Computationally-efficient Spatial Analysis of Precipitation Extremes Using Local Likelihood

Chris Paciorek
Department of Statistics
University of California, Berkeley

Michael Wehner
Lawrence Berkeley National Laboratory

Prabhat
Lawrence Berkeley National Laboratory

August 30, 2011

Research supported by DOE DE-AC02-05CH11231 .

Abstract

We develop a local likelihood approach to spatial analysis of extremes, building on location-specific GEV and point process modeling. The goal is to borrow strength spatially without the computational expense of fitting a full spatial extremes model. The work is part of a larger software development effort to develop parallel software tools for working with high-resolution climate model output, such as is being produced for CMIP5. Our work includes statistical, software development, and climate analysis threads:

Statistical: We develop a local likelihood strategy, proposing to use the bootstrap for uncertainty estimation, with potential subsequent spatial smoothing. The strategy is inherently naively parallel across space.

Software development: We build on the existing ismev package in R, improving computational efficiency of the core fitting routines and developing local likelihood versions. Software will be deployed serially in R, with parallelization through a back-end run by VisIt, a high-performance visualization software developed by the national labs.

Climate analysis: We present initial results on trends in observed precipitation extremes in the US and on tail properties of extremes in climate model runs.

Empirical Analyses

We focus on changes in seasonal precipitation as different atmospheric phenomena drive precipitation in different seasons.

Our analyses are as follows:

- Analysis of trends in US station data (the coop station network) for 1949-2010 by season.

- Analysis of tail properties in long climate model control runs (CCSM, CanESM) by season.
  - Here our goal is to understand why long tails are common empirically when short tails might be expected scientifically.
  - Are the long tails driven by mixing of different types of storms?

Statistical Modeling: Local Likelihood

The basic approach starts with the GEV for block maxima or point process (PP) model for peaks over threshold (POT). Let $L_i(y_i; \mu(t), \sigma(t), \xi(t))$ be the GEV or PP log likelihood for one spatial site, with parameters that are potentially nonstationary in time. E.g., we might model the location parameter as a polynomial or spline in time:

$$\mu(t) = \beta_0 + \beta_1 t + \cdots + \beta_p t^p$$

To estimate parameters and return levels for a focal site, local likelihood uses a weighted average of the likelihoods for all N_i sites near the focal site, j:

$$L_j(y) = \sum_{i \in \text{neigh}(j)} w_i L_i(y_i; \mu_i(t), \sigma_i(t), \xi_i(t)$$

Here we might model the location parameter as varying (locally) linearly in geographic space (x, y) and elevation (z):

$$\mu(t, x, y, z) = \beta_0 + \beta_1 t + \beta_x(x_i - x_j) + \beta_y(y_i - y_j) + \beta_z(z_i - z_j)$$

$\sigma(\cdot)$ and $\xi(\cdot)$ could be modeled similarly, but are often taken to be stationary in time and might be taken to be constant across sites within a local neighborhood as well.

Statistical Modeling: Bootstrap

The local likelihood approach will produce MLEs and site-specific asymptotic standard errors but not account for dependence across sites (which of course share observations and will be heavily dependent).

To estimate uncertainty, including dependence in uncertainty across locations, we propose the bootstrap, preserving the spatial and temporal structure.

Bootstrap options include:

1. Nonparametric resampling of data within temporal blocks (all data within a year kept together)

2. Parametric resampling of model residuals (all residuals within a year kept together).

To preserve the spatial dependence, we propose to resample the same sequence of years across sites within a bootstrap replicate. We assume limited dependence across years, but dependence within years, so resample at the yearly aggregation.

Statistical Modeling: Spatial Smoothing

The local likelihood approach is of course heavily influenced by the choice of spatial weighting. We propose a **simple Gaussian kernel**, truncated at 3 standard deviations. The kernel bandwidth might be chosen based on expert judgment about homogeneity of climate.

With the results from the local likelihood + bootstrap, we propose to **work with return levels rather than model parameters**, to give us a one-d spatial field.

We then plan to consider **further smoothing** of the field of estimated return values. Here the challenges are:

1. Choosing a spatial model that accounts for nonstationarity in space and elevation

2. Incorporating the high-dimensional bootstrap empirical covariance of the estimated return levels. This may need to be smoothed to ensure positive definiteness.

Software Development

1. We have modified the ismev gev.fit() and pp.fit() functions to handle missing data and improve computational efficiency
   - In particular for pp.fit() we work only with observations above the threshold.
2. We have extended gev.fit() and pp.fit() to do local likelihood.
3. We plan to develop R code for the bootstrap-based assessment of uncertainty.
4. As part of a larger team, we are developing the ability to run R within the VisIt parallel visualization software using VtK. The goal is to have multiple individual R instances fitting models for blocks of stations, with VisIt handling the parallelization over the spatial blocks, in particular parallel I/O from NetCDF files.

Climate Analysis Methods: Precipitation Trends by Season

- Models have linear trend in the location parameter and no trend in scale and shape. Results are given as the 40-year change in return values, where this holds for any return interval because only the location parameter changes with time.

- GEV: Require at least 40 years of data with at least 80 days in a season

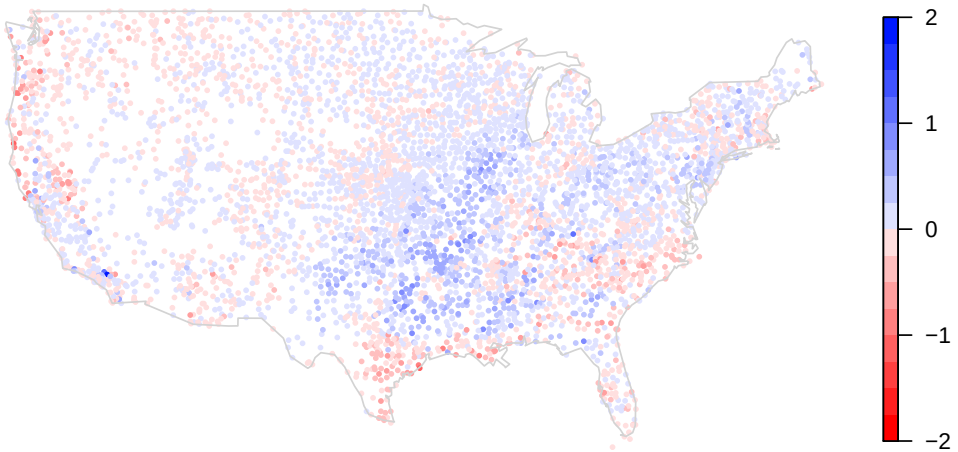
- POT: Require at least 67% of days with available data, threshold computed based on days with > 1 mm precipitation.
  - Local likelihood implementation uses Gaussian kernels with s.d. of 30 km and parameters that are constant in space. This is an initial analysis only.

Climate Analysis Results: Precipitation Trends by Season

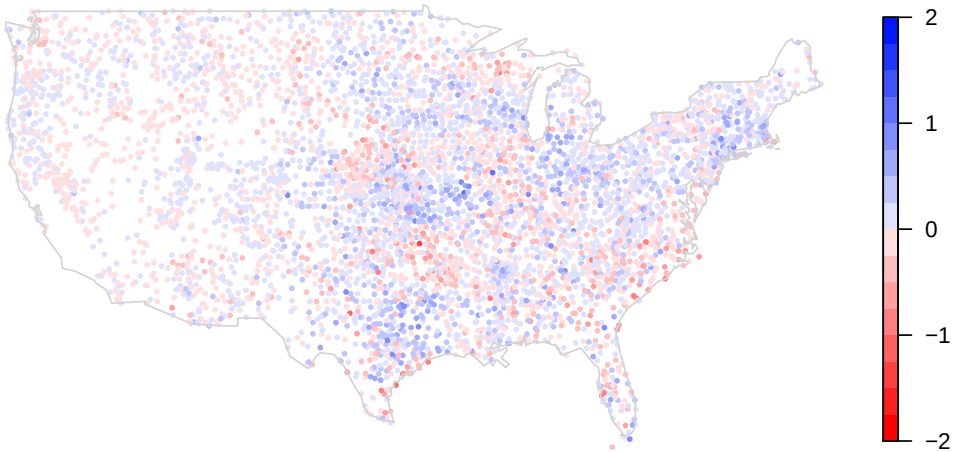
- All model fits show spatial patterns, with noticeable areas of increasing extreme precipitation, and differences between seasons.
- Model fits broadly agree on the patterns.
- Local likelihood (3rd set of panels) produces smoother, more robust, estimates.

GEV Model for Maxima

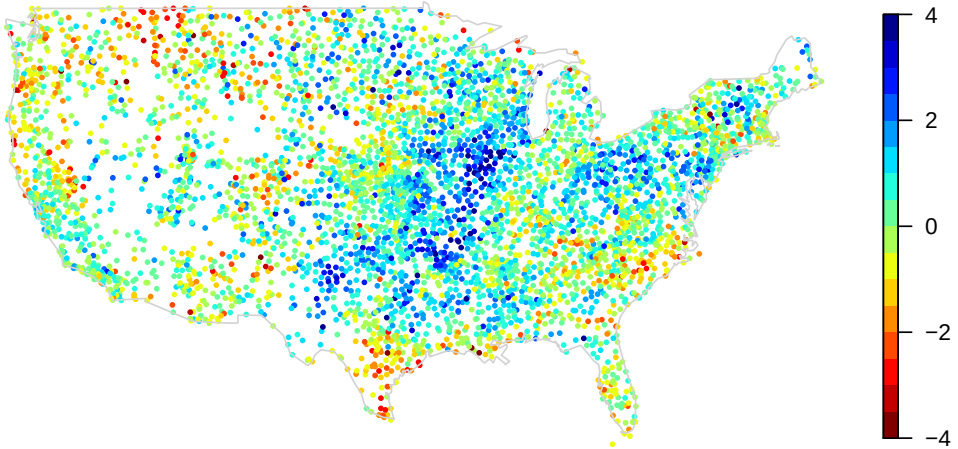
DJF − 40 yr change in return values (inches)



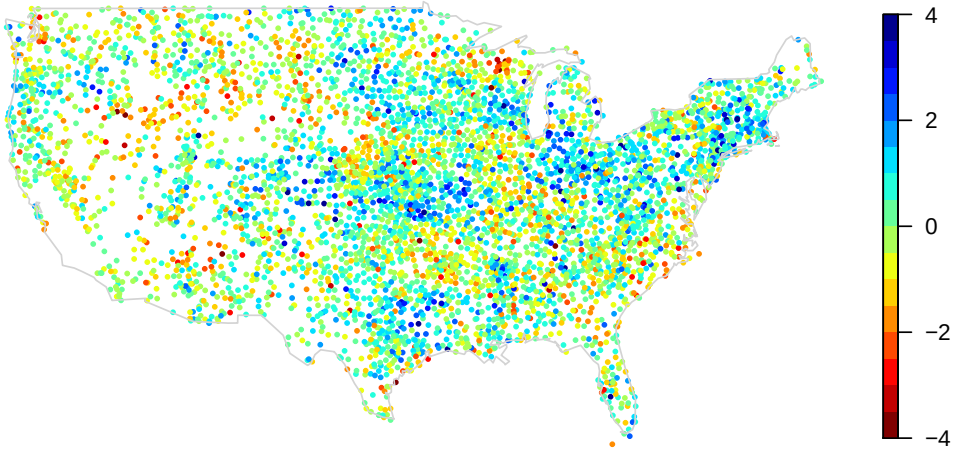
JJA − 40 yr change in return values (inches)



DJF − 40 yr Z score

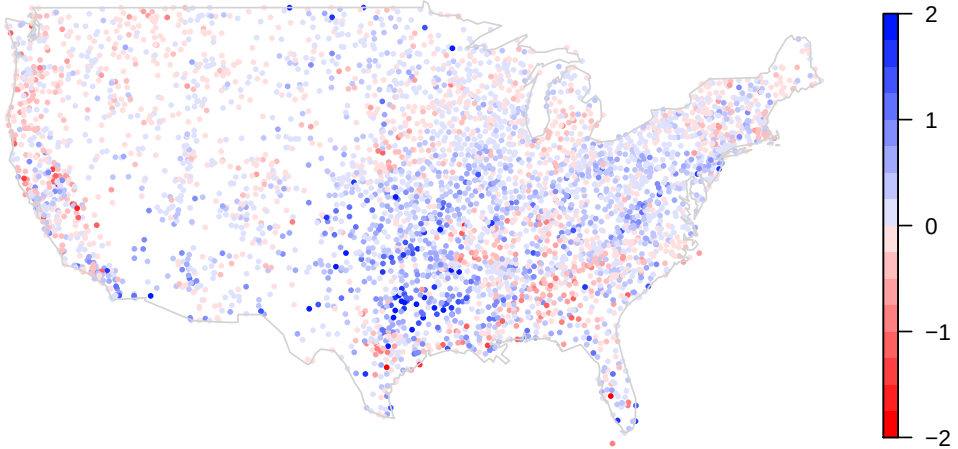


JJA − 40 yr Z score

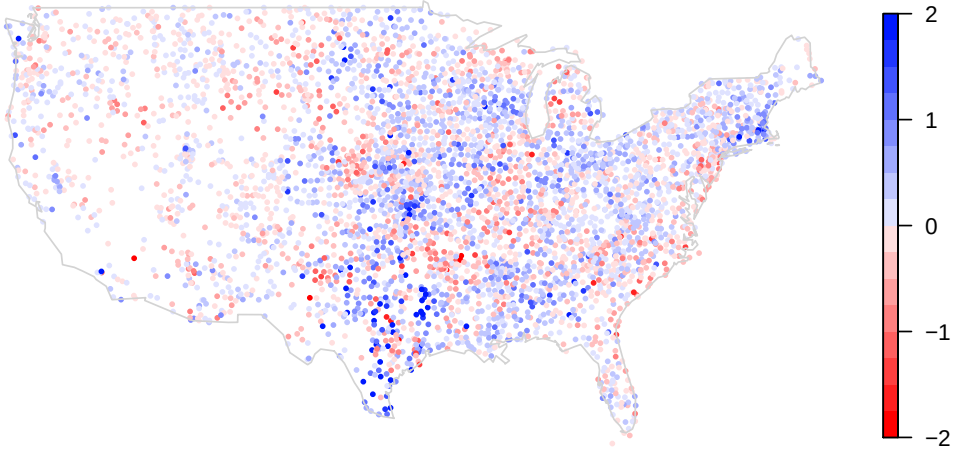


POT Model for Threshold Exceedances

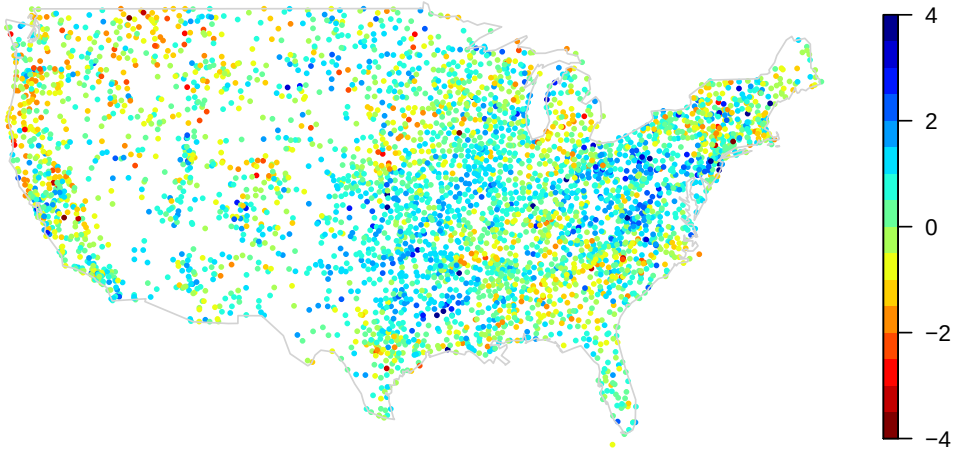
DJF − 40 yr change in return values (inches)



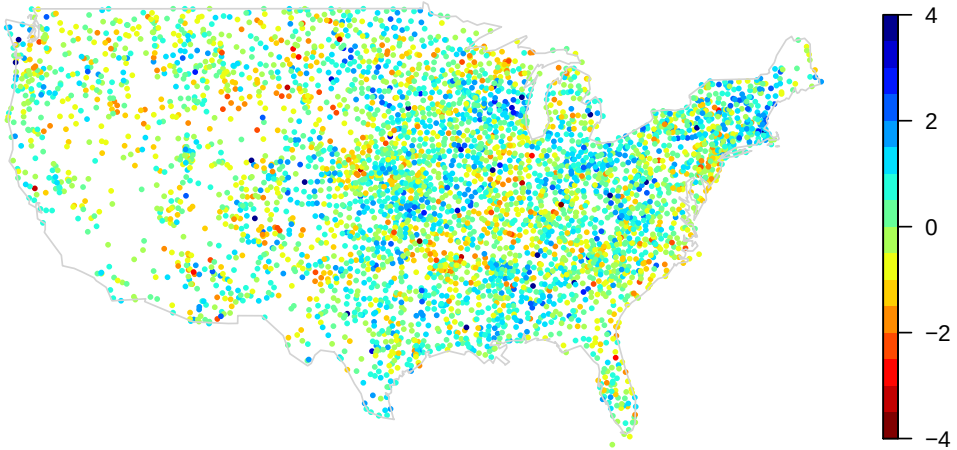
JJA − 40 yr change in return values (inches)



DJF − 40 yr Z score

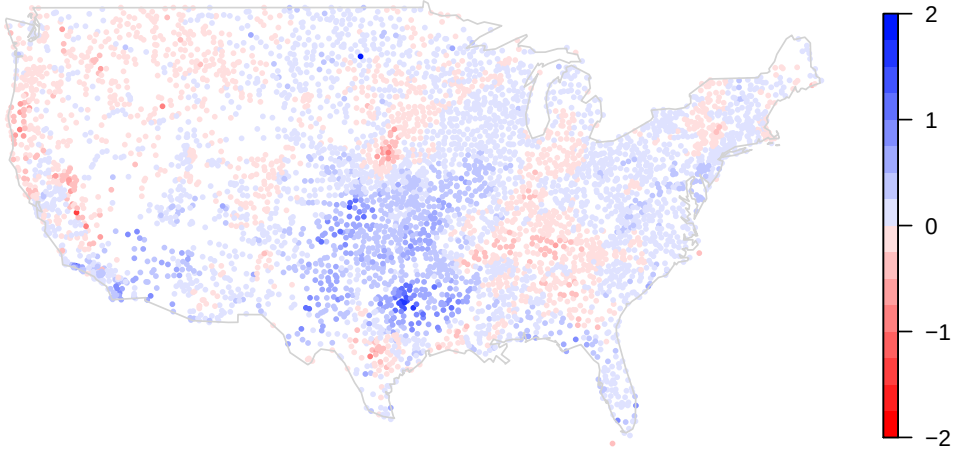


JJA − 40 yr Z score

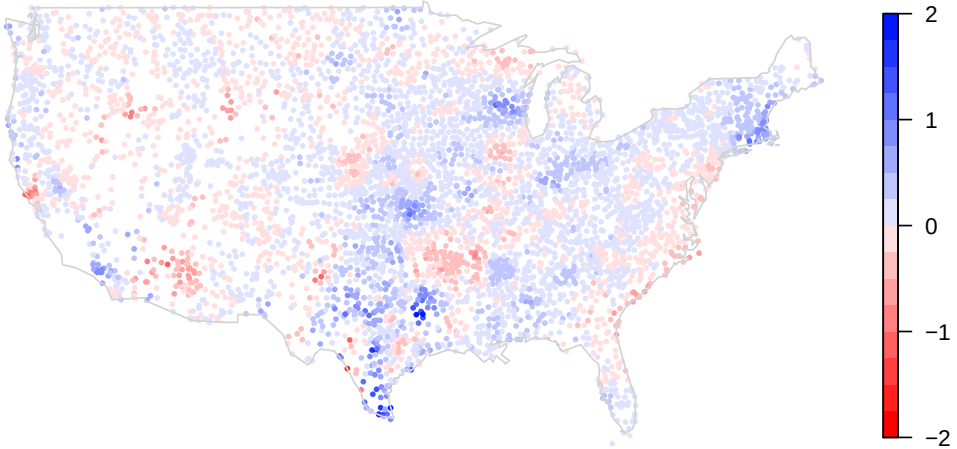


POT local likelihood

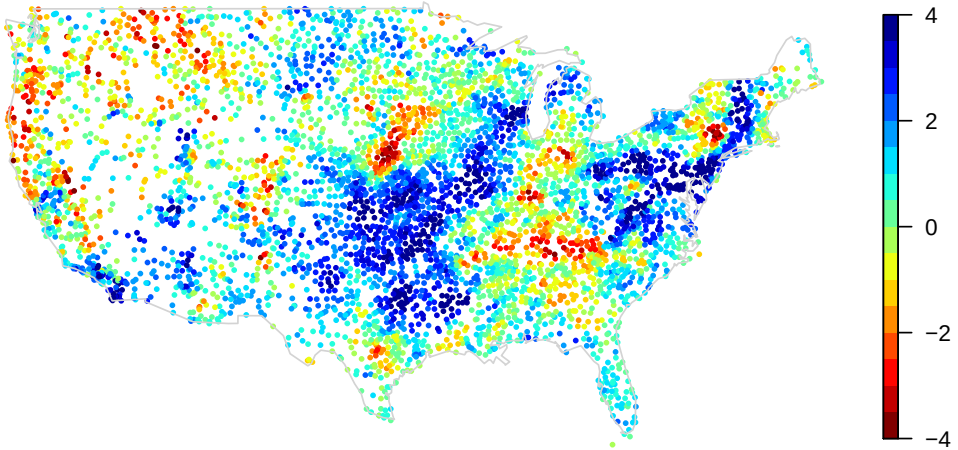
DJF − 40 yr change in return values (inches)



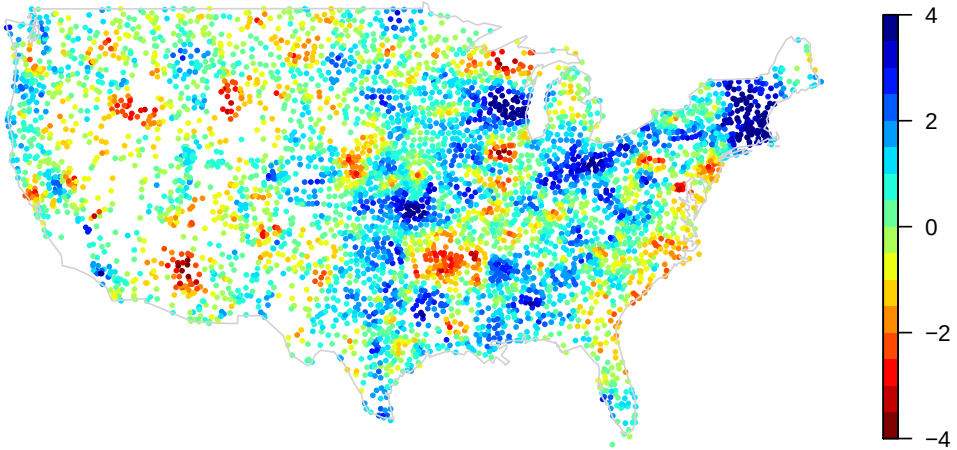
JJA − 40 yr change in return values (inches)



DJF − 40 yr Z score



JJA − 40 yr Z score



Climate Analysis: Shape Parameter Estimation

Motivation: Empirical analyses often estimate $\xi > 0$ (the Frechet distribution), corresponding to a tail that decays polynomially. Given that precipication is limited by physical constraints, one might expect $\xi < 0$ (the Weibull distribution), with a tail that decays to zero at a finite value.

Some potential explanations for long-tailed estimates:

1. We may not have enough years of data to get precise estimation of ξ

2. We may not be in the asymptotic regime (too few values in each block for GEV or too low a threshold for POT)

3. We might have precipitation resulting from different atmospheric phenomena, giving us a mixture of distributions, which may affect whether we are in the asymptotic regime.

Methods: We fit GEV and POT models for CCSM (and CanESM, not shown) control runs with 100s of years of output, with all parameters constant over time. Multi-day runs were replaced by the single day with the most precipitation.

Climate Analysis: Shape Parameter Results

1. Even with seasonal stratification and very many years of data, fat tails are common in both GEV and POT models, suggesting Explanations #1 and #3 (previous panel) cannot fully explain what is going on.

2. GEV models show many areas with fat tails ($\hat{\xi} > 0$), some of which are in dry areas (e.g., Sahara, Antarctica).

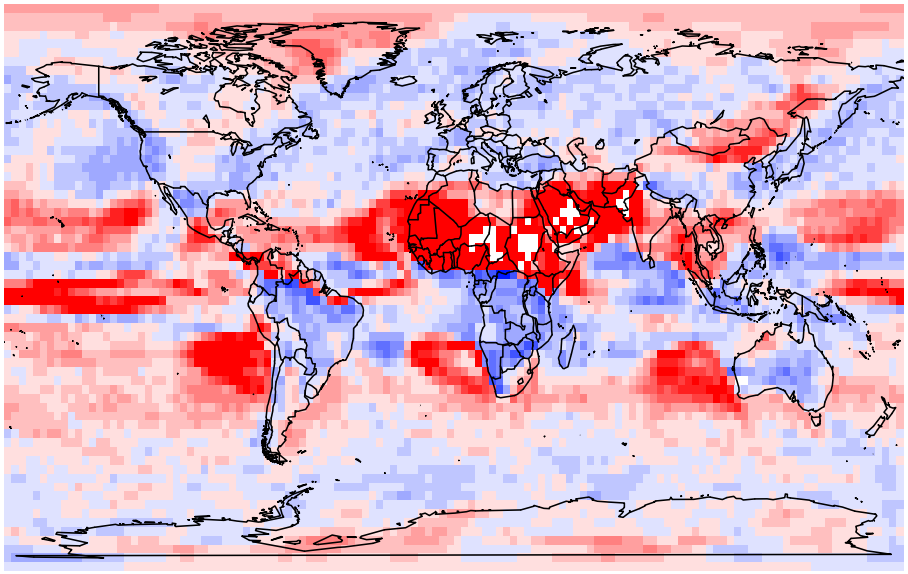
3. Under POT modeling, dry areas show less evidence of fat tails (perhaps supporting Explanation #2). However, some areas with fat tails remain with POT.

4. Use of a higher threshold in POT modeling leads to unstable estimates even with long control runs (not shown), so further assessment of Explanation #2 is difficult.

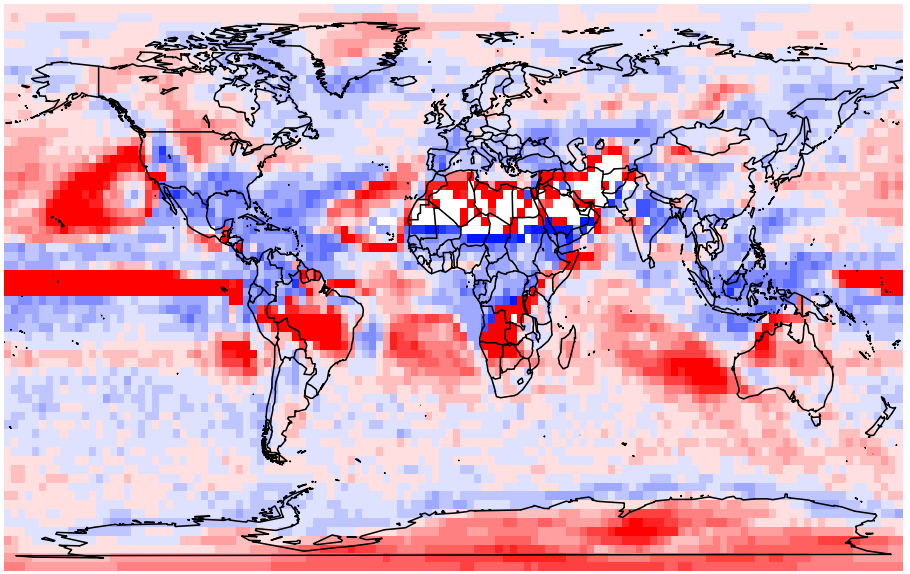
5. Fat tails are common in the station fits as well (not shown). Use of local likelihood produces shape parameter estimates that are sensitive to outliers because a site with outliers influences estimation at multiple sites.

CCSM control run, GEV

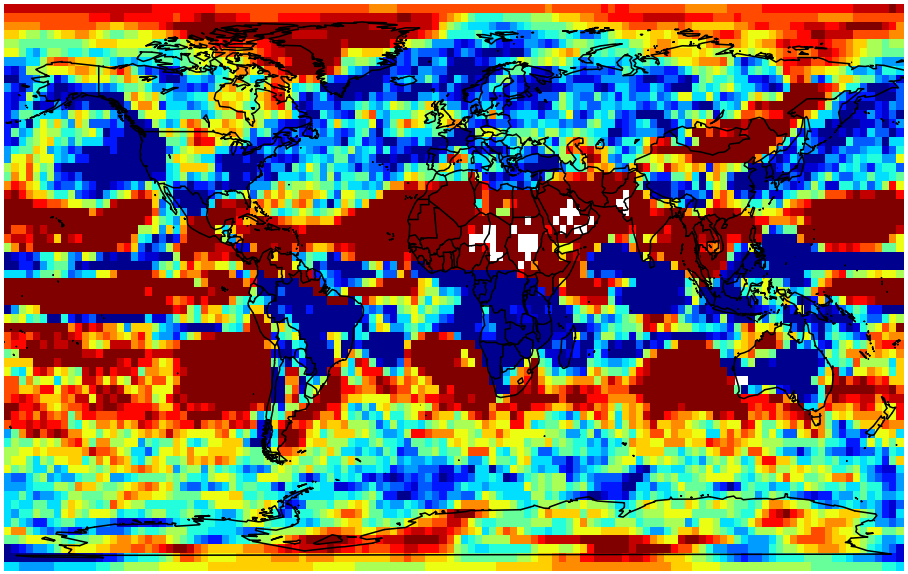
DJF − shape parameter estimate



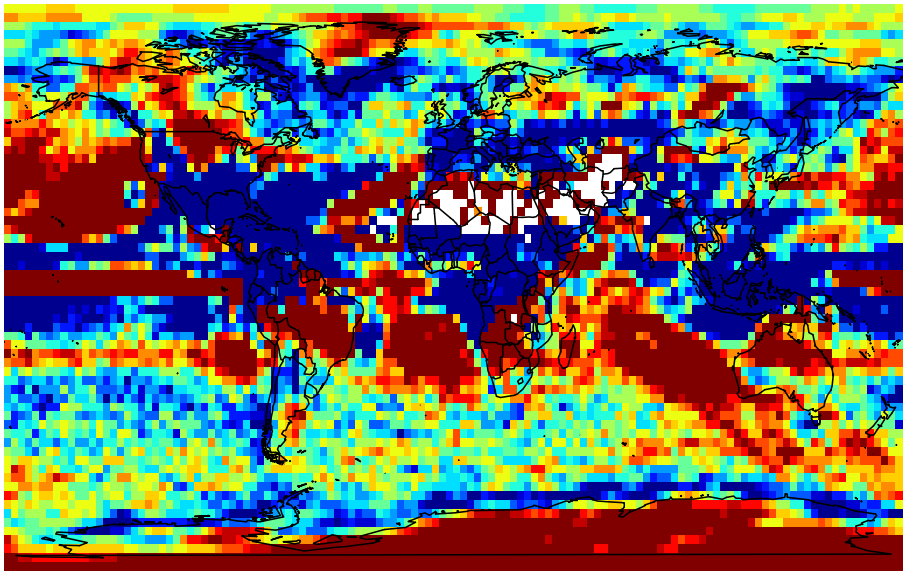
JJA − shape parameter estimate



DJF − shape estimate Z score

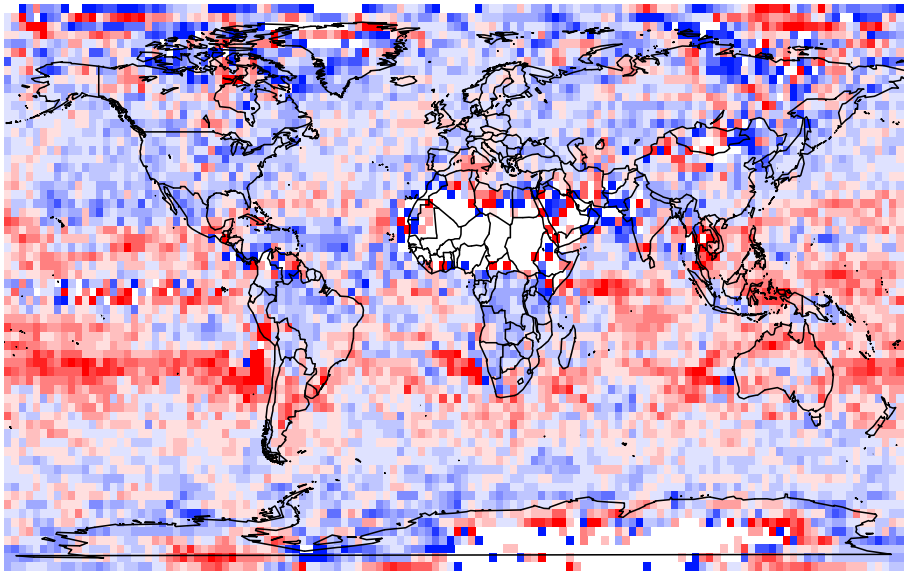


JJA − shape estimate Z score

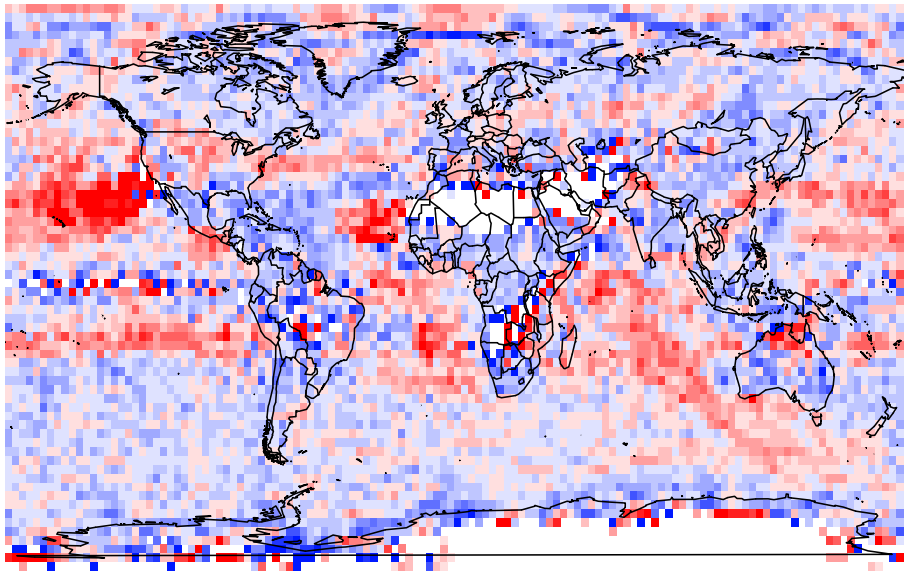


CCSM control run, POT 99%ile threshold

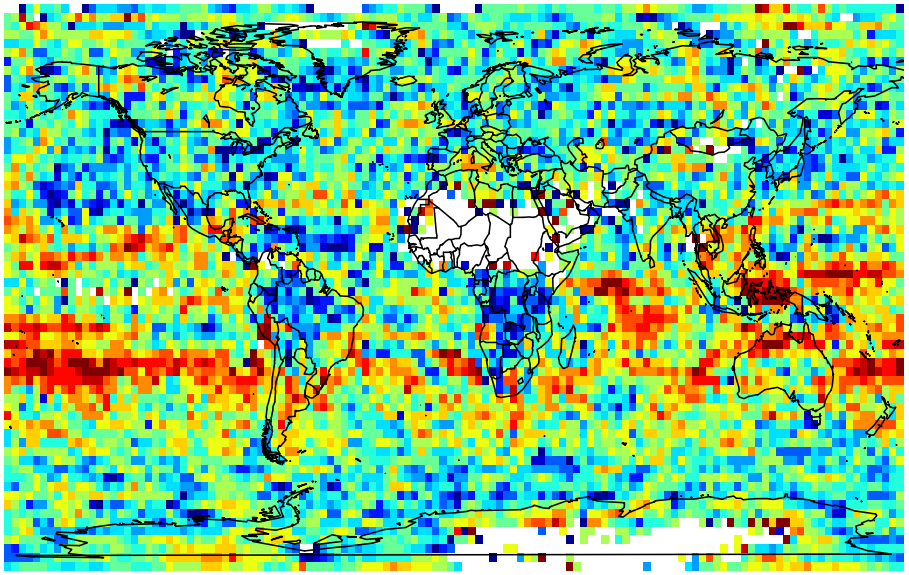
DJF − shape parameter estimate



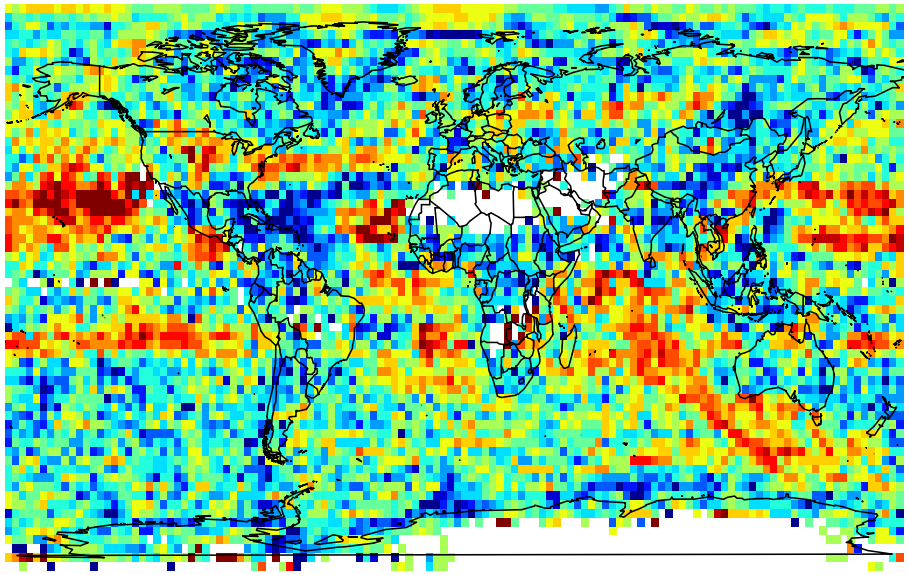
JJA − shape parameter estimate



DJF − shape estimate Z score



JJA − shape estimate Z score



Future Work

1. Explore whether mixtures of distributions lead to fat-tailed extremes, in particular if ENSO explains fat tails in climate model output in the eastern Pacific and other teleconnected regions.

2. Fit local-linear specification within local likelihood modeling of station observations.

3. Application of the methods to high-resolution GCM runs in CMIP5 within the VisIt parallel infrastructure.

4. Consider statistical methods and computational implications of spatial smoothing of local likelihood fits, with an empirical covariance matrix from bootstrapping.