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. 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 just 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.

Extreme value theory provides the framework for analyzing the stochastic behavior of a process at very large (small) values. This entails calculating the probability distribution of the maximum (minimum) of a sequence of random variabes.

Equivalently, extreme value analyses study the tails of probability distributions associated with some data generating mechanism.

In extreme value analyses, a primary interest is to understand

The standard approach is to appeal to asymtoptic arguments.

We can calculate useful quantities such as return levels.

This allows us to extrapolate beyond the span of historical data.

We compare three types of climate model simulations

The Fourth Generation Atmospheric General Circulation Model (CanCM4) from Canada

Decadal, historical, and control runs are used to obtain precipitation and temperature over California and the U.S. We have observational data from Ed Maurer. We will consider two 10-year periods: 1962–1971 and 1990–1999. We will also split these into winter months (December, January, Februrary) and summer months (June, July, August). Precipitation in summer is not analyzed.

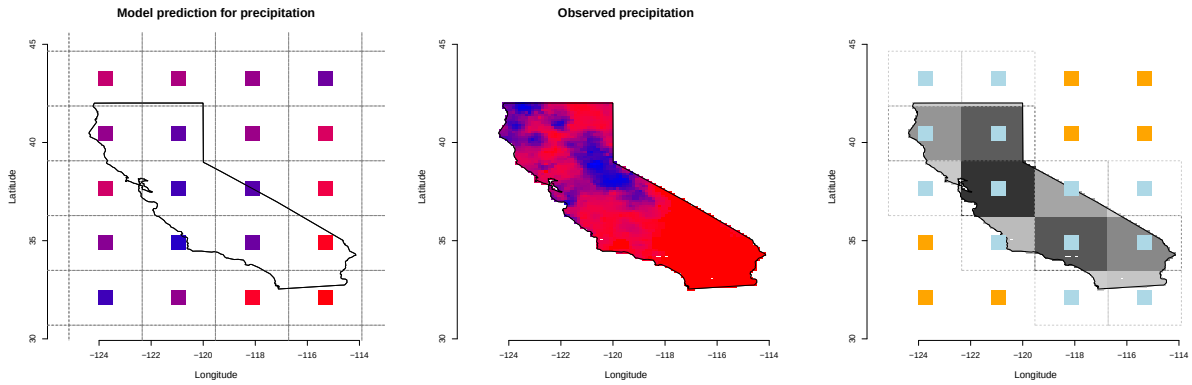

Figure 1: Left: CanCM4 simulation locations. Center: Observation locations. Right: method for computing weighted sum or average for CanCM4 to make values comparable with observations. The data shown are from a single day in January.

Purpose of the univariate analysis?

Details on the differences between decadal, historical, and control runs.

Precipitation based on the observations is summed. Precipitation from the climate model is computed with a weighted sum, based on the number of locations in the observation product.

Description of data processing

Picture of the locations CanCM4/Obs produces

Time-series plot of the variables

3 Data pre-processing

3.1 Aggregation

We will look at daily precipitation and maximum temperature, both over California during the 1990s. The quantities used in the analysis are total precipitation (a weighted sum) and average maximum temperature (weighted). (Expand). This leads to a R univariate time series for each class of simulations. To each time-series is fit a dynamic linear model (DLM) having the first two harmonics. Anomalies are computed by taking the difference of the time-series and the smoothed predictions based on the DLM. For winter we look at only December, January, and February. For summer we have June, July, and August. Finally, we treat the time-series as though there is no gap between the seasons of interest. For example, 28 February is followed immediately by 1 December in the winter analysis. This completes our pre-processing. (After the processing, the sequences are assumed stationary).

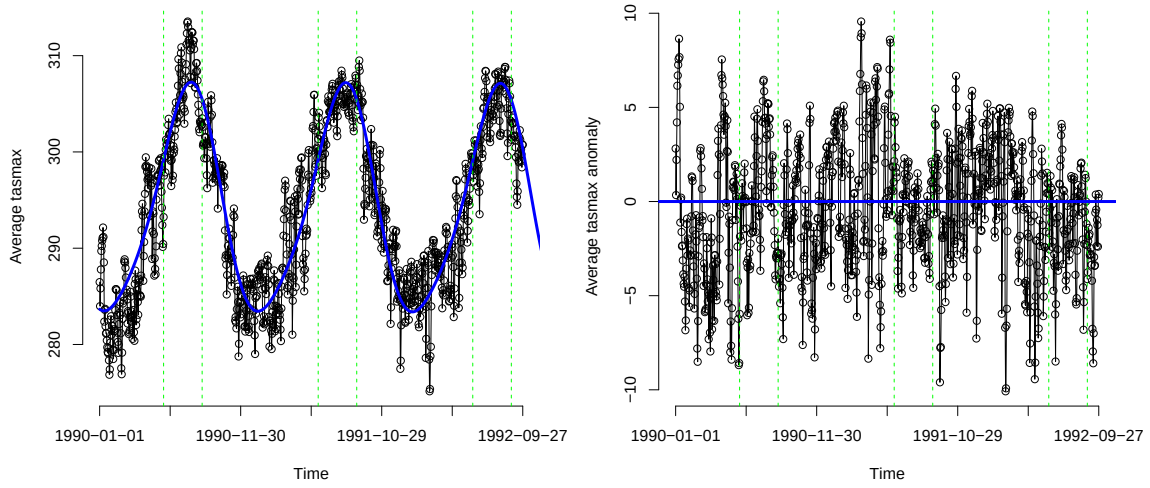

Figure 2: One of the DLMs used to calculate the anomalies. Shown is one of the decadal replicates of average tasmax in California for about the first two and one-half years of the time-series. The green dashed lines mark the beginning and the end of the summer months.

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 difference 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.

Include picture of smoothed DLM

4 Threshold exceedance model

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

4.1 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 $\zeta_i = P(X_i > u)$, 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:

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

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

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.

5.1 Estimation

Ferro and Suveges propose ways for estimating θ. It was our experience that the likelihood provided by Ferro worked better in estimating θ in a hierarchical setting than the likelihood proposed by Suveges.

Copy details from the other notes.

5.2 Declustering

6 Return levels

The m-observation return level is

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta\theta)^\xi - 1\right] \tag{5}$$

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

7 Results

8 Discussion

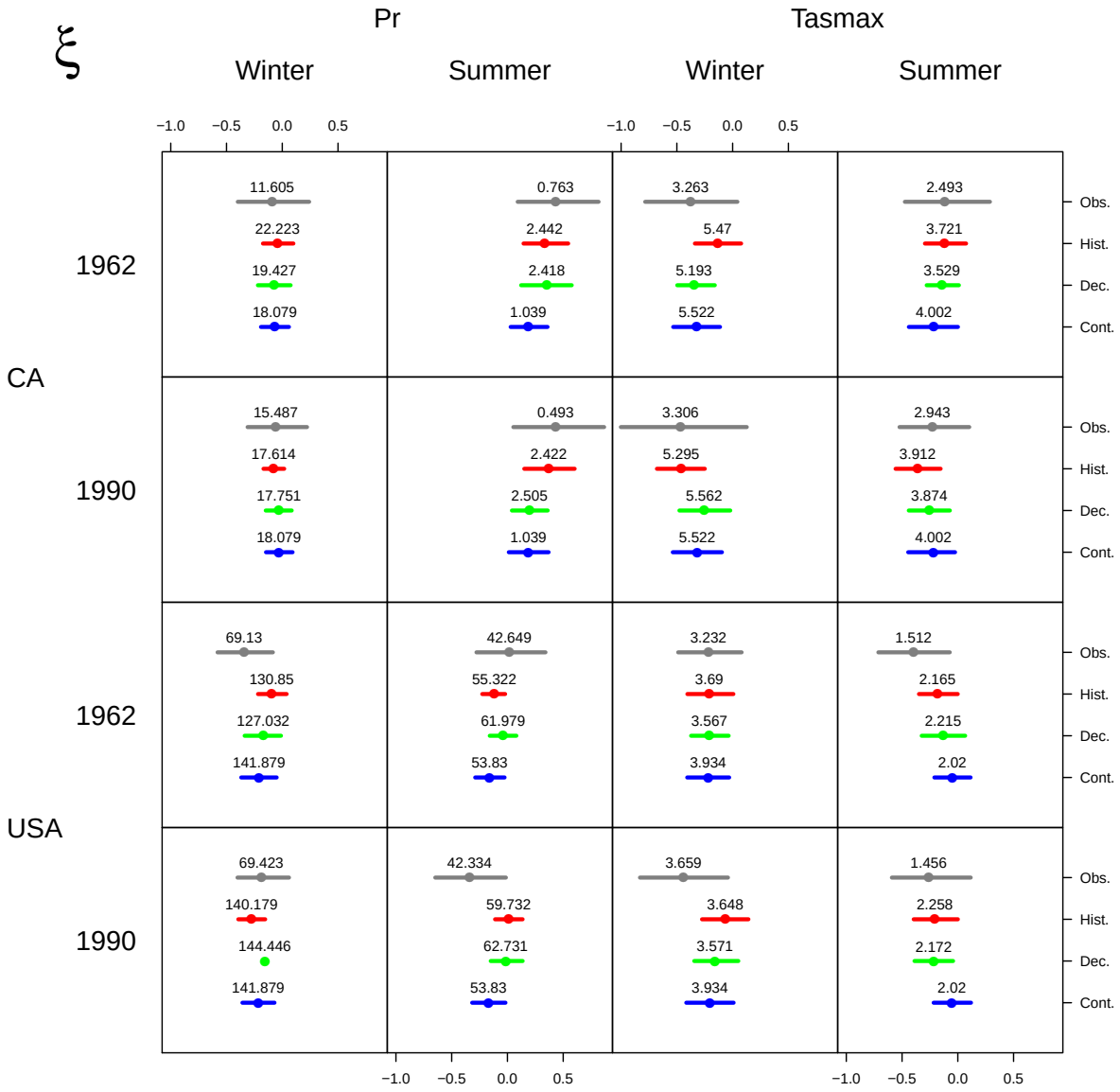

Figure 3: Posterior shape parameter, ξ, under each domain and each of the four data types. The points are the means and the lines mark the 95% h.p.d. intervals. The value above each point is the threshold used in the analysis. Note: The x-axes are the same for every plot. The y-axes (for this and all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

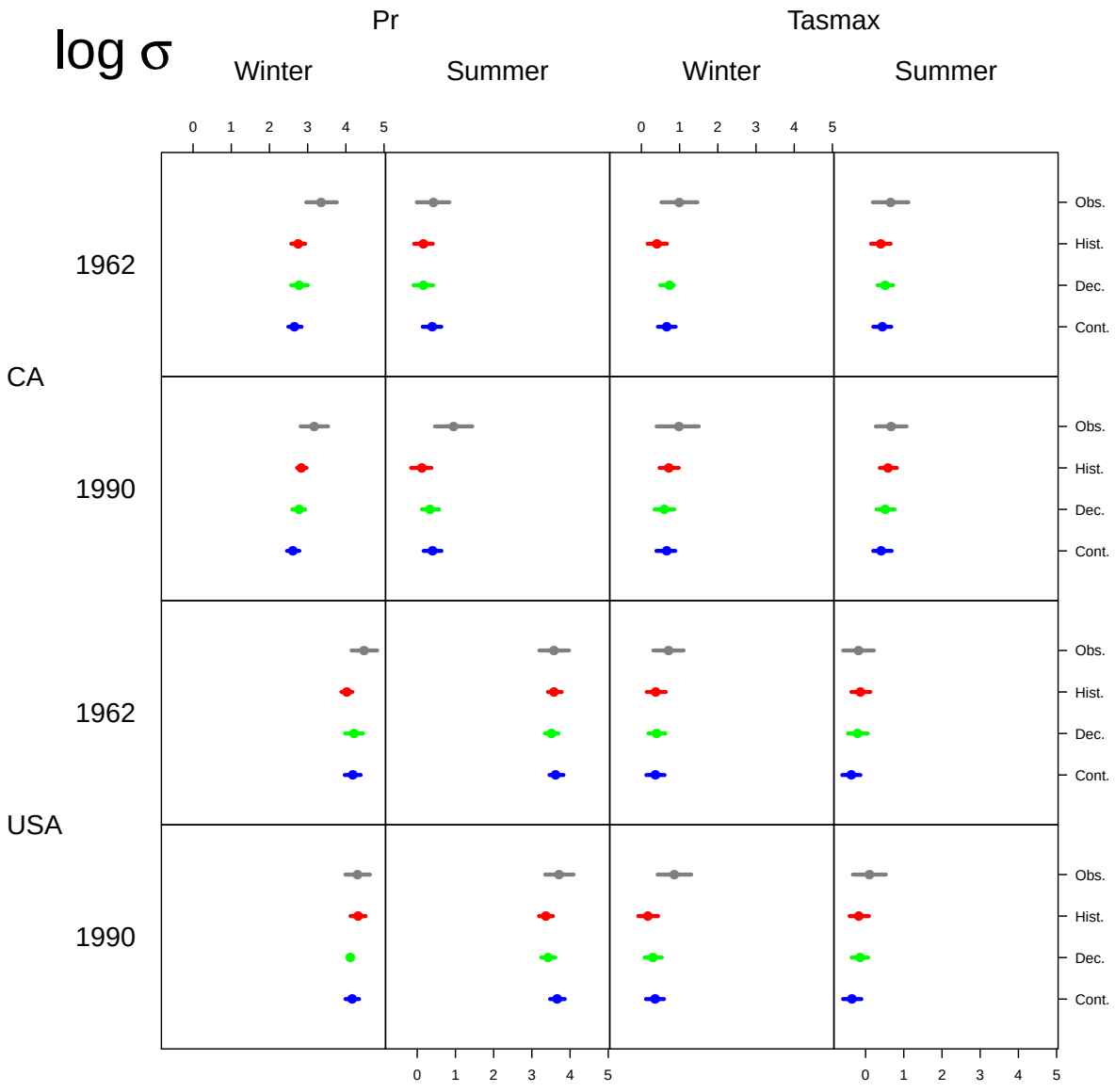

Figure 4: Natural logarithm of the posterior scale. For the CanCM4 simulations, the parameter shown is $\log(\alpha/\beta)$ (the mean scale) because σ_i follows a Gamma distribution with mean α/β. No change of variables is necessary for the observations. Note: The x-axes are the same for every plot.

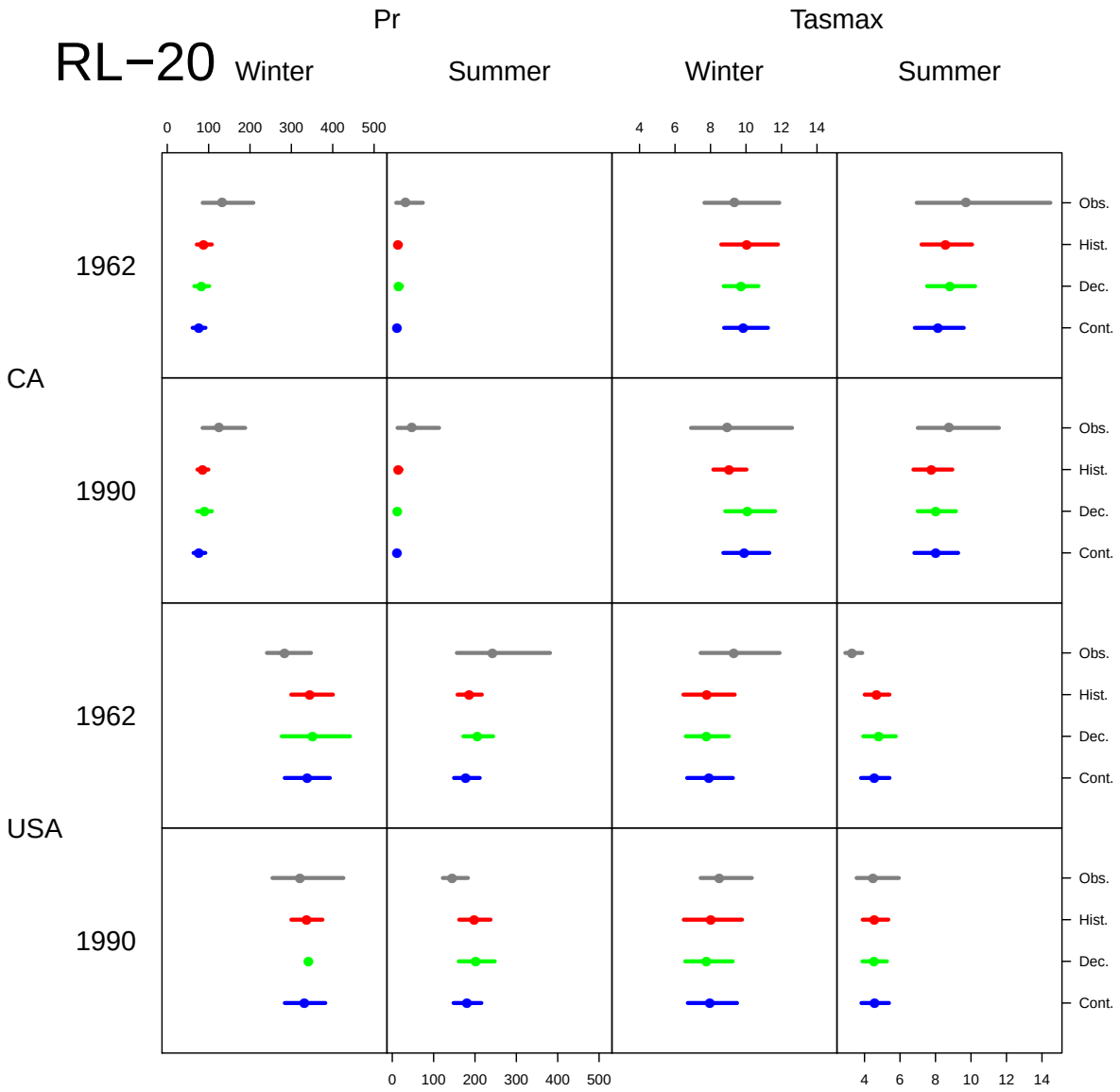

Figure 5: 20-year return levels. Note: The left two columns have the same x-axes, which are different than those in the right two columns, which have the same.

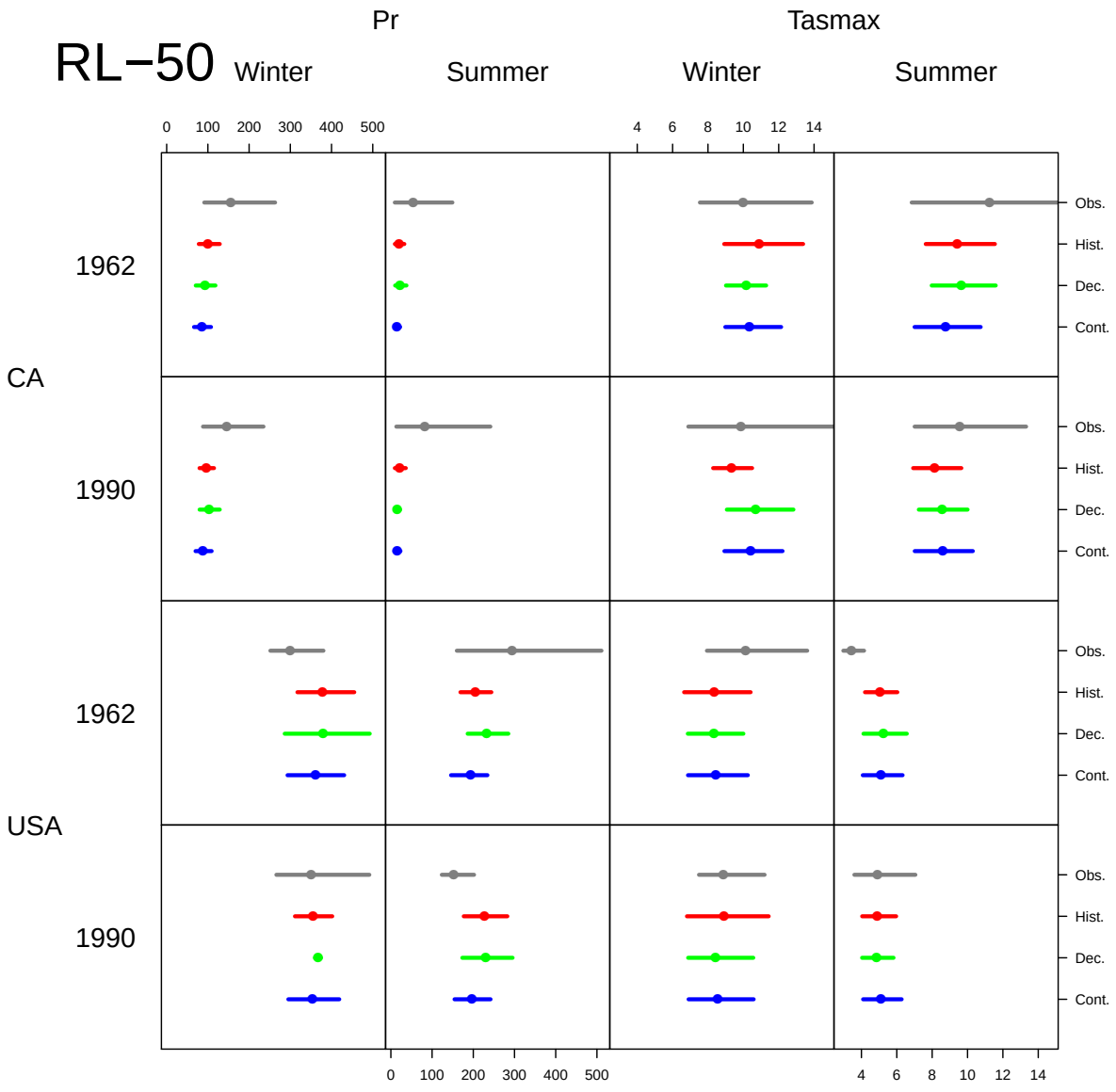

Figure 6: 50-year return levels. The x-axes are the same as those in Figure 5.

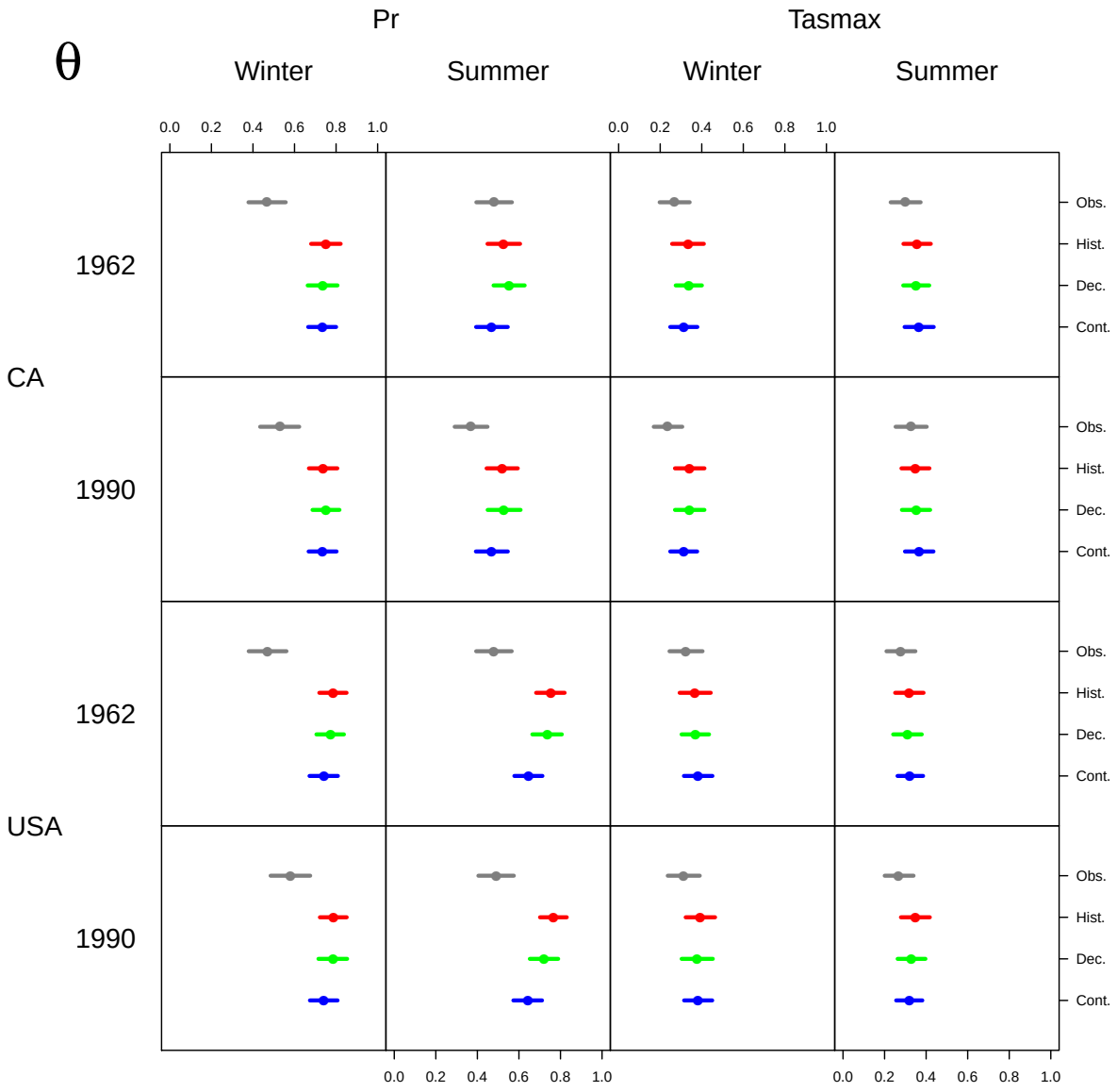

Figure 7: The mean extremal index. Like the parameters shown in Figures 3 and 4, the hierarchical mean is shown for the CanCM4 simulations.