A calibration method for projecting future extremes via a linear mapping of parameters

Jeongjin Lee[1*], Daniel Cooley[2], Anna M. Wagner[3], Glen E. Liston[4]

[1*]Department of Mathematics and Statistics, Lancaster University, Fylde College, Lancaster, LA1 4YF, United Kingdom.

[2]Department of Statistics, Colorado State University, Fort Collins, 80523, Colorado, United States of America.

[3]U.S. Army Engineer Research and Development Center, Cold Regions Research and Engineering Laboratory, Fort Wainwright, Alaska, United States of America.

[4]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, United States of America.

*Corresponding author(s). E-mail(s): j.lee58@lancaster.ac.uk;
Contributing authors: cooleyd@stat.colostate.edu;
Anna.M.Wagner@erdc.dren.mil; Glen.Liston@colostate.edu;

Abstract

In order to study potential impacts arising from climate change, future projections of numerical model output often must be calibrated to be comparable to observations. Rather than calibrating the data values themselves, we propose a novel statistical calibration method for extremes that assumes there exists a linear relationship between parameters associated with model output and parameters associated with observations. This approach allows us to capture uncertainty in both parameter estimates and the linear calibration, which we achieve via bootstrap. To focus on extreme behavior, we assume both model output and observations have distributions composed of a mixture model combining a Weibull distribution with a generalized Pareto distribution for the tail. A simulation study shows good coverage rates. We apply the method to project future daily-averaged river runoff at the Purgatoire River in southeastern Colorado.

Keywords: climate projections, downscaling, flooding, extremes, calibration

1 Introduction

Numerical models are widely used in the Earth sciences to study atmospheric and ocean dynamics, hydrology, atmospheric chemistry, and other processes. Because they are driven by the known physics of the studied system, numerical models are the best tools available to produce projections under possible climate scenarios. However, numerical model output can have notable discrepancies from observations, which can have implications for quantifying potential climate change impacts. These discrepancies (or 'model bias' in the climate literature) can occur because the model's spatial support (often a grid cell) differs from that of point-referenced observations and/or because the model's physics are necessarily a simplification and cannot capture the full complexity of the Earth system. To better understand the behavior of quantities of interest in relation to observations, there is a need to calibrate (or downscale) model output. Various statistical calibration and bias correction methods have been developed to address these issues. We provide an overview of widely applied methods in Section 2.2. As part of their review, Teutschbein and Seibert (2012) compare different correction methods specifically for regional climate model (RCM) simulations in hydrological impact studies.

The calibration problem can be visualized as in Figure 1. The boxes with green checkmarks have data (either model output or observations) available. The red 'X' indicates there are no observations under the projected climate. Quantities of interest related to projected observations must be estimated based on a modeled relationship between model output and observations learned from the historical period, which is then applied to the projected period.

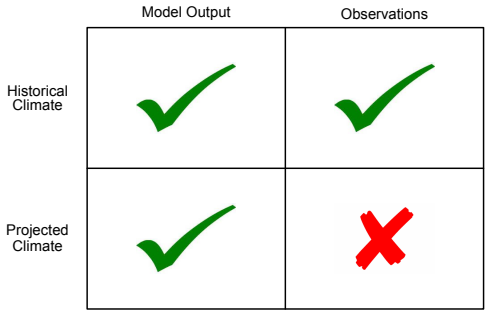


Fig. 1: Illustration of the calibration method applied to projected observations under the projected climate. Boxes with green checkmarks have data available and quantities of interest can be estimated directly. The red 'X' indicates there are no observations under the projected climate, and quantities of interest must be estimated based on the relationship between model output and observations, and the relationship between historical and projected climate.

Our particular calibration study is motivated by a project which aims to estimate potential flood risk from the Purgatoire River to infrastructure at a military base

in southeastern Colorado, USA. For this river, flows are often greatest in the spring due to runoff from melting snow, and it is of interest to know how climate-induced changes to the timing and duration of snow cover could affect flood risk. Notably, discrepancies between model output and local observations are amplified when focusing on extremes, as shown in Figure 2. With the growing attention on climate change impacts, numerous studies have contributed to the development of calibration methods for extremes to better assess their effects on local extremes (e.g., Schubert and Henderson-Sellers, 1997; Vrac and Naveau, 2007; Benestad, 2010). In a comparison of advanced downscaling methods, Bürger et al. (2012) evaluated approaches for extremes, including automated regression-based statistical downscaling (ASD) (Hessami et al., 2008), bias correction spatial disaggregation (BCSD) (Wood et al., 2002), and quantile regression neural networks (QRNN) (Taylor, 2000). In the context of machine learning techniques, Campozano et al. (2016) compared statistical downscaling methods with two machine learning methods, specifically artificial neural network (ANN) and least squares support vector machines (LA-SVM), to evaluate downscaled general circulation model (GCM) estimates of monthly precipitation.

In this study, given the limited and relatively short data records, and without the inclusion of additional predictor variables, our calibration method is classified as a transfer function approach, as opposed to stochastic weather generators and weather typing methods (Vrac and Naveau, 2007). Accordingly, in Section 4.3, we focus on comparing calibration methods that emphasize the direct relationships between large-scale model output and local observations. We propose a novel univariate calibration method for extremes, with the development of a multivariate version left for future work.

Our primary aim is to provide estimates of high quantiles of projected Purgatoire River observations. In particular, we wish to provide estimates (with uncertainty) of quantiles roughly corresponding to the 1-in-10 and 1-in-100 year events, the latter of which will require extrapolation into the tail as the data records we employ are much shorter. The need for extrapolation leads us to employ a parametric model. Our model, which will be fit to the entire distribution, will rely on an extreme value model to capture the behavior in the upper tail. Unlike the advanced downscaling methods referenced in Bürger et al. (2012), where data outside the range of the fitted quantiles are extrapolated using specific distributions such as Weibull or exponential distributions, we fit a generalized Pareto distribution to the tail, allowing for more flexibility in capturing different tail behaviors.

Our calibration method assumes there exists a linear transfer function governing the relationship between the parameters of the distributions describing model output and the observations. The historical period will be used to estimate this linear relationship, and then it is applied to the parameters of the projected model output to obtain an estimate of the distribution of projected observations. In addition to accounting for the uncertainty associated with parameter estimates, a bootstrap method will allow us to also account for uncertainty associated with estimating the transfer function.

In contrast to dynamical downscaling using RCMs, statistical calibration requires thorough validation, especially for extremes, to verify the methods as noted by (Bürger et al., 2012). We perform validation using the historical period, given that observations

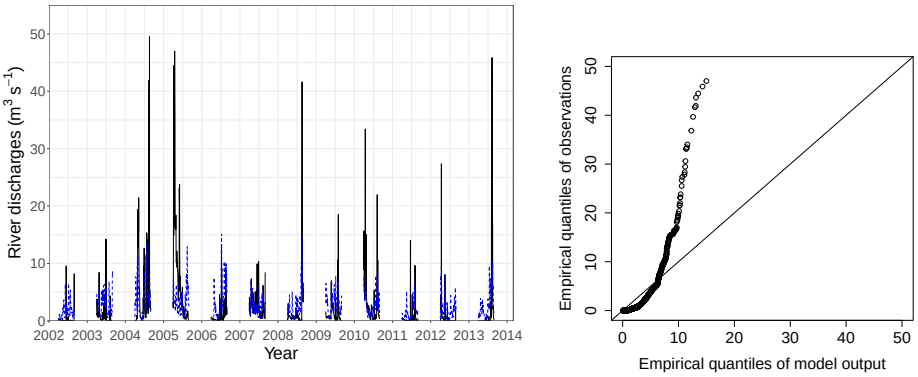
for projected climates are not available. To evaluate the performance of calibration methods for extremes, we use quantile-quantile plots (QQ-plots), a standard tool in extreme value analysis, and summary statistics derived from large values exceeding a high threshold.

Section 2 describes the data and reviews common statistical calibration and bias correction methods. In Section 3.1, we motivate and describe our model for both observed and modeled river flows. Section 3.2 describes the process of our calibration method. In Section 3.3, we describe how we obtain uncertainty estimates for the model parameters, transfer function, and estimated quantiles. We then present simulation results, case study, and method comparisons in Section 4. Finally, we conclude with a summary and discussion.

2 Data and other calibration methods

2.1 Data description

Throughout the study, we analyze daily-averaged river flow observations and river flow model output for the Purgatoire River in Colorado covering the period from 2002 to 2013. Projected river flow model output is also produced for this period under a projected climate (not shown in Figure 2). Complete measurements are taken across three datasets, resulting in a sample size of $n = 1,836$. The superimposed timeseries plot of the observations and model output for the historical period is shown in Figure 2a. Flows are only shown for the period between April and August as this is the period when the river is at risk for flooding, and river flow measurements outside these months are considered unreliable.



(a) (b)

Fig. 2: (a) Timeseries plot of daily-averaged river discharges (black solid line) and modeled river discharges (blue dashed line) for the Purgatoire River in Colorado for 2002-2013. (b) QQ-plot comparing empirical quantiles of observations to empirical quantiles of model output for the same historical period.

The process to produce modeled river flows is quite involved. Two simulations were forced with a high-resolution (4km grid) dataset obtained from the Weather Research and Forecasting (WRF) model. The current climate simulation was forced with ERA-Interim reanalysis data (Dee et al., 2011) for the period from October 2000 to September 2013. The projected pseudo-global warming simulation (PGW) (a perturbation experiment) for the same period is forced with ERA-Interim reanalysis and a climate perturbation (Rasmussen et al., 2011). This perturbation reflects the 95-year mean change signal from the Coupled Model Intercomparison Project Phase 5 (CMIP5) multi-model ensemble under the high-emissions RCP8.5 scenario. The ensemble-mean monthly climate change was derived from 19 CMIP5 models selected based on their performance, as detailed in Section 2.2. of Liu et al. (2017).

This high-resolution modeled weather output was used as input for SnowModel (Liston and Elder, 2006; Liston et al., 2020), a numerical model for the accumulation, evaporation, and melting of snow over a study area. Finally, modeled snow runoff was combined with WRF-produced meteorology and input into HydroFlow (Liston and Mernild, 2012), a numerical hydrological model which produces simulated streamflow measurements.

The modeled stream runoff is useful for understanding changes in timing and relative runoff amounts between the historical and projected periods, but there is a clear mismatch between the distributions of the modeled river flows and the observations. In particular, the extremes of the observations are not represented by the modeled runoff. Calibration is needed to use the modeled river flow output for assessment of projected flood risk.

2.1.1 Model evaluation

To assess the goodness of fit for numerical models such as SnowModel, several summary statistics can be used. We consider the determinant coefficient, $R^2 \in [0, 1]$, where values close to 1 indicate a better fit; root mean squared error, RMSE ≥ 0, with lower values implying a better fit; and Nash-Sutcliffe efficiency coefficient (Nash and Sutcliffe, 1970), NSE $\in (-\infty, 1]$, which reflects the proportion of the variance in the observations that is accounted for by the model relative to the total variance of the observations

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n} \left(x_{Mod}(t) - x_{Obs}(t)\right)^2}{\sum_{i=1}^{n} (x_{Obs}(t) - \bar{x}_{Obs})^2},$$

where $x_{Mod}(t)$ and $x_{Obs}(t)$ are the model output and observations at time $t = 1, \ldots, n$, respectively, and $\bar{x}_{Obs}$ denotes the sample mean of the observations. An NSE of 1 indicates a perfect match between model output and observations. While these measures are typically useful for evaluating mean behavior, our focus is on high-quantiles. To make these statistics more relevant for extremes, we also consider large values exceeding a high threshold. Setting the high threshold at the 0.95 quantile for each dataset of observations and model output for the historical period, the obtained statistics are summarized in Table 1. As a graphical diagnostic, we create a QQ-plot of observations versus modeled river flows in Figure 2b, showing a clear mismatch in the higher quantiles.

	R^2	RMSE	NSE
SnowModel	0.14 / 0.10	4.64 / 13.85	-0.014 / -0.524

Table 1: Summary statistics for SnowModel under the historical climate. Values before the slash are derived from all data points, while values after the slash are derived from data exceeding a high threshold.

2.2 Statistical calibration and bias corrections

We review commonly applied statistical calibration and bias correction methods. Let $\boldsymbol{x}^h_{Mod}(t)$ denote the variables from model output for the historical period at time t, let $x^h_{Obs}(t)$ be the observations of variable of interest from the historical period, and let $\boldsymbol{x}^p_{Mod}(t)$ represent the model output for the projected period. Statistical calibration methods take these available data to attempt to describe $x^p_{Obs}(t)$, the unavailable observations for the projected period. The model output used does not have to represent the projected variable of interest; $\boldsymbol{x}^p_{Mod}(t)$ might be larger-scale predictors, for example principal components of geopotential heights (e.g., Hanssen-Bauer et al., 2005). Huang et al. (2019) say that most proposed statistical calibration methods fall into three general categories: regression, a shift/scale (or delta method) approach, or quantile mapping. We note that the choice of bias correction method often depends on the specific motivations, data characteristics, and underlying assumptions of the study.

2.2.1 Regression-based method

A generic form of regression is assumed as

$$x^h_{Obs}(t) = f(\boldsymbol{x}^h_{Mod}(t), \boldsymbol{\beta}) + \epsilon(t).$$

Often the function f is standard linear regression. Calibration via regression makes the most sense when the observations and model output are synchronous; that is, the modeled weather at time t represents the actual weather at t. Climate reanalysis data is synchronous, but output from general circulation models is typically not. When the data is asynchronous, then the distribution of the model output needs to be related to the distribution of the observations. A notable study that uses the regression-based downscaling method includes (Wilby et al., 1999).

2.2.2 A shift/scale method (delta method)

A simple shift/scale approach is moment based. Letting $X^h_\cdot(t)$ and $X^p_\cdot(t)$ be the random variables representing the variable of interest in the historical period and projected period respectively and $\cdot$ can be either model output or observations, the shift/scale approach assumes

$$X^p_\cdot(t) = s(X^h_\cdot(t) + m).$$

Thus, the projected distribution function is shifted by m and scaled by s:

$$F_{X^p(t)}(x) = F_{X^h_\cdot(t)}((x-m)/s).$$

The parameters m and s are learned from the historical period and applied to the projected period. As an example, a univariate version of linear bias correction for climate model output by Bürger et al. (2012) is

$$\hat{x}^p_{Obs}(t) = \left(\frac{x^p_{Mod}(t) - \bar{x}^p_{Mod}}{\sigma^h_{Mod}}\right)\sigma^h_{Obs} + (\bar{x}^p_{Mod} - \bar{x}^h_{Mod}) + \bar{x}^h_{Obs}, \tag{1}$$

where $\bar{x}^p_{Mod}$, $\bar{x}^h_{Mod}$, and $\bar{x}^h_{Obs}$ represent the sample means of the model output for the project period, the model output for the historical period, and the observations for the historical period, respectively, and σ^h_{Obs} and σ^h_{Mod} are the standard deviations of observations and model output for the historical period. A study by Teutschbein and Seibert (2012) compares various shift and scale corrections, including linear scaling, variance scaling, power transformation, and the delta-change method.

This shift/scale method is simple and m and s are easily estimated, but is best suited for understanding changes in the center or bulk of the distribution and is of limited value for detecting changes in extremes.

2.2.3 Quantile mapping (QM)

Quantile mapping (QM) defines the transfer function that connects cumulative distribution functions (CDFs) of F^h_{Obs} and F^h_{Mod}

$$\hat{x}^p_{Obs}(t) = \widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^h_{Mod}[x^p_{Mod}(t)]\}, \tag{2}$$

where $\hat{x}^p_{Obs}(t)$ corresponds to the projected observations at time t within the projected period, $\widehat{F}^h_{Obs}$ and $\widehat{F}^h_{Mod}$ are estimated CDFs of observations and model output under historical climate. As the qunatile mapping transfer function uses only the information from the historical period, it can fail for extreme values if $x^p_{Mod}(t)$ falls outside the range of values used to estimate $\widehat{F}^h_{Mod}$. Hence, extrapolation is required for extremes.

2.2.4 Quantile delta mapping (QDM)

Instead of using the direct extrapolation, there have been methods such as equidistant and equiratio quantile mapping (Li et al., 2010; Wang and Chen, 2014), which use the information from the CDF of the projected model output, denoted by F^p_{Mod}, for the projected period. Cannon et al. (2015) verified that those quantile mapping approaches are equivalent to the quantile mapping of the 'delta change method' (Olsson et al., 2009) and termed this approach 'quantile delta mapping' (QDM).

QDM preserves relative changes (or deltas) in all quantiles between the projected model output $x^p_{Mod}(t)$ and historical model output $x^h_{Mod}(t)$. The first step involves detrending the projected model output by quantile, followed by bias correction of the projected model output to historical observations via quantile mapping. After that,

the relative changes, denoted by $\Delta_{Mod,Rel}(t)$, in the quantiles of the projected model output are applied to the bias-corrected historical values to capture the projected climate change signal (or relative changes in modeled quantiles)

$$\begin{aligned}\hat{x}^p_{Obs}(t) &= \widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\} \times \Delta_{Mod,Rel}(t) \\ &= \widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\} \times \frac{x^p_{Mod}(t)}{\widehat{F}^{h^{-1}}_{Mod}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\}}\end{aligned} \tag{3}$$

To preserve absolute changes, denoted by $\Delta_{Mod,Abs}(t)$, in quantiles (e.g., temperatures in Celsius, see Cannon et al., 2015), additive deltas are applied to historical bias-corrected values. Specifically, the adjusted quantile is given by $\widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\}$

$$\begin{aligned}\hat{x}^p_{Obs}(t) &= \widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\} + \Delta_{Mod,Abs}(t) \\ &= \widehat{F}^{h^{-1}}_{Obs}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\} + x^p_{Mod}(t) - \widehat{F}^{h^{-1}}_{Mod}\{\widehat{F}^p_{Mod}[x^p_{Mod}(t)]\}\end{aligned} \tag{4}$$

For more details, refer to Cannon et al. (2015).

Compared to the methods described above, our calibration method in Section 3.2 employs a parametric approach for extrapolation. We compare these methods and evaluate their performance in Section 4.3.2.

3 Methodology

3.1 A mixture model for modeling the distribution's bulk and tail

We create a model for the the entire distribution of projected observations, and rely on extreme value models to characterize the upper tail. To motivate our eventual model, we briefly describe modeling and calibrating only the extremes.

An aim of an extreme value analysis is to not let data in the bulk of the distribution influence estimates of tail behavior. Thus, classical extremes methods analyze only an extreme subset of the data: either block (e.g., annual) maxima or threshold exceedances. Here, our data record is insufficient to model only annual maxima, and seasonal effects would make fitting sub-annual maxima (e.g., monthly) dubious. Consider a standard peaks-over-threshold approach, in which the generalized Pareto (GP) distribution (Balkema and De Haan, 1974) is fit to data which exceed a high threshold u. The GP distribution is defined as

$$G(x; u, \xi, \sigma) = \mathrm{P}(X < x \mid X > u) = \begin{cases} 1 - \left[1 + \frac{\xi}{\sigma}(x-u)\right]_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma}\right)\right]_+, & \xi = 0, \end{cases} \tag{5}$$

where $z^+ = \max(z, 0)$. The GP density is

$$g(x; u, \xi, \sigma) = \sigma^{-1} \left\{ 1 + \frac{\xi}{\sigma} (x - u) \right\}^{-1/\xi - 1}, \tag{6}$$

on the support $\{x : x > u, 1 + \frac{\xi}{\sigma}(x-u) > 0\}$ with a scale parameter $\sigma > 0$ and a shape parameter $\xi \in (-\infty, \infty)$. The shape parameter ξ determines the fundamental nature of the tail. If $\xi < 0$, the distribution of threshold excesses has a bounded tail with an upper end point $u < x < u - \sigma/\xi$. The zero shape parameter $\xi = 0$ corresponds to a light tail (where g becomes the exponential distribution in the limit), and the positive shape parameter $\xi > 0$ corresponds to a heavy tail.

The threshold is selected (not estimated), typically using diagnostic plots. In order to perform calibration, the threshold for the observations and model output would need to have a known relationship, and one approach would be to employ thresholds defined by a common exceedance probability.

Although our primary interest is in estimating very high quantiles, we desire to perform calibration of the entire distribution so that any quantity of interest can be projected. We consider a model which flexibly fits the distribution's tail, which behaves like a generalized Pareto in the upper tail, and which restricts data in the bulk from influencing parameter estimates in the tail. Our mixture model assumes that for a threshold $u > 0$, the probability density function of the random variable X is

$$h(x; \boldsymbol{\theta}) = d(\boldsymbol{\theta})^{-1} \Big[(1 - \pi(x; u, \delta)) \frac{f(x; \beta, \lambda)}{F(u; \beta, \lambda)} (1 - \kappa_u) + \pi(x; u, \delta) g(x; u, \xi, \sigma) \kappa_u \Big], \tag{7}$$

where $\boldsymbol{\theta} = (\beta, \lambda, \xi, \sigma)$ is the parameter vector, $f(x; \beta, \lambda) = (\beta/\lambda)(x/\lambda)^{\beta-1} \exp(-(x/\lambda)^\beta)$ for $x > 0$ represents the Weibull density with scale parameter $\lambda > 0$ and shape parameter $\beta > 0$, $F(u; \beta, \lambda)$ is the Weibull CDF evaluated at u, $\kappa_u = \mathrm{P}(X > u)$, and $g(x; u, \xi, \sigma)$ is the generalized Pareto density in (6). The weight function

$$\pi(x; u, \delta) = \begin{cases} 0, & \text{if} \quad x < u \\ \frac{1}{\delta}, & \text{if} \quad u \le x \le u + \delta \\ 1, & \text{if} \quad x > u + \delta \end{cases}, \tag{8}$$

for $\delta > 0$ provides a continuous transition between the models for the bulk and tail. The normalizing constant $d(\boldsymbol{\theta}) = \int_0^\infty h(x; \boldsymbol{\theta}) \mathrm{d}x$ is needed only because of the weight function; if $\delta = 0$, then $\pi(x; u, 0)$ is defined as 0 for $x < u$ and 1 otherwise, implying $d(\boldsymbol{\theta}) = 1$ for any parameter values.

Given independent and identically distributed (i.i.d.) observations $x_1, \ldots, x_n$, inference begins by choosing an appropriate threshold u and transition range parameter δ. The threshold u can be chosen via visual diagnostics such as the mean residual life plot (Davison and Smith, 1990) and δ can be chosen qualitatively so that the transition between the bulk and the tail is satisfactory. With u and δ selected, $\hat{\kappa}_u = n^{-1} \sum_{i=1}^n \mathbb{1}(x_i > u)$, and numerical maximum likelihood (ML) can then be performed to find estimates for β, λ, ξ, and σ, where $d(\boldsymbol{\theta})$ is calculated by numerical integration.

There have been models proposed which fit the entire distribution, but which also have an upper tail which behaves asymptotically like a generalized Pareto. Often the aim of these models is to avoid the issues associated with selecting a threshold. One method, proposed by Papastathopoulos and Tawn (2013) and extended by both Naveau et al. (2016) and Stein (2021) constructs a model via a composition of the generalized Pareto distribution $G(x; u, \xi, \sigma)$ in (5) and another 'carrier' CDF, $Q(v)$ for $v \in [0, 1]$. Specifically, the key motivation in the Naveau et al. (2016) model is as follows: we can simulate from the GP distribution by applying a uniformly distributed random sample U into the GP quantile function $G^{-1}(U)$. By replacing U with a more flexible random variable $V = Q^{-1}(U)$, where Q is a continuous CDF on $[0, 1]$, the resulting random variable $Y = G^{-1}[Q^{-1}(U)]$ forms a more flexible distribution family. They propose a class of CDFs, $Q(v)$, such that the upper tail retains the behavior of the GP distribution, and the CDF of Y near zero behaves like a power function y^{κ}. Four parametric families satisfying these conditions are introduced, and for simplicity, we focus on the first two for model comparison.

1. $Q(v) = v^{\kappa}, \quad \kappa > 0,\ v \in [0, 1]$
2. $Q(v) = pv^{\kappa_1} + (1 - p)v^{\kappa_2}, \quad \kappa_1, \kappa_2 > 0,\ v, p \in [0, 1].$

Another approach, termed a 'mixture' by Scarrott and MacDonald (2012), combines a density model for the bulk with a generalized Pareto density for the tail, often smoothing the transition between them. A particular mixture model was proposed by Frigessi et al. (2002) which used a Weibull model for the bulk and a GPD for the tail. In contrast to our weight function (8), their model uses a Cauchy CDF as a weight function to transition between the bulk and the tail over the entire data range. The GP shape parameter in the Frigessi et al. (2002) model tends to be either overestimated or underestimated, as noted by Naveau et al. (2016), and the scale parameter of the Cauchy CDF, which controls the transition speed, is challenging to estimate.

We investigate the tail behavior of these parametric models without using threshold selection and assess the goodness of fit in Section 4.3, comparing it to our fixed threshold approach.

3.2 Calibration method via linear mapping of distribution parameters

Our calibration method assumes the linear relationship between parameters associated with model output and parameters associated with observations. Specifically, we will assume

$$\boldsymbol{\theta}^{\cdot}_{Obs} = A\boldsymbol{\theta}^{\cdot}_{Mod} + \boldsymbol{b}, \tag{9}$$

where $\boldsymbol{\theta}^{\cdot}_{Obs}$ and $\boldsymbol{\theta}^{\cdot}_{Mod}$ denote the parameter vectors of the distributions for the observations and the model output respectively, and '$\cdot$' is a placeholder for both h denoting the historical and p denoting the projected climate. The historical period will be used to estimate this linear relationship, and then it is applied to the parameters of the projected model output to obtain an estimate of the distribution of projected observations.

With the available observations and model output, one can obtain ML estimates for $\boldsymbol{\theta}^{h}_{Obs} = (\beta^{h}_{Obs}, \lambda^{h}_{Obs}, \xi^{h}_{Obs}, \sigma^{h}_{Obs}), \boldsymbol{\theta}^{h}_{Mod}$, and $\boldsymbol{\theta}^{p}_{Mod}$, the parameter vectors for the extreme

mixture model applied to observations in historical climate, model output under historical climate, and model output under projected climate, respectively. Furthermore, asymptotic estimates for the respective covariance matrices Σ^h_{Obs}, Σ^h_{Mod}, and Σ^p_{Mod} can be obtained using the inverse of the observed Fisher information matrix. The calibration method uses these estimates to produce an estimate of $\boldsymbol{\theta}^p_{Obs}$, the parameter vector for projected observations, for which there are no observations to analyze.

Let A be a 4×4 matrix and $\boldsymbol{b}$ be a 4×1 vector. Both A and $\boldsymbol{b}$ are assumed time-invariant; that is, this same linear relationship holds both under historical and projected climate. Essentially, this assumption says that the model 'biases' which lead to the discrepancy between the modeled river flow output and the observations do not depend on the climate state; without some similar assumption, the calibration problem is impossible.

One can use parameter estimates for both the observations and model output in the historical climate to estimate A and $\boldsymbol{b}$. Based on (9), we consider the second-order moment form, motivated by the asymptotic normality of ML estimators

$$\Sigma^h_{Obs} = A\Sigma^h_{Mod}A^\top. \tag{10}$$

Solving (10) with estimates plugged in yields $\widehat{A} = \widehat{\Sigma}^{h^{1/2}}_{Obs}\widehat{\Sigma}^{h^{-1/2}}_{Mod}$ and plugging into (9) yields $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\theta}}^h_{Obs} - \widehat{\Sigma}^{h^{1/2}}_{Obs}\widehat{\Sigma}^{h^{-1/2}}_{Mod}\hat{\boldsymbol{\theta}}^h_{Mod}$. We obtain square root matrices by spectral decomposition so that generically $\Sigma^{1/2} = P\Lambda^{1/2}P^\top$, where P is the square orthogonal matrix of Σ's eigenvectors and Λ is the diagonal matrix of eigenvalues. If the covariance matrix is not positive definite due to closely corrected ML estimates or numerical issues, it is necessary to obtain the nearest positive definite covariance matrix before performing the spectral decomposition (e.g., using the `nearPD` function from the `Matrix` R package).

With estimates $\widehat{A}$ and $\hat{\boldsymbol{b}}$, the parameter estimate for the projected observations is

$$\hat{\boldsymbol{\theta}}^p_{Obs} = \widehat{A}\hat{\boldsymbol{\theta}}^p_{Mod} + \hat{\boldsymbol{b}}. \tag{11}$$

3.3 Uncertainty quantification for projected parameter estimates

The quantity of primary interest is $\hat{\boldsymbol{\theta}}^p_{Obs}$ and we devise a bootstrap procedure to quantify its uncertainty which arises from the uncertainty of parameter estimators as well as that of the estimated linear projection in (11). Let $x^h_{Obs}(t)$, $x^h_{Mod}(t)$, $x^p_{Mod}(t)$, $t = 1, \ldots, n$, be observations in historical climate, model output under historical climate, and model output under projected climate, respectively. For each dataset, resample with replacement to obtain bootstrap ML estimates, $\hat{\boldsymbol{\theta}}^{h\,(b)}_{Obs}, \hat{\boldsymbol{\theta}}^{h\,(b)}_{Mod}, \hat{\boldsymbol{\theta}}^{p\,(b)}_{Mod}, \widehat{\Sigma}^{h\,(b)}_{Obs}, \widehat{\Sigma}^{h\,(b)}_{Mod}$, and $\widehat{\Sigma}^{p\,(b)}_{Mod}$, where $b = 1, \ldots, B$, the number of bootstrap iterations. One can obtain $\widehat{A}^{(b)}$ and $\hat{\boldsymbol{b}}^{(b)}$ as in Section 3.2 to obtain $\hat{\boldsymbol{\theta}}^{p\,(b)}_{Obs}$. Bootstrap confidence intervals can then be constructed.

To account for serial dependence in real data applications such as the river flow data for the Purgatoire River, we apply the block bootstrap method. The approach involves spliting the data into non-overlapping blocks and resamples these blocks to

preserve the temporal dependence in each block. To determine the block length, we use both the `blockboot` function from the `OBL` R package, which selects the optimal block length in terms of the minimum root mean squared error, and the autocovariance function (ACF).

4 Results

4.1 Simulation study

We assess coverage rates via a simulation study. Values for $\boldsymbol{\theta}^h_{Obs}$, $\boldsymbol{\theta}^h_{Mod}$, $\boldsymbol{\theta}^p_{Mod}$ are set to the ML estimates for our application (given in the first three rows of Table 3 below). Assuming (9) and using covariance estimates from the real data, we solve for $A, \boldsymbol{b}$, and obtain $\boldsymbol{\theta}^p_{Obs}$.

For each simulation iteration, $n = 5,000$ i.i.d. realizations are drawn by accept-reject algorithm from the extreme mixture models for historical observations, historical model output, and projected model output. For each sample, we draw $B = 1,000$ bootstrap samples and obtain bootstrap ML estimates for these three distributions. $\widehat{A}^{(b)}$, $\hat{\boldsymbol{b}}^{(b)}$, and $\hat{\boldsymbol{\theta}}^{p\,(b)}_{Obs}$ are obtained as described in Section 3.2.

Bootstrap based 95% confidence intervals are produced for all parameter estimates. Parallel computing is used to repeat the simulation 100 times, and Table 2 reports coverage rates. Coverage rates for the quantity of interest $\boldsymbol{\theta}^p_{Obs}$ appear reasonable, especially given the coverage rates for the other parameters which do not require calibration.

	Coverage Rates
$(\hat{\beta}, \hat{\lambda}, \hat{\xi}, \hat{\sigma})^h_{Mod}$	(0.97, 0.98, 0.97, 0.95)
$(\hat{\beta}, \hat{\lambda}, \hat{\xi}, \hat{\sigma})^h_{Obs}$	(0.92, 0.93, 0.93, 0.95)
$(\hat{\beta}, \hat{\lambda}, \hat{\xi}, \hat{\sigma})^p_{Mod}$	(0.97, 0.94, 0.85, 0.93)
$\boldsymbol{(\hat{\beta}, \hat{\lambda}, \hat{\xi}, \hat{\sigma})^p_{Obs}}$	$(\mathbf{0.99}, \mathbf{0.99}, \mathbf{0.95}, \mathbf{0.97})$

Table 2: Coverage rates for 95% bootstrap based confidence intervals for 100 simulations.

4.2 Case study: application to river discharges

We apply the calibration method to daily-averaged runoff model output and river discharges from the Purgatoire River in southeastern Colorado. We use complete measurements from three data sets: historical model output, projected model output, and historical river discharges for the period from April to August between 2002 and 2013, as described in Section 2.1.

To assess the tail behavior of the historical observations, we first fit a GP distribution to the observations exceeding the 0.95 quantile ($q_{0.95} = 10.27$), obtaining the shape parameter estimate of $\hat{\xi}^h_{Obs} = 0.05$. The 95% confidence interval for ξ is $(-0.21, 0.32)$, suggesting the possibility of either a light tail or a slightly heavy tail.

We then fit a mixture model $h(x;\boldsymbol{\theta})$ in (7) to each data set, obtaining the ML estimates reported in the first three rows of Table 3. The threshold u is set to $q_{0.95}$,

	ML estimates	Standard Errors
$(\hat{\beta},\hat{\lambda},\hat{\xi},\hat{\sigma})^{h}_{Mod}$	(1.31, 2.76, -0.13, 2.18)	(0.03, 0.07, 0.10, 0.39)
$(\hat{\beta},\hat{\lambda},\hat{\xi},\hat{\sigma})^{h}_{Obs}$	(0.57, 0.91, 0.05, 8.99)	(0.01, 0.05, 0.11, 1.78)
$(\hat{\beta},\hat{\lambda},\hat{\xi},\hat{\sigma})^{p}_{Mod}$	(1.27, 2.42, 0.01, 3.10)	(0.03, 0.05, 0.12, 0.58)
$\boldsymbol{(\hat{\beta},\hat{\lambda},\hat{\xi},\hat{\sigma})^{p}_{Obs}}$	$(\mathbf{0.53},\mathbf{0.57},\mathbf{0.19},\mathbf{9.92})$	$(\mathbf{0.03},\mathbf{0.11},\mathbf{0.45},\mathbf{6.30})$

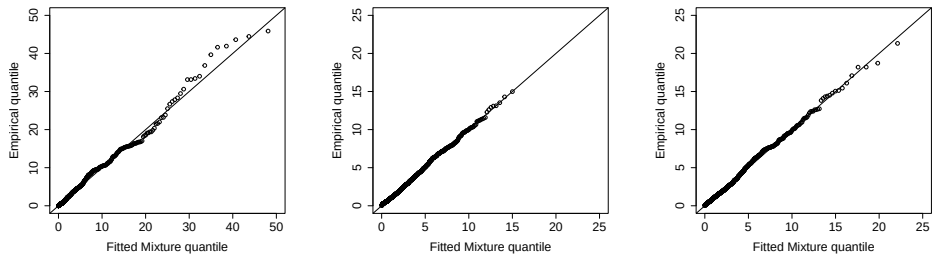
Table 3: ML estimates for parameters and their corresponding bootstrap-based standard errors.

the empirical 0.95 quantile of each data set and $\delta = q_{0.96} - u$. Figure 3 shows QQ-plots for the mixture model fit to river discharges in historical climate, model output under historical climate, and model output under projected climate. Of particular interest is the upper tail, and the QQ-plots show a reasonable fit accounting for the usual model uncertainty associated with estimating extreme behavior.

The estimate $\hat{\xi}^{h}_{Obs} = 0.05$ agrees with our preliminary generalized Pareto fit showing that the bulk data do not influence the tail estimate. The estimate of $\hat{\xi}^{h}_{Mod} = -0.13$ from the historical model output, with a 95% confidence interval of $(-0.37, 0.07)$, may suggest a bounded tail. This is because the confidence interval contains more negative plausible values when accounting for sampling uncertainty. The bounded tail for model output may illustrate the challenges that the sequence of numerical models face in replicating extreme behavior. The GP shape parameter estimate from the projected model output is $\hat{\xi}^{p}_{Mod} = 0.01$, and eight instances of modeled river flows under the projected climate exceed the maximum modeled river flow under the historical climate as shown in the center and right panels of Figure 3. To test whether the tail behavior of the model output under the projected climate is significantly greater than that under historical climate, a one-sided Wald test was conducted. The resulting p-value of 0.21 indicates that the data do not provide sufficient evidence to support a significant difference in the tail behavior of the model output between the two climates.

We next find the linear relationship between river discharges and model output in historical climate through $\widehat{A}$ and $\hat{\boldsymbol{b}}$ and then apply the linear relationship under projected climate to obtain $\hat{\boldsymbol{\theta}}^{p}_{Obs}$, reported in the fourth row of Table 3. Differences between $\hat{\boldsymbol{\theta}}^{p}_{Obs}$ and $\hat{\boldsymbol{\theta}}^{h}_{Obs}$ reflect the differences between $\hat{\boldsymbol{\theta}}^{p}_{Mod}$ and $\hat{\boldsymbol{\theta}}^{h}_{Mod}$, as both the GP shape and scale parameters have increased. The estimate $\hat{\xi}^{p}_{Obs} = 0.19$ seems large, but its standard error of 0.45 reflects the uncertainty in estimating tail indices with short data records, and the uncertainty in projecting observations. Figure 4 shows the densities associated with the fitted models for both model output and observations under both historical and projected climate, and shows kernel density estimates for the three combinations with data.

We report 95% block-bootstrap confidence intervals for parameters of $\boldsymbol{\theta}^{h}_{Obs}$ and $\boldsymbol{\theta}^{p}_{Obs}$, respectively in the first four columns of Table 4. Using the `blockboot` tool and



(a) (b) (c)

Fig. 3: (a) QQ-plot of river discharges under historical climate; (b) QQ-plot of model output under historical climate; and (c) QQ-plot of model output under projected climate.

ACF together, we set the block lengths to 48, 30, and 30 days for historical observations, historical model output, and projected model output, respectively. The 95% block-bootstrap confidence intervals for (ξ, σ) in a GP distribution are of primary interest. Not surprisingly, we observe wider bootstrap confidence intervals for $(\xi, \sigma)^p$ than ones for $(\xi, \sigma)^h$ due to the additional uncertainty of linear projection. Estimates of high quantiles are of more practical interest than the parameter estimates. We provide 95% block-bootstrap confidence intervals for the 0.9993 and 0.99993 quantiles of the historical and projected observations. These correspond to the 1-in-1500 and 1-in-15000 day event; as there are 152 observations in the April to August period, these are roughly 1-in-10 and 1-in-100 year events. The width of the confidence intervals under the historical period is wide due to the short data record we employ, and this uncertainty becomes amplified when projected. Nevertheless, despite the limited information in the data and model runs for this risk study, it seems there is potential for higher river flows and thus increased flood risk under the projected climate.

	β	λ	ξ	σ	$q_{0.9993}$	$q_{0.99993}$
$\boldsymbol{\theta}^h_{Obs}$	(0.55, 0.60)	(0.83, 1.00)	(-0.13, 0.29)	(5.64, 12.31)	(40.36, 60.37)	(57.27, 1768.95)
$\boldsymbol{\theta}^p_{Obs}$	(0.51, 0.60)	(0.53, 0.85)	(-0.70, 0.90)	(6.70, 29.73)	(44.33, 344.51)	(49.37, 4357.51)

Table 4: 95% block-bootstrap based confidence intervals for parameters associated with local observations as well as for the 0.9993 and 0.99993 quantiles under historical climate and projected climate, respectively.

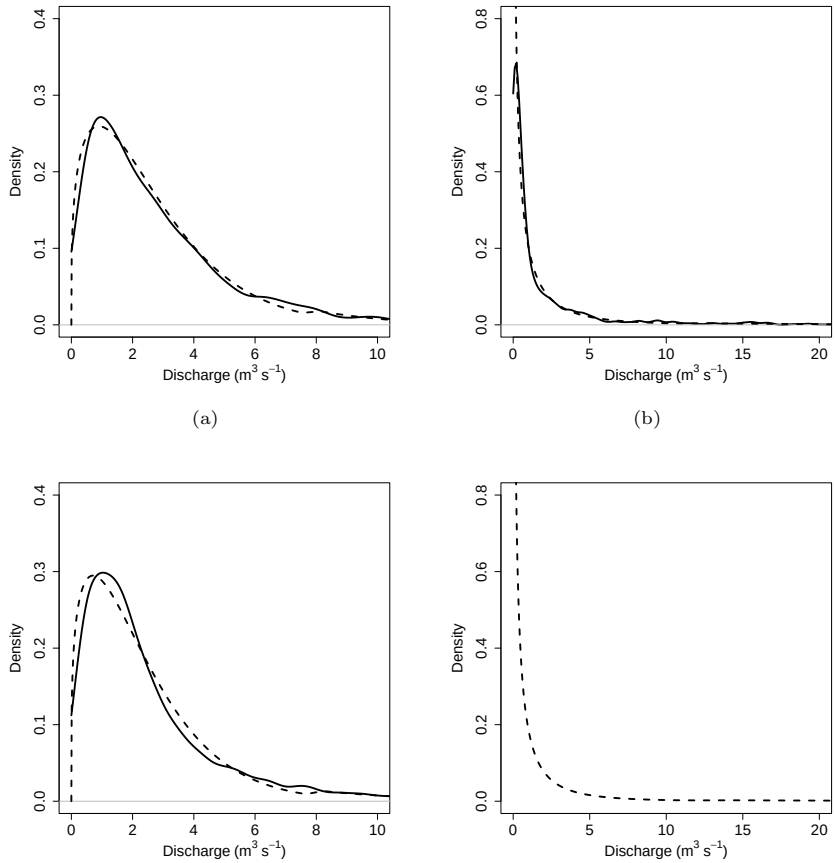


Fig. 4: Dashed line shows the fitted mixture model density (a) for model output under historical climate, (b) for observed river discharges under historical climate, (c) for model output under projected climate, and (d) for river discharges under projected climate. Kernel density estimates are shown with solid lines in (a), (b), and (c).

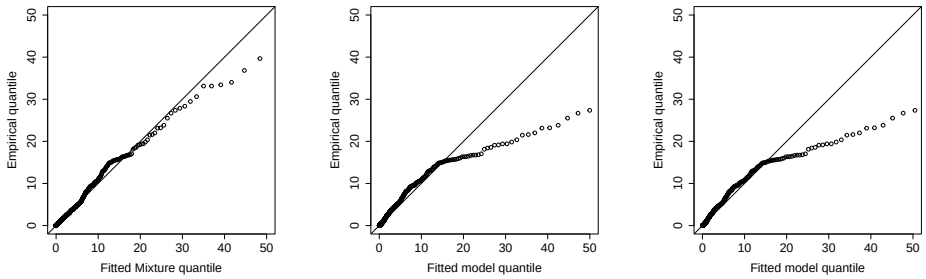
4.3 Method validation and comparative assessment with other approaches

4.3.1 Goodness of fit for other parametric models without threshold selection

We explore the tail behavior of other parametric extremes models described in Section 3.1, starting by fitting each model to the entire historical river discharges. We used the `fit.extgp` tool in the `mev` R package to fit the Naveau et al. (2016) model considering two carrier functions: $Q(v) = v^{\kappa}$ and $Q(v) = pv^{\kappa_1} + (1-p)v^{\kappa_2}$, $\kappa_1, \kappa_2 > 0$, $v, p \in$

$[0, 1]$. We also fit the Frigessi et al. (2002) model by numerical maximum likelihood estimation. Both models produced heavy tail ML estimates with $\hat{\xi}^h_{Obs} = 0.56$ for the Frigessi et al. (2002) model and $\hat{\xi}^h_{Obs} = 0.99$ for both carrier functions in the Naveau et al. (2016) model.

For goodness of fit, QQ-plots of the empirical qunatiles against the fitted model quantiles for the two carrier functions showed a clear mismatch between the modeled upper tail and the largest observations in Figure 5b and 5c. While the QQ-plot for the Frigessi et al. (2002) model in Figure 5a performed better, it still showed more discrepancies in the higher quantiles compared to our model with the fixed threshold approach in Figure 3a. A similar issue arises for the historical model output (not shown).



(a) (b) (c)

Fig. 5: QQ-plot of empirical quantiles of historical observations versus fitted model quantiles (a) for the Frigessi et al. (2002) model, (b) for the Naveau et al. (2016) model with a carrier $Q(v) = v^{\kappa}$, $v \in [0, 1]$, and (c) for the Naveau et al. (2016) model with a carrier $Q(v) = pv^{\kappa_1} + (1-p)v^{\kappa_2}$, $\kappa_1, \kappa_2 > 0$, $v, p \in [0, 1]$.

While the asymptotic tail behavior of these models follows the GP distribution in the limit, this case study indicates either the estimation is challenging, or that both models do not sufficiently separate the tail behavior from the bulk, resulting in data from the bulk unduly influencing the shape parameter estimate. Therefore, we opt to use a fixed threshold approach for the tail to ensure the proper calibration of extremes.

4.3.2 Validation for different calibration methods

Focusing on extremes, we perform validation for different statistical calibration methods outlined in Section 2.2, using only the historical period, as projected observations are unavailable. Ideally, both historical and projected datasets with sufficiently long records would be used for a more reliable evaluation. However, in the absence of such data, we split both the historical observations and model output into a calibration set (60%, a sample size $n_{cal} = 1,101$) and a validation set (40%, a sample size $n_{val} = 735$), assuming that these datasets still preserve distinct climate characteristics. The calibration set is used as the historical period to estimate parameters for the mixture

model in (7) and the transfer functions in Section 2.2, and the validation set serves as the out-of-sample data to assess goodness of fit and evaluate the performance of the calibration methods.

We compare projected observations (bias-corrected values) derived from different statistical calibration methods to actual observations (assumed to be observed) in the validation set. We consider six calibration methods: linear parameter mapping, QDM preserving absolute changes in quantiles (4), QDM preserving relative changes in quantiles (3), QM (2), the linear scaling approach (1), and a simple regression approach.

To properly evaluate these calibration methods for extremes, we use both QQ-plots and summary statistics that take datasets where both observations and projected observations exceed the 0.95 quantile, respectively. Similarly in Section 2.1.1, we report the summary statistics in Table 5. The values before the slash represent summary statistics calculated using all data points, while the values after the slash are derived from data exceeding the high threshold.

	R^2	RMSE	NSE
Linear mapping	0.97 / 0.98	0.86 / 3.58	-1e-3 / -0.02
QDM-abs	0.38 / 0.70	4.20 / 10.85	-0.02 / 0.12
QDM-rel	0.39 / 0.71	3.79 / 9.68	-0.03 / -0.01
QM	0.38 / 0.68	3.38 / 10.55	-2e-3 / -0.14
Linear scaling	0.31 / 0.67	4.32 / 10.17	-0.03 / -0.01
Simple regression	0.29 / 0.68	3.57 / 13.15	-0.04 / -0.81

Table 5: Summary statistics of R^2, RMSE, and NSE for six different calibration methods. QDM-abs indicates QDM preserving absolute changes in quantiles, and QDM-rel stands for QDM preserving relative changes in quantiles.

For the linear mapping of parameters approach, we obtain GP shape parameter estimates of $\hat{\xi}^h_{Obs} = 0.09$ for the historical climate in the training set and $\hat{\xi}^p_{Obs} = 0.12$ for the projected climate in the validation set. As the projected observations in the validation set are quantile-matched to the corresponding actual observations for the projected period, the correlation coefficient, or R^2, between projected and actual observations is effectively close to 1.

To visualize the performance of the six approaches for extremes, QQ-plots are shown in Figure 6. The high quantile values from the linear mapping approach align closely with the diagonal, indicating a satisfactory fit for extreme values. The QDM results also appear reasonable, indicating potential for further improvement through parametric estimation of the CDFs. It is important to note that this method comparison does not definitely conclude that any one method is superior in all cases. The performance of each method may vary depending on the characteristics of the real data and the available predictor variables, where different transformations and parameterizations could lead to different results.

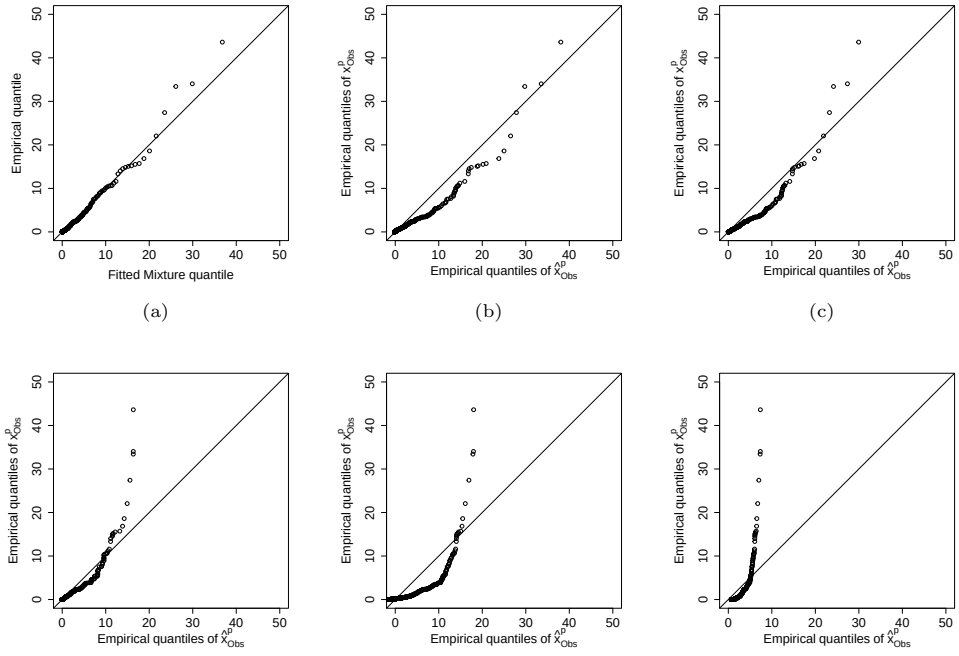


Fig. 6: (a) QQ-plot of empirical quantiles of observations versus fitted model quantiles in the validation set for a linear mapping of parameters. QQ-plots of empirical quantiles of actual versus projected observations in the validation set (b) for a QDM preserving absolute changes in quantiles, (c) for a QDM preserving relative changes in quantiles, (d) for QM, (e) for a linear scaling approach, and (f) for a simple regression approach, respectively.

5 Summary and Discussion

We develop a novel statistical calibration method focusing on extreme values by applying a linear parameter mapping approach rather than directly calibrating the model output. This method has a fundamental assumption of a linear relationship between parameters associated with model output and those with local observations. Once estimated under historical climate conditions, this linear relationship is applied to parameter estimates under projected climates. This stationarity assumption is standard to most statistical calibration methods. However, given that both model output and observations exhibit temporal dependence, the dependence could affect uncertainty estimates. Given a long data record, a possible extension to this work is to incorporate temporal dependence in the calibration method by considering non-stationarity in this linear relationship, that is, incorporating temporal dependence or possible covariates to the scale or shape parameters of the GP distribution.

To flexibly fit the tail of the distribution, we employ a mixture distribution with an extreme value model for the tail; however, the proposed calibration method itself does not require any particular distribution. We opt for a fixed threshold approach for tail approximation due to a better performance for extremes compared to automatic threshold selection methods. But, this fixed threshold approach introduces some subjectivity. Some other approach that can reduce the subjectivity of a threshold selection would be useful for improving the calibration method for extremes. Our method is currently applied in a univariate case, focusing on changes in the marginal distribution. Extending this approach to a multivariate calibration using a multivariate GP distribution could be an interesting future work.

Importantly, to account for uncertainty in both parameter estimates and linear projection, a bootstrap approach is employed. Accounting for uncertainty is at least as important as the projected quantity estimates themselves. Of course, the projections we produce are based on only one possible climate scenario, and the uncertainty of human behavior and its affect on the climate likely outweighs the estimation uncertainty we capture here. As the climate changes, there is increased need to produce projections to inform planning for potential future outcomes. Numerical models are powerful, but imperfect tools, and calibration methods such as ours can help to account for when model output does not accurately reflect the quantities needed by decision makers.

Acknowledgments JL, DC, AW, and GL's work was supported by Strategic Environmental Research and Development Program (SERDP) Resource Conservation (RC) and Resilience Project RC-2515. JL and DC were also partially supported by National Science Foundation (NSF) DMS-1811657.

References

Balkema, A.A., De Haan, L.: Residual life time at great age. The Annals of probability, 792–804 (1974)

Benestad, R.E.: Downscaling precipitation extremes. Theoretical and Applied Climatology **100**(1-2), 1–21 (2010)

Bürger, G., Murdock, T., Werner, A., Sobie, S., Cannon, A.: Downscaling extremes—an intercomparison of multiple statistical methods for present climate. Journal of Climate **25**(12), 4366–4388 (2012)

Cannon, A.J., Sobie, S.R., Murdock, T.Q.: Bias correction of gcm precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? Journal of Climate **28**(17), 6938–6959 (2015)

Campozano, L., Tenelanda, D., Sanchez, E., Samaniego, E., Feyen, J.: Comparison of statistical downscaling methods for monthly total precipitation: case study for the paute river basin in southern ecuador. Advances in Meteorology **2016**(1), 6526341 (2016)

549 Davison, A.C., Smith, R.L.: Models for exceedances over high thresholds. Journal of
550 the Royal Statistical Society: Series B (Methodological) **52**(3), 393–425 (1990)

551 Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae,
552 U., Balmaseda, M., Balsamo, G., Bauer, d.P., *et al.*: The era-interim reanalysis:
553 Configuration and performance of the data assimilation system. Quarterly Journal
554 of the royal meteorological society **137**(656), 553–597 (2011)

555 Frigessi, A., Haug, O., Rue, H.: A dynamic mixture model for unsupervised tail
556 estimation without threshold selection. Extremes **5**(3), 219–235 (2002)

557 Hanssen-Bauer, I., Achberger, C., Benestad, R., Chen, D., Førland, E.: Statistical
558 downscaling of climate scenarios over scandinavia. Climate Research **29**(3), 255–268
559 (2005)

560 Hessami, M., Gachon, P., Ouarda, T.B., St-Hilaire, A.: Automated regression-based
561 statistical downscaling tool. Environmental modelling & software **23**(6), 813–834
562 (2008)

563 Huang, Y.-N., Reich, B.J., Fuentes, M., Sankarasubramanian, A.: Complete spatial
564 model calibration. The annals of applied statistics **13**(2), 746–766 (2019)

565 Liston, G.E., Elder, K.: A distributed snow-evolution modeling system (snowmodel).
566 Journal of Hydrometeorology **7**(6), 1259–1276 (2006)

567 Liu, C., Ikeda, K., Rasmussen, R., Barlage, M., Newman, A.J., Prein, A.F., Chen,
568 F., Chen, L., Clark, M., Dai, A., *et al.*: Continental-scale convection-permitting
569 modeling of the current and future climate of north america. Climate Dynamics **49**,
570 71–95 (2017)

571 Liston, G.E., Itkin, P., Stroeve, J., Tschudi, M., Stewart, J.S., Pedersen, S.H.,
572 Reinking, A.K., Elder, K.: A lagrangian snow-evolution system for sea-ice applica-
573 tions (snowmodel-lg): Part i—model description. Journal of Geophysical Research:
574 Oceans **125**(10), 2019–015913 (2020)

575 Liston, G.E., Mernild, S.H.: Greenland freshwater runoff. part i: A runoff routing
576 model for glaciated and nonglaciated landscapes (hydroflow). Journal of Climate
577 **25**(17), 5997–6014 (2012)

578 Li, H., Sheffield, J., Wood, E.F.: Bias correction of monthly precipitation and tem-
579 perature fields from intergovernmental panel on climate change ar4 models using
580 equidistant quantile matching. Journal of Geophysical Research: Atmospheres
581 **115**(D10) (2010)

582 Naveau, P., Huser, R., Ribereau, P., Hannart, A.: Modeling jointly low, moderate, and
583 heavy rainfall intensities without a threshold selection. Water Resources Research
584 **52**(4), 2753–2769 (2016)

585 Nash, J.E., Sutcliffe, J.V.: River flow forecasting through conceptual models part i—a
586 discussion of principles. Journal of hydrology **10**(3), 282–290 (1970)

587 Olsson, J., Berggren, K., Olofsson, M., Viklander, M.: Applying climate model pre-
588 cipitation scenarios for urban hydrological assessment: A case study in kalmar city,
589 sweden. Atmospheric Research **92**(3), 364–375 (2009)

590 Papastathopoulos, I., Tawn, J.A.: Extended generalised pareto models for tail
591 estimation. Journal of Statistical Planning and Inference **143**(1), 131–143 (2013)

592 Rasmussen, R., Liu, C., Ikeda, K., Gochis, D., Yates, D., Chen, F., Tewari, M., Barlage,
593 M., Dudhia, J., Yu, W., *et al.*: High-resolution coupled climate runoff simulations
594 of seasonal snowfall over colorado: A process study of current and warmer climate.
595 Journal of Climate **24**(12), 3015–3048 (2011)

596 Schubert, S., Henderson-Sellers, A.: A statistical model to downscale local daily
597 temperature extremes from synoptic-scale atmospheric circulation patterns in the
598 australian region. Climate Dynamics **13**, 223–234 (1997)

599 Scarrott, C., MacDonald, A.: A review of extreme value threshold estimation and
600 uncertainty quantification. REVSTAT-Statistical journal **10**(1), 33–60 (2012)

601 Stein, M.L.: Parametric models for distributions when interest is in extremes with an
602 application to daily temperature. Extremes **24**, 293–323 (2021)

603 Taylor, J.W.: A quantile regression neural network approach to estimating the con-
604 ditional density of multiperiod returns. Journal of forecasting **19**(4), 299–311
605 (2000)

606 Teutschbein, C., Seibert, J.: Bias correction of regional climate model simulations
607 for hydrological climate-change impact studies: Review and evaluation of different
608 methods. Journal of hydrology **456**, 12–29 (2012)

609 Vrac, M., Naveau, P.: Stochastic downscaling of precipitation: From dry events to
610 heavy rainfalls. Water resources research **43**(7) (2007)

611 Wang, L., Chen, W.: Equiratio cumulative distribution function matching as an
612 improvement to the equidistant approach in bias correction of precipitation.
613 Atmospheric Science Letters **15**(1), 1–6 (2014)

614 Wilby, R.L., Hay, L.E., Leavesley, G.H.: A comparison of downscaled and raw gcm out-
615 put: implications for climate change scenarios in the san juan river basin, colorado.
616 Journal of Hydrology **225**(1), 67–91 (1999) https://doi.org/10.1016/S0022-1694(99)
617 00136-5

618 Wood, A.W., Maurer, E.P., Kumar, A., Lettenmaier, D.P.: Long-range experimental
619 hydrologic forecasting for the eastern united states. Journal of Geophysical Research:

620 Atmospheres **107**(D20), 6 (2002)