Modeling and Simulating Rainfall

Kenneth Shirley, Daniel Osgood, Andrew Robertson, Paul Block, Upmanu Lall, James Hansen, Sergey Kirshner, Vincent Moron, Michael Norton, Amor Ines, Calum Turvey, Tufa Dinku

October 7, 2008
Motivation

Rainfall Models

A Model Checking Example
The Goal of Index Insurance

Index Insurance reduces the variance of a farmer’s income.

(It also reduces the expected value of a farmer’s income, but this tradeoff is worthwhile for long-term economic growth).

Without Insurance:

\[ \text{Income}_i = \text{Yield}_i. \]

With Insurance:

\[ \text{Income}_i = \text{Yield}_i + \text{Payout}_i - \text{Price}_i, \]

for the \( i^{\text{th}} \) period, which is usually a growing season.
Everything is a function of rainfall

\[ \text{Income}_i = \text{Yield}_i + \text{Payout}_i - \text{Price}_i \]

- The yield is a random variable, a positive function of rainfall (up to a point), and is estimated using crop models: \( \text{Yield}_i = f(\text{Rain}_i) + \epsilon_i \).
- The payout is a random variable, and is a negative function of rainfall which the contract designer must choose. It could be linear, for example: \( \text{Payout}_i = a \times \text{Rain}_i + b \), where \( a < 0 \).
- The price is a constant that depends on the rainfall distribution, and consists of an average payout plus a premium. For example, \( \text{Price}_i = \mathbb{E}(\text{Payout}_i) + 0.06 \times (Q(\text{Payout}_i, 0.99) - \mathbb{E}(\text{Payout}_i)) \).
In a perfect world...

If we knew the distribution of Rain$ _i $ and the exact relationship Yield$ _i = f(Rain_i) + \epsilon_i $, then we could optimize the parameters of different payout functions to find the income mean-variance tradeoff that is most attractive to farmers.
In practice...

We use existing crop models to estimate the relationship \( \text{Yield}_i = f(\text{Rain}_i) + \epsilon_i \) for a given location and crop. These models often require \textit{daily} rainfall as an input variable.

We estimate the distribution of \( \text{Rain}_i \) and related statistics that pertain to the design of the payout and price function, which, in the example we’ll present later, are functions of \textit{dekadal} rainfall.

- The payout function we use is a piecewise linear function of one or more dekadadal rainfall sums.
A payout function example

Figure: (a) An example of a piecewise linear payout function, and (b) an example of the distribution of payouts.
Why use a model?

\( N \) years of rainfall data produce \( N \) yearly payouts, but \( N \times 365 \) daily rainfall observations.

- We gain statistical power by using more data. Can be more challenging to model, but potentially worthwhile because...
- Some aspects of the empirical payout distribution are sensitive to small changes in contract parameters - i.e. if the trigger is moved a little, the number of payouts that would have occurred would change a lot.
Types of models for daily rainfall

We describe 4 basic types:

1. Generalized Linear Models (GLMs)
2. Hidden Markov Models (HMMs)
3. Models using nonparametric and resampling methods
4. “Mechanistic Models”

Review articles Wilks & Wilby (1999) discuss the first 3 types, and Chandler et al. (2007) describe the first and fourth types.
Modeling Issues

- Can the model condition on seasonal forecasts and other exogenous variables?
- How is spatial correlation modeled?
- How is temporal correlation modeled?
Motivation

Rainfall Models

A Model Checking Example

GLMs

- GLMs for rainfall data were first described by Coe and Stern (1982) and Stern and Coe (1984).
- Usually consist of two components: (1) A two-state Markov chain for the rainfall occurrence process, and (2) a gamma or mixed exponential distribution for the amounts process.
- Can be adapted for multiple stations and the incorporation of covariates, like SST, seasonal forecasts, and other large-scale climate variables.
- Flexibility to incorporate long-term climate variables is an important advantage.
HMMs

- HMMs model spatial and temporal correlation using hidden states which correspond to regional weather regimes (Hughes and Guttrop, 1994; Hughes et al., 1999; Bellone, 2000; Robertson et al., 2004, 2006).

- A non-homogeneous HMM can incorporate climatological variables to allow for seasonality.

- Certain modeling choices, like how many hidden states to use, make HMMs slightly more challenging to fit than other models, but their physical interpretation is a worthwhile advantage.
Nonparametric and Resampling Methods

- Simulate rainfall by resampling from the observed data in a way that preserves spatial and temporal correlations (Young, 1994; Lall et al., 1996; Rajagopalan & Lall, 1999; Moron et al., 2008).
- Flexibility allows for better description of nonlinear relationships between variables and marginal distributions that don’t fit parametric forms.
- Incorporation of covariates is more difficult, as is interpretation.
“Mechanistic Models”

- Models that attempt to represent the physical process of a storm, in which a very large, moving ‘rain event’ occurs randomly in time and space, and within it, smaller “storms” arise randomly, each of which consists of “rain cells” which also occur randomly (Chandler, et. al 2007).

- These models require high resolution radar data, are not yet adapted to incorporate covariates, and are presently used more in flooding applications.
The Data (1)

Figure: Daily Rainfall, Catacamas, Honduras, 1960-1964.
The Data (2)

**Figure:** Boxplots of dekadal frequency, average intensity, and total rainfall in Catacamas, 1960 - 2005.
Wet Day Indicator $X = \{X_{ktd}\} = 1\{Y_{ktd} > 0\}$, the indicator of a wet day on day $d$ in year $t$ of dekad $k$.

Rainfall $Y = \{Y_{ktd}\}$ denote the amount of rainfall on day $d$ in year $t$ of dekad $k$.

We fit a model whose likelihood can be factored into two parts: the likelihood of the frequency of wet days, $X$, and the likelihood of the intensity of rainfall on wet days, $Y$:

$$P(X, Y | \theta) = P(X | \theta_1) \cdot P(Y | \theta_2).$$
The model for rainfall intensity ($Y$)

$$Y_{ktd} \mid X_{ktd} = 1 \sim \text{Gamma}(\alpha_{kt}, \beta_{kt}),$$
The model for rainfall intensity ($Y$)

\[ Y_{ktd} \mid X_{ktd} = 1 \sim \text{Gamma}(\alpha_{kt}, \beta_{kt}), \]

\[ \begin{pmatrix} \log(\alpha_{kt}) \\ \log(\beta_{kt}) \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_k \\ \Sigma_k \end{pmatrix} \right), \]
The model for rainfall intensity ($Y$)

$$Y_{ktd} \mid X_{ktd} = 1 \sim \text{Gamma}(\alpha_{kt}, \beta_{kt}),$$

$$\begin{pmatrix} \log(\alpha_{kt}) \\ \log(\beta_{kt}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_k \\ \Sigma_k \end{pmatrix}\right),$$

$$p(\mu_k, \Sigma_k) \propto |\Sigma_k|^{-\frac{3}{2}},$$

for $t = 1, \ldots, T$, and $d = 1, \ldots, n_d$. 
The model for rainfall frequency \( (X) \)

\[
X_{ktd} \mid X_{kt(d-1)} \sim P_{kt} = \begin{pmatrix}
p_{1kt} & 1 - p_{1kt} \\
1 - p_{2kt} & p_{2kt}
\end{pmatrix}
\]
The model for rainfall frequency ($X$)

$$X_{ktd} \mid X_{kt(d-1)} \sim P_{kt} = \begin{pmatrix} p_{1kt} & 1 - p_{1kt} \\ 1 - p_{2kt} & p_{2kt} \end{pmatrix}$$

$$p_{jkt} \sim \text{Beta}(\mu_{jk}, s_{jk})$$
The model for rainfall frequency ($X$)

\[ X_{ktd} \mid X_{kt(d-1)} \sim P_{kt} = \begin{pmatrix} p_{1kt} & 1 - p_{1kt} \\ 1 - p_{2kt} & p_{2kt} \end{pmatrix} \]

\[ p_{jkt} \sim \text{Beta}(\mu_{jk}, s_{jk}) \]

\[ \mu_{jk} \sim \text{U}(0, 1) \]
\[ s_{jk} \sim \text{U}(0, \infty) \]

for $d = 2, \ldots, n_d$, $m = 1, \ldots, 12$, $t = 1, \ldots, T$, and $j = 1, 2$. 
Fitting the model and simulating from it

- Fit the model using whatever methods are preferred.
- When simulating from a fitted model, one must respect two sources of variation: (1) sampling variation built into the model, and (2) variation associated with the uncertainty with which the parameters of the model were estimated.
- A nice feature of proper simulation is that shorter historical records naturally produce relatively less certainty about the rainfall process, and therefore produce more variable simulations. This should result in a higher Value-at-risk, and more expensive insurance.
Checking the fit of a model

Standard model checks include:

- Means and standard deviations of monthly rainfall sums (interannual variability).
- The distribution of the length of wet and dry spells.
Index Insurance related model checks should include:

- The distribution of dekadal sums
- The distribution of capped dekadal sums
- The onset of the growing season (the sowing dekad)
- The distribution of phase sums
- The lower quantiles of phase sums
Checking Dekadal Sums

Figure: The simulated distribution of the sum of rainfall for the first dekad of August in year 10 (1970), with a red vertical line indicating the value observed in the data.
Checking the Sowing Dekad

Figure: The simulated distributions of the percentages of years (out of 46) in which the sowing dekad occurred in the specified dekad, with a red vertical line indicating the value observed in the data.
Checking the Sowing Dekad - More Closely

**Figure:** Posterior predictive checks of the mean MAY 3 rainfall (across all 46 years) and the sd of MAY 3 rainfall. We underestimate the variability with this model.
Checking the Phase Sums

Figure: Comparison of Phase 3 sum distribution, estimated via simulation (green), vs. the empirical distribution (blue dotted).
Thoughts on the phase sum distribution

- The simulated mean is probably too high - many simulated dekads were capped at 60mm, whereas this rarely happened in the observed data.

- The lower tail is shorter than in the data. Variability of dekadal sums might be too low.

- Maybe dekadal sums are OK, but there is correlation between dekads. We modeled dekads within the same year as independent. Positive correlation would increase the variance of their sum, and negative correlation would decrease the variance of their sum.

- The exit is lower than it needs to be.
Conclusions

- The payout function and price are complicated functions of rainfall.
- It is a challenge to model rainfall, accounting for spatial and temporal correlations, and exogenous variables. (GLMs, HMMs, etc.)
- It is also a challenge to check the fit of a model - virtually nothing is automatic.
- As with any quantitative modeling, one must compromise between trusting the data and trusting a model; is it random chance or something real?