Extreme value comparison of CanCM4 simulations and observations

Mickey Warner

1 Abstract

We fit a Bayesian hierarchical model to threshold exceedances from CanCM4 climate simulations. Three simulation classes are analyzed: decadal, historical, and pre-industrial control. These are compared against on observation product. We find that in some domains, the simulations are in agreement with the observations, but in others can be quite different.

2 Introduction

The Fourth Generation Coupled Global Climate Model (CanCM4) produces a wide array of atmospheric conditions across the globe. Two variables will be analyzed: precipitation (labled `pr`, in meters) and maximum temperature (labeled `tasmax`, in Kelvin). Three experimental classes that are of particular interest are decadal, historical, and pre-industrial control runs.

The decadal simulations provide climate estimates for ten years into the future, after conditioning on weather conditions at the time. We consider two decades in this analysis: 1962–1971 and 1990–1999, which are conditioned on climate states in 1961 and 1989, respectively. Historical simulations are obtained for the years 1961–2005 and are noted for including events that affect the climate such as volcanoes. The pre-industrial control, or simply control, simulations begin at climate conditions comparable to those preceding the industrial revolution and are run over a thousand years into the future. Decadal and historical simulations are run at $R = 10$ different input settings. To obtain $R = 10$ "replicates" for the control simulations, we randomly select ten non-overlapping 10-year periods.

An observation product is obtained from Maurer et al. (2002). The observations are based on daily measurements from weather stations throughout the United States and are interpolated onto a fine grid. To make the observations comparable to the climate simulations, we take weighted sums or averages of the climate simulations and just sums or averages of the observations. See section 3 for details, along with other changes made to the data in preparation for analysis.

Being a threshold exceedance analysis, we must concern ourselves with exceedances occuring together within a short time. This is handled by studying the extremal index θ, a measure of dependence among the extremes. With an estimate for θ, we can "decluster" the exceedances to obtain independent clusters. The method for estimating θ and declustering has been generalized to the hierarchical setting, see section 4.

Having replicates of a time-series suggests the use of a hierarchical model, described in detail in section 5.2. Under such a framework we can model each series separately, while assuming these series come from a larger population. In the analysis, we will place focus on the mean of this larger population, being akin to the ensemble average in a climate study.
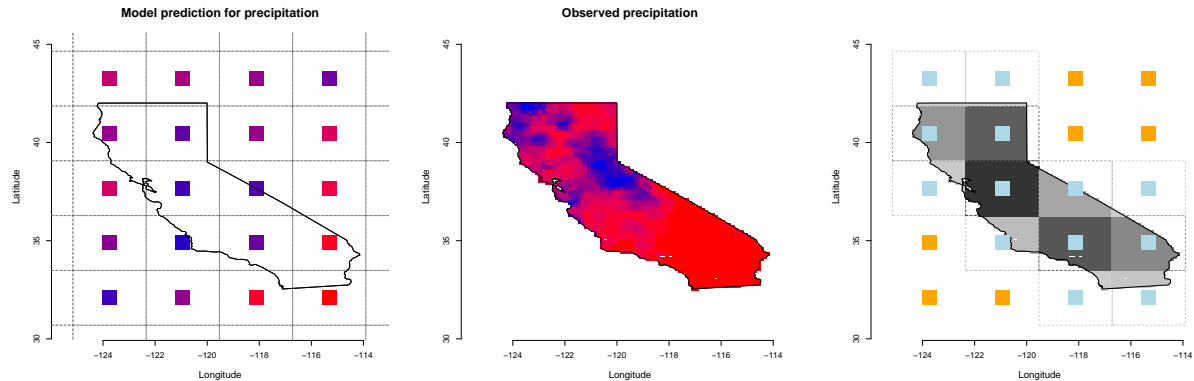
Figure 1: Left: CanCM4 simulation locations. Center: Observation locations. Right: method for computing weighted sum or average for CanCM4 to make values comparable with observations; the lighter gray points mean less weight is applied to the climate simulations and the darker gray means more weight. The data shown are from a single day in January.

The main focus of this paper is to compare the extreme values of the observation product with those of CanCM4 climate simulations. The comparison is done mostly through visualization of the differences between posterior parameters and a useful quantity called the return level (section 6).

3 Data pre-processing

3.1 Aggregation

In this subsection, we describe how the simulations and observations were made to be comparable. Figure 1 shows the spatial locations of each data source. The plots show only California, but the climate simulations were over the entire globe and the observation product over the United States.

We will analyze precipitation and temperature over both California and the United States. In each case, we take the climate locations and create non-overlapping cells, or rectangles, such that each location is roughly in the center of the cell. Then we count the number of locations from the observation product that are contained with each cell. The number of locations within the cells are used to weight the climate simulations (the right-most plot in Figure 1 shows which climate simulation locations have non-zero weight). For precipitation, we take a weighted sum and for temperature a weight average. No weighting is used for the observations. Instead, a straight sum or average of all locations within our region of interest (either California or U.S.) is used. This method places the simulations and the observations on the same scale and yields time-series on daily time scales.

3.2 De-trending

Each time-series is "de-trended" prior to declustering and parameter estimation. This is accomplished through the use of dynamic linear models (DLMs). We will review some basic concepts for DLMs, see Prado and West (2010) chapter 4 for more details.
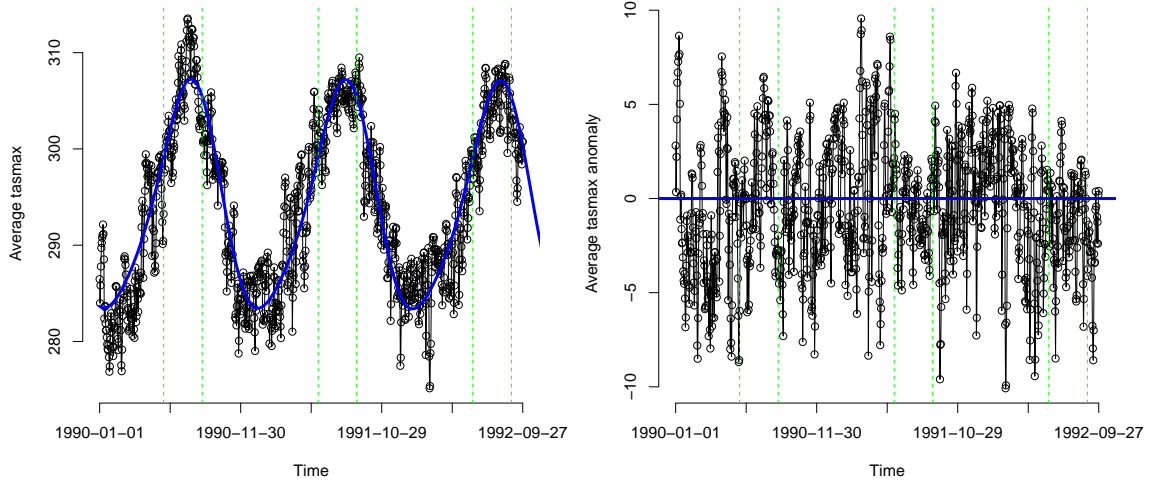
Figure 2: One of the DLMs used to calculate the anomalies. Shown is one of the decadal replicates of average `tasmax` in California for about the first two and one-half years of the time-series. The green dashed lines mark the beginning and the end of the summer months.

A normal DLM is specified by the quadruple $\{\mathbf{F}_t, v_t, \mathbf{G}_t, \mathbf{W}_t\}$ which determine how a univariate time series $y_1, \ldots, y_T$ is modeled over time. We assume

$$
\begin{aligned}
y_t &= \mathbf{F}_t^\top \boldsymbol{\theta}_t + \nu_t, & \nu_t &\sim N(0, v_t) \\
\boldsymbol{\theta}_t &= \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \mathbf{w}_t & \mathbf{w}_t &\sim N(\mathbf{0}, \mathbf{W}_t)
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\theta}_t$ is the length p state vector, $\mathbf{F}_t$ is a length p vector of known constants are regressors, ν_t is obersvation noise, $\mathbf{G}_t$ is the known $p \times p$ state evolution matrix, and $\mathbf{w}_t$ is the state evolution noise. Note that ν_s and $\mathbf{w}_t$ are independent and mutually independent.

An advantage to model (1) is its capability in yielding a smooth and flexible mean across time. After conditioning on the data up to time T, we extrapolate back over time to obtain the posterior distributions $p(\boldsymbol{\theta}_t | D_T)$ for all $t < T$, which have mean $\mathbf{a}_t$. Using these distributions, and given $\mathbf{F}_t$, the mean of y_t is simply $\mathbf{F}_t^\top \mathbf{a}_t$ (we refer the reader to Prado and West (2010) for the algorithmic details).

We must omit further details in the interest of space. Our DLM is finalized in the following way. We construct $\mathbf{F}_t$ and $\mathbf{G}_t$ such that the evolution of $\boldsymbol{\theta}_t$ has annual and semi-annual periods, i.e. the first and second harmonics. Higher harmonics did not seem to make significant contributions in modeling the time-series. A discount factor of $\delta = 0.9999$ was chosen, signifying low systematic variance. We assume the prior for v_t is an inverse gamma having sensible shape and scale parameters.

In the end, we are left with what is shown in Figure 2. The blue line in the left plot is the mean of y_t, $\mathbf{F}_t^\top \mathbf{a}_t$, given the whole time series. The interior of the vertical green lines mark the summer months. The right plot is the result of subtracting the observation y_t with the mean from the DLM, which produces a roughly stationary sequence. Thus, in our extreme value analysis we work in the anomaly space.

3

For each time-series to be analyzed, we fit a DLM having the characteristics described above to obtain the anomalies. When working within a specific season, either winter (December, January, February) or summer (June, July, August), we extract just those days contained in the season and treat the beginning of a season as the day that follows immediately after the last day. So, for example in winter, 28 February is followed immediately by 1 December.

4   Extremal Index

The threshold exceedance model described in section 5 relies on an assumption of independence which is unrealistic for a time-series. When there is dependence between the random variables, the extremes are related according to the so-called extremal index, denoted by $\theta \in (0, 1]$. The extremal index can loosely be interpreted as

$$\theta = (\text{limiting mean cluster size})^{-1}.$$

As an example, suppose $\theta = 0.5$, then we would expect exceedances of a large threshold to occur in pairs; for $\theta = 0.33$, in groups of 3.

Ferro and Segers (2003) show that the extremal index arises in the limiting distribution of the times between exceedances of a threshold. If T_θ is the random variable for interexceedance times in the limit, then T_θ is distributed according to the mixture

$$(1 - \theta)\epsilon_0 + \theta\mu_\theta \tag{2}$$

where ϵ_0 is the degenerate probability distribution at 0 and μ_θ is an exponential distribution with mean θ^{-1}. This means that the role of θ is two-fold: it is both the proportion of non-zero interexceedance times and the inverse mean of non-zero interexceedance times. This poses a challenge when estimating θ since is it impossible to observe an interexceedance time of zero in practice.

We next describe the hierarchical model used to esimate θ. This is distinct from the threshold exceedance model and is used only in getting a single estimate for θ, which is used to decluster the exceedances and to calculate return levels.

4.1   Estimation

Ferro and Segers (2003) propose estimating θ by considering the interexceedance times, the length of time between each random variable that exceeds the threshold. Suppose we have observations $X_1, \ldots, X_n$. For a threshold u, the N exceedances $Y_i = X_i - u$ given $X_i > u$ occur at times $1 \leq j_1 < \cdots < j_N \leq n$. The observed interexceedance times are given by $T_i = j_{i+1} - j_i$ for $i = 1, \ldots, N - 1$. The following log-likelihood is then provided

$$l(\theta, p; \mathbf{T}) = m_1 \log(1 - \theta p^\theta) + (N - 1 - m_1)\{\log(\theta) + \log(1 - p^\theta)\}$$
$$+ \theta \log(p) \sum_{i=1}^{N-1} (T_i - 1) \tag{3}$$

where p is the probability of not exceeding the threshold. We require this likelihood to be used in a hierarchical model.

Suppose we have R replicates from a climate model with values from replicate i denoted $X_{i,1}, \ldots, X_{i,n}$. If we assume these simulations are independent from each other, then we expect there to be R unique extremal indices $\theta_1, \ldots, \theta_R$. However, since these all come from the same climate model, we may wish to assume that the θ_i come from a common distribution,

$$\theta_i \overset{iid}{\sim} Beta\left(\theta\nu, (1-\theta)\nu\right).$$

Under model (3), we place a similar prior on the p_i,

$$p_i \overset{iid}{\sim} Beta\left(p\tau, (1-p)\tau\right).$$

The model is completed by choosing priors for θ, ν, p, and τ—the latter two parameters being required only for model (3). We assume

$$\theta \sim Beta(a_\theta, b_\theta)$$
$$\nu \sim Gamma(a_\nu, b_\nu)$$
$$p \sim Beta(a_p, b_p)$$
$$\tau \sim Gamma(a_\tau, b_\tau)$$

with the hyperparameters chosen to be

$$
\begin{array}{lll}
\theta: & a_\theta = 1 & b_\theta = 1/2 \\
\nu: & a_\nu = 1 & b_\nu = 1/10 \\
p: & a_p = 100\hat{F} & b_p = 100(1 - \hat{F}) \\
\tau: & a_\tau = 1 & b_\tau = 1/10
\end{array}
$$

where $\hat{F} = \sum_{i=1}^{R} \sum_{j=1}^{n} \mathbb{1}(X_{i,j} \leq u)$. Our parametrization for the gamma random variables are such that $X \sim Gamma(\alpha, \beta)$ has mean α/β. The prior values for θ attempt to mitigate some of the issues surrounding model (3)

By assuming indepedence between the simulations, we can construct the following log-likelihood

$$L = \sum_{i=1}^{R} l(\theta_i, p_i; \mathbf{T}^{(i)}) \tag{4}$$

where $\mathbf{T}^{(i)}$ is the vector of interexceedance times for replicate i having length N_i.

4.2 Declustering

Declustering is done as given in Ferro and Segers (2003). Each replicate is declustered separately. Let $\hat{\theta}_i$ be the posterior mean of the extremal index of each replicate. Calculate $C_i = \lfloor \hat{\theta}_i N_i \rfloor + 1$, the estimated number of independent clusters. Let T_{C_i} be the C_ith largest interexceedance time in $\mathbf{T}^{(i)}$. In the case of ties, decrement C_i by one until T_{C_i+1} is strictly greater than T_{C_i}. Clusters are formed by grouping the exceedances that are separated in

time by no more than T_{C_i}. In other words, two exceedances are in the same cluster if their interexceedance time is less than or equal to T_{C_i}.

The C_i clusters that will be formed using the above scheme are assumed to be independent. For each cluster we compute the cluster maximum, this being the ultimate quantity used in our inference.

5 Threshold exceedance model

5.1 Univariate

Under some mild assumptions, for random variable X and for large enough u, the distribution of $X - u$ (the exceedance), conditional on $X > u$ is approximately

$$P(X - u \leq y | X > u) \approx H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi} \tag{5}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\sigma) > 0\}$. $H(y)$ is the distribution function for a generalized Pareto random variable with shape paremeter $\xi \in \mathbb{R}$ and scale $\sigma > 0$.

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables and u be a high threshold. Define $Y_i = X_i - u$ for $X_i > u$ be the k exceedances. The likelihood of (ξ, σ) is derived from (5) as

$$L(y_1, \ldots, y_k; \sigma, \xi) = \sigma^{-k} \sum_{i=1}^{k} \left(1 + \frac{\xi y_i}{\sigma}\right)_+^{-1/\xi - 1} \tag{6}$$

where $z_+ = \max(z, 0)$. This provides the basis for an extreme value analysis. For example, after declustering, the cluster maxima (which are roughly independent) may be fit using likelihood (6).

5.2 Hierarchical model

Suppose we have R replicates or computer simulations, each with n_i observations, for $i = 1, \ldots, R$. Let X_{ij} denote the jth observation in replicate i. We assume

$$X_{ij} \sim F_i, \qquad i = 1, \ldots, R, \qquad j = 1, \ldots, n_i$$

and all X_{ij} are mutually conditionally independent. From (6), we deriv

For a fixed u and each i, define the following sets:

$$A_i = \{j : x_{ij} \leq u\}, \quad A_i^c = \{j : x_{ij} > u\}$$

where $|A_i| = n_i - k_i$ and $|A_i^c| = k_i$ with k_i being the number of exceedances in replicate i. We define our exceedances as

$$y_{ij} = (x_{ij} - u) \cdot \mathbb{1}_{(j \in A_i^c)}$$

so that all observations not exceeding u are marked as 0. Let $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n_i})^\top$ and $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_R^\top)^\top$.

The likelihood is given by

$$L(\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\zeta}) = \prod_{i=1}^{R} f_{Y_i}(\mathbf{y}_i | \sigma_i, \xi_i, \zeta_i)$$

$$= \prod_{i=1}^{R} \left[\prod_{j \in A_i} F_{X_i}(u) \times \prod_{j \in A_i^c} f_{X_i}(y_{ij} + u) \right]$$

$$\approx \prod_{i=1}^{R} \left[\prod_{j \in A_i} F_{X_i}(u) \times \prod_{j \in A_i^c} [1 - F_{X_i}(u)] h(y_{ij} | \sigma_i, \xi_i) \right] \qquad \text{(approximation (5))}$$

$$= \prod_{i=1}^{R} \left[\prod_{j \in A_i} (1 - \zeta_i) \times \prod_{j \in A_i^c} \frac{\zeta_i}{\sigma_i} \left(1 + \xi_i \frac{y_{ij}}{\sigma_i} \right)_+^{-1/\xi_i - 1} \right] \qquad (\zeta_i = 1 - F_{X_i}(u))$$

$$= \prod_{i=1}^{R} \left[(1 - \zeta_i)^{n_i - k_i} \zeta_i^{k_i} \prod_{j \in A_i^c} \frac{1}{\sigma_i} \left(1 + \xi_i \frac{y_{ij}}{\sigma_i} \right)_+^{-1/\xi_i - 1} \right] \qquad (7)$$

Note that the parameters describing the tail of F_i (i.e. ξ_i, σ_i) depend only on those observations which exceed u. The parameter $\zeta_i = P(X_{ij} > u)$, which is necessary for calculating return levels (section 6), is based only on the number of exceedances. This justifies the use of cluster maxima for $\mathbf{y}_i$.

We complete the hierarchical model formulation by specifying the following priors:

$$\xi_i | \xi, \tau^2 \sim Normal(\xi, \tau^2)$$
$$\sigma_i | \alpha, \beta \sim Gamma(\alpha, \beta)$$
$$\zeta_i | \zeta, \eta \sim Beta(\zeta\eta, (1 - \zeta)\eta)$$

$$(8)$$

$$\xi \sim Normal(m, s^2) \qquad\qquad \tau^2 \sim Gamma(a_\tau, b_\tau)$$
$$\alpha \sim Gamma(a_\alpha, b_\alpha) \qquad\qquad \beta \sim Gamma(a_\beta, b_\beta)$$
$$\zeta \sim Beta(a_\zeta, b_\zeta) \qquad\qquad \eta \sim Gamma(a_\eta, b_\eta)$$

By combining (7) and (8) we obtain the full posterior distribution. Samples are obtained via MCMC.

6 Return levels

A most useful quantity in an extreme value analysis is the return level. Generally, for a distribution G, the return level x_m is the solution to

$$G(x_m) = 1 - \frac{1}{m} \qquad (9)$$

and has the convenient interpretation as the quantity that is exceeded on average once every m observations.

When working with the generalized Pareto model (5), it can be shown that the m-observation return level is

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta\theta)^\xi - 1\right] \tag{10}$$

where the terms ζ and θ account for the probability of exceeding u and being within a cluster, respectively. Using MCMC samples, we can obtain a distribution for x_m, but note that the posterior mean for θ is used, not samples obtained using (4), when calculating the return level.

7   Bhattacharyya distance

Bhattacharyya (1943) proposed a means for measuring the degree of similarity between two probability distributions. For two continuous random variables on support $\mathcal{X}$ with densities p and q, the Bhattacharyya coefficient is defined as

$$BC(p, q) = \int_\mathcal{X} \sqrt{p(x)q(x)}dx. \tag{11}$$

The Bhattacharyya distance is

$$D_B(p, q) = -\log BC(p, q). \tag{12}$$

We use kernel density estimation to calculate p and q along a grid of the support and then approximate the integral in (11). If the support is different for the two random variables (as will typically be the case when comparing random variables whose parameters determine the support such as the generalized Pareto), we will integrate over the intersection of the supports.

8   Results

For each of the four data sources (i.e. the three climate simulation classes and the observation produce), there are four factors with two levels each. The factors, with their levels, are:

1. Variable — precipitation or maximum temperature

2. Season — winter or summer

3. Decade — 1962–1971 or 1990–1999

4. Region — California or U.S.A.

There are then 16 combinations of the factors to be made. For each combination, the hierarchial model described in section 5.2 is fit to the decadal, historical, and control runs; the univariate model in section 5.1 is fit to the observation product since this data source does not have replicates.

The next six figures shows posterior parameters and other quantities of interest. For the hierarchical model, we show the results of the *mean* process. For example, in Figure 3 the parameter shown is the posterior for ξ, the mean of $\xi_1, \ldots, \xi_R$. This is in opposition to inference on an unknown replicate which would require sampling, among other things, a new shape parameter ξ^*. Therefore, the intervals are more narrow than if we looked at the posterior predictive distribution for a new replicate, but the parameters will be comparable to those from the univariate model with the observations and give us a sense of how the climate simulation performs on average.

The posterior shape parameters in Figure 3 show overlapping bounds in many cases, but in some combinations of factors we can seem some departure from the observations. The numbers shown above the lines are the thresholds: 0.95 quantile for the climate simulations and 0.85 for the observations. The quantity itself is not too important since different thresholds may produce similar return levels. Figure 4 shows the logarithm of the mean scale parameter, $\log(\sigma)$ for the observations and $\log(\alpha/\beta)$ for the simulations.

Figures 5 and 6 give the 20- and 50-year return levels, respectively. In these figures, we have the same x-axes for the columns in total precipitation and for the columns under average maximum temperature. Some intervals are difficult to see given the scale, but we can still inspect how the return levels from the observations differ from those of the climate siulations. For winter precipitation, we see that the climate simulations underestimate the return level in california, but does much better when over the whole of the United States.

Figure 7 shows posterior predictive samples (based on the mean parameters for the hierarchical model) drawn from the generalized Pareto distribution, conditioned on the random variables being greater than the given threshold. There is significant overlap of the simulations with the observations, but this alone does not suggest accuracy of the simulations. Thus, the aforementinoed return levels are necessary when determining similarities and differences.

The posterior mean and 95% highest posterior density intervals for the extremal index are shown in Figure 8. The climate simulations seem to be consistent with the observations for the temperature, but not so for precipiation where they tend to overestimate θ. Though, in terms of limiting mean cluster size, there is not a great difference between $1/0.5 = 2$ and $1/0.8 = 1.25$.

Recall that these results are based on the anomalies, the difference between the daily values and their average. Predictions based on the hierarchical and univariate models proposed in section 5 can only be of practical use if the average is also given. This may be simple to calculate for the climate simulations, but more care would be needed for projecting the average of the observations. We have not considered this in our analysis.

References

Bhattacharyya, A. (1943), "On a measure of divergence between two statistical populations defined by their probability distribution," *Bull. Calcutta Math. Soc.*

Ferro, C. A. and Segers, J. (2003), "Inference for clusters of extreme values," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 545–556.
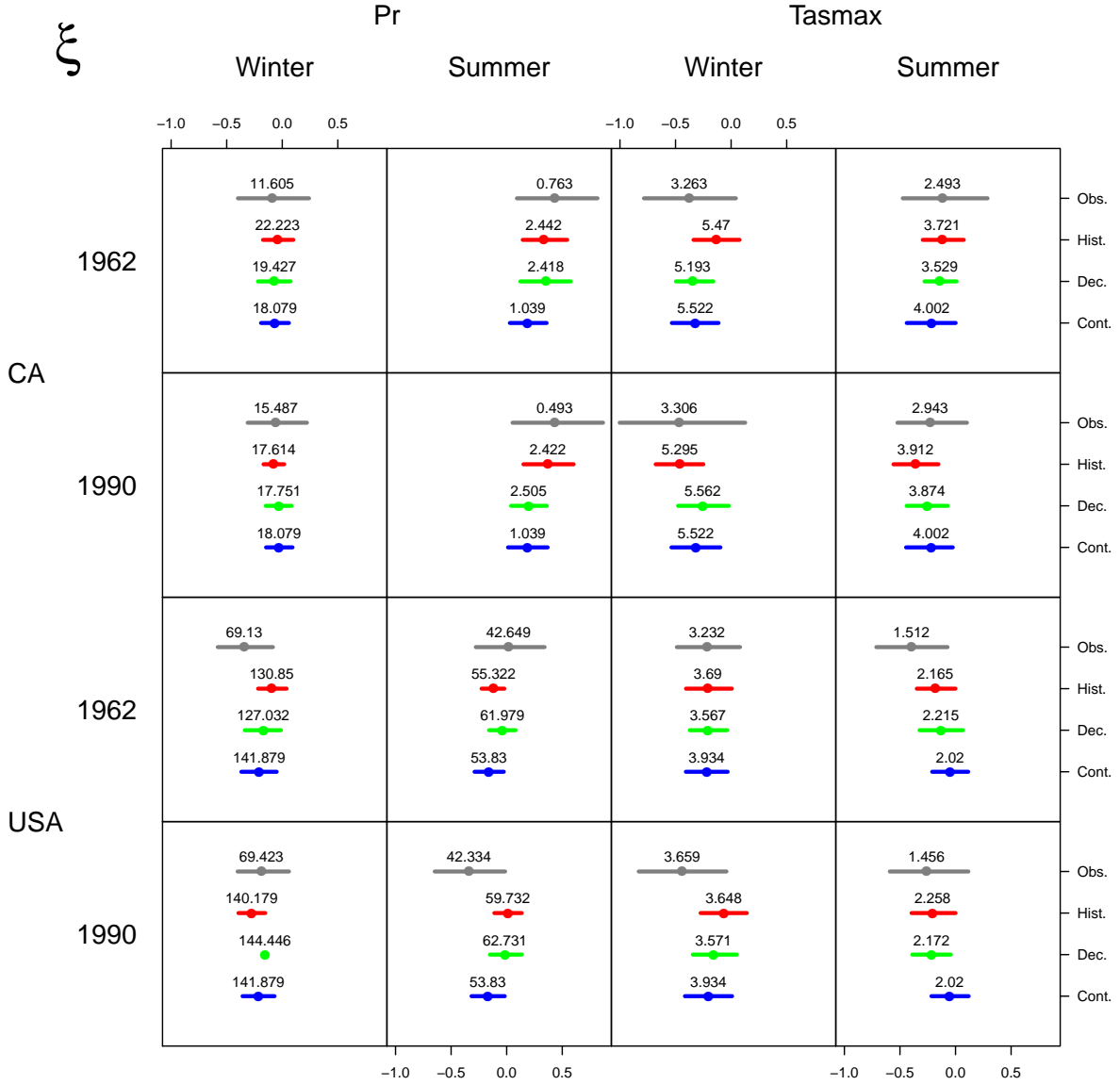
Figure 3: Posterior shape parameter, ξ, under each domain and each of the four data types. The points are the means and the lines mark the 95% h.p.d. intervals. The value above each point is the threshold used in the analysis. Note: The x-axes are the same for every plot. The y-axes (for this and all subsequent figures) denote only the data type and thus hold no quantitative meaning.
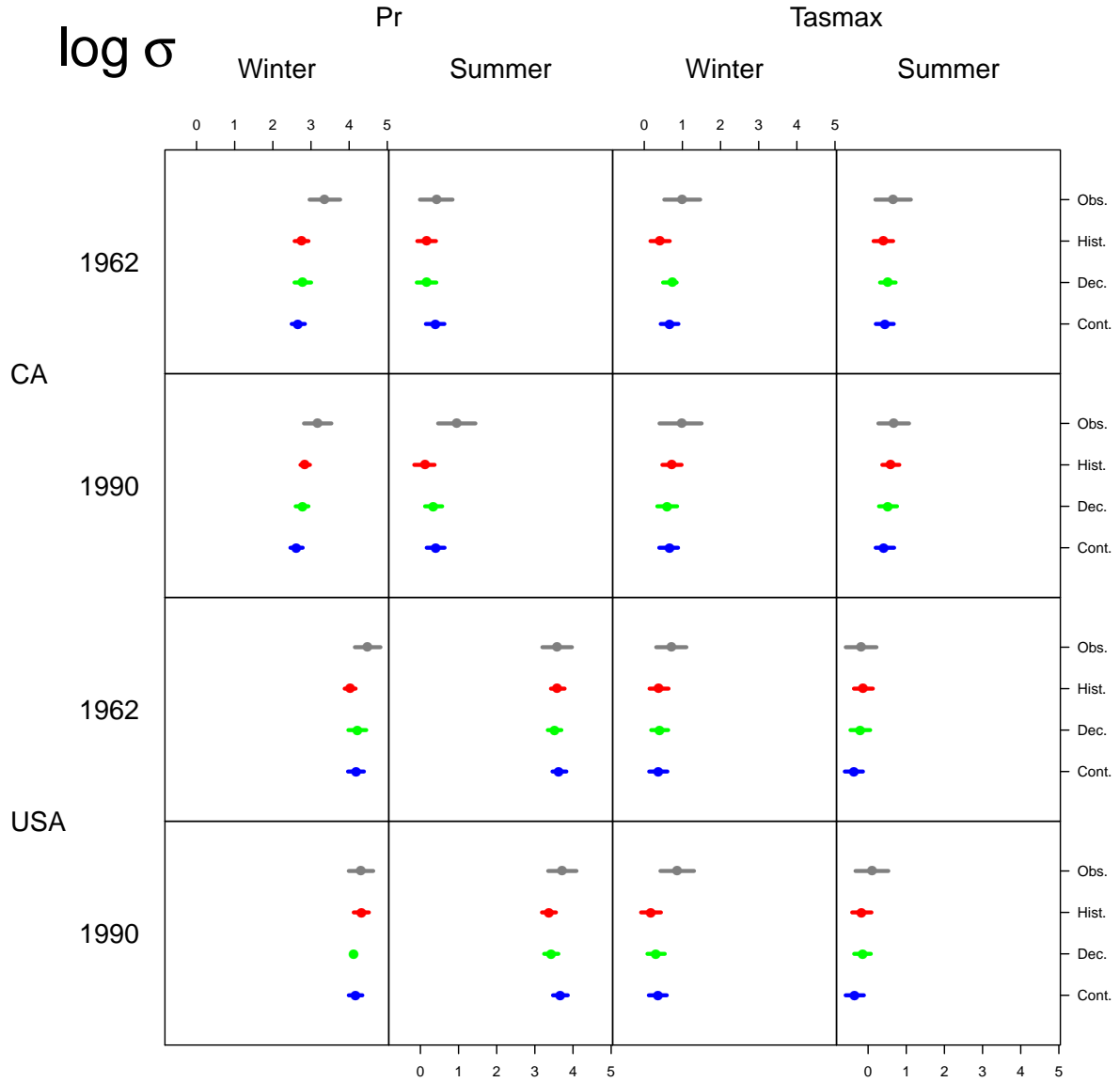
Figure 4: Natural logarithm of the posterior scale. For the CanCM4 simulations, the parameter shown is $\log(\alpha/\beta)$ (the mean scale) because σ_i follows a Gamma distribution with mean α/β. No change of variables is necessary for the observations. Note: The x-axes are the same for every plot.
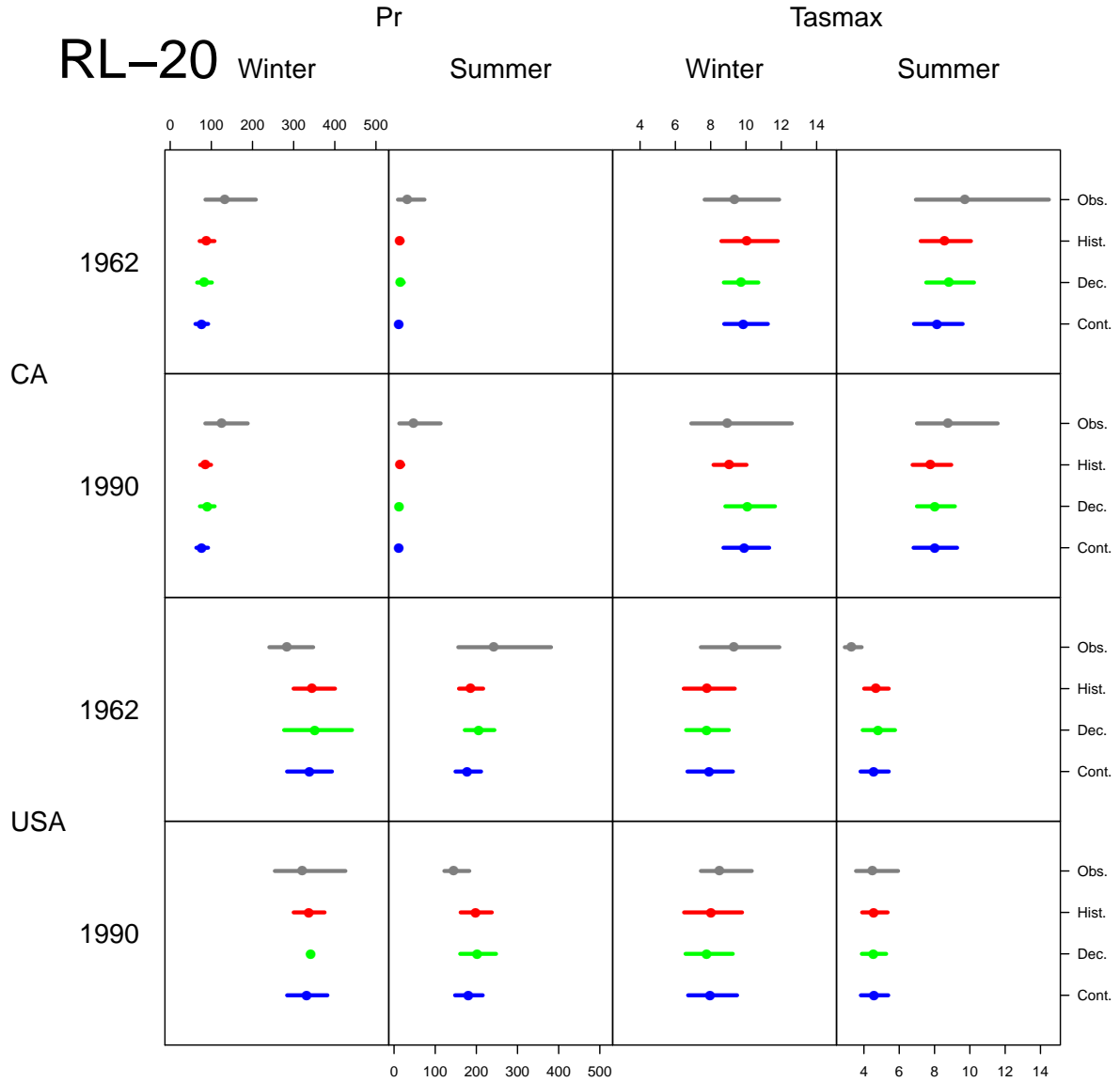
Figure 5: 20-year return levels. Note: The left two columns have the same *x*-axes, which are different than those in the right two columns, which have the same.
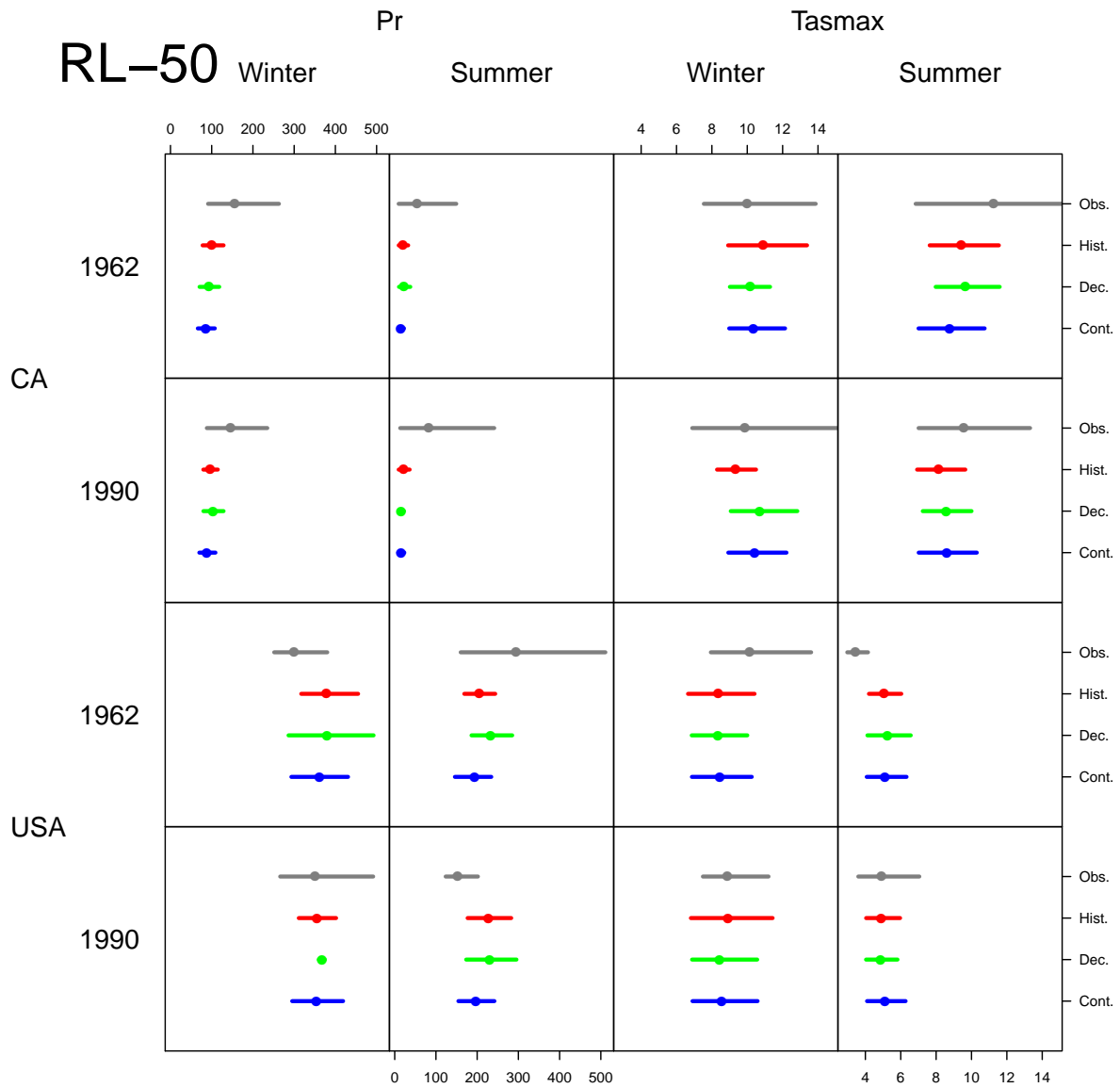
Figure 6: 50-year return levels. The x-axes are the same as those in Figure 5.
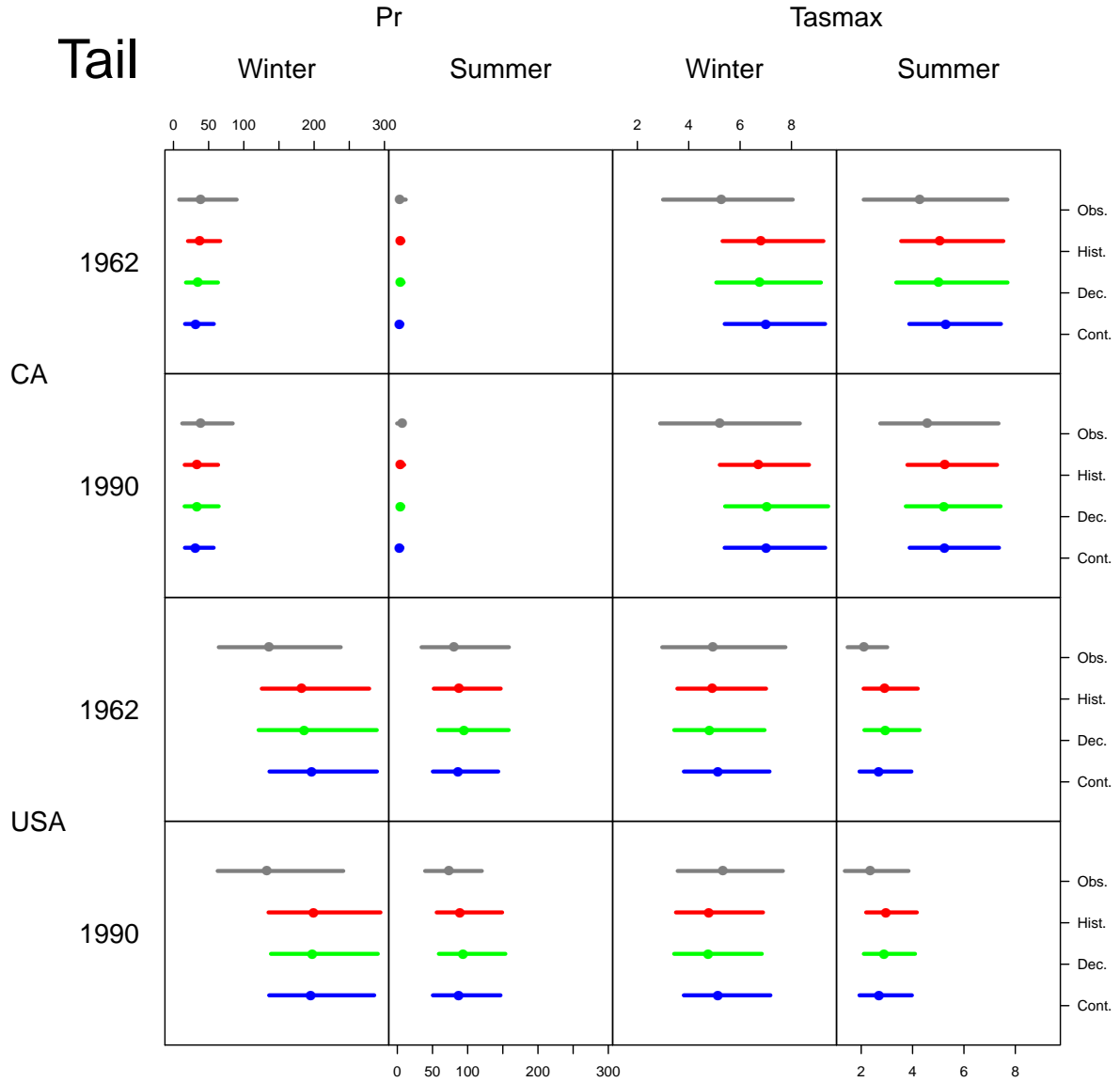
Figure 7: Mean and 95% h.p.d. for the upper tail (i.e. the generalized Pareto) of the ensemble average. Similar to Figures 5 and 6, the left two columns have the same x-axes and the right two columns have the same x-axes.
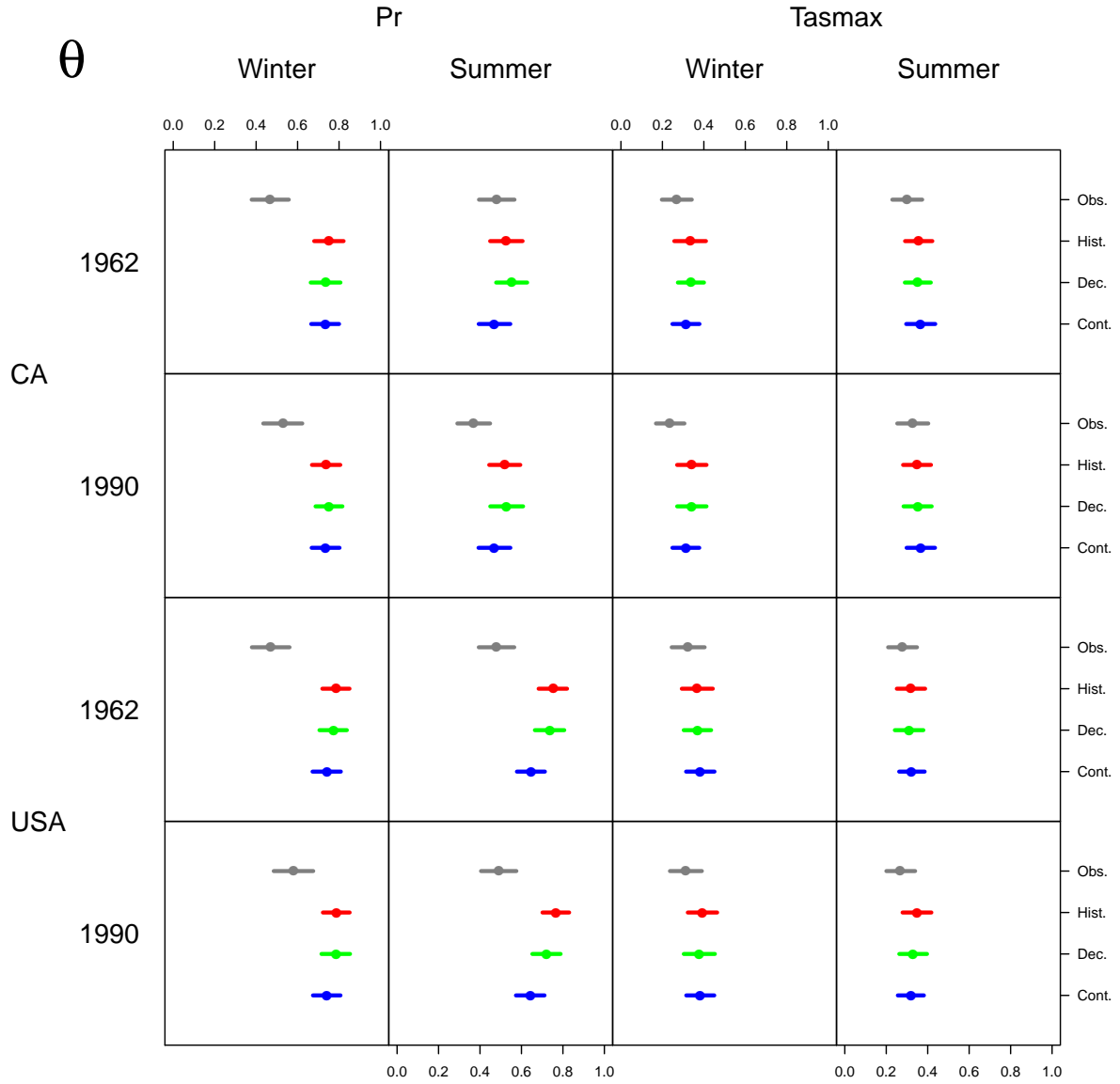
Figure 8: The mean extremal index. Like the parameters shown in Figures 3 and 4, the hierarchical mean is shown for the CanCM4 simulations.

Maurer, E., Wood, A., Adam, J., Lettenmaier, D., and Nijssen, B. (2002), "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States," *Journal of climate*, 15, 3237–3251.

Prado, R. and West, M. (2010), *Time series: modeling, computation, and inference*, CRC Press.