# Session 8: stochastic rainfall simulation (Lab Report 2)

*Natalie Nelson, PhD; Biological & Agricultural Engineering, NCSU*

*03/16/2018*

## Background

The USDA would like to estimate how rainfed (non-irrigated) crop yields could be impacted under various rainfall scenarios (drought, excessive rainfall, etc.). Seasonal rainfall impacts are typically reported in terms of changes to the total rainfall, but the timing of rainfall is often more consequential for crop yields than the total volume. The USDA wants to run rainfall simulations that account for both the timing and total amount of rainfall.

The USDA is considering using the Independent Rectangular Pulses Model (IRPM) to simulate changes in precipitation patterns across rainfed agricultural fields in the US, but they are unsure whether the IRPM can effectively simuate rainfall patterns. They have hired you as a consultant to evaluate the IRPM.

The IRPM simulates storm events stochastically (Figure 1). Specifically, it uses probabilistic relationships to simulate: (1) the total rainfall from a given storm ("storm depth"; h), (2) the storm duration time (ts), and (3) the "interarrival" time between storms (tr). The IRPM assumes that h, ts, and tr follow **exponential distributions**. This model then strings this information together to simulate rainfall over time.

You decide to evaluate the IRPM by running a case study using precipitation data collected at the NCSU Lake Wheeler agricultural research station. You will run the model to simulate rainfall data at Lake Wheeler, and present the results to the USDA in a report. In this report, you will summarize the underlying statistical structure of the model, and describe its advantages and disadvantages. Based on the advantages and disadvantages, you will also offer recommendations as to what types of applications this approach might be good for, and types of applications that this approach might not be good for (give your best guesses for these recommendations; your recommendations will be graded based on completeness, not accuracy).

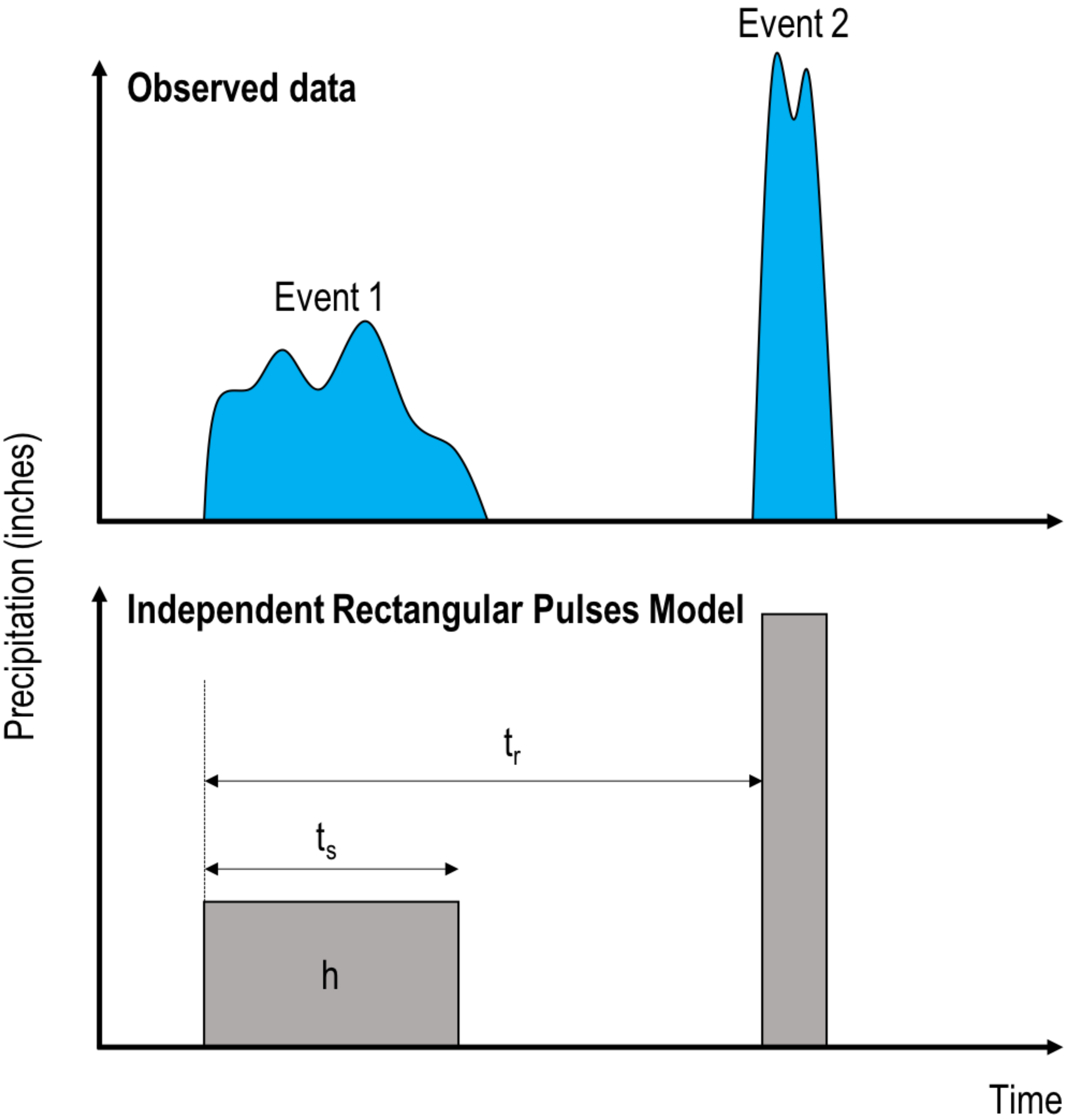

**Figure 1**. The Independent Rectangular Pulses Model

## Rainfall data

Your data are included in one data file, “LakeWheeler_precip.csv”. The data include the date and time of data collection (hourly data collected from 01/01/2018 - 03/06/2018), the corresponding rainfall measurements (inches), and the storm event. 40 storm events were observed over the study period.

## Exponential distribution

- PDF: $\lambda e^{-\lambda x}$
- Mean: $1/\lambda$
- Variance: $1/\lambda^2$

$\lambda$ is described as the "rate parameter". In the lecture slides, we parameterized the exponential distribution by the scale parameter $\beta$. The rate parameter is simply the inverse of the scale parameter ($\lambda = 1/\beta$). In R, the exponential distribution is parameterized with the rate parameter ($\lambda$).

## Analysis objectives

**All analysis should be performed in R.** It is suggested that these objectives be addressed in the order they are listed.

1. In order to run the model, you will need to fit three exponential distributions (h, ts, tr) to the Lake Wheeler data.

- Once you have fit the distributions to your data, visualize the corresponding empirical and theoretical probability densities for each of the three factors. You will create one separate ggplot for h, ts, and tr (**3 figures in total**). In each of these plots, you should have the empirical probability density, as well as a line that represents the theoretical probability density.
- Next, visualize the empirical and theoretical cumulative distributions for h, ts, and tr (**3 figures in total**)

2. Now that your probability distributions are parameterized, you need to generate "synthetic data" for h, ts, and tr. "Synthetic data" refers to the fact that we will be creating data *that are not real*. These synthetic data will take on the properties of the probability distribution that we fit to our *real observations*. So, although the data aren't real, they should ideally have similar characteristics to the observated data. You can generate synthetic data using `rexp()` for the exponential distribution (or `rnorm()` for the normal, `rlnorm()` for the lognormal, etc.).

- Visualize the synthetic data for h, ts, and tr using histograms (**3 figures in total**).

3. Run the IRPM and visualize the observed and simualted data in **1 facetted figure**. The figure should have two panels (one for observed, one for simulated) and the two panels should be stacked one on top of the other (1 column). How did the IRPM do? Do you think this is an effective tool for simulating storm events? What are the pros and cons?
4. Summarize the simulated and observed data in **a table**. The table should include (1) total, (2) mean, (3) median, (4) maximum, and (5) minimum precipitation for both the observed and synthetic data. How do the summary statistics between the observed and simulated data compare?

## New functions

Note that none of these functions require any new packages to be downloaded.

- `lag()` - creates a lagged copy of a data series (e.g., a column in a dataframe)
- `stat_function()` - include this in your ggplot code to add a line to your plot that corresponds to a provided statistical function (e.g., pnorm, dnorm, qnorm)
- `ceiling()` - round the value to the next highest integer (e.g. `ceiling(2.2)` = 3, `ceiling(5.6)` = 6)

- `geom_area()` - add this to your ggplot code to plot a line with a filled area (which is to say, the area beneath the line is filled in with a color)

## Report

The primary objective of your report is to summarize the IRPM and this case study, detail your methods, and inform the USDA as to what the advantages and disadvantages of the IRPM are, as well as what applications the IRPM is best suited for. Although you can work together to perform the analysis, the report **must be in your own words** and all material should be original. Any reports with evidence of plagiarism will receive no credit.

Be sure to address each of the analysis objectives in your report. Your report should be structured as follows:

- Title page - title (create your own title), name, date
- Executive abstract - summarize the report in 4-6 sentences; be sure to include your main results and recommendations, as this is the most important information included in abstracts.
- Introduction - provide all background information needed to understand the contents of the report. Be sure to *paraphrase* the writeup included here; do not copy directly.
- Methods - explain the statistical methods you used, both based on the theory and the specific application at hand. Demonstrate that you understand how and why you used the statistical principles (exponential distributions, probability densities, cumulative distributions) in this case study. Include equations when relevant. **You do not have to detail the steps you took in R.**
- Results - state the results objectively. Explain what the figures and numbers you've produced from this analysis are telling you.
- Recommendations - inform the USDA as to what the pros/cons are of the IRPM, and offer recommendations as to what types of applications might be best suited for the IRPM.
- Appendices - include your R code (should be neatly organized and professional).

**Additional notes on formatting:**

- All figures and tables should be captioned, numbered, and referred to in the text.
- Table captions are placed *above* the table. Figure captions are placed *below* the figure.
- Equations should be centered and included on their own lines.