Practical issues with modeling extreme Brazilian rainfall

Paulo V. C. Pereira\textsuperscript{a}, Isolde T. S. Previdelli\textsuperscript{a} and Anthony C. Davison\textsuperscript{b,1}

\textsuperscript{a}Universidade Estadual de Maringá
\textsuperscript{b}École polytechnique fédérale de Lausanne

Abstract. Accurately quantifying extreme rainfall is important for the design of hydraulic structures, for flood mapping and zoning and for disaster management. In order to produce maps of estimates of 25-year rainfall return levels in Brazil, we selected 893 shorter and 104 longer rainfall time series from the Agência Nacional de Águas (ANA), and applied the framework of extreme value theory. Care was needed to reduce the impact of poor data. Estimates of the shape parameter of the extreme-value model fitted to rainfall data are typically biased, so we discuss an empirical correction that takes into account not only the sample-size bias, but also a so-called penultimate approximation that we use to inform a Bayesian spatial latent variable model for the annual rainfall maxima. This model accounts for subtle patterns of spatial variation in the data and provides plausible return level estimates.

1 Introduction

Floods and landslides are frequent after heavy and continuous summer rainfall in Brazil. An incident on 11 and 12 January 2011 in the mountainous region of Rio de Janeiro State led to 947 deaths and is regarded as the worst natural disaster in Brazil’s history. The accumulated rainfall in 24 hours was 242 mm, with a peak of 62 mm in an hour (Dourado, Arraes and Silva, 2012). Just a year later, in the austral summer and fall of 2012, extreme drought severely affected eastern Brazil, damaging the country’s agricultural and electrical production. According to Getirana (2016), this was mostly due to unusually low precipitation, exacerbated by ineffective energy development and poor water management. At the same time there was record-breaking rainfall and flooding in Amazonia (Marengo et al., 2013). Thus the modeling of likely extreme rainfall has considerable environmental and societal relevance, which can be expected to grow alongside sustainability concerns.

\textsuperscript{1}ACD was privileged to know Bent Jørgensen for over 30 years. He was a generous friend, an admired colleague and an outstanding scientist, and is sorely missed.

Keywords and phrases. Bayesian hierarchical model, Data quality, penultimate approximation, return level map, small-sample bias, statistics of extremes
Davison and Huser (2015) comment that “in an evolving climate, changes in the sizes and frequencies of rare events, rather than changes in the averages, may be what lead to the most devastating losses of life and the greatest damage to infrastructure.” Rare events are characterized as being of low frequency, and thus have long return periods. Extreme value theory provides techniques to estimate probabilities of events that may never yet have been observed. However, due to the complexity of the underlying phenomena, naive application of this theory can give a sense of false security. For example, in Venezuela, according to Coles and Pericchi (2003), “prior to 1999, simple extreme value techniques were used to assess likely future levels of extreme rainfall, and these gave no particular cause for concern. In December 1999, a daily precipitation event of more than 40 cm, almost three times the magnitude of the previously recorded maximum, caused devastation and an estimated 30,000 deaths.” One issue is that estimated probabilities for very rare events are often based on a few, less extreme, incidents derived from short time series, so the resulting estimates for much rarer events can be quite wrong. Thus it is important to incorporate any external information into analysis, and to allow for possibly small sample sizes.

The purpose of this paper is to describe and attempt to correct for some of the practical issues that can arise when attempting to model extreme rainfall. One key issue is data quality, and a second is the effect of sample size on extremal modeling. In particular, we discuss the effect of block size when fitting extremal models to block maxima, and use the discussion to suggest suitable values for the shape parameter of the extremal distributions.

Section 2 describes and discusses the dataset used for this study, and in Section 3 we present some of its salient features. The effects of different biases are discussed in Section 4, and we discuss the illustrative output, 25-year return level maps, in Section 5.

All analysis was done in the R (R Core Team, 2020) environment.

2 Dataset

Our dataset is publicly available from the Hydrological Information System (HidroWeb), administered by Agência Nacional de Águas (ANA), a Brazilian regulatory agency. In its hydrological information system, ANA has daily rainfall totals measured in tenths of millimetres for 11,368 rain stations, almost all run by ANA, government agencies or public companies.

We removed all values tagged by ANA as doubtful or accumulated (see below), and retained two subsets of stations whose data cover the same periods in order to reflect the climate over fairly homogeneous time windows.
The first subset consists of 1216 stations with daily cumulative rainfall data for the years 1972–2011 and with fewer than 10% of values missing. This interval was chosen to maximize the number of observations in a 40-year span. The second subset contains 164 stations with record lengths over 80 years, starting from 1909 but with differing time windows.

It is essential to check data quality before starting analysis. As 70 stations in the subset with the longest rainfall series lie in the subset of shorter series, we looked individually at series from 1310 stations.

A rainfall series may be useless if it arises from inappropriate measurement conditions. Metadata, such as a station’s history and photographs showing its location and measurement conditions, are crucial to assessing data quality. As stated in the World Meteorological Organization (WMO) guidelines on “climate observation networks and systems,” an observation site should be representative of the climatic regime for which it is intended, as otherwise it represents local features only (Plummer et al., 2003). The amount of rainfall measured is very sensitive to systematic wind-field deformation, so neither completely open exposure nor large objects close to the gauge are desirable. Unfortunately, the ANA database has photos of very few stations, and those that are available usually show the station from only one direction, whereas they should show all possible obstacles around the gauge and the slope of the surrounding land (Jarraud, 2008, p. 38). As the station metadata make it hard to assess whether the WMO guidelines are met and hence whether the series can be regarded as reliable, we turn to internal evidence from the data themselves.

Standard quality control procedures may fail to detect many problems that are nevertheless visible when plotting the data (Hunziker et al., 2017). Plots of the 1310 rainfall time series highlighted the following issues.

1. None of the stations have recorded negative rainfall, but 15 and 291 stations have daily values greater than 500 mm and 200 mm, respectively. To decide whether these values were really observed would require specific knowledge about the local climate and the station. For one station in the Amazon basin the two largest measured daily rainfall values are 999.7 mm and 186.2 mm. The capacity of some measuring devices, such as standard Hellmann rain gauges which can collect only up to 200 mm, might be too small for such regions.

2. Truncation of high precipitation events, whereby only events below a certain value are recorded, or the presence of a frequency peak at specific values that appear as jumps in the empirical distribution function. Truncation can arise when the observer does not properly understand the measurement procedure (Hunziker et al., 2017), and may be hard
Figure 1 Map of Brazil showing the stations in the subset with the 893 shorter and 104 longer rainfall series (left and right panels), selected from the database available in the Hydrological Information System (HidroWeb), administered by ANA. The altitude scale is also shown (meters above mean sea level).

to detect. Even though some of these stations seem to be registering extreme values, we excluded series that showed obvious truncation.

3. Some time series sections have far too few or too many high precipitation events, relative to the norm for that station.

4. There are gaps due to neglect of low precipitation measurements. With time, the scale and the water marks of the device become faded, especially for low values, so the observer might erroneously interpret such marks as zero precipitation. The most frequent cause of this is irregular measurement, i.e., measures are made only after “substantial” rain events, with very low amounts being ignored.

5. Some stations have implausibly long stretches of zeros.

Similar problems were identified by Hunziker et al. (2017) in the Bolivian and Peruvian station networks. We removed obvious outliers from 60 stations and discarded 200 stations that had any of the other problems listed above.

Rainfall accumulates inside the measuring device when the observer does not make a measurement, for example at weekends. Such measurements should be tagged as accumulated but are often overlooked. In order to detect stations with untagged rainfall accumulation, we tested for homogeneity of the proportions of dry and wet days for each day of the week; this was rejected for about 10% of the stations. Since accumulations of two or more days may significantly increase daily extremes, we excluded these stations, leaving 893 and 104 stations remaining in the subsets with short and long series. Figure 1, which shows their locations, reveals that data, particularly long series, are unavailable for much of the interior of the country, and in Section 5 we analyse different regions separately.

Homogenisation of meteorological time series to remove artifi-
cial change points due to changes of instrumentation, observers or procedures is crucial when studying long-term trends (Venema et al., 2018). Application of specialised software for the homogenisation of rainfall data (Wang et al., 2010) showed only a few change points in the upper tails of our time series, and most of these seemed to be false positives; any changes in rainfall extremes seem to be indetectable relative to their intrinsic variability. As we do not study trends, we chose to ignore change points.

3 Extreme values

3.1 Basic distributions

Extremes are usually defined either as block maxima or exceedances over a suitably high threshold. Extreme value theory, which provides a link between these two types of extremes and their generalisations, is fully described in books such as Embrechts, Klüppelberg and Mikosch (1997), Beirlant et al. (2004), de Haan and Ferreira (2006) or Resnick (2006). Coles (2001) and Davison and Huser (2015) focus more on statistical aspects.

Consider a block of independent identically distributed random variables \(X_1, \ldots, X_n\) with parent distribution function \(F\). The distribution of the block maximum, \(Z_n = \max(X_1, \ldots, X_n)\), is \(G_n(z) = F^n(z)\), but this is not useful when \(F\) is unknown, and we instead seek approximations for large \(n\). The extremal types theorem states that if there exist sequences \(\{b_n\}\) and \(\{a_n\} > 0\) such that the renormalised variable \((Z_n - b_n)/a_n\) has a non-degenerate limiting distribution as \(n \to \infty\), then it belongs to one of three families that comprise the generalized extreme-value distribution,

\[
G(z) = \begin{cases} 
\exp \left[ - \left( 1 + \xi \frac{z - \mu}{\sigma} \right)^{-1/\xi} \right], & \xi \neq 0, \\
\exp \left[ - \exp \left\{ - \left( z - \mu \right) / \sigma \right\} \right], & \xi = 0,
\end{cases}
\tag{1}
\]

where \(a_+ = \max(a, 0)\). The support of maxima under this limiting model is \(\{z : 1 + \xi(z - \mu)/\sigma > 0\}\), where \(\mu \in \mathbb{R}\), \(\sigma > 0\), and \(\xi \in \mathbb{R}\) are the location, scale, and shape parameters. The value of \(\xi\) determines the weight of the upper tail of the parent density, providing qualitatively different behaviors for maxima, the ‘three types’ originally derived by Fisher and Tippett (1928):

- if \(\xi < 0\), yielding the reverse Weibull distribution, the support of the density is bounded above at \(\mu - \sigma/\xi\);
- if \(\xi > 0\), yielding the Fréchet distribution, the support of the density has a lower bound at \(\mu - \sigma/\xi\), and the density function decays polynomially as \(z \to \infty\);
• if $\xi = 0$, yielding the Gumbel distribution, the density is supported on
the entire real line and decays exponentially as $z \to \infty$.

Estimates of $\xi$ in environmental applications depend on the phenomenon
studied but typically lie in the interval $(-1/2, 1/2)$. Fits to temperature
maxima and minima tend to yield $\xi < 0$, for example, whereas estimates of
$\xi$ for daily rainfall extremes are typically slightly positive. See Section 4.

The set of parent distributions $F$ for which the block maxima have the
same type of limit distribution is called its max-domain of attraction. Kout-
sosiyannis (2004a) noted that “most types of parent distribution functions
that are used in hydrology, such as exponential, gamma, Weibull, normal
and lognormal, belong to the domain of attraction of the Gumbel distribu-
tion,” and this is one of the reasons for its popularity in hydrology texts.
One difficulty with fixing $\xi = 0$, or any other specific value, is that assess-
ments of uncertainty are then smaller than seems realistic, and we prefer to
allow the shape parameter to adapt to the data.

The discussion above suggests that (1) can be used to model the
maxima of blocks of observations, for $n$ sufficiently large. One issue
that arises with rainfall is that the number of days with positive
rainfall in a block is random, and may vary from very small in arid
or desert regions to very large in areas of regular precipitation. and
in some cases a single distribution will not fit block maxima well.
Eastoe and Tawn (2010) discuss how modeling of the variation in
$n$ may be used to improve standard extreme-value analysis, but
we find the fit of (1) to be adequate without the need for this.

A second issue is potential dependence of daily rainfall totals, and a third
is the assumption that the observations that contribute to the block maxima
are identically distributed. We discuss these in Sections 3.2 and 3.3.

An alternative to analysis of block maxima is that of exceedances
of a threshold $u > 0$. If $F$ lies in the max-domain of attraction
of 1 and the function $a(u) > 0$ is chosen appropriately, then as $u$
increases the distribution of $(X - u)/a(u)$, conditional on $X > u$,
converges to a generalized Pareto distribution (Pickands, 1975)

$$H(x) = 1 - (1 + \xi x/\tau)^{-1/\xi}, \quad x > 0, \quad \xi \neq 0.$$  (2)

The shape parameters of $H$ and $G(x)$ are equal and $\tau = \sigma + \xi(u-\mu)$.
Taking $\xi \to 0$ yields the exponential distribution with mean $\tau$.
The probability that $u$ is exceeded must also be estimated when
fitting threshold exceedances, so use of either (1) or (2) involves
the estimation of three parameters even in the simplest possible situation.
3.2 Dependence

Convergence of block maxima to the generalized extreme-value distribution still holds if long-range dependence at extreme levels is sufficiently weak (Leadbetter, Lindgren and Rootzén, 1983, Chapter 5), but rainy days tend to occur together, which may lead to clusters of days with heavy rainfall. Leadbetter, Lindgren and Rootzén (1983) show that if $X_1, \ldots, X_n$ is a stationary sequence of random variables, then under mild conditions the distribution of $Z_n = \max(X_1, \ldots, X_n)$ is approximately $F^n \theta$, where $\theta \in (0, 1]$, called the extremal index, may be interpreted as the reciprocal of the expected number of high values in a cluster. If we suppose that the shorter series in our dataset are realisations of stationary processes, then we can estimate $\theta$ for each station with the intervals estimator of Ferro and Segers (2003), which is based on an asymptotic result for the times between threshold exceedances; other estimators also exist. The left-hand panel of Figure 2 shows that many of the extremal index estimates equal unity, and in fact the average cluster length for extremes in the ANA rainfall data is only slightly over one day, so it appears that temporal clustering is generally limited.

Spatial dependence can be more substantial. The cross-extremogram (Davis, Mikosch and Cribben, 2012) measures extremal serial dependence between time series. For stationary bivariate time series $(X_t, Y_t)$ it can be reduced to the limiting conditional probability

$$
\rho(h) = \lim_{u \to 1} \mathbb{P}(Y_{t+h} > y_u \mid X_t > x_u) \quad h = 0, 1, 2, \ldots,
$$

where $x_u$ and $y_u$ denote the marginal $u$ quantiles of the series. An estimate
\[ \hat{\rho}_u(h) \] is obtained by replacing the ratio of probabilities in \((3)\) with frequencies for some high \(u\). The right-hand panel of Figure 2, which shows \(\hat{\rho}_{0.98}(h)\) for temporal lags \(h = 0, 1\), suggests the presence of appreciable dependence at distances up to 100 km for extreme events on the same day, but much lower dependence for consecutive days.

We conclude that while spatial dependence of extremes on the same day should be accounted for and may be helpful in borrowing strength between stations, temporal dependence seems to be limited.

### 3.3 Nonstationarity

Environmental processes are typically nonstationary due to seasonality in climate patterns. Variation may also appear due to large-scale phenomena such as the El Niño-Southern Oscillation (ENSO), though such phenomena may be difficult to relate directly to rainfall, since the teleconnection indices that measure them are not directly related to precipitation.

**One approach to include non-stationarity in extremes is by regression analysis, but this leads to difficulties with threshold methods.** The generalized Pareto distribution is threshold-stable: if a random variable follows \((2)\) above a threshold \(u_0\) with scale and shape parameters \(\tau_0\) and \(\xi\), then the excesses above another threshold \(u > u_0\) follow \((2)\) with scale and shape parameters \(\tau_u = \tau_0 + \xi(u - u_0)\) and \(\xi\). This implies that if the threshold is varied in nonstationary contexts, when performing sensitivity analysis, for instance, regression structures for \(\tau_u\) and \(\tau_0\) will be related in somewhat artificial ways; log-linear models for \(\tau_0\) and \(\tau\) will be incompatible unless \(\xi = 0\), for example. For this reason Eastoe and Tawn (2009) suggested preprocessing the data before extreme value analysis, first modeling any nonstationarity in the bulk of the data, and then fitting an extremal model to the resulting residuals. Since the extremes and bulk of the data might have different patterns of nonstationarity, Eastoe and Tawn still include covariates in the generalized Pareto model, hoping that any lack of threshold stability will be limited.

Despite our efforts to eliminate poor-quality series, we are unsure about the quality of some smaller measurements in the series we retained, and, moreover, preprocessing seems best confined to cases where the physical origin of the nonstationarity is well understood (Davison and Smith, 1990). Since the parent distribution may be nonstationary and we wanted to avoid preprocessing, we finally chose to use block maxima. This leaves the problem of choosing the block size. Taking annual maxima obviates the need to model seasonality, but could lead to a large loss of information. Taking
blocks of length one month, however, yields many short blocks, sometimes with very few wet days in the dry season, and then (1) may describe their maxima poorly.

In Section 4.2, we focus on the shape parameter of the generalized extreme-value distribution, which can be considered as constant over time. To study this we took monthly maxima in order to use as much information as possible, but we used only the six rainiest consecutive months for each station. These may be found from the map in Figure 2 of Rao et al. (2016). After a detailed study of rainfall patterns in Brazil, these authors state that the central parts of the country have the main characteristic of a monsoon region: “6 months of rain during the austral summer followed by 6 months of scanty rainfall in austral winter.” Exceptions are the northern parts of the Amazon basin, where the wet season happens in the boreal summer (mostly from June to August), the south region, where rainfall is relatively uniform across time, and northeast Brazil, where most of the rainfall occurs in a three-month period that varies from region to region.

4 Bias and approximation

4.1 Penultimate approximation and estimation bias

In their seminal paper, Fisher and Tippett (1928) pointed out that convergence of maxima of normal variables to the limiting Gumbel distribution is very slow, and that for finite $n$ the reverse Weibull distribution provides a better, so-called penultimate, approximation for the distribution of the maxima. Such an approximation will often have $\xi \neq 0$ even if the limiting distribution is of Gumbel form. One can approximate the shape parameter $\xi_n$ for finite $n$ by $r'(b_n)$, where $r(x) = \{1 - F(x)\} / f(x)$ is the reciprocal hazard function, $F(x)$ is the parent distribution function, $f(x) = F'(x)$, and $b_n = F^{-1}(1 - 1/n)$, but this requires knowledge of $F$; see Smith (1987), for example.

Wilson and Toumi (2005) express precipitation as the product of mass flux, specific humidity, and precipitation efficiency, and assume that these have independent approximate normal distributions for daily totals. Using large deviation theory, they find that under their assumptions the tail of the daily rainfall distribution is Weibull with shape parameter $k = 2/3$, i.e., for large $x$,

$$F(x) = 1 - \exp\{- (x/\lambda)^k\}, \quad x > 0, \lambda > 0.$$  \hspace{1cm} (4)

They claim that the shape of the tail should be “largely unaffected by climate change,” and “the robustness of the shape parameter may now seem unsurprising given the physical basis of moisture conservation.”
Empirical analysis of global data using the extreme-value models described in Section 3.1 shows that $\xi \approx 0.1$ on average (Koutsoyiannis, 2004b; Wilson and Toumi, 2005; Papalexiou and Koutsoyiannis, 2013; Serinaldi and Kilsby, 2014). Interestingly, if daily rainfall arose from (4) with $k = 2/3$, for $n = 90$ and $n = 365$ wet days, we would obtain $\xi_{90} = r'(b_{90}) = 0.11$ and $\xi_{365} = r'(b_{365}) = 0.08$. As annual daily rainfall maxima typically have $n$ of order $10^2$, the penultimate approximation for a distribution close to the Weibull proposed by Wilson and Toumi (2005) would therefore have $\xi_n \approx 0.1$, even though $\xi_\infty = 0$. However, the small-sample downward bias of the maximum likelihood estimator $\hat{\xi}$ (Hosking, Wallis and Wood, 1985) might yield estimates of $\xi$ around zero.

To assess how penultimate approximation and small-sample estimation bias might interact, we simulated 1,000 series of length 100 years, with each year having $n = 90$ or $n = 365$ observations, from the Weibull distribution with $k = 2/3$; here $n$ represents the number of wet days per year. We fitted the generalized extreme-value distribution by maximum likelihood to the annual maxima of the first $m = 10, 15, 20, \ldots, 100$ years of each series, and then, for each $m$, computed the average and the standard deviation of the resulting shape parameter estimates. Figure 3 shows that these empirical values closely follow curves of the form

$$\phi_n(m) = \alpha + \beta m^{-\gamma}, \quad \alpha, \beta \in \mathbb{R}, \gamma > 0,$$

where $\phi_n(m)$ denotes either the mean of the estimates, $\mu_n(m)$, or their standard deviation, $\sigma_n(m)$. Here $\alpha$ is the limit of $\phi_n(m)$ when $m \to \infty$, and $\beta$ and $\gamma$ determine the rate at which $\phi_n(m) \to \alpha$ as $m \to \infty$. The larger the value of $\beta$, the more rapidly $\phi_n(m)$ explodes for small $m$. The estimates of these parameters are shown in Table 1. When the block size $n$ equals 365, then as the number of block maxima $m$ increases, the average of $\hat{\xi}$ increases towards the penultimate approximation value $\xi_{365}$, but the negative bias of $\hat{\xi}$ is evident for small $m$. Interestingly, $\mu_{90}(m) \approx 0.1$: if daily rainfall totals can be described by a Weibull distribution with $k = 2/3$ and the frequency of wet days is around 25%, then if we fit the generalized extreme-value distribution to annual maxima, we expect to obtain $\hat{\xi} \approx 0.1$, largely irrespective of the number of years of data.

4.2 Application to the longer series

To assess the value of bias correction for the ANA data, we consider the impact of record length on the estimated shape parameter for the longer data series. For each station, we fit the generalized extreme-value distribution by maximum likelihood to the rainy season monthly maxima. To account for
Figure 3  Average and standard deviation of the maximum likelihood estimates for the shape parameter of the generalized extreme-value distribution based on subsamples of the 1,000 simulated rainfall series. The curves $\mu_{90}(m)$ and $\mu_{365}(m)$ were fitted to the average of the estimates (thick lines), while $\sigma_{90}(m)$ and $\sigma_{365}(m)$ were fitted to their standard deviation (dotted lines). Block sizes $n = 90, 365$ are represented by black and grey colours, respectively.

Table 1  Estimates with standard errors (in parentheses) for the parameters of equation (5) according to the block size $n$ (the number of wet days per year) from the Weibull distribution. There are no estimates of $\beta$ and $\gamma$ for $\mu_{90}$ because the curve is approximately a constant, denoted here as $\alpha$.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{90}$</td>
<td>0.101 (0.003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{365}$</td>
<td>0.072 (0.001)</td>
<td>-5.1 (5.8)</td>
<td>2.15 (0.48)</td>
</tr>
<tr>
<td>$\sigma_{90}$</td>
<td>0.035 (0.002)</td>
<td>2.1 (0.1)</td>
<td>0.82 (0.02)</td>
</tr>
<tr>
<td>$\sigma_{365}$</td>
<td>0.04 (0.003)</td>
<td>2.3 (0.1)</td>
<td>0.87 (0.02)</td>
</tr>
</tbody>
</table>

seasonality and possible trend, we suppose that the location and log-scale parameters vary as

$$p(t) + \beta_1 \sin(2\pi t / 365) + \beta_2 \cos(2\pi t / 365),$$

where $t$ (days) indexes when the maximum occurred, $\beta_2$ and $\beta_3$ are the amplitudes of the harmonic terms, and $p$ is a polynomial of degree four. We then proceed as in Section 4.1, splitting the series into subsamples, taking
the first \( m = 10, 15, 20, \ldots, 95 \) years, and refitting the model. Then, for each subsample, we computed the average and the standard deviation of the shape parameter estimates. Figure 4 graphs these empirical values against \( m \). We excluded nine stations with very discrepant estimates for \( \xi \), leaving 95 stations.

Let \( \mu(m) \) and \( \sigma(m) \) denote the functional curves for the average and the standard deviation of the shape parameter estimates. We set the latter to be \( \sigma(m) = \alpha_\sigma + \beta m^{-\gamma} \), whereas for the average we found that the function

\[
\mu(m) = \frac{\alpha_\mu}{1 + \exp\left\{ (\nu - m)/\delta \right\}}, \quad \alpha_\mu \in \mathbb{R}, \nu, \delta > 0, \tag{7}
\]

was more appropriate than (5); here \( \alpha_\mu \) is the horizontal asymptote as \( m \to \infty \), \( \nu \) is the inflection point of the curve, and \( \delta \) represents the distance between \( \nu \) and the point at which \( \mu(m) \approx 3\alpha_\mu/4 \). For very small \( m \) the values of \( \hat{\xi} \) are highly variable but show overall downward bias. The estimate for the asymptotic value of \( \mu(m) \) is \( \hat{\alpha}_\mu = 0.06 \), lower than the value around 0.1 found in other studies (Koutsoyiannis, 2004b; Wilson and Toumi, 2005; Papalexiou and Koutsoyiannis, 2013; Serinaldi and Kilsby, 2014). This may be because our sample is most representative of northeast Brazil (65% of the 95 stations), followed by the southeast (21%) and the south (13%). Overall it seems preferable to use the asymptotic value of 0.06 for the shape parameter as prior information in a Bayesian setting.

## 5 Spatial modeling

As mentioned above, accurate estimation of extreme quantiles of daily rainfall is important for many purposes. Extreme value theory provides a class of models that allow extrapolation to events both rarer than those previously seen, and spatial modeling extends them to sites for which no data are available, but the uncertainty due to difficulties in estimating both the shape parameter and the spatial structure can be very large. In an ideal world we would have a dense network of reliable long time series, but Figure 1 shows that the Brazilian stations with longer series are too sparse to allow reliable spatial interpolation over most of the country. On the other hand it seems plausible that one may borrow strength for spatial estimation using the shorter series.

We fit the Bayesian hierarchical model described in Davison, Padoan and Ribatet (2012) to the 893 shorter time series, and illustrate the results by computing spatial estimates of 25-year return levels. In principle we could include temporal covariates, but the current implementation of this model
in the R package \texttt{SpatialExtremes} only allows spatial covariates, so we take rainfall annual maxima, which are effectively the maxima for the rainy season, as our response.

Let $Z_t(s_d)$ denote rainfall annual maxima in year $t$ and station $d$ at location $s_d$, for $t = 1, \ldots, n$ and $d = 1, \ldots, D$. Conditional on the values of three independent Gaussian processes that describe spatial variation, we take the $Z_t(s_d)$ to be independent and distributed according to a generalized extreme-value distribution,

$$Z_t(s_d) \mid \{\eta(s_d), \tau(s_d), \xi(s_d)\} \sim \text{GEV}\{\eta(s_d), \tau(s_d), \xi(s_d)\}.$$ 

We assume that the generalized extreme-value parameters vary according to three independent Gaussian processes:

$$\phi(s) = f_\phi(s; \beta_\phi) + S(s; \alpha_\phi, \lambda_\phi), \quad \phi = \eta, \tau, \xi,$$

where $f_\phi$ is a deterministic function depending on regression parameters $\beta_\phi$, and $S_\phi$ is a zero-mean stationary Gaussian process with exponential covariance function $\alpha_\phi \exp(-\|h\|/\lambda_\phi)$ and sill and scale parameters $\alpha_\phi, \lambda_\phi > 0$.
0. We allow the means of the location and scale parameters to depend on longitude, latitude, and mean annual precipitation (MAP) via the structure

\[ f_\eta(s) = \beta_{0,\eta} + \beta_{1,\eta} \text{lon}(s) + \beta_{2,\eta} \text{lat}(s) + \beta_{3,\eta}(s) \text{MAP}(s), \quad (8) \]
\[ f_\tau(s) = \beta_{0,\tau} + \beta_{1,\tau} \text{lon}(s) + \beta_{2,\tau} \text{lat}(s) + \beta_{3,\tau}(s) \text{MAP}(s), \quad (9) \]
\[ f_\xi(s) = \beta_{0,\xi}. \quad (10) \]

Inclusion of MAP should accommodate some factors that longitude and latitude cannot capture, such as variation in altitude, that might explain variation in extreme rainfall.

A joint prior density must be defined for the regression parameters, \( \beta_\eta = (\beta_{0,\eta}, \beta_{1,\eta}, \beta_{2,\eta})^T \), \( \beta_\tau = (\beta_{0,\tau}, \beta_{1,\tau}, \beta_{2,\tau})^T \) and \( \beta_{0,\xi} \), and the parameters of the covariance functions, \( \alpha_\eta, \alpha_\tau, \alpha_\xi, \lambda_\eta, \lambda_\tau \) and \( \lambda_\xi \). Conjugate priors are used whenever possible in order to reduce the computational burden.

In order to reduce the uncertainty for extrapolation of return levels, the prior mean for the shape parameter was set to the asymptotic value 0.06 found from the long series in Section 4.2, and its standard deviation was set to 0.02, giving a fairly informative prior for this parameter. Maximum likelihood estimates were used to inform the prior means of the other parameters, with the prior variances set to large values. When estimating the parameters via maximum likelihood, we verified the suitability of the exponential covariance function. We also found that spatial correlation is almost non-existent for the shape parameter. In summary, we attributed independent

- normal priors for the regression parameters with means \( \mu^*_\eta = (80, 0, 0, 0)^T \), \( \mu^*_\tau = (20, 0, 0, 0)^T \) and \( \mu^*_\xi = 0.06 \), and covariance matrices \( \Sigma^*_\eta = \Sigma^*_\tau = \text{diag}(400, 100, 100, 100) \) and \( \Sigma^*_\xi = 0.0004 \),
- inverse gamma distributions as priors for \( \alpha \), with shape parameters \( \kappa^*_\alpha_\eta = \kappa^*_\alpha_\tau = \kappa^*_\alpha_\xi = 1/2 \), and scale parameters \( \theta^*_\alpha_\eta = 180/2, \theta^*_\alpha_\tau = 30/2, \theta^*_\alpha_\xi = 0.03/2 \), and
- gamma priors for \( \lambda \) with shapes \( \kappa^*_\lambda_\eta = \kappa^*_\lambda_\tau = 10 \) and \( \kappa^*_\lambda_\xi = 5 \), and scales \( \theta^*_\lambda_\eta = \theta^*_\lambda_\tau = 7 \) and \( \theta^*_\lambda_\xi = 3 \).

A Gibbs sampler described by Davison, Padoan and Ribatet (2012) used to fit this model requires inversion of \( D \times D \) matrices at each iteration and is slow for large \( D \). We ran 20,000 iterations of the Markov chain, thinned by a factor of 30 and preceded by a burn-in of 5,000 iterations. Due to the heavy computational burden, we fitted the model to separate regions indicated in Table 2. For each region, we used two-thirds of the stations for fitting and left the remaining third for validation, except for the state of São Paulo, where we used just one third of the many stations for the fitting. Standard
Modeling extreme rainfall

Table 2 Total and fitting number of stations per region.

<table>
<thead>
<tr>
<th>Region</th>
<th>Abbreviation</th>
<th>Total</th>
<th>Fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>North*</td>
<td>North</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>Bahia, Sergipe</td>
<td>BA, SE</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>Rio Grande do Sul</td>
<td>RS</td>
<td>71</td>
<td>47</td>
</tr>
<tr>
<td>Paraná</td>
<td>PR</td>
<td>165</td>
<td>110</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>RJ</td>
<td>52</td>
<td>35</td>
</tr>
<tr>
<td>Espírito Santo</td>
<td>ES</td>
<td>52</td>
<td>35</td>
</tr>
<tr>
<td>Minas Gerais</td>
<td>MG</td>
<td>103</td>
<td>69</td>
</tr>
<tr>
<td>São Paulo</td>
<td>SP</td>
<td>322</td>
<td>107</td>
</tr>
</tbody>
</table>

* The “north” region includes the states of Tocantins, Amapá, Pará, Piauí, and Maranhão.

diagnostics suggest no lack of convergence, except for the state of Santa Catarina, which we do not discuss below.

Figure 5 shows the pointwise posterior means for the three parameters of the generalized extreme-value distribution and a 25-year return level map for the region of Rio de Janeiro. The estimates for the metropolitan area of Rio de Janeiro contrast to those for the Serra do Mar mountain range, where the worst natural disaster in Brazil’s history occurred. Even with a quite informative prior distribution for the shape parameter, we obtained negative estimates in some parts of Serra do Mar, while for the city of Rio the estimate is close to 0.1. The city also has the largest estimate for the scale parameter in the entire country. The estimates for the city are what we would expect for the mountainous region, since both were affected by the same extremal event of January 2011. However, the map indicates smaller return levels in the inner regions of the state of Rio, where most of the stations we selected are located. If there were more stations with reliable data in the city of Rio, we might obtain lower estimated return levels for the region.

Residual-checking is awkward in a Bayesian context, but we can compare quantile plots for data from each station with such plots derived from maximum likelihood fits of the univariate (stationary) generalized extreme-value distribution. Because the latent model provides a spatially-smoothed fit, the empirical and model quantiles differ a little, unlike the maximum likelihood fits. Even so, for the most part, the QQ plots are quite close not only for both fitting and validation stations, but also for the longer series (which were not used in the fitting process). Figure 6 shows such quantile plots for three stations in Minas Gerais. For the station very close to the
border (second line), the two models provide similar estimates for all parameters, while for the other two stations, which are very close to each other, the estimated shape parameters have different signs for the two models. The latent model provides higher estimates for the 25-year return level for about two-thirds of the series. For the longer series in the north and northeast regions, and in the state of Rio Grande do Sul, the latent model return level estimates are around 20 mm higher than those for the univariate models.

Our climate effects comprise longitude, latitude, and mean annual precipitation, but we also fitted the latent model without this last covariate, and compared these nested models using the deviance information criterion (Spiegelhalter et al., 2002); see Table 3. For the states of Paraná, Rio de Janeiro, Minas Gerais, and São Paulo, the annual mean precipitation appears significant, so we used it in making the return level maps, some of which are shown in Figure 7; the others are provided in the Supplementary

Figure 5 Pointwise posterior mean for the three parameters of the generalized extreme-value distribution and the 25-year return level for daily rainfall (mm) for the region of Rio de Janeiro.
Material, which also contains maps of the upper bound of the 95% credible intervals for the return levels, in case conservative return level estimates are needed. The estimates reflect what is commonly known in Brazil, a country of continental dimensions: there is a strong difference in precipitation between the south and north regions. The patterns of variation in the latitude and longitude for the return levels are very clear for Rio Grande do Sul, Paraná, and the north region, perhaps because those
Table 3 Deviance information criterion (according to region). Model 1 only has longitude and latitude as covariates, while Model 2 also includes mean annual precipitation. Lower is better.

<table>
<thead>
<tr>
<th>Region</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>6,884</td>
<td>6,884</td>
</tr>
<tr>
<td>BA, SE</td>
<td>8,058</td>
<td>8,112</td>
</tr>
<tr>
<td>RS</td>
<td>17,691</td>
<td>17,688</td>
</tr>
<tr>
<td>PR</td>
<td>38,717</td>
<td>38,666</td>
</tr>
<tr>
<td>RJ</td>
<td>13,012</td>
<td>13,001</td>
</tr>
<tr>
<td>ES</td>
<td>12,704</td>
<td>12,699</td>
</tr>
<tr>
<td>MG</td>
<td>24,839</td>
<td>24,820</td>
</tr>
<tr>
<td>SP</td>
<td>39,224</td>
<td>39,107</td>
</tr>
</tbody>
</table>

are the regions where ENSO impacts precipitation the most. In São Paulo, return levels vary mostly between 120 mm and 150 mm, but they increase dramatically towards the Serra do Mar mountain range by the coast, where they approach 220 mm, a level also estimated for the city of Rio de Janeiro. Espírito Santo, Minas Gerais, and the northeast exhibit no obvious pattern: the return level estimates are more uniform, varying between 120 mm and 150 mm. Apart from the coastal areas of São Paulo and Rio de Janeiro (and perhaps the Serra da Mantiqueira mountain range, in Rio), high estimates, close to 170 mm, are observed in the inner regions of the states most affected by ENSO (Rio Grande do Sul, Paraná, and the north).

Neighbouring stations can generally be expected to have similar return levels, so a spatial model smooths out the marginal maximum likelihood estimates. As mentioned by Davison, Padoan and Ribatet (2012), the main advantage of the latent variable approach is that, with simple deterministic trend surfaces, it captures complex spatial variation in the return levels. However, the assumption of conditional independence given the latent process is inappropriate, because of spatial dependence. This assumption leads to implausible spatial process realisations, and mis-estimates the multivariate distribution of extremes for several sites. Since our emphasis is on return levels at single sites, including those where data are unavailable, this is not a key concern. Joint modeling would require more complex techniques and is outside the scope of the present paper.
Figure 7  Pointwise posterior mean (left panels) and standard deviation (right panels) for the 25-year return level for daily rainfall (mm). The points represent the marginal maximum likelihood estimates. Points with a cross correspond to validation stations.
6 Discussion

An ideal world would provide a large number of stations with long high-quality measurement series that were truly representative of the climatic regime of their sites. We did our best to select reliable rain stations, despite the lack of important metadata like the stations’ history and location and measurement conditions. Hunziker et al. (2017) suggest methods for the reconstruction or improvement of metadata.

Despite the large number of stations, only 104 with record lengths greater than 80 years satisfied our selection criteria. Since this subset is too sparse to produce return level maps for daily rainfall, we based our main analysis on 893 shorter series, covering the period from 1972 to 2011. The longer series were used to extract information about the shape parameter of the extreme-value model, complementing the findings of Papalexiou and Koutsoyiannis (2013) and Serinaldi and Kilsby (2014), with another large dataset. These large sample studies allow us to formulate a correction for the estimated parameters that not only takes the sample-size bias into account, but also the penultimate approximation (or the implicit block size), and can be used as prior input for Bayesian modeling. The latent variable model that we fitted seems to perform very well when used for the validation stations or the longer series. It provides a smooth spatial fit, reducing uncertainty and allowing extrapolation to some regions without reliable data. The highest estimates for the 25-year return level are found in the coastal areas of São Paulo and Rio de Janeiro, and the inland of Rio Grande do Sul, Paraná, and the north region, where the ENSO phenomenon most affects extreme precipitation.

Acknowledgements

We thank the Agência Nacional de Águas (ANA) for providing the rainfall series, and are especially grateful to Andrelina Laura dos Santos for her help. We thank Stefan Hunziker of the University of Bern for helpful input on data quality, the reviewers for their constructive remarks, and gratefully acknowledge financial or material support from the Universidade Estadual de Maringá (UEM) and the Brazilian federal agency, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).
References


