact of convection on the atmospheric state at the scales that resolves, and thus individual convective events are not simulated. Also, while 50-km-resolution models are able to simulate tropical cyclone–like features, the global model that is used to drive CanRCM4, CanESM2 with its T63 horizonal resolution, has insufficient resolution to simulate tropical cyclones. It is therefore likely that CanRCM4 significantly undersimulates the effects of tropical cyclone activity, if at all. Second, our analysis has implicitly assumed that the climate is stationary during the latter half of the twentieth century, in contrast with repeated assessments (IPCC 2013) and an extensive climate change detection and attribution literature indicating that this is not the case. Studies of observed changes in extreme 1-day and 5-day precipitation amounts suggest, when considering global land data, that human influence has intensified such events at a rate close to the Clausius–Clapeyron rate (Min et al. 2011; Zhang et al. 2013). Nevertheless, nonstationarity remains difficult to detect in local historical precipitation records (Sun et al. 2020, manuscript submitted to *J. Climate*; Westra et al. 2013) and cannot be reliably accounted for in the analysis of individual historical records (Li et al. 2019). While we recognize that the stationarity assumption is not satisfied, given the weak and difficult to discern influence of nonstationarity due to global warming on local precipitation records, departures from stationarity over the period considered in this paper may have a smaller effect on estimated long period RLs than departures from the stability assumption.

Acknowledgments. Author Ben Alaya was supported by the Climate Related Precipitation Extremes project of the Global Water Futures program. We also very much appreciate the assistance provided by Guilong Li in analyzing individual extreme precipitation events in the CanRCM4 large ensemble. We thank Dr. Francesco Serinaldi and two anonymous reviewers for their constructive and insightful comments, which helped us to improve this paper.

REFERENCES

Barth, N. A., G. Villarini, and K. White, 2019: Accounting for mixed populations in flood frequency analysis: Bulletin 17C perspective. *J. Hydrol. Eng.*, **24**, 04019002, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001762.

Bukovsky, M., 2012: Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective. Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, accessed 15 March 2018, http://www.narccap.ucar.edu/contrib/bukovsky/.

Caires, S., and A. Sterl, 2005: 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *J. Climate*, **18**, 1032–1048, https://doi.org/10.1175/JCLI-3312.1.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values.* Springer, 208 pp.

Davison, A. C., S. A. Padoan, and M. Ribatet, 2012: Statistical modeling of spatial extremes. *Stat. Sci.*, **27**, 161–186, https://doi.org/10.1214/11-STS376.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775–779, https://doi.org/10.1038/nclimate1562.

Fisher, R. A., and L. H. C. Tippett, 1928: Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Cambridge Philos. Soc.*, **24**, 180–190, https://doi.org/10.1017/S0305004100015681.

Giorgi, F., C. Jones, and G. R. Asrar, 2009: Addressing climate information needs at the regional level: The CORDEX framework. *WMO Bull.*, **58**, 175–183.

Huang, W. K., M. L. Stein, D. J. McInerney, S. Sun, and E. J. Moyer, 2016: Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **2**, 79–103, https://doi.org/10.5194/ascmo-2-79-2016.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis.* Cambridge University Press, 1535 pp., https://doi.org/10.1017/CBO9781107415324.

Kharin, V. V., and F. W. Zwiers, 2005: Estimating extremes in transient climate change simulations. *J. Climate*, **18**, 1156–1173, https://doi.org/10.1175/JCLI3320.1.

Klemeš, V., 1986: Dilettantism in hydrology: Transition or destiny? *Water Resour. Res.*, **22**, 177S–188S, https://doi.org/10.1029/WR022I09SP0177S.

——, 1987: Hydrological and engineering relevance of flood frequency analysis. *Hydrologic Frequency Modeling*, Springer, 1–18.

——, 2000: Tall tales about tails of hydrological distributions. I. *J. Hydrol. Eng.*, **5**, 227–231, https://doi.org/10.1061/(ASCE)1084-0699(2000)5:3(227).

Kotz, S., and S. Nadarajah, 2000: *Extreme Value Distributions: Theory and Applications*. World Scientific, 196 pp.

Lang, M., T. Ouarda, and B. Bobée, 1999: Towards operational guidelines for over-threshold modeling. *J. Hydrol.*, **225**, 103–117, https://doi.org/10.1016/S0022-1694(99)00167-5.

Leadbetter, M. R., G. Lindgren, and H. Rootzén, 1983: *Extremes and Related Properties of Random Sequences and Processes*. Springer, 336 pp.

Li, C., F. Zwiers, X. Zhang, and G. Li, 2019: How much information is required to well constrain local estimates of future precipitation extremes? *Earth's Future*, **7**, 11–24, https://doi.org/10.1029/2018EF001001.

Min, S.-K., X. Zhang, F. W. Zwiers, and G. C. Hegerl, 2011: Human contribution to more-intense precipitation extremes. *Nature*, **470**, 378–381, https://doi.org/10.1038/nature09763.

Rossi, F., M. Fiorentino, and P. Versace, 1984: Two-component extreme value distribution for flood frequency analysis. *Water Resour. Res.*, **20**, 847–856, https://doi.org/10.1029/WR020i007p00847.

Scinocca, J., and Coauthors, 2016: Coordinated global and regional climate modeling. *J. Climate*, **29**, 17–35, https://doi.org/10.1175/JCLI-D-15-0161.1.

Smith, R. L., 2003: Statistics of extremes, with applications in environment, insurance, and finance. *Extreme Values in Finance, Telecommunications, and the Environment*, Chapman and Hall/CRC, 20–97.

Tan, Y., F. Zwiers, S. Yang, C. Li, and K. Deng, 2019: The role of circulation and its changes in present and future atmospheric rivers over western North America. *J. Climate*, **33**, 1261–1281, https://doi.org/10.1175/JCLI-D-19-0134.1.

Waylen, P., and M. Woo, 1982: Prediction of annual floods generated by mixed processes. *Water Resour. Res.*, **18**, 1283–1286, https://doi.org/10.1029/WR018i004p01283.

Westra, S., L. V. Alexander, and F. W. Zwiers, 2013: Global increasing trends in annual maximum daily precipitation. *J. Climate*, **26**, 3904–3918, https://doi.org/10.1175/JCLI-D-12-00502.1.

Whan, K., and F. Zwiers, 2016: Evaluation of extreme rainfall and temperature over North America in CanRCM4 and CRCM5. *Climate Dyn.*, **46**, 3821–3843, https://doi.org/10.1007/s00382-015-2807-7.

Zhang, X., H. Wan, F. W. Zwiers, G. C. Hegerl, and S. K. Min, 2013: Attributing intensification of precipitation extremes to human influence. *Geophys. Res. Lett.*, **40**, 5252–5257, https://doi.org/10.1002/grl.51010.