Rainfall Simulator Technical Appendix

When pricing and design analysis is primarily based on historical burn analysis, products and prices are sensitive to particular features of one or two historical events, making it possible to overemphasize the importance of the specifics of these events. However, it is important to remember that most rainfall simulators have limitations in what they can accurately reflect, and unless utilized with caution, may lead to an inaccurate understanding of the performance of a contract. An overview of issues with rainfall simulators in index insurance is provided by Shirley 2008.

When feasible, daily simulators are preferable. However, because a dekadal simulator has the less challenging task of simulating a ten day sum of rainfall, they have some advantages. They run much more quickly, are relatively accurate for modeling ten day sums, and do not require the manual supervision by a trained operator that high quality daily simulators need.

The dekadal rainfall simulator described below is a robust simulator that can be reliably run unsupervised, is not processor intensive, and can therefore be run in a loop under software control. It has the useful feature that statistical uncertainty due to short data series is represented in the simulations but does not have mechanisms to estimate and utilize links between datasets. However, the simulator as written may produce unrealistically large dekadal sums. We are currently addressing this problem by capping these rare simulated events.

The rainfall simulator presented below is available for use as code we have developed in the freely available R statistical package. It also has been implemented in graphical user interface through the online Weather Index Insurance Educational Tool developed by IRI in partnership with the World Bank CRMG.

Designing a Dekadal Rainfall Simulator

To illustrate the potential benefits and challenges in applying rainfall simulation, we have been working on a simple rainfall simulator. This simulator represents many of the strategies utilized by industry standard products. In this section, we present the model for dekadal rainfall, we fit it and evaluate the fit for two example data sets: Adiha (Ethiopia), and Lilongwe (Malawi), where these data sets consist of 7 and 44 years of observed dekadal data, respectively. The ultimate goal is to use the model to provide realistic dekadal rainfall simulations for any site in the world.

To be useful, we want the model to have the following qualities:

1. It must not take a long time to fit the model to data. Ideally we would be able to simulate a few thousand years of data within 1-2 seconds. This restricts the class of models we can consider.
2. It will be parametric, so that we can simulate values that did not occur in the observed data, which is impossible using certain resampling strategies.
3. It will be sensitive to the amount of data with which it is fit, so that less observed data results in more uncertainty in simulated rainfall. To do this conveniently, we use the Bayesian framework.
4. Lastly, we want the procedure to be automatic, so that as little supervision and data set-up as possible is required. This might be the greatest challenge of all, because different amounts of observed data (5 years vs. 50 years, for example) and different rainfall-generating processes throughout the world may require different types of models.

Here is some general notation, and the basic outline of how we will fit and simulate rainfall from our models. Let $Y_{dt}$ be the total amount of rainfall during dekad $d$ of year $t$, for dekads $d = 1,\ldots,36,$
and years 1,..,T. We will fit a parametric model, \( p(\theta | Y) \), where \( Y \) denotes the vector of observed values of \( Y_{dt} \) for a given site and time span. In Bayesian statistics, the statistical distribution assumed (the prior distribution) is updated using the data to arrive at an updated (posterior) distribution. Given our assumed prior distribution for \( \theta \) (which requires careful construction) we will sample values of \( \theta \) from its posterior distribution, \( p(\theta | Y) \). The posterior sample of \( G \) values is denoted \( (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}) \), where \( G=1000 \) is typical. To simulate new dekadal rainfall observations, we do the following:

1. Draw a value of \( \theta^{(g)} \) from its posterior distribution.
2. Draw a value of \( Y_{\text{prod}}^{(g)} | \theta^{(g)} \) where \( Y_{\text{prod}} \) denotes a new, simulated realization of \( Y \).

Finally, we compare the distribution of \( Y_{\text{prod}} \) to the observed values of \( Y \), to make sure that the simulations are similar in character to the observations. We also check to see that we don't overfit the model by performing a cross-validation exercise. This Bayesian framework is a convenient way to simulate rainfall in a way that reflects not just the natural uncertainty of the rainfall process, but also the uncertainty associated with our parameter estimation.

**The Basic Model and Adiha Data**

Let \( Y_{dt} \) be the total amount of rainfall during dekad \( d \) of year \( t \), and let \( X_{dt} \) be the indicator of whether there was any rainfall during the specified dekad and year, such that \( X_{dt} = 1 \{Y_{dt} > 0\} \).

The model we fit has two components: a model for the frequency of rainfall, and, when rainfall occurs, a model for the intensity of rainfall.

The frequency model is

\[
X_{dt} \sim \text{Bernoulli}(p_d),
\]

for \( t = 1, \ldots, T \) and \( d = 1, \ldots, 36 \), and the intensity model is

\[
Y_{dt} | (X_{dt} = 1) \sim \text{Gamma}(\alpha_d, \beta_d),
\]

where \( \alpha_d \) and \( \beta_d \) are the shape and rate parameters of the gamma distribution with mean \( (\alpha_d / \beta_d) \). The notation \( Y_{dt} | (X_{dt} = 1) \) is a reminder that this is the model for intensity of dekadal rainfall conditional on the fact that the rainfall amount is non-zero for that dekad. The full distribution of dekadal rainfall depends on \( \alpha_d, \beta_d, \) and \( p_d \).

The probability density function of a single observation can therefore be written:
\[
\Pr(Y_{dt} = y_{dt}) = \begin{cases} 
1 - p_d & \text{if } y_{dt} = 0 \\
\frac{\beta^a_{t_d}}{T(\alpha_d)} \left(\frac{\beta^a_{t_d}}{T(\alpha_d)}\right)^{y_{dt}-1} \exp(-\beta^a_{t_d} y_{dt}) & \text{if } y_{dt} > 0
\end{cases}
\]

Since the occurrence of rainfall in a given dekad is independent across years, the frequency model can be simplified in the following way: Let \(Z_d = \sum_{t=1}^{T} X_{dt}\), the number of years out of \(T\) in which a non-zero amount of rainfall occurred. Then

\[Z_d \sim \text{Binomial}(T, p_d),\]

for \(d = 1, \ldots, 36\).

**Example: Adiha Cup-on-a-stick**

The data for the cup-on-a-stick site in Adiha consist of dekadal sums for each of the 36 dekads in the year for the years 2000 - 2007, with the exception of 2004 (whose data were missing); hence for this data, \(T = 7\).
Table 1 Annual Rainfall data from manual raingauge at Adi Ha

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>14.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>14.00</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>2.00</td>
<td>0.00</td>
<td>7.00</td>
<td>0.00</td>
<td>0.00</td>
<td>4.5</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>25.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>20.00</td>
<td>0.00</td>
<td>22.5</td>
<td>3.5</td>
</tr>
<tr>
<td>11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.00</td>
<td>7.00</td>
<td>0.00</td>
<td>29.7</td>
<td>35.7</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6.00</td>
<td>22.5</td>
<td>15.0</td>
<td>0.00</td>
<td>14.5</td>
<td>8.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>35.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>41.0</td>
<td>8.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.07</td>
<td>0.00</td>
<td>0.00</td>
<td>36.00</td>
<td>0.00</td>
<td>18.5</td>
<td>24.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>24.30</td>
<td>14.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.4</td>
<td>11.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.00</td>
<td>46.5</td>
<td>11.0</td>
<td>0.00</td>
<td>5.0</td>
<td>19.0</td>
<td>19.5</td>
<td>16.0</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18.00</td>
<td>41.0</td>
<td>61.5</td>
<td>81.0</td>
<td>33.5</td>
<td>43.0</td>
<td>89.5</td>
<td>52.5</td>
<td>25.9</td>
</tr>
<tr>
<td>19</td>
<td>45.00</td>
<td>92.0</td>
<td>60.5</td>
<td>14.5</td>
<td>72.0</td>
<td>46.5</td>
<td>52.0</td>
<td>54.6</td>
<td>24.2</td>
</tr>
<tr>
<td>20</td>
<td>56.00</td>
<td>28.0</td>
<td>45.5</td>
<td>122.5</td>
<td>32.0</td>
<td>104.0</td>
<td>59.6</td>
<td>38.4</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>64.50</td>
<td>107.5</td>
<td>82.0</td>
<td>100.0</td>
<td>57.0</td>
<td>129.5</td>
<td>88.8</td>
<td>25.3</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>65.00</td>
<td>6.0</td>
<td>58.5</td>
<td>51.0</td>
<td>76.5</td>
<td>67.0</td>
<td>103.0</td>
<td>61.0</td>
<td>29.4</td>
</tr>
<tr>
<td>23</td>
<td>57.50</td>
<td>0.0</td>
<td>72.5</td>
<td>81.5</td>
<td>43.5</td>
<td>59.0</td>
<td>94.5</td>
<td>68.1</td>
<td>18.4</td>
</tr>
<tr>
<td>24</td>
<td>42.00</td>
<td>75.0</td>
<td>26.0</td>
<td>74.5</td>
<td>101.5</td>
<td>39.5</td>
<td>70.1</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>31.00</td>
<td>10.0</td>
<td>40.0</td>
<td>41.0</td>
<td>0.0</td>
<td>44.0</td>
<td>33.2</td>
<td>13.8</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>29.50</td>
<td>0.0</td>
<td>31.0</td>
<td>10.0</td>
<td>55.0</td>
<td>0.0</td>
<td>31.4</td>
<td>18.4</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>4.00</td>
<td>0.0</td>
<td>4.0</td>
<td>27.0</td>
<td>0.0</td>
<td>0.0</td>
<td>11.7</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>6.00</td>
<td>0.0</td>
<td>3.0</td>
<td>8.0</td>
<td>0.0</td>
<td>0.0</td>
<td>5.7</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>17.50</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>17.5</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>38.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>38.0</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>23.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>23.0</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

Let's look at some graphical summaries of the data. First, Figure 1(a) plots the percentage of years in which the rainfall sum was non-zero for each dekad. Second, Figure 1(b) plots the mean amount of rainfall by dekad, for the dekads in which there was non-zero rainfall in at least one of the seven observed years. Lastly, Figure 1(c) plots the standard deviation of rainfall by dekad, this time for only the dekads in which there was nonzero rainfall in at least two of the seven observed years (so that the observed standard deviation could be calculated). In all three plots, it is clear that there is dependence between dekads, due to seasonality, that the current model doesn't include.
Figure 1 Graphical summaries of Adiha cup-on-a-stick data

A simple version of the model
A very simple model would use non-informative priors for the parameters of the binomial and gamma distributions. Ultimately, we would like it if the same priors provided realistic simulations no matter how large the data is. This is difficult, though.

The Frequency Model
For the frequency model, a beta distribution, \( p_d \sim \text{Beta}(1,1) \) for \( d = 1, \ldots, D \), is a conjugate prior for the binomial likelihood, and leads to a beta posterior:

\[
p(p_d | Z_d) \sim \text{Beta}(Z_d + 1, T - Z_d + 1).
\]

This posterior actually leads to appropriate inferences and simulations for small and large data sets, so we use it to model the frequency of rainfall by dekad without modification.

The Intensity Model
In order to fit and simulate from the intensity model quickly, we seek a conjugate prior for the parameters of the gamma distribution so that the posterior distribution is available in closed form. Unfortunately, the gamma distribution has a convenient conjugate prior only when the shape parameter, \( \alpha_d \), is known. In such a case, if we use a gamma prior for the rate parameter, \( \beta_d \), then the posterior distribution of \( \beta_d \) is also a gamma distribution. The questions, then, are how do we choose values for \( \alpha_d \) for each dekad, and what prior distribution do we use for the rate parameter, \( \beta_d \)?

In our first attempt at fitting the intensity model, we assume \( \alpha_d \) is the same for each dekad, and we use the data to estimate an average value. The value of the shape parameter, \( \alpha \), of a gamma distribution determines the relationship between the mean and the standard deviation of the distribution. To be precise, if \( Y_{d_t} \sim \text{Gamma} (\alpha_d, \beta_d) \), then \( \text{Sd}(Y_{d_t}) = \text{Mean}(Y_{d_t}) \sqrt{\alpha_d} \). In order to set a constant value of \( \alpha_d \) for all \( d = 1, \ldots, 36 \), we regress the observed standard
deviations of dekad rainfall against their means, and compute the estimated slope via weighted least squares, where the intercept is fixed at zero. Figure 2 shows this relationship and the weighted least squares regression line for Adiha, which has a slope of 0.44, which equates to a value of $\alpha_d = 1 / 0.44^2 = 5.1$. In other words, on (weighted) average, the standard deviations of dekad rainfall sums by dekad are about 44% as large as the means.

If we use a gamma prior for the rate parameter,

$$\beta_d = \text{Gamma}(\alpha_0, \beta_0),$$

then the posterior distribution of $\beta_d$ is given by

$$p(\beta_d | Y_d) \sim \text{Gamma}(\alpha_0 + Z_d \alpha_d, \beta_0 + \sum_{t} Y_{dt})$$
where $\mathbf{Y}_d$ denotes the vector $(Y_{d1}, \ldots, Y_{dT})$. We choose $(\alpha_0, \beta_0) = (1.5, 1.5)$, so as to be relatively noninformative. For dekads in which no non-zero rainfall sums are observed, this prior results in a rate parameter with a mean of one, and an sd of about 0.8.

To see the results, consider Figure 3. This figure displays samples from the posterior distributions of the gamma distributions from which dekadal rainfall sums are drawn. The procedure is as follows: For posterior simulations $g = 1, \ldots, G$,

1. Draw $p_d^{(g)}$ (for $d = 1, \ldots, 36$) and $\beta_d^{g}$ (for all $d$ such that dekad $d$ had at least one non-zero observation) from their posterior distributions.

2. Draw $X_{dt}^{(g)}$ from the Bernoulli($p_d^{(g)}$) distribution for $d = 1, \ldots, 36$, $t = 1, \ldots, T$; Draw $Y_{dt}^{(g)}$ from the Gamma($\alpha_d$, $\beta_d^{(g)}$) distribution for $t = 1, \ldots, T$, and all $d$ for which there was at least one non-zero observation, where $\alpha_d$ is the fixed value of $\alpha_d$ for all $d$ that was computed using weighted least squares. The new simulated rainfall amount for iteration $g$ is $Y_{\text{pred}(d,t)}^{(g)} = X_{dt}^{(g)} \times Y_{dt}^{(g)}$.

Figure 3 contains the densities of the gamma distributions that correspond to different draws of $\beta_d$ for four dekads of the year.
Figure 3 Densities (gray) of the gamma distributions that correspond to different draws of $Y_{di}$ from their posterior distributions, compared to kernel density estimates (green) of the densities of $Y_{di}$ from the observed values, plotted as green tick marks along the x-axis, for dekads $d=13,17,21,26$.

Comments on Figure 3

- The green kernel density estimates are not smooth enough. For example, for Dekad 17, the kernel density estimate is bimodal, with virtually zero probability mass between 30 and 40 mm. This is why we use a parametric model - because we know that the distribution of dekadal rainfall should be somewhat smooth, and probably unimodal.

- The gamma distributions have long right tails. For Dekad 13, for example, one of the gamma distributions is very flat, and puts probability mass above 150 mm, even though the only two non-zero observations for this dekad were 35 mm and 47 mm. The data can't rule out the possibility that this gamma distribution produced these two data points,
but observations from surrounding dekads suggest that this gamma distribution is not appropriate.

• For Dekad 21, the observed mean rainfall amount is much larger than the observed sd, but setting the shape parameter of the gamma distribution $\alpha_{21} = 5.1$ restricts the standard deviation to be just under half mean. From Figure 3, it looks like there is too much mass at low values, and too long a tail on the right.

Two alternative versions of the model

To find a better-fitting model, we modify the way that we estimate the shape parameter for each dekad. Instead of setting it to a fixed value for each dekad, we will allow it to vary across dekads. Ideally, we would estimate it simultaneously with the rate parameter in a Bayesian way, resulting in a fully flexible model for the intensity of rainfall. Unfortunately, there is no simple way to do this and simulate thousands of years of data in a matter of seconds (i.e. there is no joint conjugate prior for both the shape and rate parameters that yields a closed form posterior).

We propose two different alternative methods for estimating $\alpha_d$ for each dekad. First, we estimate a smoothed version of $\alpha_d$ by fitting a model to the method of moments (MOM) estimates that assumes a yearly cycle. Second, we simply use the MOM estimates themselves. The first alternative is more robust to outliers, but is also less flexible and less able to pick up patterns beyond those that follow a yearly cycle. Figure 4 shows the smoothed estimates and the MOM estimates of the shape parameter $\alpha_d$ for each dekad for the Adiha data.
Results

We evaluate the three models

1. Model 1: Fixed $\alpha_d = \bar{\alpha}$ for each dekad

2. Model 2: Smoothed $\alpha_d$ (see Figure 4)

3. Model 3: MOM $\alpha_d$

in two different ways. First we use posterior predictive checks. Second, we use cross-validation to compute out-of-sample errors.

Posterior Predictive Checks

Posterior predictive checks consist of comparing the value of an observed statistic to the posterior predictive distribution of that statistic. For example, suppose we are interested in the mean and
standard deviation of rainfall in Dekad 21. The observed values of these statistics are 89 mm and 25 mm, respectively. Figure 5 contains histograms of the posterior predictive distributions of these statistics for the three types of models:

All three models appear to be a reasonable fit, although Models 1 and 2 both tend to overestimate the standard deviation of rainfall in this dekad, as evidenced by the fact that the posterior predictive distribution of this statistic for both of these models puts much of its mass in the range 30 mm - 80 mm, whereas the observed value is about 25 mm.

For another posterior predictive check, consider the maximum dekadal rainfall value for each dekad. Figure 6 plots the observed maximum values (x-axis) against the posterior mean maximum values for each dekad (y-axis) and model. Two points in Figure 6 stand out. First, in the upper right corner, there is a green and a red point; these both reflect the fact that for Dekad 21, the gamma distributions for Models 1 (red) and 2 (green) have right tails that are too long - they overestimate the maximum dekadal rainfall. The other point that stands out is not actually in the plot, because it is too far away. In Dekad 15, the observed maximum dekadal rainfall was 36 mm, but the mean maximum dekadal rainfall in the posterior predictive distribution using Model 3 is 245! The reason for this is that the MOM estimate of $\alpha_1 = 0.56$, which results in a very long-tailed distribution ($\alpha_2 = 1$ is the exponential distribution), allowing for a few very large values of dekadal rainfall to be simulated.
Figure 6 Posterior Predictive checks of the maximum dekadal rainfall by dekad and model, where the y-axis provides the scale for posterior means of the maximum rainfall for each dekad and model.

In summary, the posterior predictive checks can provide detailed accounts of exactly where a model is fitting the data well, and where it is fitting the data poorly. If we are concerned about simulating a few huge dekads of rainfall, we would avoid Model 3, or at the very least, cap the dekadal rainfall simulations that come from this model using a reasonable maximum. For the Adiha data set, checking means and standard deviations for all dekads shows that Model 1 and 2 contain more uncertainty in the simulations than Model 3. None of the three models is clearly superior, although there is reason to be concerned about Model 3 because of some unrealistically large simulated values.

**Cross-Validation**

Another way to evaluate models is by using cross-validation to simulate out-of-sample predictions. We perform 7-fold cross validation and display the results in Figure 7. The basic idea of cross-validation is to divide the data into a training set and a test set, fit the model to the training set, and then evaluate the model's predictions on the test set. 7-fold cross-validation consists of partitioning the data into 7 sections, and for each section, using the 6/7 section of data
as the training set, and using the 1/7 section of data as the test set. This results in each data point being part of the test set exactly one time.

Figure 7 Cross-validation results for 3 models, and 2 data sets. The leftmost 3 boxplots contain results from the Adiha data set, and the rightmost 3 boxplots contain results from the Lilongwe data set.

In our final procedure, we perform 7-fold cross-validation 1000 times, for each of the 3 models, and each of the 2 data sets. That is, for each model and data set,

1. For \( g = 1, \ldots, 1000 \):
   a. For data sections \( i = 1, \ldots, 7 \):
      i. Fit the model to 6/7 of the data (Adiha or Lilongwe), leaving out section \( i \) of the data. For the Adiha, the left-out section is one year. For the Lilongwe data, it is 6 or 7 years.
      ii. Sample one value from the posterior distribution of the parameters: 
          \[ \theta^{(g)} = (p_1, p_2, \ldots, p_{36}, \beta_1, \beta_2, \ldots, \beta_{36}, \alpha_1, \alpha_2, \ldots, \alpha_{36}) \], where given the data and the model, \( \alpha_d \) is determined for all \( d = 1, \ldots, 36 \) and doesn't vary from iteration to iteration. In other words, it doesn't depend on \( g \), whereas \( p_d \) and \( \beta_d \) do vary from one iteration, \( g \), to the next, \( g + 1 \).
iii. Compute the deviance of each data point in the test set (section i of the data), where the deviance of a data point is 
\[ d_{di}^{(g)} = -2 \times P(Y_{di} | \theta_{(g)}) \].

b. Compute the average deviance across the whole data set,

\[ \bar{d}(g) = \frac{1}{36T} \sum_{g=1}^{36} \sum_{t=1}^{T} d_{di}^{(g)} \]

Recall that for each iteration, \( g \), each data point \( Y_{di} \) is a member of the test set exactly once.

Figure 7 contains boxplots of \( \bar{d} = (\bar{d}^{(1)}, \bar{d}^{(2)}, ..., \bar{d}^{(1000)}) \) for each of the three models and two data sets, where each boxplot contains \( G = 1000 \) points. Lower values of the average deviance indicate a model that makes better predictions. For the Adiha data set, Model 2 (smoothed \( \alpha \)) does the best, Model 1 (fixed \( \alpha \)) is the second best, and Model 3 (MOM \( \alpha \)) is the worst. This isn't surprising - the MOM estimates for a small data set are likely to be bad, compared to the fixed \( \alpha \) model results, because the fixed \( \alpha \) is more robust to outliers, whereas MOM estimates are sensitive to outliers.

For the Lilongwe data, the results are slightly different: the ranking of the models from lowest average deviance to highest is 2,3,1. In this case, the data set is relatively large, and the MOM estimates aren't that bad, whereas the fixed \( \alpha \) model is too conservative, and fails to pick up important differences between dekads, that the MOM estimates have the flexibility to model.

Ultimately, however, for both data sets Model 2, smoothed \( \alpha \), was the best, and we recommend it going forward. It is quick and easy to compute, and its most important advantage is that it models dependence between neighboring dekads, because the smoothing function is a regression of dekadal means and sds on the periodic elements \( x_1 = \cos(2\pi t / 36) \) and \( x_2 = \sin(2\pi t / 36) \), and then the shape parameter is estimated from these smoothed functions.

The denominator in the sine and cosine terms contains a 36 because that is the number of dekads in one year, which is the most dominant cyclical component of rainfall in virtually every site in the world. It is essentially a compromise between MOM estimates of the shape parameter for each dekad and a constant value of the shape parameter for each dekad.

**Bibliography**