

Accounting for Joined Probabilities in Nation-Wide Flood Risk Profiles



Ferdinand Diermanse, Joost V. L. Beckers, Cathy Ansell, and Antoine Bavandi

Abstract A risk profile provides information about the probabilities of event impacts of varying magnitudes. In this study, a probabilistic framework is developed to derive a national-scale flood risk profile, which can be used for disaster risk management and financial planning. These applications typically require risk profiles over a wide range of return periods. For most countries, the historical record of flood impacts is limited to a few decades, insufficient to cover the longest return periods. To overcome this limitation, we developed a stochastic model that can generate arbitrarily long synthetic time series of flood events which have the same statistical characteristics as the historical time series. This includes the joint occurrence probabilities of flood events at different locations across the country. So, the probability of each pair of locations experiencing a flood event in the same event should be the same for the synthetic series as for the historic series. To this end, a novel approach based on ‘simulated annealing’ was implemented. Results show an almost exact reproduction of the statistical properties of the historical time series.

Keywords Joint probabilities · Risk profiles · Simulated annealing

1 Introduction

In order to increase the financial resilience of ASEAN + 3 members to climate and disaster risk, the Southeast Asia Disaster Risk Insurance Facility (SEADRIF) has been established by ASEAN + 3 in partnership with the World Bank. To support this endeavor, and to increase the financial resilience of Lao PDR, Cambodia and Myanmar against large-scale floods, the World Bank commissioned the development of tools to support a rapid response financing mechanism. Flood risk profiles for

F. Diermanse (✉) · J. V. L. Beckers
Deltares, Delft, The Netherlands
e-mail: ferdinand.diermanse@deltares.nl

C. Ansell · A. Bavandi
Disaster Risk Financing and Insurance Program, World Bank, Washington, DC, USA

these three countries were required to inform the design of financial risk transfer instruments.

Disaster risk is often quantified in terms of “annual average affected population” or for financial purposes, “annual average losses”. These are the long-term averages over many years of larger and smaller disasters as well as no-event years. However, for financial risk transfer, such as catastrophe insurance, a year loss table (YLT) or event loss table (ELT) is required due to the importance of the low frequency and high severity events. Crucially, these also provide information on the temporal and spatial variance of individual events which cannot be represented by an “average year”. The historical record is often too short to act as an ELT and will not sufficiently represent the tail events, therefore a stochastic event set is required. We developed a methodology to derive a long synthetic time series of flood events to support SEADRIF countries.

2 Flood Modelling Concept

The flood modelling in this project is based on the concept of a single flood driver for a given type of flood and subarea. For example, a flood plain along a stretch of river is called a fluvial subarea. The local flood driver in this case is the river discharge. We assume the flood extent in the fluvial subarea is fully determined by the value of this flood driver. We defined four types of flooding and corresponding flood drivers: Fluvial flooding (river discharge), pluvial flooding (rainfall), tidal flooding (river water level) and coastal flooding (sea water level). The first step in the flood risk analysis is to identify and classify the subareas over the region of interest (typically a country). For each subarea, historical values for the flood driver are collected. Subsequently, the number of affected people for each flood map is calculated using the WorldPop population density grid [11].

The historical flood driver values include gauge readings from local hydrometeorological centers over the past few decades, simulated river discharges from a hydrologic model over a 35-year period (using 1979–2013 MSWEP meteorological input, see [1]), as well as storm surge levels from the Global Tide and Surge Reanalysis (GTSR) data set which also spans 35 years [9] augmented by observations and hydrodynamic simulation of historical cyclones.

3 Method for Generating Synthetic Time Series

The historical period of 35 years is sufficient for probabilistic assessment up to return periods of about 10 years, but not for the longer return periods (up to 1000 years) which are required for assessment of low frequency, high severity events. Therefore, we generate a long synthetic time series of flood events (characterized by flood driver values) that enables the analysis of higher return periods. Our methodology generates

a synthetic time series which has the same statistical characteristics as the 35-year historical time series. This includes:

1. exceedance probabilities of annual maxima for each flood driver;
2. mutual correlations of annual maxima between all pairs of flood drivers;
3. probabilities of joint occurrences for all pairs of flood drivers, i.e. the probability that annual maxima of two flood drivers occur during the same event;
4. the influence of the joint occurrence on the correlation between the values of annual maxima (because if annual maxima occur during the same event, the mutual correlation is generally much higher).

A stochastic event sampling method was developed that reproduces these four statistical properties. This method consists of four components:

Component 1: deriving probability distributions of individual flood drivers;

Component 2: sampling of annual maxima of flood drivers;

Component 3: sampling of events;

Component 4: linking of annual maxima to event numbers.

These four components are detailed in the next four subsections.

3.1 Component 1: Deriving Probability Distributions of Flood Drivers

Extreme value distributions were derived for the various flood drivers, based on the available 35-year historical time series. For each flood driver, annual maximum values were selected, and an extreme value distribution function was fitted, applying fairly “standard” techniques such as described in Coles [2]. Figure 1 shows an example of a Gumbel fit on annual maximum discharges for Nam Khan River in Laos.

3.2 Component 2: Sampling of Annual Maxima

The second component of the sampling method concerns the sampling of annual maxima. In this step, the mutual correlation between the annual maxima of different flood drivers is taken into account. The correlation coefficient is derived for all flood driver pairs from the observed annual maximum values. This results in an $n * n$ covariance matrix, C , where n is the number of flood drivers ($n = 127$ for the study area).

To reproduce these correlations in the synthetic time series, a Gaussian Copula is applied in the sampling procedure (see e.g. [4, 8]). This method requires correlation matrix, C , as input. As proven by Fang et al. [5], C should be taken equal to $\sin(\pi\tau/2)$, where τ is Kendall’s rank correlation matrix. The procedure to generate correlated samples is as follows:

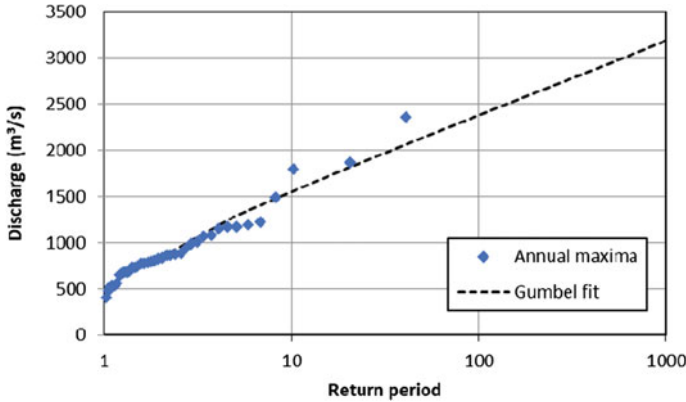


Fig. 1 Fitted Gumbel distribution function to simulated annual maximum discharges at Ban Mixay gauge station (Nam Khan River, Lao PDR)

1. Derive a matrix \mathbf{P} for which: $\mathbf{PP}' = \mathbf{C}$, through Cholesky decomposition of correlation matrix \mathbf{C} (see, e.g. [12]). Note: \mathbf{P}' is the transpose of matrix \mathbf{P} .
2. Sample values u_1, \dots, u_n from the standard normal distribution function; store the results in a $1 \times n$ vector \mathbf{u} .
3. Compute: $\mathbf{u}^* = \mathbf{uP}'$.

The \mathbf{u}^* -values are subsequently transformed to “real-world” values of flood drivers, using the probability distribution functions of the individual flood drivers as derived in component 1:

$$\Phi(u_i^*) = F_i(x_i) \tag{1}$$

Where Φ is the standard normal distribution function, u_i^* is the sampled \mathbf{u}^* -value of the i th flood driver, x_i is the “real-world” realisation of the i th flood driver and F_i is the derived extreme value distribution function of x_i .

3.3 Component 3: Sampling of Events

The sampling method for annual maxima (AM) in the previous section creates a synthetic time series with correlated annual maxima for each flood driver. The correlation between AM refers to their value, not to their timing. Within a single year, the annual maxima of the flood drivers are not expected to all occur during the same event. Typically, there are several events per year and the AM are distributed over them. Since we are interested in *event* impacts, the relative timing of the maxima within a year also needs to be part of the sampling method. The generated AM are thus assigned to events and the number of annual maxima per event should be in accordance with the historical series. More specifically: the probability of each pair

of flood drivers having an AM during the same event should be the same for the synthetic series as for the historical series. In the case study, there are 127 flood drivers, which means there are $127 * 126/2 = 8001$ joined occurrence probabilities that need to be reproduced in the sampling procedure. For this challenging objective we developed a novel approach based on simulated annealing (Kirckpatrick et al., 1983).

3.3.1 Outline of the Method

The starting point of the analysis is a time series z_t , consisting of event numbers as derived from historical data. z_t is an $Y * n$ matrix, with Y being the number of years and n the number of flood drivers. If $z_t(y,j) = 3$, this means flood driver j had its AM in the 3rd “biggest” event of year y (note: in each year, events are ordered based on the number of flood drivers that had their AM occurring during the event). As an illustration, Fig. 2 shows the five biggest events in 1979 and 1980 and the flood drivers that had their AM during one of these events. The objective of the stochastic simulation method is to create a (lengthy) synthetic time series z_{s_t} with similar characteristics as the historical time series z_t . Here, similarity refers to [1] the probability that AM of any two flood drivers, L_1 and L_2 , occur during the same event and [2] the probability distribution of the number of flood drivers having their AM occur in the biggest event in the year.

To this end, a “cost” function, $G(z_{s_t}, z_t)$, is defined that penalizes differences between the historical and synthetic series. G is formulated in such a way that it decreases if z_t and z_{s_t} are in better agreement. Thus, function G needs to be minimized to obtain the best agreement between the historical and synthetic series. The choice of function G is critical to the performance of the procedure, both in terms of computation time and the quality of the end result. Both z_t and z_{s_t} are matrices with the same number of columns, where each column represents a flood driver, but different numbers of rows, where each row represents a year. Matrix z_t has 35 rows, corresponding to the 35 years of observation, whereas z_{s_t} has a user-defined number of rows (e.g. 10,000). Both z_t and z_{s_t} contain event numbers, where 1 corresponds to the biggest event.

The event sampling procedure is based on the method of ‘simulated annealing’ (Kirckpatrick et al., 1983). Figure 3 shows the basic algorithm. The procedure starts with a randomly selected initial synthetic time series for which the cost function is evaluated. The elements of this synthetic time series are subsequently permuted in a (large) number of iteration steps, until a stop criterion is reached. In each iteration step, two elements of the time series are permuted to create a newly ‘proposed’ time series. For the simulated annealing procedure, we adopted the Matlab implementation of Joachim Vandekerckhove¹ and adapted it for our specific application. The algorithm is as follows:

¹<https://nl.mathworks.com/matlabcentral/fileexchange/10548-general-simulated-annealing-algorithm>

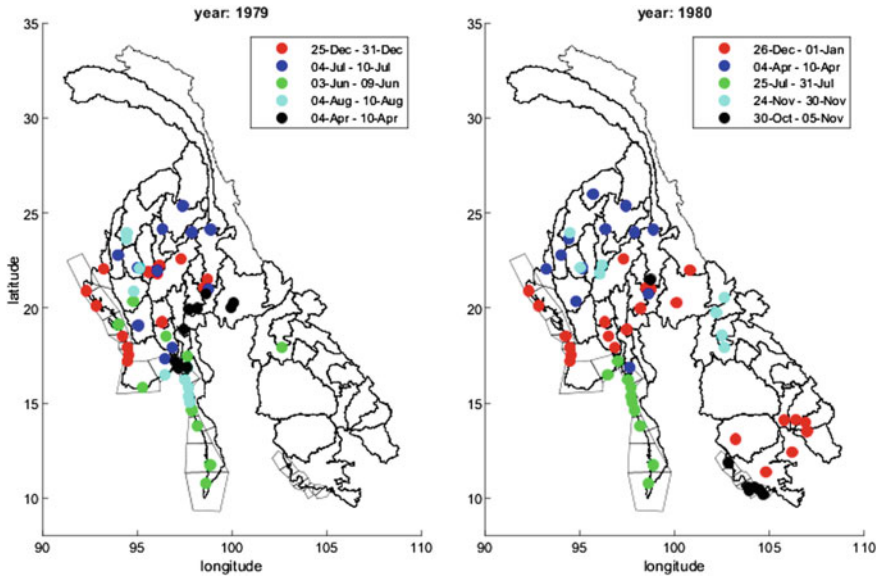


Fig. 2 Example of joint occurrences in the years 1979 and 1980. Clusters of flood drivers that experience the annual maximum within the same event have the same colour.

- [1] start with an initial time series $z_{s_t}(0)$
- [2] evaluate cost function $G(z_{s_t}, z_t)$
- [3] start with an initial 'temperature' $T = T_0$
- [4] Select an end temperature $T_e < T_0$ as stop criterion
- [5] while $T > T_e$ do: (*stop criterion 1*)
- [6] while *stop criterion 2* is not met do:
- [7] Select a new time series $z_{s_t}^*$ by randomly changing swapping 2 elements of z_{s_t}
- [8] evaluate cost function $G(z_{s_t}^*, z_t)$
- [9] accept $z_{s_t}^*$ as the new solution, i.e.: $z_{s_t} = z_{s_t}^*$, with probability $p(T)$ and reject it with probability $1-p(T)$
- [10] decrease T : $T = c * T$; with c a constant < 1

This algorithm explores the $Y * n$ -dimensional space of all possible outcomes, where Y is the number of years of the synthetic series and n is the number of flood drivers. The probability, $p(T)$, of accepting the proposal solution is decreasing to near-zero at the end of the procedure. Therefore, in the later phases of the procedure, the procedure has an increasing tendency to move into the direction of lower values of the cost function and to end up in a minimum. To prevent the procedure from ending up too early in a local minimum, a 'temperature' T is introduced, which allows for the solution to move to a higher value of the cost function. In the beginning of the procedure, the temperature is high, thereby increasing the probability of moving away

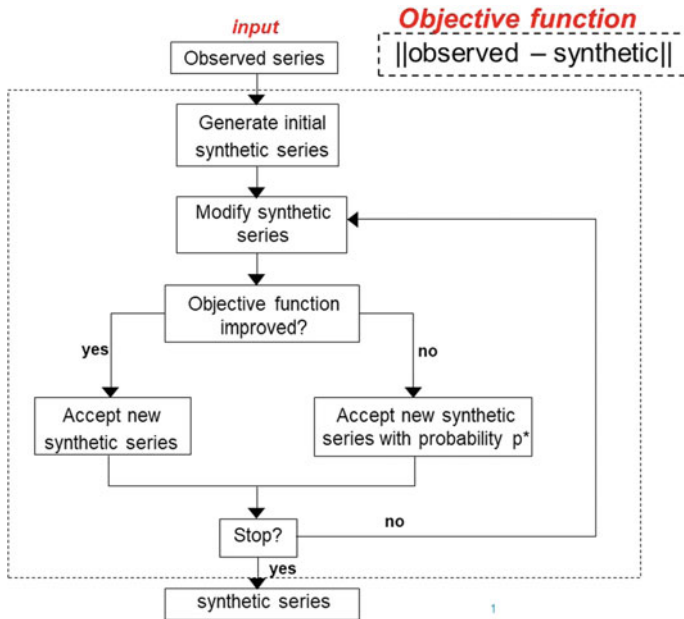


Fig. 3 Schematic view of the basic principle of simulated annealing

from a (local) minimum. The temperature is then slowly reduced, and the solution is forced to move towards a (local) minimum of the cost function. The procedure can be repeated several times with a different random seed to check if the same minimum is found each and every time. This is no guarantee for a global minimum, but it does provide more confidence.

3.3.2 Details of the Annealing Method

In this section, we describe some further details of the simulated annealing method step by step. The number below refer to the steps mentioned in Sect. 3.3.1.

[1] First, we have to define the number of years of the synthetic time series (Y) and the maximum number of events per year (N_e). The time series z_{s_t} consists of event numbers 1 to $N_e + 1$. For example: if $z_{s_t}(y,j) = 5$, this means flood driver j had its annual maximum in the 5th event of year y . Event number $N_e + 1$ represents the ‘non-event’, which means if $z_{s_t}(y,j) = N_e + 1$, the annual maximum of flood driver j in year y is assumed to have taken place in isolation. In the initialization of z_{s_t} , an integer number is randomly sampled from the range $[1, N_e + 1]$ for each combination of year y and flood driver j .

[2] The cost function $G(z_{s_t}, z_t)$ was chosen to be the sum of three functions $G_1 - G_3$. To compute these three functions, some pre-processing is required. First of all, the probability that annual maxima of any two flood drivers, L_1 and L_2 , occur during

the same event is estimated from the historic series. So, for example, if flood drivers L_1 and L_2 , had their annual maxima occur in the same event in 10 out of 35 years, the estimated probability of joint occurrence is $10/35 \approx 0.29$. This is done for each pair of flood drivers, resulting in $M = n * (n - 1)/2$ percentages ($n =$ the number of flood drivers), which are stored in an $M * 1$ vector H_o . The same computation is done for the current solution (z_{s_t}) of the synthetic time series, and the results are stored in an $M * 1$ vector H_s . Subsequently, the absolute differences between H_o and H_s are computed and stored in an $M * 1$ vector D . Function G_1 is the maximum value of D . Function G_2 is the mean value of D . To compute function G_3 , the maximum number of flood drivers with an annual maximum occurring in a single event is determined per year and stored in an $Y * 1$ vector. Subsequently, the mean over all the years is computed. In other words: the mean number of flood drivers that have their annual maximum occurring in the biggest event. This is done for the observed and synthetic series. Function G_3 quantifies the differences between the two.

[3 + 4 + 5] The starting value T_0 should be chosen in such a way that accepting a new time series with a higher (“worse”) cost function than the current time series should be relatively large, whereas T_{end} should be chosen in such a way that accepting a new time series with a higher (“worse”) cost function than the current time series should be close to 0. T_{end} should be several orders of magnitude smaller than T_0 to provide the method with a sufficient number of iterations to converge to a “good” result. The best choice of T_0 and T_{end} requires some insights in the cost function loss function $G(z_{s_t}, z_t)$ and the speed with which it converges to the (local) minimum. In our case T_0 was set equal to 1 and T_{end} equal to $1E-8$.

[6] Stop criterion 2 controls the number of iterations for a single temperature T . An obvious criterion is to set a maximum allowed number of iterations. Additional criteria can be to stop after a user-defined number of accepted new solutions and/or stop after a long successive series of rejected solutions. All three criteria have been implemented in the Matlab implementation of Joachim Vandekerckhove that was used as the basis of our method.

[7] A key step in the procedure is the selection of a new proposal time series. A straightforward method is to randomly select a specific flood driver in a specific year and to randomly generate a new event number for this flood driver. However, this approach led to very slow convergence of the procedure. To speed up the procedure, we implemented an alternative method in which we look for the combination of two flood drivers L_1 and L_2 that contribute most to the outcome of function $G_1(z_{s_t}, z_t)$. In other words, the two flood drivers L_1 and L_2 for which the difference in computed joint occurrence probability between the observed series z_t and the synthetic series z_{s_t} is the largest. We then change the event number of one of the flood drivers in such a way that the objective function is decreased. Note, however, that this approach is slightly in contrast with the concept of simulated annealing in which increases in the objective function in successive iterations should also be allowed to prevent it from converging too soon to a local minimum. Therefore, the final strategy was a mixture of both: with a probability p^* we apply the first method (random selection of flood driver) and with a probability $1 - p^*$ we apply the second method (selection of flood

driver that contributes most to the cost function). We found that a value of $p^* = 0.8$ in general gave most satisfactory results.

[8] The probability, p , of accepting a newly proposed solution, $z_{s_t}^*$, is set equal to:

$$p = \min\left(1, e^{-\frac{\Delta G}{kT}}\right); \Delta G = G(z_{s_t}^*, z_t) - G(z_{s_t}, z_t) \quad (2)$$

This means if $G(z_{s_t}^*, z_t) \leq G(z_{s_t}, z_t)$, the new solution is accepted with probability 1, whereas if $G(z_{s_t}^*, z_t) > G(z_{s_t}, z_t)$, the acceptance probability depends on the difference between $G(z_{s_t}^*, z_t)$ and $G(z_{s_t}, z_t)$, on the temperature T and constant k . To make this function generically applicable, cost function G is normalised by dividing it by G_0 , i.e. by the value of the cost function in the first iteration. Constant k should preferably be inversely proportional to the total number of elements, N (N is equal to the number of years times the number of flood drivers in our case). We chose $k = 10/N$.

[9] The decrease in the value of T is taken care of as follows: $T = c * T$; with c a constant < 1 . This means the temperature declines exponentially. We adopted a value of $c = 0.8$.

3.4 Component 4: Linking Annual Maxima to Event Numbers

The sampling procedures of component 2 and 3 are carried out independently from each other. That means the correlation between annual maxima of two flood drivers is not influenced by the fact whether these two maxima are observed during the same event. In reality, however, there is a relation between the two, as occurrence during the same event means there may be a common cause that is also likely to affect the magnitude of the annual maximum of both flood drivers. This is confirmed by an analysis of the data of the historical 35-year series. In the analysis we computed the correlation between rank numbers² of annual maxima for [A] all annual maxima occurring in the same year and [B] all annual maxima occurring in the same event. The table below shows the difference between the two is significant. Ignoring the relation between event numbers and (correlations between) annual maxima means the ‘within-event-correlation’ will be equal to the numbers shown under [A], whereas it should be equal to the numbers shown under [B].

Country	# Flood drivers	[A] Within year correlation	[B] Within event correlation
Cambodia	21	0.20	0.66
Lao PDR	22	0.24	0.35

(continued)

²Rank numbers are numbers from 1..35 indicating per flood driver the highest (1), second highest (2).. Lowest (35) annual maximum in the series of 35 years.

(continued)

Country	# Flood drivers	[A] Within year correlation	[B] Within event correlation
Myanmar	84	0.11	0.40
Combined	127	0.06	0.29

In order to take the relation between event numbers and annual maxima into account an additional simulated annealing procedure was implemented. In each iteration, a flood driver is randomly selected and the sampled annual maxima of this flood driver for two randomly selected years (as provided by component 2) are exchanged. This means the event numbers of these two annual maxima have been exchanged. This has an impact on the overall computed ‘within event correlation’. A cost function is defined that quantifies the difference between the observed and synthetic ‘within event correlation’. The iterations are carried out until the computed within event correlation is the same as the corresponding number of the observed data (see table above). It turned out that the method was capable of exactly reproducing these numbers.

4 Results

To test the applicability of the procedure, we verify if the relevant statistical properties of the historical times series are reproduced by the synthetic time series. Figure 4 shows the joint occurrence probabilities of annual maxima during an event for all flood driver pairs as derived from the historical series (vertical axis) and from a generated 10,000-year synthetic series (horizontal axis). The plot on the left shows results for Cambodia (21 flood drivers), the plot on the right shows results for the three countries combined (127 flood drivers). The figures show that the joint occurrence probabilities in the synthetic time series are in very good agreement with those in the historical time series.

The quantile plots in Fig. 5 compare the probability distributions of the number of flood drivers that had their annual maximum during the “biggest event” in the year, as derived from the historical series (horizontal axis) and from the 10,000-year synthetic series (vertical axis). The blue dots are all close to the line $y = x$. This shows the probability distribution of the number of locations in the largest event is very well captured in the synthetic series.

Figure 6 shows frequency curves of affected population that were derived from the synthetic series (red) and from the historical series (blue dots). The numbers were normalized by the 100-year return value as the actual numbers are not eligible for publication. The plot on the right is a zoomed version of the plot on the left. It shows that the derived frequency curves are well in accordance with the historical numbers, which is an essential validation of the method. The added value of the synthetic method is that it provides return values for much larger return periods, as can be seen from the left plot.

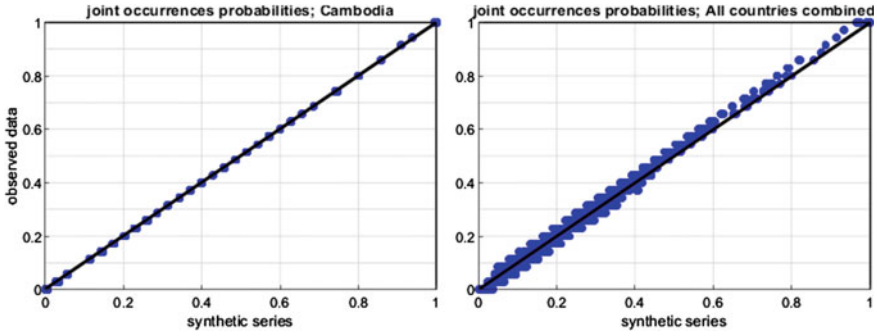


Fig. 4 Joint occurrence probabilities for all pairs of flood drivers in Cambodia (left) and all three countries combined (right). Synthetic time series (x-axis) compared to historical time series (y-axis)

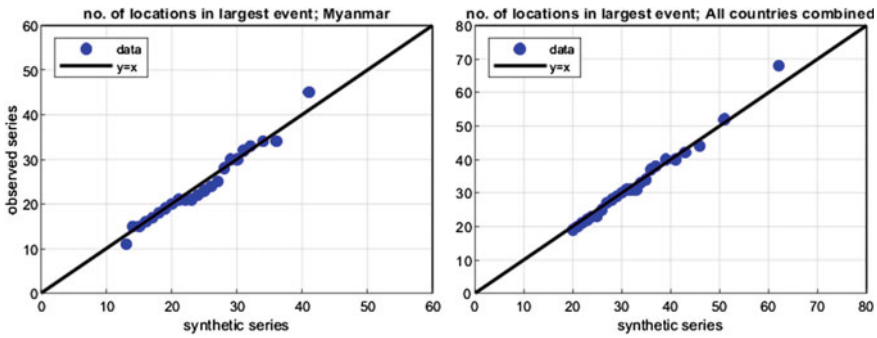


Fig. 5 Quantile plot for the number of flood drivers having their annual maximum during the largest event in Myanmar (left) and all three countries combined (right)

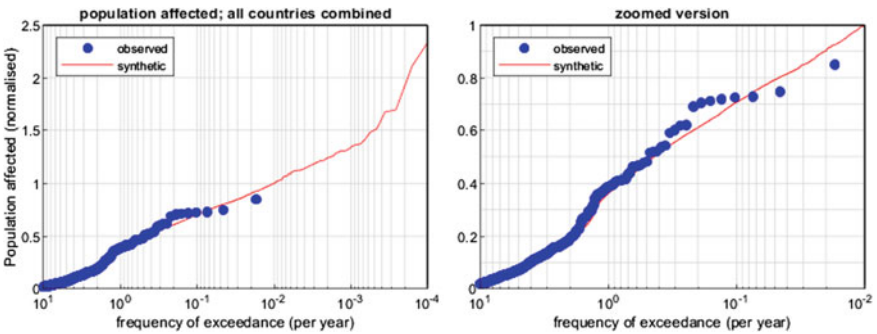


Fig. 6 Frequency curves (normalized) of population affected – empirical (blue) versus probabilistic (red). The plot on the right is a zoomed version of the plot on the left.

5 Conclusions

The results prove that the synthetic series of flood events have statistical properties that are very similar to the historical series. This shows that the stochastic sampling method performs well. The lengthy synthetic time series that can be generated with the stochastic model offers opportunities to provide an event loss table and detailed risk profile for various applications. The challenge of reproducing joint occurrence probabilities of ~8000 pairs of flood drivers was tackled by a novel approach based on simulated annealing (Kirckpatrick et al., 1983). One of the attractive features of this method is that multiple objective functions can be optimised simultaneously. This enabled the reproduction of several relevant statistical features of the historical time series in the synthetic time series. In this study, we have focused on population affected by flood events, but the methodology can easily be generalized to economic losses and other types of disasters.

In this paper, the objective was to generate a synthetic time series with similar statistics as the historic time series. However, the method can also be applied to create synthetic time series that account for climate change projections. It is possible to choose/design virtually any set of statistics (for example perturbing the annual maxima frequency and correlations due to climate change) and to subsequently generate a synthetic time series which will match these statistics. That potential is very valuable to the risk modelling community.

References

1. Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21, 589–615. <https://doi.org/10.5194/hess-21-589-2017>
2. Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer. ISBN 1-85233-459-2.
3. Deltares. (2018). Southeast Asia flood monitoring and risk assessment for regional DRF mechanism. Component 2 report. March 2018.
4. Diermanse, F. L. M., Geerse, C. P. M. (2012). Correlation models in flood risk analysis, reliability engineering and system safety (RESS). 64–72.
5. Fang, H., Fang, K., & Kotz, S. (2002) The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1), 1–16.
6. Holland, G. (1980). An analytical model of the wind and pressure profiles in hurricanes. *Monthly Weather Review*, 108, 1212–1218 (3, 11, 40)
7. Holland, G. (2008). A revised hurricane pressure–wind model. *Monthly Weather Review* 136, 3432–3445 (v, 17, 18)
8. Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 27(2), 179.
9. Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C., & Ward, P. J. (2016). A global reanalysis of storm surges and extreme sea levels. *Nature Communications*, 7.
10. Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of statistics*, 6(2), 461–464, <https://doi.org/10.1214/aos/1176344136>.

11. Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE*, 10, e0107042.
12. Strang, G., (1982). *Linear algebra and its applications*, San Diego: Harcourt, Brace, Jovanovich, Publishers. ISBN 0-15-551005-3