UPPSALA UNIVERSITET

# TripleM

## User manual

Code release V 1.0

Korbinian Breinl*, Marc Girons Lopez and Giuliano Di Baldassarre

University of Uppsala, Department for Earth Sciences, LUVAL, Villavägen 16, 752 36 Uppsala, Sweden

23 February 2017

*korbinian.breinl@geo.uu.se

# Document information

| | |
|---|---|
| Model name | TripleM precipitation algorithm |
| Document title | TripleM User manual |
| Document version | 1.0 |
| Release data | 23 February 2016 |

# Executive summary

This document is the quick user manual for the open-source daily multi-site precipitation generator TripleM (Multisite Markov Model) version 1.0. The code is a scientific tool for the fast simulation of synthetic daily precipitation time series at one or multiple sites as required for many different studies in earth sciences such as atmospheric sciences, hydrology, hazard and risk assessment or climate impact assessment. When simulating a single site, the code corresponds to a classic 'Richardson type model', where precipitation occurrences are first simulated with a Markov chain and precipitation amounts are randomly sampled, either from a parametric distribution [1] or from the observations. When two or more sites are simulated, daily snapshots of precipitation occurrences at all sites – also called 'occurrence vectors' - are clustered into similar precipitation patterns, which are then simulated in a Markovian process. The model has been first described in Breinl, et al. [2] for a small number of rain gauges, and has been further improved for modelling a larger number of rain gauges. In case of multiple sites, the model can either be used as a pure resampling (bootstrap) algorithm, meaning that the simulated precipitation is built from the observations, or in a parametric version, where the resampled observed precipitation amounts are replaced by synthetic precipitation amounts, which allows for the simulation of unobserved precipitation extremes. The model can be set up in a monthly or seasonal setup depending on the requirements of the application.

# History of model versions

| Version | Date | Information |
|---|---|---|
| 1.0 | February 2017 | Latest version created by Korbinian Breinl and Marc Girons Lopez in MATLAB R2016a, new clustering method implemented for modelling larger gauge networks, full flexibility through soft-coding of key elements, model renamed into TripleM (Multisite Markov Model), first public release of precipitation algorithm |
| 0.7 | June 2014 | Advanced version created by Korbinian Breinl in MATLAB R2013b, called the RSS (Reduced State Space) model, under affiliation with Salzburg University, Department of Geoinformatics Z_GIS, Salzburg, Austria |
| 0.5 | August 2013 | First version created by Korbinian Breinl in MATLAB R2012a [2], called the FSS (Full State Space) model applicable for a very small number of rain gauges, under affiliation with Salzburg University, Department of Geoinformatics Z_GIS, Salzburg, Austria |

# Table of Contents

# 1. Data requirements and input files

## 1.1. Data requirements

The only data requirement for TripleM is a matrix with daily precipitation records and their dates of recording, for one or multiple sites. The unit of precipitation amounts is in the metric system (mm), dates have to be defined in the format YYYY-MM-DD.

## 1.2. Input files

### 1.2.1. Rain file (rain.csv)

This file contains the precipitation records used in TripleM. The file is a comma separated .csv file. The first column of the file contains the dates of precipitation records in the format YYYY-MM-DD. The second and/or following columns contain the precipitation records in mm. The header is “date” for the date column and “S1”, “S2”, “S3” etc. for the precipitation records, depending on the number of sites modelled. The file can be prepared in MS Excel or in other spreadsheet applications. The precipitation records do not have to complete, i.e. (complete) days can be missing. An example precipitation input file rain.csv is given below (for five rainfall gauges):

```
date,S1,S2,S3,S4,S5
1986-01-01,0,0,0,0,0
1986-01-02,0,0,0,0,0
1986-01-03,0,0,0,0,0
1986-01-04,14.5,0.5,4.6,0,3.8
1986-01-05,38.1,7.1,2.5,0,5.1
1986-01-06,2,0,0,0,0.5
1986-01-07,0,0,0,0,0
1986-01-08,0,0,0,0,0
1986-01-09,0,0,0,0,0
1986-01-10,0,0,0,0,0
1986-01-11,0,0,0,0,0
1986-01-12,0,0,0,0,0
1986-01-13,0,0,0,0,0
1986-01-14,0,1.3,1.5,0,0
(…)
```

If a single site is modelled, TripleM is essentially similar to a generic ‘Richardson type’ precipitation generator, which has been thoroughly discussed in the literature [1,3].

### 1.2.2. Model parameters (param.yml)

This file defines all model parameters required for the simulations. The required values have to be written between the colon and the hashtack of each line. The file can be altered in any text editor. An example model parameter file param.yml is given below (Figure 1):

```
# ----------------------
% Copyright (c) 2017, Korbinian Breinl
% All rights reserved.
#
# Model parameter file
# ----------------------

p_thr      : 0.1     # precipitation threshold
start_sim  : 2017    # start of the simulation period
length     : 30      # length of the generated precipitation series
max_dup    : 1       # maximum number of duplicates
mc_order   : 1       # order of the Markov chain
cl_period  : season  # clustering period ('month' or 'season')

param_p        : off      # parametric precipitation sampling ('on' or 'off')
  min_sample   : 100       # minimum number of samples per season/month for precipitation distributions
  corr_rand    : off       # correlated random numbers ('on' or 'off')
  dist         : Weibull-GPD    # distribution ('Weibull', 'Gamma' or 'Weibull-GPD')
    pareto_thr : 3        # threshold for GPD distribution when using Weibull-GPD model

```

*Figure 1. Exemplary parameter file*

In the following, all parameters are explained in detail.

### p_thr

This is the threshold that defines a day with recorded precipitation amounts below this threshold as dry. The ideal value depends on the precision of the precipitation recording equipment. Typical values are 0.1 or 0.2. The same specified threshold is applied to all observations in the input matrix rain.csv.

### start_sim

Defines the start year of the simulations, which always starts with the 1st of January of that year.

### length

Defines the length of the simulated time series in years specified with an integer number. The user has to be aware that reliable long extrapolations require a solid data basis of usually 30 years or more of precipitation observations.

### max_dup (ignored for single-site simulations)

Defines the maximum duplication rate of the simulated time series in percent. The model TripleM has the inherent characteristic of producing duplicated time series. A large duplication rate improves several statistical metrics of the simulated time series but at the expense of duplicated fragments in the simulated time series. For this reason, the duplication rate can be fixed. We recommend a value between 1% and 5%, which has turned out to produce sufficient time series with still low duplication rates. Ideally, the user should conduct a sensitivity analysis to compare the trade-off between the duplication rate used and the quality of the simulated time series. A larger duplication rate means higher computational costs.

If a single site is modelled, *max_dup* is ignored as single site simulations always have a 0% duplication rate.

**mc_order**

Defines the order of the Markov chain used. In its current setup, the same order is applied to all months or seasons, depending the model type (see *cl_period*).

**cl_period**

Defines the simulation setup of the model. Two options are available: A monthly model routine where all simulation steps are conducted on a monthly basis ('month') and a seasonal setup where all steps are conducted on seasonal basis ('season'). In the monthly model setup, a new amount vector is randomly drawn on each first day of a month and this also applies to the seasonal setup with random draws after each season. The seasons are defined as Spring (March, April, May), Summer (June, July, August), Autumn (September, October, November) and Winter (December, January, February).

**param_p**

The user can choose whether to simulate unobserved synthetic precipitation amounts from parametric distribution functions ('on') or not ('off'). If parametric precipitation sampling is switched off TripleM works like a bootstrap by purely resampling observed precipitation amounts.

**min_sample** (ignored when param_p = 'off')

When using parametric precipitation sampling, this value defines the minimum number of non-zero precipitation observations used for fitting the parametric precipitation distributions. A reasonable threshold avoids the fitting parametric distributions to too few precipitation observations, for example in arid climates. The minimum sample size refers to all sites and periods (i.e. months or seasons). If the observations in a period only provide a number of non-zero precipitation amounts below this threshold, the code simply resamples observations for that period and does not conduct parametric distribution fitting.

**corr_rand** (ignored when param_p = 'off')

The parametric sampling of precipitation amounts over several sites can be conducted with or without correlated random numbers and be switched on ('on') or off ('off'). Precipitation correlations are derived from the observations. We recommend using this option when simulating short time series of only few years. The effect is much less pronounced when simulating long time series. In some cases, the derivation of correlated random numbers by a Cholesky factorization can lead to errors when running the model, because of ill-defined correlation matrices due to noise in the observed data see for example [4]. We recommend simulating without correlated random numbers in such cases.

**dist** (ignored when param_p = 'off')

Defines the parametric distribution type to model synthetic (unobserved) precipitation amounts. Currently available are the Gamma distribution ('Gamma'), the Weibull distribution ('Weibull') or a compound distribution with a Weibull distribution for lower and a General Pareto distribution for higher precipitation amounts ('Weibull-GPD'). In the current setup of the code, all sites and periods (months or seasons) are simulated with the same parametric distribution as defined. The value is ignored if parametric precipitation sampling is switched off.

**pareto_thr** (ignored when param_p = 'off')

This value defines the threshold between the Weibull distribution for lower and the Generalized Pareto distribution (GPD) in mm. In the current setup of the code, the same threshold is applied to all sites and periods (i.e. months or seasons). Suitable thresholds can for example be derived by 'mean residual life plots' or 'parameter stability plots', for instance with the package 'extRemes' in the statistical programming language 'R' [5]. Studies suggest fairly low thresholds of only a few millimetres [6]. In an Alpine study area modelled in Breinl, et al. [7] the average threshold was 3.7mm. The value is ignored when other parametric distributions are chosen, or when the parametric

# 2. Preparing and running a simulation

The following files are required for a simulation. The two files param.yml and rain.csv have to be prepared by the user (see above).

| File | Description | Comment |
|---|---|---|
| data_cluster.m | Routine for clustering of the amount vectors | |
| param.yml | Main input file for model parameters | To be prepared by user |
| rain.csv | Input matrix with precipitation observations | To be prepared by user |
| read_file.m | Routine to read rain.csv | |
| read_param.m | Routine to read parameter values from the file param.yml | |
| trans_count.m | Routine to count transitions probabilities for Markovian simulation | Copyright (c) 2014, Shawn Pethel (https://se.mathworks.com/matlabcentral/fileexchange/40188-whittle-surrogate/content/trans_count.m) |
| TripleM.m | Main application | |

All files as described in section 2 have to be copied into a common folder. The simulation is run by executing the file TripleM.m. The simulated precipitation time series are written out into the same folder (sim.csv).

# 3. Recommendations

From our experience, we can give the following general recommendations for applying TripleM:

- The maximum duplication rate (*max_dup*) should be fixed between 1% and 5%, which turned out to produce good results. Higher duplication rates increase the computing time significantly and turned out to lead to only minor improvements of the modelling results.
- The order of the Markov chain (*mc_order*) should not be larger than 1, unless a single site is modelled. If more than a single site is modelled, the state space of the Markov chain increases exponentially with higher orders, which reduces the quality of the simulations for the same duplication rate.
- Extrapolating with a weather generator requires a solid database. We recommend at least 25 observations years but ideally more (30 years or more).
- When using parametric precipitation sampling, a solid number of precipitation observations is required for stable parametric fits. A sufficient value is required for *min_sample*. If the observation time series do not provide enough precipitation observations, for example in arid climates, the workaround is either a seasonal instead of a monthly model setup or the application of a pure bootstrap without parametric precipitation sampling.

# 4. References


1 Richardson, C. W. Stochastic Simulation of Daily Precipitation, Temperature, and Solar-Radiation. *Water Resour Res* **17**, 182-190 (1981).

2 Breinl, K., Turkington, T. & Stowasser, M. Stochastic generation of multi-site daily precipitation for applications in risk management. *J Hydrol* **498**, 23-35, doi:http://dx.doi.org/10.1016/j.jhydrol.2013.06.015 (2013).

3 Katz, R. W. Precipitation as a Chain-Dependent Process. *J Appl Meteorol* **16**, 671-676 (1977).

4 Brissette, F. P., Khalili, M. & Leconte, R. Efficient stochastic generation of multi-site synthetic precipitation data. *J Hydrol* **345**, 121-133, doi:10.1016/j.jhydrol.2007.06.035 (2007).

5 Gilleland, E. & Katz, R. W. New Software to Analyze How Extremes Change Over Time. *Eos, Trans Amer Geophys Union* **92**, 13-14, doi:10.1029/2011eo020001 (2011).

6 Deidda, R. A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series. *Hydrol Earth Syst Sc* **14**, 2559-2575, doi:DOI 10.5194/hess-14-2559-2010 (2010).

7 Breinl, K., Turkington, T. & Stowasser, M. Simulating daily precipitation and temperature: a weather generation framework for assessing hydrometeorological hazards. *Meteorol Appl* **22**, 334-347, doi:10.1002/met.1459 (2015).