Hidden variable approach to precipitation modeling: how to deal with lots of 0's in your data

Oleg V. Makhnin and Devon McAllister
Mathematics Department, New Mexico Tech, Socorro, NM
olegm@nmt.edu

Objectives:

* understand spatio-temporal properties of precipitation
* unify "mean" calculations and probability calculations for quantities that frequently equal 0

Idea:

Gauge precipitation R is related to a *hidden* variable W so that [1]:

$$\begin{cases} W = R^{1/\beta}, & R > 0 \\ W < 0, & R = 0 \end{cases}$$

Thus, W is power-transformed and assumed to be *negative* when the observed precipitation is 0.
Practically, $W < 0$ is treated as *missing* and is imputed in our model fit.

The transformation and imputation of W are done to insure W has *normal* distribution (we have used $\beta = 2$). Normality enables us to use traditional techniques like kriging.

W will be called *precipitation potential*.

Model

The W values are fitted into a spatio-temporal model, which is a combination of purely temporal (autoregressive) for *region-wide* precipitation potential θ_t, and spatial (range/nugget) model for gauges

i = Station, t = Day,
d = Day of year (1,...,365) corresp. to t

Level 1 $\quad W_{it} = Z_{it} + \nu_{it}$
$\quad$ *"nugget"* $\nu_{it} \sim Normal(0, \tau^2)$

Level 2 $\quad Z_{it} = \theta_t + \epsilon_{it}$
$\quad$ *vector* $\boldsymbol{\epsilon}_t \sim Normal(0, \sigma_\epsilon^2 \mathbf{V})$

Level 3 $\quad \theta_{t+1} = \mu_{d+1} + r(\theta_t - \mu_d) + \xi_t$
$\quad$ *autoregression* $\xi_t \sim Normal(0, \sigma_\xi^2)$

Level 4 $\quad \mu_d = \sum a_k \cos\left(\frac{2\pi dk}{365}\right) + \sum b_k \sin\left(\frac{2\pi dk}{365}\right)$
$\quad$ *seasonal via Fourier series*

where:

* **V** is covariance matrix corresponding to exponential model with range ϕ:
$$V_{ij} = exp(-Dist_{ij}/\phi)$$
* ν_{it} is the nugget (white-noise) variation, and ϵ_{it} is the spatially correlated variation in precip. potential.

* The model distinguishes between "season normal" μ_d mean potential, and year-specific potential θ_t.

Estimation: Bayesian approach

Computation using *Gibbs sampler* requires finding *full conditional posteriors* (FCP)
Let $(X_1, X_2, ..., X_n)$ be the vector of all unknown quantities and parameters.

Markov Chain Monte-Carlo algorithm through Gibbs sampling:

1. Draw a sample from FCP of one X_j given all other parameters and the data:
$$p(X_j \mid X_1, ..., X_{j-1}, X_{j+1}, ..., X_n, \texttt{data})$$
2. Cycle repeatedly for $j = 1, ..., n$

The algorithm obtains (correlated) samples from parameters of interest.

For example, FCP for W_{it} is just $R_{it}^{1/\beta}$ when $R_{it} > 0$, and is truncated Normal(mean = Z_{it}, st.dev.= τ) when $R_{it} = 0$.

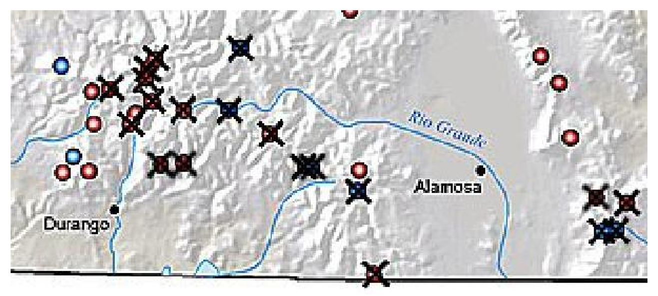
Use for prediction: to predict probability of precipitation at site j, time t, generate samples of W_{jt} and count the proportion of samples with $W_{jt} > 0$.

Benefits: accounts for all sources of incertainty, allows multi-level models, expandable/modular code

Challenges: computationally intensive

Results

Study Area: southwestern Colorado, 20 SNOTEL gauges
Data: daily precipitation for 16 years (1990-2006)

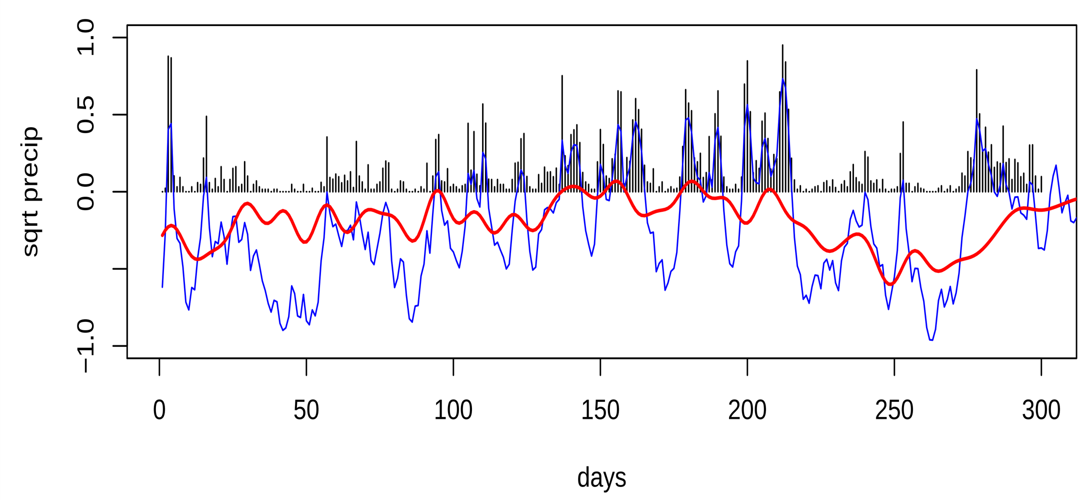


Provides a tough test for our model since only 35% of observations are non-zero.

Estimates (percentiles) of model parameters:

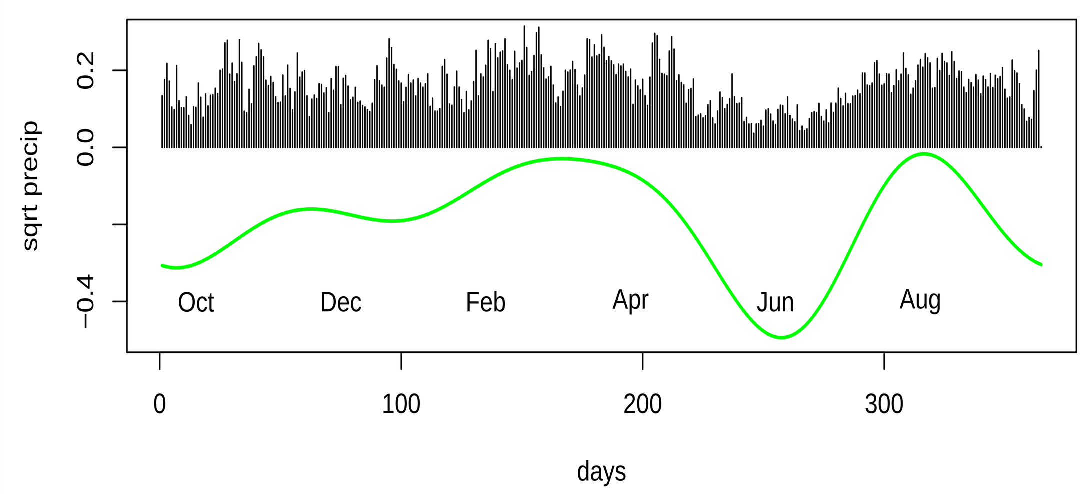
	2.5%	25%	50%	75%	97.5%
τ	0.335	0.338	**0.339**	0.340	0.343
σ_ϵ	0.290	0.294	**0.297**	0.299	0.304
σ_ξ	0.214	0.220	**0.223**	0.226	0.232
ϕ (range)	81.0	85.6	**87.9**	89.9	96.6
r (autocorr.)	0.761	0.776	**0.782**	0.789	0.803

Fitted values of θ_t (blue) and seasonal μ_d (red), first 300 days; against average transformed precipitation sqrt(R) (black lines).



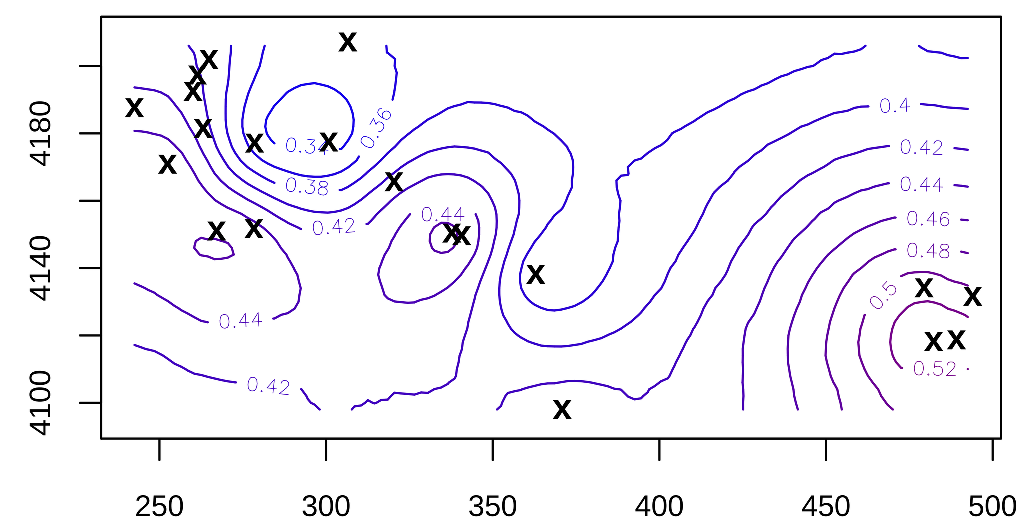
During wet periods, θ_t tracks sqrt(R). During dry periods (precipitation probability $< 50\%$), however, θ_t becomes negative.

Fitted values of seasonal potential μ_d (green) against average transformed precipitation sqrt(R).

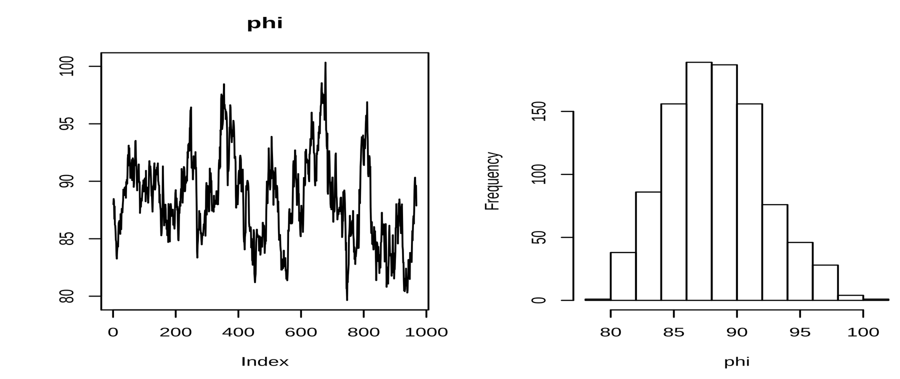


This plot illustrates the main difference between simply averaged precipitation, and the use of hidden model. In a dry season (June), the precipitation potential is large negative, which results in very rare actual precipitation. Much bigger discrepancy than the one seen in sqrt(R) (black lines)!

Estimated probability of precipitation (via Bayesian kriging), day 95



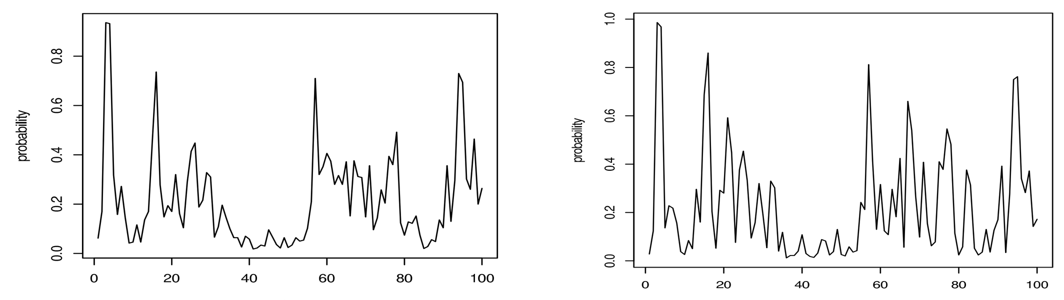
Markov Chain output: ϕ (range parameter, km)

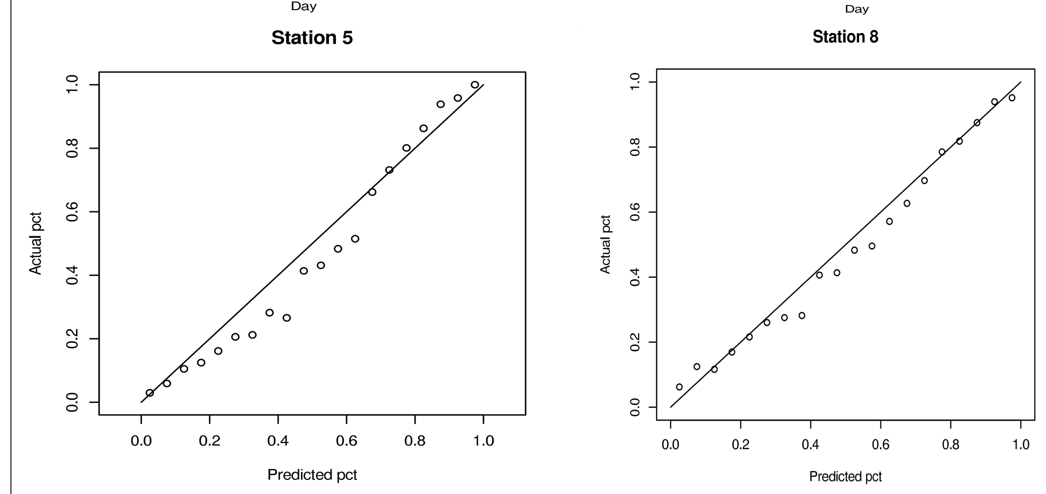


Cross-validation

Predicted precipitation probabilities (first 100 days):

Station 5 $\qquad$ Station 8





Predicted vs. Actual precipitation frequency.

Conclusions

* Simply computing average precipitation, although convenient, may not be the best choice for spatio-temporal models.
* No need for separate models: one for the occurrence and one for the observed amount.

Future work

* Larger region/ use of multivariate θ_t
* Incorporate radar & satellite data

References

1. Sansó, B. and Guenni, L. (2000) *A non-stationary multisite model for rainfall* J. of Amer. Statistical Association, **95**, pp.1089-1100