Direct Sampling for extreme weather generation

Jorge Guevara
IBM Research[*]

Maria Garcia
IBM Research

Priscilla Avegliano
IBM Research

Debora Lima
IBM Research

Dilermando Queiroz
IBM Research

Maysa Macedo
IBM Research

Daniela Szwarcman
IBM Research

Bianca Zadrozny
IBM Research

Leonardo Tizzei
IBM Research

Campbell Watson
IBM Research

Abstract

Direct Sampling is an algorithm that can generate synthetic data using only one training image and a set of conditioning points. This algorithm implicitly learns the conditional distribution of the probable values the data could take given a set of conditioning points and the training image. This algorithm does not learn an internal state, like parametric Machine Learning algorithms, but instead, it contains a pattern-matching algorithm that implicitly learns such conditional distribution. Thus, it is a non-parametric Machine learning algorithm that resembles the KNN approach. In this work, we explore the application of Direct Sampling for generating extreme precipitation events, which are precipitation weather fields with out-of-sample precipitation values. To this end, we propose to conditioning Direct Sampling not only in the training image and the conditioning points but also in a set of control points and a return precipitation level map to guide the out-of-sample precipitation value generation. We validate our approach with statistical metrics and connectivity metrics.

1 Introduction

Extreme weather analysis is important due to the effects that extreme weather events could cause on the economy, society, and the environment. Thus, there is a necessity to have tools for supporting the risk analysis of downstream tasks by considering extreme weather scenarios. In this study, we investigate the use of the generation of extreme precipitation events via the use of the Direct Sampling Algorithm, which is an algorithm that implicitely learns the conditional distribution of the probable values the data could take given a set of conditioning points and only one training image. The main advantage of this approach is that we do not have to rely on a big dataset for constructing a generative model, instead, we use a reference training image for constructing the simulations. The Direct Sampling algorithm can be understood as a non-parametric machine learning algorithm — i.e., differently that parametric machine learning algorithms that learn an internal state of the model (the model's parameters) from data, and then discard the data and only rely on the learned model parameters to do inference and generation (e.g. neural networks, linear regression, etc) — it is a model that does not have an internal state but it has an internal pattern-matching algorithm that helps the generation process, thus it strongly rely on the training image for generating new samples. It resembles the KNN approach but for data generation. In order to produce coherent (out-of-sample) extreme precipitation weather fields with Direct Sampling, we propose to use a set of control points whose locations are arbitrarily chosen and whose values are correlated with return precipitation level

[*] `jorgegd@br.ibm.com`

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

maps, i.e, a map of extreme precipitation values based on the analysis provided by Extreme Value Theory. We validate our results with a set of statistical metrics and connectivity metrics.

1.1 Direct Sampling

Direct Sampling is an algorithm that arose from the Multi-point Geostatistics community [Mariethoz et al., 2010, Mariethoz and Caers, 2014], and whose main use is to provide statistically coherent simulations with a structure mimicking the one provided by a training image. Traditional applications of direct sampling are those that require simulations coherent with a set of measurements, and coherent with the physics provided by the training image. In the weather domain, some applications of Direct Sampling are on conditional stochastic rainfall [Wojcik et al., 2009], downscaling [Jha et al., 2013, 2015], resampling extremes [Opitz et al., 2021], rainfall series generation [Benoit and Mariethoz, 2017, Oriani et al., 2014, 2018], conditional weather fields [Oriani et al., 2017].

The Direct Sampling algorithm begins by considering a Training Image (TI) which is a multi-dimensional signal providing single or multivariable information. Thus, given a TI and a simulation grid (SG), the algorithm starts by assigning conditional data points to the SG if any. The simulated value for a random location $\mathbf{x}$ in the SG is computed by doing pattern matching between the data event for $\mathbf{x}$ and all the similar data events in the TI. A data event for $\mathbf{x}$ is the set of n already simulated nodes around $\mathbf{x}$ and it is denoted by $d_n(\mathbf{x}, L) = \{Z(\mathbf{x} + \mathbf{h}_1), \ldots, Z(\mathbf{x} + \mathbf{h}_n)\}$, where L is a set of lag vectors $\{\mathbf{h}_1, \ldots, \mathbf{h_n}\}$ defining a neibourghood around $\mathbf{x}$, and $Z(\mathbf{x})$ is the simulated value for location $\mathbf{x}$. The pattern matching algorithm uses a distance function between the TI and SG data events, i.e., $D(d_n(\mathbf{x}, L), d_n(\mathbf{y}, L))$, where $\mathbf{y}$ is a random location in the TI. Once a match is found, the value of $\mathbf{y}$ in the TI is assigned to location $\mathbf{x}$ in the SG. The algorithm's stop criterion is to threshold D to a small value t or assign the lowest D value if $D \leq t$ is not satisfied.

An interesting property of this algorithm is that the pattern matching between the TI and SG data events looks for a very diverse set of structures at different scales — it scrutinizes for a myriad of patterns of different sizes without using predefined templates. Also, it allows the use of continuous and discrete variables, co-simulation, conditioning points, multi-variables, and parallel algorithms. Furthermore, the quality of the simulations depends on the quality of the TI and the parameter settings such as the distance between data events, threshold distance value, and data event size. All of that will require a first round of sensitivity analysis to calibrate the algorithm with the best set of parameters. Unfortunately, there is a positive correlation between the best parameter configuration and a high computational cost, however, some solutions based on GPU and parallel implementations exist [Huang et al., 2013, Cui et al., 2021].

1.2 Direct sampling on a target weather field conditioned on a return level map and control points

One way to generate extreme weather fields is by conditioning the Direct Sampling on a set of control points defining the locations for generating extreme weather values. In this sense, we set the control points values to be defined from the return level map values associated to a given return period. Section 2.3.1 describes how the return level map is estimated. Main advantages of this approach is that we don't need to perform quantile mappings or assume any parametric distributions and, the use of control points and return level maps to condition the generation of extreme precipitation weather values in areas of interest can be very valuable for downstream applications, e.g., flood risk analisys, risk management for disaster countermeasures, etc. We propose the following procedure:

1. Identify a target weather field $\mathbf{W}'$ in the dataset $\mathbf{W}_t, (t = 1, \ldots, T)$ with the best raking based on some criteria.
2. Set the weather field $\mathbf{W}'$ as a training image.
3. Set the conditioning data matrix matrix $\hat{\mathbf{C}}$ to the weather values of random locations within the region of interest that we do not want to generate extreme precipitation, i.e., $\{\hat{\mathbf{C}}_i = \mathbf{W}'_i\}_{i \in \hat{I}}$ and $\{\hat{\mathbf{C}}_i = \phi\}_{i \notin \hat{I}}$, where $\hat{I}$ is an index set of random locations in the spatial domain of interest.
4. Set the control points data matrix $\mathbf{C}$ to (out-of-sample) extreme weather values at locations where we want to generate extreme precipitation events. i.e., $\{\mathbf{C}_i = f_i\}_{i \in I}$ and $\{\mathbf{C}_i = \phi\}_{i \notin I}$, where f is a random process depending on a provided return precipitation level map

$\mathbf{M}$ and a specific location, that is, $f_i = f(i, \mathbf{M})$ and I is an index set of locations in the spatial domain of interest where we want to generate extreme weather. We call all the points $\mathbf{C}_i \neq \phi$ as *control points*.

5. Assign to the simulation grid $\mathbf{S}$ all the points in the conditioning and the control points data matrices, i.e, $\{\mathbf{S}_i = \hat{\mathbf{C}}_i\}_{i \in \hat{I}}$ and $\{\mathbf{S}_i = \mathbf{C}_i\}_{i \in I}$
6. Run the Direct Sampling algorithm with Training Image $\mathbf{W}'$ and Simulation grid $\mathbf{S}$, and appropriate parameters as usual but if in the process of simulating a location $\mathbf{x}$ in the simulation grid $\mathbf{S}$ it is found a value in the data event belonging to the simulation grid greater than the maximum value found in the training image, i.e., $\max d_n(\mathbf{x}, L) > \max \mathbf{W}'$, do the following update: $\mathbb{S}_{\mathbf{x}} = \mathbf{W}'_{\mathbf{y}} - \overline{d_n(\mathbf{y}, L)} + \overline{d_n(\mathbf{x}, L)}$, where $\mathbf{W}'_{\mathbf{y}}$ is a random point in the training image located at $\mathbf{y}$, the overline notation denotes de average and $d_n(\mathbf{y}, L)$ and $d_n(\mathbf{x}, L)$ are the data events of the training image $\mathbf{W}'$ and simulation grid $\mathbf{S}$, respectively.

Observe that the main purpose of conditioning data matrix $\hat{\mathbf{C}}$ is to guarantee spatial coherence and honor some values of the training image data. Also, the locations in $\hat{\mathbf{C}}$ and $\mathbf{C}$ can be selected by users or by analyzing some precipitation statistics within the region of interest. The update for $S_{\mathbf{x}}$ mentioned in step six, was mentioned in [Mariethoz et al., 2010] in the context of non-stationary distances. Figure 1 depicts the proposed pipeline showcasing the use of a 100-year return precipitation level map, a target weather field, a region of interest where extreme values will be generated, the control and the conditional points, and the generated weather filed with the extreme precipitation event.

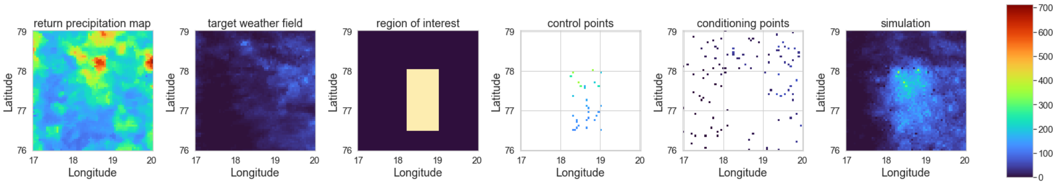

Figure 1: Generation of an extreme precipitation event using Direct sampling conditioned on a return precipitation level map via control points. The algorithm selects a target weather field as a training image and receives two external inputs: a return precipitation level map and a set of control points within a region of interest. The control points couple the locations in the target weather field with values from the return precipitation level map. Further, a set of conditioning points are used as a way to hold the non-stationarity and connectivity properties from the original target weather field. Direct Sampling uses such information to generate a simulation of an extreme precipitation weather field.

2 Experimental Evaluation

2.1 Dataset and region of interest

We used the *IMERG* dataset [Huffman et al., 2019], which contains daily precipitation values in *mm* and a spatial resolution of 0.1° × 0.1°, corresponding to approximately to 10km × 10km. We selected the precipitation data from 2001 to 2020 within a bounding box defined by the coordinates 16.2 ° N, 73.9° E, 22.2° N, 79.9° E, which roughly correspond to 360000 km^2 and contains 3600 latitude and longitude pairs. The region of interest correspond to the Maharastra state, India.

2.2 Metrics

We used a series of quantitative metrics to validate how well the simulations are reproducing the structural and statistical properties of the Training Images. We employed the following metrics to quantify the reproduction of statistical properties:

- Quantile-quantile plot between the pixel values of the training image and the simulation
- Comparison between the empirical cumulative distribution functions (eCDF) between the pixel values of the training image and the simulation

- Comparison between variograms estimated from the training image and the simulation

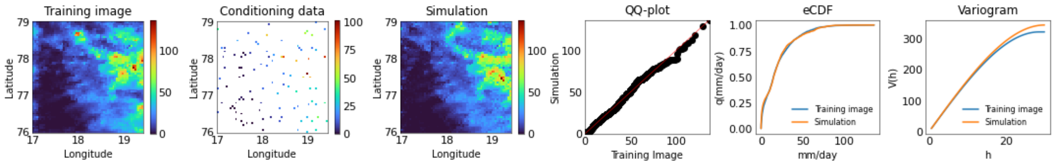

Figure 2: The quantile-quantile, eCDF, and Variogram metrics show how the simulation preserves some statistical properties of the training image: the distribution of pixel values and the variation of the standard deviation as a function of a lag.

Figure 2 shows from left to right a training image $\mathbf{W}$, the conditioning points $\mathbf{C}$, the simulation grid $\mathbf{S}$ and the quantile-quantile, eCDF, and Variogram plots. For this simulation, those metrics agree that the simulations preserve the statistical properties of the training image's pixel values.

We used the following metrics to quantify the reproduction of connectivity properties:

- Two-point probability function [Renard and Allard, 2013, Torquato et al., 1988, Torquato and Haslach Jr, 2002]— It assumes that the input is a binary image $\mathcal{I}$, and it measures the probability that, given a lag $\mathbf{h}$, two pixels located at $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ contain color one, i.e., $S_2(\mathbf{h}) = P\{\mathcal{I}(\mathbf{x}) = 1, \mathcal{I}(\mathbf{x}+\mathbf{h}) = 1\}$, observe that if $\mathbf{h} = 0$ then $S_2(\mathbf{h})$ equals to the fraction of pixels within the image with color one, i.e., $S_2(\mathbf{h}) = \mathbb{E}\{\mathcal{I}(\mathbf{x} = 1)\} = \varphi$, on the other extreme, if $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ are uncorrelated, it assumes that both points have the same probability φ, thus $S_2(\mathbf{h}) = \varphi^2$
- Two-point connectivity function [Renard and Allard, 2013, Torquato et al., 1988, Torquato and Haslach Jr, 2002] — It assumes that the input is an image containing clusters, and given a lag $\mathbf{h}$, it measures the probability that two pixels located at $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ pixels are connected, i.e., they belong to the same cluster $C(\mathbf{x}) = C(\mathbf{x}+\mathbf{h}) \neq 0$. That is, $C_2(\mathbf{h}) = P\{C(\mathbf{x}) = C(\mathbf{x}+\mathbf{h}) \neq 0\}$. Observe that if $\mathbf{h} = 0$ then $C_2(\mathbf{h}) = \varphi$ by the same arguments as beforementioned, on the other extreme, if $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ are uncorrelated, they are not connected, i.e., $C_2(\mathbf{h}) = 0$

Figure 3, shows the connectivity metrics: the two-point probability and connectivity functions. The first row from left to right shows the results of using the two-point probability function S_2 for the training image and simulation from Figure 2. In this case, the input to the procedure is a binary image where all the points of interest are labeled as one and zero otherwise. We used three criteria for creating binary images depending on the quantile range of pixel values, for instance, the first criteria aim to analyze if the simulation is reproducing the connectivity properties of the training image in regions with high precipitation values, in this case, we binarize both the training image and the simulation such the respective binary images has a value of one in the locations with high precipitation values and zero otherwise. The first row of Figure 3 shows this first criterion where we label the binary images as one in regions where the precipitation values exceed the 0.9 quantiles of pixel values. The other two criteria were to analyze if the simulation is reproducing the connectivity properties of the training image in the middle and lower precipitation values, respectively. In those cases, we provide binary images with label one in regions within 0.1 and 0.9 quantiles of pixel values (middle case) and label one in regions with less than 0.1 quantiles of pixel values (lower case). The first row of Figure 3 also shows the two-probability functions for the training image and the simulation labeled as $S_2(TI)$ and $S_2(Sim)$, respectively. Observe that the lag parameter h varies from 0 to 15 pixels, which corresponds to $0.1°$ (10km approximately). We also show the difference : $S_2(TI) - S_2(Sim)$. Furthermore, we show for completeness the profile average between the training image and the simulation, which we computed by selecting the S_2 values in the X-axis, Y-axis, XY-axis, and YX-axis, starting from the center coordinates, and averaging it out.

The second row from Figure 3 shows the results for the two-point connectivity function C_2 for the training image and simulation from Figure 2 labeled as $C_2(TI)$ and $S_2(Sim)$, respectively. The input to this procedure is an image with clusters labels in the region of interest for the analysis (A binary image describes the region of interest, the computation was likewise the S_2 metric case). The figure also shows the difference image: $C_2(TI) - S_2(Sim)$ and the profile average.

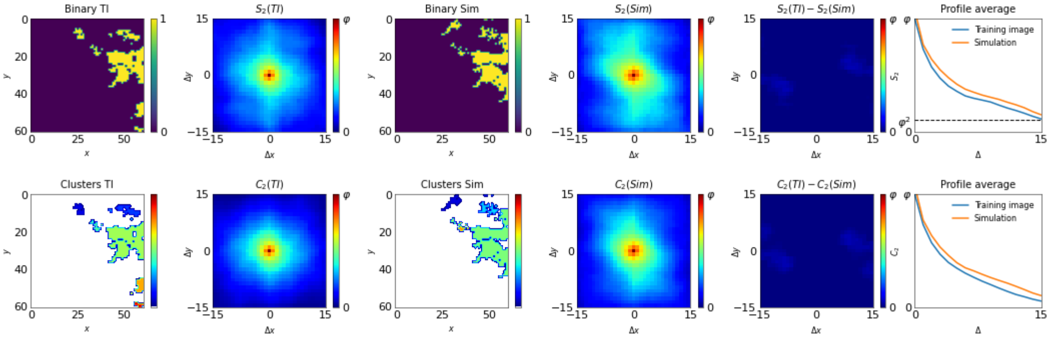

Figure 3: The two-point probability and conectivity metrics show how the simulation preserves some structural properties of the training image.

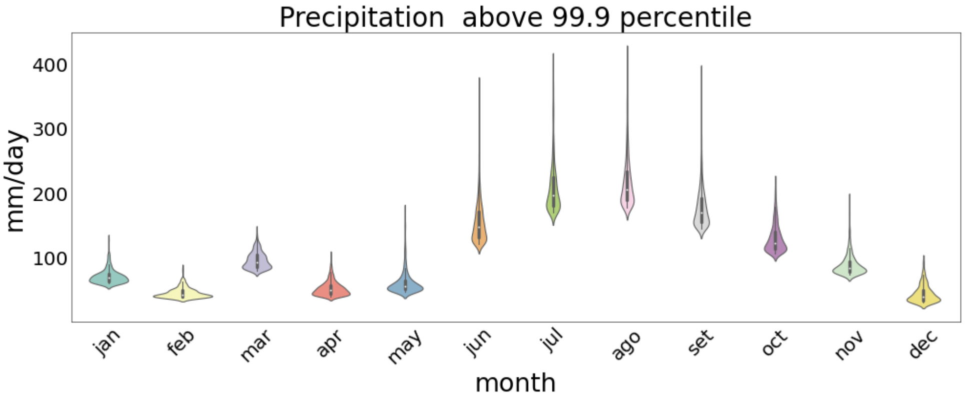

Figure 4: Distribution of the extreme daily precipitation grouped by month in the region under analysis, for the period 2001-2020.

For this simulation, those metrics agree that the simulation preserves the connectivity properties found in the training image.

2.3 Generation of precipitation weather fields with extreme events

We generated in this experiment precipitation weather fields with extreme events. We also estimated the return precipitation level maps for 100, 250, and 500-year return period events. Figure 4 shows the empirical distribution of daily precipitation values above the 99.9 percentile and grouped by month for the period 2001-2020 — observe that historical extreme precipitation values surpass the 400 mm/day for July and August months, which correspond to the Moonson period in central India.

2.3.1 Return precipitation level map estimation

The return period can be seen as a risk measure [Brunner et al., 2016] widely adopted in the climate domain because it expresses the likelihood of extreme events and consequently the likelihood of failures and large losses. The mth return period estimates the magnitude of the phenomena whose probability of exceedance in one year is equal to $\frac{1}{m}$ [Vogel and Castellarin, 2017].

For safety and structural analysis purposes, frequently we should rely on events whose return periods are larger than the period of available data, making the empirical estimation of such events impractical. Extreme Value Theory [De Haan and Ferreira, 2007] provides the theoretical statistical tools to perform such extrapolation to calculate higher quantiles based on a limited set of samples. Our approach is based on Extreme Value Theory to generate the return levels maps of 100, 250, and 500 years return periods, as we had only 20 years of available data.

We opted for the block-maxima sampling approach, collecting the most extreme precipitation events of each year and then using these samples to calibrate the parameters of a Generalized Extreme Value distribution (GEV) with the Maximum Likelihood method. Figure 5 shows the resulting return precipitation level maps for the region of interest for 100, 250, and 500-year return period events.

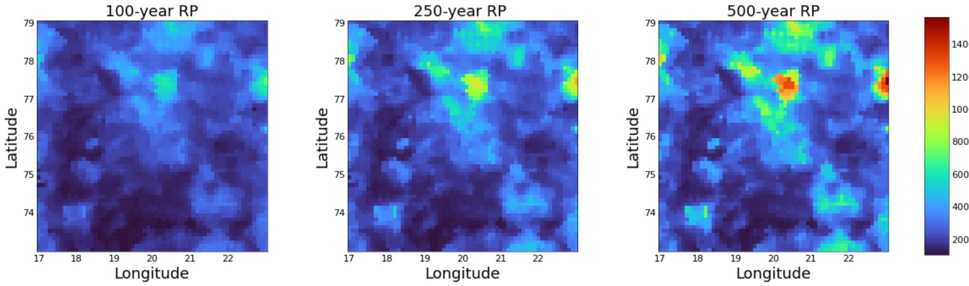

Figure 5: Return precipitation level maps for 100, 250 and 500 years of return periods.

2.3.2 Conditioning the Direct Sampling on a return precipitation level map and control points

In this section, we show the use of the Direct Sampling conditioned on control points and return level maps as described in Section 1.2 to produce extreme precipitation weather fields. Figure 6 shows the generation of extreme precipitation weather fields: the first-left column of images are the target weather fields, in this case, we arbitrarily selected those figures. Each row of Figure 6 contains the results of conditioning the Direct Sampling on the three precipitation level maps — 100, 250, and 500 -year return precipitation levels — depicted in Figure 5. The yellow areas in the second, fourth, and sixth columns of images depict the ROIs: a rectangle, two rectangles, and a more complex shape estimated by selecting the areas from the original target image with more than 0.9-quantiles of precipitation values. Each ROI contains a set of control points, in practice, users can define arbitrary locations for those control points within the ROI. In the case of Figure 6 the control points are randomly located. Also, the value that each control point could take is defined by f_i (Section 1.2). In this experiment, we defined f_i by sampling a value uniformly distributed from the interval given by the maximum value presented in the training image, and the return precipitation level map value for the location of the control point. Of course, other configurations are possible for instance sampling from a normal with a location parameter given by the mean of the beforementioned interval, and a standard deviation related to the full width, and half maximum of the normal distribution, i.e., $\sigma = 2\sqrt{2\log 2}$, or even control points can take directly its values from the return precipitation level map. Those cases are depicted in Figure 7, the first row shows the case of f_i defined by sampling from a normal distribution, and the second row, the case where f_i takes the values directly from the return precipitation level map.

Observe that the statistical and structural properties of the resulting simulation in the region outside the ROI will be correlated with those in the training image. To analyze how the extreme generation alters the statistical and connectivity metrics of the original training image, we show in Figure 8 the results of comparing one-hundred simulations vs. the original statistical and connectivity properties of the training image. The qq-plot and eCDF metrics inform how the distribution of precipitation values is shifted upwards, reflecting the fact that the simulation contains out-of-sample precipitation values. The variogram shows an increase in the standard deviation as a function of lag, this alteration is due to the ROI and the extreme precipitation values generated. The connectivity metrics show that the connectivity properties of the middle precipitation values are well preserved by the simulation. It shows an increase in the connectivity metric values for high precipitation values, i.e., the probability of having two random points within the area of high precipitation is bigger, and there is a decrease in the probability of connectivity of low precipitation values, that makes sense because the algorithm is increasing the precipitation values within the ROI. We show also the mean and the standard deviation of the simulations on the top row of the Figure.

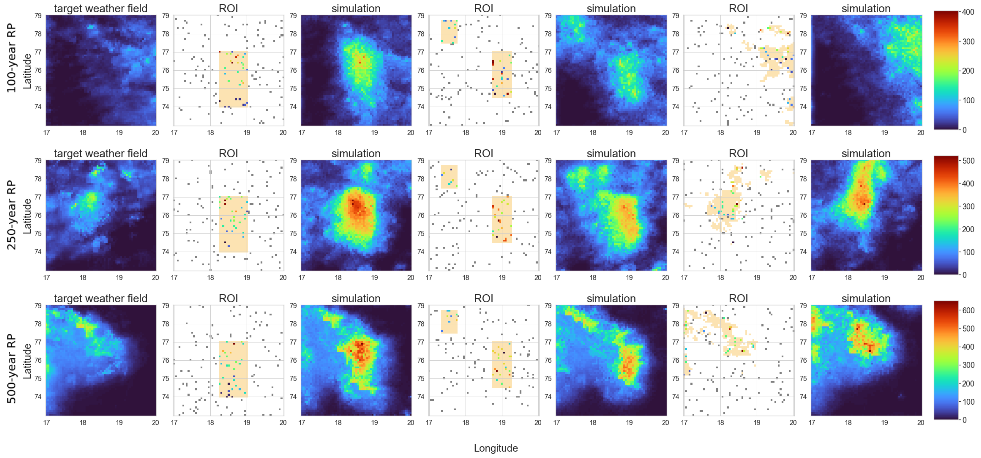

Figure 6: Extreme precipitation generation using the Direct Sampling and control points located in the yellow area of the ROI (region of interest) image, conditioned on a 100-year return precipitation level map.

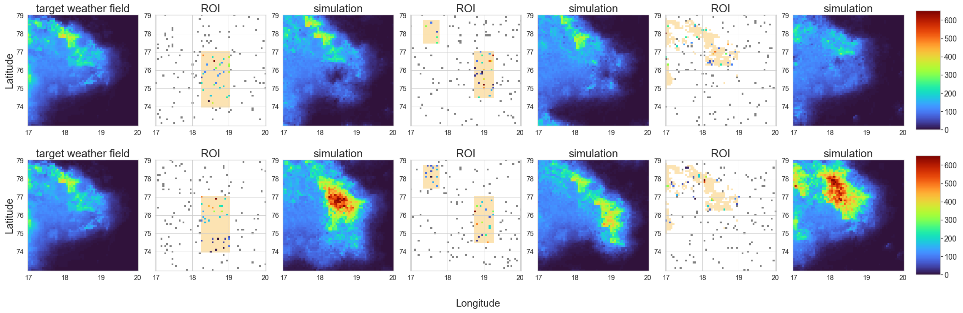

Figure 7: Extreme precipitation generation using the Direct Sampling and control points taking its values from a normal distribution or directly from the return precipitation level map

3 Conclusions

We presented how the Direct Sampling algorithm — an algorithm for data generation that can be cast as a non-parametric ML generative model — can be used for extreme precipitation generation. For this, we condition the generation process of Direct Sampling on a set of control points and return precipitation level maps. The set of control points can be defined arbitrarily by users or applications, and the return precipitation level map was estimated using Extreme Value Theory. We validated our approach with the IMERG precipitation dataset and a set of statistical (quantile-quantile, empirical cumulative distribution function, and variogram) and connectivity (two-point probability and connectivity)metrics. Future work includes solving the problem of extreme weather generation but including several weather variables at the same time and embedding the time as an additional variable.

References

L. Benoit and G. Mariethoz. Generating synthetic rainfall with geostatistical simulations. *Wiley Interdisciplinary Reviews: Water*, 4(2):e1199, 2017.

M. I. Brunner, J. Seibert, and A.-C. Favre. Bivariate return periods and their importance for flood peak and volume estimation. *Wiley Interdisciplinary Reviews: Water*, 3(6):819–833, 2016.

Z. Cui, Q. Chen, G. Liu, G. Mariethoz, and X. Ma. Hybrid parallel framework for multiple-point geostatistics on tianhe-2: A robust solution for large-scale simulation. *Computers & Geosciences*,

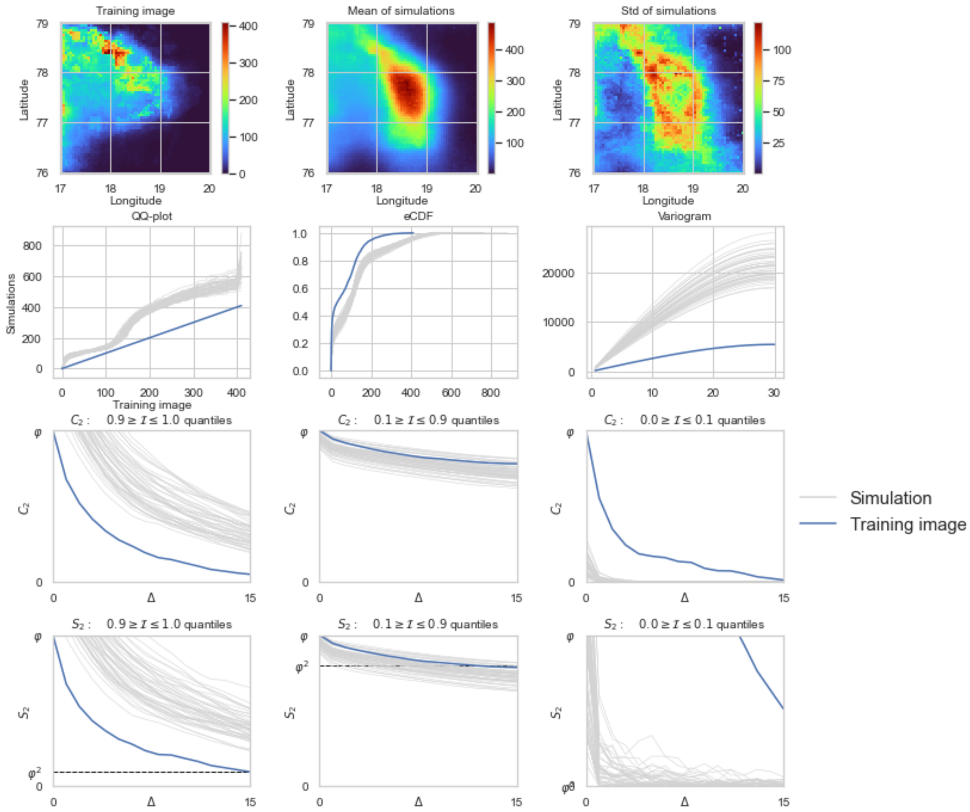

Figure 8:

157:104923, 2021.

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

T. Huang, X. Li, T. Zhang, and D.-T. Lu. Gpu-accelerated direct sampling method for multiple-point statistical simulation. *Computers & Geosciences*, 57:13–23, 2013.

G. Huffman, E. Stocker, D. Bolvin, E. Nelkin, and J. Tan. Gpm imerg final precipitation l3 1 day 0.1 degree x 0.1 degree v06, edited by andrey savtchenko, greenbelt, md, goddard earth sciences data and information services center (ges disc). *doi. org/10.5067/GPM/IMERG DF/DAY/06*, 2019.

S. K. Jha, G. Mariethoz, J. P. Evans, and M. F. McCabe. Demonstration of a geostatistical approach to physically consistent downscaling of climate modeling simulations. *Water Resources Research*, 49(1):245–259, 2013.

S. K. Jha, G. Mariethoz, J. Evans, M. F. McCabe, and A. Sharma. A space and time scale-dependent nonlinear geostatistical approach for downscaling daily precipitation and temperature. *Water Resources Research*, 51(8):6244–6261, 2015.

G. Mariethoz and J. Caers. *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons, 2014.

G. Mariethoz, P. Renard, and J. Straubhaar. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46(11), 2010.

T. Opitz, D. Allard, and G. Mariethoz. Semi-parametric resampling with extremes. *Spatial Statistics*, 42:100445, 2021.

F. Oriani, J. Straubhaar, P. Renard, and G. Mariethoz. Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrology and Earth System Sciences*, 18(8): 3015–3031, 2014.

F. Oriani, N. Ohana-Levi, F. Marra, J. Straubhaar, G. Mariethoz, P. Renard, A. Karnieli, and E. Morin. Simulating small-scale rainfall fields conditioned by weather state and elevation: A data-driven approach based on rainfall radar images. *Water Resources Research*, 53(10):8512–8532, 2017.

F. Oriani, R. Mehrotra, G. Mariethoz, J. Straubhaar, A. Sharma, and P. Renard. Simulating rainfall time-series: how to account for statistical variability at multiple scales? *Stochastic environmental research and risk assessment*, 32(2):321–340, 2018.

P. Renard and D. Allard. Connectivity metrics for subsurface flow and transport. *Advances in Water Resources*, 51:168–196, 2013.

S. Torquato and H. Haslach Jr. Random heterogeneous materials: microstructure and macroscopic properties. *Appl. Mech. Rev.*, 55(4):B62–B63, 2002.

S. Torquato, J. Beasley, and Y. Chiew. Two-point cluster function for continuum percolation. *The Journal of chemical physics*, 88(10):6540–6547, 1988.

R. M. Vogel and A. Castellarin. Risk, reliability, and return periods and hydrologic design. *Handbook of Applied Hydrology; Singh, VP, Ed.; McGraw-Hill Book Company: New York, NY, USA*, 2017.

R. Wojcik, D. McLaughlin, A. G. Konings, and D. Entekhabi. Conditioning stochastic rainfall replicates on remote sensing data. *IEEE transactions on geoscience and remote sensing*, 47(8): 2436–2449, 2009.