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 obtain from Ed Mauer's website. 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.

Having replicates of a time-series suggests the use of a hierarchical model, described in detail in section 4.2. 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.

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 5.

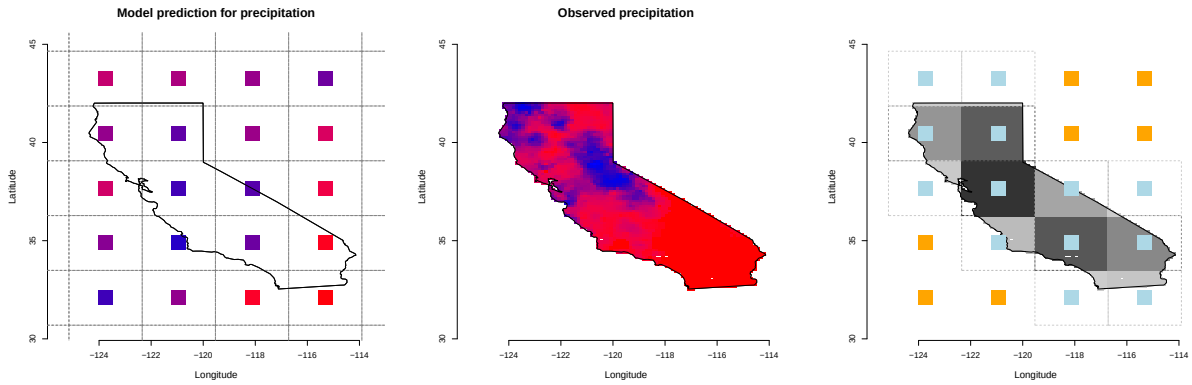

Figure 1: Left: CanCM4 simulation locations. Center: Observation locations. Right: method for computing weighted sum or average for CanCM4 to make values comparable with observations; the lighter gray points mean less weight is applied to the climate simulations and the darker gray means more weight. 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. For each time-series, we fit a dynamic linear model (DLM) with annual and semi-annual periods. From the DLMs we obtain a smoothed version of the time-series, and then take the differ-

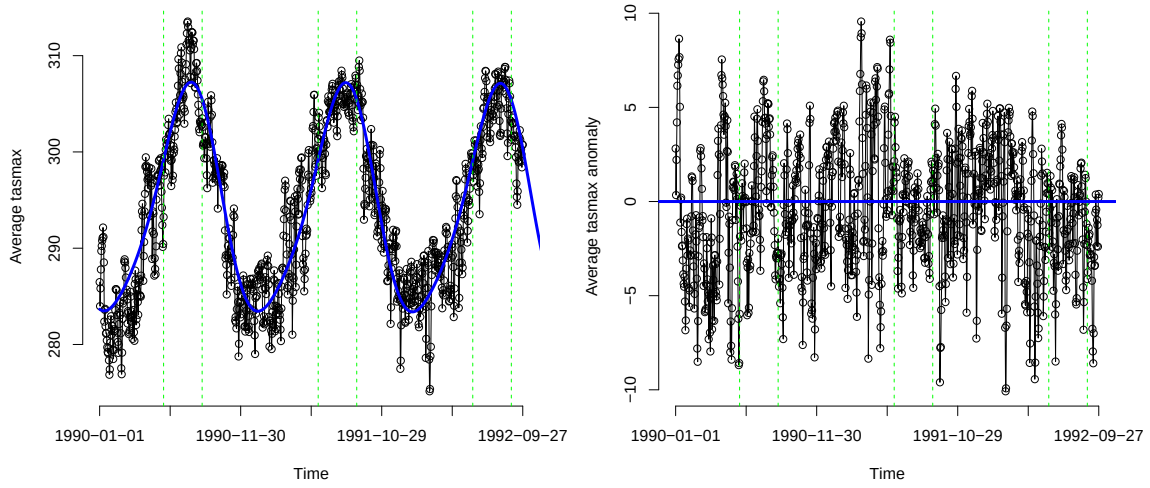

Figure 2: One of the DLMs used to calculate the anomalies. Shown is one of the decadal replicates of average `tasmax` in California for about the first two and one-half years of the time-series. The green dashed lines mark the beginning and the end of the summer months.

ence between the original series with the smoothed version. The differences are called the anomalies and the subsequent analyses are performed on them.

Each analysis is confined to a specific season, either winter or summer. Winter months are defined as December, January, and February, and summer months are June, July, and August. After de-trending based on the whole time-series, we remove all observations that do not belong to the season of interest. The remaining observations are concatenated so that, for example in winter, 28 February is followed immediately by 1 December.

An example of the method is shown in Figure 2 for the first two and one-half years of one of the decadal replicates. The end result is a roughly stationary sequence, which we assume will be valid for an extreme value analysis.

4 Threshold exceedance model

4.1 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

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

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 (1) as

$$L(y_1, \ldots, y_k; \sigma, \xi) = \sigma^{-k} \sum_{i=1}^{k} \left(1 + \frac{\xi y_i}{\sigma}\right)_+^{-1/\xi - 1} \tag{2}$$

where $z_+ = \max(z, 0)$. This provides the basis for an extreme value analysis.

In many cases, the assumption of independence in observations may be too strong. When we have dependent random variables, which is likely the case in a time series, we employ a declustering scheme to obtain independent clusters, discussed in section 5

4.2 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 (2), 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

$$\begin{aligned}
L(\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\zeta}) &= \prod_{i=1}^{R} f_{Y_i}(\mathbf{y}_i | \sigma_i, \xi_i, \zeta_i) \\
&= \prod_{i=1}^{R} \left[\prod_{j \in A_i} F_{X_i}(u) \times \prod_{j \in A_i^c} f_{X_i}(y_{ij} + u) \right] \\
&\approx \prod_{i=1}^{R} \left[\prod_{j \in A_i} F_{X_i}(u) \times \prod_{j \in A_i^c} [1 - F_{X_i}(u)] h(y_{ij} | \sigma_i, \xi_i) \right] \qquad \text{(approximation (1))} \\
&= \prod_{i=1}^{R} \left[\prod_{j \in A_i} (1 - \zeta_i) \times \prod_{j \in A_i^c} \frac{\zeta_i}{\sigma_i} \left(1 + \xi_i \frac{y_{ij}}{\sigma_i}\right)_+^{-1/\xi_i - 1} \right] \qquad (\zeta_\mathrm{i} = 1 - \mathrm{F}_{\mathrm{X_i}}(\mathrm{u})) \\
&= \prod_{i=1}^{R} \left[(1 - \zeta_i)^{n_i - k_i} \zeta_i^{k_i} \prod_{j \in A_i^c} \frac{1}{\sigma_i} \left(1 + \xi_i \frac{y_{ij}}{\sigma_i}\right)_+^{-1/\xi_i - 1} \right]
\end{aligned} \tag{3}$$

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.

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

$$\begin{aligned} \xi_i|\xi,\tau^2 &\sim Normal(\xi,\tau^2) \\ \sigma_i|\alpha,\beta &\sim Gamma(\alpha,\beta) \\ \zeta_i|\zeta,\eta &\sim Beta(\zeta\eta,(1-\zeta)\eta) \end{aligned} \tag{4}$$

$$\begin{aligned} \xi &\sim Normal(m,s^2) & \tau^2 &\sim Gamma(a_\tau,b_\tau) \\ \alpha &\sim Gamma(a_\alpha,b_\alpha) & \beta &\sim Gamma(a_\beta,b_\beta) \\ \zeta &\sim Beta(a_\zeta,b_\zeta) & \eta &\sim Gamma(a_\eta,b_\eta) \end{aligned}$$

By combining (3) and (4) we obtain the full posterior distribution. Samples are obtained via MCMC, after a burn-in of $200,000$ iterations, $100,000$ samples are thereafter obtained.

5 Extremal Index

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

5.1 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 $1 \le j_1 < \cdots < j_N \le n$. 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

$$\begin{aligned} l(\theta,p;\mathbf{T}) =& m_1\log(1-\theta p^\theta) + (N-1-m_1)\{\log(\theta)+\log(1-p^\theta)\} \\ &+ \theta\log(p)\sum_{i=1}^{N-1}(T_i-1) \end{aligned} \tag{5}$$

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 replications 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 (5), 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 (5). We assume

$$\begin{aligned}
\theta &\sim Beta(a_\theta, b_\theta) \\
\nu &\sim Gamma(a_\nu, b_\nu) \\
p &\sim Beta(a_p, b_p) \\
\tau &\sim Gamma(a_\tau, b_\tau)
\end{aligned}$$

with the hyperparameters chosen to be

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

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 (5)

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

$$L = \sum_{i=1}^{R} l(\theta_i, p_i; \mathbf{T}^{(i)}) \tag{6}$$

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

5.2 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 number of independent clusters.

6 Return levels

The m-observation return level is

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta\theta)^\xi - 1\right] \tag{7}$$

The posterior mean for θ is used, not samples, when obtaining the return levels.

7 Results

8 Discussion

References

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.

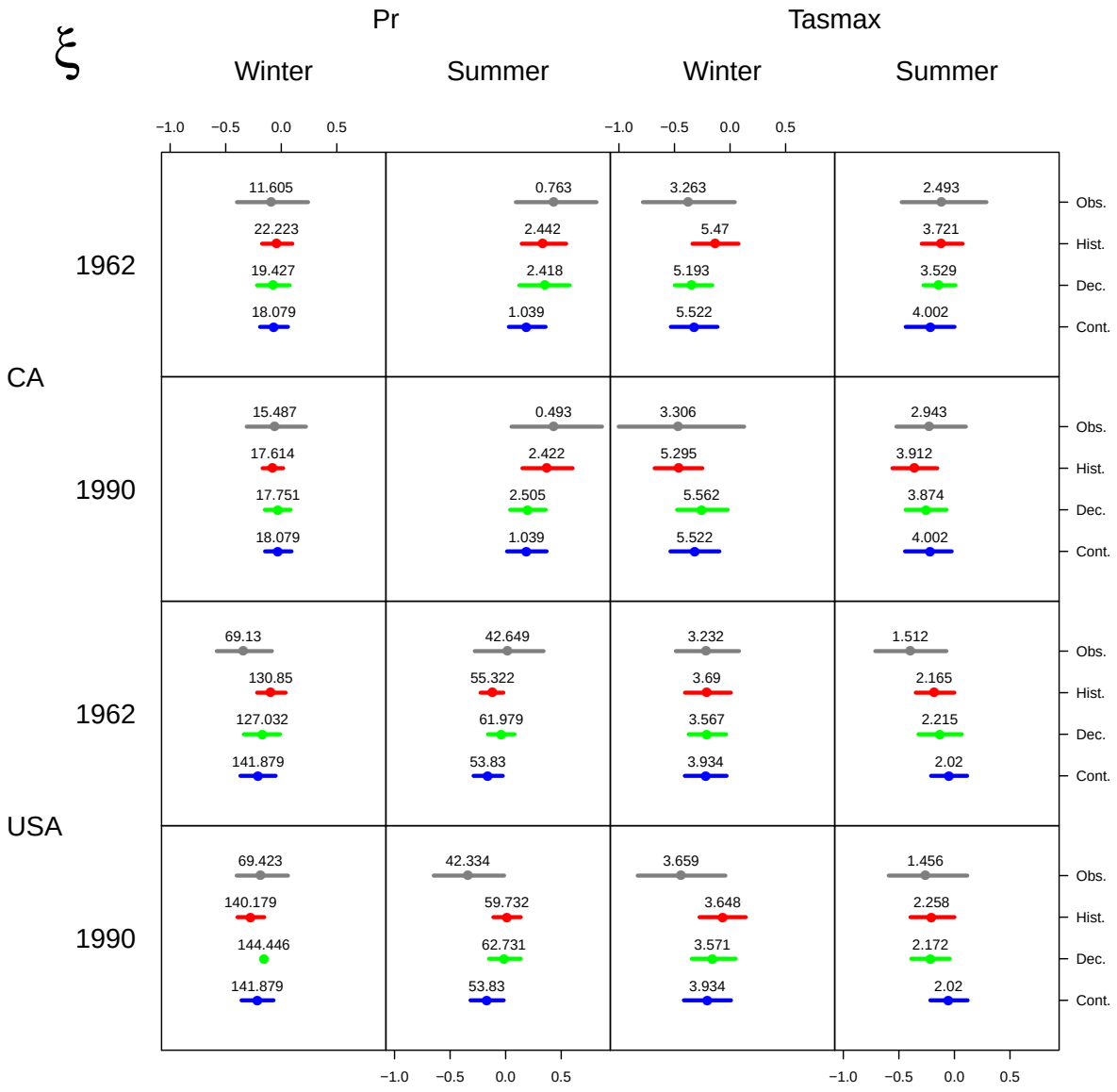

Figure 3: Posterior shape parameter, ξ, under each domain and each of the four data types. The points are the means and the lines mark the 95% h.p.d. intervals. The value above each point is the threshold used in the analysis. Note: The x-axes are the same for every plot. The y-axes (for this and all subsequent figures) denote only the data type and thus hold no quantitative meaning.

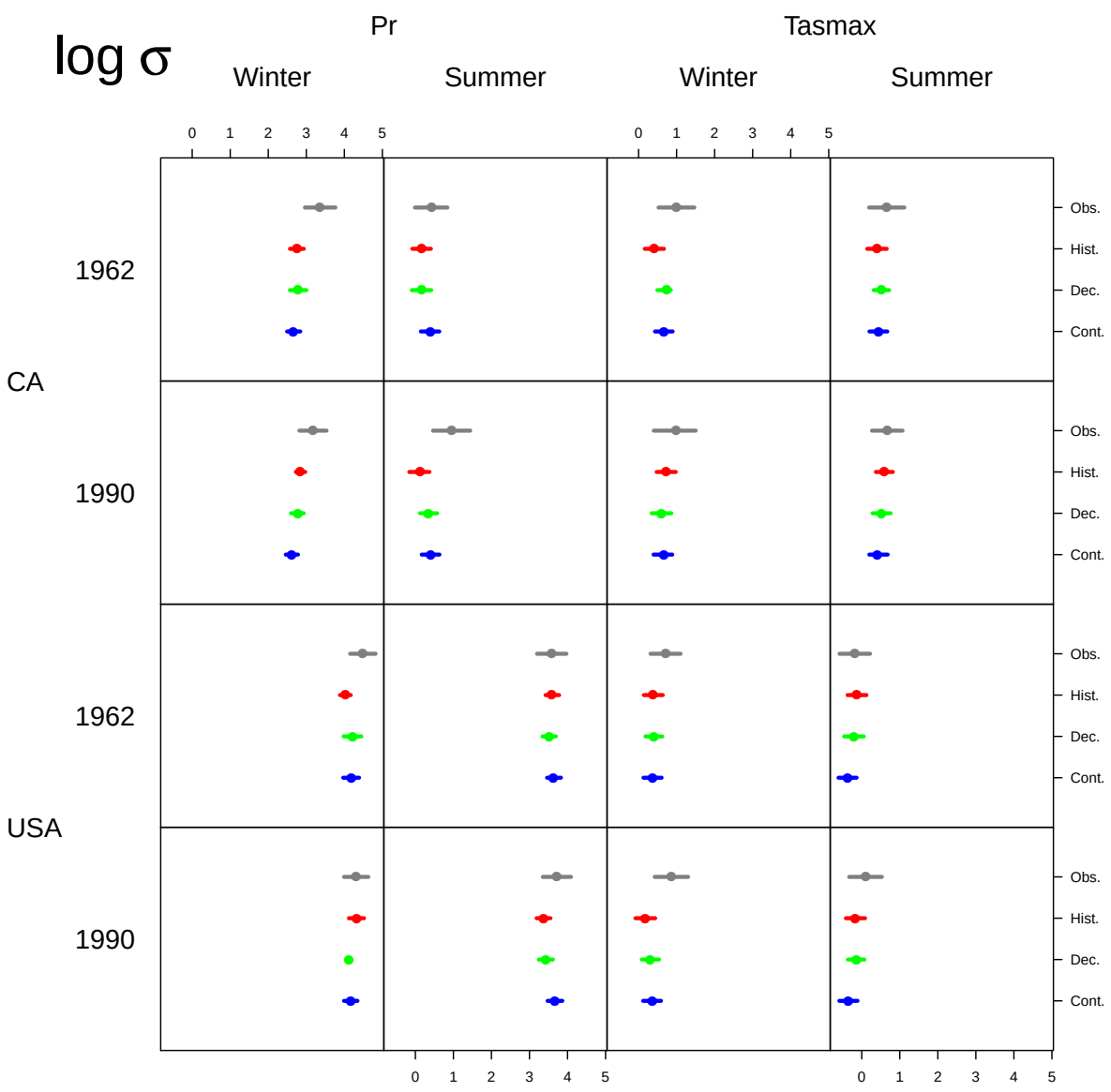

Figure 4: Natural logarithm of the posterior scale. For the CanCM4 simulations, the parameter shown is $\log(\alpha/\beta)$ (the mean scale) because σ_i follows a Gamma distribution with mean α/β. No change of variables is necessary for the observations. Note: The x-axes are the same for every plot.

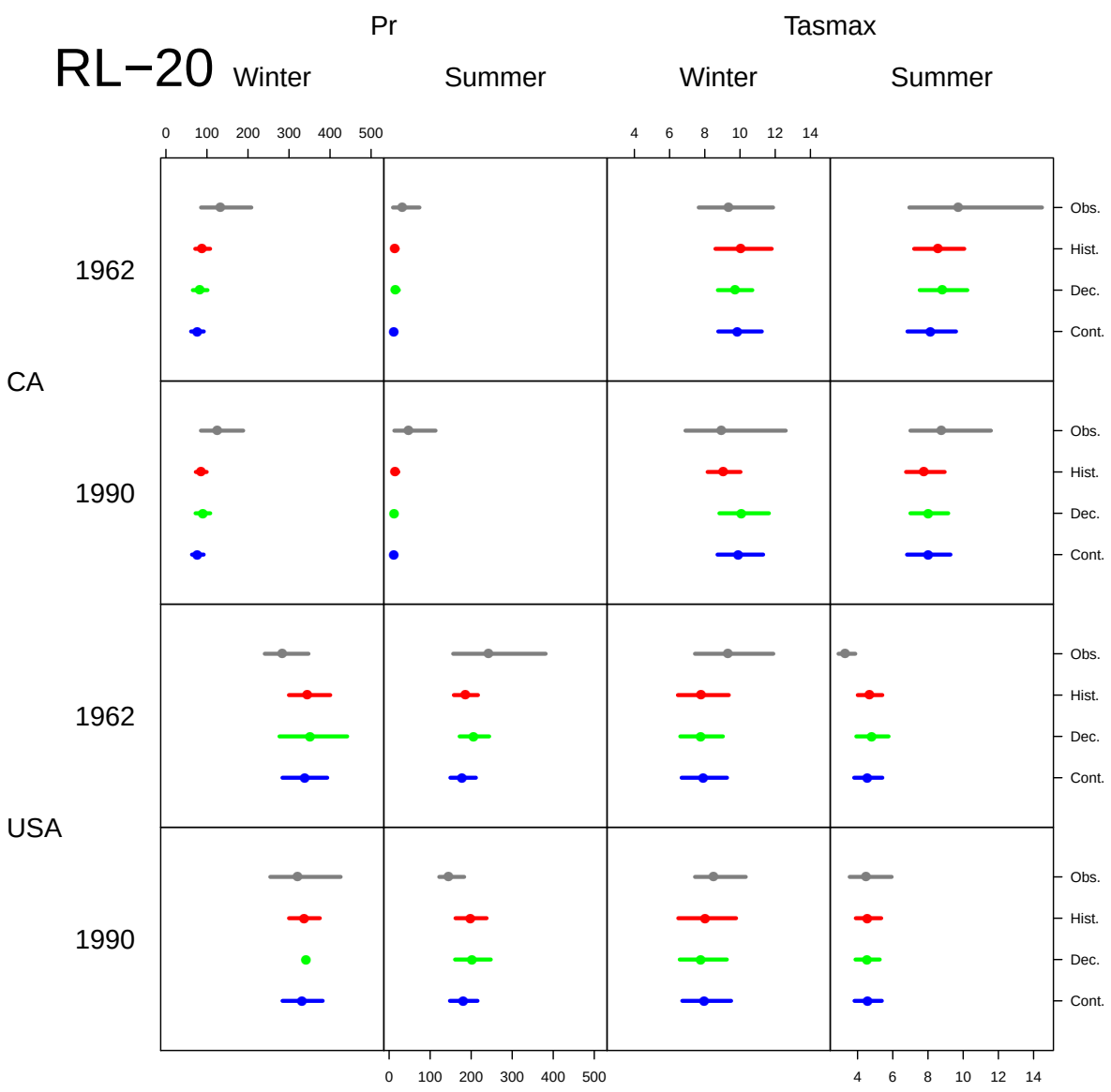

Figure 5: 20-year return levels. Note: The left two columns have the same x-axes, which are different than those in the right two columns, which have the same.

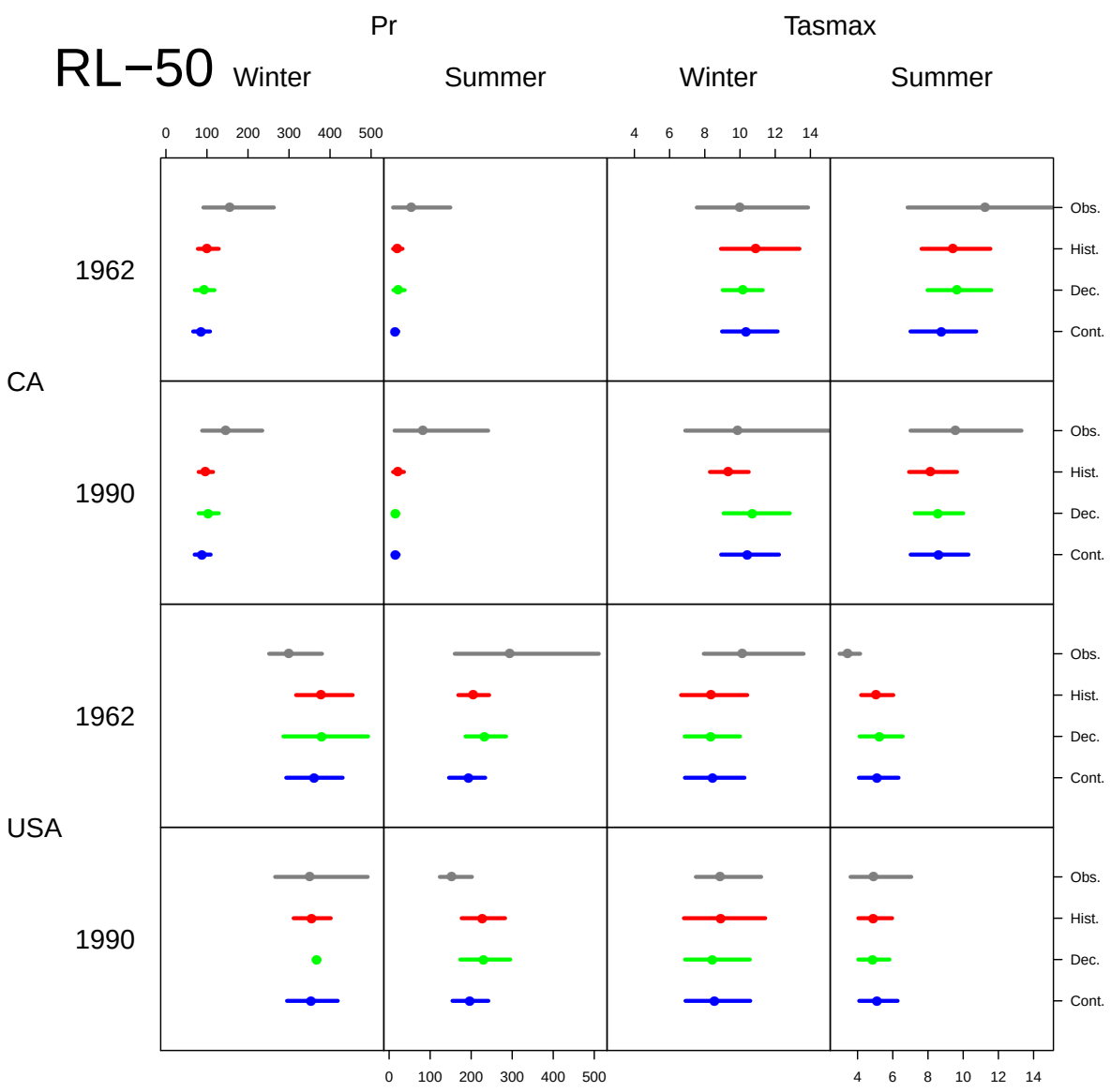

Figure 6: 50-year return levels. The x-axes are the same as those in Figure 5.

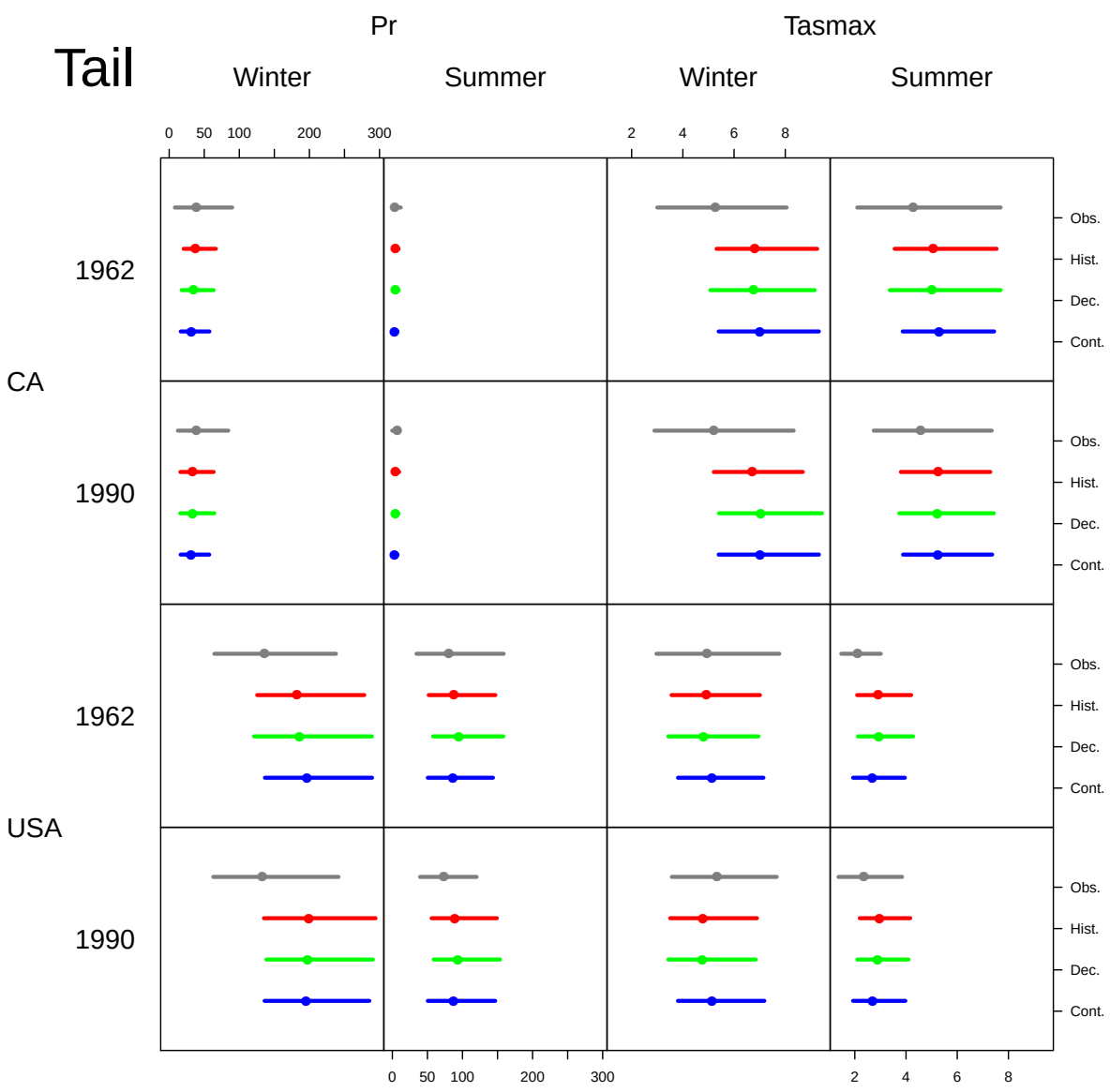

Figure 7: Mean and 95% h.p.d. for the upper tail (i.e. the generalized pareto) of the ensemble average. Similar to Figures 5 and 6, the left two columns have the same x-axes and the right two columns have the same x-axes.

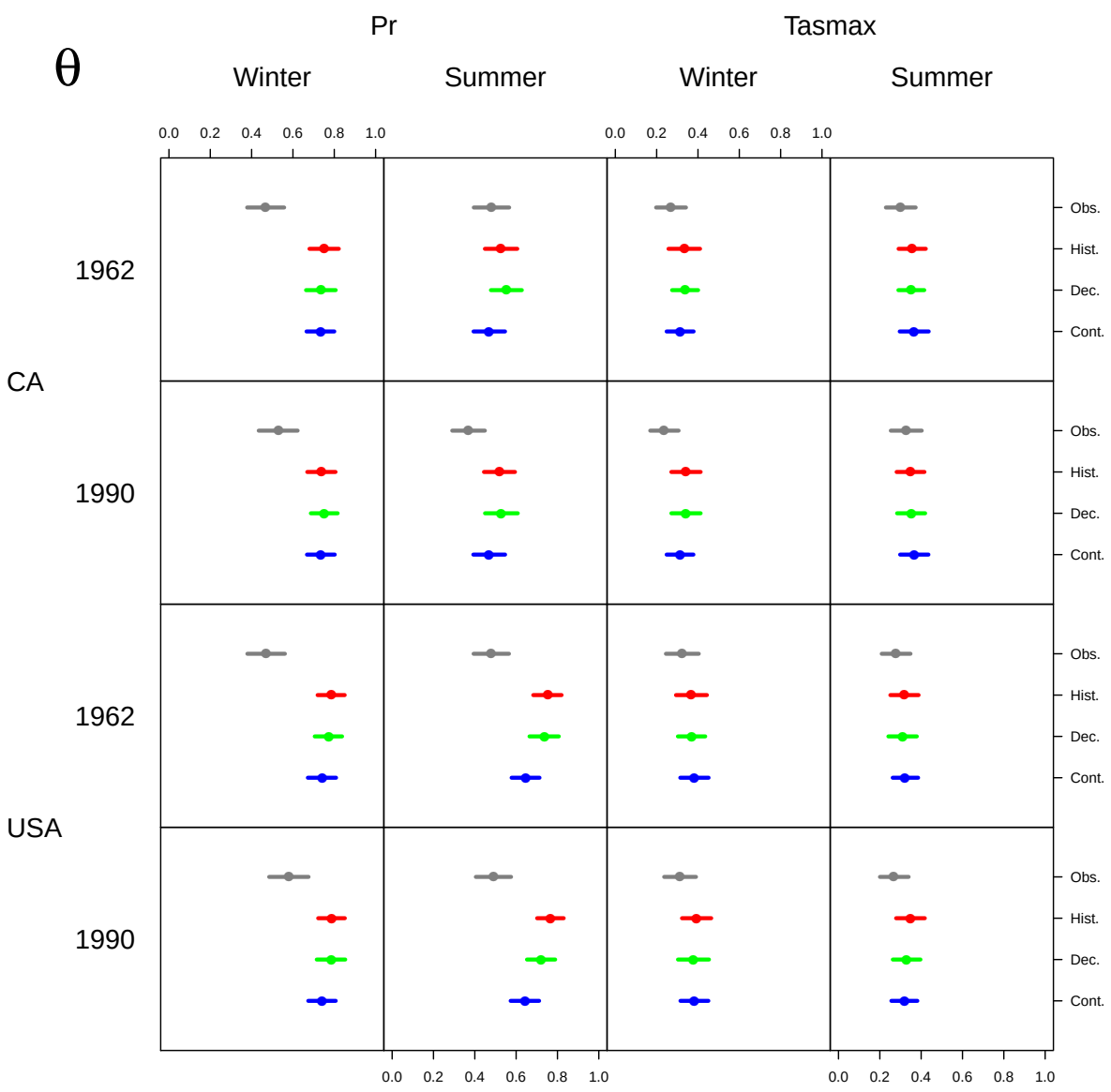

Figure 8: The mean extremal index. Like the parameters shown in Figures 3 and 4, the hierarchical mean is shown for the CanCM4 simulations.