Trends and patterns in extreme precipitation using extreme value analysis

Chris Paciorek
(Department of Statistics; University of California, Berkeley)
Michael Wehner (LBNL)
Prabhat (LBNL)

www.stat.berkeley.edu/~paciorek

Research supported by DOE DE-AC02-05CH11231

August 2013

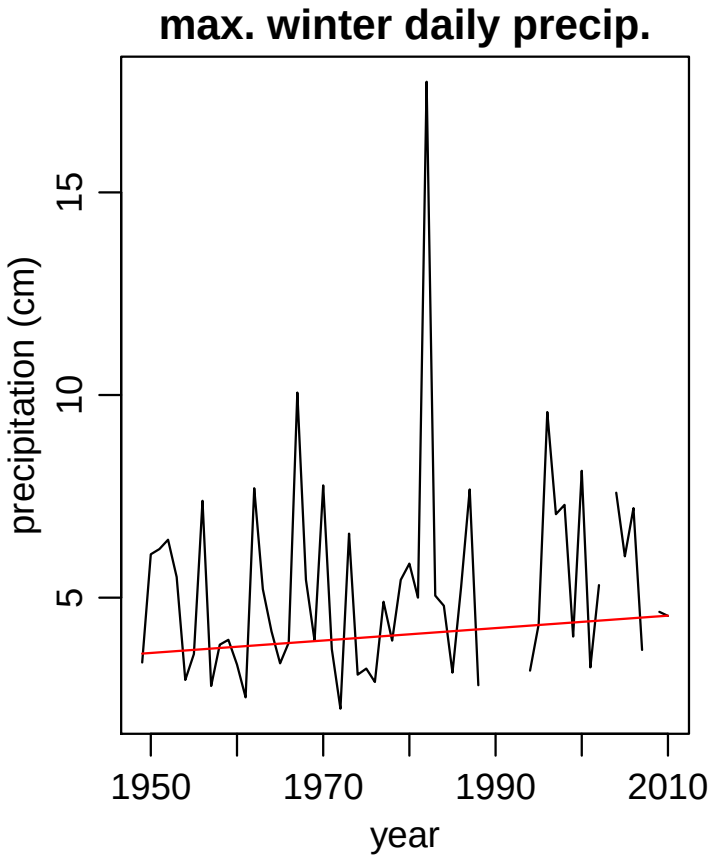
Statistical extreme value theory

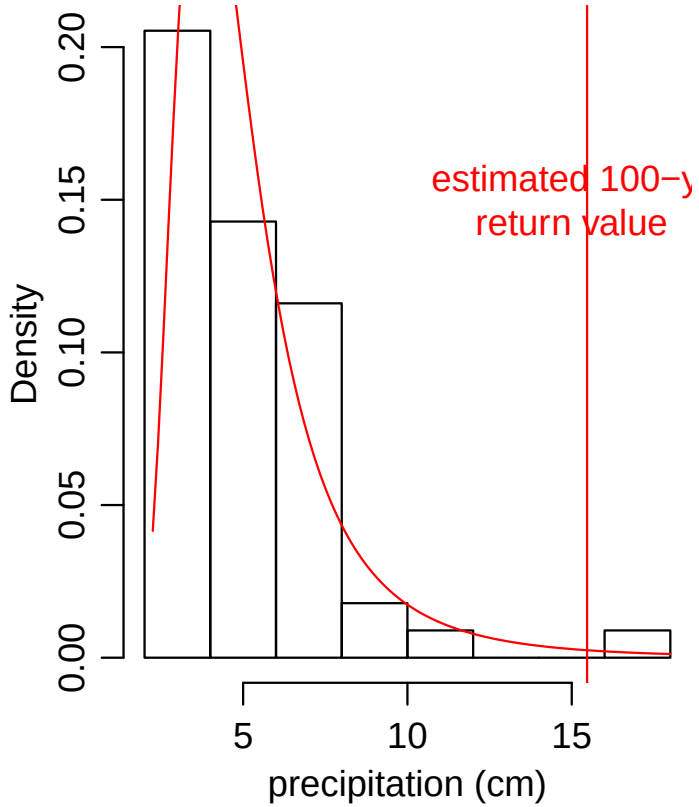
- The Generalized Extreme Value (GEV) distribution:

$$F(x) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)$$

- Location parameter μ, scale parameter σ, shape parameter ξ, generally fit via maximum likelihood.
- Unites the bounded Weibull distribution ($\xi < 0$), exponential-tailed Gumbell distribution ($\xi = 0$), and heavy-tailed Frechet distribution ($\xi > 0$)
- Asymptotic theory says that the distribution of block maxima (or minima) converges to the GEV distribution as the block size goes to infinity.
- By the quantiles of the GEV distribution, the MLE for the $1/p$-year return level is: $\hat{z}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1-(-\log(1-p))^{-\hat{\xi}}\right)$

Example: Berkeley winter precipitation





Nonstationary extreme value analysis

Extremes may also vary by season, by time, and with covariates (in particular teleconnections such as ENSO).

- A basic strategy:
  - Fit separate models by season
  - Fit nonstationary models with respect to time and ENSO:

$$F(x_t) = \exp\left(-\left[1 + \xi_t\left(\frac{x_t - \mu_t}{\sigma_t}\right)\right]^{-1/\xi_t}\right)$$

- One might have all three parameters vary with time and ENSO (linearly, polynomially, or based on splines).
- Analyses often find little evidence (based on likelihood ratio tests) that ξ (and even σ) are varying with time, though ξ in particular is hard to estimate even in a stationary model.
- A basic model is linear in time (and possibly ENSO) in μ only, as a first-order estimate of the trend over time.

Peaks over threshold (POT) analysis

- An alternative is to model all the exceedances over a high threshold, c (e.g., the 95%ile or 99%ile of all rainy days in the data). Why?
  - Don't 'waste' extreme observations that are not the block maximum
  - Readily allow for missing data when data are missing for reasons unrelated to weather
  - For precipitation, not clear that block maxima are appropriate in dry regions/seasons when there are few wet days (asymptotic conditions may not be satisfied)
- A disadvantage is requiring the raw daily data, whereas block maxima can use data summaries/indices (e.g., HadEX2 available only as indices)

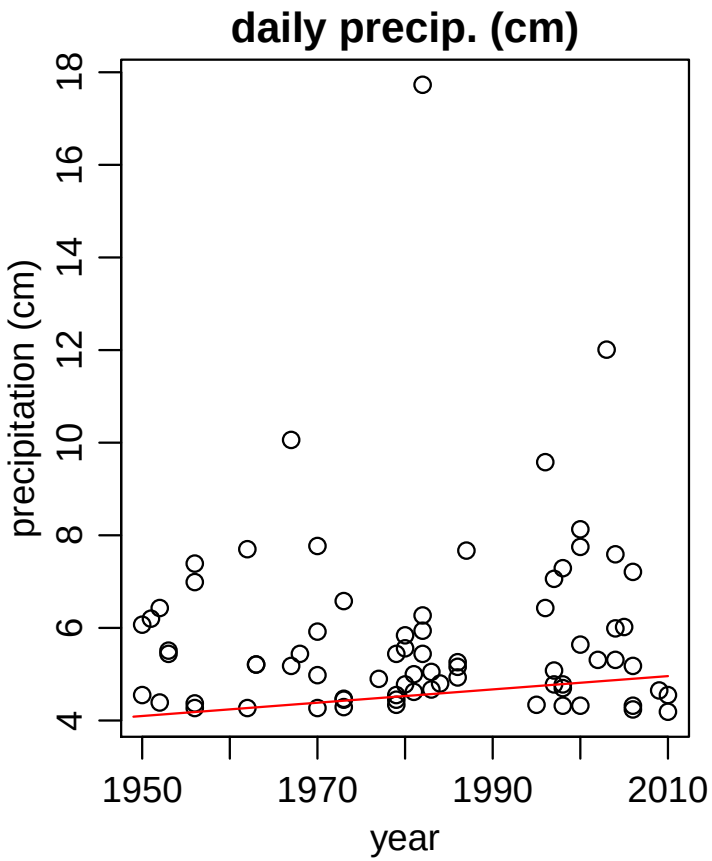
Point processing modeling

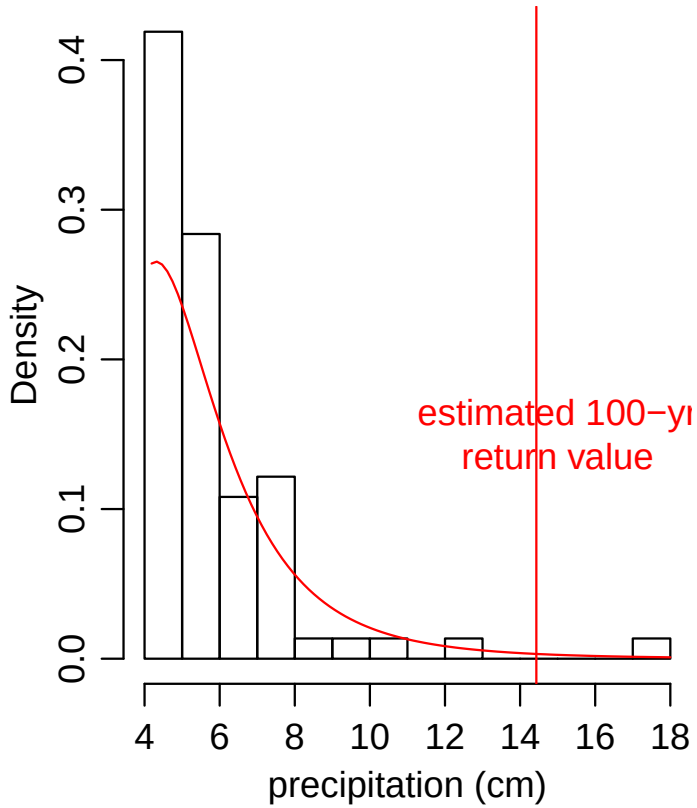
- The point process model implements the peaks-over-threshold approach by specifying the probability of the number of exceedances (the intensity measure) and the likelihood of the actual exceedances (the intensity function). The stationary version is:

$$L(\mu, \sigma, \xi; x_1, \ldots x_n) \propto \exp\left(-n_y \left[1 + \xi\left(\frac{c-\mu}{\sigma}\right)\right]^{-1/\xi}\right) \cdot \prod_{i=1}^{N(A)} \frac{1}{\sigma}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-1/\xi - 1}$$

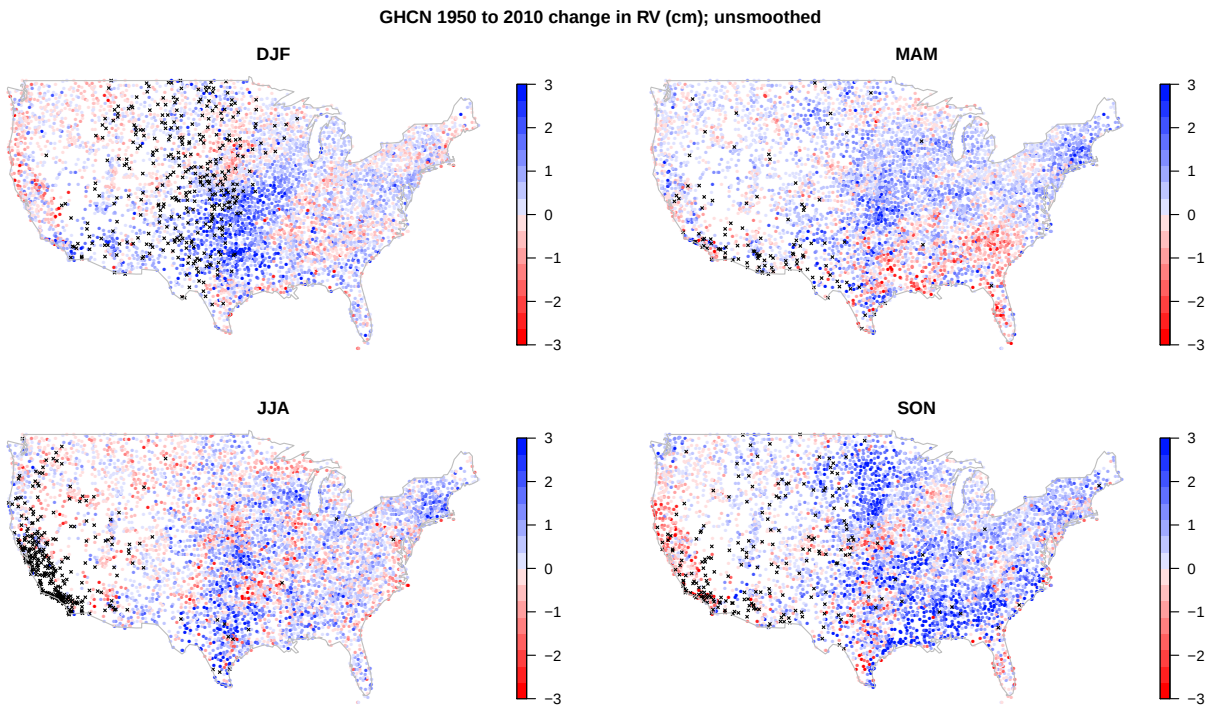
- The parameters are equivalent to the GEV parameters and can be used to compute return levels.
- Asymptotics are with respect to the threshold getting larger.

Example: Berkeley winter precipitation





US extreme precipitation



black x's are locations where statistical fit is unstable

- (1) estimated patterns are noisy because of statistical uncertainty and (2) pointwise uncertainty is large.
- We can mitigate these problems by borrowing strength spatially and (for climate model output) by fitting to initial condition ensembles.

Spatial extreme value analysis

- Given the sparsity of data and the spatial structure of climate/weather, an obvious goal is to do a spatial analysis of multiple locations, borrowing strength to:
  - Better estimate spatial patterns
  - Reduce uncertainty
- Standard spatial analyses have assumed spatially-correlated parameters, but conditionally IID observations.
  - Hierarchical Bayesian approaches have been a common approach: Cooley, Gelfand, Sang, Shaby, and others
  - Computation is a big hurdle and MCMC performance can be poor
- Analysts often remove consecutive exceedances to reduce temporal autocorrelation
- Some recent work on models that allow for spatially-correlated observations.

Our perspective

- Given the size of observation and climate model output datasets and the increasing spatial resolution of models, a hierarchical modeling strategy fit by MCMC is not practical for most large-scale and production-mode climate analysis.
- Our focus:
  - Location-specific analysis (embarrassingly parallel)
  - Basic models for temporal change and associations with teleconnections (linear)
  - Stratify by season rather than modeling seasonality
  - Assess uncertainty via bootstrapping to deal with temporal and spatial structure
  - Development of parallel software

Bootstrapping

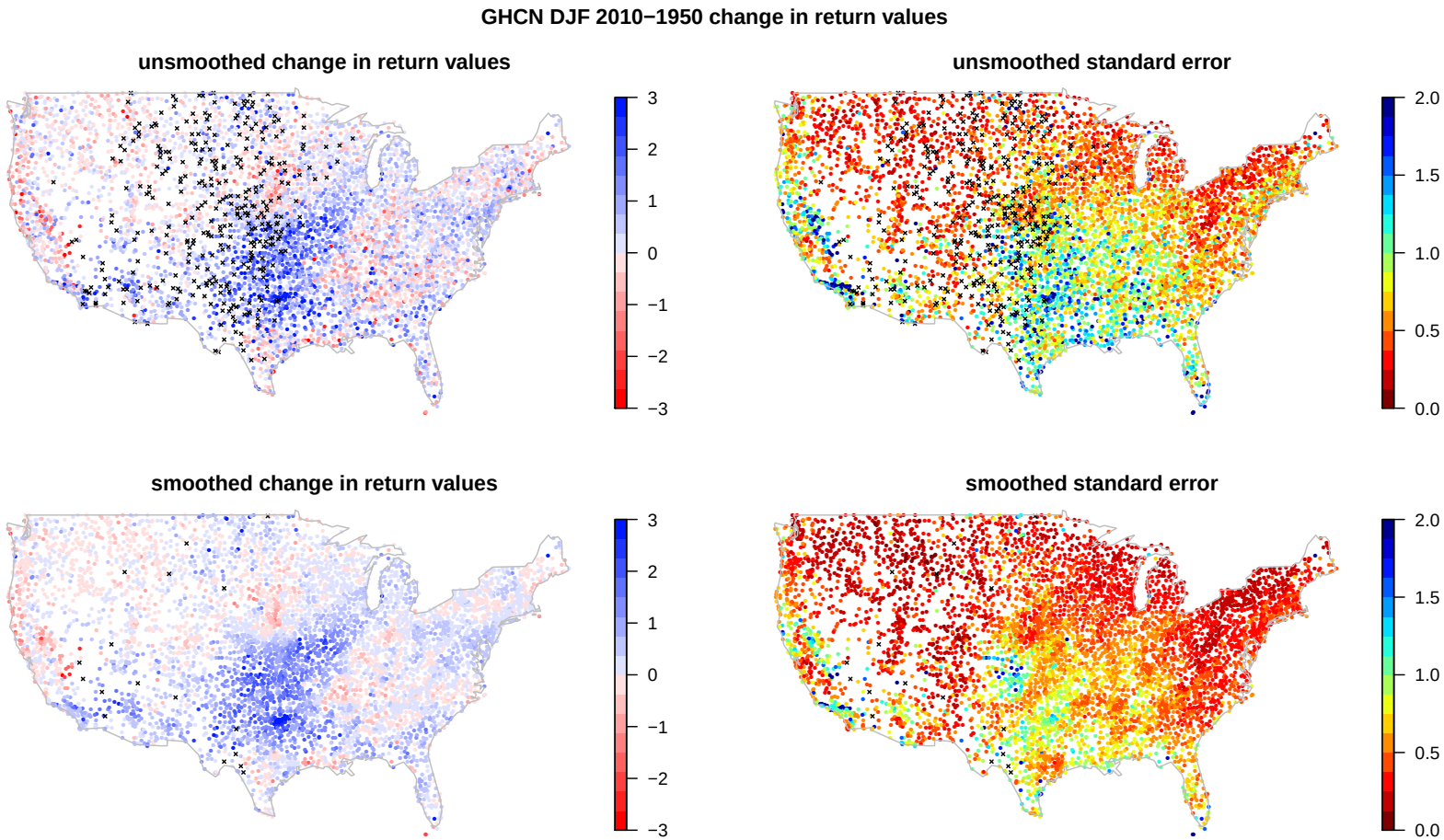
- Advantages:
  - Avoids asymptotic assumptions/approximations
  - Embarrassingly parallel
  - Rather than reducing temporal autocorrelation by choosing the maximum daily precipitation within blocks of days or runs of extreme precipitation, the bootstrap can directly account for this:
    - Basic approach is to bootstrap in year-long blocks
    - Inclusion of teleconnections as covariates and stratifying by season further reduce temporal autocorrelation
- Disadvantages
  - Computationally-intensive
  - Optimization can fail for some of the bootstrap resamples; unclear how to address this in a formal statistical fashion

Spatial smoothing: Local likelihood

- Ramesh & Davison (2000; JRSSB) suggest to use local likelihood to smooth in time
- Here we propose to use local likelihood to smooth in space, using cross-validation of the log-predictive density to choose the bandwidth
  - Fit for each location, borrowing strength in a neighborhood
  - Normal density smoothing kernel, truncated at 2σ or 3σ to reduce computation
  - Common threshold for analysis at each location based on quantile of focal location
  - Common parameter values, but one might consider locally linear parameters
- Bootstrap can again provide uncertainty estimates, accounting for the spatial dependence by resampling the same blocks (i.e., years) at each location

Effects of spatial smoothing

Change in US DJF return levels over 1950 to 2010

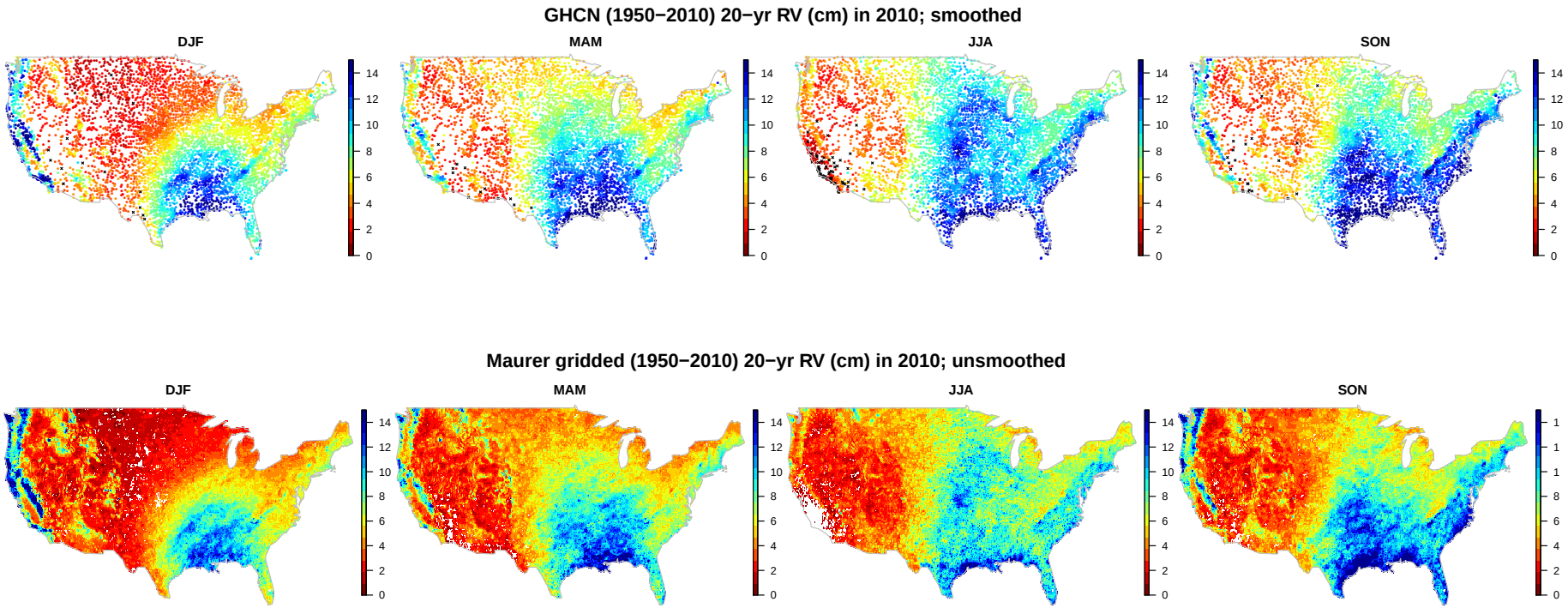


Analysis details

- Data sources:
  - US station-specific analyses of Global Historical Climatology Network-Daily (GHCND), 1950-2010
  - US gridded (1/8 deg.) observational data from Maurer et al. (2002), with elevation correction, 1950-2010
  - Global CCSM4, run #1 from the CMIP5 archive, 1950-2005 (more runs and models to be analyzed)
- Details (analyses still in progress):
  - Threshold: 95th percentile of daily precipitation greater than 1 mm, location-specific
  - GHCND data fit using local likelihood with Gaussian kernels, $\sigma = 25$km, truncated at 50km
  - Maurer and CCSM4 data fit by grid point without smoothing
  - Bootstrap-based uncertainty

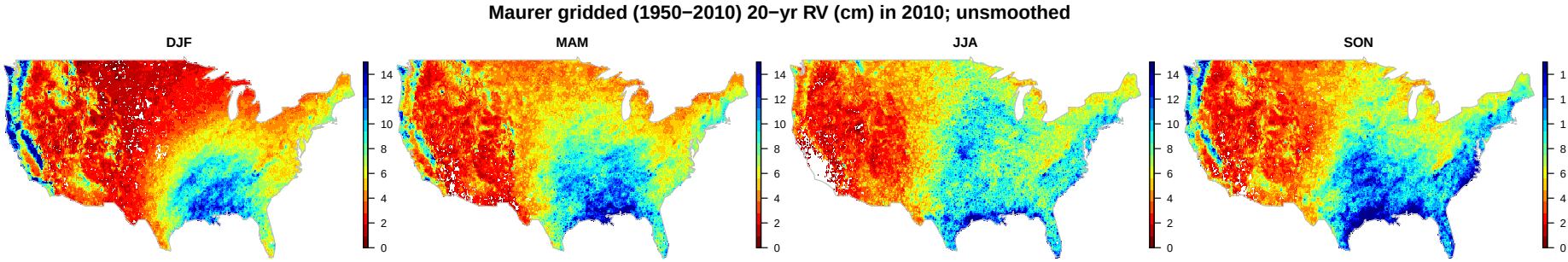
Climatology of US extreme precipitation

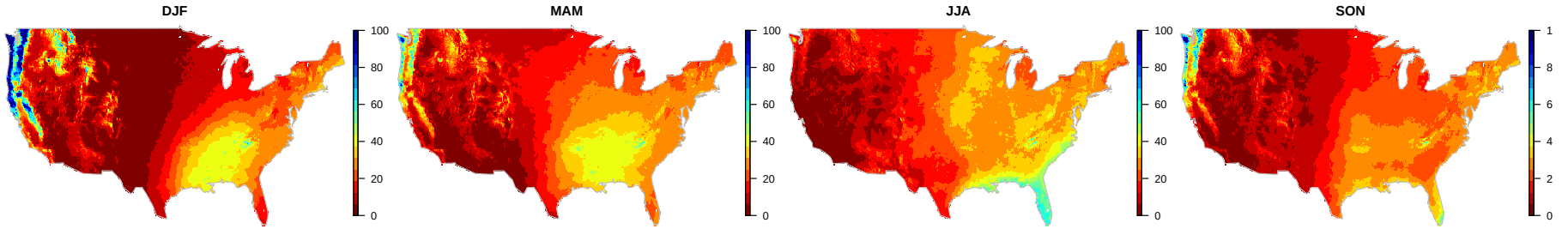
20-year return levels, 2010



- The Maurer gridded product mostly retains the point-based features.
- Summer precipitation in the eastern US is the exception.
- Note the effect of the Appalachians in summer and fall.

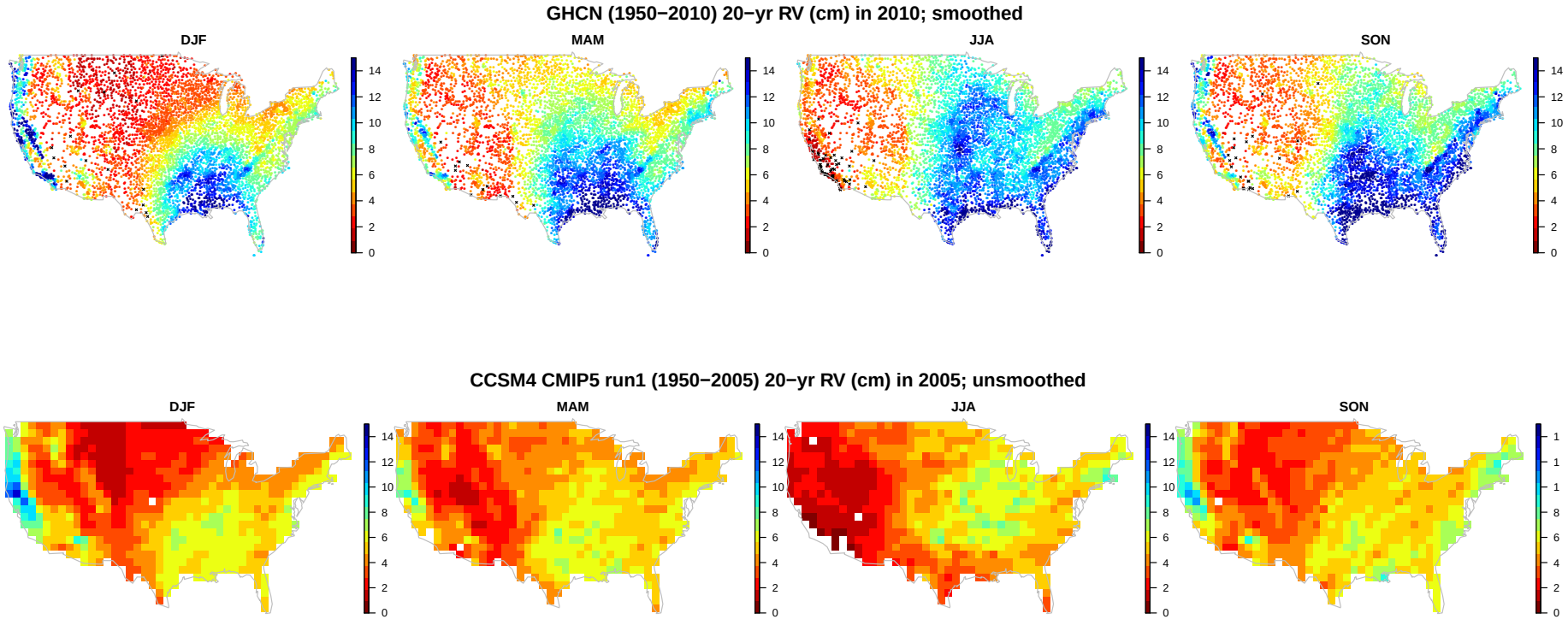
Climatology: Extremes vs. Means





CCSM4 climatology fidelity

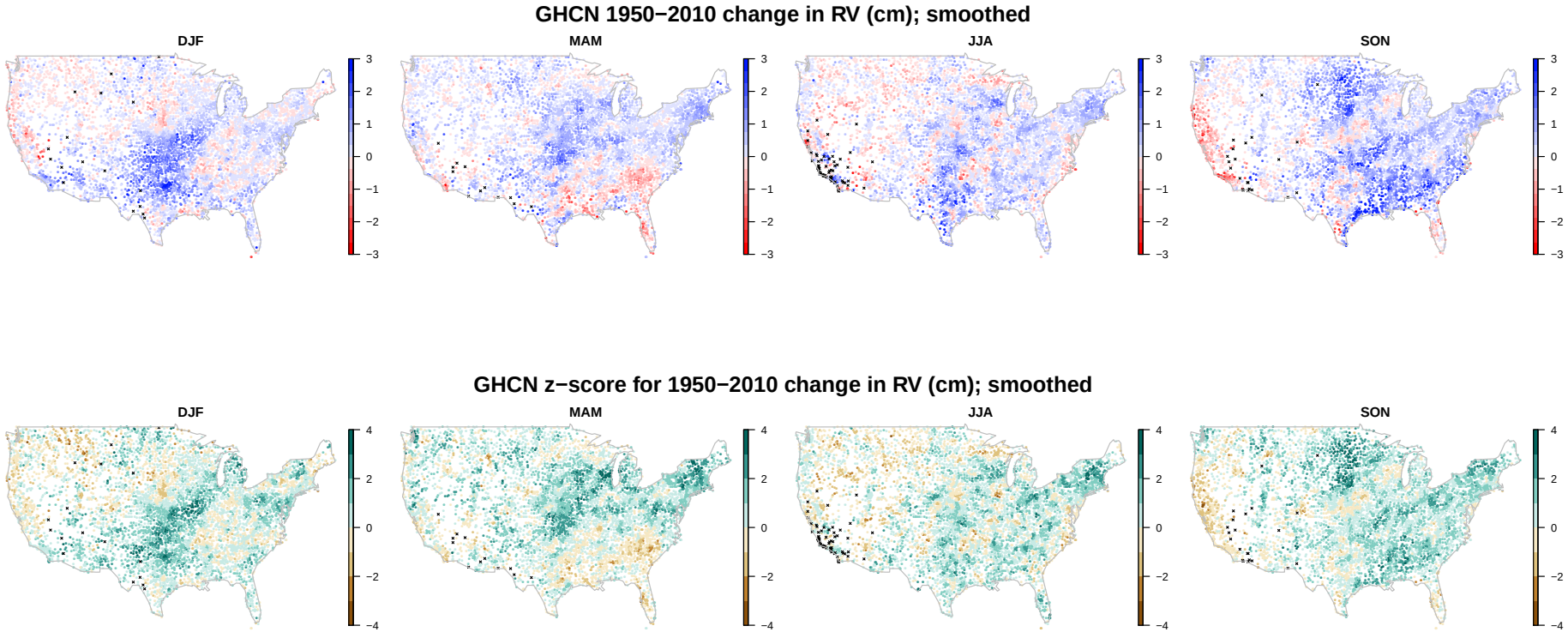
20-year return levels, 2005/2010



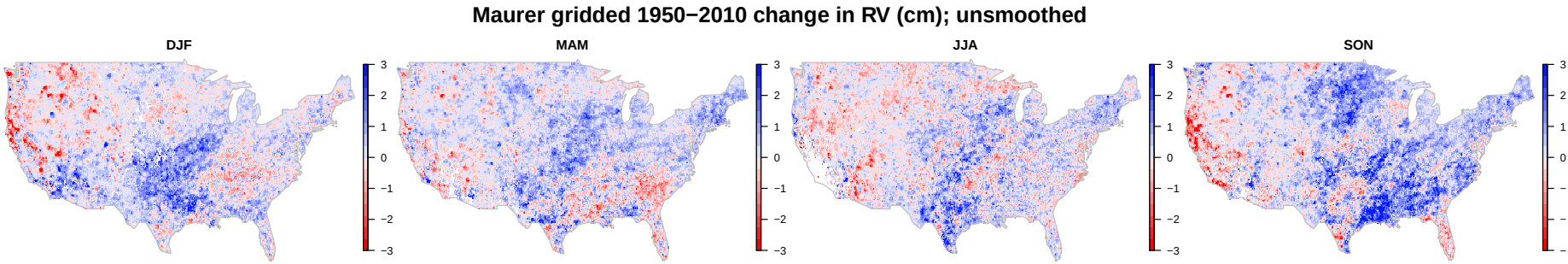
- Patterns in CCSM4 are reasonable, but magnitude in wetter areas, particularly in the summer, is far too low.

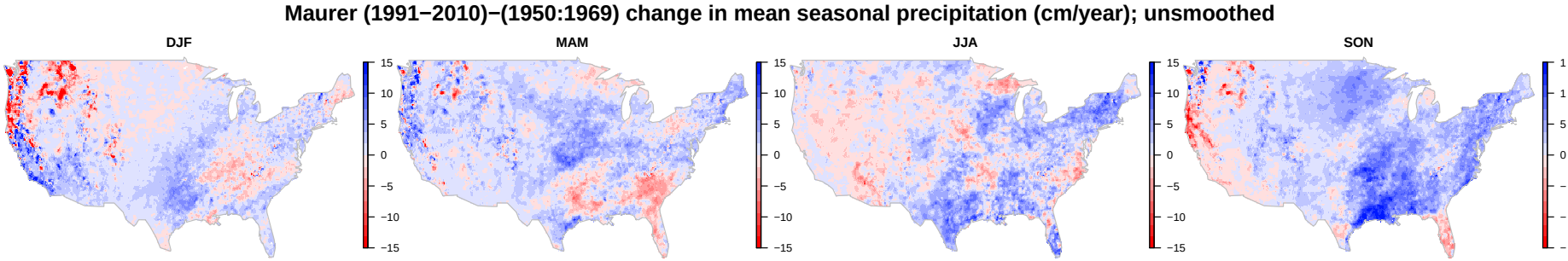
Trends in US extreme precipitation, 1950-2010

2010 return levels minus 1950 return levels



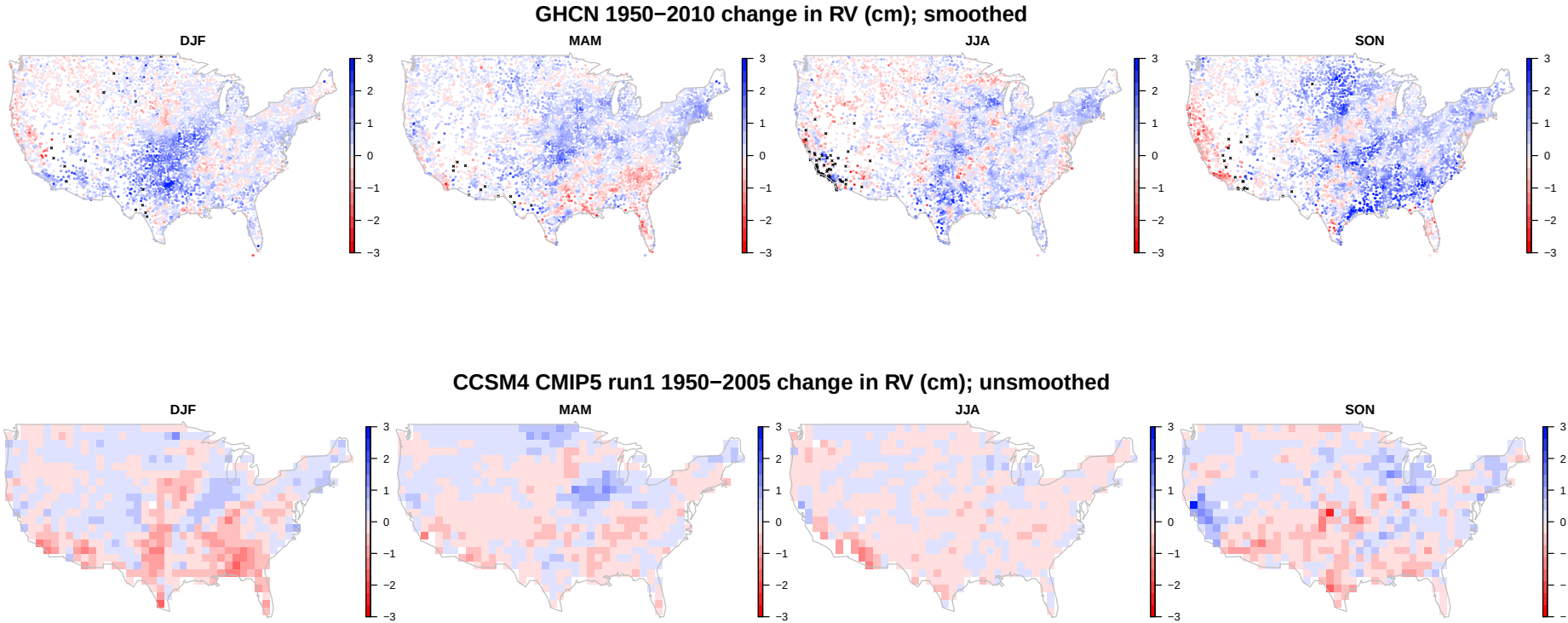
Trends: Extremes vs. Means





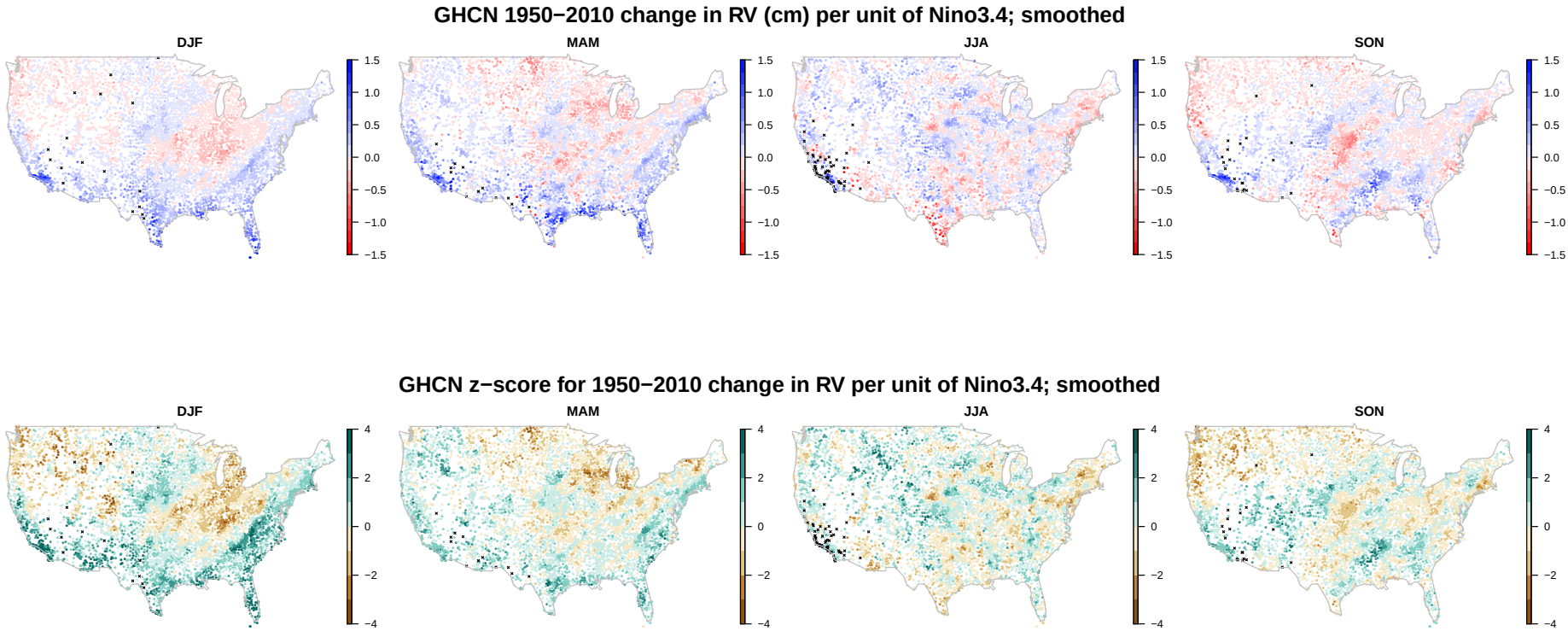
CCSM4 trends fidelity

2010/2005 return levels minus 1950 return levels



Association with ENSO

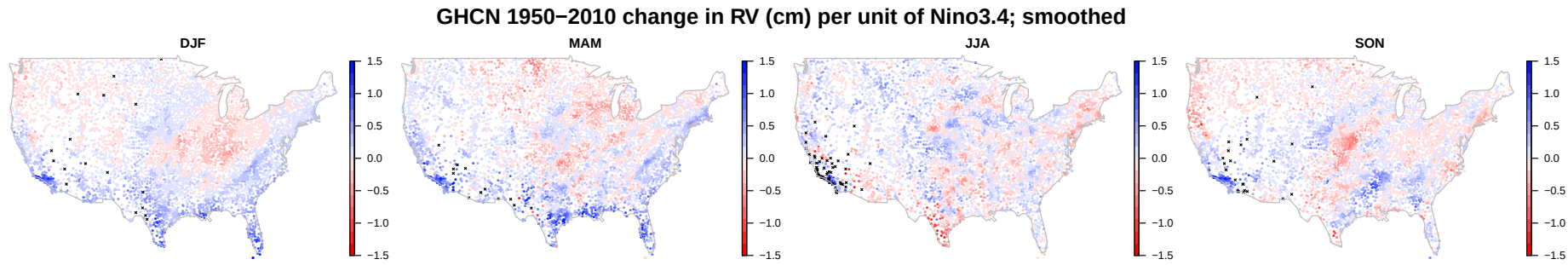
Effect of one-unit change in Nino3.4 on return levels

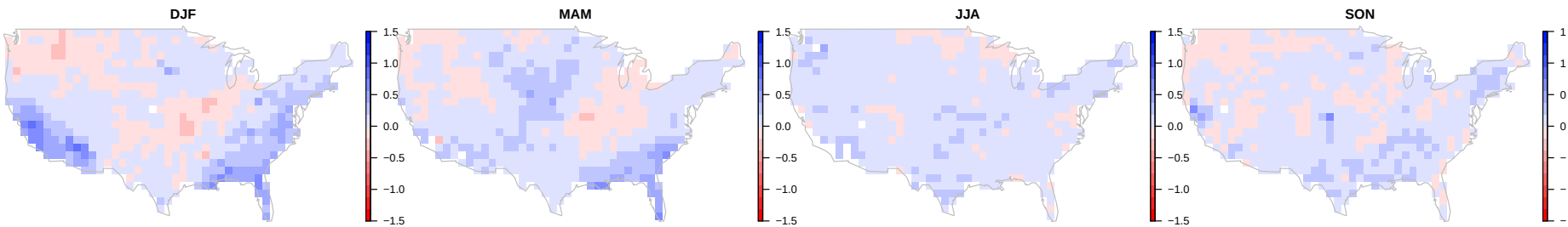


Blue indicates a positive association with El Nino conditions and red a positive association with La Nina conditions.

CCSM4 ENSO association fidelity

Effect of one-unit change in Nino3.4 on return levels





Note the fidelity for winter, which is the season with the most robust estimates in the observational data.

Open questions

- How should we display and assess joint uncertainty? Note the small-scale, but 'significant' patterns in many results
- Is field significance at all useful given that it says nothing about which patterns are robust?
- Would the False Discovery Rate approach be helpful for assessing the collection of z-scores, particularly given the larger bootstrap standard errors?

R software development

- In collaboration with Eric Gilleland (NCAR), I have built the following capabilities either into the *extRemes* R package or as add-on functionality in the *Ilex* R package under development.
  - Handling missing values in point process modeling (common in observational data), assuming MAR missingness
  - Fitting point process models given only the exceedances; this greatly speeds computation
  - Including delta-method-based uncertainty for return values and differences in return values in nonstationary models
  - Including block bootstrap capability
  - Allowing local likelihood fitting

Parallel software deployment

- In collaboration with Dave Pugmire at ORNL and Hari Krishnan at LBNL, we are developing tools within the VisIt parallel visualization software (developed at the national labs) to use extreme value analysis for gridded climate data.
- Goal is to allow analysis in VisIt (and also in the new UV-CDAT software) by calling R functionality, with VisIt handling parallel I/O, initializing processes, collecting results, and visualization and R handling the statistical model fitting.
- Two core tools:
  - VisIt operators for parallelized GEV and POT analysis with linear time trends in location, scale, shape parameters, as desired. Ensemble analysis is possible.
  - General purpose VisIt capability to run Python and R scripts. This will allow for general extreme value analyses using R, including arbitrary covariates, spatial smoothing, etc.