Extreme value comparison of climate model simulations and observations

Mickey Warner

27 Feb 2018

Introduction

CanCM4 simulation classes (with $R = 10$ replicates each):

1. Decadal
2. Historical
3. Control

Observations over U.S. interpolated from weather stations

Factors:

1. Variable — Total Precipitation (`pr`) or Average Maximum Temperature (`tasmax`)
2. Season — Winter or Summer
3. Decade — 1962–1971 or 1990–1999
4. Region — California or USA

Locations

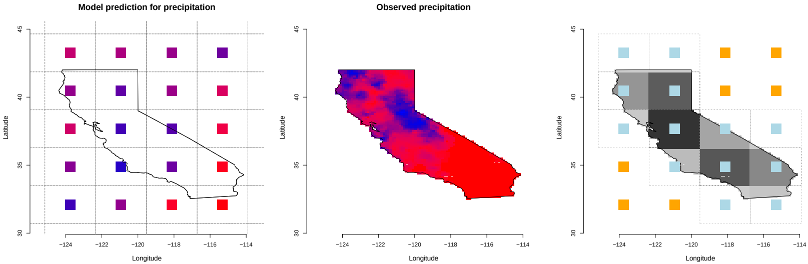
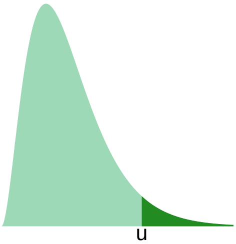


Figure: Left: CanCM4 simulation grid cells. Center: Observation locations. Right: method for computing weighted sum or average for CanCM4 to make values comparable with observations.

Extremes

For r.v. X and large threshold u, the exceedance $Y = X - u$, for $X > u$, approximately follows the generalized Pareto distribution (GPD), which has density

$$f_Y(y) = \frac{1}{\sigma}\left(1 + \xi\frac{y}{\sigma}\right)_+^{-1/\xi - 1}$$



Data processing

Two objectives before performing the analysis:

1. Make climate simulations comparable to observations
2. Get near-independent random variables for model fitting

These are accomplished by

1. Taking weighted sums (`pr`) or weighted averages (`tasmax`)
2. Computing anomalies based on DLMs, and
3. Declustering

Weighted sum or average

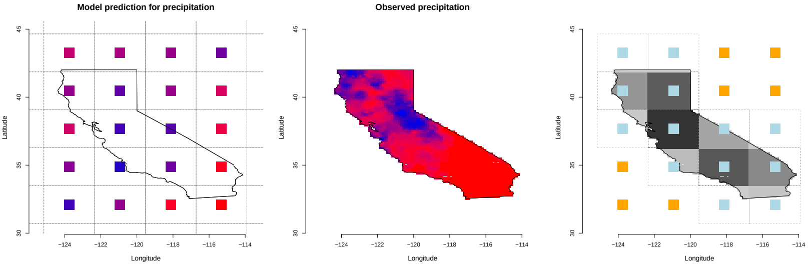
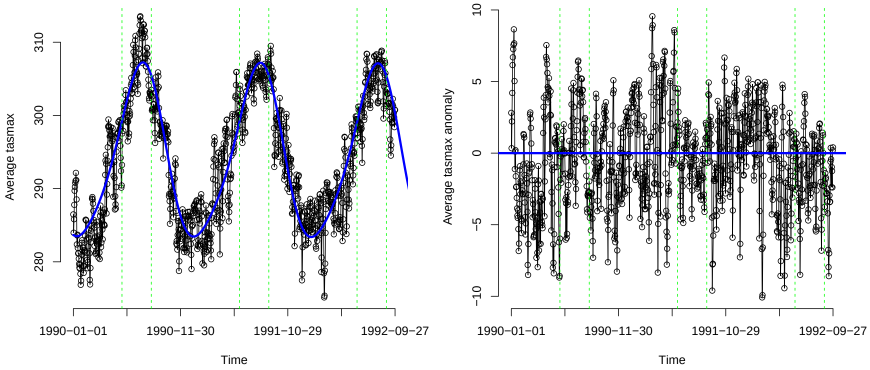


Figure: Left: CanCM4 simulation grid cells. Center: Observation locations. Right: method for computing weighted sum or average for CanCM4 to make values comparable with observations.

DLM-based anomaly



Extremal index (declustering)

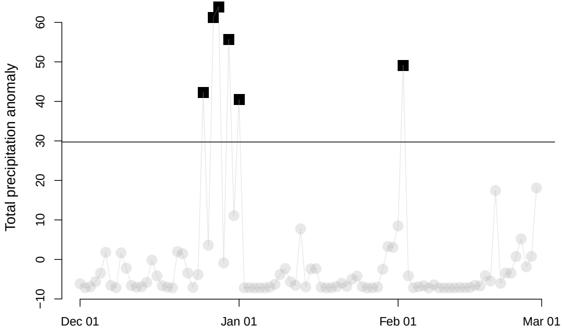
The extremal index θ is the inverse of the limiting mean cluster size

It can be estimated using interexceedance times, $T_i = S_{i+1} - S_i$, with a log-likelihood of

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

p is the probability of not exceeding the threshold

Declustering



Likelihood

Replicate i, observation j, exceedances $Y_{ij} = X_{ij} - u$, and keep only those $Y_i > 0$. These have likelihood

$$L(\mathbf{y}; \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\zeta}) = \prod_{i=1}^{R} \left[(1 - \zeta_i)^{n_i - k_i} \zeta_i^{k_i} \prod_{j=1}^{k_i} \frac{1}{\sigma_i} \left(1 + \xi_i \frac{y_{ij}}{\sigma_i} \right)_+^{-1/\xi_i - 1} \right]$$

n_i is the number of X_{ij}'s
k_i is the number of Y_{ij}'s
ζ_i is the probability of exceeding the threshold

Priors

These priors complete the hierarchical model formulation. Greek letters are random variables while English letters are fixed.

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

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

Return level

For a distribution G, the return level x_m is the solution to

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

The value x_m is exceeded on average once every m observations.

For the GPD, the return level is given by

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

Bhattacharyya distance

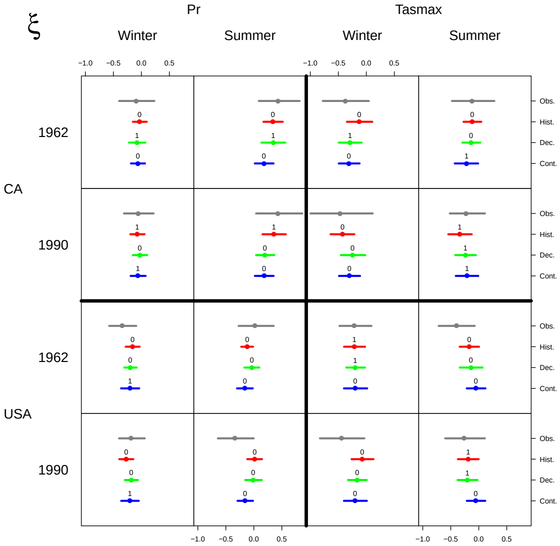
Bhattacharyya coefficient

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

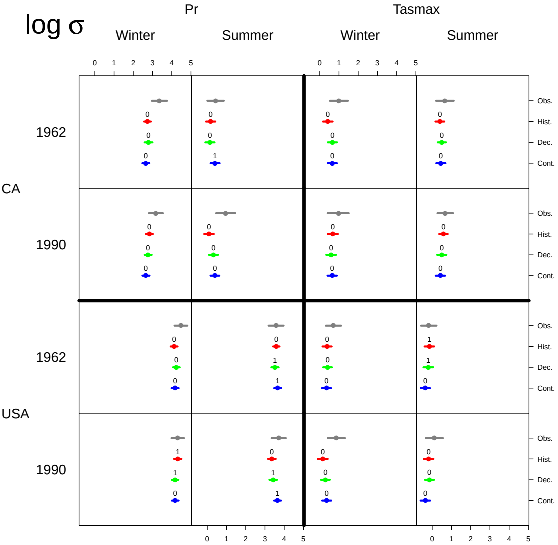
Bhattacharyya distance

$$D_B(p,q) = -\log BC(p,q).$$

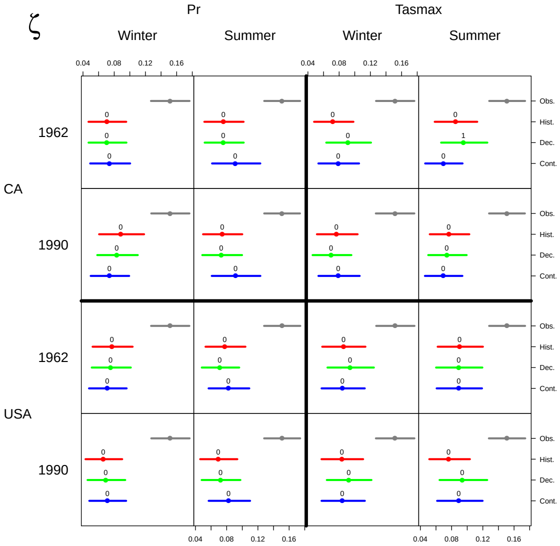
D_B is computed between parameters in the replicates (and observations) and parameters in the hierarchy.

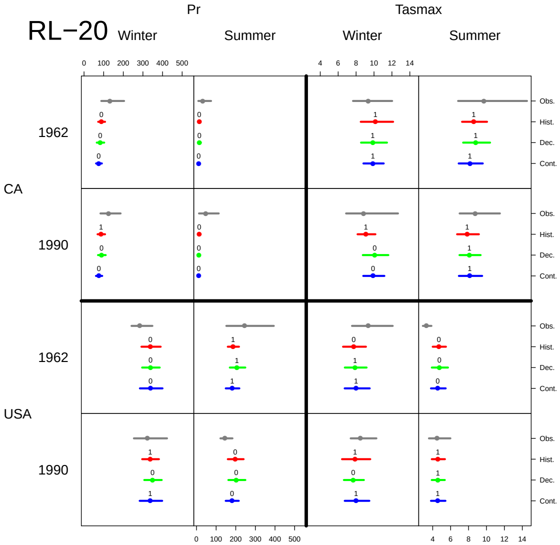


ξ
Pr
Tasmax
Winter
Summer
Winter
Summer
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.
−1.0
−0.5
0.0
0.5

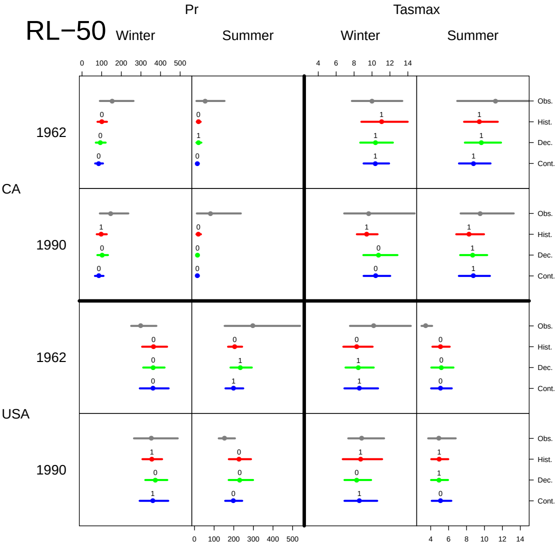


log σ
Pr
Tasmax
Winter
Summer
Winter
Summer
1962
1990
CA
1962
1990
USA
Obs.
Hist.
Dec.
Cont.


ζ
Pr
Tasmax
Winter
Summer
Winter
Summer
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.
0.04
0.08
0.12
0.16

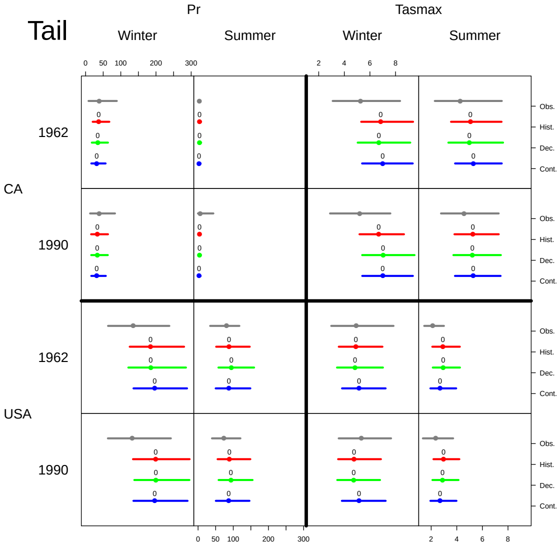


RL−20
Pr
Tasmax
Winter
Summer
Winter
Summer
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.

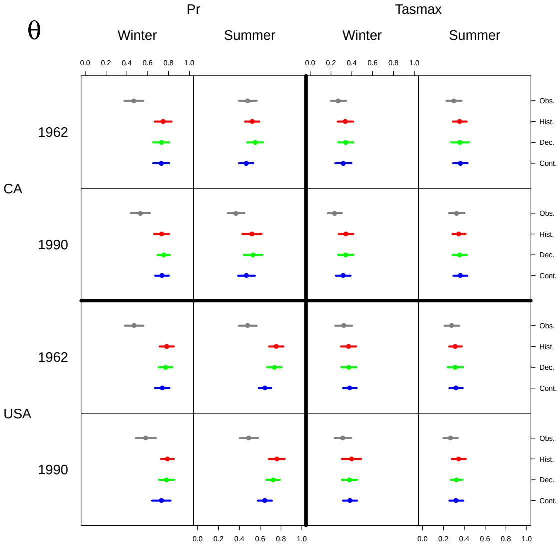


RL−50
Pr
Tasmax
Winter
Summer
Winter
Summer
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.

Tail



Pr
Tasmax
Winter
Summer
Winter
Summer
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.
0
50
100
200
300
2
4
6
8


θ
Pr
Tasmax
Winter
Summer
Winter
Summer
0.0 0.2 0.4 0.6 0.8 1.0
CA
USA
1962
1990
1962
1990
Obs.
Hist.
Dec.
Cont.

Bivariate analysis

The univariate analysis allows us to make comparisons between simulations and observations, but not to model their extremal relationship.

A key concept in multivariate extreme value analysis is asymptotic tail dependence, described by the following quantity

$$\chi = \lim_{z \to z^*} P(X > z | Y > z)$$

where X and Y share a common marginal distribution and z^* is the (possibly infinite) right end-point of X and Y.

Note: even for normal distributions with correlation $\rho < 1$, $\chi = 0$.

Simple Pareto process

For stochastic process X, define

$$T_t X = \left(1 + \xi \frac{X - u_t}{\sigma_t}\right)_+^{1/\xi}.$$

Under certain conditions,

$$\lim_{t \to \infty} P\left(T_t X \in A \middle| \sup_{s \in S} T_t X(s) > 1\right) = P(W \in A)$$

where W is a simple Pareto process (SPP).

Note: for SPP, we cannot have $\chi = 0$.

Simple Pareto process, continued

Climate simulations and observations are transformed with

$$W_i(s_j) = T_t X_i(s_j) = \left(1 + \xi(s_j)\frac{X_i(s_j) - u_t(s_j)}{\sigma_t(s_j)}\right)_+^{1/\xi(s_j)}.$$

using posterior means for ξ and σ, where s_j denotes the data source and i denotes individual observations.

Let $\mathbf{W}(s) = (W_1(s), \ldots, W_{n(s)}(s))^\top$, and form the bivariate vector $\mathbf{W}_{12} = (\mathbf{W}(s_1), \mathbf{W}(s_2))$ for some s_1, s_2 pair.

Rows of $\mathbf{W}_{12}$ are considered samples from a simple Pareto process.

Bivariate data

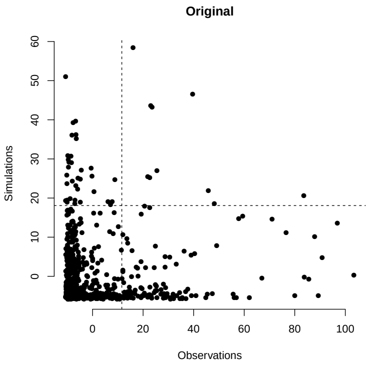


Figure: Untransformed data for CA winter precipitation, observations against the first control replicate. Dashed lines mark the thresholds.

Bivariate data

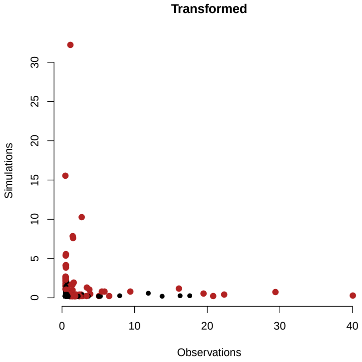


Figure: Data transformed to have Pareto marginals in the exceedances. Red dots mark the points that are kept after declustering.

Asymptotic tail dependence for SPP

Each row of $\mathbf{W}_{12}$ can be written as

$$(Y_i V_i(s_1),\ Y_i V_i(s_2)),$$

where Y_i is a standard Pareto random variable and $V_i(s_j) \geq 0$ with $V_i(s_1) \vee V_i(s_2) = 1$, for all i.

It can be shown that

$$\chi = E\left(\frac{V(s_1)}{E(V(s_1))} \wedge \frac{V(s_2)}{E(V(s_2))}\right)$$

Bivariate data

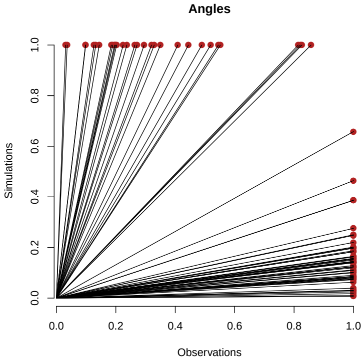


Figure: The red dots mark the points $(V_i(s_1), V_i(s_2))$. These are constrained to lie on the unit supremum cone.

Asymptotic tail dependence for SPP, continued

Given the supremum constraint for $V_i(s_1)$ and $V_i(s_2)$, we can write rows of $\mathbf{W}_{12}$ in terms of Y_i and the angle

$$\phi_i = \frac{2}{\pi} \arctan\left(\frac{V_i(s_2)}{V_i(s_1)}\right) \in [0, 1].$$

The angles $\phi_1, \ldots, \phi_n$ are modeled with a Bernstein-Dirichlet prior (BDP), a flexible model for density estimation.

Posterior samples for ϕ are back-transformed to $(V(s_1), V(s_2))$ and χ is estimated.

Bivariate data

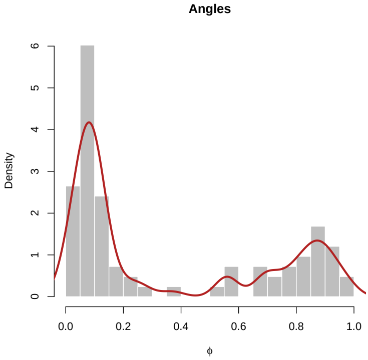
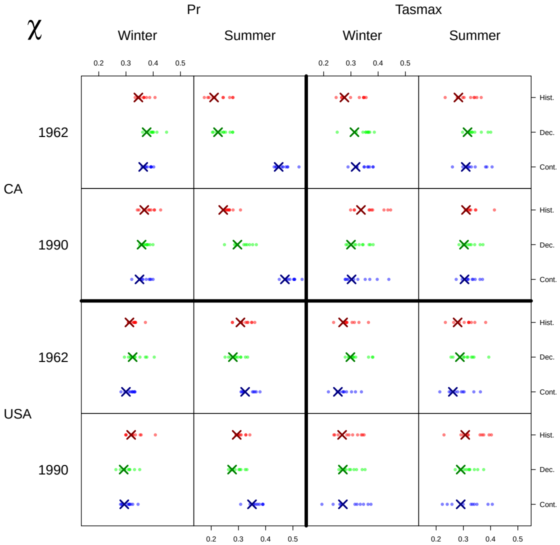


Figure: $(V_i(s_1), V_i(s_2))$ transformed to ϕ_i.


χ
Pr
Tasmax
Winter
Summer
Winter
Summer
0.2
0.3
0.4
0.5
1962
1990
CA
1962
1990
USA
Hist.
Dec.
Cont.