Stochastic simulation of rainfall and climate variables using the direct sampling technique

thesis presented at the

Faculty of Sciences

Center for Hydrogeology and Geothermics (CHYN)

University of Neuchâtel, Switzerland

for the degree of

Doctor of sciences

presented by

Fabio Oriani

accepted in the recommendation of

Prof. Philippe Renard
Dr. Julien Straubhaar
Prof. Grégoire Mariethoz
Dr. Denis Allard
Prof. Alexis Berne
Prof. Olivia Romppainen-Martius

Defended on September 17^{th}, 2015

UNIVERSITÉ DE NEUCHÂTEL

Faculté des sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel - Suisse
Tél: + 41 (0)32 718 2100
E-mail: secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Fabio ORIANI

Titre:

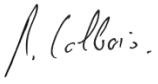
“Stochastic simulation of rainfall and climate variables using the Direct Sampling technique”

sur le rapport des membres du jury composé comme suit:

- Prof. Philippe Renard, Université de Neuchâtel, directeur de thèse
- Dr Julien Straubhaar, Université de Neuchâtel
- Prof. Grégoire Mariethoz, Université de Lausanne
- Prof. Alexis Berne, EPF Lausanne
- Dr Denis Allard, INRA, Avignon, France
- Prof. Olivia Romppainen-Martius, Université de Berne

Neuchâtel, le 19 octobre 2015

Le Doyen, Prof. B. Colbois



Acknowledgements

The research work for this thesis was funded by the Swiss National Science Foundation (project #134614) and supported by the National Centre for Groundwater Research and Training (University of New South Wales). The data sets used are courtesy of Meteoswiss, the Swiss Environmental Federal Office (OFEV), the Australian Bureau of Meteorology (BOM), the Israel Meteorological Service (IMS) and Survey of Israel (SOI).

Fabio would like to thank...

My supervisor Prof. Philippe Renard, my co-supervisors Prof. Grégoire Mariethoz and Dr. Julien Straubhaar: a big thanks for your efforts, your brilliant ideas and absolute willingness.

The other colleagues that have contributed to this research work (uniformly distributed in a random order!): Dr. Raj Mehrotra, Prof. Ashish Sharma, Dr. Andrea Borghi, Noa Ohana-Levi, Dr. Francesco Marra, Dr. Amir Givati, Prof. Efrat Morin, Prof. Arnon Karnieli, Dr. Simon Stisen, Dr. Christian Moeck, Jérémie Chappuis, Louis Fournier, Prof. Reinhard Furrer. The other members of the thesis examination board for their precious advice: Prof. Olivia Romppainen-Martius, Dr. Denis Allard, Prof. Alixis Berne. The colleagues from the University of New South Wales: Prof. Bryce Kelly, Prof. Andy Baker, Dr. Sanjeev Jha. The colleagues and friends of the Ensemble Project: Emanuel, Peter, Cline, Niklas, Bastien, Tobias, Jef, David, Laureline, Ivan. The guys from Philippe's stochastic group: Andrea and Alessandro (my mentors in the art of programming), Amine, Axa, Cécile, Christoph, Cybèle, Damian, Guillaume, Jaouher, Martin, Philip. All the friends from the CHYN for their everyday support and company. Maiann and Valeria for the wonderful period passed together as house mates in Neuchâtel. Linda and Ale for having kindly hosted me during my stay in Sydney. The musical friends from Azyne. The guys from the B3 Jazz Orchestra. My friends from Neuchâtel, Lausanne and from Italy, in particular the GPL crew and Thomas.

Last but not the least, all my family, in particular my parents Silvana and Angelo and my girlfriend Greta: without your support this would not have been possible.

My heartfelt thanks to all of you.

Keywords

stochastic, simulation, rainfall, climate, multiple-point, geostatistics, non-parametric.

Mots-clés

stochastique, simulation, pluie, climat, multi-point, géostatistique, non-paramétrique.

Abstract

An accurate statistical representation of hydrological processes is of paramount importance to evaluate the uncertainty of the present scenario and make reliable predictions in a changing climate. A wealth of historic data has been made available in the last decades, including a consistent amount of remote sensing imagery describing the spatio-temporal nature of climatic and hydrological processes. The statistics based on such data are quite robust and reliable. However, to explore their variability, most stochastic simulation methods are based on low-order statistics that can only represent the heterogeneity up to a certain degree of complexity.

In the recent years, the stochastic hydrogeology group of the University of Neuchâtel has developed a multiple-point simulation method called Direct Sampling (DS). DS is a resampling technique that allows the preservation of the complex data structure by simply generating data patterns similar to the ones found in the historical data set. Contrarily to the other multiple-point methods, DS can simulate either categorical or continuous variables, or a combination of both in a multivariate framework.

In this thesis, the DS algorithm is adapted to the simulation of rainfall and climate variables in both time and space. The developed stochastic weather or climate generators include the simulation of the target variable with a series of auxiliary variables describing some aspects of the complex statistical structure characterizing the simulated process. These methods are tested on real application cases including the simulation of rainfall time-series from different climates, the variability exploration of future climate change scenarios, the missing data simulation within flow rate time-series and the simulation of spatial rainfall fields at different scales. If a representative training data set is used, the proposed methodologies can generate realistic simulations, preserving fairly well the statistical properties of the heterogeneity. Moreover, these techniques result to be practical simulation tools, since they are adaptive to different data sets with minimal effort from the user perspective. Although leaving large room for improvement, the proposed simulation approaches show a good potential to explore the variability of complex hydrological processes without the need of a complex statistical model.

Contents

Chapter 1

Introduction

For thousands years, man has been trying to control and make predictions about Nature. Logical knowledge lead us to identify different natural processes and link them by mean of observational models based on repetitiveness of the phenomena and cause-and-effect relationships. The aim of this approach is mainly to make forecasts in practical situations to prevent dangerous consequences and adapt to future scenarios. This thesis is about the development of tools to make predictions in the domain of hydrology, regarding the quantitative representation of rainfall and some climate variables. These phenomena are characterized using mathematical variables, i.e. abstract objects that represent with numbers some specific aspect of a natural process. For example, the rain amount fallen at one location in a certain period, the mean measured temperature or humidity, etc. All these objects offer, from a limited perspective, a sharp description of a phenomenon, containing sometimes useful information to make future predictions.

1.1 Motivation

One main topic of current research in hydrology is the quantification of rainfall and climate variables in space and time. The hydro-meteorological processes exhibit non-stationarity and a complex distribution in time and space which is difficult to represent with traditional parametric models that can handle complexity only up to a certain degree. Moreover, the interaction between climate and the basin heterogeneity (land coverage, soil moisture, geology) can lead to an extremely variable hydrological and hydrogeological response [216, 219, 50, 11, 197, 170, 71], therefore a faithful representation of spatio-temporal patterns at the sub-catchment scale in consistency with larger scales is needed. Finally, all these processes are occurring in a changing climate, for which non-stationarity and anomaly are on the agenda for the next decades. Handling this complexity requires a representative amount of data and a convenient way to extract the relevant information, make the required predictions and quantify their uncertainty. Today more than ever, we need prediction tools that should be reliable, adaptive to different data sets and easily applicable on real cases.

1.2 Current state of the research

The main objective of rainfall quantification techniques is to generate a field or an ensemble of fields representing rainfall and related variables in space and time, used as input for hydrological models or distributed (spatial) models of other type. Methods to represent the spatio-temporal distributions of these variables can be broadly divided in three categories: i) techniques based on the interpolation and analysis of available data, ii) physically-based simulation methods and iii) stochastic simulation methods.

The first category involves the use of ground measurements and remote sensing products as input variables for hydrological models. In the most simple and approximated form, the hydrological model can be reduced to a one-dimensional problem: one or more equations representing the hydrological process of the overall basin. In this case, the climatic and meteorological input is usually a time-series representative of the historical climate record of the basin. These models, called lumped, are sometimes a convenient solution allowing approximate predictions with minimal effort that are in some cases comparable to more complex models [41, 156, 204]. Conversely, when a more detailed spatial description of the hydro-climatic processes is required, e.g. as input for distributed hydrological models, the punctual measurements of climate stations are usually interpolated to achieve a spatial distribution of the variable of interest. Today, this operation is done with the help of radar and satellite products, the latter being available from several international missions (for example TRMM [178], CHMORPH [86] or PERSIANN [75]). Remote sensing imagery gives

information about the spatial structure of rainfall, but it needs to be corrected using ground measurements. This correction called data merging is generally based on geostatistical techniques [101, 39, 13, 72, 171, 60, 202, 172, 179, 205, 177]. This type of approach, originally developed to model the heterogeneity as a multiGaussian field, is limited to first- or second-order moments, which are insufficient to describe the rainfall heterogeneity [see e.g. 17]. As a consequence, the spatial interpolation of the error shows an overly simple structure. In addition, since the spatial density of ground measurements is insufficient to represent the small-scale heterogeneity of the variable of interest, these corrections/estimations usually show very large uncertainty. In conclusion, even the most recent products often present local underestimation of the variability or consistent bias [140, 224, 99, 154, 150]. Combined data sets of this kind are the standard products available from governmental institutions and international data sets, still remaining the most used today. To overcome the limited information given by the available data, physical and stochastic simulation methods are often used to obtain a more efficient of uncertainty estimation related to rainfall and climate processes.

The second category comprises the Physically-based simulation methods, that generate synthetic fields by solving equations based on mass, momentum and energy conservation (see for example the widely used WRF climate model [184]) and obtain different realizations by perturbing the initial and/or boundary conditions. Although this is the classical method for weather prediction which can give a reliable representation of the mean behavior of rainfall, this kind of models often under-represent the local variability [27, 227, 10, 25] that can significantly affect the hydrological basin response in terms of water balance and output discharge. Moreover, the computational demand of such approaches for solving coupled physical equations and the complexity of the parameter calibration make it impractical to generate multiple synthetic rainfall outputs for uncertainty quantification. For these reasons, stochastic approaches are often preferred.

The third category is constituted by stochastic approaches that mimic the observed space-time structure of the target variable without simulating the physical processes responsible for this structure. This is achieved by generating random values from a probabilistic model based on the statistical information extracted from the available historical data. One advantage of stochastic methods is to provide error bounds on the prediction, which is crucial for risk analysis [116]. Ideally, the newly generated rainfall should show the same spatial and temporal variations as the historical data. In addition, the statistical technique should include the possibility of being conditioned to other measured variables which influence the target one, for example elevation in the case of rainfall. This way, the variability of the generated heterogeneity would be reduced to a range of possible values and space structures relative to the specific situation described by the auxiliary variables. There exist two main types of statistical techniques: parametric and non-parametric.

Parametric techniques are based on a statistical law defined a priori, which is adapted to historical data by calibrating its parameters. Each spatial value is seen as the outcome of a random variable responding to a statistical law. Although being based on rigorous probabilistic laws, the representation of the natural processes given by these models is often oversimplified and additional complexity has to be added to the model structure with the aim of generating more realistic simulations. For example, one has to use nested models for different space and time scales, and often a combination of different methods that need to be customized for the application to rainfall simulation [36, 59]. The intermittent behavior of rainfall (rain/no-rain) and the very asymmetrical probability distribution of the rainfall amount are two main issues that brought to the formulation of more complex statistical models. Rainfall intensity is often modeled with a gamma distribution [79]. Some techniques are based on Markov chains to simulate the succession of wet and dry days [214]. Others

simulate the precipitation occurrence as a Poisson process [160, 69]. Abundant work has been published on the multi-scale nature of rainfall, with multi-fractal [119, 118, 123, 43, 44] and multiplicative cascade models [144, 145, 58]. Other methods for reproduction of multi-scale precipitations use wavelets [104, 102] and clustered point processes [38, 103, 199] or censored Gaussian models [5, 201, 98, 3]. The problem of representing the spatial rainfall is usually approached using geostatistics, which aim at analyzing the covariance between couple of points at different distance (two-point statistics) for generating spatially distributed synthetic fields. Geostatistics offers a well-established multivariate framework that allows obtaining rainfall data conditioned to different attributes such as elevation or wind (see for example [57, 83, 105]) and rainfall measurements [112]. This approach is also used in the treatment of rainfall data such as data homogenization [37], reconstruction of missing sequences [52] and data interpolation [65, 133]. However, the multiGaussian assumption underpinning most of the geostatistical techniques does not hold for rainfall. As a consequence, the complexity of the simulated spatial structure is limited even for the most advanced of these methods [see e.g. 29].

An alternative research direction widely developed in the recent decades regards the family of statistical non-parametric techniques, which do not impose a prior probabilistic model, but they use a generation technique which is mainly data-driven. Such statistics can for example take the form of kernel functions [110, 125] or Markov processes [190, 128]. One type of non- or semi-parametric algorithm are resampling techniques [223, 110, 109, 155, 24, 217, 33], that sample with replacement the historical data set conditionally to the recent past simulated values. These methods can represent the fundamental characteristic of the heterogeneity and, in their most recent versions [e.g. 127], the long-term behavior of rainfall [127]. Nevertheless, they are limited to low-order statistical dependencies on some specifically chosen low-frequency of fluctuation, therefore they require an accurate data analysis to adapt the prior statistical dependence model to the current data set.

1.3 Multiple-point statistics: the world is not multiGaussian, but it reveals patterns!

A family of non-parametric techniques devoted to spatial simulation is multiple-point statistics (MPS), which aims at inferring the probability of occurrence of an event at one location (for example the occurrence of rainfall) by considering complex data patterns in its neighborhood (conditional probability of occurrence). The computation of this high-order statistical measure is based on the analysis of the training data set (called Training Image): a data set or a conceptual model which is representative of the heterogeneity that is simulated. MPS is considered as the non-parametric evolution of two-point statistics, allowing more realistic simulations than its predecessor [196, 228, 4, 12, 73, 193, 198]. The main limit of MPS techniques is that they require a sufficiently complete and reliable picture of the heterogeneity: for example, this is not available for mall-scale spatial rainfall, since ground measurements are only sparse data sets. To our knowledge, only one work in literature used MPS for spatial meteorological data simulation, in particular satellite rainfall images [218]. However, because it uses traditional MPS techniques, this approach can only simulate events as a binary variable (rain/no-rain). MPS has been also applied to downscaling of different climate variables using a collection of images generated from a small-scale climate model [81], with the possibility of increasing the spatial field resolution, merging different data set and simulating missing data. The method proposes a series of steps that comprise the generation of the rainfall occurrence clusters and their population with rainfall intensities using traditional multi-scale methods. Although the method gives realistic results, it does not use high-order statistics from the training image to simulate the rainfall intensity. With

MPS, the idea of reproducing realistic data patterns seems to be a convenient way to make predictions about a natural process without the need of an overly complex statistical model. But how to extend this approach to continuous variables like rainfall amount?

1.4 The Direct Sampling method: a convenient approach to rainfall simulation?

In the recent years, the stochastic hydrogeology group of the University of Neuchâtel has developed a non-parametric MPS simulation method called Direct Sampling (DS). Originally developed to simulate geological heterogeneity [122, 121], it is then applied to the simulation of environmental variables [82] and missing data reconstruction [121, 120]. This technique brings a step forward the multiple-point concept: to make a simulation, instead of evaluating the conditional probability distribution of the variable for a catalog of possible conditioning patterns and then drawing a random value from it, DS directly samples the training data set where a sufficiently similar pattern occurs. This implies an intensive random scan of the training data set, but avoids the need of computing any probability measure and extends the application of MPS to continuous variables. Another new fundamental feature introduced with DS is that the simulated values are generated in an absolute random order. Consequently, the conditioning pattern formed by the already simulated data vary during the simulation from large-scale sparse to small-scale dense neighbor patterns. This principle is in contrast with the main time-series modeling tradition, where simulation is causal: i.e. the simulation path follows the temporal sequence and it is conditioned by recent past values at series of fixed time lags. The strong point of DS, which makes it a potentially efficient algorithm for the simulation of rainfall and climate variables, is the capability of preserving high-order statistical dependencies with minimal parameterization and relying on data observations with no prior statistical model imposed. Nevertheless, this technique has to be appropriately adapted to the domain of hydrology. In fact, there are some main statistical aspects about hydrological processes, for example the recurrence of extremes, that were not a primary goal for previous DS applications, but they will be decisive in this research work.

1.5 Objectives

The aim of this thesis is to explore the potential of DS applied to the simulation of rainfall, climate and hydrological variables in time and space. The research work is mainly oriented to the development of operative tools directly applicable to real cases. The choice of the treated problems, the simulated variables and the spatio-temporal scales is done according to the common needs of application and ranges from daily time-series to spatial high-resolution fields, including stationary simulations and future predictions. The weak and strong points of the developed methodologies are put in evidence through different tests on real and synthetic data sets and, where possible, the algorithm is compared to other latest-generation algorithms from the domain of hydrology.

1.6 Thesis structure

Each of the following chapters presents one specific application of the DS algorithm and, even if linked to the others, it can be read separately.

Chapter 2

In this chapter, a standard setup for daily rainfall time-series simulation is presented and tested on data sets from three different climatic regions of Australia. A detailed description of the DS algorithm is given, with an in-depth discussion about its main parameters and a comparison with other resampling techniques. This chapter is based on the published paper by F. Oriani[1], J. Straubhaar[1], P. Renard[1] and G. Mariethoz[3] *Simulation of rainfall time series from different climatic regions using the direct sampling technique* [143].

Chapter 3

This chapter, contains a comparison between the time-series simulation technique developed in chapter 2 and one last-generation Markov-chain based technique, which is one of the traditional approaches to rainfall time-series simulation. The test is done using a synthetic signal, revealing the strong and weak points of the two methods and suggesting possible improvements. This chapter is based on the paper by F. Oriani[1], R. Mehrotra[2], G. Mariethoz[3], J. Straubhaar[1], A. Sharma[2] and P. Renard[1] *Simulating the complexity of rainfall time-series: a comparison between the Markov-chain approach and multiple-point statistics* in preparation.

Chapter 4

In this chapter, a DS setup for missing data simulation of high-resolution flow rate time-series is presented and tested using a historical data set of karst sources. The performance is analyzed using different missing data scenarios and auxiliary information. This chapter is based on the paper by F. Oriani[1], A. Borghi[4], J. Straubhaar[1], G. Mariethoz[3] and P. Renard *Missing data simulation inside flow rate time-series using the direct sampling technique* in preparation.

Chapter 5

In this chapter, DS is used to make a stationary multivariate simulation of daily climate variables including temperature, humidity and global radiation. The technique is then integrated in a framework to assess the uncertainty of future climate projections given by a simple delta-change approach. This chapter is based on the poster presentation by F. Oriani[1], J. Straubhaar[1], P. Renard[1], G. Mariethoz[3] *Modeling future climatic time-series according to climate change scenarios* [142].

Chapter 6

In this chapter, a DS setup for the simulation of spatial daily rainfall fields using radar images as training data set is presented. The methodology, which includes conditioning using elevation, is tested on different groups of images from the coastal and central mountainous region of Israel. This chapter is based on the research work in collaboration with Noa Ohana-Levi[5], J. Straubhaar[1], Amir Givati[6], P. Renard[1], Arnon Karnieli[5], G. Mariethoz[3], Efrat Morin[7] and Francesco Marra[7].

Chapter 7

This chapter presents the early development of a methodology based on DS to simulate high-resolution rainfall spatial fields using time-series and satellite images. A preliminary test is performed using a synthetic spatio-temporal rainfall field. This chapter is based on the poster presentation by F. Oriani[1], J. Straubhaar[1], G. Mariethoz[3] and P. Renard[1] *Spatial Rainfall Simulation: Trading Time for Space with Multiple Point Statistics* [141].

Chapter 8

This chapter contains the concluding remarks about all the research work and a brief discussion about future perspectives.

Appendices

In this last part of the thesis, some content related to the subjects treated in the previous chapters is attached under form of appendix:

- appendix A shows the result of a sensitivity analysis of some of the main DS parameters applied to daily rainfall time-series simulation;
- appendix B presents the results of the preliminary tests done for the formulation of the DS setup for daily rainfall time-series;
- appendix C shows some preprocessing treatment techniques for time-series, applied in the course of the thesis;
- in appendix D, the results of a preliminary test on the simulation using multiple training data sets, an optional feature of DS, is shown;
- appendix E contains some notes on the type of deformation applied to the reference rainfall field used in chapter 7.

[1]Centre for Hydrogeology and Geothermics, Université de Neuchâtel, Neuchâtel, Switzerland.
[2]School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia.
[3]Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland.
[4]École Nationale Supérieure de Géologie, Vandoeuvre-lès-Nancy, France.
[5]Jacob Blaustein Institutes of Desert Research, Ben Gurion University of the Negev, Midreshet Ben-Gurion, Israel.
[6]Water Authority, Israeli Hydrology Service, Jerusalem, Israel.
[7]Department of Geography, Hebrew University of Jerusalem, Jerusalem, Israel.

Chapter 2

Simulation of rainfall time-series from different climatic regions

Abstract

The Direct Sampling technique, belonging to the family of multiple-point statistics, is proposed as a non-parametric alternative to the classical autoregressive and Markov-chain based models for daily rainfall time-series simulation. The algorithm makes use of the patterns contained inside the training image (the past rainfall record) to reproduce the complexity of the signal without inferring its prior statistical model: the time-series is simulated by sampling the training data set where a sufficiently similar neighborhood exists. The advantage of this approach is the capability of simulating complex statistical relations by respecting the similarity of the patterns at different scales. The technique is applied to daily rainfall records from different climate settings, using a standard setup and without performing any optimization of the parameters. The results show that the overall statistics as well as the dry/wet spells patterns are simulated accurately. Also the extremes at the higher temporal scale are reproduced adequately, reducing the well known problem of over-dispersion.

2.1 Introduction

The stochastic generation of rainfall time-series is a key topic for hydrological and climate science applications: the challenge is to simulate a synthetic signal honoring the high-order statistics observed in the historical record, respecting the seasonality and persistence from the daily to the higher temporal scales. Among the different proposed techniques, exhaustively reviewed by [175], the most commonly adopted approach to the problem since the '60 is the Markov-chain (MC) simulation: in its classical form, it is a linear model which cannot simulate the variability and persistence at different scales. Solutions to deal with this limitation consist of introducing exogenous climatic variables and large-scale circulation indexes [70, 18, 89, 220, 77, 208, 212, 93, 78], lower-frequency daily rainfall covariates [213, 23, 84, 90] or an index based on the short-term daily historical or previously generated record [68, 67, 127, 126] as conditioning variables for the estimation of the MC parameters. By doing this, non-linearity is introduced in the prior model, the MC parameters changing in time as a function of some specific low-frequency fluctuations. An alternative proposed method is model nesting [209, 187, 188, 189], that implies the correction of the generated daily rainfall using a multiplicative factor to compensate the bias in the higher-scale statistics. These techniques generally allow a better reproduction of the statistics up to the annual scale, but they imply the estimation of a more complex prior model and cannot completely capture a complex dependence structure.

In this chapter, we propose the use of some lower-frequency covariates of daily rainfall in a completely unusual framework: the Direct Sampling (DS) technique [122], which belongs to multiple-point statistics (MPS). Introduced by [61] and widely developed during the last decade [196, 4, 228, 12, 73, 195, 198], MPS is a family of geostatistical techniques widely used in spatial data simulations and particularly suited to pattern reproduction. MPS algorithms use a training image, i.e. a data set to evaluate the probability distribution (pdf) of the variable simulated at each point (in time or space), conditionally to the values present in its neighborhood. In the particular case of the Direct Sampling, the concept of training image is taken to the limit by avoiding the computation of the conditional pdf and making a random sampling of the historical data set where a pattern similar to the conditioning data is found. If the training data set is representative enough, these techniques can easily reproduce high-order statistics of complex natural processes at different scales. MPS has already been successfully applied to the simulation of spatial rainfall occurrence patterns [218]. In this chapter, we test Direct Sampling on the simulation of daily rainfall time-series. The aim is to reproduce the complexity of the rainfall signal up to the decennial scale, simulating

the occurrence and the amount at the same time with the aid of a multivariate data set. Similar algorithms performing a multivariate simulation had been previously developed by [223] and [155] using a bootstrap-based approach. As discussed in details in section 2.2.3, the advantage of Direct Sampling with respect to the mentioned techniques is the possibility to have a variable high-order time-dependence, without incurring excessive computation since the estimation of the n-dimensional conditional pdf is not needed. Moreover, we propose a standard setup for rainfall simulation: an ensemble of auxiliary variables and fixed values for the main parameters required by the Direct Sampling algorithm, suitable for the simulation of any stationary rainfall time-series, without the need of calibration. The technique is tested on three time-series from different climatic regions of Australia. The chapter is organized as follows: in section 2.2 the DS technique is introduced and compared with the existing resampling techniques. The data set used, the proposed setup and the method of evaluation are described in section 2.3. The statistical analysis of the simulated time-series is presented and discussed in section 2.4 and section 2.5 is dedicated to the conclusions.

2.2 Methodology

In this section we recall the basics of multiple-point statistics and we focus on the Direct Sampling algorithm. The data set used is then presented as well as the methods of evaluation.

2.2.1 Background on multiple-point statistics

Before entering in the details of the DS algorithm, let us introduce some common elements of MPS. The whole information used by MPS to simulate a certain process is based on the *Training Image* (TI) or *training data set*: the data set constituted of one or more variables used to infer the statistical relations and occurrence probability of any datum in the simulation. The TI may be constituted of a conceptual model instead of real data, but in the case of the rainfall time-series it is more likely to be a historical record of rainfall measurements. The *Simulation Grid* (SG) is a time referenced vector in which the generated values are stored during the simulation. Following a simulation path which is usually random, the SG is progressively filled with simulated values and becomes the actual output of the simulation. The *conditioning data* are a group of given data (e.g. rainfall measurements) situated in the SG. Being already informed, no simulation occurs at those time-steps. The presence of conditioning data affects, in their neighborhood, the conditional law used for the simulation and limits the range of possible patterns. MPS, as well some MC based algorithm for rainfall simulation (see section 2.1), may include the use of *auxiliary variables* to condition the simulation of the target variable. Auxiliary variables may either be known (fully or partially) and used to guide the simulation, or they may be unknown but still co-simulated because their structures contains important characteristics of the signal. For rainfall time-series, it could be for example: covariates of the original or previously simulated data (e.g. the number of wet days in a past period), a correlated variable for which the record is known, a theoretical variable that imposes a periodicity or a trend (e.g. a sinusoid function describing the annual seasonality over the data). Finally, the *search neighborhood* is a moving window, i.e. the portion of time-series located in the past and future neighborhood of each simulated value, used to retrieve the *data event*, i.e. the group of time-referenced values used to condition the simulation.

2.2.2 The Direct Sampling algorithm

Classical MPS implementations create a catalog of the possible neighbors patterns to evaluate the conditional probability of occurrence for each event with respect to the considered

neighborhood. This may imply a significant amount of memory and always limits the application to categorical variables. On the contrary, Direct Sampling generates each value by sampling the data from the TI where a sufficiently similar neighborhood exists. The DS implementation used in this chapter is called *DeeSse* [193]. The following is the main workflow of the algorithm for the simulation of a single variable. For the multivariate case see the last paragraph of this section.

Let us denote $\vec{x} = [x_1, ..., x_n]$ the time vector representing the SG, $\vec{y} = [y_1, ..., y_m]$ the one representing the TI and $Z(\cdot)$ the target variable, object of the simulation, defined at each element of $\vec{x}$ and $\vec{y}$. Before the simulation begins, all continuous variables are normalized using the transformation $Z \longmapsto Z \cdot (max(Z) - min(Z))^{-1}$ in order to have distances (see step 3) in the range $[0, 1]$. During the simulation, the uninformed time-steps of the SG are visited in a random order. The random simulation path $t \in \{1, 2, ..., n\}$ is obtained by sampling without replacement the discrete uniform distribution $U(1, n)$ where n is the SG length. At each uninformed x_t, the following steps are executed:

1. The data event $\vec{d}(x_t) = \{Z(x_{t+h_1}), ..., Z(x_{t+h_k})\}$ is retrieved from the SG according to a fixed neighborhood of radius R centered on x_t. It consists of at most N informed time-steps, closest to x_t. This defines a set of lags $H = \{h_1, ..., h_k\}$, with $|h_i| \leq R$ and $k \leq N$. The size of $\vec{d}(x_t)$ is therefore limited by the user-defined parameter N and the available informed time-steps inside the search neighborhood.

2. A random time-step y_i in $\vec{y}$ is visited and the corresponding data event $\vec{d}(y_i)$, defined according to H, is retrieved to be compared with $\vec{d}(x_t)$.

3. A distance $D(\vec{d}(x_t), \vec{d}(y_i))$, i.e. a measure of dissimilarity between the two data events, is calculated. For categorical variables (e.g. the dry/wet rainfall sequence), it is given by the formula:

$$D(\vec{d}(x_t), \vec{d}(y_i)) = \frac{1}{k} \sum_{j=1}^{k} a_j, \qquad a_j = \begin{cases} 1 & if \quad Z(x_j) \neq Z(y_j) \\ 0 & if \quad Z(x_j) = Z(y_j) \end{cases} \tag{2.1}$$

   while for continuous variables the following one is used:

$$D(\vec{d}(x_t), \vec{d}(y_i)) = \frac{1}{k} \sum_{j=1}^{k} |Z(x_j) - Z(y_j)| \tag{2.2}$$

   where k is the number of elements of the data event. The elements of $\vec{d}(x_t)$, independently from their position, play an equivalent role in conditioning the simulation of $Z(x_t)$. Note that, using the above distance formulas, the normalization applied before the simulation start is not needed for categorical variables, while for the continuous ones it ensures distances in $[0, 1]$.

4. If $D(\vec{d}(x_t), \vec{d}(y_i))$ is below a fixed threshold T, i.e. the two data events are sufficiently similar, the iteration stops and the datum $Z(y_i)$ is assigned to $Z(x_t)$. Otherwise, the process is repeated from point 2 until a suitable candidate $\vec{d}(y_i)$ is found or the prescribed TI fraction limit F is scanned.

5. If a TI fraction F has been scanned and the distance $D(\vec{d}(x_t), \vec{d}(y_i))$ is above T for each visited y_i, the datum $Z(y_i^*)$ minimizing this distance is assigned to $Z(x_t)$.

This procedure is repeated for the simulation at each x_t until the entire SG is covered. Figure 2.1 illustrates the iterative simulation using Direct Sampling and stresses some of its peculiarities. First, simulating $Z(x_t)$ in a random order allows $\vec{x}$ to be progressively populated

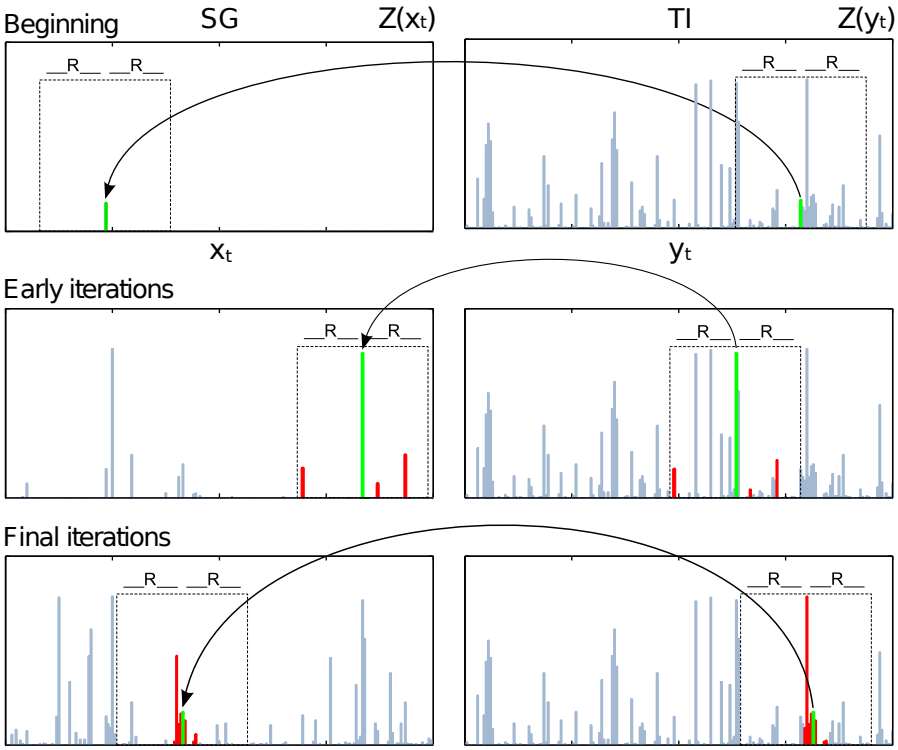


Figure 2.1: Sketch of the sequential simulation of a rainfall time-series performed by Direct Sampling: the dashed rectangle represents the search neighborhood of radius R, the datum being simulated is in green and the ones composing the data event are in red. Note the non-exact match between the data event in the SG and the one in the TI.

at non-consecutive time-steps. Therefore, the simulation at each x_t can be conditioned on both past and future, as opposed to the classical Markov-chain techniques, that use a linear simulation path starting from the beginning of the series, allowing conditioning on past only. In the early iterations, the closest informed time-steps used to condition the simulation are located far from x_t and its number is limited by the search window, i.e. conditioning is mainly based on large past and future time lags. On the contrary, the final iterations dispose of a more populated SG, conditioning is thus done on small time lags since only the closest N values are considered. This variable time-lag principle may not respect the auto-correlation on a specific time-lag rigorously, but it should preserve a more complex statistical relationship, which cannot be explored exhaustively using a fixed-dependence model.

DS can simulate multiple variables together similarly to the univariate case, dealing with a vector of variables $\vec{Z}(x_t)$ and considering a data event $\vec{d}_k$ different for each k-th variable, defined by N_k and R_k. Unlike the implementation presented in [122], *DeeSse* also uses a specific acceptance threshold T_k for each variable. Point 3 of the algorithm is repeated until a candidate with a distance below the threshold for all variables is found. If this condition is not met, the scan stops at the prescribed TI fraction F and the error for each candidate y_i and k-th variable is computed with the following formula: $E_k(y_i) = (D(\vec{d}_k(x_t), \vec{d}_k(y_i)) - T_k)T_k^{-1}$, where $D(\cdot,\cdot)$ is defined as in Point 3. Finally, the candidate minimizing $max(\vec{E}(y_i))$ is assigned to $\vec{Z}(x_t)$. Note that the entire data vector $\vec{Z}(x_t)$ is simulated in one iteration, reproducing exactly the same combination of values found for all the variables at the sampled time-step, excluding the conditioning data, already present in the SG. This feature, called Variable Vector Mode (VVM), although reducing the variability in the simulation, has been

adopted to accurately reproduce the correlation between variables (see the preliminary tests in appendix B).

2.2.3 Comparison with existing resampling techniques

The resampling principle is at the base of some already proposed techniques for rainfall and hydrologic time-series simulation. There exist two principal families of resampling techniques: the block bootstrap [207, 191, 136], which implies the resampling with replacement of entire pieces of time-series with the aim of preserving the statistical dependence at a scale minor than the blocks size, and the k-nearest neighbor bootstrap (k-NN), based on single value resampling using a pattern similarity rule. This latter family of techniques, introduced by [46] and inspired to the jackknife variance estimation, has seen several developments in hydrology [223, 110, 109, 155, 24, 217, 33]. Having different points in common with Direct Sampling, its general framework is briefly presented in the following. Each datum inside the historical record is characterized by a vector $\vec{d_t}$ of predictor variables, analogous to the data event for DS. For example, to generate $Z(x_t)$ one could use $\vec{d_t} = [Z(x_{t-1}), Z(x_{t-2}), U(x_t), U(x_{t-1})]$, meaning that the simulation is conditioned to the 2 previous time-steps of Z and the present and previous time-steps of U, a correlated variable. In the predictor variables space $\mathbb{D}$, the historical data as well as $Z(x_t)$, which still has to be generated, are represented as points whose coordinates are defined by $\vec{d_t}$. Consequently, proximity in $\mathbb{D}$ corresponds to similarity of the conditioning patterns. $Z(x_t)$ is simulated by sampling an empirical pdf built on the k points closest to $Z(x_t)$; the closer the point is, the higher is the probability to sample the corresponding historical datum. Proposed variations of the algorithm include transformations of the predictor variables space, the application of kernel smoothing to the k-NN pdf to increase the variability beyond the historical values, and different methods to estimate the parameters of the model, e.g. k and the kernel bandwidth.

Going back to Direct Sampling, the similarities with the k-NN bootstrap are: i) they both make a resampling of the historical record conditioned by an ensemble of auxiliary/predictor variables; ii) they both compute a distance as a measure of dissimilarity between the simulating time-step and the candidates considered for resampling. Nevertheless, there are several points of divergence in the rationale of the techniques: i) in the k-NN bootstrap, the distance is used to evaluate the resampling probability, while in DS it is used to evaluate the resampling possibility. This means that, using the k-NN resampling, the conditional pdf is a function of the distance, while in DS the distance is only used to define its support. In fact, using DS, the space $\mathbb{D}$ is not restricted to the k nearest neighbors but it is bounded by the distance thresholds: outside the boundary, the resampling probability is zero, while inside, it follows the occurrence of the data in the scanned TI fraction, without being a function of the pattern resemblance. Only in case of no candidate found, the closest neighbor outside the bounded portion of $\mathbb{D}$ is chosen for resampling. The latter can be considered as an exceptional condition which usually does not lead to a good simulation and seldom occurs using an appropriate setup and training data set (see appendix A). ii) Using DS, the conditional pdf remains implicit, its computation is not needed: the historical record is randomly visited instead and the first datum presenting a distance below the threshold T is sampled. This is an advantage since it avoids the problem of the high-dimensional conditional pdf estimation which limits the degree of conditioning in bootstrap techniques [175]. iii) The k-NN technique considers a fixed time-dependence, while it varies during the simulation in the case of DS. iv) Finally, the simulation path (in the SG) is always linear in the k-NN technique, while it is random using DS, allowing conditioning on future time-steps of the target variable.

Table 2.1: Summary of the data set used.

Location	Station	Period [years]	Record length [days]	Missing data [days]
Alice Springs	A.S.Airport	1940-2013	26347	305
Sydney	S.Observatory Hill	1858-2013	56662	184
Darwin	D.Airport	1941-2013	26356	0

2.3 Application

The data set chosen for this study is composed of three daily rainfall time-series from different climatic regions of Australia: Alice Springs (hot desert), with a very dry rainfall regime and long droughts, Sydney (temperate), with a far wetter climate due to its proximity to the ocean, and Darwin (tropical savannah), showing an extreme variability between dry and wet seasons. Table 2.1 presents the data set used: the chosen stations provide a considerable record of about 70 years for Darwin and Alice Springs and 150 years for Sydney. Any gaps or trends have been explicitly kept to test the behavior of the algorithm with incomplete or non-stationary data sets. Direct Sampling treats gaps in the time-series in a simple way: each data event found in the TI is rejected if it contains any missing data. This allows incomplete training images to be dealt with in a safe way, but, as one could expect, a large quantity of missing data, especially if sparsely distributed, may lead to a poor simulation. [121] and chapter 4 show how Direct Sampling can be used for data reconstruction.

Since rainfall is a complex signal exhibiting not only multi-scale time dependence but also intermittence, the classical approach is to split the daily time-series generation in two steps: the occurrence model, where the dry/wet daily sequence is generated using a Markov-chain, and the amount model, where the rainfall amount is simulated on wet days using an estimation of the conditional pdf [e.g., 34]. The simulation framework proposed here is radically different: we use Direct Sampling to generate the complete time-series in one step, simulating multiple variables together. In particular, the TI used is based on the past daily rainfall record and composed of a group of variables (table 2.2). These are all time-series ($f(t)$), but we simplify the notation where possible by omitting the term (t): 1) the average rainfall amount on a 365 days centered moving window ($365MA$) [mm], 2) the sum of the current and the previous day amounts ($2MS$) [mm], 3) and 4) two out-of-phase of triangular functions ($A1$ and $A2$) with frequency 365.25 days, similar to trigonometric coordinates expressing the position of the day in the annual cycle, 5) the dry/wet sequence (W), i.e. a categorical variable indicating the position of a day inside the rainfall pattern (1 = wet, 0 = dry, 2 = solitary wet, 3 = wet day at the beginning or at the end of a wet spell), 6) the daily rainfall amount, which is the target of the simulation. The first two auxiliary variables are covariates used to force the algorithm to preserve the inter-annual structure and the day-to-day correlation, which are known to exist a priori. The other ones are used to reproduce the dry/wet pattern and the annual seasonality accurately. Moreover, any unknown dependence in the daily rainfall signal is generically taken into account in the simulation by using a data event of variable length as explained in section 2.2.2. It has to be remarked that, apart from 3) and 4), which are known deterministic functions imposed as conditioning data, the rest of the auxiliary variables are transformations of the rainfall datum, automatically computed on the TI and co-simulated with the daily rainfall.

To summarize, the main parameters of the algorithm are the following: the maximum scanned TI fraction $F \in (0, 1]$, the search neighborhood radius R, the maximum number of neighbors N, both expressed in number of elements of the time vector, and the distance threshold $T \in (0, 1]$. Recall that, apart from F, each parameter is set independently for each simulated variable. The setup shown in table 2.2 is used together with $F = 0.5$ and proposed as a standard for daily rainfall time-series. The sensitivity analysis, shown in appendix A

and a trial-and-error refining phase led to this setup which is not the result of a numerical optimization on a specific data set, but it is rather in accordance to the criteria used to define the order and extension of the variable time-dependence, as shown below. Applying it to any type of single-station daily rainfall data set, the user should obtain a reliable simulation without needing to change any parameter or give supplementary information. An additional refinement of the setup is also possible, keeping in mind the following general rules:

- R limits the maximum time-lag dependence in the simulation and should be set according to the length of the largest sufficiently repeated structure or frequency in the signal that has to be reproduced. Being interested to condition the simulation upon the inter-annual fluctuations (visible in the 10-year MA time-series in figure 2.9), we set $R_{365MS} = R_Z = 5000$ for the $365MS$ and daily rainfall variables. Regarding dry/wet pattern conditioning, we prefer limiting the variable time dependence within a 21-day window ($R_W = 10$). This window should be set between the median and the maximum of the wet spell length distribution, in order to properly catch the continuity of the rainfall events over multiple days.

- N controls the complexity of the conditioning structure but also influences the specific time-lag dependence. For instance, if one increases N, higher-order dependencies are represented, but the weight accorded to a specific neighbor in evaluating the distance between patterns becomes lower. This leads to a less accurate specific time-lag conditioning, but a more complex time-dependence is respected on average. For the rainfall amount and $365MA$ variables, $N \ll R$ follows the same setup rule as for R_W. In this way, in the initial iterations, the conditioning neighbors will be sparse in a 10001 days window ($R = 5000$) to respect low-frequency fluctuations, whereas, in the final iterations, they will be contained in a N-day window to respect the within-spell variability. The standard value proposed here ($N_{365MA} = N_Z = 21$) corresponds approximately to the spell distribution median of the Darwin time-series, remaining in the appropriate range for the other considered climates. Conversely, N_W is kept lower in order to focus the conditioning on the small-scale dry/wet pattern. $N_W = 5$ gave in general the best result in terms of dry/wet pattern reproduction.

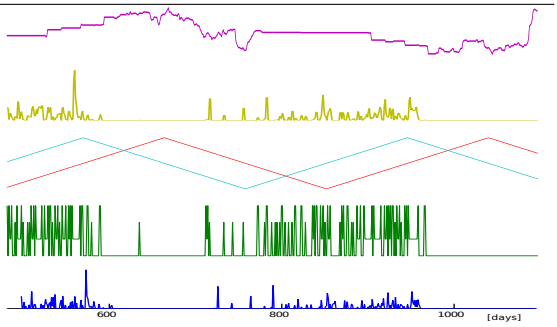
- For $2MS$, $A1$ and $A2$, the time-dependence is limited to lag 1 by using $N = R = 1$. This combination should not be changed since we have no interest in expanding or varying the time lag-dependence for the mentioned variables.

- T determines the tolerance in accepting a pattern. The sensitivity analysis done until now on different types of heterogeneity (see appendix A and [124]) confirmed that the optimum generally lies in the interval $[0.01, 0.07]$ (1 to 7% of the total variation). Higher T values usually lead to poorly simulated patterns, but lower ones may induce a bias in the marginal distribution and increase the phenomenon of verbatim copy, i.e. the exact reproduction of an entire portion of data by oversampling the same pattern inside the TI. For these reasons, we recommend keeping the proposed standard value $T = 0.05$ for all the variables.

- F should be set sufficiently high to have a consistent choice of patterns but a value close to 1, i.e. all the TI is scanned each time, may lower the variability of the simulations and increase the verbatim copy. Using a training data set representative enough, the optimal value corresponds to a TI fraction containing some repetitions of the lowest-frequency fluctuation that should be reproduced. Considering the randomness of the TI scan, the value $F = 0.5$ chosen in this case is sufficient to serve the purpose.

Table 2.2: Standard setup proposed for rainfall simulation. The parameters are: search window radius R, maximum number of neighbors N and distance threshold T. The variables are: 1) the 365 days Moving Average ($365MA$), 2) the Moving Sum of the current day and the one before ($2MS$), 3) and 4) annual seasonality triangular functions ($A1$ and $A2$), 5) the dry/wet sequence W and 6) the daily rainfall amount as the target variable. On the right, a portion of multivariate TI is given as example.

Variable	R	N	T
1) $365MA$	5000	21	0.05
2) $2MS$	1	1	0.05
3) $A1$	1	1	0.05
4) $A2$	1	1	0.05
5) W	10	5	0.05
6) Z	5000	21	0.05



2.3.1 Imposing a trend

As already shown in [32, 121, 73, 76], in case of a non-stationary target variable, the simulation can be constrained to reproduce the same type of trend found in the TI by making use of an auxiliary variable . The one proposed here is the integer vector $L = [1, 2, ..., \mathrm{M}]$, where M is the length of the time-series, tracking the position of each datum inside the TI. L is assigned to the SG as conditioning datum with the following parameters: $R_L = 1$, $N_L = 1$ and $T_L = 0.01$. According to the threshold T_L, the sampling is therefore constrained to a neighborhood of the same time-step inside the TI: for example, in the Darwin case, being $\mathrm{M} = 26356$ and $T_L = 0.01$ (1% of the total variation allowed), the sampling to simulate $Z(x_t)$ is constrained to the interval $y_t \pm 263$ [days]. In this way, the marginal distribution is respected, but the local variability is restricted to the one found inside the training data set, reproducing the same trend. The following remarks are noteworthy: i) to avoid an unnecessary restriction of the sampling, T_L should correspond to the maximum time interval for which the target variable can be considered stationary; ii) the simulation should not be longer than the training data set, having no basis to extrapolate the trend in the past or future; iii) the local variability is not completely limited by L: a pattern outside the tolerance range (i.e. with a distance over the threshold) could be sampled if no better candidate is found.

2.3.2 Validation

To test the proposed technique the visual comparison of the generated time-series with the reference as well as several groups of statistical indicators are considered. The empirical cumulative probability distributions, obtained using the Kaplan-Meier estimate [87], of the daily, the annual and decennial rainfall time-series, obtained by summing up the daily rainfall, are compared using quantile-quantile (qq-) plots. Moreover, the minimum moving average, i.e. the minimum value found on the moving average of each time-series, is computed using different running window lengths up to 60 years to assess the efficiency of the algorithm in preserving the long-term dependence characteristics of the rainfall.

The daily rainfall statistics have been analyzed separately for each month considering the average value of the following indicators: the probability of occurrence of a wet day and the mean, standard deviation, minimum and maximum on wet days only. For instance, the

standard deviation is computed on the wet days of each month of January, then the average value is taken as representative of that time-series. We therefore obtain a unique value for the reference and a distribution of values for the simulations represented with a box-plot.

Another used validation criterion is the comparison of the dry and wet spell length distributions. Each series is transformed into a binary sequence with zeros corresponding to dry days and ones to the wet days. Then, counting the number of days inside each dry and wet spell, we obtain the distributions of dry and wet spell length, that can be compared using qq-plots. This is an important indicator since it determines, for example, the efficiency of the algorithm in reproducing long droughts or wet periods.

Since DS works by pasting values from the TI to the SG, it is straightforward to keep track of the original location of each value in the training image. If successive values in the TI are also next to each other in the SG, then a patch is identified. A multiple box-plot is then used to represent the number of patches found in each realization as a function of the patch length to keep track of the verbatim copy effect.

The last group of indicators considered is the sample Partial Autocorrelation Function (PACF) [22] of the daily, monthly and annual rainfall. Given a time-series $\vec{X}_t$, the sample PACF is the estimation of the linear correlation index between the datum at time t and the ones at previous time-steps $t-h$, without considering the linear dependence with the in-between observations. For a stationary time-series the sample PACF is expressed as a function of the time-lag h with the following formula:

$$\hat{\rho}(X_t, h) = Corr[X_t - \hat{E}(X_t|\{X_{t-1}, ..., X_{t-h+1}\}), X_{t-h} - \hat{E}(X_{t-h}|\{X_{t-h+1}, ..., X_{t-1}\})] \quad (2.3)$$

where $\hat{E}(X_t|\{X_{t-1}, ..., X_{t-h+1}\})$ is the best linear predictor of X_t knowing the observations $\{X_{t-1}, ..., X_{t-h+1}\}$. $\hat{\rho}(\cdot, h)$ varies in $[0, 1]$, with high values for a highly autocorrelated process. This indicator is widely used in time-series analysis since it gives information about the persistence of the signal. The autocorrelation function could be used instead, but PACF is preferred here since it shows the autocorrelation at each lag independently. In the case of daily rainfall, the partial autocorrelation is usually very low, while the higher-scale rainfall may present a more important specific time-lag linear dependence. As usually done in the absence of any prior knowledge about X_t, the $5-95\%$ confidence limits of an uncorrelated white noise are adopted to assess the significance of the PACF indexes. An accurate way to detect a significant autocorrelation at a certain lag, is to compare it with an IID $\sim N(0, \sigma^2)$ white noise. Such a signal is totally non-autocorrelated and presents a sample PACF ($\hat{\rho}_{\mathrm{AN}}$) near zero for any $h > 1$. Moreover, $\hat{\rho}_{\mathrm{AN}}$ follows the asymptotic normal distribution $\mathrm{AN}(0, \mathrm{n}^{-1})$, n being the number of observations in the considered sample. The 95% confidence interval of this distribution can be used to test the significance of any $\hat{\rho}(h)$. That is, in the estimation of the autocorrelation for $\vec{X}_t$, at all the $\hat{\rho}(h)$ values within $0 \pm 1.96\mathrm{n}^{-1/2}$ [Bartlett's formula 19] can be considered negligible, being of the same magnitude as $\hat{\rho}_{AN}$. Conversely, the values outside these boundaries are probably the expression of a significant autocorrelation and should be reproduced by the simulation. Since the time-series used in this case are not necessarily stationary, any sample PACF is computed from the standardized signal X_t^s, obtained by applying moving average estimation $\hat{m}_t$ and standard deviation $\hat{s}_t$ filters with the following formula:

$$\begin{aligned} X_t^s &= \frac{X_t - \hat{m}_t}{\hat{s}_t}, \qquad \hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t+j} \\ \hat{s}_t &= [(2q+1)^{-1} \sum_{j=-q}^{q} (X_{t+j} - \hat{m}_t)^2]^{-\frac{1}{2}}, \qquad q+1 \leq t \leq n-q \end{aligned} \quad (2.4)$$

where $q = 2555$ (15 years centered moving window). It is important to note that this operation may exclude from the PACF computation a consistent part of the signal ($q+1 \leq t \leq n-q$), especially on the higher time-scale. In the case of the data sets used, the annual time-series is reduced to less than 60 values for Alice Springs and Darwin: a barely sufficient quantity, considering that the minimum amount of data for a useful sample PACF estimation suggested by [22] is of about 50 observations.

2.4 Results and discussion

To evaluate the proposed technique, a group of 100 realizations of the same length as the reference is generated for each of the 3 considered data sets to obtain a sufficiently stable response in both the average and the extreme behavior. The setup used is the one presented in section 2.3 with the fixed parameters values shown in table 2.2. The obtained results are shown and discussed in the following section.

2.4.1 Visual comparison

Figure 2.2 shows the comparison between random samples from both the simulated and the reference time-series. For each data set, the generated rainfall looks similar to the reference: the extreme events inside the 10-year samples are reproduced with an analogous frequency and magnitude. The annual seasonality, particularly pronounced in the Darwin series, is accurately simulated as well as the persistence of the rainfall events, visible in the 100-day samples. These aspects are evaluated quantitatively in the following sections.

2.4.2 Multiple-scale probability distribution

The qq-plots of the rainfall empirical distributions are presented in figure 2.3, where all the range of quantiles is considered. The distribution of the daily rainfall (computed on wet days only) is generally respected, although some extremes that are present only once in the reference and, in particular, at the start or end of the time-series, may not appear in the simulation. It is the case of the Darwin series, with a mismatch of the very upper quantiles. Moreover, DS being an algorithm based on resampling, the distribution of the simulated values is limited by the range of the training data set: this is shown in the Alice Springs and Sydney qq-plots, where the distribution of the last quantiles is clearly truncated at the maximum value found in the reference. This result is normally expected using this type of technique: Direct Sampling is of course not able to extrapolate extreme intensities higher than the ones found in the TI at the scale of the simulated signal. On the contrary, the distribution of the rainfall amount on the solitary wet days is accurately respected, with some realizations including higher extremes than the reference. More importantly, the annual and 10-year rainfall distributions are correctly reproduced and do not show over-dispersion. The phenomenon, common among the classical techniques based on daily-scale conditioning, consists in the scarce representation of the extremes and underestimation of the variance at the higher scale. This problem is avoided here because a variable dependence is considered, up to a 5000-day radius on the $365MA$ auxiliary variable, that helps preserving the low-frequency fluctuations. We also see that, at this scale, DS is capable of generating extremes higher than the ones found in the reference, meaning that new patterns have been generated using the same values at the daily scale. This result is based on the reproduction of higher-scale patterns: the acceptance threshold value chosen for the $365MA$ auxiliary variable allows enough freedom to generate new patterns although maintaining an unbiased distribution. Nevertheless, this approach is not meant to replace a specific technique to

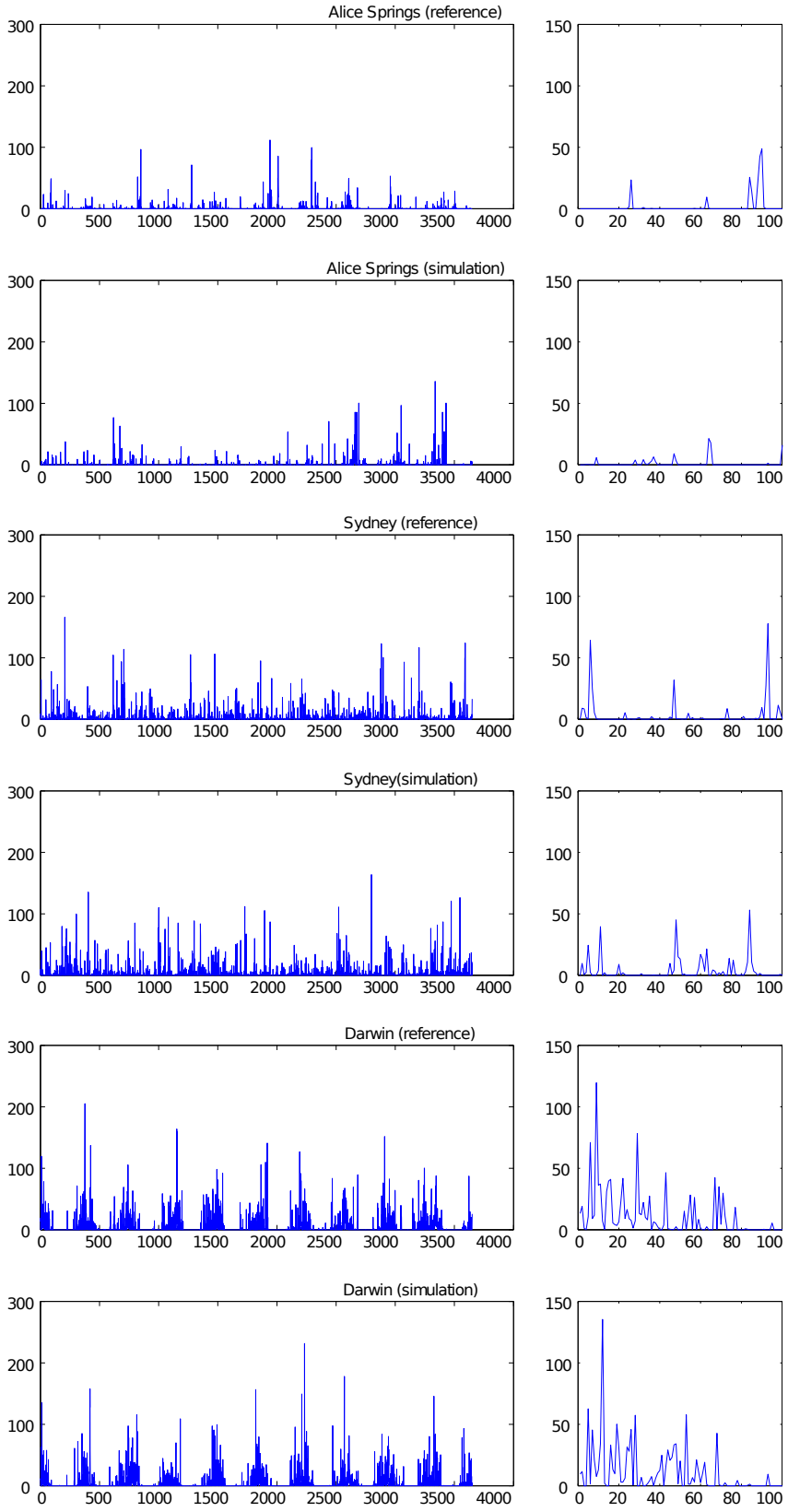


Figure 2.2: Visual comparison between the simulated and the reference daily rainfall [mm] time-series: 10-year (left column) and 100-day (right column) random samples.

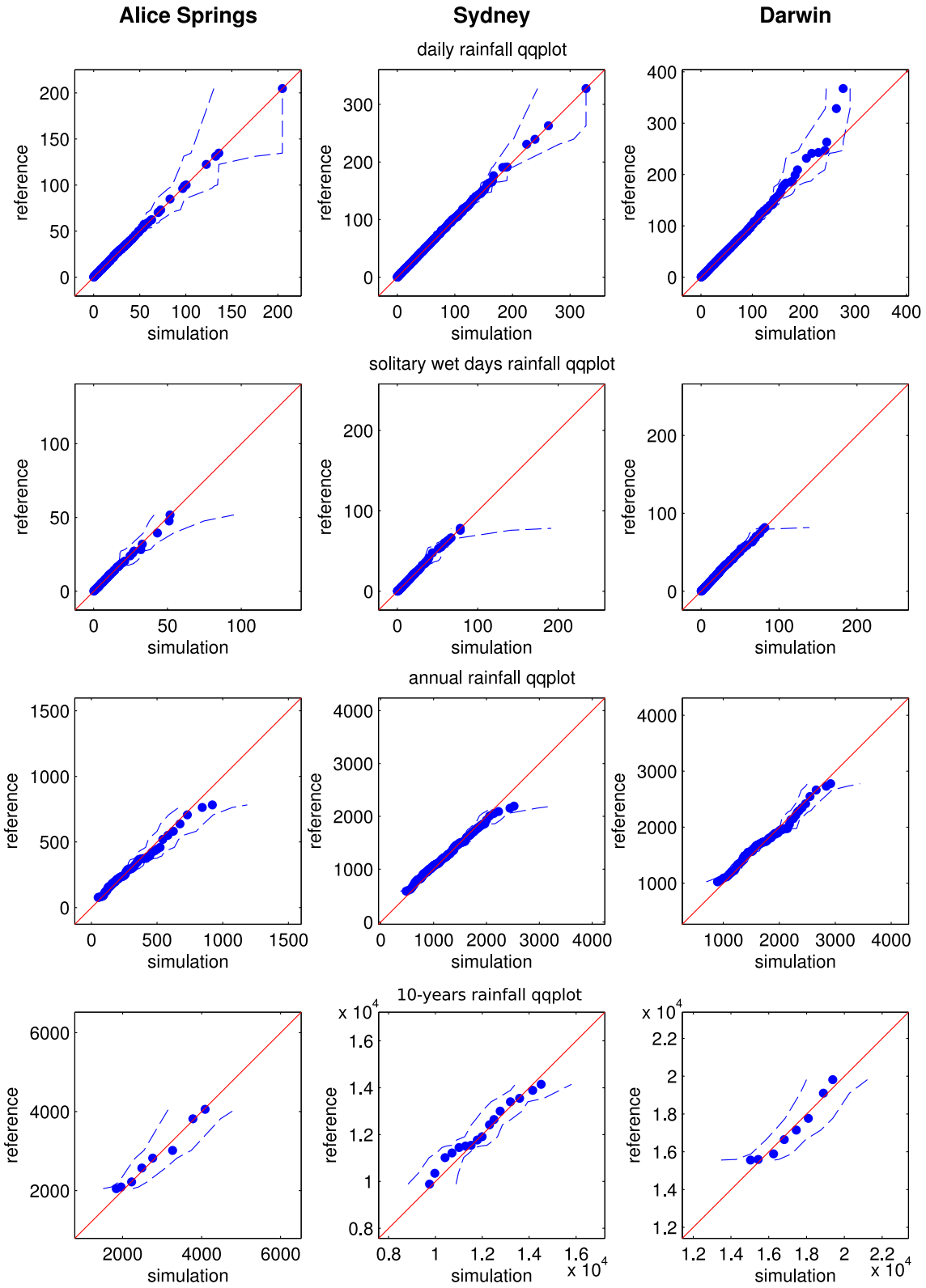


Figure 2.3: qq-plots of the empirical probability rainfall amount [mm] distributions: median of the realizations (blue dots), 5th and 95th percentile (dashed lines). The bisector (solid line) indicates the exact quantile match.

predict long recurrence-time events at any temporal scale, since it is not focused on modeling the tail of the probability distribution.

2.4.3 Annual seasonality and extremes

Figure 2.4 shows the principal indicators describing the annual seasonality of the reference and the generated time-series: each different season is accurately reproduced by the algorithm, with almost no bias. The probability of having a wet day, usually imposed by a prior model in the classical parametric techniques, is indirectly obtained by sampling from the rainfall patterns of the appropriate period of the year. This goal is mainly achieved using the auxiliary variables $A1$ and $A2$ as conditioning data (see section 2.3). The simulation of the average extremes, shown in figure 2.5, also follows the reference rather accurately.

2.4.4 Rainfall patterns and verbatim copy

The statistical indicators regarding the dry/wet patterns shown in figure 2.6 demonstrate the efficiency of the proposed DS setup in simulating long droughts or wet periods according to the training data set: the dry and wet spell distributions are preserved and extremes higher than the ones present in the TI are also simulated.

The verbatim copy box-plots show the distribution of the time-series pieces exactly copied from the TI as a function of their size for the ensemble of the realizations: the number of patches decreases exponentially with their size. The phenomenon is mainly limited to a maximum of few 8-day patches, with isolated cases up to 14 days.

The 10-year rainfall moving sum, shown at the bottom of figure 2.6, illustrates the low-frequency time-series structure: the quantiles of the simulations at each time-step confirm that the overall variability is correctly simulated, but the local fluctuations do not match the reference. For example, the Darwin reference series shows a clear upward trend which is not present in the superposed randomly-picked DS realization. Generally, the TI is supposed to be stationary or the non-stationarity should be at least described by an auxiliary variable. If it is not the case, as for the Darwin time-series, the algorithm honors the marginal distribution of the reference, but it does not reproduce a specific trend. This problem is treated separately in section 2.4.6.

The minimum moving average on different window lengths up to 60 years (figure 2.7) gives information about the long-term structure of rainfall. The zero values are in accordance with the dry spell distribution shown in figure 2.6: for example, Alice Springs presents a zero minimum moving average until 5 months, meaning that it contains dry spells of this length. Alice Springs and Sydney show a very different long-term structure: the former with long dry spells, the latter with a wider range of minimum values. Darwin presents the peculiarities of both climates with a sharp rising from the annual to the 60 years scale. According to this indicator, the simulation of the long-term structure is fairly accurate. The negative bias, lower than 0.5 mm, shows a modest tendency to underestimate the minimum moving average from the annual to the decennial scale for wet climates as Sydney and Darwin.

2.4.5 Linear time-dependence

The specific linear time-dependence of the generated and reference signals has been evaluated at different scales using the sample Partial Autocorrelation Function (PACF, figure 2.8, Equation 2.4). At the daily scale, the data show the same level of autocorrelation at lag-1 and a low but significant linear dependence until lag 3 for Alice Springs and Sydney, while Darwin presents a longer tailing that asymptotically approaches the confidence bounds of an uncorrelated noise. The DS simulation shows a tendency to a slight underestimation of the

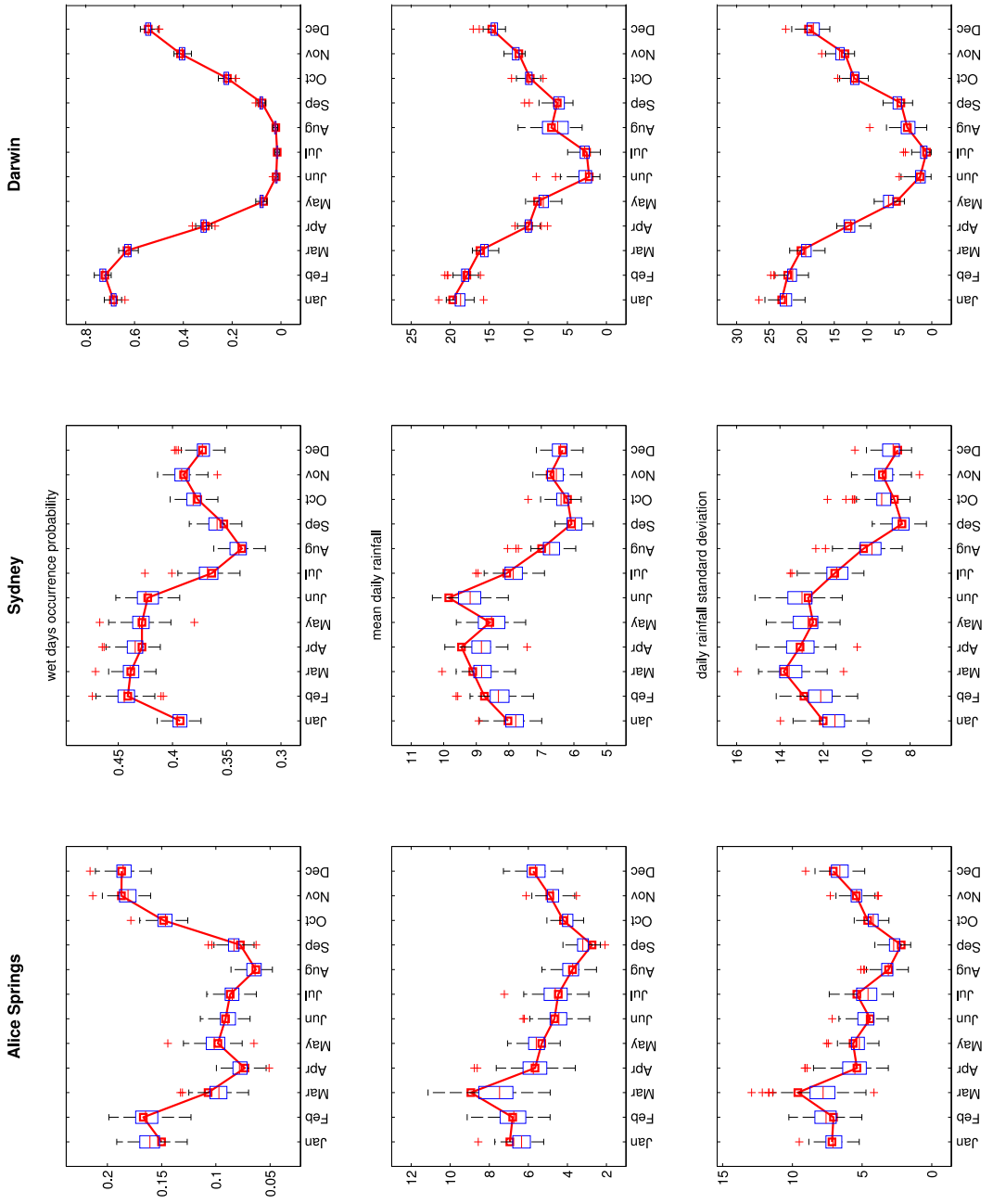


Figure 2.4: Box-plots of the average wet day occurrence probability, mean daily rainfall amount [mm] and its standard deviation per month. The solid line indicates the reference.

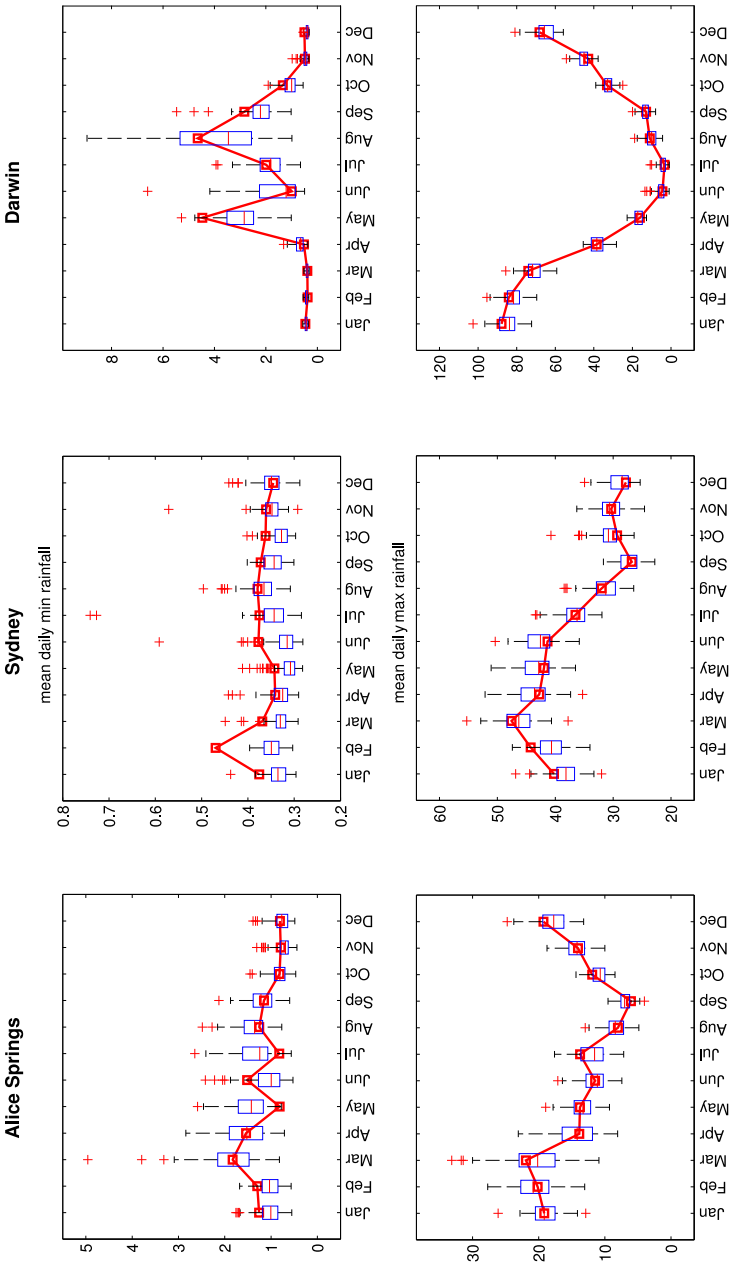


Figure 2.5: Box-plots of the average extremes per month [mm]. The solid line indicates the reference.

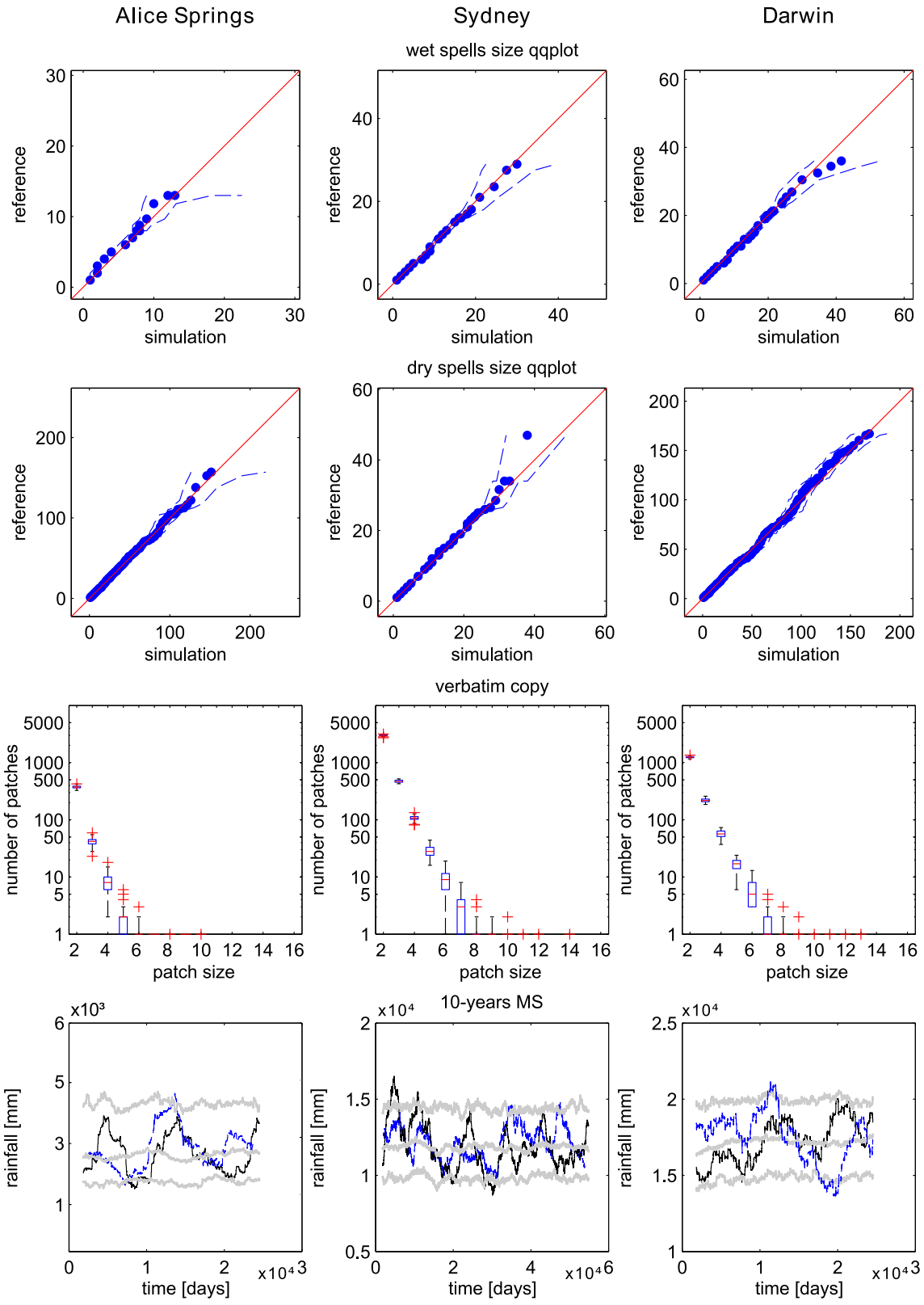


Figure 2.6: Main indicators describing the rainfall pattern: qq-plots of the dry and wet spell [days] distributions, verbatim copy box-plots as function of the patch size [days] and daily 10-year Moving Sum (MS) time-series [mm] of the reference (black line), median, 5-th and 95-th percentile of the realizations (gray lines) and a randomly picked simulation (dashed blue line).

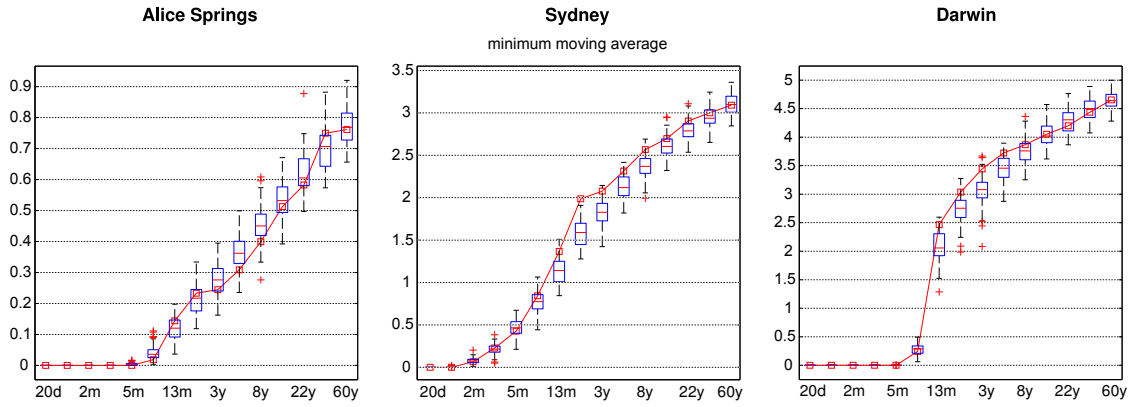


Figure 2.7: Minimum moving average of daily rainfall [mm] for different running window lengths (days, months or years). The solid line indicates the reference.

lag-1 PACF, with a maximum error around 0.1 for Sydney. Since the algorithm operates in a non-parametric way and imposes a variable time-dependence, the eventuality of modifying the structure of the daily signal cannot be excluded a priori, for this reason the PACF has been calculated up to the 20th lag, assuring that no extra linear-dependence has been introduced.

At the monthly scale, the linear time-dependence structure is clearly related to the annual seasonality, with a negative autocorrelation around lag 6 and a positive one around lag 12. The climate characterization is also evident: from Alice Springs to Darwin we see a more marked seasonality reflected in the PACF. The simulation follows the reference fairly well, with a maximum error around ± 0.1.

At the annual scale, the limited length of the time-series leads to wider confidence bounds for the non-significant values (see section 2.3.2). The reference does not show a clear linear time-dependence structure which is not similarly reproduced by the simulation. Some more relevant discrepancy is present in the Darwin series, showing a more discontinuous structure. However, using such a limited data set for the time scale considered here, it is difficult to determine if the reference PACF is really indicative of an effective linear dependence.

2.4.6 Non-stationary simulation

Figure 2.9 shows the Darwin time-series simulation preserving the same non-stationarity contained in the reference by using the technique proposed in section 2.3.1. The 10-year moving sum plot shows that the trend and low-frequency fluctuation present in the reference are accurately simulated: the median of the realizations follows the reference and a variability of about 4 m between the 5-th and 95-th percentile is present. Regarding the other considered statistical indicators, the performance appears to be essentially the same as for the stationary simulation: the only remarkable difference is a modest positive bias in the maximum wet periods length. The fact that, to impose the trend, the sampling is restricted to a local region of the reference reduces the local variability with respect to the stationary simulation. Consequently, a modest increase of the verbatim copy effect occurs.

This technique can find application in cases where a specific non-stationarity extended to high-order moments should be imposed, e.g. exploring the uncertainty of a given past or future scenario, where a simple trend or seasonality adjustment is insufficient and an overly complex parametric model would be necessary to preserve the same long-term behavior.

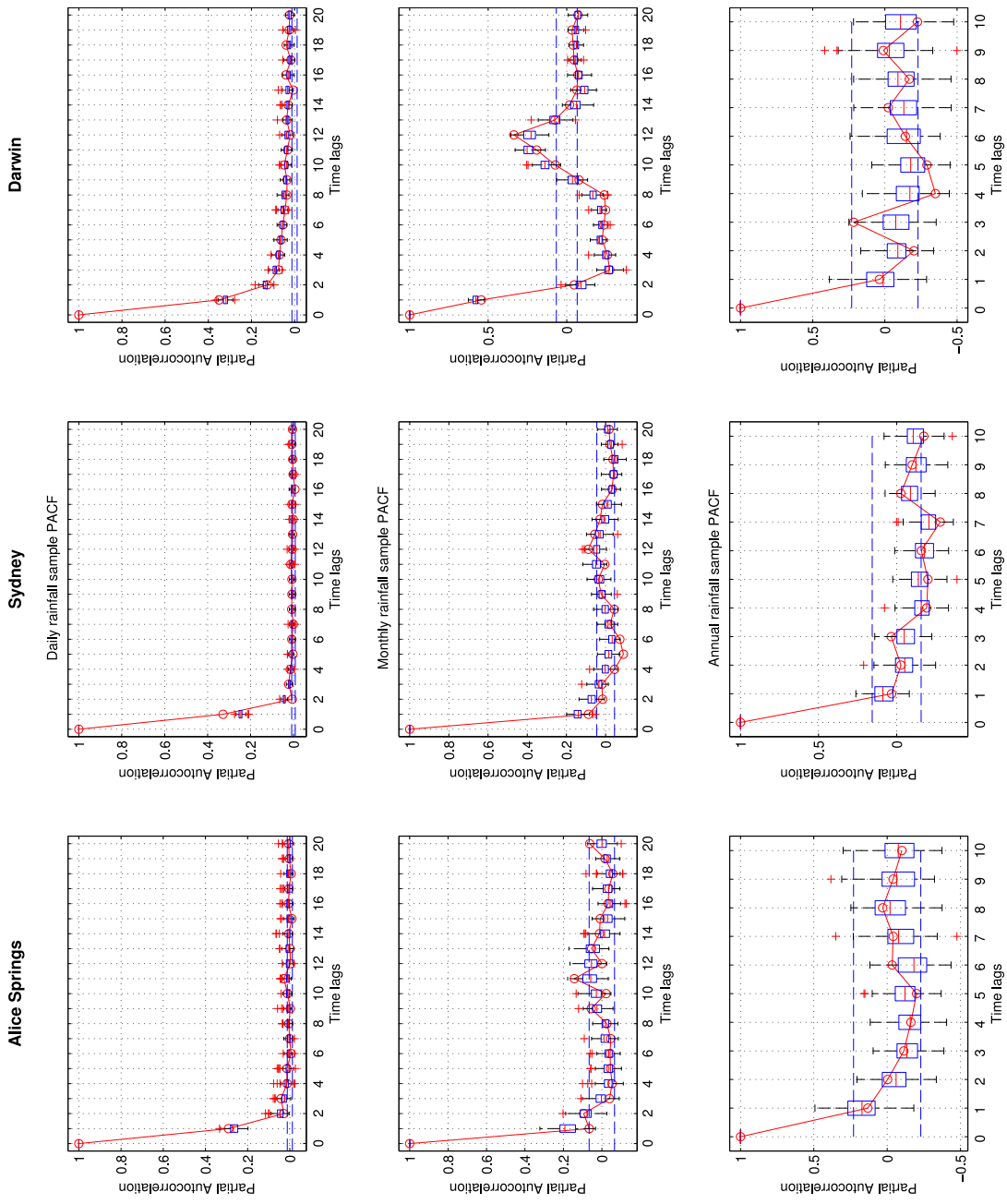


Figure 2.8: Sample Partial Autocorrelation Function (PACF) of the daily, monthly and annual rainfall signal: the reference (solid line), 100 DS simulations (box-plots), and confidence bounds for the negligible autocorrelation indexes (dashed lines).

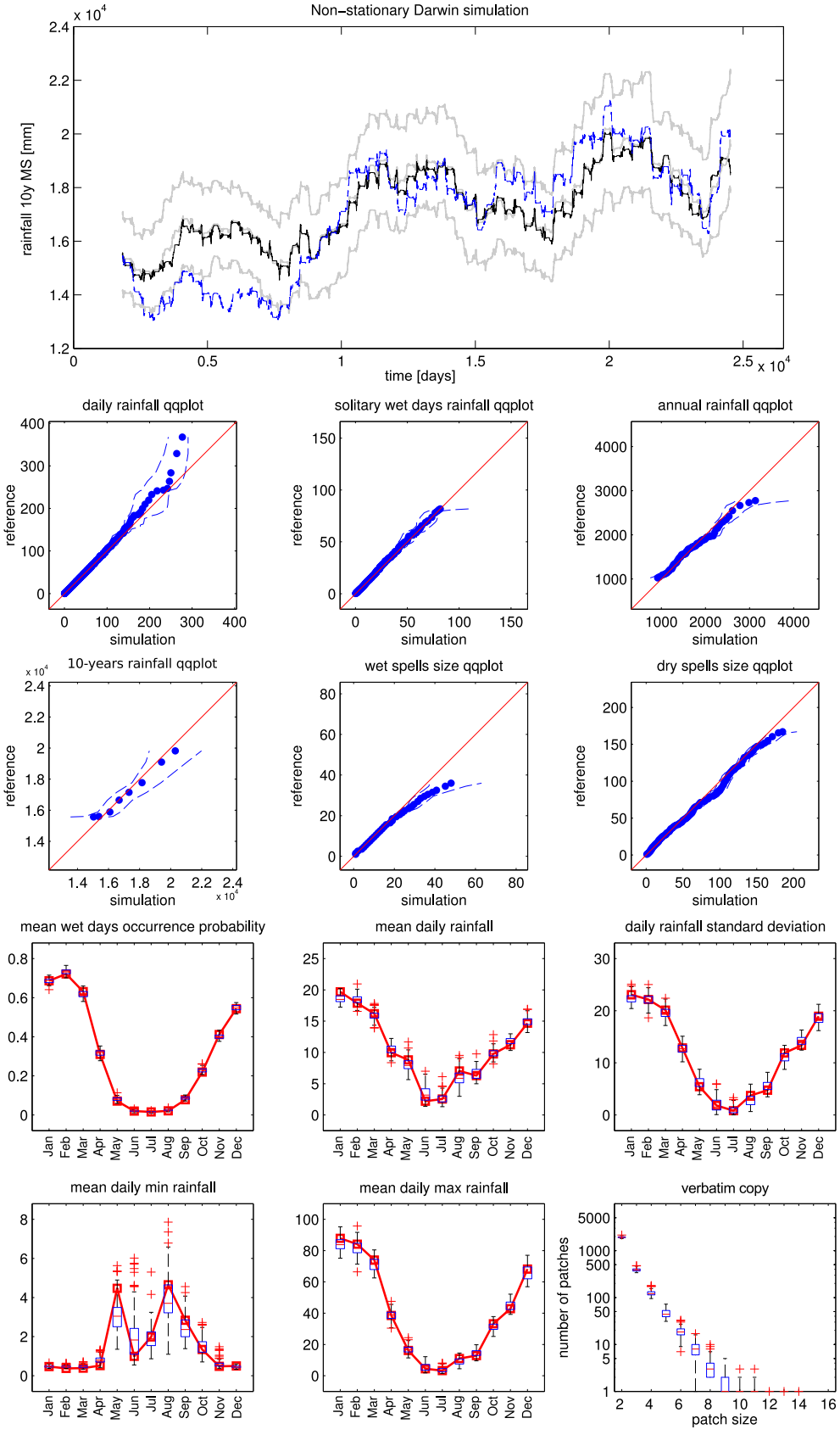


Figure 2.9: Darwin daily rainfall non-stationary simulation: 10-year Moving Sum time-series (top) of the reference (black line), median, 5-th and 95-th percentile of the realizations (gray lines) and a randomly picked simulation (dashed blue line); main quantile-comparisons (center); main seasonal indicators and verbatim copy box-plot (bottom).

2.5 Conclusions

The aim of the chapter is to present an alternative daily rainfall simulation technique based on the Direct Sampling algorithm, belonging to multiple-point statistics family. The main principle of the technique is to resample a given data set using a pattern-similarity rule. Using a random simulation path and a non-fixed pattern dimension, the technique allows imposing a variable time-dependence and reproducing the reference statistics at multiple scales. The proposed setup, suitable for any type of rainfall, includes the simulation of the daily rainfall time-series together with a series of auxiliary variables including: a categorical variable describing the dry/wet pattern, the 2 days moving sum which helps respecting the lag-1 autocorrelation, the 365 days moving average to condition upon inter-annual fluctuations and two coupled theoretical periodic functions describing the annual seasonality. Since all the variables are automatically computed from the rainfall data, no additional information is needed.

The technique has been tested on three different climates of Australia: Alice Springs (desert), Sydney (temperate) and Darwin (tropical savannah). Without changing the simulation parameters, the algorithm correctly simulates both the rainfall occurrence structure and amount distribution up to the decennial scale for all the three climates, avoiding the problem of over-dispersion, which often affects daily-rainfall simulation techniques. Being based on resampling, the algorithm can only generate data which are present in the training data set, but they can be aggregated differently, simulating new extremes in the higher-scale rainfall and dry/wet pattern distributions. The technique is not meant to be used as a tool to explore the uncertainty related to long recurrence-time events, but rather to generate extremely realistic replicates of the datum, to be used as inputs in hydrologic models.

Reproducing the specific trend found in the data is also possible by making use of an additional auxiliary variable which simply restricts the sampling to a local portion of the TI. This way, any type of non-stationarity present in the TI is automatically imposed on the simulation. The Darwin example demonstrates the efficiency of this approach in reproducing 100 different realizations showing the same type of trend and marginal distribution. This setup can be useful to simulate multiple realizations of a specific non-stationary scenario regardless of its complexity.

In conclusion, the Direct Sampling technique used with the proposed generic setup can produce realistic daily rainfall time-series replicates from different climates without the need of calibration or additional information. The generality and the total automation of the technique makes it a powerful tool for a routine use in scientific and engineering applications.

Chapter 3

Simulating the complexity of rainfall: a comparison between the Markov-chain approach and multiple-point statistics

Abstract

Daily rainfall is a complex signal exhibiting intermittence, regular fluctuations but also a chaotic behavior at multiple scales that cannot be preserved by stationary statistical models. In this chapter, we compare two different techniques of stochastic rainfall simulation that aim at preserving this complexity: the first one is the modified Markov model, belonging to the latest generation of Markov-chain based techniques, which introduces non-stationarity in the chain parameters to preserve the long-term behavior of rainfall. The second one is direct sampling, based on multiple-point statistics, which aims at simulating a complex statistical structure by reproducing the same data patterns found in the training data set. The two techniques are tested against each other in the simulation of a synthetic daily rainfall time-series showing an irregular alternation of two regimes with a specific statistical signature. The results cast a new light on the efficiency of the two approaches in reproducing a complex signal for which the prior structure is unknown, under different data availability scenarios. Direct sampling can automatically capture high-order statistical relations but it needs a sufficiently informed training data set to avoid under-representation of daily extremes. Conversely, the modified Markov model can only simulate time dependences contained in its prior structure but it can extrapolate extremes by using a kernel based technique and approach the reference distribution even in case of scarce training data. The complementarity of the two techniques suggests the potential of a future hybrid approach.

3.1 Introduction

It has been observed that daily rainfall can have a chaotic behavior [20, 80, 182, 181, 131, 85, 183], requiring high-order statistical or deterministic models [165, 92] to generate realistic simulations and reliable short- and long-term predictions.

The Markov-chain (MC) family of techniques is the most common approach to simulate daily rainfall since the 60's [e.g., 54]; it treats rainfall occurrence and amount separately as two joint random variables. The simulation of both quantities is usually sequential and conditional to recent past (low-order time dependence). The drawback of classical daily Markov models is the under-representation of the variance of monthly and annual historical wet days and rainfall totals. Even the use of higher-order dependence underestimates higher time-scale variances [88], while dramatically increasing the number of parameters. A key improvement has been brought to the latest generation of MC based algorithms, introducing non-stationary parameters consistent with the underlying recent past variations. In particular, the daily rainfall occurrence probability is conditioned using either exogenous climatic variables, for example large-scale atmospheric indicators [70, 18, 89, 220, 77, 208, 212, 93, 78, 24], change factors available from climate model projections [94], lower-frequency daily rainfall covariates [213, 23, 84, 90] or indexes based on the recent past rainfall behavior [68, 67, 127, 126] to reproduce the low-frequency fluctuations observed in the training data set. An alternative strategy allowing the preservation of the mean and variance at multiple scales is model nesting [209, 187, 188, 189], which involves the correction of the generated daily rainfall using a multiplicative factor to compensate the low-order moments bias at the monthly and annual scales.

In this chapter, we compare two of the few techniques proposed in literature that can preserve rainfall statistics up to the decennial scale without any additional information apart from the historical daily rainfall time-series. The first one, representative of the latest generation of MC based algorithms, is the modified Markov model [126], which conditions the Markov chain parameters on the number of past wet days to impart the low frequency fluctuations. The second one is direct sampling [122], belonging to multiple-point statis-

tics (MPS), a family of geostatistical techniques widely used in spatial data simulation [61, 196, 4, 228, 12, 73, 195, 198]. MPS algorithms are based on the concept of training data set (TI): a data set representative of the simulated variable used to infer the occurrence probability of each event conditionally to multiple neighbors points. This high-order conditioning allows respecting a complex covariance structure by reproducing the same type of patterns as found in the TI at multiple scales. The direct sampling algorithm takes this principle further: instead of defining a conditional pdf, the simulation is generated by sampling with replacement the TI where a pattern similar to the conditioning data is found. MPS has already been successfully applied to the simulation of spatial rainfall occurrence patterns [218]. Direct sampling has been tested as a rainfall time-series generator on data sets from different climate settings [143] and has shown to be a relatively simple and efficient tool for simulating daily rainfall without the need of calibration. On the other hand, the modified Markov model, expressing a non-stationary Markovian dependence, is able to honor the essential aspects of the variability at multiple scales [126].

Both techniques have already been tested on the simulation of rainfall data presenting a stationary seasonality and regular inter-annual behavior. Nevertheless, recent studies confirm that irregular seasonality and non-stationary behavior at different scales are observed in reality. For example, irregular rainfall patterns for different climate types [55, 91, 47], seasonal anomalies correlated to atmospheric circulation indexes [135, 31, 203, 51, 225] and highly variable occurrence of storms [8] are observed. In these cases, a stationary model adapted to different periods of the year is not sufficient to correctly represent this complexity. The aim of this chapter is therefore to focus on the irregular aspects of a complex rainfall signal that become critical to make reliable predictions in some real applications, where, in addition, data and prior information are often limited. Both techniques are tested on the simulation of a synthetic signal composed of an irregular alternation of two regimes, each of which shows a specific Markovian dependence and an extremely variable duration. The synthetic model, for which the prior structure is known in the validation phase, allows testing the efficiency of both techniques in simulating: i) the particular statistical signature of two different rainfall regimes, ii) their irregular alternation and iii) the asymptotic behavior of the simulation under different data availability conditions. The results provide information to guide the choice of the most suitable technique for specific applications and give insight about methodological improvements.

The chapter is organized as follows: in Section 3.2 the two techniques are described as well as the reference model and the method of evaluation, the statistical analysis of the simulated time-series is presented and discussed in Section 3.3 and Section 3.4 is dedicated to the conclusions.

3.2 Methodology

In this section, a brief introduction to the considered simulation techniques is given, focusing on the distinctive elements of each one. We refer to the cited publications for a more detailed description of the algorithms. In the last two subsections, the generation of the time-series used as reference and the methods of evaluation are presented in details.

3.2.1 The modified Markov model

In the fashion of the MC based techniques, the modified Markov model (MMM) is split into two sub-models: for rainfall occurrence R_t and amount Z_t respectively at time t, following a sequential simulation path generating each subsequent day from the beginning to the end of the time-series. R_t is simulated using a variation of an order-1 Markov model where the

probability of having a wet day $P(R_t = 1)$ is conditioned by the previous day state R_{t-1} and a predictor variable vector $\vec{X}_t$, expressing long-term variability and persistence. [126] identified two variables composing $\vec{X}_t$ appropriate for daily rainfall simulation: the 30- and 365-day wetness indexes ($30W$ and $365W$), i.e. the number of wet days found on the past 30- and 365-day running window, allowing conditioning on monthly and annual fluctuations.

The occurrence time-series R_t is simulated with the following procedure:

1. For all calendar days of the year, calculate, on the historical record, the transition probability of the standard first-order Markov model using the observations falling within the moving window of 31 days centered on each day. Denote the transition probability as $p_{11} = P(R_t = 1|R_{t-1} = 1)$ for previous day being wet and $p_{10} = P(R_t = 1|R_{t-1} = 0)$ for previous day being dry.

2. Also estimate the mean, variance and covariance of the higher time scale predictor variables (the elements of $\vec{X}_t$) separately for occasions when the current day is wet/dry and the previous day is wet/dry.

3. To simulate R_t, consider R_{t-1}, ascertain the appropriate critical transition probability to the day t based on the value of R_{t-1}. If $R_{t-1} = 1$ (wet), assign the critical transition probability p as p_{11}; otherwise, assign p_{10}.

4. Calculate the values of $30W$ and $365W$ for t from the available generated sequence. If t is at the beginning of the simulation, without enough days already generated, randomly pick the matching calendar day of a random year from the historical record and calculate the historical $30W$ and $365W$.

5. Modify the critical transition probability p of step 3 using the following equation:

$$p = P(R_t = 1|R_{t-1} = i, \vec{X}_t) = p_{1i}\frac{\frac{1}{\det(V_{1,i})^{1/2}}\exp\{-\frac{1}{2}(\vec{X}_t - \mu_{1,i})V_{1,i}^{-1}(\vec{X}_t - \mu_{1,i})'\}}{\sum_{j=0,1}\frac{1}{\det(V_{j,i})^{1/2}}\exp\{-\frac{1}{2}(\vec{X}_t - \mu_{j,i})V_{j,i}^{-1}(\vec{X}_t - \mu_{j,i})'\}p_{ji}} \quad (3.1)$$

where X_t is the predictor set for R_t, $\mu_{1,i}$ parameters represent the mean $\mathrm{E}(\vec{X}_t|R_t = 1, R_{t-1} = i)$ and $V_{1,i}$ is the corresponding variance-covariance matrix estimated on the historical record. Similarly, $\mu_{0,i}$ and $V_{0,i}$ represent, respectively, the mean vector and the variance-covariance matrix of X when $R_{t-1} = i$, and $R_t = 0$. p_{1i} parameters represent the baseline transition probabilities of the first-order Markov model defined by $P(R_t = 1|R_{t-1} = i)$ and $det(\cdot)$ represents the determinant operation.

6. Compare p with the random variate u_t generated from the standard uniform distribution. If $u_t \leq p$, assign rainfall occurrence $R_t = 1$; otherwise, $R_t = 0$.

7. Move to the next date in the generated sequence and repeat steps 3 to 6 until the desired length of generated sequence is obtained. The underlying hypothesis of normality regarding the joint probability distribution of $\vec{X}_t$ holds in general for the chosen wetness indexes applied to daily rainfall.

The amount Z_t is simulated on wet days of the generated sequence using the kernel density estimation technique proposed in [176]. For each Z_t, a conditional probability density function $f(Z_t|\vec{C}_t)$ conditioned on a predictor variable vector $\vec{C}_t$ is used. For example, $\vec{C}_t$ can be composed by the rainfall amount on previous days or by some other correlated variables. $f(Z_t|\vec{C}_t)$ is built as a sum of weighted kernels, each one associated to an historical datum Z_i. As stated in [126], in the kernel density estimation procedure, the Gaussian reference bandwidth [169] is considered as optimal if the underlying probability density is Gaussian.

However, when dealing with skewed variables, like daily rainfall, using local bandwidths, also known as adaptive kernel, provides flexibility in reducing the variance of the estimates in areas with few observations, and reducing the bias of the estimates in areas with many observations. In this case, an adaptive or local bandwidth estimation procedure as mentioned in [169] is used. The local bandwidth is estimated for each observation in the historical record. This gives a better estimation of the conditional probability density function, especially at the lower boundary of the distribution. As kernel density estimate can lead to rainfall amounts that are less than the threshold amount of 0.3 mm (the minimum non-zero rainfall amount considered in the algorithm), a minimum rainfall amount of 0.3 mm is assigned to such days. A distinctive feature of this non-parametric method is that each kernel weight w_i varies in function of the respective predictor variable vector $\vec{C}_i$ following the formula:

$$w_i = \frac{\exp(-\frac{1}{2\lambda^2}\{[\vec{C}_t - \vec{C}_i]\psi\}[\mathbf{S}_{\mathrm{CC}}]^{-1}\{[\vec{C}_t - \vec{C}_i]\psi\}^T)}{\sum_{j=1}^{N}\exp(-\frac{1}{2\lambda^2}\{[\vec{C}_t - \vec{C}_j]\psi\}[\mathbf{S}_{\mathrm{CC}}]^{-1}\{[\vec{C}_t - \vec{C}_j]\psi\}^T)} \tag{3.2}$$

where by definition all the w_i sum to 1, $\vec{C}_t$ is the conditioning vector computed on the simulated day, ψ is a diagonal matrix of influence weights to give different importance to each element of $\vec{C}_t$ in conditioning the simulation, $\mathbf{S}_{\mathrm{CC}}$ is the covariance matrix of $\vec{C}_t$ (estimated form the historical record) and N the number of considered training data. In this chapter $\vec{C}_t$ is reduced to one variable (see section 3.2.4), consequently $\psi = 1$. According to Equation 3.2, higher weight is given to data presenting $\vec{C}_i$ similar to $\vec{C}_t$. This technique allows generating extremes not contained in the historical data set since, with the use of kernels, the support of $f(Z_t|\vec{C}_t)$ extends beyond the considered historical data.

3.2.2 The direct sampling technique

Direct sampling (DS) is a non-parametric resampling technique from the MPS family based on a pattern-similarity rule. We use the *DeeSse* implementation [193] which allows generating the rainfall occurrence and amount at the same time. The simulation follows a random path which visits a time referenced empty vector $\vec{x}$ called simulation grid (SG), becoming progressively populated until rainfall at all time steps is simulated. The target variable $Z(x_t)$ is generated by sampling with replacement of the training data set (TI) $\vec{y}$ composed of historical data. The sampled data are chosen conditionally to a neighborhood varying throughout the course of the simulation. We report here the main workflow of the algorithm, for a detailed description of the application to rainfall simulation and comparison with other resampling techniques see [143]:

1. Select a random position x_t of the SG that has not yet been simulated.
2. To simulate the rainfall amount (and occurrence) $Z(x_t)$: retrieve a data event $\vec{d}(x_t)$, i.e. a group of already simulated or given neighbors of x_t, according to a fixed time interval $t \pm R$. $\vec{d}(x_t)$ consists of at most the N informed time steps closest to x_t inside the mentioned interval. The size and configuration of $\vec{d}(x_t)$ is therefore limited by the user-defined parameters N and R, and the number of already informed neighbors inside the considered window.
3. Visit a random time-step y_i in the TI, and retrieve the corresponding data event $\vec{d}(y_i)$.
4. Compute a distance $D(\vec{d}(x_t), \vec{d}(y_i))$, i.e. a measure of dissimilarity between the two data events. For categorical variables (e.g. the dry/wet rainfall sequence) the proportion of non-matching elements of $d(\cdot)$ is used as criterion, while for continuous variables the choice is the mean absolute error.

5. If $D(\vec{d}(x_t), \vec{d}(y_i))$ is smaller than a fixed threshold T, assign the value of $Z(y_i)$ to $Z(x_t)$. Otherwise repeat from step 3. to 5. until the value is assigned or a prescribed TI fraction F has been scanned. T is expressed as a fraction of the total variation shown by Z in the TI. For example, $T = 0.05$ allows $D(\vec{d}(x_t), \vec{d}(y_i))$ up to 5% of this total variation. In case of a categorical variable, $T = 0.05$ allows a mismatch between $\vec{d}(x_t)$ and $\vec{d}(y_i)$ for 5% of the composing neighbors.

6. If the prescribed TI fraction F has been covered by the scan, assign to $Z(x_t)$ the scanned datum $Z(y_i*)$ that minimizes D.

7. Repeat the whole procedure until all the SG is informed.

The same process is applicable to a multivariate data set, where a vector $\vec{Z}(x_t)$, composed of rainfall amount and some auxiliary variables, is simulated instead. The parameters N_k, R_k and T_k allow defining different pattern dimensions and acceptance threshold for each k-th variable.

As explained in [143], the technique departs from the resampling or bootstrap algorithms previously proposed for hydrological application [223, 110, 109, 155, 24, 217, 33] for the following main aspects: i) The classical resampling techniques apply a fixed and generally low-order time-dependence, while DS uses a variable and high-order time-dependence. This is achieved by considering multiple-neighbor patterns which change size and configuration during the simulation (see section 3.2.3). ii) The simulation path (in the SG) is always linear in resampling techniques, while it is random using DS, allowing conditioning on sparse past and future time-steps of the target variable. iii) Using DS, the conditional pdf remains implicit since its computation is not needed: the historical record is randomly visited instead and the first datum presenting a distance below the acceptance threshold is sampled. This allows extending the order of the time dependence without the problem of the high-dimensional conditional pdf estimation which limits the degree of conditioning in bootstrap techniques [175].

3.2.3 Fixed versus variable time dependence

In this section, we emphasize the different ways in which the two algorithms deal with time dependence (figure 3.1), since this is a crucial aspect regarding the simulation of rainfall heterogeneity at multiple scales. Both techniques operate in a multivariate framework where it is possible to have conditioning variables describing large scale fluctuations, e.g. the wetness indexes (see section 3.2.1), contained in the predictor variable vectors $\vec{X}_t$ and $\vec{C}_t$ for MMM and among the auxiliary variables for DS. This helps the preservation of the distribution at larger scales. On both the target and these conditioning variables, MMM applies a fixed time dependence, meaning that the conditioning time lags are rigorously defined by the order of the Markov chain and remain constant throughout the course of the simulation. Conversely, DS makes use of a variable conditioning pattern: following a random simulation path where the SG becomes more and more populated in a random order, the conditioning neighborhood changes progressively from a large-scale, sparse-neighbors to a small-scale, close-neighbors pattern by considering the closest N informed time steps inside the search window of length $2R + 1$. The conditioning time lags vary for each simulated datum and they cannot be defined a priori, but the order of the conditioning is limited by the parametrization of the search window. Although not being precisely defined as the fixed one used by MMM, this type of conditioning may allow a better preservation of high-order statistics contained in the historical data by generating more realistic multiscale patterns. On the other hand, MMM is focused on a specific choice of long-term statistical indicators contained in the vector $\vec{X}_t$

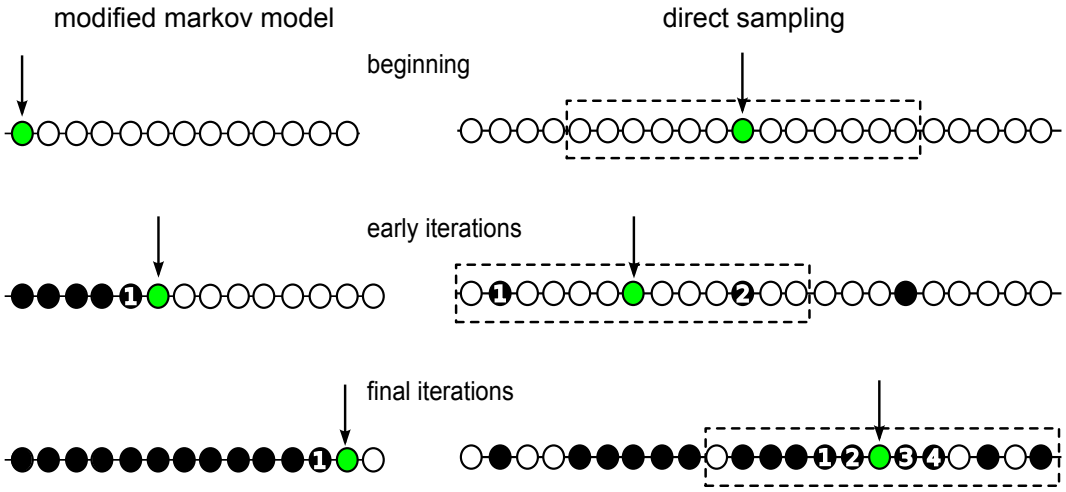


Figure 3.1: Comparison of the two considered simulation techniques: the MMM, using a linear simulation path together with an order-1 time dependence, and DS, using a random simulation path and a variable time dependence. The arrow indicates the time step being simulated, the black time steps are already simulated and the numbered ones are used for conditioning. For DS, the sketch illustrates the behavior using $N = 4$ and $R = 6$, the search window being represented by the dashed line.

and on the order-1 Markov-chain time dependence (see section 3.2.1). This choice may be more appropriate to describe the essential aspects of the heterogeneity.

3.2.4 Experiment setup

In order to test the performance of both methods in capturing the complex properties of a rainfall signal, a synthetic time-series is generated and used as a reference. The desired properties of this time-series are that it presents a pseudo-chaotic behavior, variable auto-correlation properties and significant high-order variability (see section 3.2.5). The term pseudo-chaotic is used here to indicate a stochastic model that mimics a dynamical system unpredictable and highly sensitive to initial conditions, reflecting a complexity similar to chaotic systems. This behavior can characterize the rainfall heterogeneity in some real cases (see section 3.1). Random samples from the reference time-series are used as training data sets for the modified Markov model and DS. The two techniques are tested on the simulation of the reference signal under four different scenarios (table 3.1), each of which considers a different training data amount: 1 million, 10'000, 1000 and 100 days. This allows analyzing the algorithms performance under different data availability conditions. The experiment for groups 2, 3 and 4 are repeated 3000 times considering each time a different random sample of the reference as training data to assure independence of the results with respect to the used training data. A preliminary convergence test (not shown here) on the variance of the simulated data showed that this number of realizations is sufficient to cover exhaustively the variability of the simulation with respect to the considered reference data. Conversely, for the first group, 10 realizations are generated using the whole reference time-series as training data set. For all groups, the simulated time-series are 1-million-day long as the reference signal. The arbitrary choice of the realizations number for the first group is justified by the fact that, using the whole reference as TI and being of the same size, the realizations do not show a significant statistical variability.

The two algorithms use the same training data and the same auxiliary variables used in

Table 3.1: Summary of the training data sets and auxiliary variables used in the setup of the two algorithms. The auxiliary variables listed are: the 30- and 365-day wetness indexes ($30W$ and $365W$), the dry/wet sequence (W_d), the historical rainfall amount (Z). In the MMM algorithm, the auxiliary variables compose the vectors $\vec{X}_t$ and $\vec{C}_t$ to condition rainfall occurrence and amount simulation respectively.

Group	Training data amount [days]	Number of data sets used	Realizations per data set	MMM setup	DS setup
1	1 million	1 (reference)	10	$\vec{X}_t = [30W, 365W]$,$\vec{C}_t = Z_{t-1}$	$365W, W_d, Z$
2	10'000	3000	1	$\vec{X}_t = [30W, 365W]$, $\vec{C}_t = Z_{t-1}$	$365W, W_d, Z$
3	1000	3000	1	$\vec{X}_t = [30W, 365W]$, $\vec{C}_t = Z_{t-1}$	$365W, W_d, Z$
4	100	3000	1	$\vec{X}_t = 30W$, $\vec{C}_t = Z_{t-1}$	$30W, W_d, Z$

Table 3.2: The multivariate setup used with DS. The parameters are: search time interval radius R [days], maximum number of neighbors considered N [days] and distance threshold T [-].

Variable	R	N	T
1) $365W$	5000	21	0.05
2) W_d	10	5	0.05
3) Z	5000	21	0.05

real application, apart from the theoretical variables that describe the position of the day in the year in the standard direct sampling setup for rainfall. These are not used here since the regular annual seasonality is not present in the reference. All auxiliary variables are automatically computed from the simulated and training rainfall data without using any additional information. As shown in table 3.1, the predictor variable vector $\vec{X}_t$ used in the MMM occurrence model is composed of the 30- and 365-day wetness indexes ($30W$ and $365W$). In the MMM amount model, the conditioning vector $\vec{C}_t = Z_{t-1}$ (i.e., the generated rainfall amount in the previous day) is used, therefore applying an order-1 dependence. The Gaussian kernel with adaptive bandwidth is also used as explained in section 3.2.1.

Regarding the DS technique, a multivariate TI is used including the following variables: 1) $365W$, 2) The dry/wet sequence (W_d) (i.e., a categorical variable indicating the position of a day inside the rainfall pattern, taking on the labels: 0 dry day, 1 wet day with wet day either side, 2 solitary wet day, and 3 wet day at the beginning or at the end of a wet spell) and 3) rainfall amount (Z, mm). The DS parameters used with each variable are shown in table 3.2. For example, for the variable Z, $R = 5000$ and $N = 21$, meaning that the pattern considered for conditioning will be composed of at most the 21 already simulated data closest to the time step being simulated, at a time distance of at maximum 5000 days (about 14 years) in the past or future. Moreover, the distance threshold value $T = 0.05$, used to compare patterns between the TI and the SG (see section 3.2.2 point 5.), corresponds to 5% of the total variation of the simulated variable.

It is worth noting that $30W$ is only used with DS in the last simulation group, where the 100-day TI do not allow the computation of $365W$. In the other groups the use of $30W$ is not needed, since the high-order conditioning applied at the daily scale is generally sufficient to preserve the fluctuations at the monthly scale.

3.2.5 The reference signal

The rainfall time-series used as reference is a synthetic daily time-series generated with an occurrence model composed of two alternating regimes showing a different Markovian time dependence structure. The transition from one regime to the other depends on the sum of

wet days in the previous 200 days, creating a pseudo-chaotic regime alternation. Using a synthetic model allows disposing of a very long reference time-series for which the exact prior structure is known in the validation phase. This way, it is possible to illustrate how each of the two considered techniques can detect and reproduce a highly irregular time dependence structure and preserve its asymptotic behavior. The procedure used to build the reference time-series is described in the following, with l_i being the model parameters.

A starting sequence of 200 days is generated using the model $Z_t = 10N_t\mathbb{I}_{N_t>1}$ where N_t is a random number from the standard normal distribution and $\mathbb{I}_{N_t>1} = 1$ if $N_t > 1$ and 0 otherwise. For the successive time steps, the rainfall occurrence time-series U_t (being $U_t = 1$ for a wet day and $U_t = 0$ for a dry one) is simulated by following two possible regimes. In regime A the probability of having a wet day is conditioned by the MC rule:

$$P(U_t = 1 \mid U_{t-l_1}, U_{t-l_2}) \tag{3.3}$$

with the following conditional probability values:

$$\begin{array}{l} P(U_t = 1 \mid U_{t-l_1} = 0, U_{t-l_2} = 0) = l_5 \\ P(U_t = 1 \mid U_{t-l_1} = 0, U_{t-l_2} = 1) = l_6 \\ P(U_t = 1 \mid U_{t-l_1} = 1, U_{t-l_2} = 0) = l_7 \\ P(U_t = 1 \mid U_{t-l_1} = 1, U_{t-l_2} = 1) = l_8 \end{array} \tag{3.4}$$

Regime A is active if the sum of wet days in the previous 200-day period meets the condition:

$$\sum_{i=1}^{200} U_{t-i} > l_3 \tag{3.5}$$

otherwise regime B takes place, with the rule:

$$P(U_t = 1 \mid U_{t-l_4}) \tag{3.6}$$

with:

$$\begin{array}{l} P(U_t = 1 \mid U_{t-l_4} = 1) = l_9 \\ P(U_t = 1 \mid U_{t-l_4} = 0) = l_{10} \end{array} \tag{3.7}$$

Since defining the parameters of such a model is not straightforward, we formulated an objective function and used it to find an appropriate parametrization. The parameter vector $\vec{L} = (l_1, l_2, ..., l_{10})$ is in part optimized such that the model presents a highly irregular alternation of both regimes with a distinct time dependence signature and a finite spell duration. To contain the computational burden, the following elements of $\vec{L}$ are arbitrarily defined: $l_4 = 1$, $l_6 = 0.51$, $l_7 = 0.45$, $l_8 = 0.64$, $l_9 = 0.65$. The other elements of $\vec{X}$ are calibrated to minimize the following objective function:

$$O(\vec{X}) = \begin{cases} -10|ac(Z_\mathrm{A}) - ac(Z_\mathrm{B})| - |\bar{S}_\mathrm{A} - \bar{S}_\mathrm{B}| & \text{if } 10 < \bar{S}_\mathrm{A} < 70 \text{ and } 0 < \bar{S}_\mathrm{B} < 70 \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

where Z_A and Z_B are the portions of the generated signal belonging to the regime A and B respectively, $ac(\cdot)$ is the lag-1 autocorrelation coefficient and $\bar{S}$ is the mean length of the time-series segments belonging to one regime. Such numerical optimization, constituting a mixed integer problem, is solved with a genetic algorithm [30], which is often used to find the global minimum of highly non-linear or non-continuous functions. The values $l_1 = 6$, $l_2 = 12$, $l_3 = 95$, $l_5 = 0.30$, $l_{10} = 0.31$, obtained by minimizing $O(\vec{X})$, lead to a similar wetness level for the two regimes ($P(U_t = 1) \approx 0.46$) but a different time dependence: regime B presents a higher day-to-day persistence since U_t is correlated to U_{t-1}, while, in regime A, U_t is

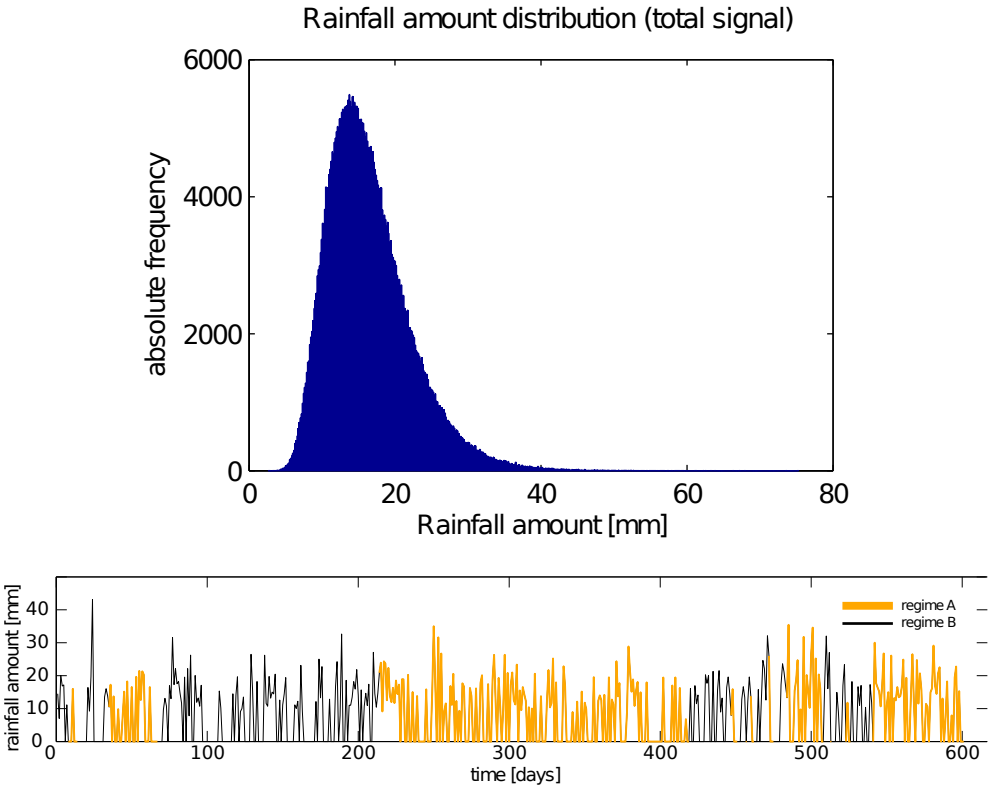


Figure 3.2: Rainfall amount histogram and a random sample of the synthetic reference signal, with alternating regimes: A=lag-6 and -12 Markov chain and B=lag-1 Markov chain. The rainfall amount on all rainy days have been generated from the same log-normal probability distribution.

correlated to lags U_{t-6} and U_{t-12}, resulting in a more discontinuous dry/wet pattern. A random sample of the reference is shown in figure 3.2.

The rainfall amount on wet days of both regimes (histogram in figure 3.2) is simulated by randomly sampling the log-normal distribution $\ln \mathcal{N}(2.74, 0.34)$, which is fitted on the starting sequence (first 200 days, then eliminated from the reference signal). A time-series of 1 million days is so obtained by using the explained model and considered as the reference for the tests.

3.2.6 Evaluation

The statistical indicators used to test the performance of both techniques describe the overall time-series as well as the specific statistical signature of the two regimes. The purpose of studying the regime-specific statistics allows verifying whether the specific statistical signature of each regime is detected and preserved in the simulation. The probability distribution of daily rainfall amount on wet days, the annual and decennial amount of the total signal are compared using qq-plots. The two-regime sequence is then reconstructed inside each simulated time-series with the original criterion used to generate the reference: the number of wet days over the previous 200 days. This allows separating the two regimes inside the simulations, to study their alternation and their statistics separately. A qq-plot is used to compare the quantiles of the two-regime spell length distribution and the dry/wet spell distribution. To verify the accuracy in the preservation of the time dependence, the sample autocorrelation function (ACF) is computed separately for both regimes as well as on the total signal. Finally, the minimum moving average (MMA), i.e. the minimum value found in the moving average filter using different moving window sizes, is computed on the total signal to compare the simulation of the long-term behavior up to the centennial scale.

3.3 Results

The results about the mentioned statistical indicators are analyzed in the following and a summary is given at the end of the section.

3.3.1 Multiple scale distribution

The comparison with the reference distribution for each simulation group and technique is shown in figure 3.3. At the daily scale and using the whole reference as training data set (1 million-day group), both techniques can accurately preserve the marginal distribution: the realizations median shows virtually no bias and the narrow region between the 05-95 percentiles of the simulations indicates very low uncertainty. Since the reference exhibits a skewed distribution (see figure 3.2), training data sets smaller than the reference lack information about the distribution tail, as indicated by a progressive under-representation of the extremes in the other simulation groups. MMM, based on conditional kernel density estimation, resolve in part this issue by extrapolating extreme values not present in the training data set and approaching the reference distribution even for extremely small available data amounts (figure 3.3,a). Conversely, DS, being based on resampling, is limited to the range of values found in the TI and cannot represent properly the asymptotic behavior at the daily scale if the TI is not sufficiently informative. The truncation of the distribution at a series of specific values visible in the qq-plot clearly illustrates this phenomenon (figure 3.3,b).

The truncation of the distribution does not affect larger temporal scales, for which both techniques can preserve an unbiased distribution up to the decennial rainfall, showing a modest over-dispersion only when using a 100-day training data set(figure 3.3,c,d,e,f). This result is achieved with the help of the wetness indexes, tracking the low-frequency fluctuations

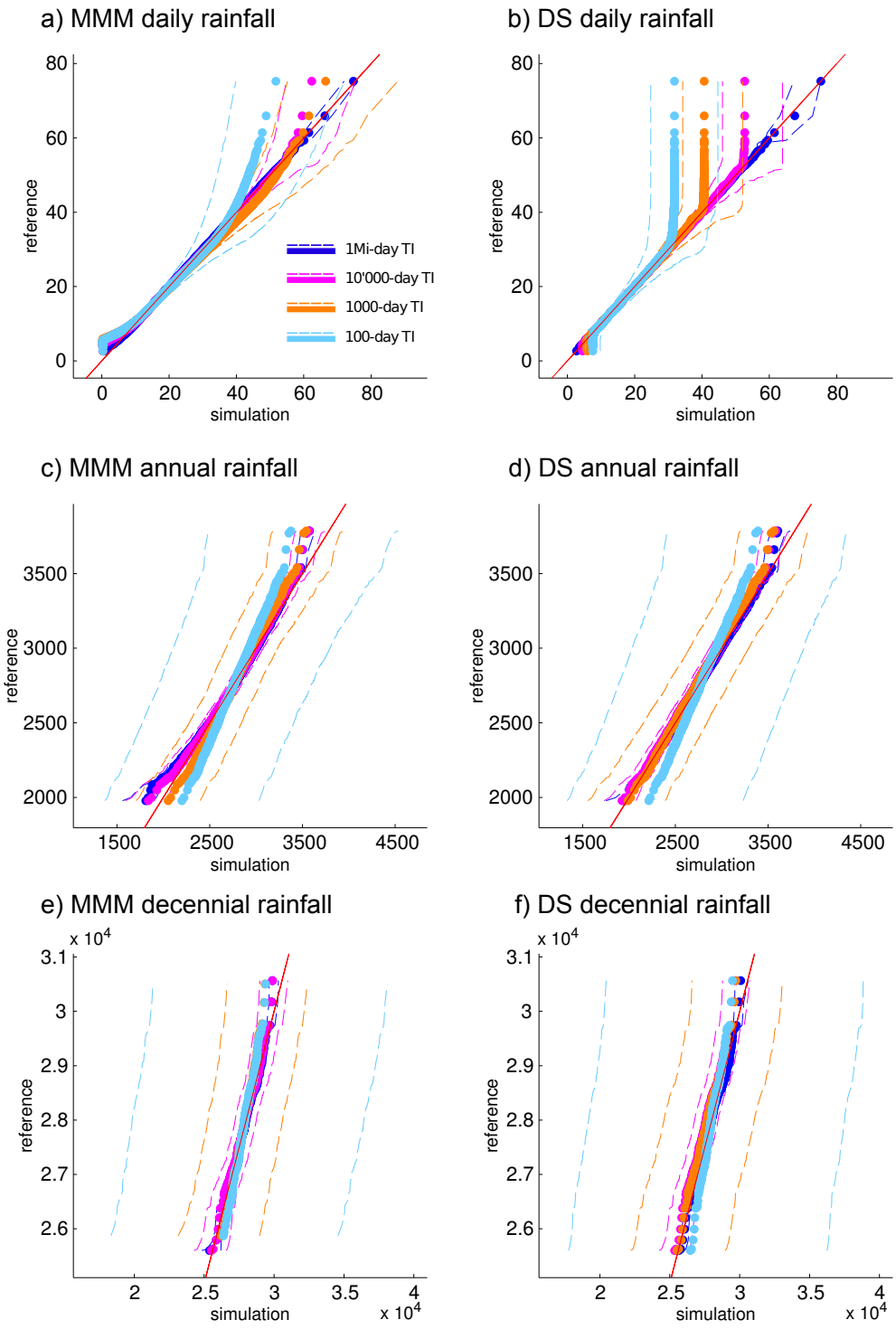


Figure 3.3: qq-plots of the empirical probability rainfall amount distributions (mm), showing for each quantile: the median of the realizations (dots), 5th and 95th percentiles (dashed lines). The bisector (solid straight line) indicates the exact quantile match.

(see section 3.2). Both techniques have a comparable performance: we observe only a slight tendency to overestimate low annual rainfall values by MMM when using large training data sets. DS is able to re-aggregate the sampled TI values in different ways and correctly explore the uncertainty at large scales. Nevertheless, with both techniques, this uncertainty is large when extremely limited training data sets are used: the 05-95 percentile boundaries of the realizations are very wide when using 100-day or 1000-day training data groups, meaning that the used training data sets of this size present a variable statistical content. This result suggests that, as expected, a longer historical record is needed to represent the large scale variability and reduce the uncertainty of the simulations.

3.3.2 Regime alternation

As mentioned in section 3.2.6, the alternation between regimes A and B is reconstructed inside the generated time-series and the spell length distribution of each one is shown in figure 3.4. Even if the reference model is calibrated to assure a continuous regime alternation, the very skewed spell distribution still suggests an irregular behavior presenting large variability, with a maximum spell duration up to about 1500 days for both regimes. The region delimited by the 05-95 percentiles of the simulations indicates that the uncertainty increases when reducing the training data amount. The 100-day group constitutes the degenerate case for which the regime transition rule, based on the 200-day wetness index, is not observable and the regime alternation cannot be exhaustively represented. In fact, the 05-95 percentile boundaries (not visible for this group) correspond to null and infinite spell duration respectively, meaning that the whole generated time-series belongs to one single regime. For large training data sets (1 million- and 10'000-day groups) the uncertainty is significantly smaller using both techniques and the distribution is preserved fairly well, with a modest under-representation of the very upper quantiles for regime A (figure 3.4,a,b). The main difference in their performance is observed in the 1000-day group, where MMM shows a negative bias larger than the one obtained in the 100-day group (figure 3.4,a,c). This indicates that a larger data set is needed to calibrate the parameters of MMM using the 30- and 365-day wetness indexes. On the contrary, using the 30-day wetness index only and a 100-day training data set results in a smaller bias but larger uncertainty. Representativeness of the training data set plays again a fundamental role: since no information about the irregular regime alternation is contained in the prior structure of both simulation techniques, the distribution is simulated accurately only when the training data set contains a sufficient repetition of the two-regime transition.

3.3.3 Time dependence structure

The specific short-term time dependence structure of the total signal as well as the two separate regimes is analysed using the sample autocorrelation function (ACF, figure 3.5). According to its occurrence model, the reference signal shows a distinctive autocorrelation level for lags 6 and 12 in regime A and for lag 1 in regime B (red lines) with the total signal presenting a mixture of both.

The two simulation techniques show a different behavior: on the total signal, MMM simulates the lag-1 dependence correctly, but does not preserve the lag-6 and lag-12 autocorrelation, with a subsequent underestimation of the persistence (figure 3.5,a). This is due to the fact that only the lag-1 dependence is considered in the MMM occurrence model. For this reason, the model is weak to any other high-order time dependence observable in the training data. Conversely, DS can reproduce the whole time dependence structure of the total signal with no need for any prior information about it (figure 3.5,b). This is achieved by applying a random simulation path and a variable conditioning pattern composed of mul-

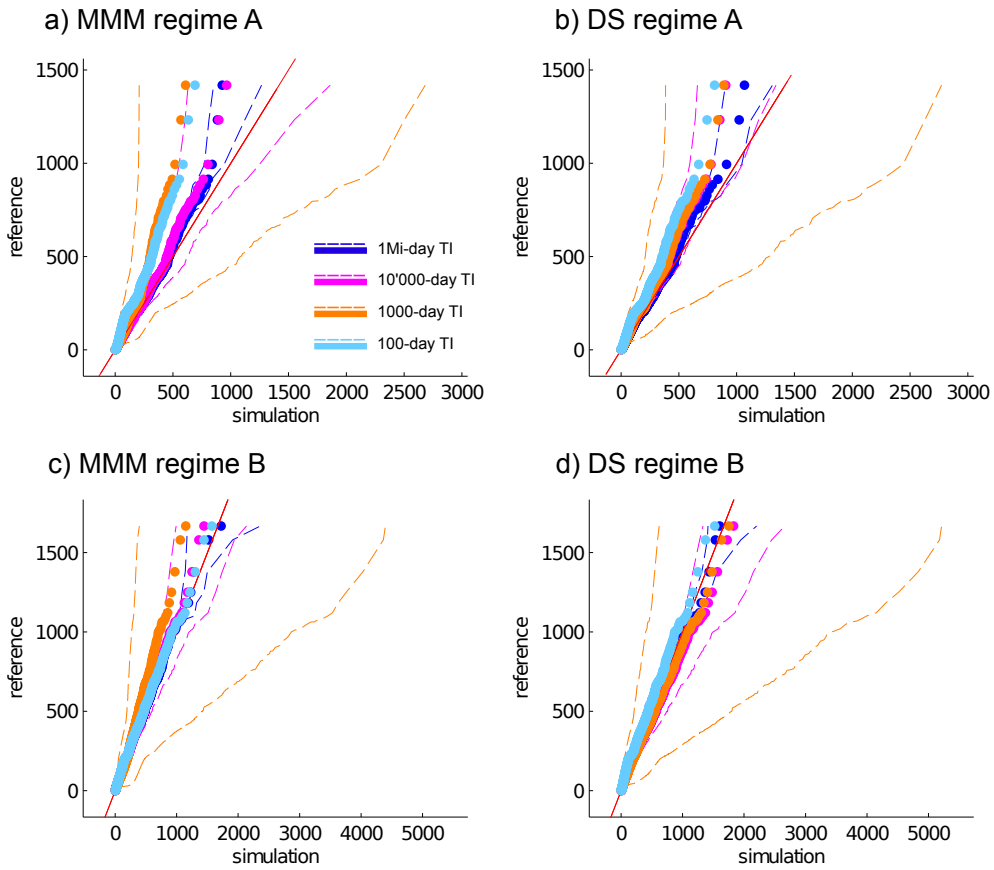


Figure 3.4: qq-plots of the two-regime-spell length distributions (days) for each simulation group (different colours), showing for each quantile: the median of the realizations (dots), 5th and 95th percentiles (dashed lines). The bisector (solid straight line) indicates the exact quantile match.

tiple neighbors (see section 3.2.2). In other words, a variable high-order time dependence is considered during the simulation, which allows preserving on average the autocorrelation at any lags. The advantage of this feature is that complex non-linear time dependencies are simulated more easily than using a parametric technique. It is worth noting that, for highly autocorrelated signals, the autocorrelation is not exactly preserved using DS. Even with the most appropriate setup, the resampling process adds a small noise to the data, which is detectable on very smooth signals and may need a post-processing treatment. In case of daily rainfall this effect is negligible since the signal presents a very low autocorrelation.

Both techniques fail in simulating the time dependence for the two separate regimes. In particular, MMM always reproduces the lag-1 autocorrelation estimated from the total training data set (figure 3.5,c,e) and DS does the same with the overall dependence structure (figure 3.5,d,f). This means that non-stationarity in the time dependence, due to the regime alternation, cannot be automatically captured and preserved in the simulation. Designing an ad-hoc model structure based on the analysis of the training data set is therefore necessary with both techniques in order to correctly simulate the non-stationarity. For example, information about an irregular regime shift can be incorporated in the DS setup using a discrete auxiliary variable as it is done with the dry/wet sequence (see section 3.2.2). The mentioned variable would be simulated together with the rainfall helping the simulation of the two-regime alternation. Conversely, using a parametric approach like MMM, a regime switch could be accommodated in the MC structure as it has been done for the reference signal generation (see section 3.2.5). In both cases, the main point is to catch the relevant non-stationary features from the available data in a preliminary analysis, which may not be straightforward in case of highly irregular fluctuations.

3.3.4 Dry/wet pattern and long-term behavior

The dry/wet spell length distribution of the total signal is compared with the reference using qq-plots (figure 3.6,a,b,c,d). When using large training data sets (1 million- and 10'000-day groups), both techniques can preserve the dry/wet spell distribution accurately. Reducing the available training data, we observe in both cases a large uncertainty in the upper quantiles, but DS tends to underestimate the extreme dry and wet spells more than MMM. The latter preserves, in fact, a fairly accurate distribution of the simulation median.

The minimum moving average (MMA, figure 3.6,e,f), i.e. the minimum value found on the moving average filter of the rainfall amount, expressed as a function of different moving window lengths, gives information about the accuracy in preserving the variability of the signal at different scales. According to the maximum dry spell length, the reference shows a zero-valued MMA until about a 30-day moving window length. Then it progressively increases up to a rainfall amount of about 7 mm, showing little variation after the 12-years window length. This suggests that the signal may present an effective stationarity at the decennial scale. Both techniques preserve the reference behavior quite accurately with a modest over-estimation in the 100-day group. The performance of DS is superior in case of large training data sets but MMM remains more reliable in case of limited data availability.

3.3.5 Summary

As shown in the previous sections, the two considered algorithms present a different behavior with respect to various characteristics of the signal and training data amounts considered. The relative error $\Delta = (s - r)/r$ (r = reference, s = simulations median) is calculated on a selection of indicators (table 3.3) to summarize the average performance of the two techniques. Positive values indicate overestimation and negative ones underestimation: for example ΔQ95=-0.50 indicates that the 95-th percentile has been underestimated by 50%.

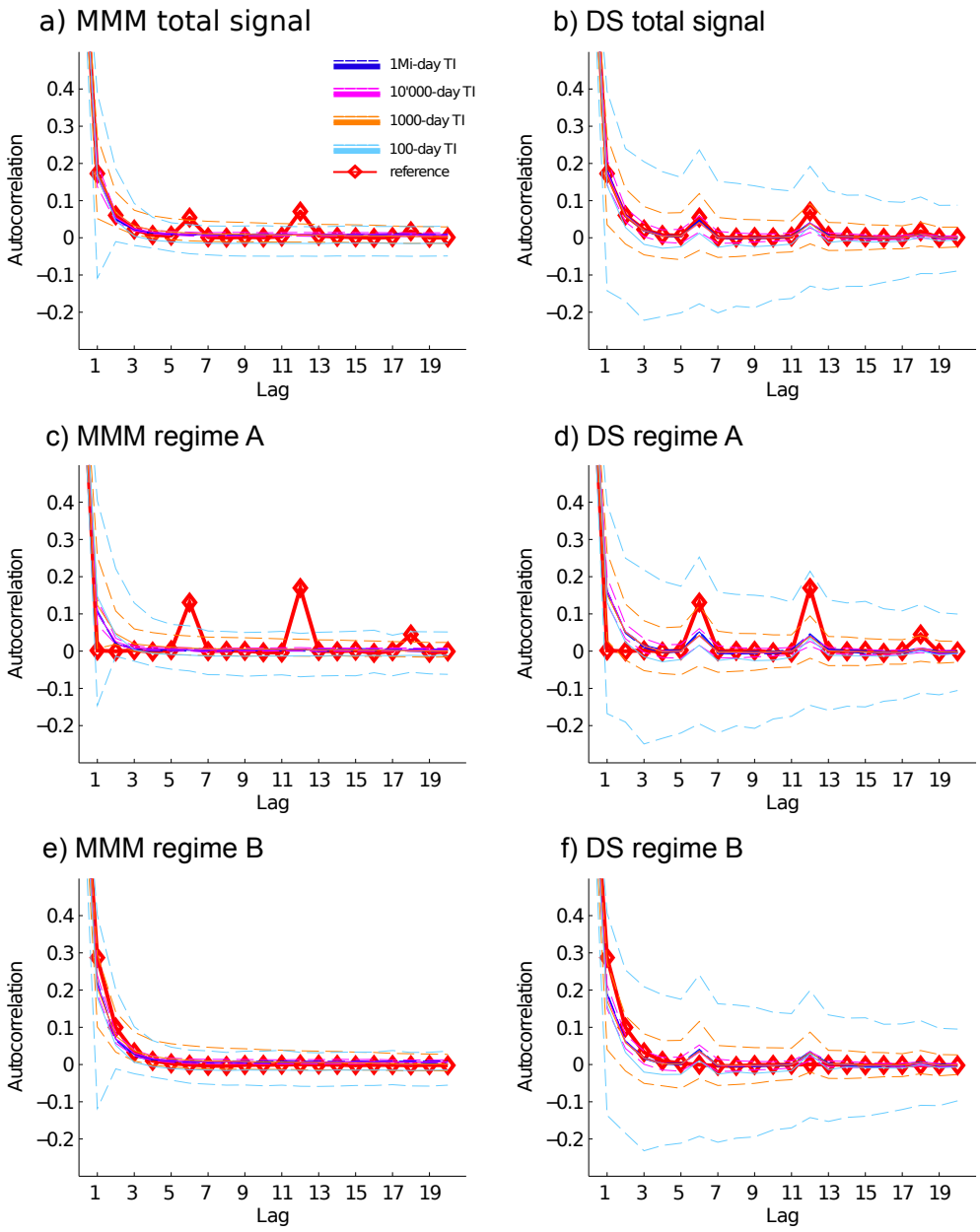


Figure 3.5: Sample ACF of the total signal and the two regimes for each simulation group (different colours). Median of the realizations (solid lines), 5th and 95th percentiles (dashed lines). The red line indicates the reference.

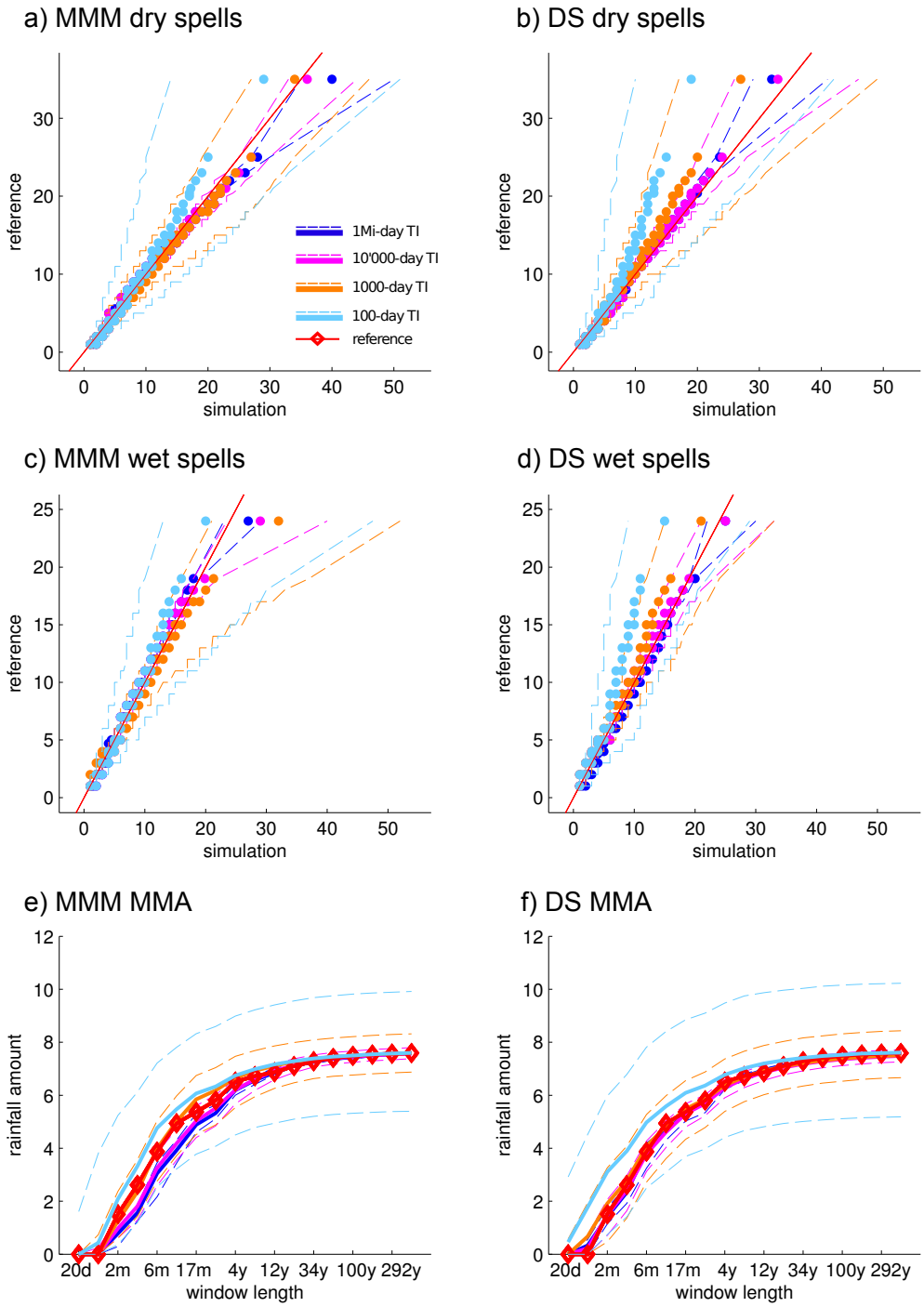


Figure 3.6: Top: qq-plots of the dry and wet spell length distributions (days) for each simulation group (different colours), showing for each quantile: the median of the realizations (dots), 5th and 95th percentiles (dashed lines). The bisector (solid straight line) indicates the exact quantile match. Bottom: minimum moving average (MMA) of the rainfall amount (mm) for different moving window lengths (d-days, m-months, y-years). Median of the realizations (solid line), 5th and 95th percentiles (dashed lines) and the reference (red line).

Table 3.3: Selection of indicators summarizing the average performance of the techniques. The relative error of the simulations median is considered for: the i-th quantile of the distribution (ΔQi), the i-th lag of the autocorrelation function (ACF, $\Delta \text{lag}i$) and the minimum moving average using a n-months-long moving window (MA, Δnm). For each error indicator, a couple of values referring to the two algorithms is given. Couples highlighted in green indicate a superior performance by MMM and in cyan a superior performance by DS.

sim. group		daily amount		annual amount		10-y amount		regime A spells		regime B spells	
		ΔQ99	Δ Q100	ΔQ05	ΔQ95	ΔQ05	ΔQ95	ΔQ99	ΔQ100	ΔQ99	ΔQ100
MMM	1Mi	0.00	-0.01	-0.01	-0.01	-0.01	0.00	-0.03	-0.35	-0.06	0.03
DS	1Mi	-0.00	-0.00	0.00	-0.00	-0.01	0.00	-0.01	-0.25	-0.00	-0.04
MMM	10'000	0.03	-0.17	-0.00	-0.01	-0.01	-0.00	-0.06	-0.32	-0.10	-0.13
DS	10'000	0.00	-0.30	-0.01	-0.02	-0.02	-0.01	-0.13	-0.36	0.11	0.10
MMM	1000	0.06	-0.12	0.03	-0.02	0.01	-0.01	-0.35	-0.57	-0.27	-0.31
DS	1000	-0.01	-0.46	0.00	-0.02	-0.01	-0.01	-0.14	-0.37	0.07	0.05
MMM	100	0.02	-0.31	0.05	-0.04	0.02	-0.01	-0.35	-0.51	-0.08	-0.06
DS	100	-0.09	-0.58	0.06	-0.04	0.01	-0.01	-0.26	-0.43	-0.13	-0.09

sim. group		ACF total signal			ACF regime A			ACF regime b		
		Δlag1	Δlag6	Δlag12	Δlag1	Δlag6	Δlag12	Δlag1	Δlag6	Δlag12
MMM	1Mi	-0.00	-0.85	-0.88	56.89	-0.98	-0.98	-0.24	-23.70	20.78
DS	1Mi	0.05	-0.12	-0.44	88.27	-0.60	-0.75	-0.34	-123.79	99.04
MMM	10'000	-0.02	-0.82	-0.87	55.72	-0.96	-0.97	-0.25	-27.24	22.34
DS	10'000	0.03	-0.27	-0.63	88.78	-0.69	-0.84	-0.36	-108.58	66.59
MMM	1000	0.00	-0.77	-0.89	67.35	-0.94	-0.97	-0.28	-38.78	17.08
DS	1000	-0.04	-0.23	-0.46	82.51	-0.67	-0.77	-0.40	-116.63	99.69
MMM	100	-0.05	-1.25	-1.21	80.44	-1.08	-1.08	-0.38	43.44	-49.60
DS	100	-0.18	-0.78	-0.55	71.49	-0.88	-0.80	-0.45	-34.07	89.27

sim. group		dry spells		wet spells		MMA		
		ΔQ99	ΔQ100	ΔQ99	ΔQ100	Δ2m	Δ6m	Δ17m
MMM	1Mi	0.00	0.14	-0.10	0.12	-0.48	-0.21	-0.09
DS	1Mi	0.00	-0.09	0.00	0.04	-0.06	0.01	0.00
MMM	10'000	0.00	0.03	-0.10	0.21	-0.36	-0.16	-0.06
DS	10'000	0.00	-0.06	-0.10	0.04	-0.02	-0.02	-0.01
MMM	1000	0.00	-0.03	0.00	0.33	-0.17	0.03	0.09
DS	1000	-0.08	-0.23	-0.10	-0.12	0.27	0.05	0.02
MMM	100	-0.08	-0.17	-0.10	-0.17	0.38	0.23	0.13
DS	100	-0.25	-0.46	-0.30	-0.38	1.06	0.29	0.13

In accordance with the results shown in previous publications [126, 127, 143], it is shown here that both techniques can generate replicates of the same size as the training data set preserving the rainfall variability at multiple scales. The error in the daily rainfall amount distribution up to the decennial scale is in fact negligible for the 1-million group. Reducing the available data amount, MMM can extrapolate extremes by using a conditional kernel smoothing technique and approach the reference distribution at the daily scale even for very scarce data sets. DS, on the contrary, remains limited to the range of data found in the TI, showing a large error in the upper quantiles of the daily rainfall distribution. At higher scales, both techniques can preserve an unbiased distribution even when using a small amount of daily data. Nevertheless, the uncertainty shown by the 05-95 percentile boundaries of the realizations (figure 3.3) suggests that, for a reliable simulation of the considered reference signal, a 10000-day training data set should at least be used. This principle is confirmed by all the results shown in the previous sections.

Both techniques have a comparable performance regarding the regime A and B spell length distribution: the pseudo-chaotic reference model leads to an extremely variable regime duration and a highly skewed distribution which can be preserved when using large training data sets only. In addition, MMM shows a considerable error in the 1000-day group: this may indicate that the model based on the 30- and 365-day wetness indexes needs a larger data set to be calibrated.

The time dependence structure of the total signal is simulated quite accurately by DS as confirmed by a small ACF error on all relevant lags. Conversely, MMM is structured to preserve accurately the lag-1 autocorrelation. To avoid the underestimation of persistence using MMM it is therefore necessary to include the appropriate information in the time dependence structure of the model. This is not needed using DS since it can automatically reproduce complex time dependencies by generating multiscale patterns similar to the ones found in the training data set. Large error shown by both techniques in the ACF of the separate regimes is due to their inability to capture the non-stationarity of the two-regime alternation in absence of prior information about it.

Finally, the error on the dry/wet spell length distributions and on the minimum moving average confirms the same tendency: we observe a better performance of DS when sufficient training data are available, while MMM is more reliable in case of scarce data availability.

3.4 Discussion and conclusions

In this chapter we show a comparison between two recent techniques for daily rainfall simulation: the Markov-chain based modified Markov model (MMM) and the direct sampling technique (DS) belonging to the multiple-point statistics family. The two algorithms use the same type of information under the form of variables computed from the rainfall amount, namely: the rainfall state (dry or wet) and the wetness indexes, i.e. the number of wet days number in a past period, informing about low-frequency fluctuations. MMM is a semi-parametric model where the rainfall state generation is conditioned on a fixed order-1 time dependence and low-frequency fluctuations. The rainfall amount is generated using an order-1 conditional kernel density estimation. This way, non-stationarity is introduced in the parameters of both the occurrence and amount models allowing the preservation of the essential small- and large-scale characteristics of rainfall. Conversely, DS is a fully non-parametric resampling technique based on a pattern-similarity rule. Using a random simulation path and a variable conditioning pattern, DS simulates the same type of patterns found in the training data set at multiple scales. Consequently, high-order statistics contained in the training data are indirectly preserved in the simulations without the need of a complex prior structure.

The tests performed are based on the simulation of a synthetic daily rainfall time-series exhibiting a pseudo-chaotic alternation of two regimes. Each regime shows a specific time dependence structure and an extremely variable regime spell duration. The asymptotic behavior of the two techniques is tested by generating a 1 million-day long reference and considering a variable amount of training data.

The results confirm the efficiency of both techniques in reproducing the rainfall amount distribution at the daily and higher scales when the full 1-million-day reference is available as a training data set. Reducing the available data amount, MMM can extrapolate extreme rainfall amount values by using a conditional kernel smoothing technique and approach the reference distribution at the daily scale even for short data sets. DS, being based on resampling, remains limited to the range of data found in the training data set and thus underestimates the extremes. Therefore DS in its present form is not a suitable tool to explore the asymptotic behavior of rainfall at the daily scale. For large temporal scales, both techniques can preserve an unbiased marginal rainfall amount distribution of wet days even using a limited amount of daily data, avoiding the problem of over-dispersion that commonly affects daily rainfall simulation techniques. The highly irregular two-regime alternation of the reference signal is preserved fairly well by both techniques using a large training data set. Moreover, the specific high-order temporal correlation contained in the whole signal is automatically captured and reproduced by DS, while MMM underestimates the persistence

since it is limited to the lag-1 time conditioning contained in its prior structure. These results confirm that, using a Markov-chain based approach, a preliminary analysis is necessary to include the salient high-order time dependence features in the prior structure of the model. The autocorrelation function of the two separate regimes, showing a different time dependence signature, is not correctly preserved by either the approaches meaning that information about this kind of non-stationarity should be incorporated in the prior structure of both models. In case of MMM, this is possible by implementing a regime switch in the Markov-chain conditioning structure, while, using DS, a regime indicator can be calculated on the training data set and simulated together with the rainfall signal. Finally, the dry/wet spell distribution and the minimum moving average of the rainfall amount confirms the higher accuracy of DS in simulating long wet periods and the multiple-scale features when a sufficient training data set is available, while MMM is more reliable in case of scarce data availability, where DS underestimates the length of both the dry and wet extreme spells.

The considered techniques constitute in sum two valid approaches for daily rainfall simulation, preserving the distribution up to the decennial scale. With both algorithms, the uncertainty shown by the 05-95 percentile boundaries of the realizations suggests that, for real applications, the training data set should cover multiple decades to correctly represent the variability of the relevant statistical features up to the annual and decennial scales. The semi-parametric approach conditioned by low-frequency indicators implemented in the MMM technique simulates more efficiently the rainfall amounts in case of limited training data, where an effective extrapolation at the daily scale is needed to correctly explore the tail of the distribution. Nevertheless, in case of a complex time-dependence at multiple scales, this technique requires an adequate preliminary data analysis for the parametrization of its prior structure. Conversely, DS is a more adaptive tool for generating realistic replicates of the observed data since it can capture complex multiple scale features without the need of a high-order prior structure. It is however limited to the rainfall amount values observed in the training data set, requiring therefore a very large amount of data to avoid under-representation of the extremes.

To take the best of both approaches, future research could see the development of a semi-parametric or kernel based amount model inside the DS framework to perturb the sampled historical values. This idea has been already proposed for k-nearest neighbor resampling techniques [110, 155] and applied to some stochastic hydrological models of the same family: inspired by traditional autoregressive models, they consider the non-linear regression $m(H_i)$ to describe the relationship between the training data Z_i and a predictor variable vector H_i. The simulated value $Z_t = m(H_t) + e_t$ is the sum of the deterministic conditional mean $m(H_t)$ and an innovation term e_t, generated by sampling from the local residuals of $m(H_t)$ [151] or calibrating a random noise on them [180, 174]. These works show that the introduction of an innovation term is a promising path to increase the prediction skills of resampling techniques. Moreover, the parametric framework of these algorithms present a limited degree of conditioning, which, on the contrary, is extended much further in the non-parametric approach of DS. For these reasons, the development of a perturbation stage of the sampled values in the direct sampling framework may lead to an improved model. A possible simulation workflow may be the following:

1. apply a variable transformation to the training data set to obtain a uniformly distributed sample in the range [0,1] – one way to achieve this task is to compute the cumulative probability distribution function using a Generalized Extreme Value (GEV) distribution model fitted on the training data set;

2. choose a random non-informed time step in the simulation grid;

3. begin the training data set random scan;

4. sample a value according to the canonical DS algorithm;
5. define a probability distribution function with mean centered on the sampled value and support limited, but not necessarily extended, to the interval [0,1];
6. draw a random value from the defined distribution to generated a perturbation of the sampled data;
7. assign the value to the simulated time step in the simulation grid;
8. repeat from step 2 to 7 until all the simulation grid is informed;
9. apply to the simulated values a back transformation using the same distribution model used in step 1.

This proposed algorithm should allow the simulation of new data not contained in the TI and the generation of long recurrence time extremes, according to the size of the simulated data. These would be no more limited by the TI size. Nevertheless, this prototype algorithm leaves many open questions: for example, which distribution should be used to draw the random perturbation of step 5? What should be the magnitude of the random perturbation? Future research will be devoted to this topic.

Chapter 4

Missing data simulation inside flow rate time-series

Abstract

In this chapter, the Direct Sampling technique is proposed as a non-parametric method to simulate missing data within hydrological flow rate time-series. The algorithm makes use of the patterns contained inside the training data set (the available data) to reproduce the complexity of the signal in the missing data simulation. The proposed setup is tested in the reconstruction of a flow rate time-series, considering several missing data scenarios and using different auxiliary information. The results show that, most of the times, the statistical content of the missing data is entirely recovered and the predictive power of the technique is much increased when a correlated flow rate time-series measured from a nearby location is used as conditioning variable. Even when the auxiliary time-series is incomplete, the algorithm preserves a similar accuracy in the estimation.

4.1 Introduction

The reconstruction of missing data portions inside time-series is a critical topic in applied hydrology since a large part of the numerical simulation techniques used to model the hydrological processes need as input continuous data records. Sometimes technical failures of measure instruments produce missing or unreliable data for long time periods, for which the uncertainty about the observed phenomena is high. For this reason, a technique capable of generating realistic simulations of the missing data, reflecting the complex structures of the signal and possibly making use of auxiliary information, is needed.

Many different approaches have been proposed for time-series gap filling in the earth sciences: for example techniques based on mean diurnal variation or regression [48, 132], autoregression [21], singular spectrum analysis [168, 100], self-organizing maps [210, 111], look-up tables [15], rough sets [45], and artificial neural networks, widely used in the recent years [6, 42, 137, 14, 139, 45]. In this chapter, we propose a non-parametric method to simulate missing data inside flow rate time-series based on the Direct Sampling (DS) technique [122] belonging to multiple-point statistics (MPS). Already tested on gap filling in multivariate data sets representing natural heterogeneities [120] and on rainfall time-series simulation [143], DS can simulate the outcome of a complex natural process by reproducing similar patterns to the ones found in the available data, without imposing a specific statistical model. In particular, missing data are simulated by sampling with replacement the available data set where a sufficiently similar pattern is found. High-order statistical relations in the variable of interest are preserved by respecting the similarities in the neighborhood at multiple scales. The approach is almost entirely data-driven and fairly simple, but its efficiency largely depends on finding the good ensemble of auxiliary variables, suitable to the current application. We present a multivariate standard setup for missing data simulation inside hydrological flow rate time-series using a correlated time-series as auxiliary variable. The setup is tested on the gap filling of a high-resolution karst flow rate time-series using different auxiliary variables. To make the test systematic and relevant for real application, a gap size varying from some hours to 20 days and a total missing data percentage up to 30% are considered. The general methodology, the setup as well as the data set used are illustrated in section 4.2, the results are presented in section 4.3 and section 4.4 is dedicated to the conclusions.

4.2 Methodology

4.2.1 The data set

The data set used to test the proposed technique is the 1990-2013 flow rate record from two karst springs of the Jura mountains (Switzerland), provided by the Swiss Federal Office for the Environment (FOEN). Three high resolution (10-min) time-series are used: the Areuse creek measured at St.Sulpice station (Ar) is used as target variable, while the same water flow measured at Boudry station (Ar2) and the Seyon creek measured at Valangin station (Se), in a nearby basin, are used as auxiliary variables. Ar and Ar2 are highly correlated (Pearson's correlation coefficient PCC=0.96), whereas Ar and Se show a medium to weak correlation (PCC=0.72). The considered time-series do not present in origin any missing data, but Ar1 and Ar2 show isolated sharp fluctuations around the local trend due to instrumental errors. To remove this kind of artifact, the following preprocessing treatment described in appendix C, section C.3 is applied.

4.2.2 The Direct Sampling technique

Multiple-point statistics (MPS) techniques are based on the concept of training data set (TI): a representative sample of the target variable or conceptual model which is used to estimate the probability of occurrence of each event inside the simulation. MPS methods [61, 196, 4] generally consider a catalog of neighbor patterns found in the TI to impose high-order conditioning in the simulation and thus reproduce similar structures to the ones found in the TI. This requires the estimation of the conditional probability density function for each pattern and limits the application of the method to categorical variables. The Direct Sampling technique [122] avoids this preliminary step by sampling the TI where a sufficiently similar neighborhood occurs in the TI. This principle also extends the application to continuous variables and multivariate data sets. In case of missing data simulation, only the non informed time steps are simulated and the rest of the data set is used as conditioning data (CD). We refer to chapter 2 and [195] for a detailed description of the actual algorithm implementation. In general, the required input for a DS simulation are: a data set used as TI and the simulation grid (SG): a time vector hosting the CD together with the simulated data. In case of missing data simulation, the TI and the CD may be the same data set: this means that the gaps are filled using the data already present in the SG. The main DS parameters, related to each simulated variable are: i) the search neighborhood radius R, that defines the time interval $t_i \pm R$, used to retrieve the conditioning pattern for the simulation at time step t_i; ii) the maximum number of neighbors N used to form the conditioning pattern; and iii) the distance threshold T, a scalar value used to accept or reject the pattern scanned inside the TI. One last parameter defined once for all variables is the maximum TI fraction (F) scanned at each algorithm iteration. The value F=0.5 is used in all the tests presented in this chapter. This is a standard value used in previous applications that, in case of a representative training data set, allows scanning a sufficient training data amount at each iteration. The scanning of the total TI is avoided since it can lead, in some cases, to the oversampling of the same TI regions and the reproduction of entire data set portions.

As explained in section 2.2.3, the main difference of this approach with respect to the existent resampling techniques for time-series simulation [e.g. 155, 24, 217, 33] is the combined use of: i) a random simulation path and ii) a variable conditioning scheme, using the N informed neighbors closest to the simulated time step. These two elements allow considering large-scale patterns at the beginning of the simulation and denser small-scale patterns toward the end of the simulation. For instance, by setting R=100 and N=10, the conditioning pattern for the simulation of the first random time steps will be formed by 10 or less sparse

neighbors in the time $t \pm R$, while, for the last simulated time steps, it will be composed by 10 time steps much closer to the simulated one, since at this stage, the SG is more densely informed. This imposes a variable time-dependence, which allows preserving the statistical structure at multiple scales without the formulation of a high-dimensional prior statistical model. Moreover, multiple variables can be simulated together by applying the same procedure and using a multivariate TI and SG: at each time step t, the variables are simulated beside each other by considering a multivariate data event. A suitable candidate pattern presenting a distance below the threshold for all variables is found by scanning the TI. At this point, the whole vector of variables at the center of the multivariate pattern is assigned to t (see the VVT mode in appendix B), except for the variables that are already informed (conditioning data). This method allows preserving not only the coherence in the multivariate patterns, but also the statistical correlation of all variables. For example, one can use a sufficiently informed variable to guide the simulation of large missing data portions inside the target variable (see section 4.2.3). It is worth remembering that, since DS samples the data found in the TI, the use of a representative TI is crucial to obtain a reliable simulation (see section 4.3).

4.2.3 The DS setup for flow rate time-series

In this section, we present a standard multivariate DS setup for the simulation of missing flow rate time-series data, composed by a series of variables and the main DS parameter values (table 4.1). A flow rate time-series ($Z(t)$), is simulated in its missing data parts together with a group of auxiliary variables, allowing the preservation of the temporal structure contained in the original data set. The auxiliary variables include: two deterministic periodic functions describing the annual seasonality ($A_1(t)$ and $A_2(t)$), a correlated flow rate time-series ($Q(t)$) measured from a nearby location (if available), informing about the occurrence of floods, and an indicator variable that describes the hydrographic structure as an alternation of rising and recessing hydrograph portions ($H(t)$). This multivariate data set is defined at the same temporal resolution of $Z(t)$, that may vary according to the considered data set. A more detailed description of these variables and the respective DS parameterization are given in the following. For simplicity, we omit from the notation the temporal reference where possible.

- The flow rate time-series Z is the target variable, presenting missing data portions that are generated in the simulation and informed time-series portions that are used as conditioning data. A variable high-order conditioning is applied to Z by extending the search neighborhood radius (R) to 10'000 10-min time steps and considering a maximum (N) of 15 neighbors. The distance threshold value T=0.002 allows the 0.2% of total variation on the conditioning pattern. The chosen values for R and N for all variables can be related to the correlation length of their temporal structure. The user is recommended to change the value of these parameters according to the time-series resolution. Conversely, the proposed T values, lacking a physical meaning related to the variable, are manually set up by trying a limited set of values (0.002, 0.01, 0.05, 0.07) and using the indicators presented in section 4.2.5 as optimization criterion. A sensitivity analysis of the DS parameters [124] showed that, for the majority of the application cases considered until now, the optimal T values lie in the range [0.001,0.1] with lower values suitable for highly autocorrelated, smooth signals and higher values for low-correlated, more noisy signals. In case of lower resolution or more noisy data sets, a higher T value may be more appropriate (see e.g. the setup for daily rainfall in chapter 2).
- Two out-of-phase periodic triangular functions (A1 and A2) with period 365.25 days, indicate the position of each datum inside the annual cycle. A1 and A2 are given as CD

to help the simulation respecting the annual seasonality. Since high-order conditioning is not necessary for this purpose, R and N are set to 1. The distance threshold T set to 0.07 allows sampling from the same period of the year with a maximum 7% of the total variation of the variables. As already observed for rainfall and climate variables, T varying between 0.05 and 0.07 allows imposing the annual seasonality without over-conditioning the simulation.

- A flow rate time-series (Q) from a nearby located station, is given as CD, but it is not necessarily fully informed. Any missing data inside Q will be co-simulated with the target variable. If Q is correlated to Z, its conditioning helps restricting the uncertainty around the missing data, e.g. indicating a flood occurrence if a peak is present in Q. The same DS parametrization as Z is applied to Q.

- An indicator variable called recession indicator (H), takes values H=1 to indicate a recessing hydrograph limb and H=0 for a rising hydrograph limb observed in Q. H is necessary to simulate a more realistic flood pattern in the target variable. Since the flow rate time-series is a complex signal, showing abrupt fluctuations of different magnitude, computing the sign of the derivative in Q is not sufficient to identify the effective succession of rising and recessing limbs, corresponding to the main flood pattern. For this reason, H is computed with a more complex procedure described in the following. H is a deterministic function of time t and the user defined parameters (w, v). First, the local extremes (minimum and maximum) of Q inside a moving temporal window $[t \pm w]$ are identified. Moreover, each extreme is considered only if: i) it shows a variation greater than v with respect to the previously considered extreme and ii) the next extreme found is not of the same type (minimum or maximum). Finally, H is obtained by applying a logical test on the selected local extremes: a local minimum activates a rising limb (H=0) until a local maximum occurs activating a recessing limb (H=1), assuring a continuous alternation of the two categories. If Q is incomplete, H also presents missing data at the corresponding time steps. The procedure for the computation of H is implemented with the following algorithm:

  1. give the input arguments:
     $Q(t_i)$ with $i = 0, \ldots, n$;
     $v \in R | v \geq 0$
     $w \in N | 0 < w < n/2$
  2. define the initial values:
     $\{H(t_0), \ldots, H(t_n)\} = 0$
     $t_p = 0$
     $L_{max} = \max(Q(t_i))$
     $L_{min} = \min(Q(t_i))$
  3. starting from $t_i = t_w$:
     if $Q(t_i) = \varnothing$, where $\varnothing$ denotes a missing value, **then**
         $H(t_i) = H(t_{i-1})$
     else if $Q(t_i) = \min\{Q(t_{i-w}), \ldots, Q(t_{i+w})\} \wedge L_{max} - Q(t_i) > v$ **then**
         $H(t_i) = 0$
         if $H(t_p) = H(t_i)$ **then**
             $\{H(t_p), \ldots, H(t_{i-1})\} = 1$
             $t_p = t_i$

$L_{min} = Q(t_i)$

else if $Q(t_i) = \max\{Q(t_{i-w}), \ldots, Q(t_{i+w})\} \wedge Q(t_i) - L_{min} > v$ **then**

$H(t_i) = 1$

if $H(t_p) = H(t_i)$ **then**

$\{H(t_p), \ldots, H(t_{i-1})\} = 0$

$t_p = t_i$

$L_{max} = Q(t_i)$

else

$H(t_i) = H(t_{i-1})$

4. repeat for the successive t_i up to $i = n - w$
5. put $H(t_i) = \varnothing$ for $i|Q(t_i) = \varnothing$

This technique allows, with a reasonable approximation, the automatic detection of the hydrographic structure and can be applied to different flow rate time-series by setting the parameters w and v. The values w=50 and v=2 mm for Q=Ar2 and w=50 v=0.3 for Q=Se have been chosen since they allowed an adequate visual representation of the hydrographic structure. The DS parameter values R=10'000 time steps, N=20 and T=0.05 are applied to H. This N value, set up by trial and error, allows in this case an adequate representation of the hydrographic structure.

The proposed DS setup makes use of Q as source of additional information. The rest of the variables are in fact derived from it (H) or known a priori ($A1$ and $A2$). If Q is not available, the simulation is still possible with H computed on the informed part of Z. Since the most adequate parameter values for DS and H may vary in function of the flow rate characteristics and the time-series sample rate, the user should not consider the suggested values as fixed but rather as a starting point for optimization to a specific application.

4.2.4 Experiment design

The proposed technique is tested by artificially creating gaps in the multivariate data set and simulating the corresponding missing data. Ar is considered as target in 5 simulation tests presenting different time-series as Q : 1) In test Ar no Q variable is used and H is computed on the informed part of Ar. 2) In test Ar-Ar2, Ar2 is used as complete Q variable, highly correlated to Ar. 3) An incomplete version of Ar2 (Ar2*) is used in test Ar-Ar2*. 4) In test Ar-Se, Se is used as Q to represent a case where the auxiliary time-series is poorly correlated with Ar. 5) Finally, in test Ar-Se*, Q=Se*, representing a case where Q is incomplete and poorly correlated with Z. To test the sensitivity of the method performance to different gap size and missing data quantity, random groups of non touching and equally sized gaps are generated inside Z according to different missing data scenarios: as shown in table 4.2, three different classes of missing fraction up to 30% and three different classes of gap size up to 3000 time steps per gap are considered for a total of 9 fraction-size combinations. Since the time step is 10 minutes, the generated gap size vary between 8 hours and 20 days. This time range can represent the data loss due to small mechanical failures or large breakdowns comprehending entire wet periods (see figure 4.1). For each fraction-size combination, 10 gap scenarios and 10 DS realizations per scenario are generated for a total of 900 runs per test. For test Ar-Ar2* and Ar-Se*, gap scenarios for Q are generated independently from those for Z and present always a 20% missing fraction with 300-time-step gaps.

Table 4.1: DS setup proposed for flow rate time-series simulation. The parameters are: search window radius R, maximum number of conditioning neighbor data N and distance threshold T. The variables are: 1-2) the annual seasonality triangular functions ($A1$ and $A2$), 3) the recession indicator (H) and 4) a flow rate time-series (Q) from a nearby located station and 5) The flow rate time-series (Z), being the target variable of the simulation. The variables Q and Z are expressed in m^3/s. A sample of the multivariate TI is shown on the right.

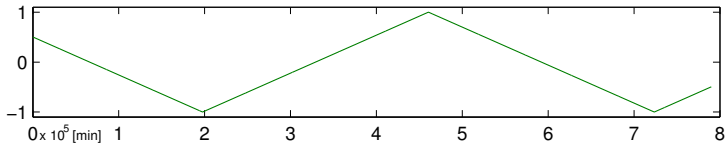
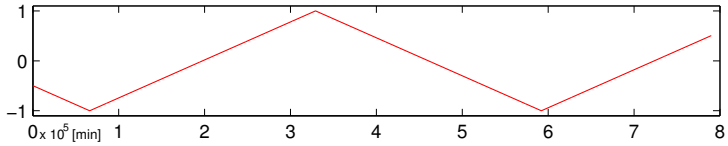
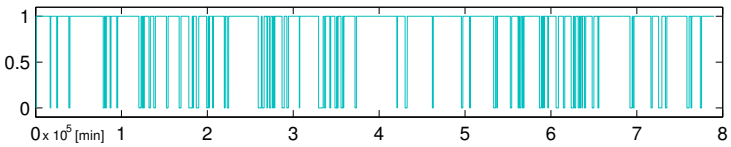
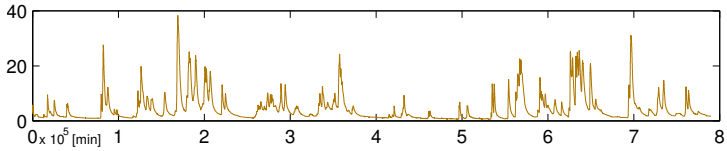
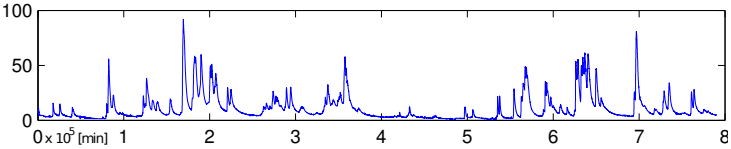
Variable	R	N	T	TI example
1) $A1$	1	1	0.07	
2) $A2$	1	1	0.07	
3) H	10'000	20	0.05	
4) Q	10'000	15	0.002	
5) Z	10'000	15	0.002	

Table 4.2: Simulation schedule for each test: 9 missing fraction-gap size combinations, 10 gap scenarios per combination, 10 realizations per gap scenario, for a total of 900 realizations.

missing fraction → gap size (num. time steps) ↓	5%	10%	30%
1) 50 (∼ 8 hours)	10 scenarios ×10 real.	10 × 10	10 × 10
2) 300 (∼ 2 days)	10 × 10	10 × 10	10 × 10
3) 3000 (∼ 20 days)	10 × 10	10 × 10	10 × 10

4.2.5 Evaluation

The performance of the proposed simulation technique is analyzed separately for each test and fraction-size combination by considering a visual comparison between the generated and reference time-series as well as a group of statistical indicators. To test the efficiency in simulating the statistical content of the missing data, the probability distribution of the simulated and reference missing time-series portions are compared using quantile-quantile (qq-) plots. To show the behavior of the simulation ensemble, the median, 5th and 95th percentile of the realizations are plotted for each quantile. The predictive power of the technique is tested using some classical goodness-of-fit measures: the Pearson's correlation coefficient (PCC) between the simulated and reference missing data, the root mean square error (RMSE) and the Nash-Sutcliffe model efficiency coefficient (NSE).

4.3 Results and discussion

4.3.1 Visual comparison

Figure 4.1 shows a time-series portion of about 100 days presenting two simulated gaps and the corresponding missing data together with the auxiliary variables used. Among the considered gap scenarios, one presenting 30% missing fraction and 3000-time-step gaps has been chosen for visual comparison, since it better illustrates how the signal is reconstructed by the algorithm. In general, the algorithm generates hydrographic structures similar to the one found in the reference by sampling each datum from the TI where a similar neighborhood is found: the simulated flood events are in the same magnitude range and the asymmetric shape of the hydrograph looks realistic although containing a modest noise. In test Ar (figure 4.1, a), where no Q variable is used and H is computed on the informed part of Z, the flood occurrence in the simulation does not match the reference flood structure but present flow rate values in the same range. This is expected since, within the missing data portion, the simulation is only conditioned by A_1 and A_2 informing about the annual seasonality. Therefore, the algorithm explores a larger variability, generating different types of flood structures. Conversely, when Q is present and highly correlated to Z as in test Ar-Ar2 (figure 4.1, b), the uncertainty is restricted around the reference flood structure. Local extremes are estimated quite accurately as shown by the simulation mean (red line). When Q is poorly correlated to Z (test Ar-Se, figure 4.1, d), the simulation mean follows the main reference shape, but the simulation ensemble shows larger uncertainty. This suggests that Q and its derived variable (H) are highly informative about the hydrograph structure and play an important role in conditioning the simulation. When this auxiliary information is incomplete as in test Ar-Ar2* (figure 4.1, c), the algorithm shows a similar performance: even if some portions of the auxiliary variables (Q and H) are missing in correspondence to peaks, DS can efficiently simulate the local extremes. Normally, this result cannot be achieved with a parametric technique based on fixed time dependence, since it requires the predictor variables to be fully informed. Conversely, DS uses a variable conditioning pattern adapted to the data available in a temporal range defined by the parameter R (see section 4.2.3). The missing data are simulated in a random order which allows first defining the large-scale structure of the missing portion. Then the data sequence is completed by considering the already simulated values as conditioning data. This leads to a realistic reconstruction even if the auxiliary time-series used is incomplete.

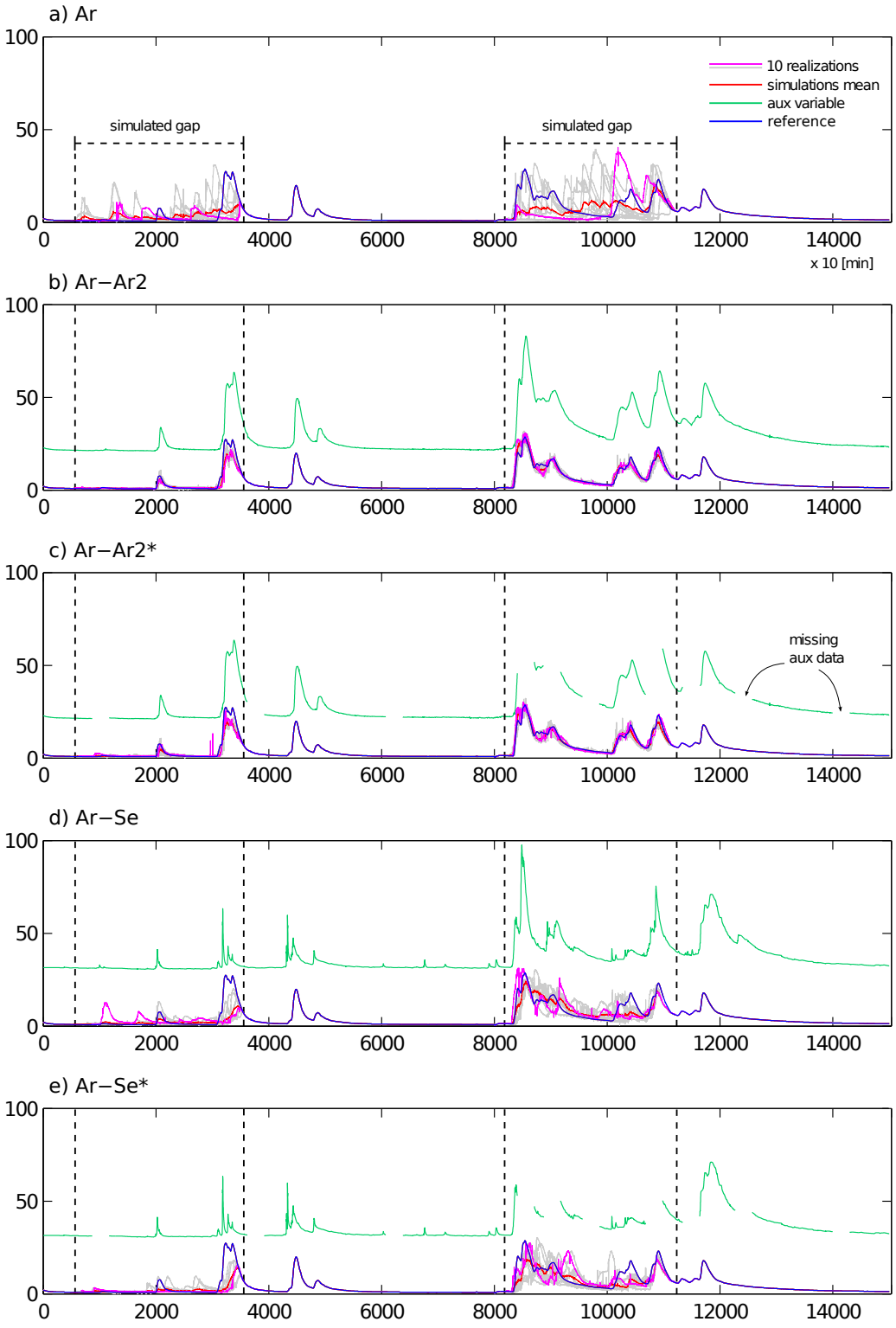


Figure 4.1: A flow rate time-series portion (m^3/s, 10-min average, about 100 days) showing two simulated gaps, the reference and the auxiliary variable used for all the tests. The shown examples belong to one scenario presenting 30% missing fraction and 3000-time-step long gaps. A randomly chosen realization (pink color) is put in evidence over the simulation ensemble (gray color). The auxiliary variable (green line) is shifted with respect to the vertical scale for illustration purpose.

4.3.2 Statistical comparison

In figure 4.2, the probability distributions of the missing reference and simulated data are compared by mean of qq-plots. This allows testing the efficiency of the technique in recovering the statistical content lost with the missing data. The results may vary significantly depending on the missing time-series portions. For this reason, 10 scenarios for each fraction-gap combination and test have been considered (see section 4.2.4). For all tests, the median of the simulations (solid lines) mainly lies on the bisector of the graph, indicating that the simulation preserves the reference distribution on average. Nevertheless, a tendency to underrepresent data between 40 and 60 m^3/s is observed when the simulated data quantity is limited (5-10% of the data set, figure 4.2 left and center column). This suggests that the algorithm is not an appropriate tool to represent the extremal behavior of the target variable at the temporal scale of the data since it may underrepresent the extreme values in some cases. Moreover, being based on resampling, it is not capable of generating values not observed in the training data set. Nevertheless, the underrepresented data correspond to very rare events: for example, in this case, they constitutes only 0.0014% of the observations and occur sparsely in the time-series with a negligible impact on the hydrological regime. Therefore, their underrepresentation is not a main issue unless the user is studying the extremal behavior of the process at the 10-minute scale. Increasing the missing data quantity, this bias tends to disappear (30% of total data set, figure 4.2 right column). In this case, the algorithm scans more deeply the training data set to generate a larger number of data patterns, with a higher chance to sample extreme values. The 5th and 95th percentile boundaries (dashed lines) of the realizations indicate the uncertainty on the recovered statistical distribution. This is larger when a little percentage of the data is missing since it is dependent on the statistical content of the missing data, different for each scenario (figure 4.2 left column). Conversely, with a larger missing data amount (figure 4.2 right column), the missing statistical content varies much less depending on the scenario and also the uncertainty of its estimation is lower. In summary, these results show that the algorithm can efficiently recover the main statistical content even when no auxiliary information is used (test Ar).

4.3.3 Predictive power

In this section, the predictive performance of the technique is analyzed by mean of some goodness-of-fit indicators computed for each test and gap fraction-size combination. The box-plot of the root mean squared error (RMSE) between the simulation and the reference missing data is shown in figure 4.3: the results are grouped hierarchically by test, missing percentage and gap size (indicated by different colors).

The most important influence on the prediction is played by the gap size: for all simulation groups, RMSE is lower than 1 m^3/s in case of small-sized gaps (50 time steps, red color) and does not present any substantial change in function of the missing percentage and test type. This can be explained by the fact that the variable is highly autocorrelated and does not show big variations in 50 missing time steps (see for example figure 4.1). The variability of the missing data in gaps of this size is very limited and can be efficiently performed without using any auxiliary information. A comparable result may be achieved with a reliable method of interpolation. When the gap size is larger, the variability of the possible data patterns is higher and the simulation is much more dependent on the setup used. For test Ar, where Q is not used, RMSE increases rapidly with the gap size: between 2-4 m^3/s for 300 time-step gaps (blue) and between 4-10 m^3/s for 3000 time-step gaps (black), with the median of the realizations between 7 and 8 m^3/s. For all tests, the variability of the performance is dependent on the gap size and missing percentage: as shown by the interquartile

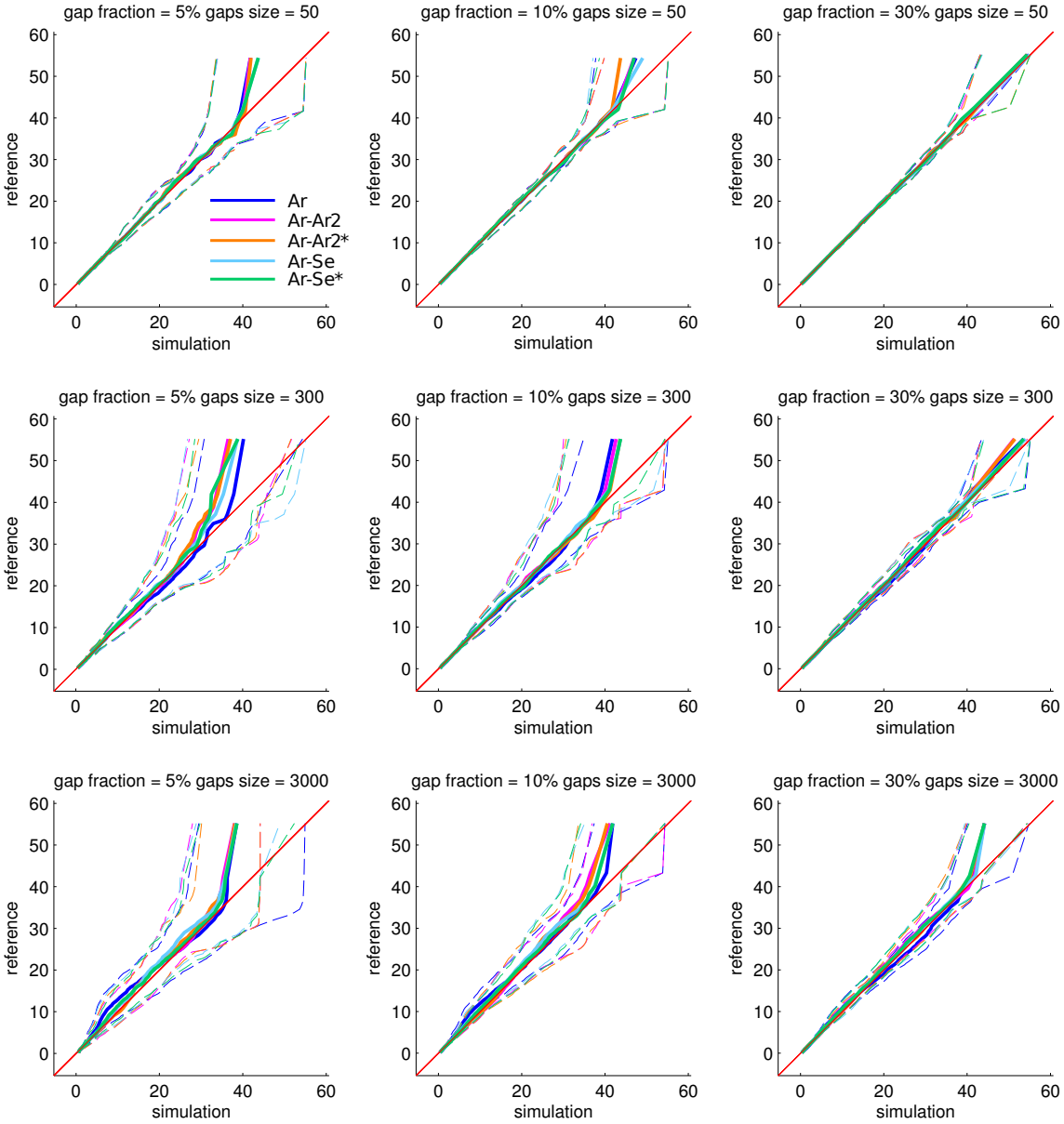


Figure 4.2: QQ-plot of the reference distribution against the simulation ensemble for all tests and gap scenario classes. Each graph contains the simulation of a specific fraction-size combination (10 realizations by 10 scenarios): the missing fraction augment from left to right and the gap size from top to bottom. Each test is indicated with a different color, solid lines indicate the realization median and dashed lines the 5-95th percentile boundary.

range (thick part of the box-plots), the error variability is maximized for largest gap sizes and smallest missing percentage. In this case, the performance is much dependent on the content of the random missing portion, while, augmenting the missing percentage or reducing the gap size, this effect is statistically compensated and the RMSE of the simulation ensemble converges towards its median. The error on large gaps is limited below 2 m^3/s if a highly correlated Q variable is used (Ar-Ar2). This confirms the efficiency of a highly correlated auxiliary variable in reducing the uncertainty of the simulation. By imparting the right flood sequence, the predictive power of the technique is augmented. Conversely, only a modest augmentation of the RMSE is observed when Q is incomplete (Ar-Ar2*). Test Ar-Se and Ar-Se*, using a lowly correlated Q, present a performance in-between tests Ar and Ar-Ar2: the RMSE is between 1 and 3 m^3/s for 300 time-step gaps and between 2 and 6.5 m^3/s with median around 4.5 m^3/s for 3000 time-step gaps.

The Pearson's correlation coefficient (PCC, figure 4.4) and the Nash-Sutcliffe model efficiency coefficient (NSE, figure 4.5) confirms the same results shown by the RMSE. Remind that PCC can be interpreted as the fraction of the reference variability predicted by the simulation: for example PCC=0.7 the model can predict the 70% of the variability. NSE is of less immediate interpretation: NSE=0 indicates that the simulation has the same predicting power as the estimated mean, NSE=0.7 indicates that the RMSE is equal to the 30% of the observed variance [115] and NSE=1 corresponds to a perfect prediction. For example, NSE values between 0.75 and 1 are considered as indicator of a very good prediction performance for monthly time-series by Moriasi et al. [134]. Referring to figures 4.4 and 4.5, we can consider the predictive performance of the simulation very good (PCC $>$ 0.9 and NSE $>$ 0.8) in case of small gaps (red color) for all considered tests, in case of medium gaps (blue) when Q is used and in case of large gaps when Q is highly correlated to Z (Ar-Ar2 and Ar-Ar2*). In absence of Q (test Ar), the prediction remains reliable for gaps up to the medium size (red and blue colors). The prediction is not efficient in case of large gaps and absent or lowly correlated Q variable. In this cases, the flood sequence generated by the algorithm may still be a realistic estimation, but it does not correspond to the one present in the reference, since the algorithm explores a larger uncertainty space.

4.4 Conclusions

The aim of this chapter was to propose and test a novel stochastic methodology for missing data simulation inside hydrological flow rate time-series based on the Direct Sampling technique. Its rationale is fairly simple: without imposing a statistical model for the process of interest, the missing data are simulated by resampling data patterns of the variable of interest together with a group of auxiliary variables. By scanning the available data, a similar pattern is found and the datum at its center is assigned at the simulated time step. The process is repeated until all the data set is complete. Since multiple neighbors and different pattern size are considered for conditioning, realistic structures at multiple scales can be generated.

A standard setup for flow rate time-series is proposed, including the variable of interest (Z) and a series of auxiliary variables: a couple of periodic theoretical functions describing the annual seasonality, a predictor variable (Q), which is a correlated flow rate time-series, and a categorical variable computed on Q describing the hydrographic structure as a succession of rising and recessing flood limbs. The setup can be adapted to any type of flow rate time-series, with or without the use of auxiliary variables, but it may require an adjustment of the parameters.

The model is tested on the gap filling of a high-resolution (10-min) karst flow rate time-series from the Areuse St.Sulpice station (Jura mountains, Switzerland) by using as Q dif-

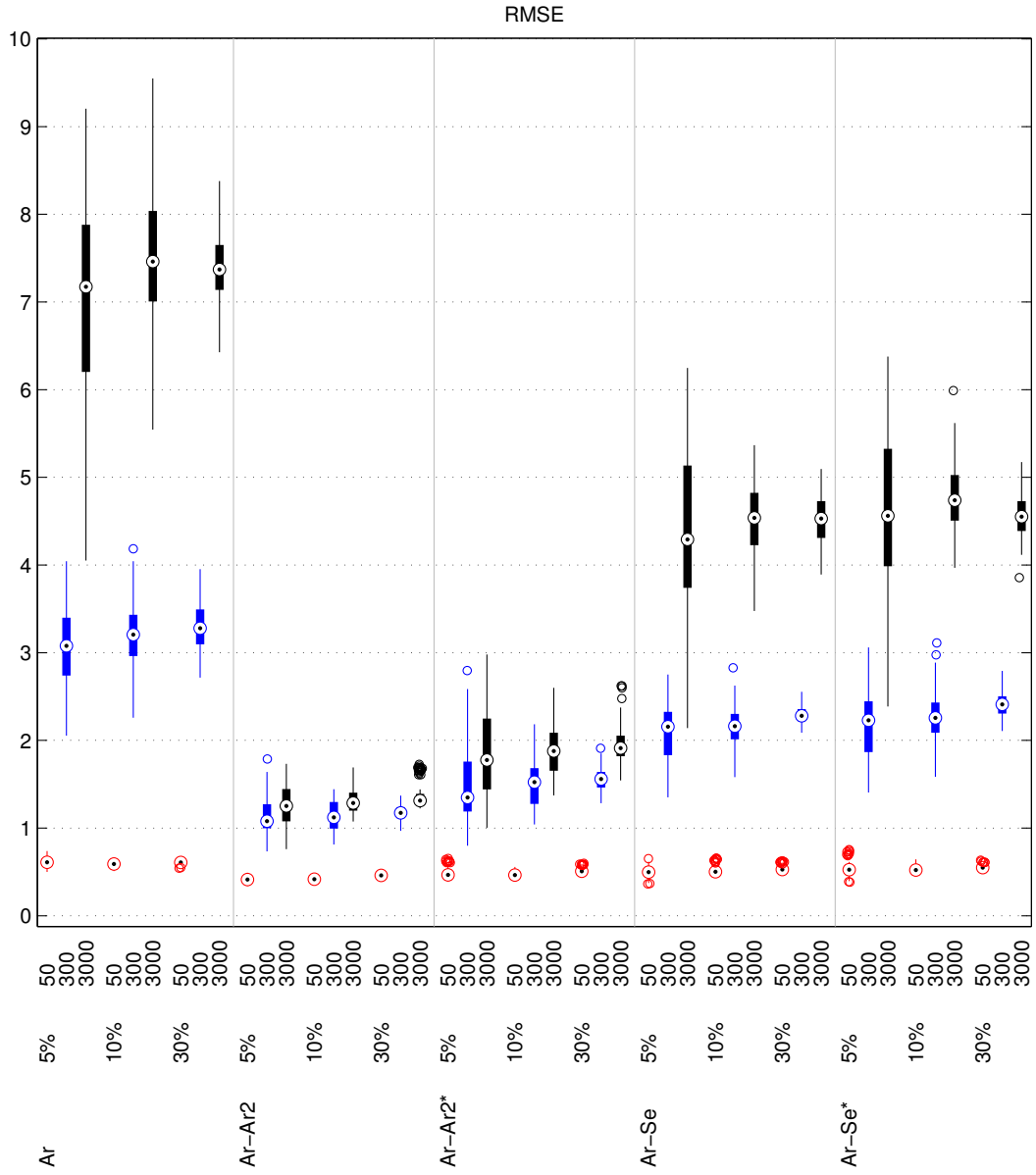


Figure 4.3: Boxplot of the root mean square error (RMSE) computed on the simulated missing flow rate data [m^3/s] for all tests and gap scenario classes. Colors indicate the gap size class: 50 (red), 300 (blue) and 3000 (black) time steps.

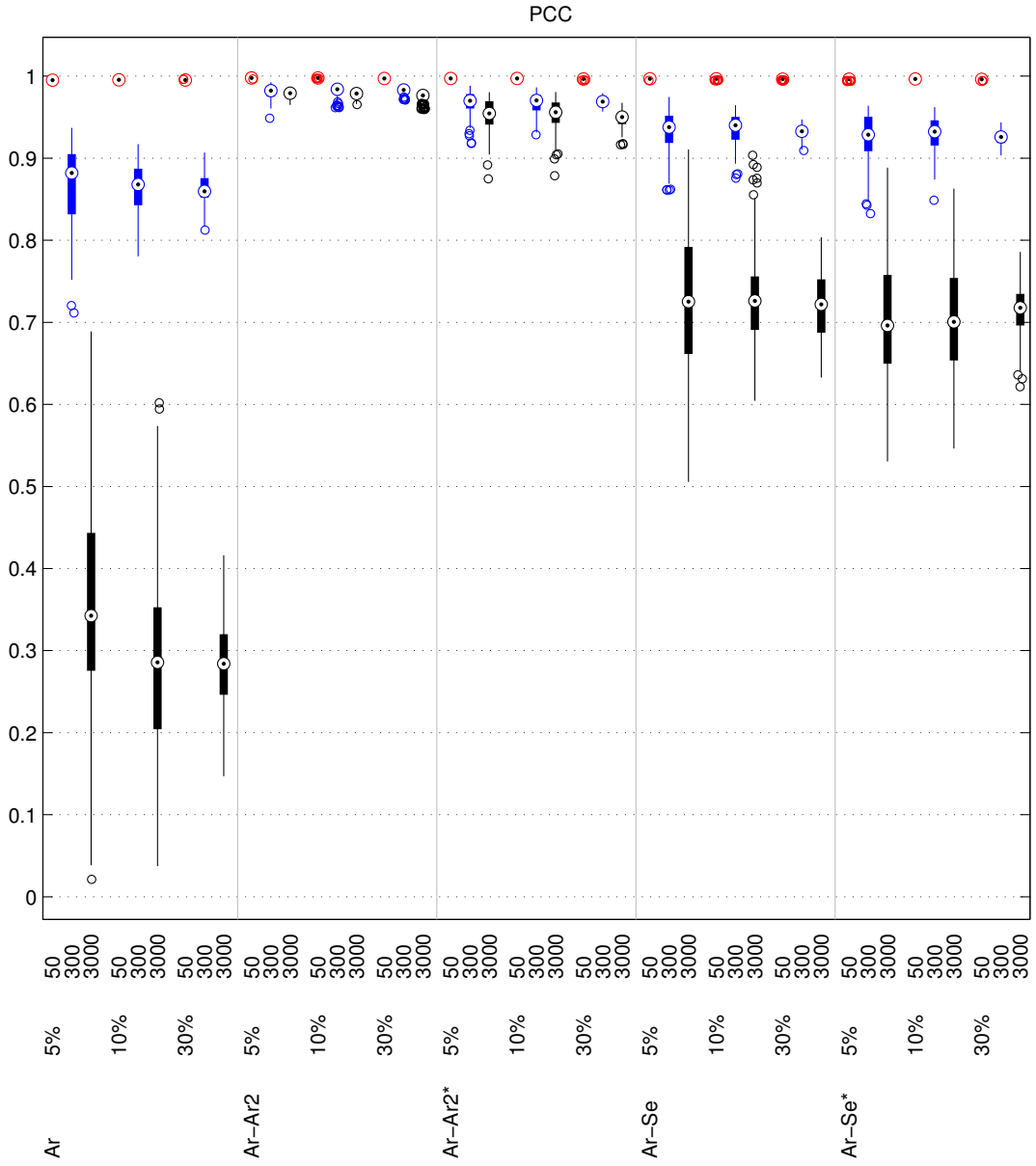


Figure 4.4: Boxplot of the Pearson's correlation coefficient (PCC) computed on the simulated missing data portions for all tests and gap scenario classes. Colors indicate the gap size class: 50 (red), 300 (blue) and 3000 (black) time steps.

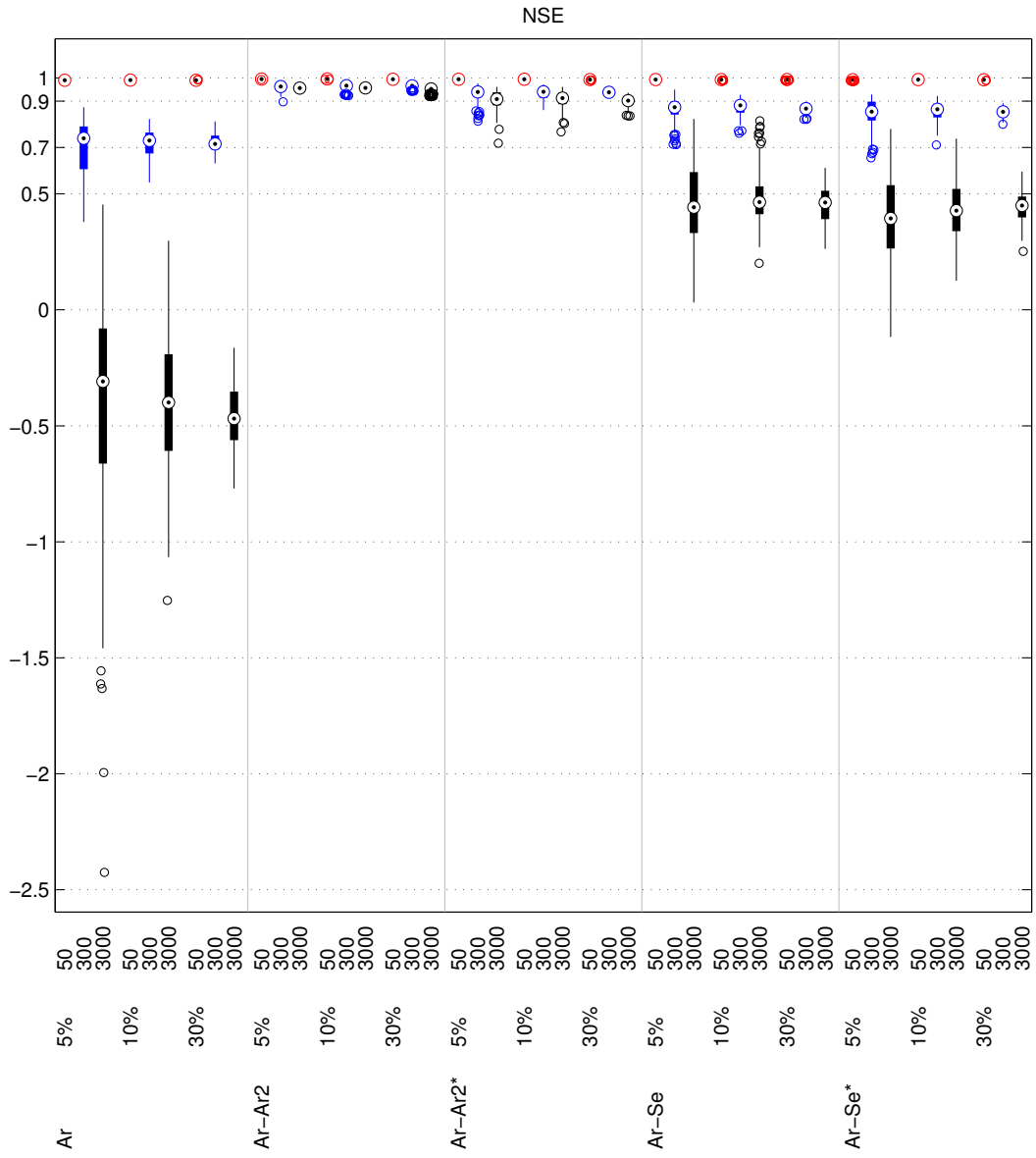


Figure 4.5: Boxplot of the NashSutcliffe model efficiency (NSE) coefficient computed on the simulated missing data portions for all tests and gap scenario classes. Colors indicate the gap size class: 50 (red), 300 (blue) and 3000 (black) time steps.

ferent time-series. The performance of the technique is analyzed by considering a variable missing percentage up to 30% and multiple gaps of size up to 3000 time steps, corresponding to about 20 days. The generated missing data portions show realistic asymmetrical hydrographic structures similar to the one found in the reference even when large gaps are simulated and no Q variable is used. The statistical content lost with the missing data is mainly recovered even when these constitute large portions of the data set, but very rare events may be underrepresented in some cases. Finally, the predictive power of the technique, measured by classical goodness-of-fit measures and compared to the reference values of these indicators, is very high when Q, even if incomplete, is highly correlated to Z. If Q is absent or lowly correlated to Z, the prediction is more uncertain since the variability of the possible data patterns within large gaps is much higher.

The strong point of the proposed approach is its capability of reconstructing the missing data patterns in a very realistic way even if the auxiliary information given is incomplete. Parametric techniques that define a specific cross-correlation structure between the simulated and auxiliary time-series cannot easily deal with partially informed auxiliary variables. This makes the proposed method a convenient alternative for gap-filling in the everyday professional practice, with the only requirement of a representative training data set. Future developments may include the use of other types of auxiliary source of information, like for example a rainfall amount time-series.

Chapter 5

Modeling future climatic time-series according to climate change scenarios

Abstract

This chapter illustrates the application of the Direct Sampling technique to the multivariate simulation of climatic variables. A setup for the simulation of temperature, radiation and humidity using the historical record of a weather station as training data set is proposed. In the first part of the chapter, the technique is tested on the simulation of a stationary climate record, while in the second part, a methodology to incorporate the information of a climate change projection in the simulation is proposed. It is shown that the probability distribution of the variables at multiple scales as well as their complex statistical relation are preserved inside the simulation. It is also shown that, with a simple methodology, the technique can explore the uncertainty related to a climate change scenario.

5.1 Introduction

Stochastic modeling of climatic variables is often required to estimate the uncertainty about present and future climate scenarios. Stochastic simulation algorithms aim at generating synthetic data series representing several possible outcomes of a natural process, honoring the reference statistics as well as respecting the seasonality and the persistence from the daily to the annual scale. Several types of technique have been proposed climatic time-series, the classical weather generators are often based on multivariate autoregressive first order Markov chain [158, 215, 28]. These techniques are based on the linear dependence between the considered climate variables and random components described with a specific family of probability distribution. The main drawback in this case is that they do not allow the preservation of the high-order statistics observed in the data, for example the long-term fluctuations and the non-linear correlation between variables. Increasing the Markov-chain order does not generally bring any substantial improvement to preserve the interdiurnal and interannual variability [107] which is critical for the application. Moreover, as explained in [97], the appropriate probability distribution of the weather variables can be highly site dependent, which requires each time an accurate data analysis and makes the technique less adaptive.

Non-parametric algorithms have been proposed as an alternative to overcome these limits, for example the resampling techniques as the multivariate k-nearest neighbor resampling [114] or the block bootstrap [97]. Another strategy, proposed by [66], uses Principal Component Analysis to identify the most significant statistical dependencies and a technique based on neural networks for the simulation. These approaches are essentially based on pattern recognition and sampling of historical data. They can preserve non-linear relations and the essential statistical dependency with minimal prior statistical assumptions. Nevertheless, they may require a complex formulation and still not preserve the variability at multiple scales. A recent attempt [192] to preserve the interannual variability is a more complex statistical model using a combination of different techniques comprising wavelet decomposition, Markov chain and k-nearest neighbor resampling.

The Direct Sampling technique (DS) [122], belonging to the family of multiple point geostatistics (MPS) methods, is proposed here as a simple resampling strategy to reproduce complex statistical relations between multiple climate variables and preserve their interannual variability. The rationale of DS is the use of a search window to find and compare patterns between the multivariate training data set (the past record) and the simulated data iteratively. Patterns are searched by randomly scanning the historical data. When a similar pattern is found, the datum at its center is directly sampled and assigned to the simulation grid. The main difference with respect to block bootstrap and k-NN resampling is that, with DS, the data are simulated in a random order and the neighbors considered for conditioning

are always the closest informed ones to the simulated time step (see chapter 2, section 2.2.3 for a more detailed comparison with the other resampling techniques). Proceeding in this way, the size and configuration of the conditioning neighborhood changes during the simulation: for the first simulated data, the conditioning neighborhood is sparse and contains points that are distant from each other in time; conversely, in the final iterations, the conditioning neighborhood is dense, containing data that are closer to each other and to the simulated datum. This mechanism, mainly data driven, allows considering the statistical dependency at multiple scales without the need of a complex statistical model. The aim of this chapter is to propose a Direct Sampling setup for the multivariate simulation of climatic variables. This multivariate climate generator is tested on the simulation of the climate record from a weather station in the Aargau region (Switzerland) including temperature, solar radiation and humidity.

The second goal is to propose a simulation framework to assess the uncertainty of a climate change scenario for the same location. One widely used approach to estimate the future local climate change is to use the information downscaled from climate models output and given under form of delta change, which represents the absolute mean variation of climate variables in future. This information is usually added to the present data or to stationary simulated time-series to estimate the future climate variability. According to [7, 26], one main problem is that there are no clear guidelines in literature to estimate the delta change for all the variables required for impact assessment studies in hydrology, for example wind speed, solar radiation, humidity and snow-related quantities. For the Buchs station data set, considered in this study, the Swiss Climate Change CH2011 scenarios [26] give the estimated mean daily temperature change for the period 2012-2085: we propose a methodology that uses this sole information to explore the variability of the climate change for daily temperature, global radiation and humidity. This is done by using the linear relation among these variables to impose the mean climate change, while the complex structure of the residuals to this relation is simulated using DS.

The chapter is structured as follows: a DS setup for the multivariate simulation of daily temperature, global radiation and humidity is presented in section 5.2 and tested on a real data set of the Swiss Plateau in section 5.3. Moreover, a simple technique to incorporate in the methodology the downscaled information of a climate change scenario is proposed in section 5.4. Section 5.5 is dedicated to the conclusions.

5.2 A DS setup for multivariate climate time-series

The rationale of DS is based on the reproduction of patterns similar to the ones found in a representative historical record used as training data set (TI). The simulation, taking place in an empty or incomplete vector called Simulation Grid (SG), is based on the *DeeSse* implementation [193], described in chapter 2, section 2.2.2. The technique is used to simulate multiple time-series together, using multivariate conditioning patterns. This way, both the statistical relation between variables and their heterogeneity are preserved in the simulation. Each variable is a function of time $f(t)$, but, where possible, we simplify the notation by omitting (t). The multivariate setup shown in table 5.1 is composed by the following variables: 1-6) the 365-day Moving Average ($365MA$) and the Moving Sum of the current day and the one before ($2MS$) for each climate variable are used to preserve the interannual variability and the day-to-day correlation observed in the data. 7-8) Two periodic triangular functions ($A1$ and $A2$) describe the annual seasonality. 9-11) The target variables are daily temperature (Z), global radiation (S) and humidity (H). The main parameters of the algorithm are the search time interval length R, the maximum number of conditioning data N, the maximum scanned TI fraction F and the distance threshold T. These parameters are manu-

ally set according to the available data size and the desired temporal dependence imposed in the simulation. The user may change the setup according to the following guidelines. Being interested in conditioning the simulation upon the interannual fluctuations and disposing of a TI of less than 15 years, R is fixed to 500 days for $365MA$ and the target variables. The user may increase this value if the available historical record allows the observation or larger repeated structures, for example interdecadal fluctuations. The value $N = 15$ results to be sufficient to describe the heterogeneity without increasing the computation time. To adapt the technique to a different data set, this value can easily be reset by trial end error: N takes integer values, the suggested range is [5-25] (see for example the sensitivity analysis on rainfall data in appendix A). Conversely, for all $2MS$, $A1$ and $A2$, the time-dependence is limited to lag 1 by using $N = R = 1$, since we have no interest in expanding or varying the time lag-dependence for the mentioned variables. The threshold T has been kept at the same value of 0.05 for all variables as used for daily rainfall time-series (see chapter 2). It is appropriate to decrease this value if the simulation shows a considerable added noise with respect to the training data. This may happen in case of highly autocorrelated variables (smoother signals).

Table 5.1: Standard setup proposed for rainfall simulation. The parameters are: search window radius R, maximum number of conditioning neighbors N and distance threshold T. The variables are: 1-6) the 365-day Moving Average ($365MA$), the 2-day Moving Sum ($2MS$) for each climate variables (Z, S and H), 7-8) triangular functions describing the annual seasonality ($A1$ and $A2$) and 9-11) daily temperature (Z), global radiation (S) and humidity (H).

Variable	R	N	T
1) $365MA_Z$	500	3	0.05
2) $2MS_Z$	1	1	0.05
3) $365MA_S$	500	3	0.05
4) $2MS_S$	1	1	0.05
5) $365MA_H$	500	3	0.05
6) $2MS_H$	1	1	0.05
7) $A1$	1	1	0.05
8) $A2$	1	1	0.05
9) Z	500	15	0.05
10) S	500	15	0.05
11) H	500	15	0.05

The maximum scanned TI fraction F is set to 0.5, meaning that maximum half of the TI is scanned in each iteration to find a suitable candidate to sample. This limitation is necessary to reduce the oversampling of a specific region of the TI and the subsequent reproduction of an extended data portion.

Finally, an important feature of the algorithm when applied to a multivariate data set is the use of the Variable Vector Mode (VVM). When VVM is used, the simulation of all variables at the same time step of the SG is done simultaneously by looking for a similar multivariate pattern in the TI and sampling the whole variable vector at the chosen time step. This way, the values of the different variables remain associated inside the simulation, which would not be the case if the values were simulated in random order for each variable separately. Even if VVM tend to reduce the variability in the simulation, it is preferred since it helps respecting the complex statistical relation between different variables (see the preliminary tests on rainfall data, appendix B).

5.3 Stationary simulation

In this section, we show the results of a stationary simulation, composed of 100 realizations of the same length as the training data. The aim of this experiment is to test if the technique is capable to generate a realistic stationary simulation of a multivariate climatic record, preserving the statistical structure and relationship among the variables. The proposed DS setup is used to simulate the 26-year (1985-2011) daily climatic record from the Buchs station (Aargau region, Switzerland, Meteoswiss database), including daily mean air temperature [°C], global radiation [W/m^2] and humidity [%] measured 2 m above ground level. Apart from $A1$ and $A2$ that are known a priori, all the auxiliary variables that complete the setup are computed from the data. The training data set presents no missing value, but their possible presence does not compromise the simulation, as explained in chapter 2, section 2.3, if the available data are still representative of the heterogeneity. The mentioned historical record constitutes both the training and reference data sets. In the following, the performance of the algorithm is analyzed by mean of some visual and statistical indicators describing the probability distribution and the correlation among the climate variables.

The visual comparison between two portion of reference and simulated data is shown in figure 5.1. The time span of about 3 years shows that global radiation (S) has strong seasonal and interdiurnal fluctuations. Humidity (H) also presents strong day-to-day fluctuations but a weaker seasonality. Temperature (Z) is more autocorrelated from one day to the next and shows a quite clear seasonal variation following S. The algorithm generates realistic structures (figure 5.1, b) preserving the annual fluctuations in the range of values shown by the reference. A modest added noise is visible in the simulated temperature time-series.

The preservation of the seasonality is confirmed by the boxplot of figure 5.2: the monthly mean, average maximum and minimum daily values is preserved for all variables. The bias in the simulation is very small and the variability shown by the simulation ensemble is of about 2 °C for Z, 20 W/m^2 for S and 3-5 % for H.

The probability distributions at the daily and annual scales are analyzed using qq-plots considering all the observed values (figure 5.3). At the daily scale, all variables show an accurate match between the simulation and the reference distribution. Nevertheless, as also stated in the previous chapters, the daily values generated are limited to the ones found in the TI since the algorithm is based on resampling. For this reason, under-representation of the extreme values may occur in case of scarce training data. At the annual scale the distribution is preserved fairly well: the median of the realizations (dotted line) shows only a modest bias for low S values (figure 5.3, d). The 05-95-th percentile boundary of the realizations show that, at the annual scale, new values beyond the range observed in the training data are simulated. This is possible by generating new daily data patterns using the same data present in the TI. This aspect of the uncertainty is explored more efficiently at the annual scale than at the daily one. It is also worth considering that the statistical indicators computed at the annual scale are more prone to noise with respect to daily data statistics, since they are based on a smaller amount of data (26 years).

The correlation between the variables observed in the simulated and reference data is analyzed by mean of the empirical joint probability density function (JPDF, figure 5.4). The complex relation between temperature and both radiation and humidity is accurately preserved in the simulation. This result is achieved with the reproduction of complex patterns and the use of the Variable Vector Mode, which allows sampling all the variables at the same time step of the TI (see section 5.2).

In summary, the daily and annual distributions and the complex statistical relation between the variables are well preserved. The fact that the estimated uncertainty sometimes is very low (see e.g. the confidence boundary of the qqplots of figure 5.3) raises the question of

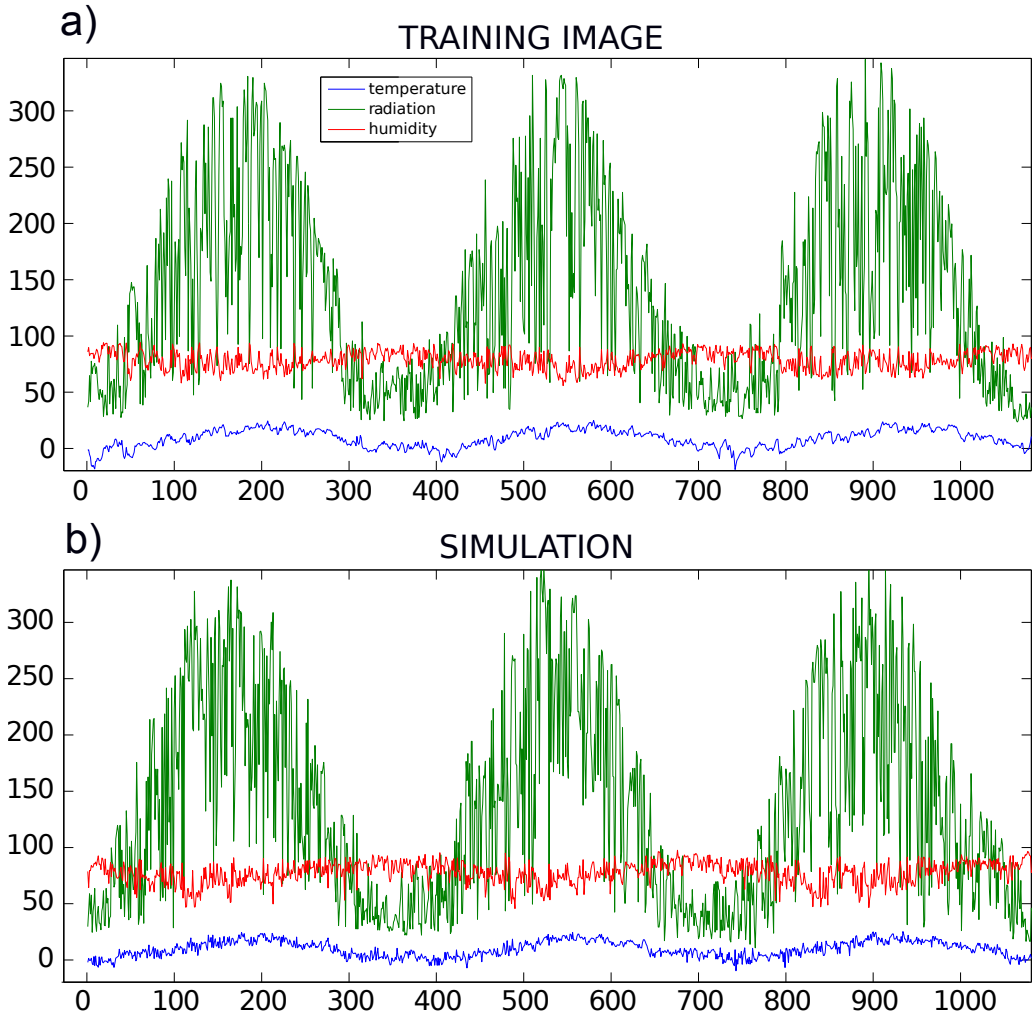


Figure 5.1: Visual comparison between two portions of the reference (a) and a simulated (b) multivariate time-series: daily temperature (Z, °C), global radiation (S, W/m^2) and humidity (H [%]).

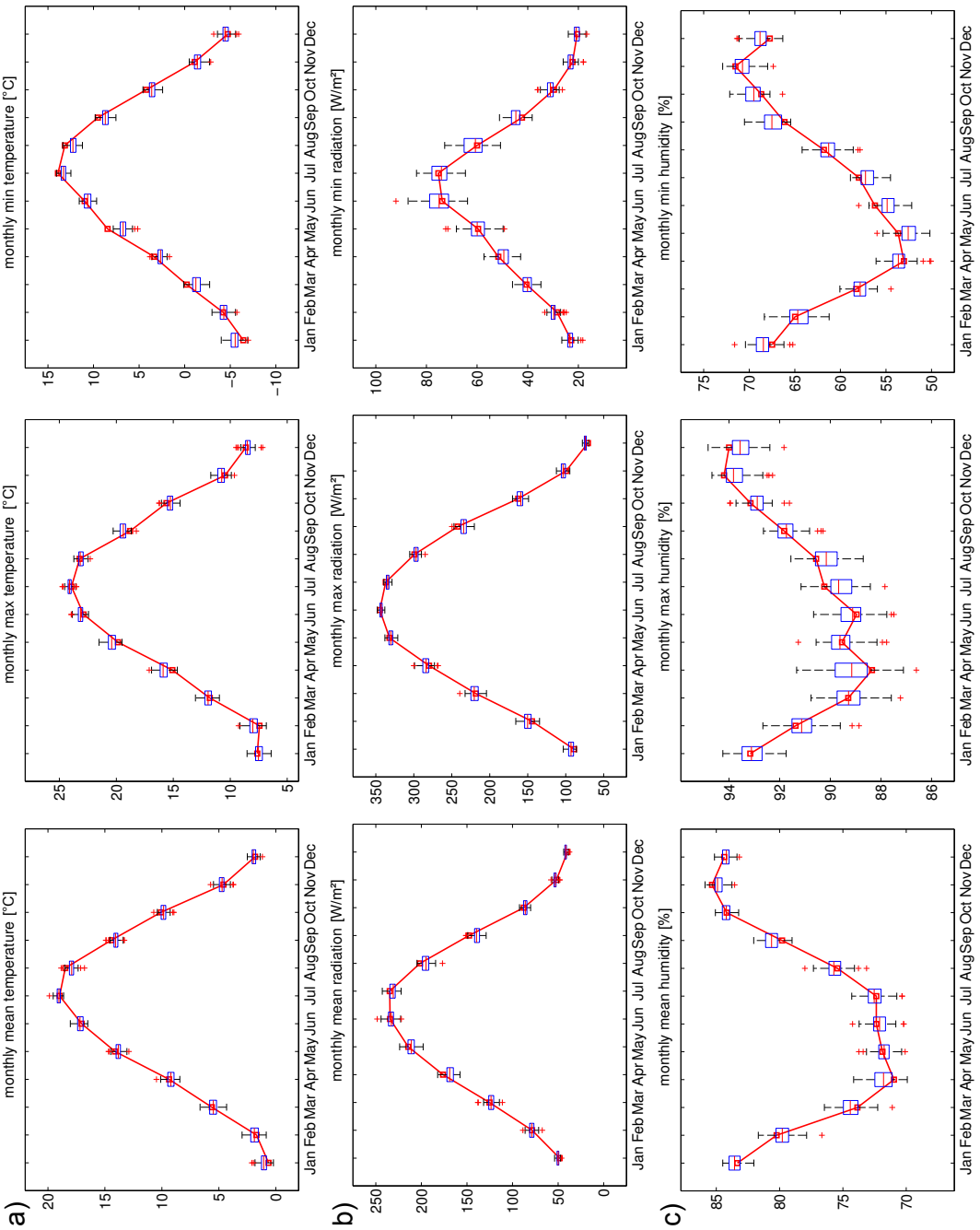


Figure 5.2: Boxplot of the daily mean, average maximum and minimum values of the reference and simulated time-series for each month (period 1985-2011): a) temperature, b) global radiation and c) humidity. The solid line represents the reference.

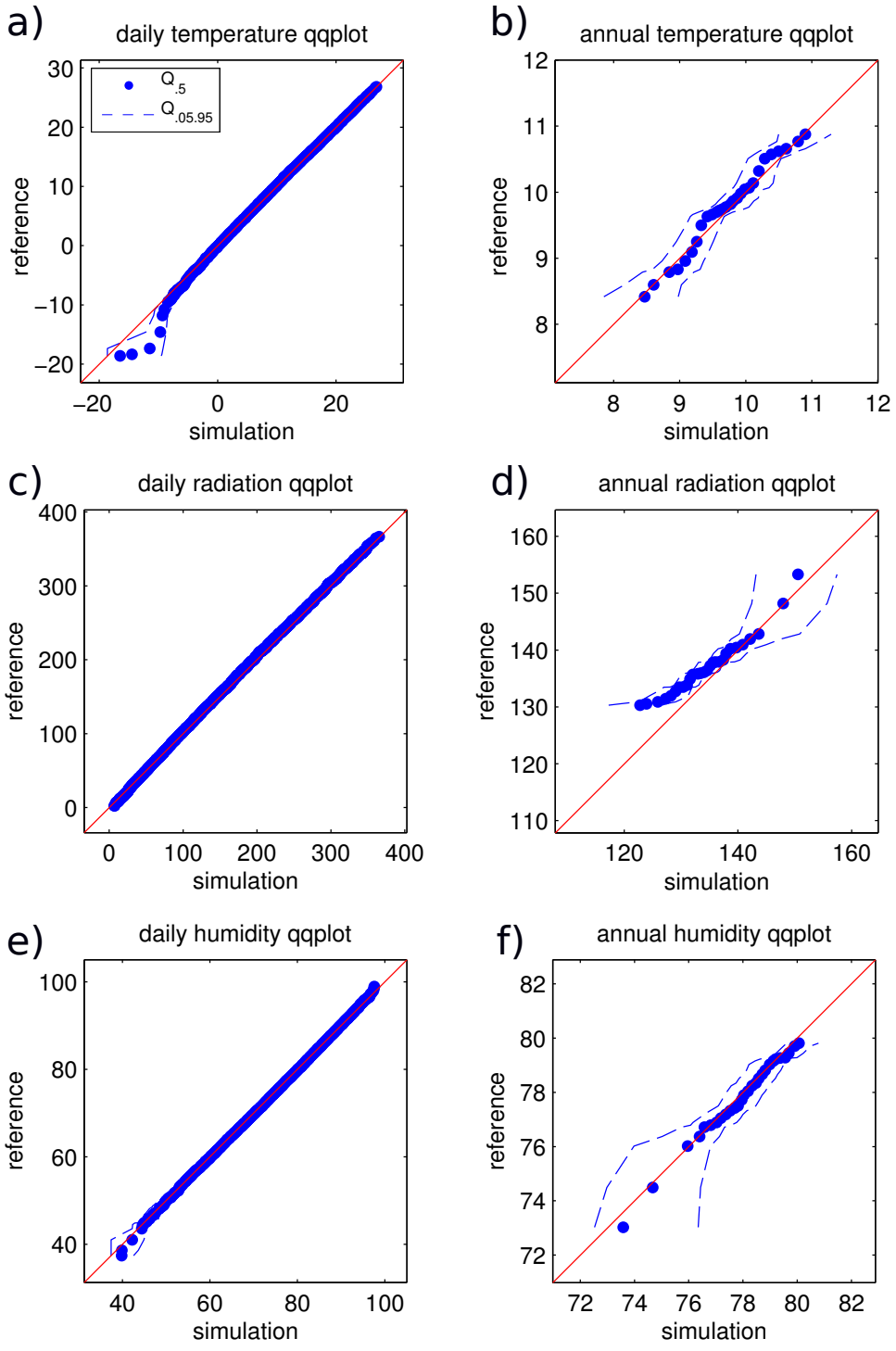


Figure 5.3: QQ-plot comparing the empirical distribution of the reference data with the simulation ensemble at the daily (left column) and annual scale (right column) for all variables. Dotted line indicates the median ($Q_{.5}$) and dashed lines the 5-th and 95-th percentiles of the realizations.

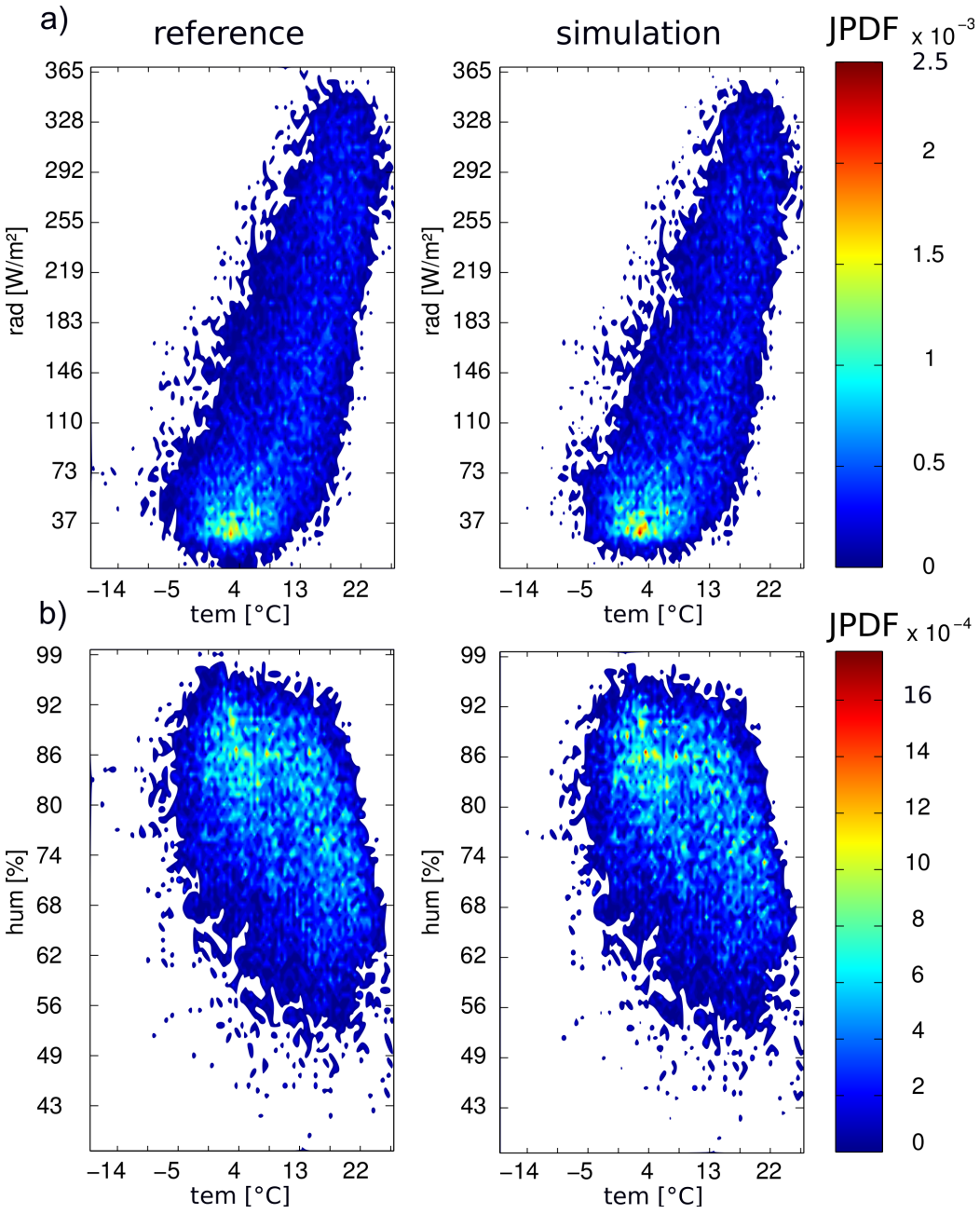


Figure 5.4: Empirical joint probability density functions (JPDF) for temperature-radiation (a) and temperature-humidity (b), observed in the reference and simulation.

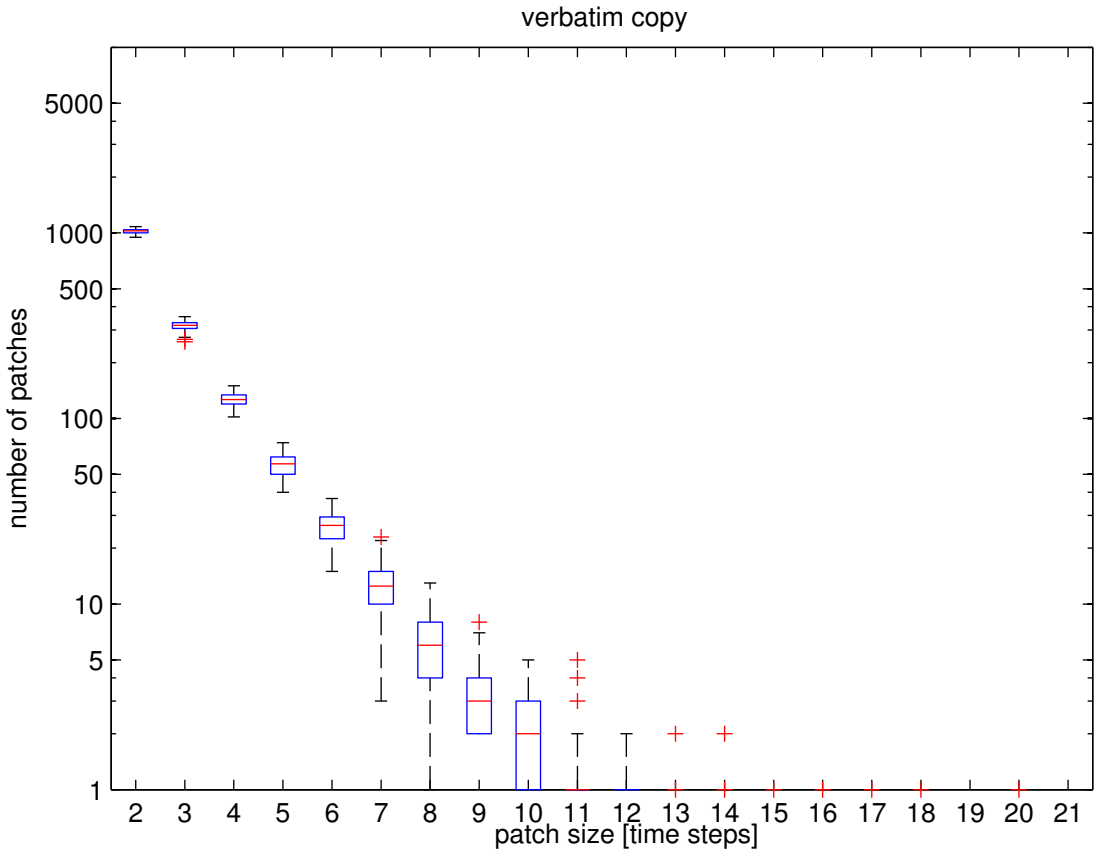


Figure 5.5: Boxplot of the verbatim copy (number of patches) observed in the simulations in function of the patch size.

whether the algorithm generates the exact reproduction of entire portions (patches) of the training data set by sampling, iteration by iteration, always at the same location of the TI (verbatim copy). By keeping track of the time stamp of each sampled datum, it is possible to analyze how much this phenomenon takes place in the simulation. The boxplot of figure 5.5 shows the frequency distribution of the number of data patches found in the simulation as a function of their area, the uncertainty given by the multiple realizations considered. The majority of the patches is formed by less than 5 subsequent time steps, with few cases up to 20 steps. This indicates that, in this case where the simulations have the same length as the training data set, the verbatim copy does not dominate the simulation. Nevertheless, this phenomenon may increase if the simulated time-series are significantly longer than the training data.

5.4 Exploring the variability of a climate change scenario

In this section, we propose a simple methodology using the proposed simulation technique to assess the uncertainty of a climate change scenario. The climate change information for the Buchs station is provided with the widely used approach of the delta change factor. The large-scale mean change for temperature until 2085 is estimated by the ALADIN project (CNRM, France). The downscaled mean temperature change $\Delta Z(t)$ for the Buchs station [26] is used here. $\Delta Z(t)$ is given for each day of the year (DOY) of 2035, 2060 and 2085 and it is interpolated using the cubic spline method for all the intermediate years (figure 5.6).

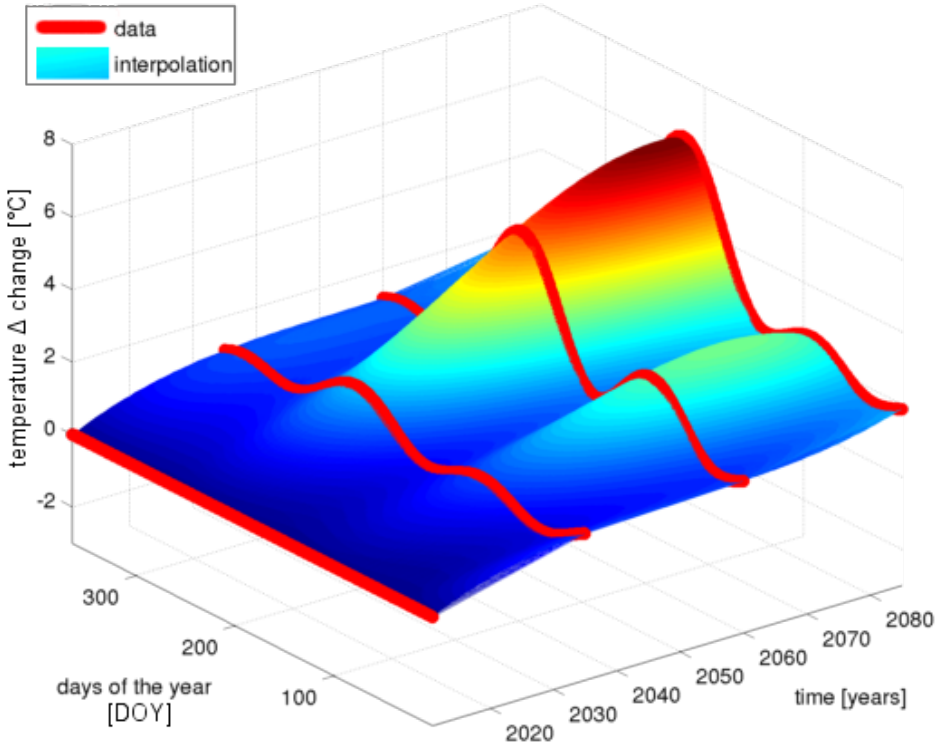


Figure 5.6: The temperature delta change scenario downscaled for the climate station of Buchs and the period 2012-2085. The delta change index is given for all the days of the year (DOY) of 2035, 2060 and 2085 and interpolated for the remaining days.

The mean change in temperature is imposed by adding $\Delta Z(t)$ to an ensemble of simulated time-series using the DS setup proposed in section 5.3. Then, since for the other variables the delta change is not available, the mean temperature change is converted to a mean change in radiation and humidity using a simple linear regression model that relates the variables. The residuals, describing the more complex part of the variability, are simulated using DS. This way, the mean change is imposed as a deterministic component with the delta change approach, while its residual is stochastically simulated to explore the variability of the climate change scenario. The proposed simulation workflow is the following.

Pre-processing

- Eliminate any important trend or shift visible in the available time-series data set, using the techniques presented in appendix C. In this case, only a large-scale trend has been observed and eliminated from the radiation historical record.
- Let us consider the stationary daily temperature $Z(t)$, radiation $S(t)$ and humidity $H(t)$ time-series, where $t = t_0, ..., t_n$; is the time reference.
- Applying the equation C.1 in appendix C, compute the linear regression $\hat{S}(Z(t))$. For each DOY d, compute a different regression parameters α_d and β_d, estimated using the data from all available years in the DOY time span $d \pm 30$ days.
- Similarly, estimate the linear regression $\hat{H}(S(t))$.
- From the training data set (Z, S, H) compute the regression residuals $\epsilon_S(t) = S(t) - \hat{S}(Z(t))$ and $\epsilon_H(t) = H(t) - \hat{H}(S(t))$. This way, a new training data set $(Z, \epsilon_S, \epsilon_H)$ is

obtained.

Simulation

- Using the DS setup previously introduced for stationary simulations (section 5.3), simulate the group of variables $(Z', \epsilon'_S, \epsilon'_H)$ using as training data set $(Z, \epsilon_S, \epsilon_H)$.
- Finally, the simulated variables (Z^*, S^*, H^*) representing the future climate scenario are computed as follows:

$$\begin{aligned} Z^*(t) &= Z'(t) + \Delta Z(t) \\ S^*(t) &= \hat{S}(Z^*(t)) + \epsilon'_S(t) \\ H^*(t) &= \hat{H}(S^*(t)) + \epsilon'_H(t) \end{aligned} \tag{5.1}$$

  where $\Delta Z(t)$ is the delta change factor given for temperature.

This approach, inspired by the classical moving average time-series models, imposes the climate change information to S^* and H^* as a deterministic component function of a predictor variable ($\hat{S}(Z^*)$ and $\hat{H}(S^*)$), while a stochastic component (ϵ'_S and ϵ'_H) is simulated using DS. Figure 5.7 shows the scatter plot of the training data, their projection using the linear regression and the residuals: the linear regression estimated separately for each DOY represents the main trend of the statistical relationship among the variables, but a large part of the variability is left to the residuals. The difference with a canonical moving average model is that, here, the residual components ϵ'_S and ϵ'_H, instead of being simulated as a simple random noise, are co-simulated with Z' using DS, preserving their high-order statistical structure and correlation (see section 5.3). Moreover, since Z' is a component of Z^*, which is at the base of the two linear regressions, there is also an indirect dependence between the deterministic and stochastic components of S^* and H^*. This method relies on the following hypothesis: the linear relations among the climate variables used to compute $\hat{S}(Z^*)$ and $\hat{H}(S^*)$ and the residual components ϵ'_S and ϵ'_H are supposed to be stationary in time.

The technique is applied by generating 100 realizations of the multivariate time-series for the period 2012-2085. In absence of a reference data set, the results are compared with the 1988-2011 historical record to see the magnitude of the mean change and the variability. In figure 5.8 the boxplot of the mean, average maximum and minimum values are analyzed for each month. The main trend observed is an increase of mean and extremes for temperature and radiation in summer together with a decrease of humidity. Conversely, the simulated residuals allow the exploration of the uncertainty.

At the annual scale (figure 5.9), we observe a mean change reflecting the given daily delta change but also an important year-to-year variability, especially for humidity. The variability simulated with DS at the annual scale allows exploring the variability around the mean imposed by the climate model.

5.5 Conclusions

In this chapter, we proposed a methodology for the multivariate simulation of climate variables using the Direct Sampling technique. The technique is based on the use of a multivariate training data set containing the target variables together with two types of auxiliary variables: the 365-day Moving Average, which helps preserving the interannual variability, and the 2-day moving sum to preserve the day-to-day correlation inside the time-series. The setup is applied to the simulation of the 1985-2011 daily climatic record from Buchs (Aargau, Switzerland), including temperature, global radiation and humidity (100 realizations). This setup can generate realistic replicates of the multivariate data set by simply respecting the

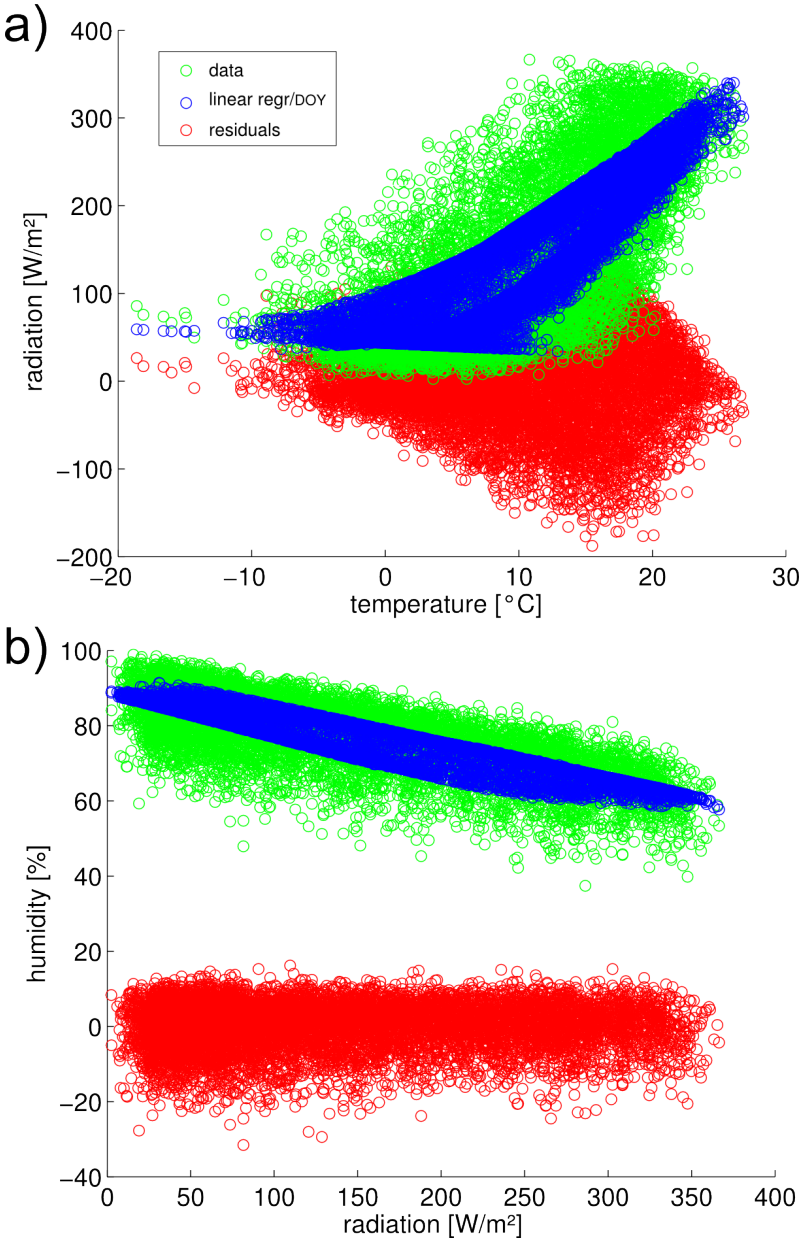


Figure 5.7: Scatterplot between temperature-radiation (a) and radiation-humidity (b) showing: the training data (green), their projection using the regression coefficients computed on each DOY (blue) and the residuals (red).

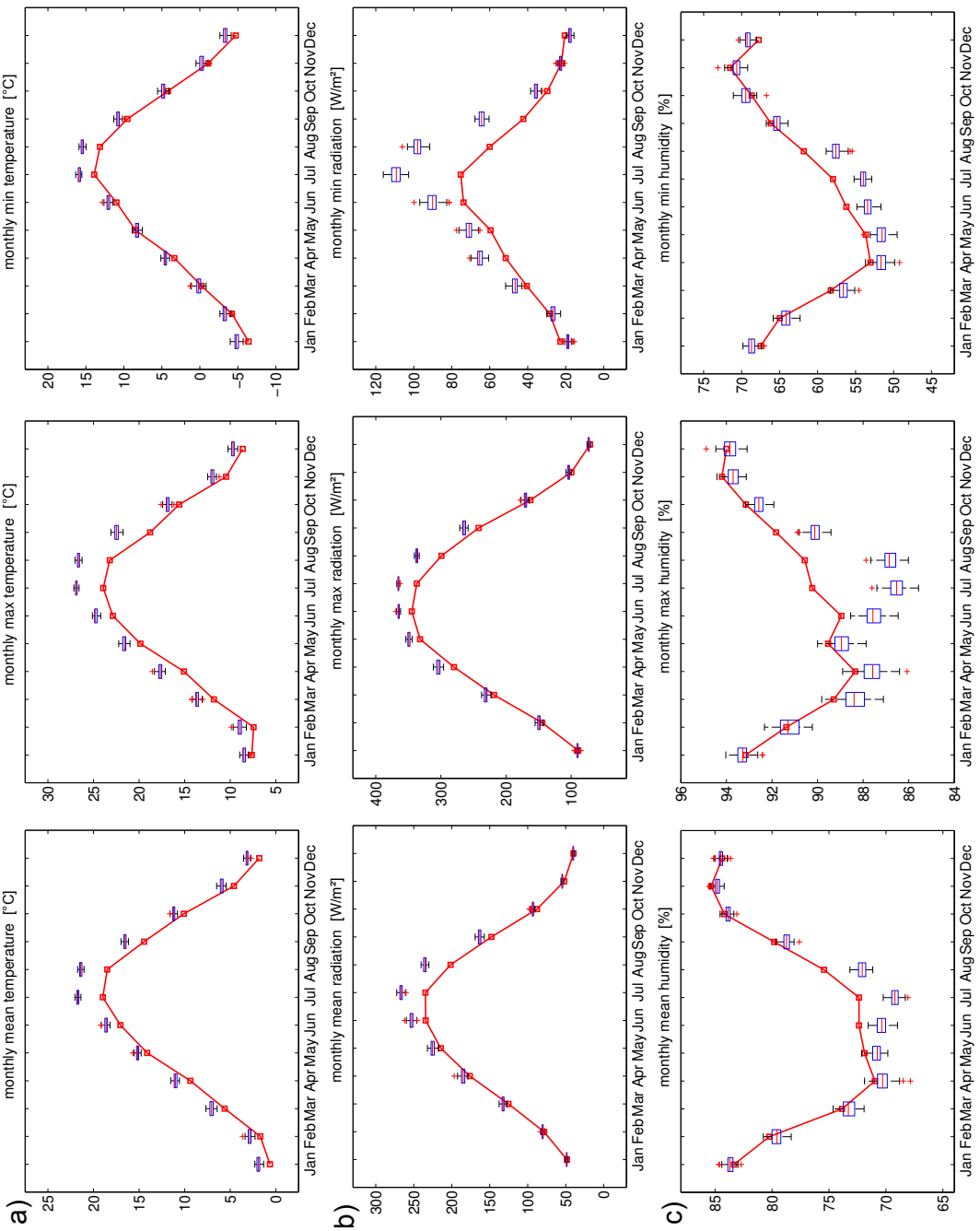


Figure 5.8: Boxplot of the daily mean, average maximum and minimum values of the simulated time-series for each month (period 2012-2085): a) temperature, b) global radiation and c) humidity. The reference line refers to the historical values of the period 1985-2001.

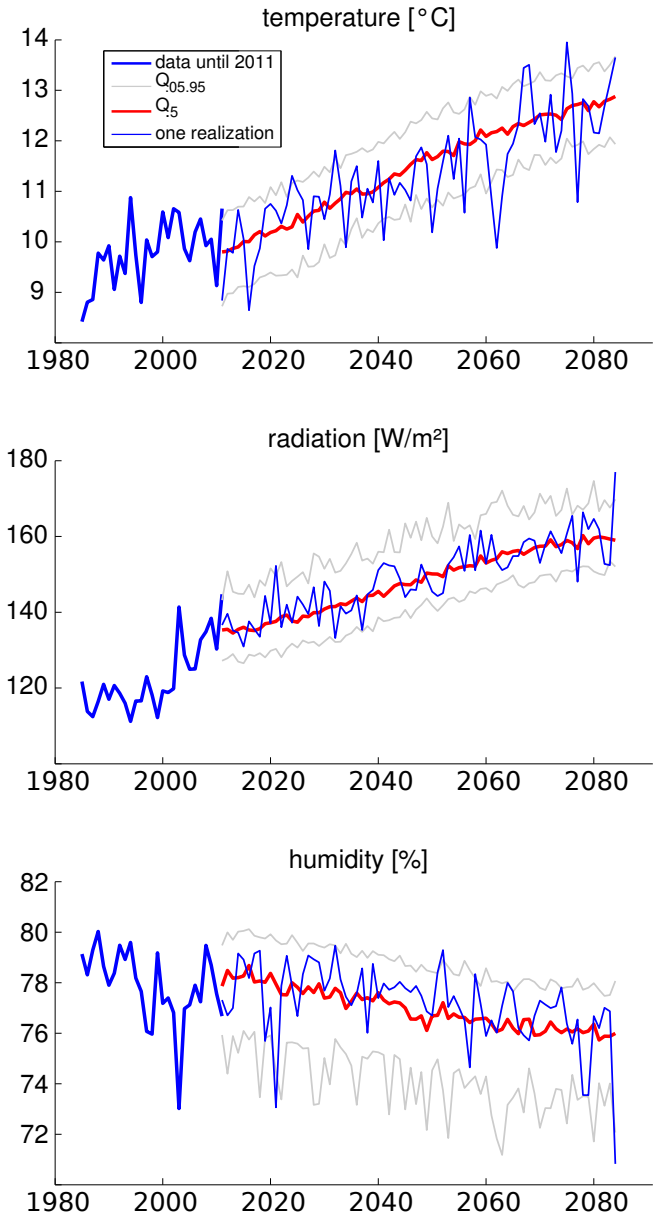


Figure 5.9: The annual time-series of the given and simulated data: the red line indicates the median ($Q_{.5}$) of the realizations, the grey lines the 5-th and 95-th percentile boundary and the blue line is one example of realization.

similarity of multiscale patterns observed in the training data. In particular, the probability distribution at the annual scale is preserved for all variables, a result difficult to achieve with traditional parametric simulation techniques. For this reason, Direct Sampling used with the proposed setup, can be a practical tool to explore the variability of multivariate climate time-series if the used training data set is sufficiently long and representative. The generated multivariate data sets can be used as input of hydrological models to study the propagation of the uncertainty to the basin hydrological response. Nevertheless, since DS cannot generate daily values not contained in the training data set, the proposed technique is not suitable to explore the long-term extremal behavior of the process at the daily scale.

In the second part of the chapter, a methodology to simulate the daily variability and evaluate the corresponding uncertainty around a climate change scenario is presented. The mean change for temperature is given for the period 2011-2085 and extended to radiation and humidity using the linear correlation between these variables, while the complex structure of the residuals is simulated using DS. The results show that the simulated climate change has a specific seasonal structure and a significant interannual variability. The technique allows estimating the climate change for radiation and humidity which is normally not given in the climate change assessment scenarios, but it relies on the hypothesis of stationarity of the linear relation between the variables that still needs to be proven on a physical basis. Moreover, the residual component shows a complex structure but it is simulated as a stationary signal, which may be an oversimplification.

Nevertheless, some improvements to this technique are possible: for example, the deterministic component imposing the delta change to radiation and humidity could include the dependence with other climate variables to estimate a more realistic climate change. Moreover, multiple training data sets, representing the present and future climatic heterogeneity, can be used to simulate a more complex climate change, regarding not only the mean values but also the residuals. In fact, with the current DS implementation [193], it is possible to change progressively the training data set used inside the simulation as function of time, to mimic a transition to different climate conditions (see the application to rainfall time-series in appendix D). A training data set used to represent the future climate could be, for example, the historical record of a nearby location, presenting the climatic conditions expected in future for the location of interest. This method relies on the concept of climate zone transition, which is not easy to predict for the small scale. Nevertheless, it may be applicable to locations presenting a small change in latitude or altitude and a very similar geographical setting, including the distance from the sea, the wind regime and exposure, the influence of large-scale circulation phenomena and the type of human activity.

In conclusion, if the user disposes of a representative training data set, the DS algorithm can be a convenient tool to assess the uncertainty about current climate data by simulating an entire stationary multivariate data set. Moreover, in absence of a more complete information about the climate change, the simple delta change approach can be integrated with the proposed technique to simulate more realistic future climate scenarios.

Chapter 6

Simulation of daily rainfall radar images conditioned by elevation.

Abstract

Nowadays weather radar is the sole current instrumentation giving continuous high-resolution images of the rainfall heterogeneity in space. Simulating rainfall radar images is useful to study the uncertainty related to spatial rainfall fields and its propagation to distributed hydrologic models. The Direct Sampling technique is proposed here as a stochastic generator of spatial daily rainfall fields based on the simulation of radar imagery. A standard Direct Sampling setup for radar image simulation is presented and tested on the simulation of three groups of images from a 10-year historical record from the central region of Israel. If correctly trained, the algorithm can generate realistic rainfall fields for different rainfall types, preserving the variability and the covariance structure of the reference reasonably well. Moreover, conditioning the simulation using the digital elevation model allows preserving the complex relation between rainfall amount and elevation.

6.1 Introduction

As already explained in chapter 1, the small-scale variability of rainfall can lead to unpredictable and extremely variable hydrological response [216, 219, 50, 11, 197, 170, 71]. A realistic representation of spatial rainfall at the sub-catchment scale is therefore important to make reliable predictions about the basin recharge and extreme hydrological events. Spatial rainfall fields at kilometric scale are primarily needed as input for distributed hydrological models to make detailed predictions of the hydrological response in space and time. Last generation weather radars are normally the sole data source of this kind. Moreover, weather radar images constitute an indirect measure of the rainfall amount carrying different types of systematic errors reflecting in a biased statistical content. In the last decades, conspicuous efforts have been devoted to the development of methodologies for the correction and improvement of weather radar estimations. Some of them are mainly based on geostatistics, like the data merging techniques [101, 39, 202, 179, 205, 177]. These methods aim at applying a correction to a specific group of radar fields by interpolating or simulating the error between the ground measurements and the corresponding radar values. Conversely, some techniques simulate, with stochastic or physical models, the response of a radar apparatus given a known synthetic rainfall field [117, 173, 49, 152, 40]. This approach aims at quantifying the magnitude and type of systematic errors associated to a specific radar instrumentation, to perform an appropriate correction of real radar data or plan the best location of a new radar station.

Once a radar data set has been corrected, it can be used to train a statistical model to study the uncertainty of the actual rainfall heterogeneity represented in space. There exist several stochastic methods to simulate an ensemble of spatial rainfall fields. Some techniques perturb the given radar images with a random noise, generated as a multi-Gaussian field correlated to the radar rainfall amount [2] or showing a more complex covariance structure [56, 95]. A more sophisticated approach is represented by object-based algorithms [63, 138, 162, 146, 149] that generate new rainfall fields. Each rainfall cell is simulated as a Poisson-point process or using a Markov-chain approach and then its shape and evolution is modeled as a random process following a series of parametric statistical laws. This type of approach can lead to realistic fields but underestimates in some cases the extreme values and usually requires a large amount of parameters. Among the recent geostatistical techniques, Leblois et al. [113] propose a model taking into account the anisotropy, advection and turbulence of rainfall cells and test it using some reference radar images. Another recent geostatistical method is the dry-drift model [166], which imposes a deterministic structure on the rainfall amount, augmenting with the distance from the dry/wet limit. Both works propose as future

goal the possibility to introduce non-stationarity in the model conditioned by topographic elevation. This is considered as one primary influence factor on rainfall amount [see for example 74, 222, 159, 163] and has been already employed as conditioning variable in geostatistical techniques for rainfall data interpolation [164, 153, 16].

We propose here a radar image simulation method based on the Direct Sampling technique (DS) [122], from multiple-point statistics (MPS). MPS is based on the concept of training data set: a data set that represents the heterogeneity of the variable of interest, used to estimate the probability of occurrence of each event conditionally to its neighbor pattern. This way, realistic data patterns are generated in the simulation, preserving the high-order spatial dependence that characterizes the heterogeneity. To our knowledge, there has been only one case of application of MPS to spatial rainfall simulation: the simulation of rainfall occurrence (dry/wet pattern) using satellite images as training data [218]. With respect to previous MPS algorithms, DS allows simulating continuous variables. The methodology presented here benefits from this feature to simulate the rainfall occurrence and amount at the same time conditioned by the topographic elevation. The technique is tested on the simulation of three groups of daily weather radar images from the central region of Israel, each of which represents events of different magnitude. The aim is to test the capability of the method in generating a group of spatial fields (not correlated in time), presenting a realistic rainfall heterogeneity.

The chapter is organized as follows: section 6.2 describes the data set used, section 6.3 presents the methodology and the DS setup, section 6.4 the statistical tools used to analyze and compare the simulated fields to the reference, the results are shown in section 6.5 and section 6.6 is devoted to the conclusions.

6.2 The data set

The study area covers a surface of 125 km$\times$124 km comprehending the western coastal region from Tel Aviv to Gaza and the northern part of the Nagev desert, the central mountain chain of Jerusalem and the beginning of the depression zone associated to the Jordan Valley on the east. The rainfall radar data (Israel Meteorological Service IMS) are hourly images obtained from a non-Doppler C-band radar station located at the north of the study region, for the periods 1991-1995 and 2001-2005. The data are corrected using ground measurements and the hourly images are cumulated to the daily scale. The spatial resolution is of 1 km. To condition the rainfall simulation using elevation, the Digital Elevation Model (DEM) provided by the Survey of Israel has been used. The original 25-m resolution is reduced to 1-km using the nearest neighbor interpolation technique. As seen from figure 6.1 c, the study region presents a range of altitude gradually increasing from the sea level along the Mediterranean coast on the west to about 1000m in the central mountains and rapidly descending to negative values in the eastern depression (more than 200 m below the sea level). To create different simulation cases, three groups of 30 images each are randomly selected from the radar database in function of the mean daily rainfall amount mZ: group A contains images presenting $0.5 \leq mZ < 5$ mm, group B presents $5 \leq mZ < 15$ mm and in group C $mZ \geq 15$ mm. This allows obtaining almost stationary training data sets and testing the techniques on different rainfall types.

6.3 Methods

Direct sampling (DS) is a non-parametric multiple-point algorithm presenting one main difference from previous MPS techniques: instead of computing the conditional occurrence probability of each event to generate a random value, DS randomly scan the training data

set until a similar neighbor pattern is found. The value at its center is assigned to the simulated location, without making use of any estimated conditional probability measure. The *DeeSse* implementation [193] used here allows the simulation of multiple variables at the same time. The simulation follows a random path which visits a space referenced empty array $\mathbf{X}$ called simulation grid (SG), becoming progressively populated until rainfall at all locations is simulated. The target variable $Z(x)$ is generated by sampling with replacement the training data set (TI) $\mathbf{Y}$, composed of historical radar images and any eventual auxiliary variable. The following is the main workflow of the algorithm, for a comparison with other resampling techniques see [143]:

1. Select a random position x of the SG that has not yet been simulated.

2. To simulate the rainfall amount (and occurrence) $Z(x)$: retrieve a data event $\vec{d}(x)$, i.e. a group of already simulated or given neighbors of x, according to a fixed circular spatial window of radius R. $\vec{d}(x)$ consists of at most the N informed time steps closest to x inside the mentioned window. The size and configuration of $\vec{d}(x)$ is therefore limited by the user-defined parameters N and R, and the number of already informed neighbors inside the considered window.

3. Visit a random time-step y in the TI, and retrieve the corresponding data event $\vec{d}(y)$.

4. Compute a distance $D(\vec{d}(x), \vec{d}(y))$, i.e. a measure of dissimilarity between the two data events. For categorical variables (e.g. the dry/wet rainfall sequence) the proportion of non-matching elements of $d(\cdot)$ is used as criterion, while for continuous variables the choice is the mean absolute error.

5. If $D(\vec{d}(x), \vec{d}(y))$ is smaller than a fixed threshold T, assign the value of $Z(y)$ to $Z(x)$. Otherwise repeat from step 3. to 5. until the value is assigned or a prescribed TI fraction F is scanned. T is expressed as a fraction of the total variation shown by Z in the TI. For example, $T = 0.05$ allows $D(\vec{d}(x), \vec{d}(y))$ up to 5% of this total variation. In case of a categorical variable, $T = 0.05$ allows a mismatch between $\vec{d}(x)$ and $\vec{d}(y)$ for 5% of the composing neighbors.

6. If the prescribed TI fraction F is covered by the scan, assign to $Z(x)$ the scanned datum $Z(y^*)$ that minimizes D.

7. Repeat the whole procedure until all the SG is informed.

The same process is applicable to a multivariate data set, where a vector $\vec{Z}(x)$, composed of rainfall amount and some auxiliary variables, is simulated instead. The parameters N_k, R_k and T_k allow defining different pattern dimensions and acceptance threshold for each k-th variable. Where a variable in the SG is already informed, it is used as conditional datum together with the already simulated neighbors.

A primary feature of the algorithm is that $\vec{d}(x)$, the neighborhood used to scan the TI, varies throughout the course of the simulation. This allows to consider large-scale patterns of sparse data at the beginning of the simulation and small-scale patterns of data close to the simulated point towards the end of the simulation. This way, the multiple-scale spatial dependence of the heterogeneity is preserved in the simulated field without needing a complex statistical model or excessive parametrization. The main parameters of the model are the following: the maximum scanned TI fraction $F \in (0, 1]$ at each scanning iteration, the search neighborhood radius R, the maximum number of neighbors N, both expressed in number of pixels, and the distance threshold $T \in (0, 1]$. Recall that, apart from F, each parameter is set independently for each simulated variable. In the setup presented here, the rainfall

amount Z [mm] is simulated together with two auxiliary variables (we simplify the notation by omitting the space reference (x)): 1) The M [m] variable, informing about the elevation of each point. This is completely informed in the SG and given as conditional data and 2) W [-], a categorical variable describing the dry/wet and extreme-value pattern (0 if $Z \leq 0.2$ mm, 1 if $0.2 < Z \leq Q_{0.9}$ and 2 if $Z > Q_{0.9}$, where $Q_{0.9}$ is the 0.9-quantile of Z measured on the current TI). W is cosimulated with Z. An example of one radar image together with the auxiliary variables used are given in figure 6.1. The setup shown in table 6.1 is used together with $F = 0.5$ for all simulation groups. Since the correlation length of the rainfall structures are longer than the domain size, R is set to the maximum value allowed by the current TI size. N and T are set manually by trial and error starting in the range of values indicated by the sensitivity analysis for daily rainfall time-series shown in appendix A. It is worth noting that the distance threshold (T) values, which control the rule of acceptation, are about one order of magnitude lower than the ones used for daily rainfall time-series. This can be explained by the fact that daily rainfall in space is much more correlated than in time: the spatial variation is smoother and requires a more strict rule for pattern acceptation to avoid adding small-scale noise brought by the resampling procedure.

Table 6.1: Standard setup proposed for rainfall radar image simulation. The parameters are: search window radius R, maximum number of conditioning neighbors N and distance threshold T. The variables are: 1) The digital elevation model (M), 2) the dry/wet and extreme rainfall pattern (W) and 3) the rainfall amount (Z).

Variable	R	N	T
1) M	60	10	0.05
2) W	60	10	0.001 (0.002 for group B)
3) Z	60	10	0.001 (0.002 for group B)

6.4 Evaluation

To test and illustrate the proposed methodology, an image series of the same size as the training data set (124×125 pixels $\times$ 30 images, not in temporal continuity) is generated for each one of the three groups described in section 6.2. Moreover, each simulation is repeated 10 times (10×30 images generated for each group). The training data sets are also used as reference. Since the north-west portion of the study area, occupied by the Mediterranean sea, presents missing data for M, the simulation is masked at the corresponding zone, to keep the same number of simulated locations as the informed locations in the training data.

The simulated and reference fields are first visually compared and then analyzed using different statistical indicators. The qq-plot of the rainfall amount is used to compare the rainfall probability distribution. The experimental variogram is used to analyze the covariance structure of the fields and the variability at different scales. For the estimation of the experimental variogram, 5000 randomly chosen points are considered for each image (150000 points considering the 30 images in each series). The qq-plot of the rainfall cell area is used to test whether the areal extent of the simulated rainfall cells is correctly preserved. For this purpose, daily rainfall values lower than 0.2 mm are considered as dry. The empirical joint probability density function (JPDF) is used to analyze the statistical relation between elevation and rainfall amount. Moreover, the internal morphology of wet areas is analyzed by censoring the rainfall field with a variable threshold, ranging from 0.2 to 100 mm in a series of 100 equally distant values. This way, for each rainfall field, 100 different binary images are generated, showing a specific spatial distribution of rainfall cells. Then, the area and the mean rainfall for each cell is computed. Finally, the empirical JPDF of the mean

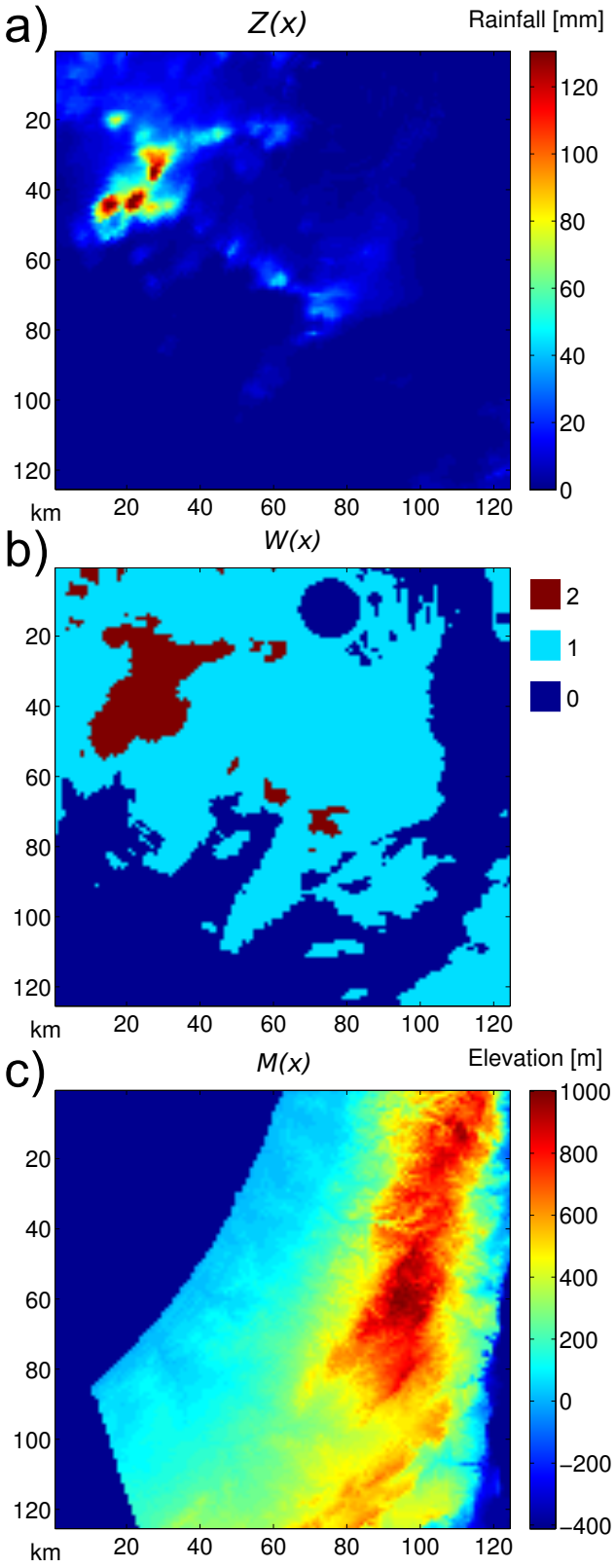


Figure 6.1: One Z image (a) with related W (b) and M (c) variables. The circular zone visible in W is a missing data region surrounding the radar station.

rainfall amount and area is used as statistical indicator describing the internal structure of wet areas.

DS, if not appropriately parametrized, may be prone to verbatim copy, i.e. the reproduction of entire training data set parts (patches). This phenomenon is unwanted since it reduces the small-scale variability in the simulation and it is often associated with biased statistics. Verbatim copy can be monitored by keeping track of the position of each sampled datum inside the TI: a patch is formed if at least two adjacent pixels (sharing one face) are found in the same relative position in both the TI and the SG. The phenomenon is quantified using two indicators computed over the whole simulation ensemble: the maximal generated patch size [%], expressed in percentage with respect to the whole simulation ensemble, and the total patched fraction [%] over all the simulation ensemble. The latter is computed as the sum of all pixels occupied by any patch, divided by the total number of pixels composing the simulation ensemble and multiplied by 100.

6.5 Results

Figure 6.2 shows some examples of the reference fields for each group and, in figure 6.3, some simulated fields are shown. As explained in section 6.4, the top-left corner of the image, corresponding to the area occupied by the sea, is masked in all fields since it presents missing data for the M auxiliary variable. The location of the radar station presenting a circular missing data region in its proximity is visible at the top of the reference fields. As expected, the three groups present a different range of daily precipitation amount. It is worth remembering that the simulations are only conditioned by elevation, so they do not aim at reproducing any specific reference field of figure 6.2, but they should show a similar type of heterogeneity.

The reference field of group A shows two types of events: one of weak intensity covering the coastal and central regions (figure 6.2, a), probably caused by a cold front coming from the sea, and another type showing a more abrupt transition to medium rainfall values, related to the topographic relief of the central region (figure 6.2, b). The simulated fields (figure 6.3, group A) show the same range of values and similar structures. The rainfall intensity inside the rainfall cells follows an ascending trend departing from the dry/wet limit, that we can call dry-drift following [166]. This phenomenon is also observed in the reference fields of all groups.

For group B, the reference shows a more intense event concentrated in some areas of limited extension, with a sharp transition to extreme values (figure 6.2, c) and a second type of field showing a more extended front of moderate intensity coming from the coast (figure 6.2, d). The simulation shows very similar structures in both shape and intensity (figure 6.3, group B). The dry-drift is well preserved for medium and extreme values, but it is less present for low rainfall zones (light blue color), presenting a less clear shape than in the reference images. This may be due to the fact that the training data set of group B contains images representing rainfall events of different types. In this case, the algorithm may simulate a rainfall field by sampling regions presenting a different type of heterogeneity, leading to an unrealistic simulation. A clear example of this phenomenon is given in figure 6.4, showing a simulated field from group B. The left part of the field shows a rainfall heterogeneity typical of an extended front, similar to figure 6.2 d, while, on the right part, a more sharp transition to local extremes suggests a resemblance to localized intense rainfall events, as the ones observed in figure 6.2 c. One main condition required by the technique is not satisfied here: as explained in chapter 2, to avoid the simulation of an unrealistic heterogeneity, the training data set should be stationary, or its non-stationarity should be indicated by an auxiliary variable. In this case, it is necessary to make a more accurate classification of the

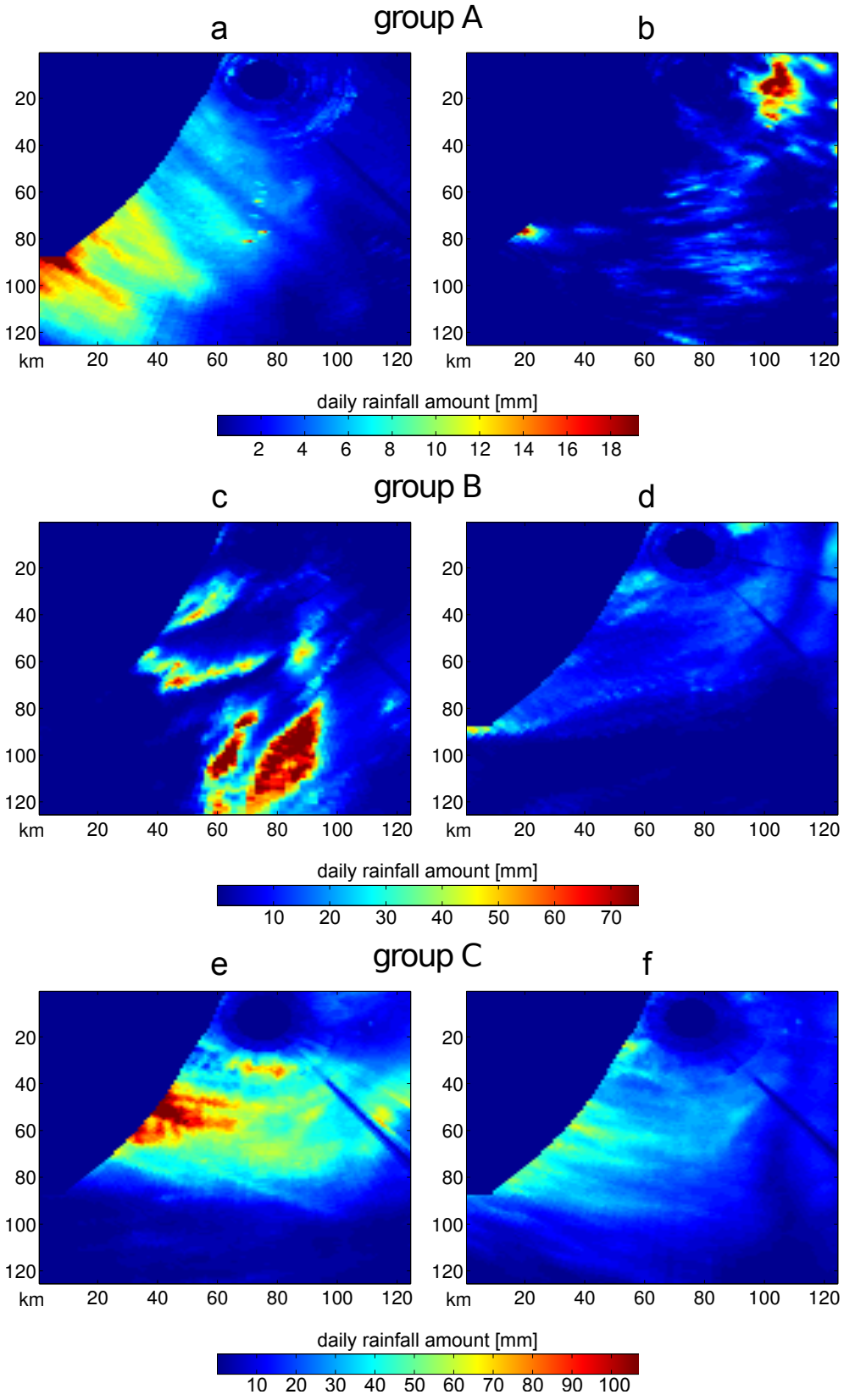


Figure 6.2: Examples of reference rainfall fields for groups A, B and C.

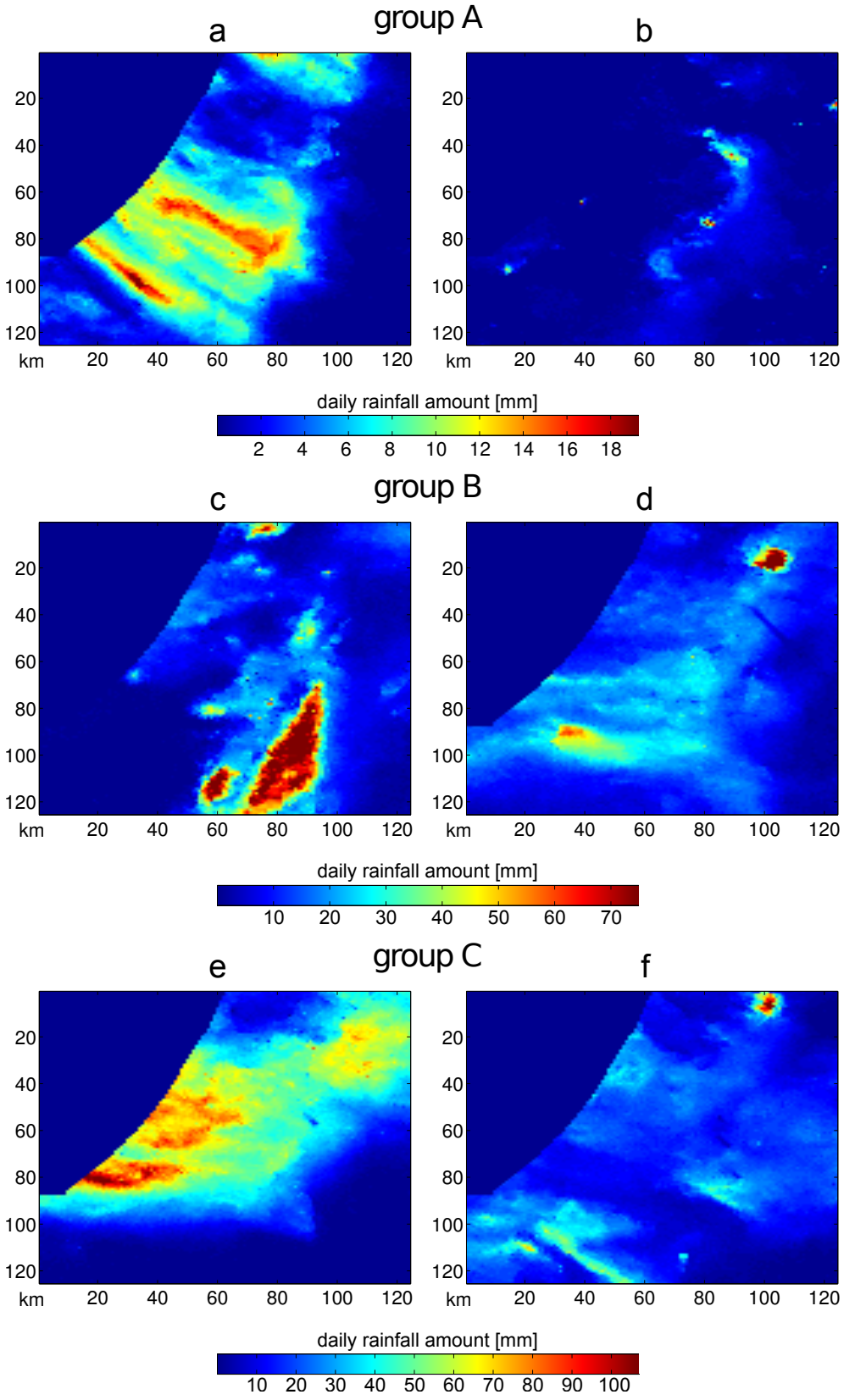


Figure 6.3: Examples of simulated rainfall fields for groups A, B and C.

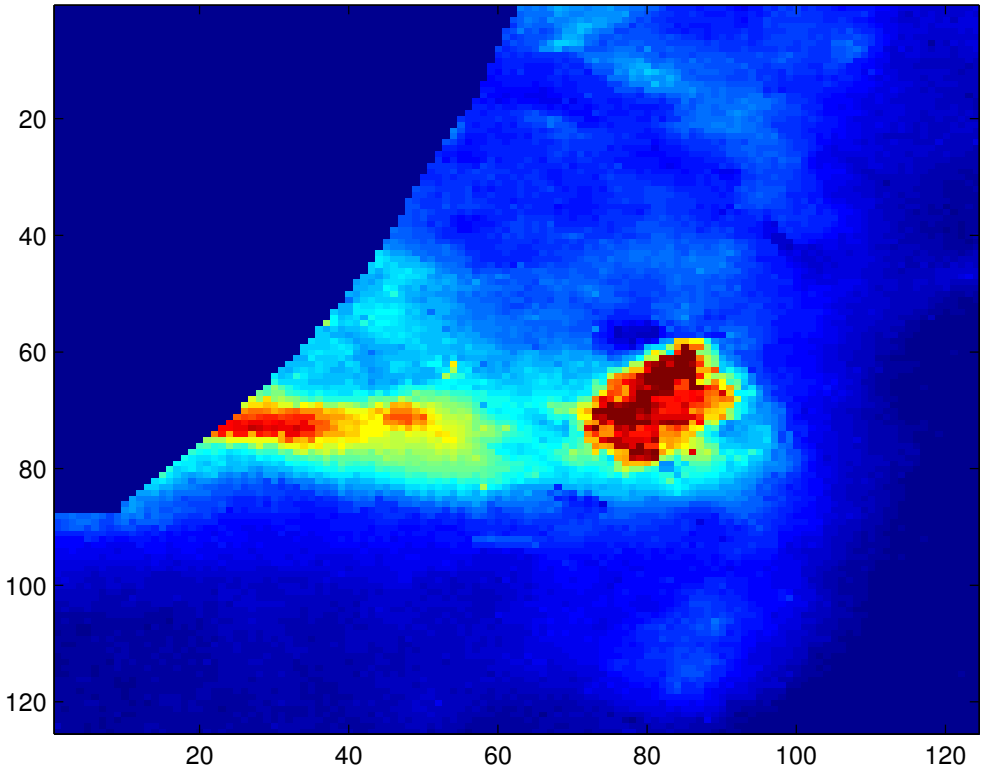


Figure 6.4: An example of simulated rainfall field from group B, presenting an unrealistic heterogeneity. In this case, the patterns of two different types of rainfall may have been sampled.

rainfall events according to the different rainfall synoptic types.

Group C presents the most intense rainfall fields observed in the region. These events takes form of a large front coming from the sea. In this case, the simulated structures are quite similar to the ones shown in the reference. Finally, for all groups, the small-scale noise observed in the simulation is of very small entity and we do not observe the radial artifacts related to the radar measurement, which, on the contrary, are present in the reference images.

Figure 6.5 shows the qq-plot comparing the rainfall amount probability distribution of the reference and simulated fields. The marginal rainfall distribution is preserved quite accurately for all groups, taking into account that the highest extremes shown in the graphs are representative of a few pixels in the whole training data set. A more significant negative bias is present for values from 30 to 45 mm in group A. In this case, a larger training data set may be required to enhance the representation of extremes.

The spatial variability of the fields is analyzed using the experimental variogram (figure 6.6). For all groups, the simulated fields approach the covariance structure of the reference with a modest tendency to underestimate the variability for large lags (50-70 km). This happens systematically for all cases but it is accentuated for group A and C, that also present a more pronounced underestimation of the extremes. The relation between the two issues is discussed later in this section. Conversely, the maximum variogram value is reached at the same distance lag (variogram range) of the reference, suggesting that the correlation length of the simulated rainfall heterogeneity is the same as the one observed in the reference.

The morphology of the rainfall fields is analyzed by comparing the empirical frequency

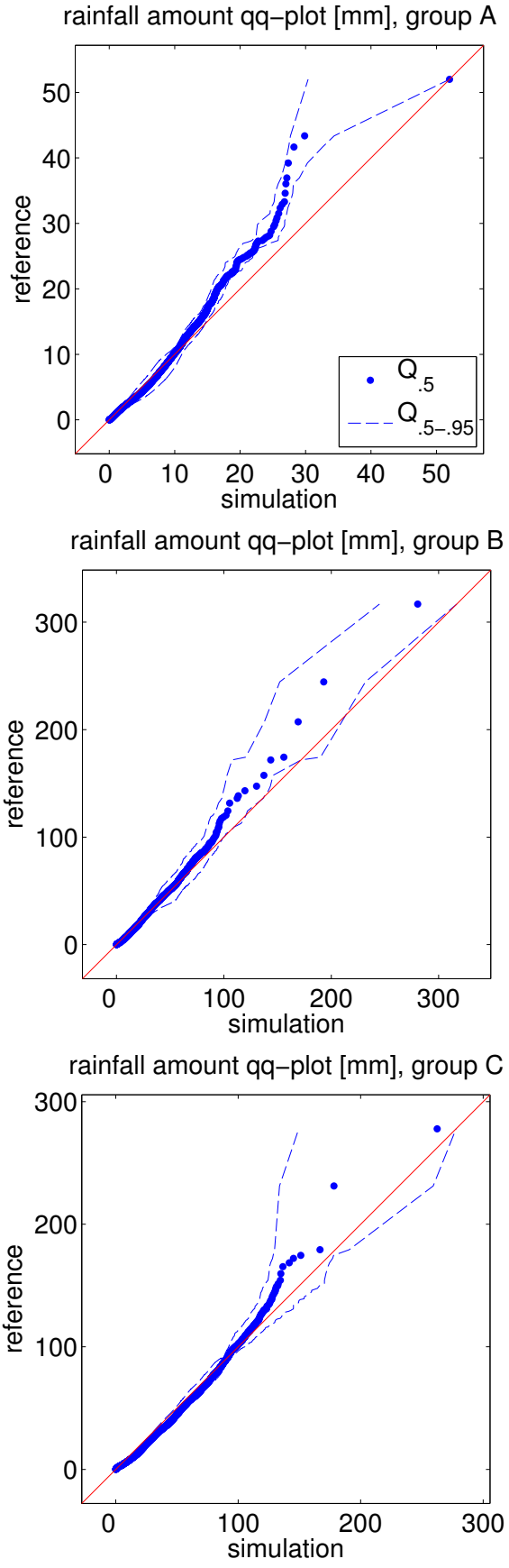


Figure 6.5: QQ-plot comparing the probability distribution of the reference and simulated rainfall fields for all groups.

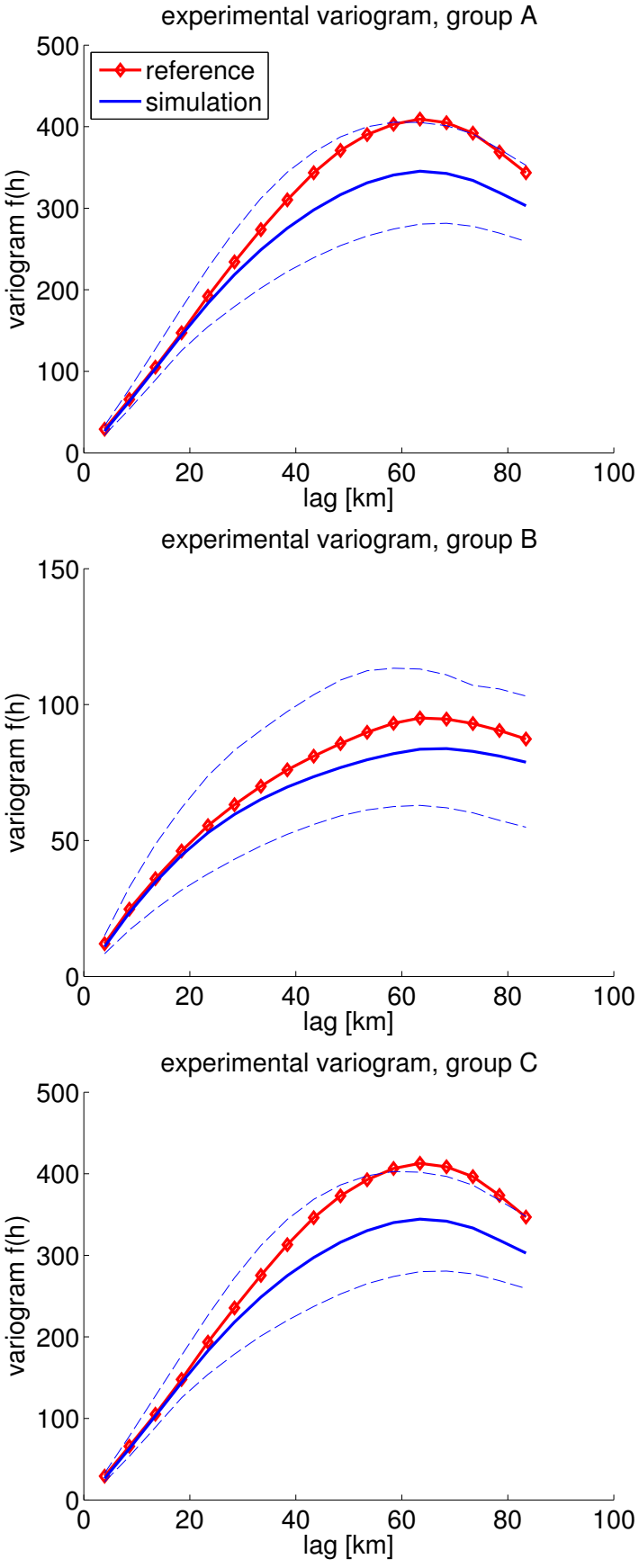


Figure 6.6: Experimental variogram of the reference and simulated rainfall fields for all groups. Continuous line represents the median and dashed lines the .05 and .95 quantiles of the simulation ensemble.

distribution of the rainfall cell area (figure 6.7). In general, the frequency distribution is preserved fairly well for all groups, but a modest over-representation of the small rain cells (1 to 4 km^2 equivalent to 1 to 4 pixels) is observed. The phenomenon is probably linked to the small-scale noise created in the simulation. This aspect is examined further in this section.

The joint probability density function of the rainfall amount and the elevation is shown in figure 6.8. The rainfall amount is linked to the elevation profile through a complex relation, that is also strongly affected by the distance from the sea. For the three groups, the JPDF is preserved fairly well: the sharp transition to low rain values associated to the depression in the eastern part of the study area (negative altitudes) is represented in the simulations, as well as the extremal behavior associated with the mountainous region. Nevertheless, the underestimated rainfall extremes observed in figure 6.5 are associated to near-zero altitudes: in particular, group A clearly show that the missing part of the distribution tail (rainfall values above 26 mm) is mainly associated to this altitude range.

This result bring us to consider that the extremes and variability underestimation is primarily related to a poor simulation of the heterogeneity in the coastal zone. Indeed, figures 6.2 a, e and f show that the extended rainfall events coming from the sea present the highest rainfall values along the coastline, just beside a large missing data zone (top left corner of each field). Moreover, we recall from chapter 2 section 2.3 that, in the TI scanning procedure, the DS algorithm rejects any pattern containing missing data. As a consequence, the large part of data patterns along the coastline are rejected since they contain missing data and the heterogeneity is simulated by sampling patterns in the nearby regions, that do not contain the same range of rainfall values. For this reason, we observe a consequent underrepresentation of the extremes and underestimation of the variability in the simulations. This particular case, where a massive presence of missing data influences the preservation of the statistics, illustrates the importance of using a complete data set, especially when dealing with non-stationary heterogeneity.

The joint probability density function of the mean rainfall amount and the rainfall cell area is shown in figure 6.9. As explained in section 6.3, different threshold values are used to censor the rainfall field. This allows identifying rainfall cells of different mean rainfall amount, describing the internal structure of the wet areas. This complex relation is fairly well preserved in the simulation: the overall shape and the range of values are covered with a modest underrepresentation of the mean-to-high area values, associated to low and medium rainfall values. Moreover, it is shown that the over-representation of the small rainfall cells detected in figure 6.7 is related to low rainfall values (bottom-left corner of the JPDF graph in figure 6.9), suggesting that this statistical bias has a minor quantitative impact on the spatial rainfall estimation.

Finally, the verbatim copy table 6.2, shows that the verbatim copy minimally affects the simulation. The total patch constitutes only a little fraction of the total generated ensemble and the maximum generated patch is of very small size.

Table 6.2: Table showing the total verbatim copy percentage and the size of the maximum patch found in all the simulation ensemble.

Group	Total Patch [%]	Max patch size [%]
A	3.96	0.65
B	3.23	0.52
C	5.33	1.92

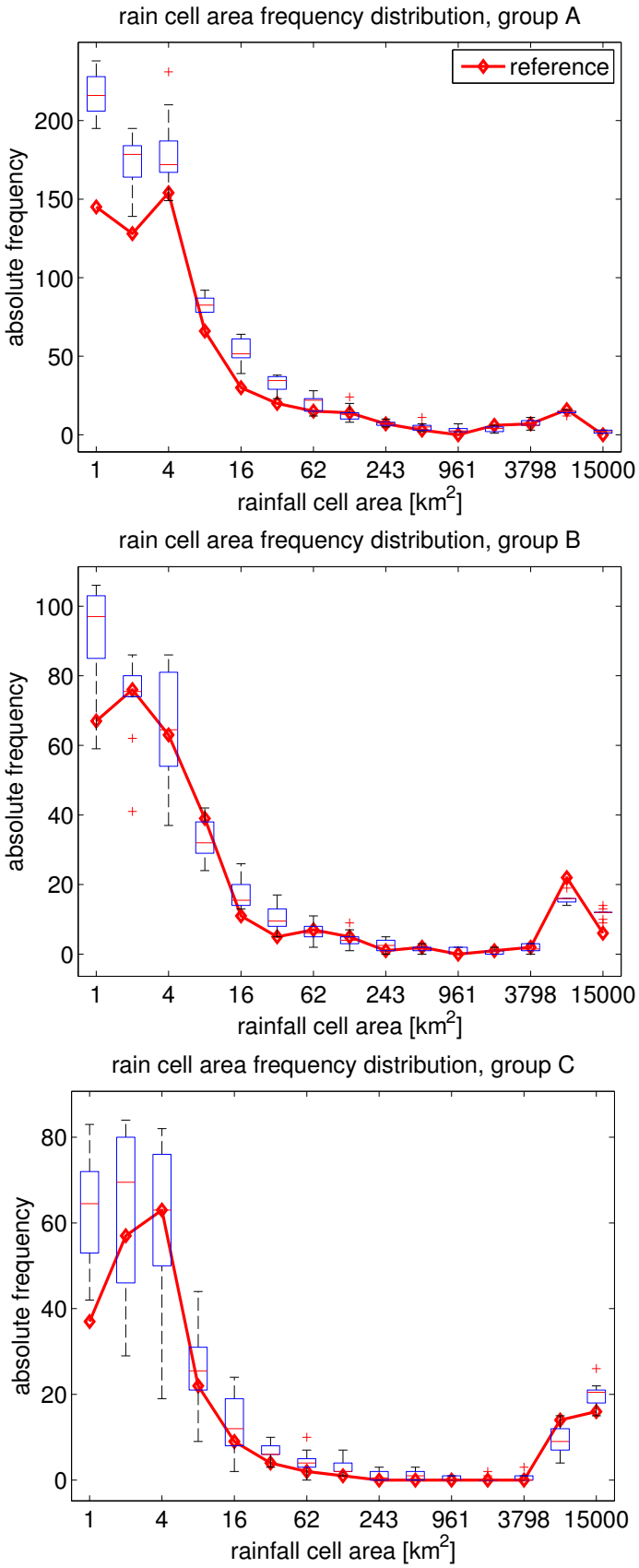


Figure 6.7: Plot of the rainfall cell area empirical frequency distribution: the red line indicates the reference and the boxplots represent the simulation ensemble.

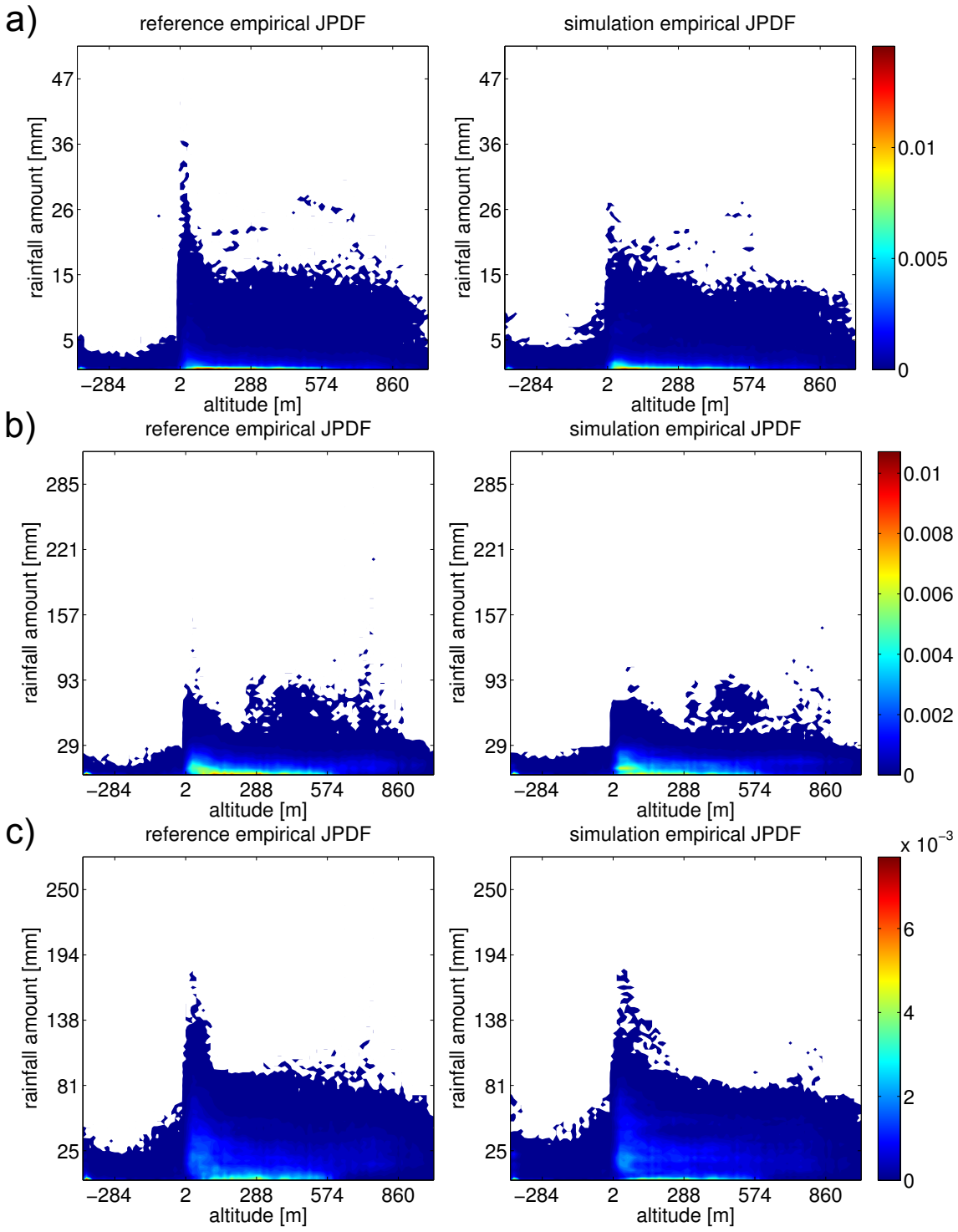


Figure 6.8: Empirical joint probability density function (JPDF) of rainfall amount and elevation. Comparison between the reference (left) and one randomly chosen simulated field (right) for groups A (a), B (b) and C (c).

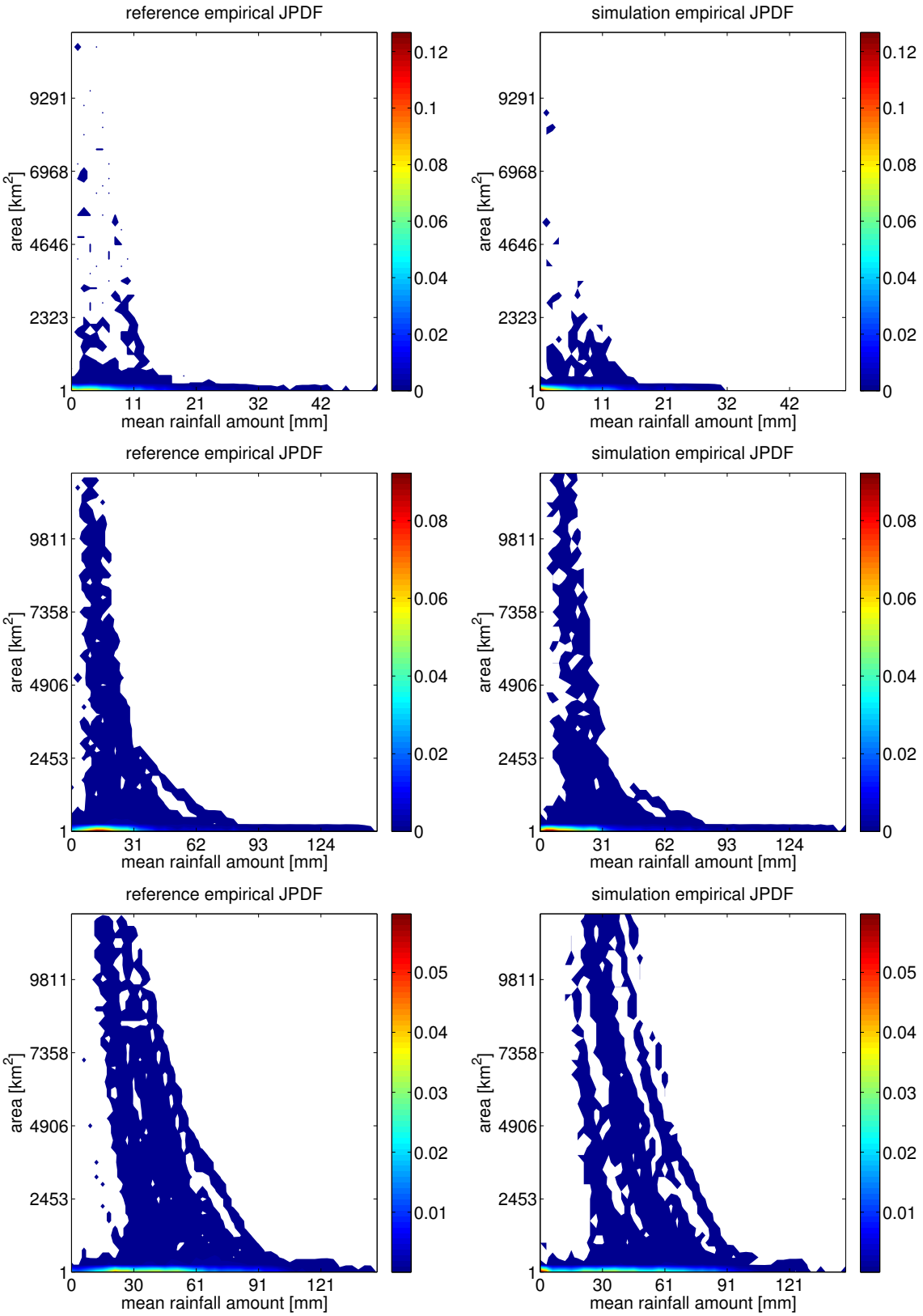


Figure 6.9: Empirical joint probability density function (JPDF) of the mean rainfall amount and area of each rainfall cell. Comparison between the reference (left) and one randomly chosen simulated field (right) for groups A (a), B (b) and C (c).

6.6 Conclusions

In this chapter, a DS setup for daily rainfall radar images is presented. DS simulates the rainfall field by scanning a catalog of historical radar images used as training data set (TI): when a datum presenting a similar pattern of neighbor data is found in the TI, the datum is sampled and assigned to the simulated image. By considering patterns of different size throughout the course of the simulation, DS is capable of generating realistic structures at multiple scales.

In the setup proposed for rainfall radar images, the daily rainfall amount is simulated together with two auxiliary variables: i) a categorical variable indicating the dry/wet pattern and the regions occupied by extreme rainfall and ii) the topographic elevation. The latter variable is given as conditioning data to allow respecting the influence of altitude on the rainfall heterogeneity.

The methodology, tested on the simulation of three groups of images presenting different ranges of rainfall amount, generates rainfall fields similar to the training data, presenting the same type of anisotropic structure and correlation length. The marginal probability distribution of the rainfall amount and its covariance structure are approached with a modest underrepresentation of the spatial variability and extreme values. It has been observed that this problem is mainly related to the poor simulation of the patterns along the coastline, that hosts the highest rainfall values near a large missing data region. The proximity to missing data prevents the algorithm to sample the patterns, with a resulting underestimation of the extremes. A better preservation of the extremes and variability may be possible by using a more complete training data set. Nevertheless, the complex relation between elevation and rainfall amount is accurately represented by the simulation. This also helps respecting the influence of the distance from the shore line. Finally, the technique can preserve quite adequately the probability distribution of the rain cell area and its complex relation with the rain cell mean rainfall amount.

The strong point of the technique is that it can generate realistic rainfall fields conditioned by elevation, with minimal complexity in the parametrization. Since the algorithm mainly relies on the quality of the training data set, it is important to use as TI a historical record representative of the specific type of rainfall event that one wants to simulate. As shown in the results, if the TI contains different types of rainfall events, the algorithm cannot distinguish this non-stationarity in time and it may simulate unrealistic heterogeneity. One main requirement is therefore an accurate classification of the historical radar record to obtain one TI representative of each different rainfall type. This phase may require the help of an expert.

In future developments, the aim is to include the technique in a more complete simulation framework to generate long time rainfall records: for example, a daily time-series of rainfall event types can be simulated (as already done in [149]). Then, for each day, the correspondent rainfall field is generated with the proposed DS setup and using the training data set related to that type rainfall event. This two-step strategy aims at simplifying the resampling procedure, since the algorithm has direct access to the appropriate training data instead of scanning the whole available catalog of rainfall events. Moreover, the simulation setup can easily accommodate large-scale synoptic variables as satellite rainfall observations or climate model output images. These variables can condition the simulation to impose the large-scale spatial distribution of rainfall or a climate change scenario. There are two possible ways to use this type of information with DS. The first one is to use these images as auxiliary variables, like it is done in this chapter with elevation. The second one is to use the block conditioning feature, available in the latest DS implementation [194]. Block conditioning allows imposing a specific local mean value, defined for different regions of the

simulated field. For example, a coarse-gridded climate model output image, giving the mean estimated rainfall on a large-scale grid, can be used to impose the local mean over all the field. In this case, the information is used to constrain the small-scale simulation to preserve the mean value given by the variable. Conversely, if the same information is used as auxiliary variable, it associates small-scale rainfall values between the simulation grid and the training data set without imposing the local mean. Block conditioning is a reliable tool when a precise large-scale information is available.

In conclusion, if a representative and complete training data set is available, the presented technique constitutes a fairly efficient and handy tool to simulate realistic spatial rainfall fields based on radar imagery. Further testing on different climates and event types is necessary, with the possibility to include the method in a more complete spatio-temporal simulation framework.

Chapter 7

Small-scale spatial rainfall simulation using spatialized time-series: a preliminary test

Abstract

This chapter shows the early development of a methodology based on the Direct Sampling multiple-point statistical method for the simulation of high-resolution spatial rainfall using ground measurements and satellite data. The Direct Sampling technique for time-series is integrated in a more complex simulation framework based on: i) remote sensing imagery as auxiliary variable to impose the morphology and movement of large-scale rainfall patterns, ii) high-resolution ground measurements contained in time-series to simulate the small-scale heterogeneity instead of using only the data present in space. In order to achieve a training data set representative of the small-scale spatial heterogeneity, the high-resolution rainfall time-series are projected into space by changing the data spacing as a function of the rainfall cell velocity. The preliminary results on a synthetic rainfall field show the efficiency of the methodology in reproducing the small-scale heterogeneity of rainfall in consistency with the large-scale picture given by satellite images. Being based on the accuracy of ground measurements and on freely available satellite images, this approach shows a strong potential at a low cost but still needs to be adapted to the complexity of real cases showing anisotropic and non-stationary spatial distributions.

7.1 Introduction

In the last decades, the availability of high-resolution remote sensing imagery completely changed the perspective in modeling the dynamics of the Earth and other planets. Recent satellite missions by NASA (American National Aeronautics and Space Administration), JAXA (Japan Aerospace Exploration Agency) and ESA (European Space Angency) produced a wealth of data that revealed the spatio-temporal complexity of climatic and hydrological processes all over the Earth. At the same time, high-resolution radar data allowed a fine-scale quantification of these processes in certain regions. This rich information at multiple scales opened a broad horizon for hydrological modeling. Since the small-scale interaction between climate and basin characterization (land coverage, soil moisture, albedo) can lead to an extremely variable hydrological response [216, 219, 50, 11, 197, 170, 71], a faithful representation of spatio-temporal patterns at the sub-catchment scale in consistency with larger scales is needed. The estimation of rainfall amount using satellite and radar instrumentation, while efficiently detecting the arrival of big events and their morphology, is far from the level of small-scale accuracy necessary for medium- and long-term water balance predictions [71, 99]. In particular, radar images present multiple sources of error, which, in the most advanced approaches, are treated separately (see e.g. [71]): i) a systematic bias related to the empirical law used to estimate the variable of interest using the remote sensing measurement – for example, rainfall amount can be estimated from reflectivity, while electromagnetic radiation in the thermal infrared band is used to estimate land surface temperature – these empirical relations vary in function of the local climatic conditions, showing a variable error structure that should require continuous recalibration; ii) a range-dependent error in relation to the distance between the radar station and the measured location; and iii) a spatially structured random error correlated to different variables like elevation. In addition, regions with a complex topography often present physical obstacles to the radar investigation, with subsequent artifacts and missing data regions that requires the installation of multiple radar stations in order to cover the entire region of interest. Nevertheless, radar constitutes nowadays the best source of information about the small-scale spatial aggregation of rainfall, providing in some cases a fairly accurate rainfall estimation when combined to a dense ground measurement network. The correction and uncertainty quantification on remote sensing products is generally based on stochastic techniques used to interpolate the error

structure (data merging) or simulate an ensemble of realization of the field. As explained in chapter 1, the most common spatial rainfall simulation and data merging techniques belongs to the geostatistics family [62, 57, 148, 105, 83, 106, 200, 206, 113]. The strong point of geostatistical techniques is that they offer a well established mathematical framework to quantify the local uncertainty and to condition the simulation upon auxiliary variables such as elevation, wind or larger-scale rainfall estimation (satellite observations or global circulation model outputs). Nevertheless, even using variable transformation to approach the very asymmetric probability distribution of rainfall, the underlying multiGaussian assumption in the simulation phase limits these techniques to the second-order moments describing the heterogeneity. As a consequence, the realism of the results produced by these methods is limited. Moreover, their uncertainty quantification is generally dependent on the density of available ground measurements (see e.g. [29]). A recent step forward in this direction is represented by the geostatistical dry drift model ([166, 167]). This technique, calibrated on radar images, impose a deterministic trend on the rainfall amount, increasing with the distance from the rainfall cell boundary. The simple approach of the dry drift can generate small-scale rainfall fields very similar to the radar images in terms of visual comparison, mean and variance, but for now its evaluation is limited to few simulation cases, analyzed using low-order moments, that are parameters of the algorithm itself. Some other recent techniques relying on the analysis of radar images are the object-based rainfall fields generators [146, 149]. These statistical algorithms take into account the probability distribution of different spatio-temporal characteristics of the rainfall cells: area, shape, density, movement and life time in a composite simulation framework calibrated using radar observations. The advantage of this approach is the capability of generating very heterogeneous rainfall fields at a high temporal resolution and the possibility of conditioning the simulation by imposing the synoptic rainfall type, which can be derived by large-scale observations or global circulation models. Its main shortcomings are the limitation to convective rainfall types, a certain tendency to underestimate the extremes and a complex parameterization. One last family of techniques are based on the hypothesis of the scale-invariant character of rainfall: the multi-fractal or multiplicative cascade techniques [64, 129, 130, 43, 147]. One of the latest algorithm of this type [161] is based on ground measurements only and uses a power law to estimate the probability distribution moments at scales lower than the density of the observation network. By tuning a user-defined parameter of the scaling law, the method can generate very fine-scale simulations that preserve the variogram estimated on real data. This approach is convenient when no other auxiliary information is available apart from ground measurements, but it needs calibration on a sufficiently informed rain gauge network.

The Direct Sampling simulation framework for time-series presented in chapter 2 demonstrates its efficiency in simulating rainfall in consistency with multiple scales. The objective of this chapter is to extend this simulation approach to the small-scale spatial rainfall simulation. The main problem in this case is that ground measurements (from climate stations) are sparse in space and cannot provide a continuous spatial training data set. Nevertheless, the rainfall stations record the transit of rainfall cells and their internal spatio-temporal variability. The main idea of the proposed technique is to project the time-series into space, by knowing the mean velocity of rainfall cell displacement. This way, in case of moderate displacement velocity and high temporal resolution, the projected "space-series" can represent an approximate example of the small-scale variability in space. This can be used as training data set with the Direct Sampling technique.

The hypothesis that rainfall in time exhibits the same covariance structure in space has been verified for limited amount of time (40-60 minutes) with empirical observations by Zawadzki [226] and in a more complex statistical framework by Waymire et al. [211]. This hypothesis is relaxed here since the spatio-temporal rainfall field does not have to be

stationary for long time periods or large areas. In fact, satellite imagery is used as auxiliary variable to impose the non-stationarity related to the morphology and movement of large-scale rainfall patterns and to estimate the mean displacement velocity. Conversely, high-resolution ground measurements contained in time-series will be used to simulate the small-scale heterogeneity, conditioned by the satellite information. The methodology is tested with a preliminary experiment involving a synthetic rainfall field. The chapter is organized as follows: in section 7.2 the methodology is presented with its underlying hypotheses, section 7.5 shows the synthetic reference model used for the preliminary test and the statistical indicators used to analyze the simulations, in section 7.6 the results are analyzed and section 7.7 is dedicated to the conclusions and future developments.

7.2 Methods and hypotheses

The proposed simulation approach allows simulating the small-scale spatio-temporal rainfall using satellite rainfall images to impose the morphology and movement of large-scale structures, while ground measurements contained in the time-series are used to simulate the small-scale heterogeneity. The main methodology workflow is described in the following.

1. **Preliminary data analysis**. Satellite images are analyzed to estimate the mean displacement velocity of rain cells. The estimation can be achieved with several techniques. One simple way to keep track of rainfall cell displacement is to compute the cross-correlation between consequent images: the cross-correlation matrix describes the correlation of two images superimposed and shifted. Each shift in one direction corresponds to a correlation value inside the matrix. The position of the maximum value inside this matrix determines the estimated displacement magnitude and direction. This value (defined in [0,1]) also informs about the reliability of the estimation: if it is too low, it means that the displacement of rain cells cannot be clearly detected. For example, this may happen if the rainfalls cells show a considerable modification of their shape with the displacement. In that case, further analysis using a more sophisticated movement tracking technique (see e.g. [108]) may be used or that data set portion may be discarded. Once a reliable displacement magnitude estimation has been achieved, it suffices to divide it by the time span between the two considered satellite images to have an estimation of the mean displacement velocity.

2. **Time-series transformation** (figure 7.1 b). Rainfall time-series are transformed in function of the estimated velocity to obtain one-dimensional spatial series representing the small-scale spatial variation in the direction of the cell displacement. This transformation is obtained with one simple operation. Let be the original time-series $Z(t)$ a function of time t and v the mean velocity of rain cell displacement. Assuming uniform motion for the amount of time intercurrent between two subsequent satellite images, the space covered by the rainfall field displacement is $\Delta x = v\Delta t$. The transformed spatial series is defined as $Z_s(x) = Z(v_x t)$. As a consequence, the spatial resolution of $Z_s(x)$ is directly proportional to v_x and the resolution of $Z(t)$. For example, with a displacement velocity $v_x = 18$ km/h and a time-series resolution $\Delta t = 1$ min, we obtain a spatial series resolution $\Delta x = 300$ m. The spatial resolution obtained can change over time as a function of v_x and can be estimated separately for each time-series portion between the subsequent satellite images. $Z_s(x)$ is linked to a time-series of satellite measurements, projected into space with the same transformation.

3. **Sequential simulation** (figure 7.1 c). The data set composed by the transformed time-series together with the corresponding satellite data are used as training data set

to simulate a spatio-temporal field. The methodology, inspired by [35], fills up a 2D spatial field with a sequence of one-dimensional simulations realized using the Direct Sampling (DS) technique. Each simulated 1D spatial series covers the entire field length in one direction and crosses the previously simulated series. At each crossing point, the already simulated datum is retained to condition the ongoing simulation. This way, the data simulated in series will progressively form a coherent structure in 2D until the whole field is completed. At the same time, a given satellite image is used as conditioning data to impose the large-scale structure of the simulation. In the simulation scheme shown in figure 7.1, the 1D spatial series are simulated in a random order along two perpendicular directions. For simplicity, this is the scheme used until now but others are possible.

It is important to note that, in the simulation stage, satellite data are just used as markers to associate small-scale values between the training data set and the simulation grid, for example to distinguish high or low rainfall regions and their morphology. This means that the absolute rainfall values estimated by the satellite do not necessarily correspond to the local mean value of the simulated small-scale heterogeneity. Therefore, any possible systematic bias of the satellite data does not affect the simulation since it is observed in both the training data set and the conditioning data. Conversely, a more complicated or non-stationary error structure may affect the quality of the simulation.

There are several hypotheses on which the methodology at this stage of development is based. Let us define the spatio-temporal rainfall as a space referenced stochastic variables $Z(x)$ where x is the spatial coordinate. We keep for simplicity of notation a 1D coordinate, but the principle can be applied to a spatio-temporal space specified by a vector $\vec{x}$. The corresponding conditioning satellite variable is $S(x)$. The joint rainfall probability distribution of the rainfall pattern composed by multiple neighbors and conditional to a certain satellite value is supposed to be strongly stationary, i.e. $f(\vec{Z}(x,\vec{l})|S(x)) = f(\vec{Z}(x+h_x,\vec{l})|S(x+h_x))$, where $\vec{Z}(x,\vec{l})$ is a variable vector representing any pattern allowed by the DS parameters (see section 7.3) and defined by a lag vector $\vec{l} = l_1, l_2, \ldots, l_n$ such that $\vec{Z}(x,\vec{l}) = \{Z(x+l_1), Z(x+l_2), \ldots, Z(x+l_n)\}$. Moreover, $f(\cdot)$ is the probability density function and h_x is any space lag. Therefore its covariance $C(h_x, S(x))$ is only function of the space/time lags and the synoptic conditions imposed by the satellite image. The use of $S(x)$ as conditioning variable links the small-scale variability to the large-scale patterns, therefore it should allow respecting the non-stationarity due to different types of rainfall events.

Finally, the spatio-temporal field $\mathbf{Z}$ defined in (x,t) is supposed to be isotropic in space but anisotropic in space-time: in fact, the conditional probability density function of the rainfall pattern in space $f(\vec{Z}(x,\vec{l})|S(x))$ is approximated to $f(\vec{Z}(vt,v\vec{l})|S(vt_i))$, where v is the mean displacement velocity of rainfall cells. In other words, the complex covariance structure in space is estimated by applying a linear transformation to the one in time, which is a direct result of the projection of time-series in space, shown above in this section.

7.3 The Direct Sampling setup for sequential 2D simulation

The sequential simulation phase is performed with the following DS setup. DS operates in the same way defined in chapter 2 for time-series rainfall simulation: the main difference is that here the simulation grid (SG) is a 2D field, while the training data set (TI) remains a 1D vector. Similarly to time-series simulation, the simulation setup comprises the following variables (table 7.1): 1) the satellite image variable (S) given at the same resolution as the target variable – S is a spatio-temporal variable associated to the time-series used as TI and given as conditioning data in the simulation grid; 2) the dry/wet sequence (W), i.e. a

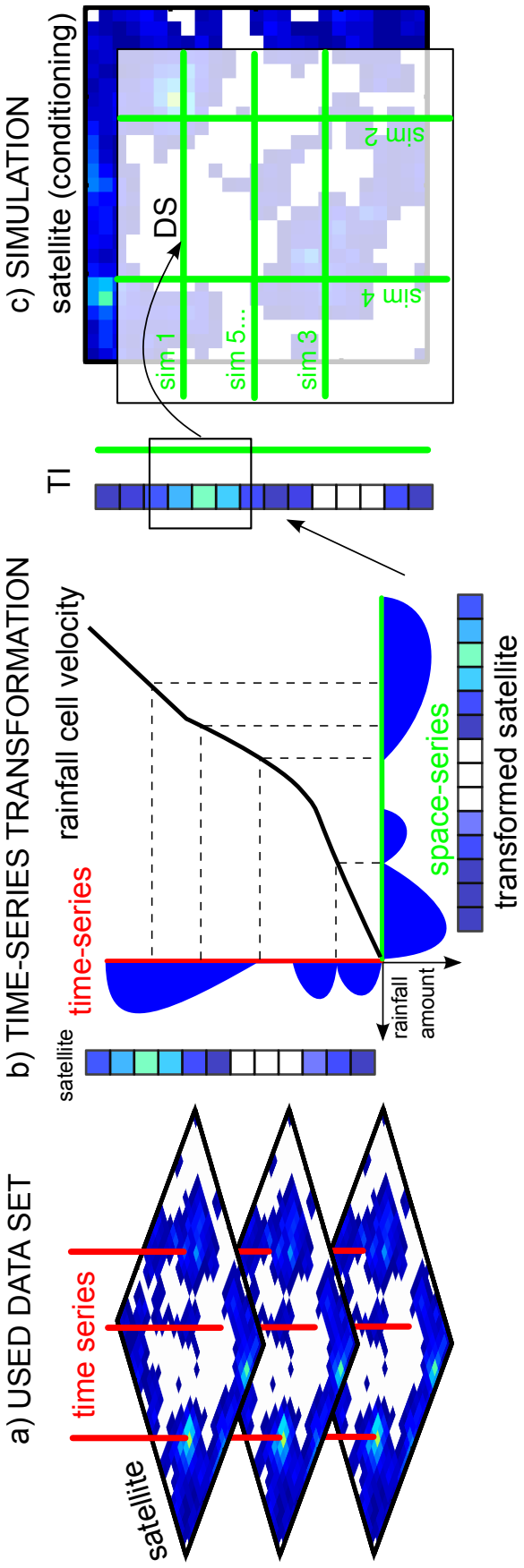


Figure 7.1: Sketch of the proposed sequential simulation technique: a) data set used, b) time-series transformation and c) sequential simulation using DS.

categorical variable indicating the position of a datum inside the rainfall pattern (1 = wet, 0 = dry, 2 = solitary wet, 3 = wet value at the beginning or at the end of a wet spell); 3) the rainfall amount (Z) constituting the target variable. W is automatically computed inside the TI and co-simulated with Z. The iterative simulation technique, represented in figure 7.1 c, involves the co-simulation of W and Z as time-series filling up the 2D space and conditioned by the satellite image (S variable) given. It is worth noting that, to bring S to the same spatio-temporal resolution of the other variables, the nearest neighborhood interpolation technique is applied. This interpolation, applied on the regular small-scale grid, does not change the structure or the content of S. The main parameters of the algorithm,

Table 7.1: Standard setup used for spatial small-scale rainfall simulation. The parameters are: search window radius R, maximum number of conditioning neighbors N and distance threshold T. The variables are: 1) The satellite image (S), 2) the dry/wet sequence (W), the rainfall amount (Z).

Variable	R	N	T
1) S	300	1	0.002
2) W	300	10	0.002
3) Z	300	10	0.005

defined for each variable, are the search radius R, maximum number of considered neighbors N and distance threshold T (see chapter 2 for a detailed description). The values used in this setup are shown in table 7.1: R limits the extension of the conditioning pattern and should be set equal or higher to the correlation length of the structures observed inside the training data set. Here R =300 approximates the size of the simulated field, meaning that from the first iterations the simulation will be conditioned by the already simulated nodes. N, the maximum number of conditioning neighbors, controls the detail of the conditioning pattern: in case of S, it is sufficient to set N=1 since the variable does not show small-scale variations and it is completely informed, the datum considered for conditioning is at the same location of the simulated datum. For W and Z, N is tuned by trial and error. In general, a value between 5 and 20 is sufficient to describe the rainfall patterns. The distance threshold is set about one order of magnitude lower than the standard value suggested for daily rainfall time-series (see chapter 2), lower values are preferred here since the signal at this temporal resolution presents higher autocorrelation. Therefore its structure requires a more strict choice of the sampled values in order to be correctly simulated. One last parameter of the simulation is the maximum scanned TI fraction F, that is set to 0.5, meaning that, at each scan iteration, the algorithm covers at maximum 50% of the TI visiting random locations. The partial scan of the TI avoids the risk to consequently sample multiple times the same patterns.

7.4 Reference model

The proposed methodology is tested on a synthetic case where the spatio-temporal field meets the hypothesis presented in section 7.2. The reference model (shown in figure 7.2), generated in the fashion of the censored Gaussian model [5, 201, 98, 3], is obtained with the following steps:

1. Using the Fast Fourier Transform algorithm (FFTW implementation [53]), a multi-Gaussian field $\mathbf{G}$ with standard normal distribution and exponential covariance is generated in the space (x', y', t'). The field generated is of size 120 km $\times$ 120 km $\times$ 7 days and resolution $res_x \times res_y \times res_t$ = 300 m $\times$ 300 m $\times$ 1 minute. The correlation length

is $l_{x'} = l_{y'} = 60$ km, while in time it is:

$$l_{t'} = \frac{l_{x'}}{res_x}\sqrt{2} \text{ minutes} \tag{7.1}$$

The value of $l_{t'}$ is chosen to approach the same correlation length in space and time after the deformation introduced in step 6.

2. To generate zero-rainfall regions, the field is censored with the following rule:

$$G^T(x', y', t') = \begin{cases} dry & \text{if } G(x', y', t') < q_p \\ G(x', y', t') & \text{otherwise} \end{cases} \tag{7.2}$$

where dry is the censoring flag, $q_p = F_G^{-1}(p)$, $F_G^{-1}(\cdot)$ is the inverse cumulative probability density function (inverse CPDF) of G and p is the proportion of dry values inside a considered historical rainfall data set. The data set used here is the 2005-2006 10-minute rainfall record of 40 climate stations in the temperate central region of Switzerland (Meteoswiss database). Since 1-minute time-series for this data set are not available, the 10-minute time-series are transformed to one-minute values by simply dividing them by 10. This method underestimates the one-minute rainfall extremes, but this effect is compensated by the procedure shown in step 4.

3. The CPDF $F_{G^T}(\cdot)$ is computed for G^T.

4. To approach a realistic 1-minute rainfall distribution, a piece-wise Pareto tail model is fitted on the wet-day values of the time-series used in step 2. The piece-wise Pareto tail model is initially composed of a non-parametric distribution obtained using a kernel smoothing technique. The obtained distribution is considered up to the 90-th percentile and a Pareto tail distribution model is fitted on the rest of the support to generate the tail. The obtained heavy-tailed distribution h is suitable to represent the occurrence of rare extreme events. Its CPDF $F_K(\cdot)$ is computed.

5. A distribution transformation is applied to the non-censored part of G^T using the previously computed CPDFs: $H^T = F_K^{-1}(F_{G^T}(G^T))$, while zero is assigned to the censored values (for $G^T = dry$, $H^T = 0$) to generate the dry regions.

6. Finally, $H^T(x', y', t')$ is deformed by the application $H^T(x', y', t') \rightarrow H^T(x, y, t)$, where $x = x' + v_x t$, to mimic a rain cell displacement along x with constant velocity $v_x =$ 300m/min (=18km/h), which corresponds to the model resolution in space and time. As a consequence, the correlation length $l_{t'}$ is subjected to the deformation $l_{t'} \rightarrow l_t$, with $l_t = 0.82 l_x$. This parametrization facilitates the time-series projection as explained in section 7.5. For more details about the applied deformation see appendix E.

As synthetic variable mimicking a large-scale remote sensing rainfall estimation, the S variable is used: S is computed as the spatio-temporal mean over large-scale blocks of size 4.8 km × 4.8 km × 16 minutes.

7.5 Evaluation

The methodology proposed in section 7.2 is only partially tested here since the following simplifications of the problem are adopted. The velocity is supposed to be constant and known a priori. This eliminates the need of analyzing the S images in order to track the rainfall cell movement. Moreover, since the field movement imposed in the reference model

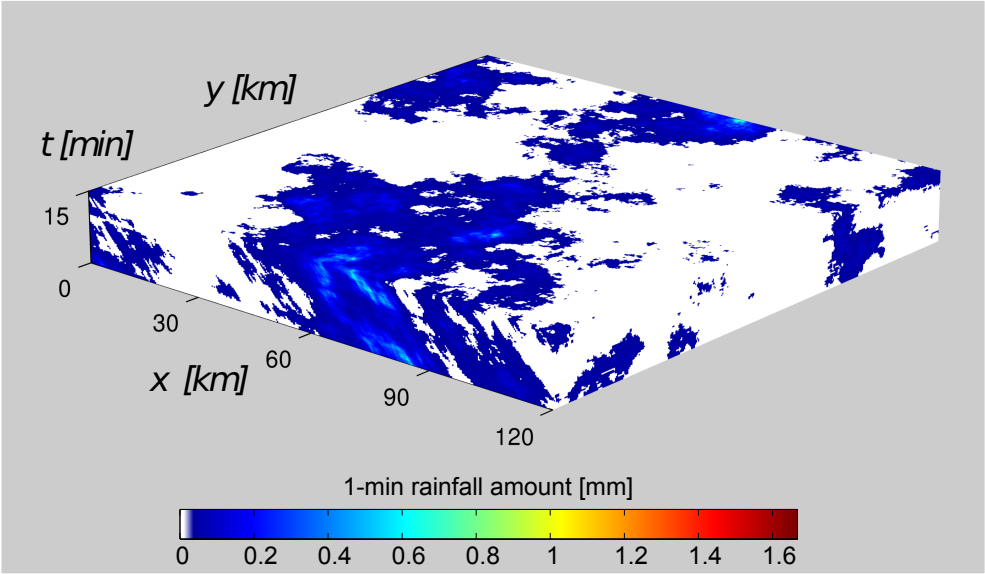


Figure 7.2: Portion of the synthetic reference model $\mathbf{H}^T$, of size 120 km × 120 km × 15 minutes.

is unitary with respect to the spatio-temporal resolution ($v_x = 1$), the time-series projection in space $S(x) = Z(tv_x)$ does not alter the original time-series spacing. This eliminates any need of interpolation to adapt the projected time-series to the spatial resolution desired for the simulation. This allows focusing on the performance of the methodology in the simulation phase, which is the aim of this preliminary test. Possible complications regarding the technique in real applications are discussed in section 7.7.

The sequential simulation technique is tested on the simulation of a randomly chosen spatial 1-minute Z field extracted from the field $\mathbf{H}^T$ and used as reference. Conversely, the time-series at 10 random locations are extracted to be used as training data set. According to the generated field, the time-series are 7-day long and present a temporal resolution of 1 minute. It should be noted that, since the model is stationary and the simulation technique does not rely on the data available in space, the same statistical content would be obtained by generating, for example, a 70-day long record and considering one time-series only. 100 realizations are generated using the simulation technique described in section 7.2 with the DS setup shown in section 7.3. No small-scale Z datum is used for conditioning, only the S variable is given as conditioning variable over the whole field. The results are visually compared with the reference and analyzed by computing the qq-plot of the rainfall amount. The spatial variability of the field is analyzed using the experimental variogram function.

7.6 Results and discussion

Figure 7.3 shows the visual comparison between the reference and one simulated field together with the S image used to condition the simulation: the simulated heterogeneity looks fairly similar to the reference. The local extremes and variability looks to be efficiently preserved. This is a relevant result considering that it has been achieved without using any information about the prior statistical model and any 2D small-scale image. Nevertheless, two kind of artifacts in the generated field are visible: the blocky texture of the S image is visible in some areas of the simulated field and some noise is present on the whole simulation. The blocky texture artifact may be corrected by applying a moving average filter to the S image,

smoothing out its blocky structure. This operation, if applied to both the conditioning and training S data, may be a viable solution since, in this case, DS does not impose the absolute S values over large-scale blocks, but uses S only as synoptic variable to associate Z patterns between the TI and the simulation grid. Conversely, the small-scale noise is intrinsic to the DS technique: it can be reduced to some extent by lowering the acceptation threshold, with the risk of degrading the statistics of the simulated fields. In this case, we accept a modest small-scale noise to avoid over-conditioning the choice of patterns and preserve a more accurate representation of the marginal distribution. Whether these artifacts have a significant quantitative impact on real case predictions, for example the recharge estimation over a basin, is an aspect that still needs to be investigated, for example using the generated rainfall fields as input data for hydrologic models. The simulation mean over 100 realizations (figure 7.4, a), confirms that the simulated small-scale heterogeneity respects the large-scale structure imposed by the S variable (figure 7.3, a): the two images show a similar blocky texture and absolute values, which makes sense since, in this synthetic model, S is generated by computing the actual local mean of the reference field. If the simulated fields look very similar to each other (figure 7.3, c and d), conversely the small-scale variability of the simulation ensemble can be considerable in the zones of intense rainfall, as confirmed by the interquartile range (figure 7.4, b) with values of up to 0.1 mm.

The skewed rainfall amount probability distribution is efficiently preserved as shown by the qq-plot that compares the PDFs of the reference and the simulation ensemble (figure 7.5 a), showing only a small positive bias for the upper extremes. Since the simulation is conditioned by S, the uncertainty over the estimated PDF is very limited as shown by the 5-95-th percentile boundary of the simulation ensemble. The experimental variogram (figure 7.5, b) shows a reasonable match between the reference and the simulation, indicating that the covariance structure of the field is fairly well preserved at different scales.

7.7 Conclusions

Assessing the uncertainty about the small-scale heterogeneity of rainfall is a an important goal in hydrology, since the variability at this scale can have a significant impact on predictions about the distributed basin recharge and the impact of extreme events. The main problem in this case is the sparsity of ground measurements, which cannot provide a complete picture of the small-scale heterogeneity used to make an accurate statistical analysis. The method proposed in this chapter is an alternative approach to bypass this obstacle. The technique, based on the Direct Sampling algorithm (DS), allows the simulation of rainfall field with high spatio-temporal resolution by using the ground measurements contained in time-series and the large-scale information of satellite images. The main rationale is that a climate station at a fixed location and measuring at a high temporal resolution, can record the spatial structure of a rainfall cell moving across it. The high-resolution time-series is then projected into space by multiplying its spacing by the mean rainfall cell velocity: the obtained spatial series constitutes an approximate example of the rainfall variation in space. If the rainfall cell velocity is moderate, this data set can be much more informative than the sparse measurements available in space. The satellite imagery is used to estimate the mean rainfall cell velocity and to condition the simulation imposing the large scale structure, while DS uses the transformed time-series to simulate the small-scale variability.

The preliminary test shown in this chapter is focused on the simulation phase using the DS algorithm. The test consists in simulating a synthetic high-resolution (300 m) field showing a stationary and isotropic spatial rainfall distribution, exhibiting intermittency and a heavy-tailed probability distribution of the rainfall amount. The simulation is conditioned by a synthetic satellite image obtained by averaging the reference field over a large-scale

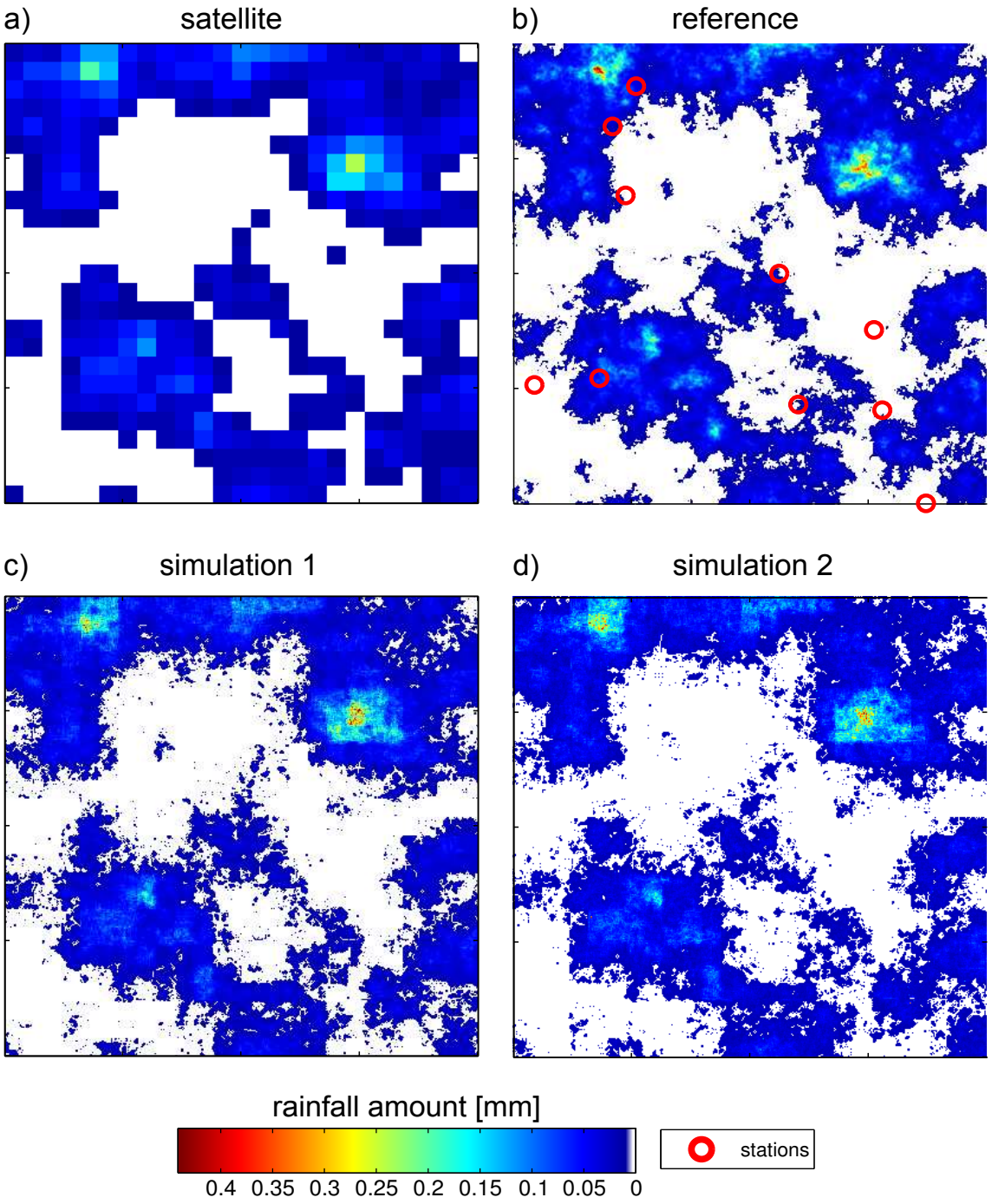


Figure 7.3: Visual comparison of the simulated and reference fields: a) the S variable field used for conditioning, b) the reference small-scale field with location of the time-series used as training data (red circles), c) and d) Two examples of simulated field.

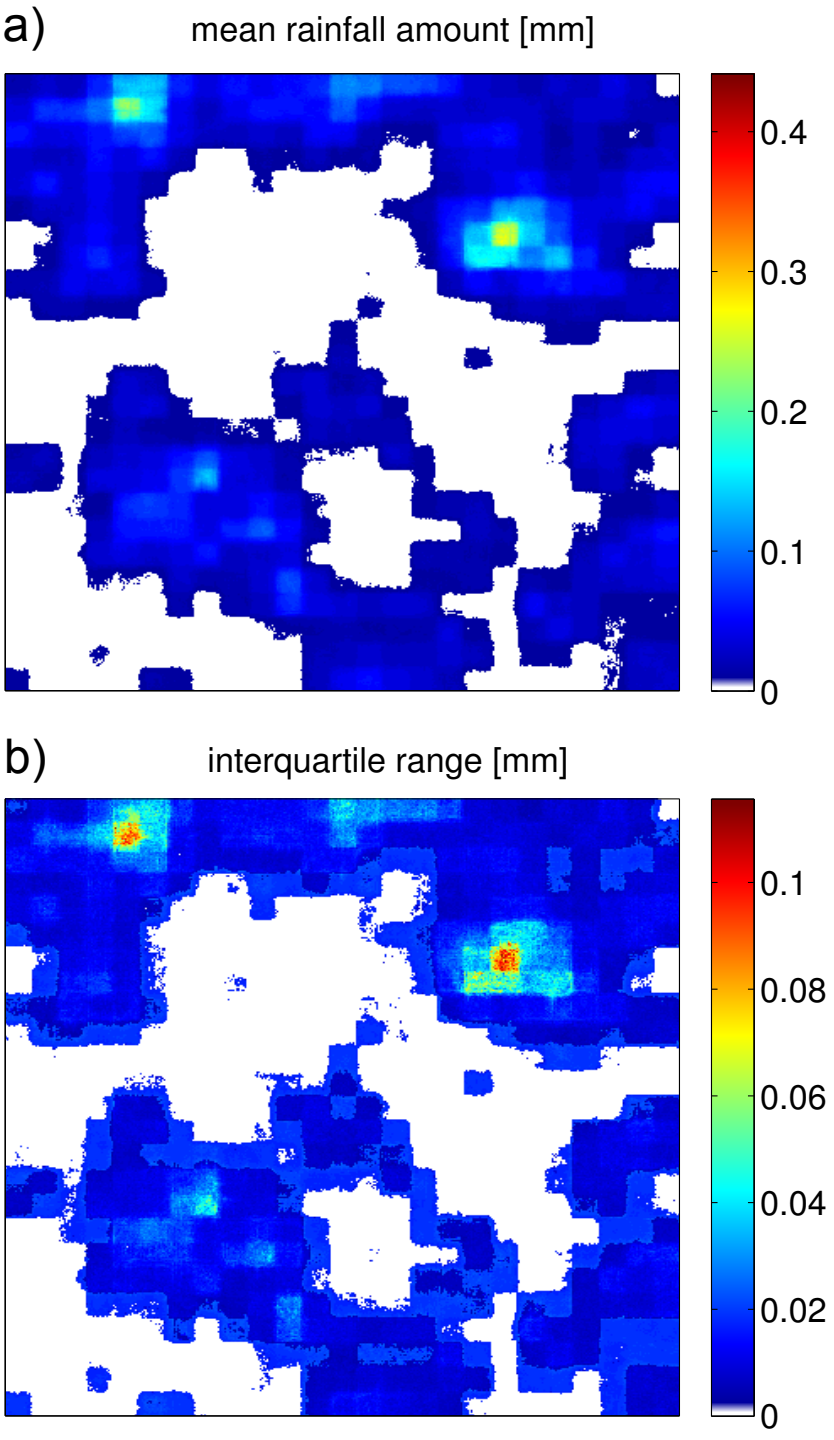


Figure 7.4: Mean value (a) and interquartile range (b) of the simulated rainfall field ensemble (100 realizations).

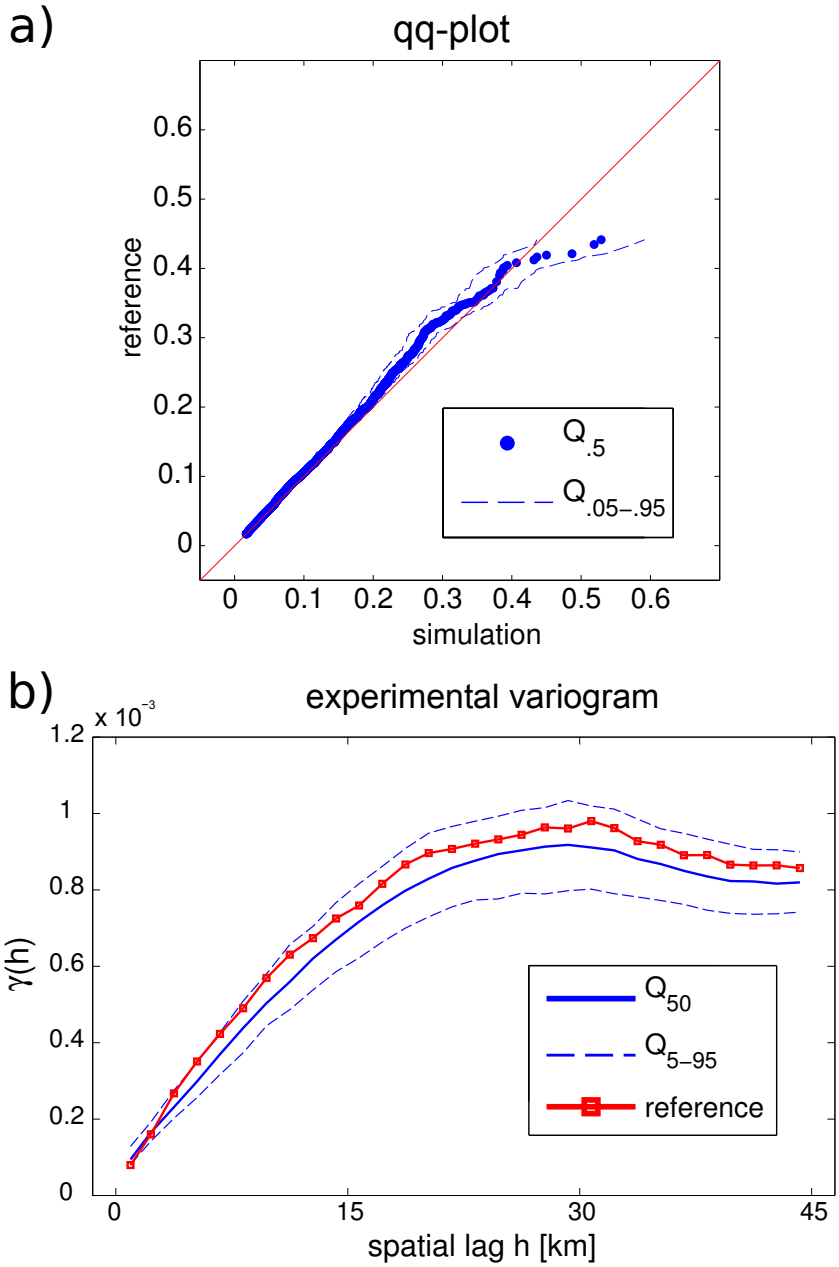


Figure 7.5: Graphs of the statistical indicators used to compare the simulated fields with the reference: a) qq-plot b) experimental variogram. For each value of the functions, the quantiles 0.5, 0.95 (dashed lines) and the median (solid line or dots) of the simulation ensemble are shown.

(5 km) regular grid. The results are promising: the simulated heterogeneity shows rainfall patterns fairly similar to the reference. This is relevant considering that the fields have been simulated without disposing of any 2D high-resolution image. The skewed rainfall intensity distribution and the covariance structure of the field have been approached reasonably well. Nevertheless, a modest noise and a weak blocky texture is visible in the simulated fields. These artifacts are imputable to the conditioning and the resampling procedures. Further testing is necessary to assess their influence on the quality of the prediction.

At this stage of development, this approach requires some strong hypotheses to be met: for the same values of the conditioning satellite variable, small-scale rainfall should exhibit strong stationarity in space and time. Moreover, even if anisotropic large-scale structures are imposed by the satellite images, the technique does not take into account the anisotropic character of small-scale rainfall in the time-series transformation. Finally, the rainfall field should present a moderate displacement velocity, otherwise the transformed time-series would not reach the desired spatial resolution. In the real application, these hypotheses may be met only for limited time and a small flat region. The satellite variable may allow distinguishing different types of rainfall event and admitting non-stationary rainfall in space and time, but this hypothesis has still to be confirmed with a test on real data.

Nevertheless, the technique can be further developed to gain more generality. The estimated mean cell velocity $v = |\vec{v}|$ is the magnitude of $\vec{v}$, the mean rainfall cell displacement vector. For an anisotropic field, the space-series $Z_s(x)$ would be representative of the variability in the direction of the displacement only. In this case, one can estimate the anisotropy in space at a certain time by analyzing the relative satellite image. A space-series for each i-th direction can be generated by applying the transformation $Z(t) \rightarrow Z_s^i(a_i v_x t)$, where a_i is the anisotropy ratio between the i-th direction and the direction of $\vec{v}$. In other words, a_i is used as a multiplying factor to change the spacing of $Z(t)$ and adapt it to different simulation directions. For example, the simulation scheme shown in figure 7.1 c includes the horizontal and vertical directions, needing two different transformation of $Z(t)$. This approach is based on the hypotheses that the satellite image can give enough information to estimate the mean anisotropy and that the covariance structure can be approximated for any direction by multiplying the spacing of $Z_s(x)$ by a factor a_i. Moreover, when simulating a rainfall field covering a large region, the hypothesis of stationarity may be relaxed in space by defining subregions using different training data sets. For this purpose, the transition between regions can be simulated by using the methodology presented in appendix D, where, at each point, a training data set is randomly selected according to a user-defined probability distribution. At least one time-series representative of each subregion should be available.

Beside the possible developments of the technique, one main drawback is that the spacing of $Z_s(x)$ depends on the temporal resolution of the original time-series $Z(t)$ and the displacement velocity v_x. This leads to two consequences: 1) v_x can easily change over time, leading to an unevenly spaced $Z_s(x)$ – transforming $Z_s(x)$ into an evenly spaced space-series that would require an interpolation stage or it may lead to a resolution lower than the desired one; 2) the temporal resolution in space is directly linked to the resolution of time-series used as training data: in practice, to reach the small-scale variability in space, a high temporal resolution in the training data and in the simulated field is mandatory. This time resolution is unnecessary for many applications and makes unpractical to realize simulations for long time periods. For these reasons, the technique is more suitable to generate hyper-resolution spatio-temporal models of a single rainfall event observed by the satellite, which can lead to better predictions in the domain of urban hydrology. Another convenient use of the technique would be to generate one realization of a sufficiently long period of time, that can be upscaled in time and used itself as training data set to simulate rainfall fields for long time periods preserving the high-spatial resolution.

The main advantage of the technique, which makes it a valid alternative to the ones already existent, is that its efficiency does not strictly depend on the spatial density of the rain gauge network: one time-series sufficiently long and representative of each subregion should be sufficient to realize the simulation. This feature makes the technique suitable to scarce data regions, where a small-scale interpolation of spatial data is not possible. Moreover, the sole use of time-series and satellite data extends its applicability to developing countries where high-resolution remote sensing imagery are generally not available (the World Meteorological Organization have recorded about 850 weather radar nowadays and more than 76% of them are located in developed countries [1]). On the contrary, large-scale weather images covering large part of the Earth are freely available from the last-generation satellite missions databases (see for example the TRMM project [178]).

In conclusion, the presented approach shows a relevant potential, but the whole methodology still needs to be tested on a real case study.

Chapter 8

Conclusions

In this thesis, we present some simulation tools based on the Direct Sampling (DS) technique to explore the variability in the domain of hydrology. In the following, a summary of the results is given together with a discussion about the possible improvements and the possible future developments.

8.1 Main results

Since it is the first time that DS is used for the simulation of hydrological and climate variables, the research has been oriented towards the development of techniques for different applications, with the aim of exploring the capabilities of the algorithm and providing as much as possible ready-to-use simulation tools.

For its large use in hydrological modeling, one main subject treated is time-series simulation. For daily rainfall (chapter 2) the proposed DS setup can generate realistic replicates for different climate types without changing its parameters. As a consequence of considering a variable time-dependence inside the simulation, the variability and structure up to the decennial scale is preserved. This is an important result since the majority of the daily-scale simulation techniques cannot preserve the multiple-scale variability. Conversely, one main limit of DS, as shown in chapter 3, is that, being entirely based on resampling of the training data, it is not capable of generating new values in the simulation. This is not an issue for categorical variables and for variables not showing significant variations in the extremes. Conversely, in the applications considered here, the generation of extremes is one primary goal since the treated variables mainly follows a hard-tailed statistical law, where longer time-series normally show higher extremes. Consequently, DS, at this stage of development, is not suitable to explore the asymptotic behavior of one process since it cannot represent long recurrence time events not present in the training data. To make realistic predictions, the time-series generated with DS have to be limited to a length of the same order of magnitude as the training data set used. On the other hand, the algorithm preserves the high-order statistical dependency contained in the data, a result that is normally not achieved with other simulation approaches. This means that DS better preserves the complex patterns characterizing the heterogeneity, which can be a significant advantage in case of highly irregular processes (see chapter 3). As for its previous applications, DS offers the possibility of conditioning using known values at certain time steps or some auxiliary variables. The application of chapter 4 takes advantage from this feature to simulate missing data portions of different sizes inside a flow rate time-series. This is a strong point of the DS approach considering that most of the used simulation techniques, e.g. the ones based on Markov-chains, simulate a time-series from the beginning to the end without the possibility of considering conditioning data.

The DS technique, tested on the joint simulation of different climate variables (chapter 5), confirmed its capability to generate realistic patterns in a multivariate framework. In addition, by preserving the multi-variate patterns, also the joint probability distribution of the simulated variables is correctly preserved without needing to build a complex statistical model. Moreover, a simple approach to evaluate the uncertainty of a climate change scenario is proposed: the climate variables trend is imposed using a simple delta-change approach and accounting for the linear relation between the variables, while the more complex structure of the trend residuals is simulated using DS. This method allows exploring the uncertainty of the future scenario by considering the complex structure of the variables and their relationship. On the other hand, it relies on strong hypotheses that may be not met in future: the stationarity of both the linear relation between variables and the probability distribution of the residuals. Nevertheless, this test shows that DS can be incorporated in a simulation framework of this kind to simulate one component of the signal which shows a complex

structure.

Regarding spatial simulation, two types of methods to generate spatial rainfall fields are proposed. The first one (chapter 6) is a simulation technique using daily rainfall radar images as training data set, capable of generating daily rainfall fields at the kilometric scale, approaching reasonably well the spatial structure of the reference fields. Moreover, by giving the digital elevation model as conditioning variable, the complex statistical relation between rainfall and altitude is efficiently preserved. The strong point of this simulation setup is that it can generate structures similar to the ones observed in the reference and conditioning on elevation is done by simply using this information as auxiliary variable. Nevertheless, the experimental variogram, representing the covariance structure of the field, is not accurately preserved as done with variogram-based methods. While the correlation length of the structures is correctly represented, we observe a modest underrepresentation of the extremes that reflects on the underestimation of the variogram sill. In this particular case, the problem is due to the proximity of a missing data zone to the highest rainfall values, that lie mainly along the coastline. The presence of missing data prevents the algorithm from sampling these patterns, with a subsequent underrepresentation of the extremes. This result stress the importance of using a complete data set, especially when the heterogeneity is non-stationary and the extremes are not evenly distributed.

The second and last DS application proposed for spatial simulation regards the generation of high-resolution spatial rainfall fields using satellite images and ground measurements (chapter 7). Where high-resolution remote sensing measurements are missing or not reliable (e.g. calibration problems, limited coverage) simulation techniques cannot rely on a continuous high-resolution training data set. This DS application is proposed to answer this issue: satellite images are used as conditioning variables to impose the large-scale structure of rainfall, while the small-scale heterogeneity is simulated using rainfall time-series from ground stations. Instead of interpolating the sparse data available in space, the technique make use of the rich time-series record available as a training data set. To approach the spatial covariance structure, the time-series are transformed by multiplying their data spacing by the mean rainfall cell velocity. The technique is evaluated with a preliminary test on a synthetic rainfall field. Even if the used reference model is overly simplified with respect to real rainfall, since it is based on strong hypotheses of stationarity, the preliminary results are encouraging: the probability distribution is efficiently preserved and the experimental variogram of the reference is reasonably approached. Several improvements are necessary and possible: for example, the technique can be adapted to anisotropic rainfall fields by applying an anisotropy study based on satellite images. This approach can be a convenient solution to simulate hyper-resolution models of a rainfall event without the need of expensive instrumentation, since satellite images are mainly available for free and only one or few high-resolution pluviometers are required.

8.2 Overall recommendations

In order to achieve reliable simulations using Direct Sampling, one main condition is required: the training data set used should provide a representative example of the studied heterogeneity, covering its variability exhaustively. Moreover, the represented heterogeneity should be stationary. When this last requirement is not met, an auxiliary variable describing the non-stationary structure can be used to include this information in a multivariate training data set. This may include periodic structures that should be repeated different times (e.g. the annual seasonality of a time-series) or more complex types of non-stationarity (e.g. the rainfall non-stationarity depending on elevation). The auxiliary variables can be co-simulated with the target variable to explore the uncertainty on the non-stationary structure or given as

conditioning data. This implies a certain creative effort from the user perspective to choose or build a specific auxiliary variable, describing an important aspect of the heterogeneity: for example, a categorical variable distinguishing the wet from the dry zones and the ones occupied by extremes values is used for spatial rainfall simulation in chapter 6, an indicator variable describing the hydrographic structure of flow-rate time-series is used in chapter 4, etc.

Another fundamental requirement to adapt the algorithm to a specific application is the careful choice of the main parameter values: the search neighborhood radius (R) limits the size of conditional patterns and can be related to the correlation length of the simulated heterogeneity, while the maximum number of considered neighbors (N) and acceptance threshold (T) applied to the comparison of the patterns, do not have a clear physical or statistical relation with the heterogeneity. T controls the strictness of acceptance of the patterns and the subsequent small-scale noise added to the simulation. Therefore, it could be related to the sample autocorrelation value measured on the training data. N is of less clear interpretation but it can be easily adjusted by trial an error since it takes integer values in a limited range (usually 5-25, see the recommendations of chapter 2 and appendix A). The DS parameterization is far from being automatic but, when an adequate setup is found for a type of heterogeneity, it can be used for different data sets, as shown for rainfall time-series (chapter 2) and spatial fields (chapter 6).

One last requirement regards extreme values: as already mentioned, DS cannot generate values not present in the training data set. For a process presenting large-recurrence-time events, the simulated data amount should not be larger than the training data set used. This is a limit of the technique which makes it not suitable to explore the long-term extremal nature of a process.

8.3 Parametric and non-parametric techniques: towards a hybrid approach

In the last decades, parametric and non-parametric techniques have followed two parallel pathways in the domain of hydrology: giving often comparable results and coming from different schools of thought, they have been seldom looked as approaches complementary to each other. This direction has been taken only recently, with some research works showing how rainfall and climate variables can be modeled using a semi-parametric method: the non-parametric component is most of the time a kernel smoothing technique incorporated in a parametric simulation framework to estimate a conditional probability distribution [186, 9, 221], a joint probability distribution [157] or to model its inner part, while a parametric model is used for the tail [185]. In some cases, an artificial neural network is coupled to parametric techniques to model a complex dynamical system [e.g. 96]. In general, non-parametric techniques (such as kernel smoothing, resampling and artificial neural network) can represent more efficiently complex statistical relations and non-linear behaviors, mainly relying on the available data. Conversely, the parametric techniques (such as regression-based, Bayesian and classic geostatistical methods), represent the statistics only up to a certain complexity, but with a better control of the distribution tail, allowing the extrapolation of extremes and the imposition of trends, based on an analytical probabilistic model. In this scenario, it is natural to glimpse the potential of a hybrid approach, taking advantage of both parametric and non-parametric methods.

We can find some examples in this thesis, where the Direct Sampling technique has been proposed as non-parametric ingredient of a hybrid simulation recipe. DS has been coupled to linear regressions to extend the climate change information to multiple variables and simulate a multivariate non-stationary time-series (chapter 5). The use of parametric

distributions is proposed to allow the extrapolation of extremes not contained in the training data set (chapter 3). Finally, linear transformations have been applied in chapter 7 to obtain a training data set used with DS for the simulation of spatial fields. It has been shown that the use of parametric techniques is of paramount importance to extend the DS applicability to complex application cases. We believe that the semi-parametric approach opens up the way for a multitude of effective simulation strategies, therefore it should be fostered in future methodological research.

8.4 Possible improvements and future developments

The presented simulation tools are just prototypes that can be developed in several directions. Without imposing a statistical model, it is difficult to allow the extrapolation of extremes not present in the training data set. As proposed in chapter 3, a possible improvement of the DS algorithm may include a perturbation stage to generate new values. This implies the choice of a limit distribution for the simulated variable, but it would allow generating new extremes associated to long recurrence-time events.

Another future development may include a statistical analysis of the relationship between the autocorrelation properties of a large number of different simulated heterogeneities and the T value chosen by the user to correctly simulate them. This may allow defining an empirical relation to estimate a prior starting T value and speed up the parameterization of the technique.

In the case of daily rainfall time-series simulation, the proposed setup is already applicable on real cases, but it is not adapted to the joint simulation of data from multiple stations together. A multi-station setup could be a natural development of the technique. Nevertheless, simulating a high number of variables may degrade the statistics related to each single station. This can be a possible issue requiring a reconsideration of the simulation setup.

Another possible development regards the generation of long-term spatio-temporal fields using radar images: the problem in this case is that the non-stationarity of the whole historical record is much more complex than in a time-series, since we can have a specific structure inside each different type of rainfall event and a seasonal fluctuation presenting in some cases long dry periods with zero-rainfall all over the spatial domain. In this context, it is not practical to simply describe the non-stationarity with one or more auxiliary variables and scan the whole training data set for the generation of each datum. A more convenient approach would be to divide the simulation in two stages, as explained in chapter 6: 1) A time-series simulation of a categorical variable specifying the synoptic weather conditions for each day (e.g. 0 = dry day, 1 = weak rainfall event, 2 = extended humid front, 3 = storm, etc...) and 2) a spatial simulation for each daily field using a dedicated radar image catalog as training data set. This would be acceptable at the daily scale, since the day-to-day correlation of rainfall amount at the same location is normally very weak. The advantage of this approach would be an increased simulation speed and an easier reproduction of rainfall patterns, since each training data set would represent a specific rainfall type. The setup can be further developed by adding the data of a climate model simulation as variable to explore the uncertainty of a specific climate change scenario (see chapter 6).

Finally, the developed tools need to be compared to the already existent techniques. We think the most effective way to have a real feedback about their performance would be to examine the distributed hydrological model response to the simulated fields. This can also inform about the degree of complexity really needed in the simulated heterogeneity and help the parameterization of the techniques.

8.5 Some personal remarks

I see this work as a little step towards the introduction of multiple-point statistics in the domain of hydrology. I tried to face the subject of stochastic simulation with a practical attitude: most of the applications proposed in this thesis are an attempt to answer some colleague's request, needing a simple technique to generate synthetic data, to fill in the gaps of incomplete sequences or to condition the simulation using correlated variables. This approach offers a much less rigorous mathematical framework compared to classical techniques, but, if integrated in the adequate simulation framework, it can really improve the quality of the predictions by approaching the space of realistic outcomes of a process. The algorithm implementations are freely available for scientific purposes to be tested, compared and developed. I encourage anyone to play with them and give us a feedback.

Appendix A

A sensitivity test on rainfall data

This short report shows the result of a sensitivity analysis on the simulation of rainfall time-series using the Direct Sampling technique, inspired by the more general work of [124]. In this case, we are interested to explore the parameter space where a more strict conditioning is applied in order to obtain a better preservation of the probability distribution, which is more critical for rainfall time-series simulation with respect to the usual application of the Direct Sampling algorithm. The sensitivity is analyzed for the parameters N, the maximum number of conditioning neighbor data, and T, the distance threshold used in the pattern comparison. As explained in chapter 2, N represents the maximum order of conditioning applied in the simulation. Since the neighbor configuration varies through the course of the simulation, this order is variable. T is the distance threshold below which a scanned pattern is considered compatible to the conditioning one. This quantity is therefore linked to the precision applied to reproduce the heterogeneity. Since there is not a clear way to set up these parameters on the basis of a preliminary analysis of the heterogeneity, the analysis of the algorithm response is critical to define a range of suitable values.

A 50-year daily rainfall time-series is simulated using as training data the 1951-2000 historical record from Neuchâtel (Switzerland, Meteoswiss database). The same data are used as reference and 100 realizations are generated. The experiment is repeated using different combination of values for the mentioned parameters among the following ones: $N = \{2, 5, 20, 200, 500, 100\}$ and $T = \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$. The rainfall amount is simulated without using any auxiliary variable. In the following, the statistical indicators used together with the relative results of the sensitivity analysis are presented (figure A.1).

The logarithm of the total computation time for all realizations [sec] is used to have an idea of the impact on computational burden obtained by setting different values of N and T. The results (figure A.1, a) show that for $N < 20$, i.e. when a small group of neighbor data are used for conditioning, no sensible change in the computation time is observed as a function of T. The computational time increases gradually for values of $N > 20$ when $T > 0.01$. This situation corresponds to the case where a complex neighborhood is considered to condition the simulation but the accepted variability inside it is large. In other words, a high-order conditioning is applied but with a low precision on each conditioning datum. Conversely, we observe a rapid augmentation of the computational time, up to 7 order of magnitude higher, with values of $N > 200$ and $T \leq 0.01$. In this range of values, the algorithm looks for complex patterns and accepts only a small mean variation with respect to the conditioning data. The simulation time is therefore maximized.

The percentage of non-matching patterns (figure A.1, b), i.e. the percentage of times where no pattern is found presenting a distance below the threshold T, is used to quantify the difficulty in reproducing the heterogeneity with a certain precision (related to T) and complexity (related to N). The results show that, for the considered heterogeneity, non-matched patterns are present only for $N > 20$ and $T \leq 0.01$. This critical range, that we can call *no-matching zone*, is correlated to the previously mentioned increase of the computational time. It corresponds to the case where a large neighborhood is considered and T is too low to find a matching pattern in the majority of the cases. This parametrization is therefore not functional to the natural pattern reproduction and should be avoided.

The lag-1 autocorrelation (AC, figure A.1, c) absolute error is used to quantify the correlation loss between subsequent data, i.e. the small-scale noise added in the simulation with respect to the reference. AC can go from 0 for totally uncorrelated signals to 1 for maximal autocorrelation, the measured noise takes values in the same interval. We observe that, for low N and T values, the added noise is minimized since only well matching patterns are accepted and the pattern distance, which is an average quantity, is evaluated over few neighbors. When the conditioning neighborhood is increased the noise begins to appear and is maximized if the condition of acceptance on the pattern is relaxed ($T \geq 0.05$). In the

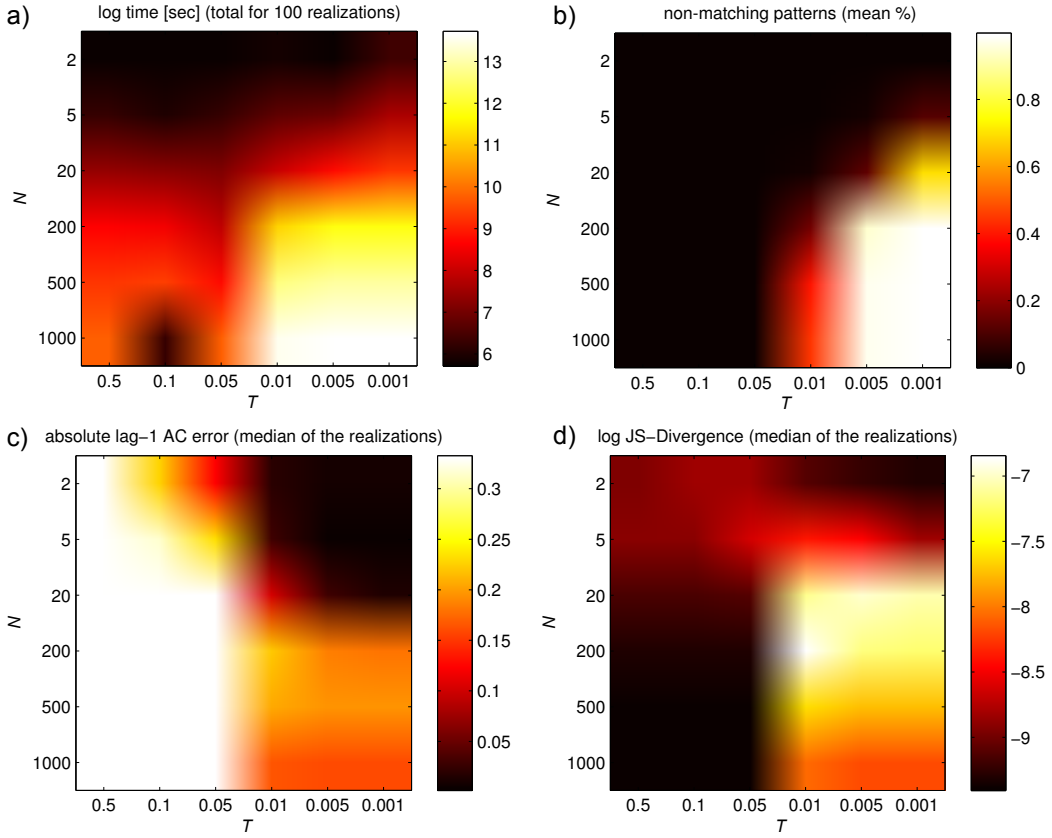


Figure A.1: Sensitivity maps showing the values of the considered statistical indicators in function of the parameters T and N: a) logarithm of the total computation time for all realizations [sec], b) mean percentage of non-matching patterns, c) absolute lag-1 autocorrelation error (median of the realizations) and d) logarithmic Jensen-Shannon (JS-) divergence (median of the realizations).

no-matching zone the error noise remains modest.

The fact that the small-scale noise is minimized does not guarantee that all the fundamental aspects of the heterogeneity are preserved. For this reason, the logarithmic Jensen-Shannon (JS-) divergence is computed. This indicator describes, with values ranging from $-\infty$ to 0, the general dissimilarity between the reference probability distribution and the one obtained from the simulation. The value shown in figure A.1, d) considers the median of the simulation and can be interpreted as a bias measure with respect to the reference. The results show that the bias is maximized in the no-matching zone. For an unclear reason, the bias is partially reduced for high N values.

The considered indicators describe different aspects of the simulation and the generated heterogeneity. In conclusion, a unique optimal combination for the parameters N and T does not exist, but a range of values representing a compromise to preserve different aspects of the heterogeneity can be defined: the no-matching zone should be avoided to avoid bias in the probability distribution and excessive computation time. For this purpose, $T \approx 0.05$ may be an appropriate starting value for daily rainfall time-series. The value of N should allow taking into account a certain complexity of the rainfall pattern, while containing the computational time: $N \approx 20$ may be an appropriate value. To better preserve certain aspects of the heterogeneity, the use of auxiliary variables may be a viable solution. For example, the co-simulation of the 2-day moving sum of the rainfall amount together with the daily

rainfall amount may help to reduce the small scale noise without changing the suggested parameter values.

Appendix B

Preliminary tests for rainfall time-series simulation: the auxiliary variable choice

Abstract

The following is a quick report about the preliminary test performed to find a suitable set of auxiliary variables for daily rainfall simulation using the Direct Sampling technique. According to the need of respecting some essential statistical aspects of the rainfall heterogeneity and the information used in stochastic algorithms proposed in literature, 10 combinations of auxiliary variables are chosen as candidate setups. A simulation of a 50-year daily rainfall time-series from a temperate region of Switzerland has been performed to test the performance of the mentioned combination of auxiliary variables. The selected variables are used as base for the development of the standard setup proposed in chapter 2.

B.1 Introduction

Until now, the Direct Sampling (DS) technique has shown the ability of generating realistic replicates of natural patterns relying on the information contained in the training data. Nevertheless, according to the nature of the simulated heterogeneity, the choice of the auxiliary variables used to condition the simulation plays a critical role on the performance of the algorithm. These variables are simulated together with the target variable as a multivariate data set. In this report, the preliminary tests on a series of auxiliary variable combinations for the simulation of daily rainfall time-series is shown.

The considered auxiliary variables are presented in the following. Three of them are deterministic functions of time used to impose the annual seasonality. They are not simulated but imposed as conditioning data:

- **month index** is a categorical variable indicating the current month in which the current day falls;
- **sincos** is a couple of periodic functions (sinus and cosine) with one-year period, indicating the exact position of each day in the year;
- **tr** is a couple of periodic triangular functions with one-year period and phase opposite to each other. They are analogous to sincos with the difference of being linear.

The rest of the variables describe other specific aspects of the heterogeneity and are co-simulated with the daily rainfall:

- **d/w** is a categorical variable indicating if the day is dry, wet, solitary wet, or wet day at the end or at the beginning of a wet spell;
- **rain dist** is a continuous variable computed on the rainfall occurrence time-series, which measures the number of days separating the current day from the closest dry/wet transition in past or future. The sign is positive for dry days and negative for wet days. For example, let us denote the rainfall occurrence time-series with R_t, being $R_t = 0$ for dry days and $R_t = 1$ for wet ones. The rain distance R_{dist} for R_t=[1 0000000100111111] is R_{dist}=[001232100000-1-2-3-4-5].
- **365MA** is the 365-day moving average, imposing inter-annual fluctuations.

The main DS parameters, namely search radius R, maximum number of considered neighbors N and distance threshold T (see chapter 2 for a detailed description), are not optimized, but a series of fixed values have been adopted for each variable (table B.1): R=1 and N=3 are adopted where no specific pattern is supposed to be generated, for example for theoretical periodic variables given as CD. Higher R and N values are given in specific cases,

Table B.1: Summary of the DS parameter values used for the target and candidate auxiliary variables.

variables	R	N	T
rainfall	50	60	0.05
month index	1	3	0.1
sincos	1	3	0.05
d/w	10	21	0.05
rain dist	10	21	0.05
365MA	1	3	0.07
tr	1	3	0.05

Table B.2: Summary of the auxiliary variable combination used for each setup.

setup n°	**month index**	**sincos**	**tr**	**d/w**	**rain distance**	**365MS**	**VVM**
1	yes	-	-	yes	-	-	-
2	-	yes	-	yes	-	-	-
3	-	yes	-	-	yes	-	-
4	yes	-	-	yes	-	-	yes
5	-	yes	-	yes	-	-	yes
6	-	yes	-	-	yes	-	yes
7	yes	-	-	-	yes	-	-
8	yes	-	-	-	yes	-	yes
9	-	yes	-	yes	-	yes	-
10	-	-	yes	yes	yes	-	yes

where these values are related to the typical size of the pattern that should be simulated. This quantity can be estimated visualizing the heterogeneity or it can be related to some statistical indicator as the variogram range or the integral scale of connected components (for categorical variables). The estimated values are then refined by trial and error. For the distance threshold T, the initial value of 0.05, estimated for the rainfall amount with the sensitivity analysis presented in appendix A, is generally kept for all the variables. Only in some cases (month index and 365MA) it is augmented to reduce the strength of the conditioning, since a bias on the rainfall distribution was observed in the simulation.

Different combinations of the mentioned variables constitute the setups considered (table B.2). An optional simulation mode is also tested: the **variable vector mode (VVM)**. In the standard mode, each variable at the same time step is generated separately by sampling the training data where a similar multivariate pattern occurs. Conversely, using the variable vector mode, the whole variable vector is generated at one time by sampling all the variables at a given location of the multivariate training data set.

The experiment is composed of the simulation of a 50-year daily rainfall time-series, using as training data the 1951-2000 historical record from Neuchâtel (Switzerland, Meteoswiss database). The same data are used as reference. 10 realizations for each setup are generated.

B.2 Results

The marginal distribution of the simulation and the reference are compared using qq-plots of the rainfall distribution at different scales: daily, monthly and annual. The qq-plots (figures B.1, B.2 and B.3) show that the reference distribution is preserved up to the annual scale for all the considered setups. No significant improvement by changing the auxiliary variable combination is observed for these indicators. The 10-year scale distribution, not investigated here for the limited size of the historical data used, may be better preserved by using setup

9, where 356MA is used as auxiliary variable (see chapter 2).

The mean monthly rainfall boxplot (figure B.4) shows that all the setups can preserve the annual seasonality. The different auxiliary variables used for this purpose, namely the month index, sincos and tr, show therefore an equivalent performance. Conversely, the mean daily rainfall on solitary wet days (figure B.5) is better preserved when d/w (dry/wet pattern) is used (setups 4,5 and 10) together with VVM (Variable Vector Mode). By using VMM, the correlation between rainfall amount and rainfall pattern is preserved more efficiently, which is necessary to respect the solitary wet day distribution. The efficiency of d/w and VVM is also seen for the dry and wet spell size distribution (figures B.6 and B.7), where the other combinations, although presenting a seasonality similar to the reference, tend to underestimate both the dry and wet spell length.

Finally, the Patching Index (PI, figure B.8), i.e. the size of the longest data sequence present in both the simulation and the reference, is measured to analyze the efficiency of the algorithm in generating new data without reproducing too long sequences of the training data set. The same indicator can be applied to the reference itself, to measure the longest data sequence repeated two times. In general, all the setups present a modest PI value, close to the one found in the training data itself.

B.3 Conclusions

This report shows the preliminary tests made to delineate a group of auxiliary variables necessary to obtain realistic simulation of daily rainfall time-series using the Direct Sampling technique. The analysis of the simulation, realized using different variable combinations, puts in evidence the essential need of the following auxiliary variables:

1. A theoretical periodic function used as conditioning datum to impose the annual seasonality. Although the three considered alternatives (month index, sincos and tr) bring a similar contribution to the simulation, tr may be the most appropriate since it describes the annual cycle with more continuity than the month index and better approaches a linear behavior with respect to sincos.

2. A categorical function describing the rainfall occurrence pattern. In this case d/w gave the best results in terms of mean solitary wet days rainfall amount and dry/wet spell size.

3. A function describing the rainfall long-term behavior. 365MA could be a good candidate, but a test on longer historical records is necessary to confirm its efficiency.

This results constitute the base on which a standard setup for daily rainfall time-series is developed and tested on data sets from different climates (chapter 2).

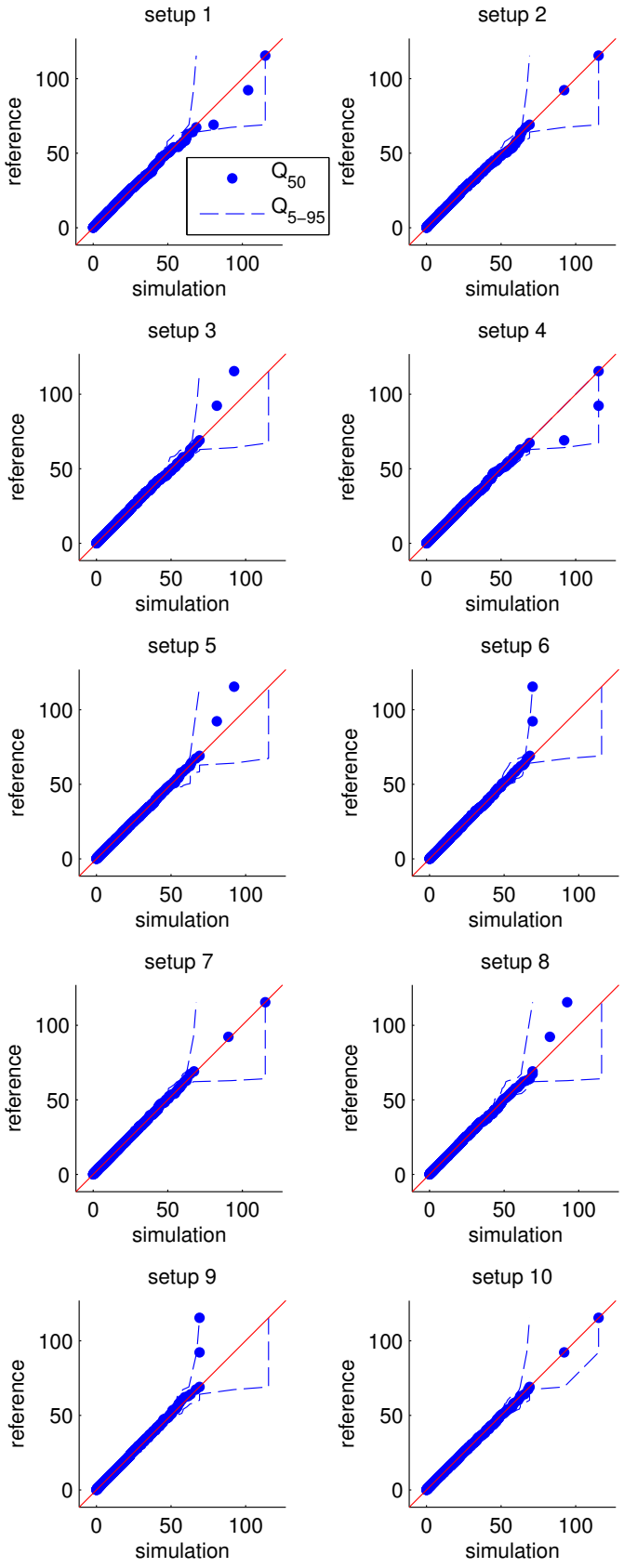


Figure B.1: qq-plots of the daily rainfall amount [mm] distribution: median of the realizations (blue dots), 5th and 95th percentile (dashed lines). The bisector (solid line) indicates the exact quantile match.

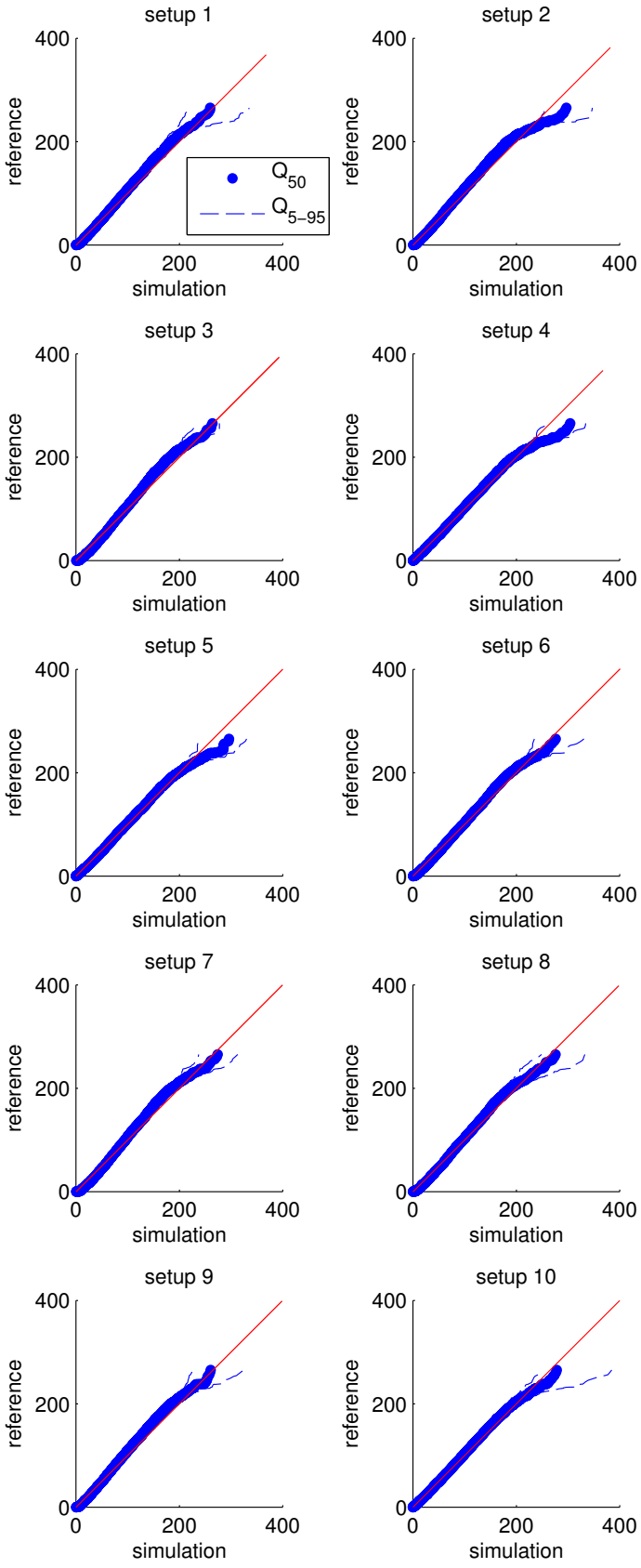


Figure B.2: qq-plots of the monthly rainfall amount [mm] distribution: median of the realizations (blue dots), 5th and 95th percentile (dashed lines). The bisector (solid line) indicates the exact quantile match.

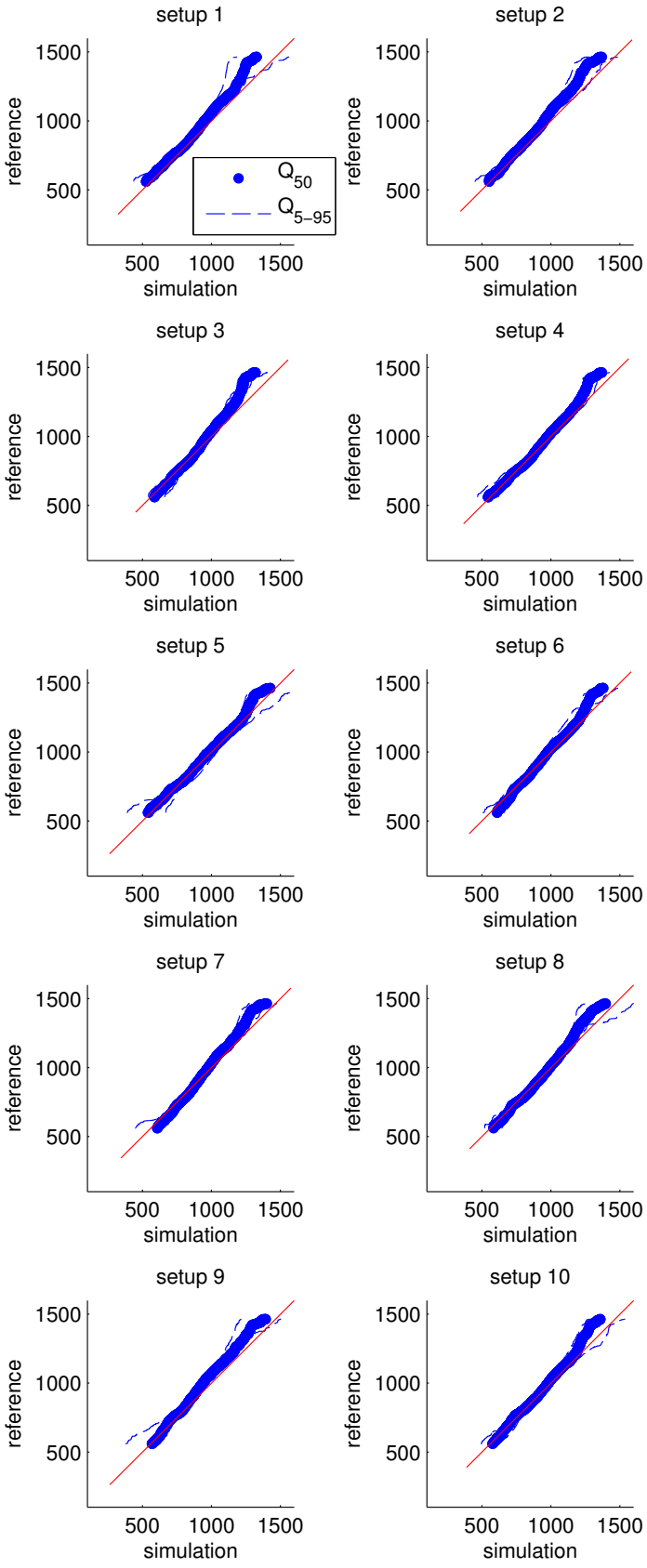


Figure B.3: qq-plots of the annual rainfall amount [mm] distribution: median of the realizations (blue dots), 5th and 95th percentile (dashed lines). The bisector (solid line) indicates the exact quantile match.

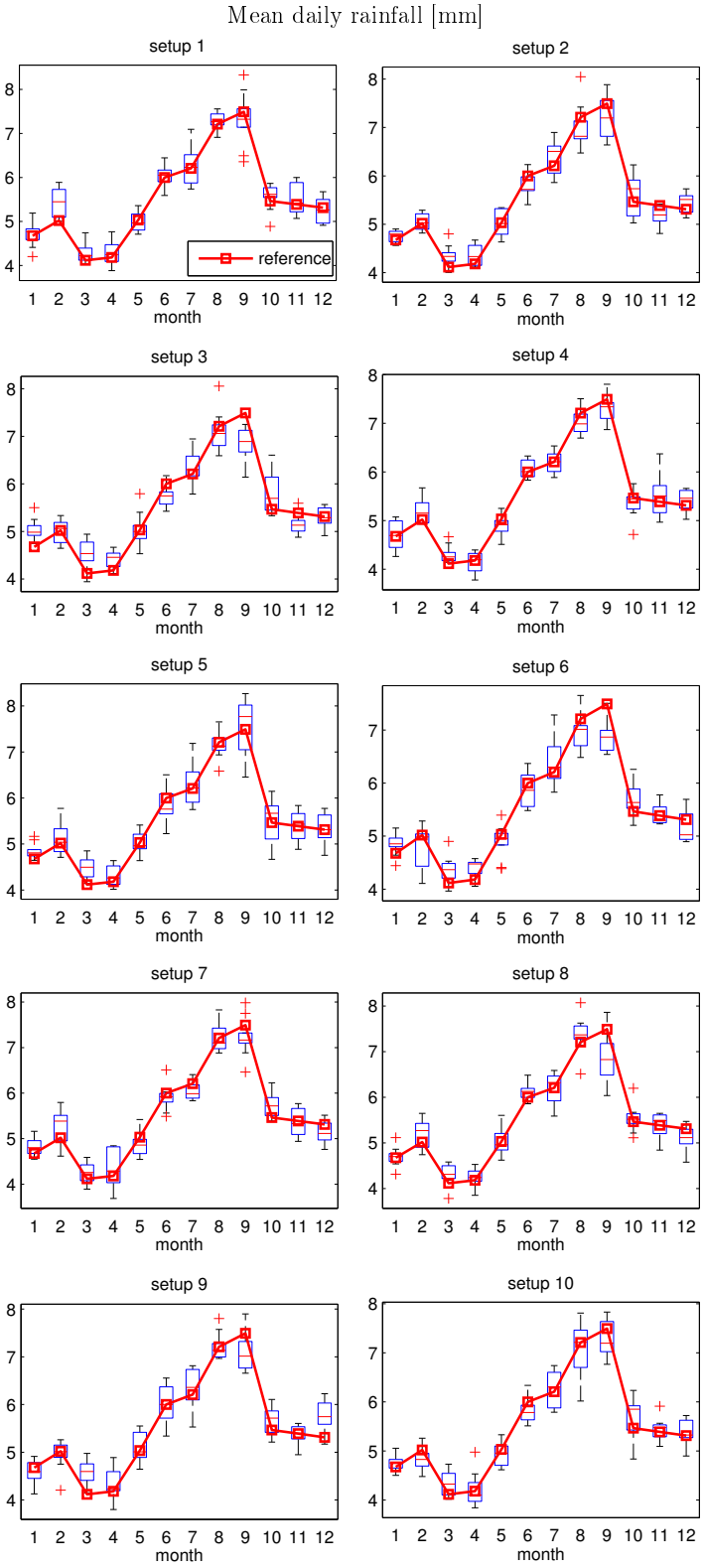


Figure B.4: Box-plots of the mean rainfall per month [mm]. The solid line indicates the reference.

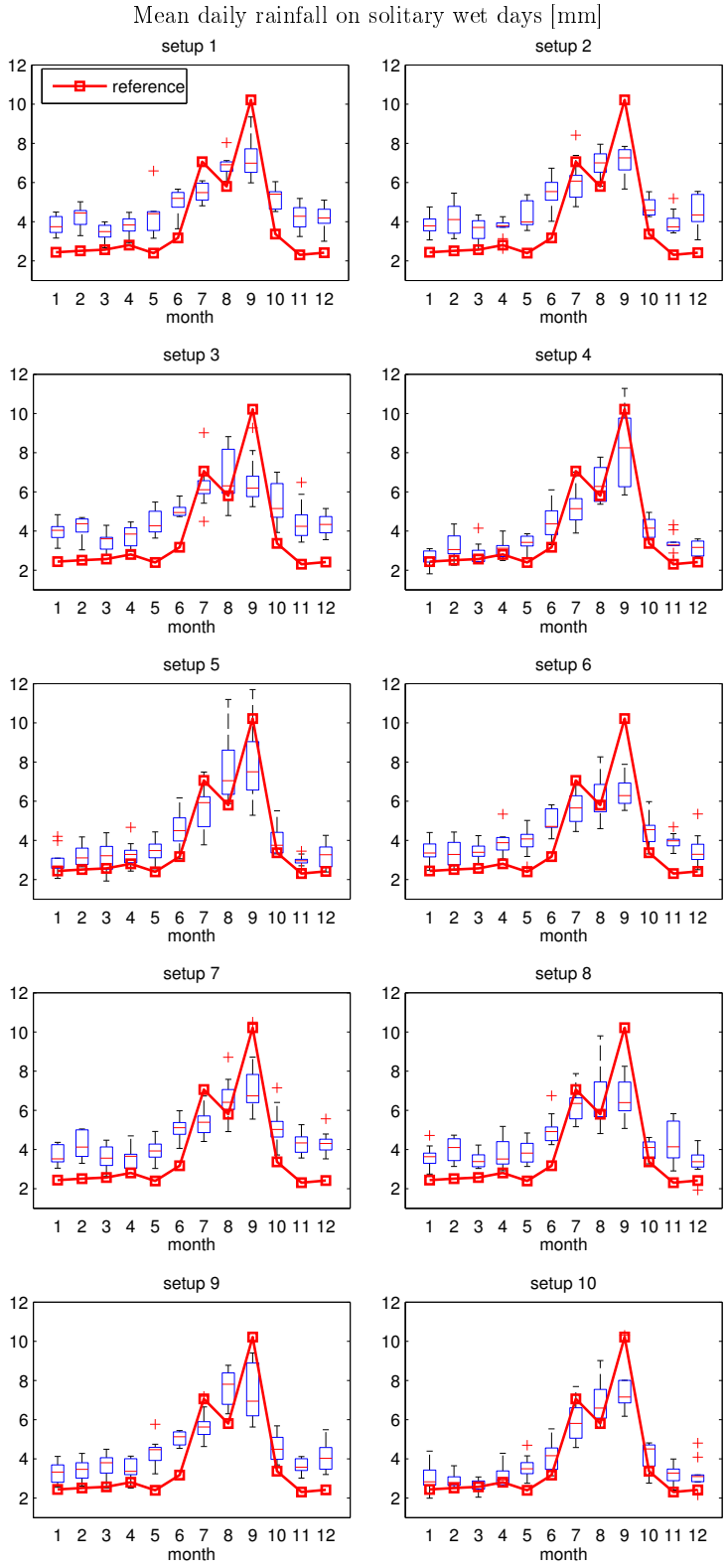


Figure B.5: Box-plots of the mean rainfall on solitary wet days per month [mm]. The solid line indicates the reference.

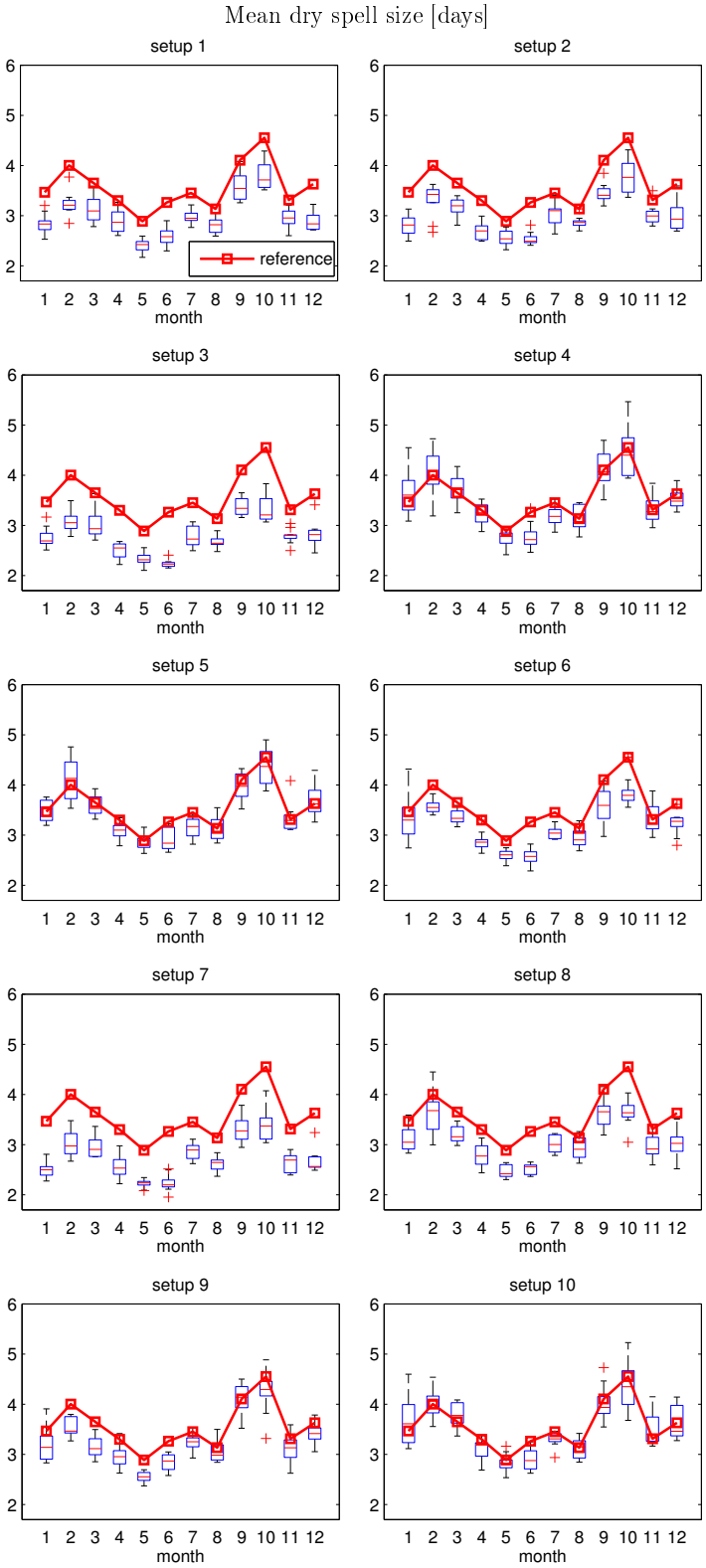


Figure B.6: Box-plots of the average dry spell duration per month [days]. The solid line indicates the reference.

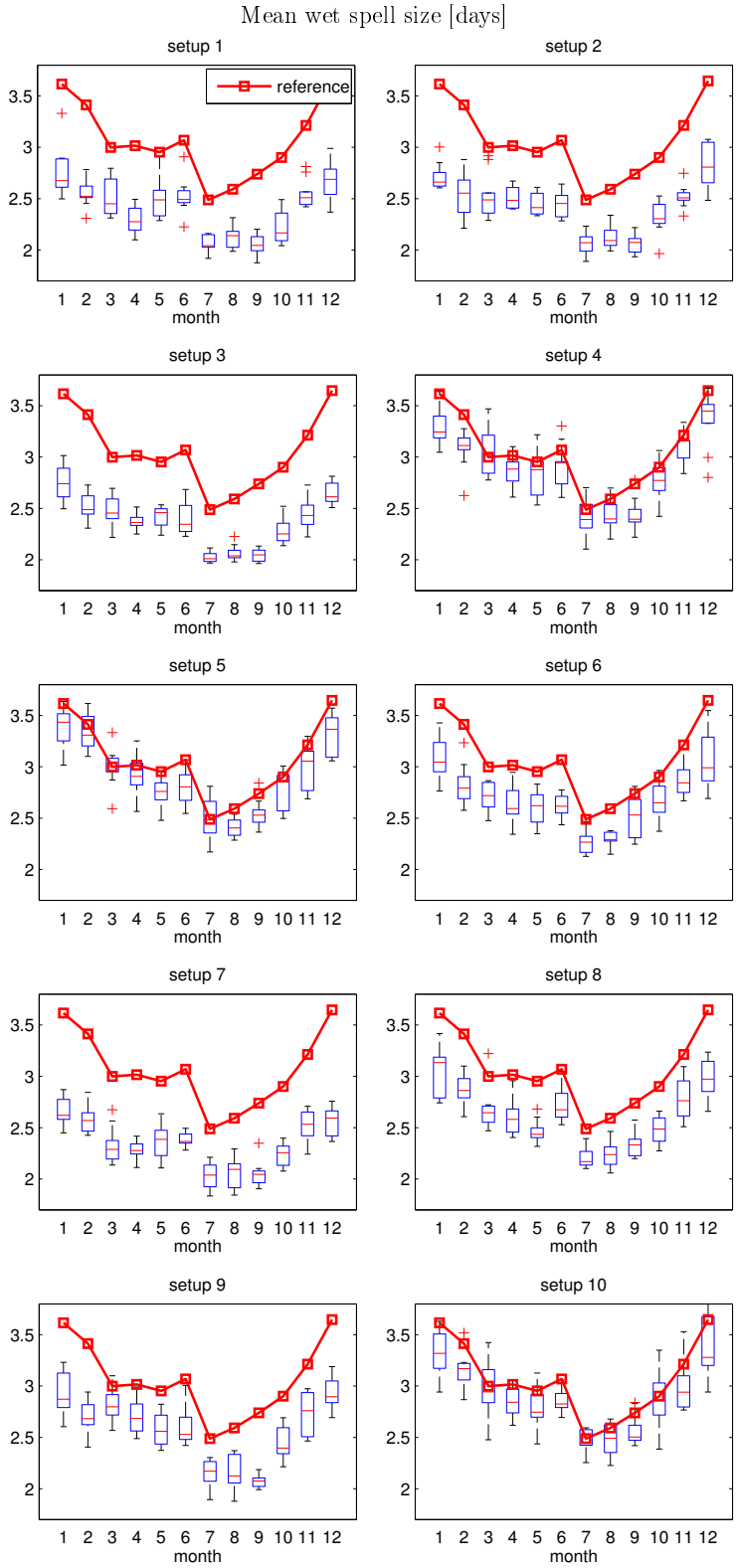


Figure B.7: Box-plots of the average wet spell duration per month [days]. The solid line indicates the reference.

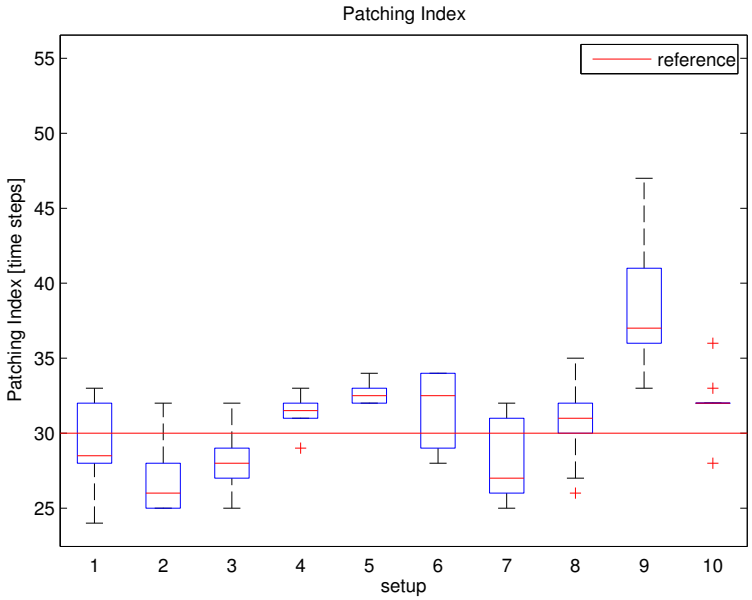


Figure B.8: Box-plots of the Patching Index. The horizontal line indicates the values found in the reference.

Appendix C

Preprocessing treatment techniques for time-series

In this appendix, some preprocessing operation applied to time-series used in the thesis are described. Let us consider a time-series $Z(t)$, where $t = t_0, ..., t_n$; is the measurement time. It is not unusual that the temporal signal presents some artifacts: for example some sharp fluctuations due to measurement errors or some abrupt changes of the mean value of a signal portion (shift). It can also happen that the signal is not stationary and presents a trend. In the following, we describe some techniques to eliminate these components.

C.1 Removing a trend

The trend for $Z(t)$ can be computed with the least-square linear regression:

$$\hat{Z}(t) = \alpha + \beta t \tag{C.1}$$

where α and β are the regression coefficients and t the predictor variable, solving:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} [Z(t_i) - \alpha - \beta t_i]^2 \tag{C.2}$$

The detrended variable is obtained with $Z(t) - \hat{Z}(t)$.

C.2 Removing a shift

A shift present in the time-series portion $Z(M)$ with $M = \{t_a, ..., t_b\}$ and $\{a, b\} \in \{0, ..., n\}$, can be removed with:

$$Z(M) - [\bar{Z}(M) - \bar{Z}(M')] \tag{C.3}$$

where $\bar{Z}(\cdot)$ is the average operator on $Z(\cdot)$ and $M' = \{t_i \notin M\}$.

C.3 Removing sharp fluctuations

The following preprocessing treatment is developed to eliminate isolated sharp fluctuations waving around the local trend due to instrumental errors. Given a time-series $Z(t)$ and computing the differential operator $\Delta Z(t) = Z(t) - Z(t-1)$, the artifacts are identified with the portions of $Z(t)$ presenting $\sigma(t, a) > b$, where $\sigma(t, a)$ is the local standard deviation of $\Delta Z(t)$, computed on the time interval $[t \pm a]$ and b is a user-defined threshold. The appropriate value of a and b depending on the smoothness of the signal and the magnitude of the artifacts, can be manually setup by visually checking the result. In the application of chapter 4, the chosen values are $a = 19$, $b = 0.3$ for $Z(t) = \mathrm{Ar}$ and $b = 0.05$ for $Z(t) = \mathrm{Ar2}$. The data $Z(t)$ identified by the procedure are replaced by a cubic spline interpolation. As shown in figure C.1, this method can localize and correct the artifacts efficiently, bringing no essential alteration of the information contained in the time-series.

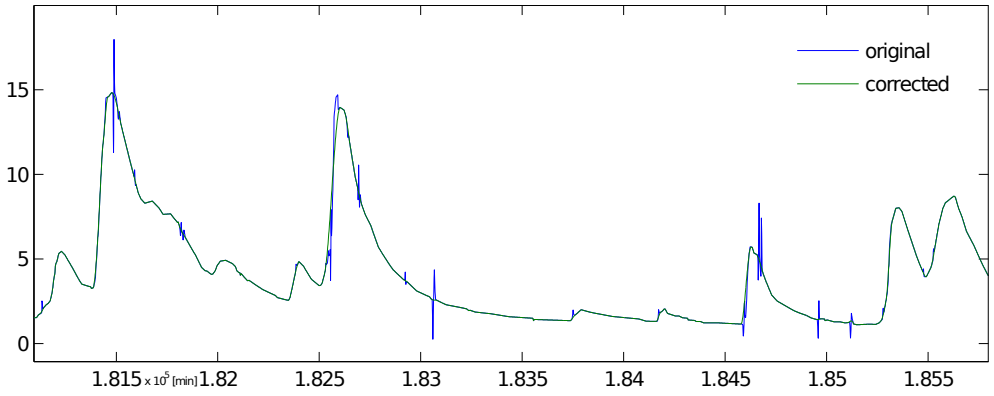


Figure C.1: Example from time-series Ar1 used in chapter 4, showing the preprocessing treatment to remove sharp fluctuations.

Appendix D

Daily rainfall simulation using multiple training images

This report shows an example of daily rainfall simulation using multiple training images, a feature available in the *DeeSse* implementation of the Direct Sampling algorithm [193]. Having multiple training data sets (TI) presenting different statistical features, it is possible to realize a simulation that represent a hybrid heterogeneity by sampling the different TI with a certain probability. The probability of sampling each TI is defined for each time step of the simulation grid and the sampled TI is then chosen randomly for each time step according to the specified probability distribution. This technique may be applied to rainfall and climate data in different situations, for example the simulation of a climatic time-series for a station located among multiple stations for which the time record is available and presenting a climatic transition among them. Another possible application could be the simulation of a climate shift at one station toward another climatic signature for which a training data set is available.

In this case, we show an example based on the 1941-2013 daily rainfall time-series from the Darwin station (BOM database) also used in chapter 2. The 10-year Moving Average (10yMA) (red line figure D.1) shows the long-term upward trend on the whole record of about 70 years. The exercise consists of using the first 22-year part of the record as a first training image (TIa), the last 22-year part as a second one (TIb) and using both TIs to simulate the in-between record of about 30 years. The two TIs present, at the daily scale, a compatible range of values but TIb shows a pronounced bias toward higher extremes with respect to TIa (qq-plot in figure D.2, a) which determines a completely different range of values at the 10-year scale. By looking at the reference record, we consider as prior hypothesis that the patterns contained in the two TIs are compatible and change with a linear trend from TIa to TIb along the the simulated time span. According to this, the probability of sampling TIa changes linearly from 1 to 0 as a function of time and vice versa for TIb. The rainfall amount at each time step will be simulated in a random order but the scanned TI will be randomly chosen according to the probability assigned. The known 33 year time-record is used as reference.

The results of 100 realizations show a behavior at the 10-year scale similar to the reference (one realization, blue line figure D.1). The simulation ensemble explore the uncertainty around the linear trend (median and 05-95-th percentile boundaries, green line figure D.1). The marginal probability distribution is preserved by the simulation ensemble (qq-plot in figure D.2, b) and the uncertainty related to the extremes is explored without introducing any bias.

The presented technique can be a handy tool to model a continuous transition or a hybrid state between multiple stationary heterogeneities by sampling multi training data sets with a given probability. The method allows imposing a linear or non-linear transition without adding much complexity to the algorithm or increasing the computational time. Nevertheless, further tests are necessary to assure its efficiency on more complex study cases. Moreover, it relies on strong hypothesis that the user should verify before applying the technique: it is intended to simulate a gradual transition and not sharp variations or shifts, therefore the heterogeneity in the different TIs should be compatible to each other in order to permit it. Moreover, if the reference data are missing the hypothesis of having a certain transition should be based on observations (for example, from nearby stations) and should be coherent with the underlying physics of the natural process that generates the heterogeneity.

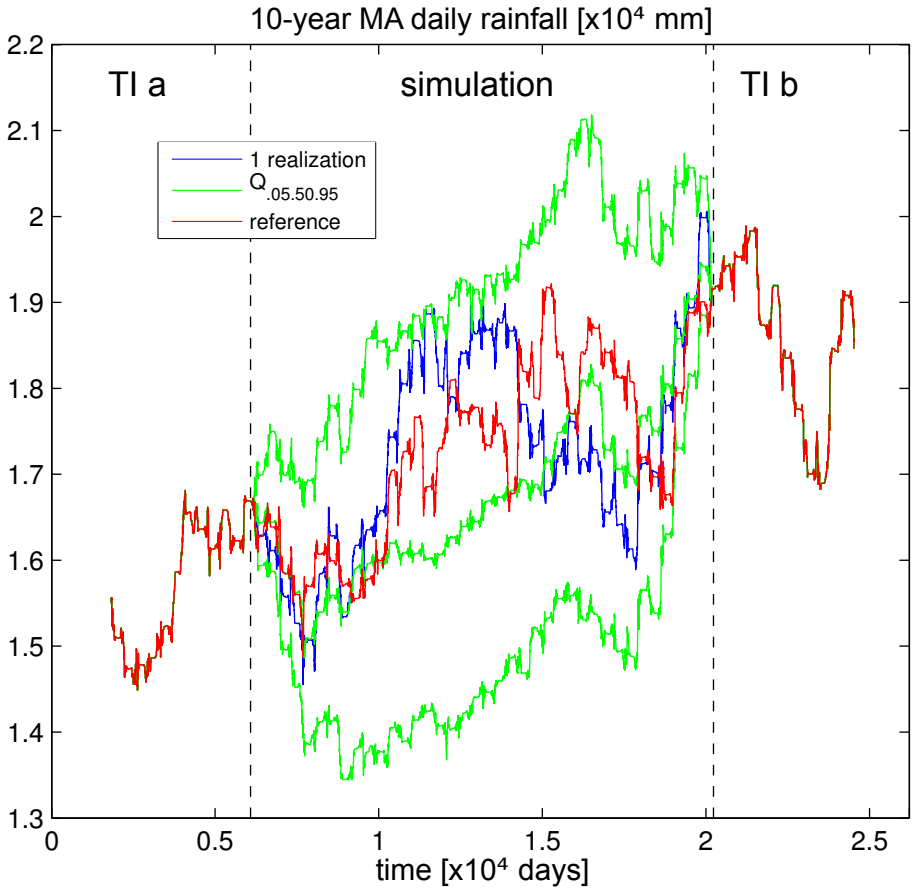


Figure D.1: Time-series of the 10-year Moving Average (MA) [mm]: reference data (red line), median, 5th and 95th percentile of 100 realizations (green line) and one example of realization (blue line). The dashed lines divide the data used as training image (TIa and TIb) from the simulated time span.

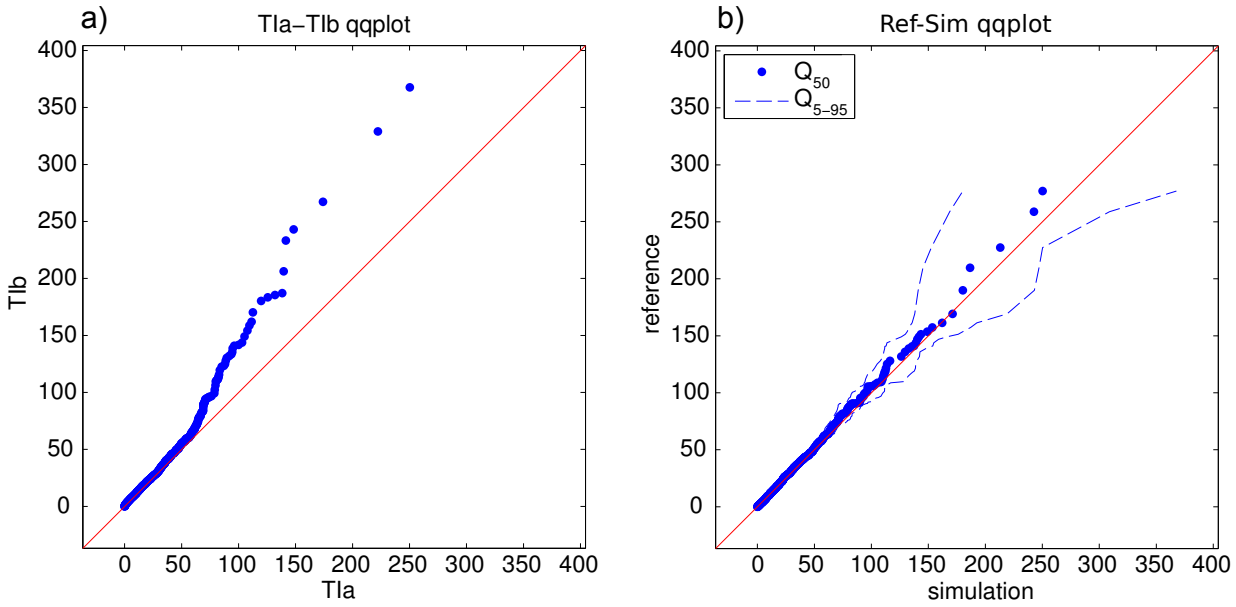


Figure D.2: qq-plots of the daily rainfall amount [mm] between: a) TIa and TIb and b) The simulation ensemble and the reference. In b) it is shown: the median of the realizations (blue dots), 5th and 95th percentiles (dashed lines). The bisector (solid line) indicates the exact quantile match.

Appendix E

Simple shear deformation of a spatio-temporal field

In this appendix, the deformation applied to the spatio-temporal rainfall field used as reference model in chapter 7 is described, focusing on the anisotropy ellipse derived by the deformation. The starting field is defined in the space (x', y', t') and the deformation applied is $(x', y', t') \rightarrow (x, y, t)$, whith $x = x' + v_x t$, $y' = y$ and $t' = t$. Since the field is isotropic in space and the deformation involves x and t, we consider for simplicity the subspace (x', t') deformed into (x, t). The aim of applying the deformation is to simulate a rainfall cell displacement along x with a constant velocity v_x. Figure E.1 depicts the original anisotropy ellipse describing the correlation length of the stochastic field. The deformation is a simple shear along x' affecting the correlation length along t: the correlation length along t' (l'_t) becomes l_t in the deformed space (figure E.1). Since the deformation is uniform along x', the correlation length along this direction remains constant and unvaried $(l_x = l_{x'})$.

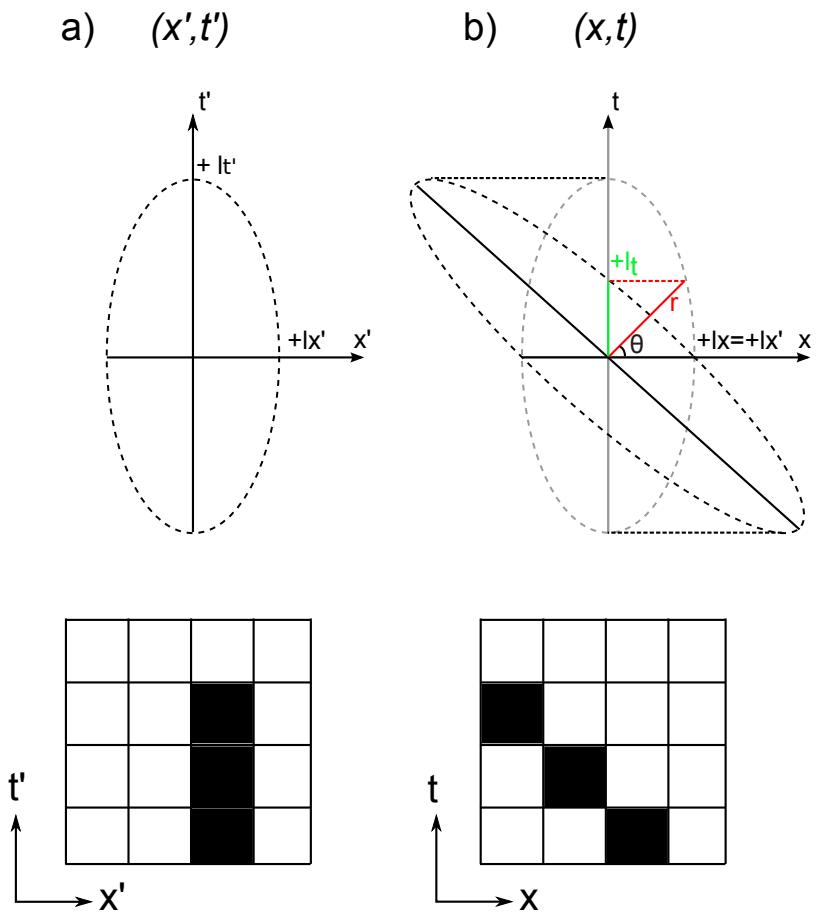


Figure E.1: Drawing of the anisotropy ellipse showing the principal correlation axis for the original (a) and deformed space (b). On the bottom, a sketch of the heterogeneity before and after the deformation.

The parameterization of the original space (x', t') is functional to its application, as explained in the following. In the simulation technique presented in chapter 7, a rainfall time-series $R(t)$ is extracted from the field (x, t) and projected into space by applying the transformation $R_s(x) = R(tv_x)$. To facilitate this operation, the velocity is set unitary with respect to the model discretization $(v_x = 1)$. Since, after the shear deformation, l_t is desired to approach l_x, the original correlation length $l_{t'}$ is opportunely set to achieve this goal. Referring to figure E.1, l_t is represented by the intersection of the deformed ellipse with the t axis. l_t is computed by projecting on t the semi-axis r which draws the shear angle θ with

x. Using the formula to compute the polar coordinate of the ellipse, r is computed as:

$$r = \frac{ab}{\sqrt{a^2 \sin^2 \theta + b^2 \cos^2 \theta}} \tag{E.1}$$

where a and b are the principal semi-axis. Referring to the field ellipse before the deformation, we have $a = l_{x'}$, $b = l_{t'}$ and $\theta = 45°$, which is the shear angle created with unitary velocity v_x. Considering $l_t = r \sin \theta = r/\sqrt{2}$ and substituting r with equation E.1, we obtain:

$$l_t = \frac{l_{x'} l_{t'}}{\sqrt{l_{x'}^2 + l_{t'}^2}}. \tag{E.2}$$

Since $l_{x'} = l_x$ and $l_{t'} < \sqrt{l_{x'}^2 + l_{t'}^2}$, we have $l_t < l_x$ and $\lim_{l_{t'} \to \infty} l_t = l_x$. Therefore, to approach $l_t = l_x$, l'_t should be set much greater than l_x. In practice, this corresponds to an increase of the rainfall cell life time. The value $l_{t'} = l_{x'}\sqrt{2}$ chosen for the application in chapter 7 is a compromise to have rainfall cells with a realistic life-time and a similar correlation length in time and space ($l_t = 0.82 l_x$).

Bibliography

[1] World Meteorological Organization - WMO radar database http://wrd.mgm.gov.tr/statistics/countries.aspx?l=en, 2015.

[2] A. Aghakouchak, E. Habib, and A. Bardossy. Modeling radar rainfall estimation uncertainties: Random error model. *Journal of Hydrologic Engineering*, 15(4):265–274, Apr. 2010.

[3] D. Allard and M. Bourotte. Disaggregating daily precipitations into hourly values with a transformed censored latent gaussian process. *Stochastic Environmental Research and Risk Assessment*, 29(2):453–462, Feb. 2015.

[4] D. Allard, R. Froidevaux, and P. Biver. Conditional simulation of multi-type non stationary markov object models respecting specified proportions. *Mathematical Geology*, 38(8):959–986, 2006.

[5] D. J. Allcroft and C. A. Glasbey. A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society Series C-applied Statistics*, 52:487–498, 2003.

[6] K. Aminian and S. Ameri. Application of artificial neural networks for reservoir characterization with limited data. *Journal of Petroleum Science and Engineering*, 49(3-4):212–222, Dec. 2005. WOS:000234302300009.

[7] A. Anandhi, A. Frei, D. C. Pierson, E. M. Schneiderman, M. S. Zion, D. Lounsbury, and A. H. Matonse. Examination of change factor methodologies for climate change impact assessment. *Water Resources Research*, 47:W03501, Mar. 2011.

[8] C. Andrade, R. M. Trigo, M. C. Freitas, M. C. Gallego, P. Borges, and A. M. Ramos. Comparing historic records of storm frequency and the north atlantic oscillation (NAO) chronology for the azores region. *Holocene*, 18(5):745–754, Aug. 2008.

[9] S. Apipattanavis, G. Podesta, B. Rajagopalan, and R. W. Katz. A semiparametric multivariate and multisite weather generator. *Water Resources Research*, 43(11):W11401, Nov. 2007.

[10] D. Argueso, J. Manuel Hidalgo-Munoz, S. Raquel Gamiz-Fortis, M. Jesus Esteban-Parra, and Y. Castro-Diez. Evaluation of WRF mean and extreme precipitation over spain: Present climate (1970-99). *Journal of Climate*, 25(14):4883–4897, July 2012.

[11] P. Arnaud, C. Bouvier, L. Cisneros, and R. Dominguez. Influence of rainfall spatial variability on flood prediction. *Journal of Hydrology*, 260(1-4):PII S0022–1694(01)00611–4, Mar. 2002.

[12] G. Arpat and J. Caers. Conditional simulation with patterns. *Mathematical Geology*, 39(2):177–203, 2007.

[13] A. Azimizonooz, W. F. Krajewski, D. S. Bowles, and D. J. Seo. Spatial rainfall estimation by linear and non-linear co-kriging of radar-rainfall and raingage data. *Stochastic Hydrology and Hydraulics*, 3(1):51–67, 1989.

[14] J. Bahrami, M. R. Kavianpour, M. S. Abdi, A. Telvari, K. Abbaspour, and B. Rouzkhash. A comparison between artificial neural network method and nonlinear regression method to estimate the missing hydrometric data. *Journal of Hydroinformatics*, 13(2):245–254, Apr. 2011. WOS:000289376800008.

[15] I. Bamberger, L. Hoertnagl, M. Walser, A. Hansel, and G. Wohlfahrt. Gap-filling strategies for annual VOC flux data sets. *Biogeosciences*, 11(8):2429–2442, 2014. WOS:000335374200023.

[16] A. Bardossy and G. Pegram. Interpolation of precipitation under topographic influence at different time scales. *Water Resources Research*, 49(8):4545–4565, Aug. 2013.

[17] A. Bardossy and G. G. S. Pegram. Copula based multisite model for daily precipitation simulation. *Hydrology and Earth System Sciences*, 13(12):2299–2314, 2009.

[18] A. Bardossy and E. J. Plate. Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, 28(5):1247–1259, May 1992.

[19] M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):pp. 27–41, 1946.

[20] S. Basu and H. I. Andharia. The chaotic time-series of indian monsoon rainfall and its prediction. *Proceedings of the Indian Academy of Sciences-earth and Planetary Sciences*, 101(1):27–34, Mar. 1992.

[21] S. Bennis, F. Berrada, and N. Kang. Improving single-variable and multivariable techniques for estimating missing hydrological data. *Journal of Hydrology*, 191(1-4):87–105, Apr. 1997. WOS:A1997XG47200005.

[22] G. E. Box and G. M. Jenkins. *Time series analysis, control, and forecasting.* San Francisco, CA: Holden Day, 1976.

[23] W. M. Briggs and D. S. Wilks. Estimating monthly and seasonal distributions of temperature and precipitation using the new cpc long-range forecasts. *Journal of Climate*, 9(4):818–826, Apr. 1996.

[24] T. A. Buishand and T. Brandsma. Multisite simulation of daily precipitation and temperature in the rhine basin by nearest-neighbor resampling. *Water Resources Research*, 37(11):2761–2776, Nov. 2001.

[25] J. Bullock, O. Russell, K. Alapaty, J. A. Herwehe, M. S. Mallard, T. L. Otte, R. C. Gilliam, and C. G. Nolte. An observation-based investigation of nudging in wrf for downscaling surface climate information to 12-km grid spacing. *Journal of Applied Meteorology and Climatology*, 53(1):20–33, Jan. 2014.

[26] C2SM, MeteoSwiss, ETH, N. Climate, and OcCC. Swiss climate change scenarios ch2011. Zurich, Switzerland, 2011.

[27] P. Caldwell, H.-N. S. Chin, D. C. Bader, and G. Bala. Evaluation of a wrf dynamical downscaling simulation over california. *Climatic Change*, 95(3-4):499–521, Aug. 2009.

[28] F. Castellvi, C. O. Stockle, I. Mormeneo, and J. M. Villar. Testing the performance of different processes to generate temperature and solar radiation: A case study at lleida (northeast spain). *Transactions of the ASAE*, 45(3):571–580, May 2002.

[29] A. Chappell, L. J. Renzullo, T. H. Raupach, and M. Haylock. Evaluating geostatistical methods of blending satellite and gauge data to estimate near real-time daily rainfall for australia. *Journal of Hydrology*, 493:105–114, June 2013.

[30] A. Chipperfield, P. Fleming, and C. Fonseca. Genetic algorithm tools for control systems engineering. In *Proc. Adaptive Computing in Engineering Design and Control*, pages 128–133. Citeseer, 1994.

[31] C. Chou, J. Y. Tu, and J. Y. Yu. Interannual variability of the western north pacific summer monsoon: Differences between ENSO and non-ENSO years. *Journal of Climate*, 16(13):2275–2287, July 2003.

[32] T. L. Chugunova and L. Y. Hu. Multiple-point simulations constrained by continuous auxiliary data. *Mathematical Geosciences*, 40(2):133–146, Feb. 2008.

[33] M. P. Clark, S. Gangopadhyay, D. Brandon, K. Werner, L. Hay, B. Rajagopalan, and D. Yates. A resampling procedure for generating conditioned daily weather sequences. *Water Resources Research*, 40(4):W04304, Apr. 2004.

[34] R. Coe and R. D. Stern. Fitting models to daily rainfall data. *Journal of Applied Meteorology*, 21(7):1024–1031, 1982.

[35] A. Comunian, P. Renard, and J. Straubhaar. 3D multiple-point statistics simulation using 2D training images. *Computers & Geosciences*, 40:49–65, Mar. 2012.

[36] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, Sept. 2007.

[37] A. C. Costa and A. Soares. Trends in extreme precipitation indices derived from a daily rainfall database for the south of Portugal. *International Journal of Climatology*, 29(13):1956–1975, Nov. 2009.

[38] P. S. P. Cowpertwait, P. E. OConnell, A. V. Metcalfe, and J. A. Mawdsley. Stochastic point process modelling of rainfall .2. regionalisation and disaggregation. *Journal of Hydrology*, 175(1-4):47–65, Feb. 1996.

[39] J. Creutin, G. Delrieu, and T. Lebel. Rain measurement by raingage-radar combination: a geostatistical approach. *Journal of Atmospheric and Oceanic Technology*, 5(1):102–115, 1988.

[40] Q. Dai, D. Han, M. A. Rico-Ramirez, and T. Islam. Modelling radar-rainfall estimation uncertainties using elliptical and archimedean copulas with different marginal distributions. *Hydrological Sciences Journal*, 59(11):1992–2008, Sept. 2014.

[41] T. Das, A. Bardossy, E. Zehe, and Y. He. Comparison of conceptual model performance using different representations of spatial variability. *Journal of Hydrology*, 356(1-2):106–118, July 2008.

[42] M. T. Dastorani, A. Moghadamnia, J. Piri, and M. Rico-Ramirez. Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment*, 166(1-4):421–434, June 2009.

[43] R. Deidda. Rainfall downscaling in a space-time multifractal framework. *Water Resources Research*, 36(7):1779–1794, July 2000.

[44] R. Deidda, M. G. Badas, and E. Piga. Space-time multifractality of remotely sensed rainfall fields. *Journal of Hydrology*, 322(1-4):2–13, May 2006.

[45] G. Dumedah, J. P. Walker, and L. Chik. Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. *Journal of Hydrology*, 515:330–344, July 2014. WOS:000338605900030.

[46] B. Efron. Bootstrap methods - another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[47] M. H. Elsanabary, T. Y. Gan, and D. Mwale. Application of wavelet empirical orthogonal function analysis to investigate the nonstationary character of ethiopian rainfall and its teleconnection to nonstationary global sea surface temperature variations for 1900-1998. *International Journal of Climatology*, 34(6):1798–1813, May 2014.

[48] E. Falge, D. Baldocchi, R. Olson, P. Anthoni, M. Aubinet, C. Bernhofer, G. Burba, R. Ceulemans, R. Clement, H. Dolman, A. Granier, P. Gross, T. Grnwald, D. Hollinger, N.-O. Jensen, G. Katul, P. Keronen, A. Kowalski, C. T. Lai, B. E. Law, T. Meyers, J. Moncrieff, E. Moors, J. W. Munger, K. Pilegaard, . Rannik, C. Rebmann, A. Suyker, J. Tenhunen, K. Tu, S. Verma, T. Vesala, K. Wilson, and S. Wofsy. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology*, 107(1):43–69, Mar. 2001.

[49] D. Faure, G. Delrieu, P. Tabary, J. P. Du Chatelet, and M. Guimera. Application of the hydrologic visibility concept to estimate rainfall measurement quality of two planned weather radars. *Atmospheric Research*, 77(1-4):232–246, Sept. 2005.

[50] J. Faures, D. Goodrich, D. Woolhiser, and S. Sorooshian. Impact of Small-Scale Spatial Rainfall Variability on Runoff Modeling. *Journal of Hydrology*, 173(1-4):309–326, Dec. 1995. WOS:A1995TG71300016.

[51] X. Feng and L. ChangZheng. The influence of moderate enso on summer rainfall in eastern china and its comparison with strong enso. *Chinese Science Bulletin*, 53(5):791–800, Mar. 2008.

[52] P. Fiorucci, P. La Barbera, L. G. Lanza, and R. Minciardi. A geostatistical approach to multisensor rain field reconstruction and downscaling. *Hydrology and Earth System Sciences*, 5(2):European Geophys Soc, June 2001.

[53] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[54] K. Gabriel and J. Neumann. A markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375):90–95, 1962.

[55] L. Garcia-Barron, M. Aguilar, and A. Sousa. Evolution of annual rainfall irregularity in the southwest of the iberian peninsula. *Theoretical and Applied Climatology*, 103(1-2):13–26, Jan. 2011.

[56] U. Germann, M. Berenguer, D. Sempere-Torres, and M. Zappa. Real - ensemble radar precipitation estimation for hydrology in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 135(639):445–456, Jan. 2009.

[57] P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1-2):113–129, Feb. 2000.

[58] I. P. Gorenburg, D. McLaughlin, and D. Entekhabi. Scale-recursive assimilation of precipitation data. *Advances In Water Resources*, 24(9-10):941–953, Nov. 2001.

[59] D. I. F. Grimes and E. Pardo-Iguzquiza. Geostatistical analysis of rainfall. *Geographical Analysis*, 42(2):136–160, Apr. 2010.

[60] D. I. F. Grimes, E. Pardo-Iguzquiza, and R. Bonifacio. Optimal areal rainfall estimation using raingauges and satellite data. *Journal of Hydrology*, 222(1):93–108, 1999.

[61] F. Guardiano and R. Srivastava. Multivariate geostatistics: beyond bivariate moments. *Geostatistics-Troia*, 1:133–144, 1993.

[62] G. Guillot and T. Lebel. Disaggregation of sahelian mesoscale convective system rain fields: Further developments and validation. *Journal of Geophysical Research-atmospheres*, 104(D24):Amer Geophys Union; Amer Meterol Soc; Mauna Lani Bay Hotel & Bungalows, Dec. 1999.

[63] V. K. Gupta and E. C. Waymire. Stochastic kinematic study of subsynoptic space-time rainfall. *Water Resources Research*, 15(3):637–644, 1979.

[64] V. K. Gupta and E. C. Waymire. A statistical-analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology*, 32(2):Amer Meteol Soc; Amer Geophys Union; Nasa; Natl Sci Fdn; Texas Insteoleoloceanog, Feb. 1993.

[65] U. Haberlandt. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. *Journal of Hydrology*, 332(1-2):144–157, Jan. 2007.

[66] M. Han and Y. Wang. Analysis and modeling of multivariate chaotic time series based on neural network. *Expert Systems With Applications*, 36(2):1280–1290, Mar. 2009.

[67] T. I. Harrold, A. Sharma, and S. J. Sheather. A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, 39(12):1343, Dec. 2003.

[68] T. I. Harrold, A. Sharma, and S. J. Sheather. A nonparametric model for stochastic generation of daily rainfall occurrence. *Water Resources Research*, 39(10):1300, Oct. 2003.

[69] M. M. Hasan and P. K. Dunn. A simple poisson-gamma model for modelling rainfall occurrence and amount simultaneously. *Agricultural and Forest Meteorology*, 150(10):1319–1330, Sept. 2010.

[70] L. E. Hay, G. J. Mccabe, D. M. Wolock, and M. A. Ayers. Simulation of precipitation by weather type analysis. *Water Resources Research*, 27(4):493–501, Apr. 1991.

[71] X. He, F. Vejen, S. Stisen, T. O. Sonnenborg, and K. H. Jensen. An operational weather radar-based quantitative precipitation estimation and its application in catchment water resources modeling. *Vadose Zone Journal*, 10(1):8–24, Feb. 2011.

[72] J. A. Hevesi, J. D. Istok, and A. L. Flint. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. *Journal of applied meteorology*, 31(7):661–676, 1992.

[73] M. Honarkhah and J. Caers. Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences*, 42(5):487–517, 2010.

[74] K.-O. Hong, M.-S. Suh, and D.-K. Rha. Temporal and spatial variations of precipitation in south korea for recent 30 years(1976-2005) geographic environments. *The Journal of The Korean Earth Science Society*, 27(4):433–449, 2006.

[75] Y. Hong, K. L. Hsu, S. Sorooshian, and X. G. Gao. Improved representation of diurnal variability of rainfall retrieved from the tropical rainfall measurement mission microwave imager adjusted precipitation estimation from remotely sensed information using artificial neural networks (persiann) system. *Journal of Geophysical Research-Atmospheres*, 110(D6):D06102, Mar. 2005.

[76] L. Y. Hu, Y. Liu, C. Scheepens, A. W. Shultz, and R. D. Thompson. Multiple-point simulation with an existing reservoir model as training image. *Mathematical Geosciences*, 46(2):227–240, Feb. 2014.

[77] J. Hughes and P. Guttorp. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research*, 30(5):1535–1546, 1994.

[78] J. Hughes, P. Guttorp, and S. Charles. A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30, 1999.

[79] G. J. Husak, J. Michaelsen, and C. Funk. Use of the gamma distribution to represent monthly rainfall in africa for drought monitoring applications. *International Journal of Climatology*, 27(7):935–944, June 2007.

[80] a. w. Jayawardena and f. z. lai. Analysis and prediction of chaos in rainfall and stream-flow time-series. *Journal of Hydrology*, 153(1-4):23–52, Jan. 1994.

[81] 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, Jan. 2013.

[82] S. K. Jha, G. Mariethoz, and B. F. J. Kelly. Bathymetry fusion using multiple-point geostatistics: Novelty and challenges in representing non-stationary bedforms. *Environmental Modelling & Software*, 50:66–76, Dec. 2013.

[83] B. Johansson and D. L. Chen. The influence of wind and topography on precipitation distribution in sweden: Statistical analysis and modelling. *International Journal of Climatology*, 23(12):1523–1535, Oct. 2003.

[84] P. G. Jones and P. K. Thornton. Spatial and temporal variability of rainfall related to a third-order markov model. *Agricultural and Forest Meteorology*, 86(1-2):127–138, Aug. 1997.

[85] V. Jothiprakash and T. A. Fathima. Chaotic analysis of daily rainfall series in koyna reservoir catchment area, india. *Stochastic Environmental Research and Risk Assessment*, 27(6):1371–1381, Aug. 2013.

[86] R. J. Joyce, J. E. Janowiak, P. A. Arkin, and P. P. Xie. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology*, 5(3):487–503, June 2004.

[87] e. Kaplan and p. Meier. Non-parametric estimation from incomplete observations. *Journal of the american statistical association*, 53(282):457–481, 1958.

[88] R. Katz and M. Parlange. Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11(4):591–601, 1998.

[89] R. W. Katz and M. B. Parlange. Effects of an index of atmospheric circulation on stochastic properties of precipitation. *Water Resources Research*, 29(7):2335–2344, July 1993.

[90] R. W. Katz and X. G. Zheng. Mixture model for overdispersion of precipitation. *Journal of Climate*, 12(8):2528–2537, Aug. 1999.

[91] O. A. Kelley. Where the least rainfall occurs in the sahara desert, the trmm radar reveals a different pattern of rainfall each season. *Journal of Climate*, 27(18):6919–6939, Sept. 2014.

[92] S. Khan, A. R. Ganguly, and S. Saigal. Detection and predictive modeling of chaos in finite hydrological time series. *Nonlinear Processes In Geophysics*, 12(1):41–53, 2005.

[93] G. Kiely, J. D. Albertson, M. B. Parlange, and R. W. Katz. Conditioning stochastic properties of daily precipitation on indices of atmospheric circulation. *Meteorological Applications*, 5(1):75–87, 1998.

[94] C. G. Kilsby, P. D. Jones, A. Burton, A. C. Ford, H. J. Fowler, C. Harpham, P. James, A. Smith, and R. L. Wilby. A daily weather generator for use in climate change studies. *Environmental Modelling & Software*, 22(12):1705–1719, Dec. 2007.

[95] S. Kim, Y. Tachikawa, T. Sayama, and K. Takara. Ensemble flood forecasting with stochastic radar image extrapolation and a distributed hydrologic model. *Hydrological Processes*, 23(4):597–611, Feb. 2009.

[96] T. W. Kim and H. Ahn. Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. *Stochastic Environmental Research and Risk Assessment*, 23(3):367–376, Mar. 2009.

[97] L. M. King, A. I. McLeod, and S. P. Simonovic. Simulation of historical temperatures using a multi-site, multivariate block resampling algorithm with perturbation. *Hydrological Processes*, 28(3):905–912, Jan. 2014.

[98] W. Kleiber, R. W. Katz, and B. Rajagopalan. Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes. *Water Resources Research*, 48:W01523, Jan. 2012.

[99] M. Knoche, C. Fischer, E. Pohl, P. Krause, and R. Merz. Combined uncertainty of hydrological model complexity and satellite-based forcing data evaluated in two data-scarce semi-arid catchments in ethiopia. *Journal of Hydrology*, 519:2049–2066, Nov. 2014.

[100] D. Kondrashov, R. Denton, Y. Y. Shprits, and H. J. Singer. Reconstruction of gaps in the past history of solar wind parameters. *Geophysical Research Letters*, 41(8):2702–2707, Apr. 2014. WOS:000335809800006.

[101] W. F. Krajewski. Cokriging radar-rainfall and rain-gauge data. *Journal of Geophysical Research-atmospheres*, 92(D8):9571–9580, Aug. 1987.

[102] P. Kumar and E. Foufoula-georgiou. A multicomponent decomposition of spatial rainfall fields .1. segregation of large-scale and small-scale features using wavelet transforms. *Water Resources Research*, 29(8):2515–2532, Aug. 1993.

[103] P. Kumar and E. Foufoula-georgiou. A new look at rainfall fluctuations and scaling properties of spatial rainfall using orthogonal wavelets. *Journal of Applied Meteorology*, 32(2):209–222, Feb. 1993.

[104] H.-H. Kwon, U. Lall, and A. F. Khalil. Stochastic simulation model for nonstationary time series using an autoregressive wavelet decomposition: Applications to rainfall and temperature. *Water Resources Research*, 43(5):W05407, May 2007.

[105] P. C. Kyriakidis, J. Kim, and N. L. Miller. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. *Journal of Applied Meteorology*, 40(11):1855–1877, 2001.

[106] P. C. Kyriakidis, N. L. Miller, and J. Kim. A spatial time series framework for simulating daily precipitation at regional scales. *Journal of Hydrology*, 297(1-4):236–255, Sept. 2004.

[107] J. Kysely and M. Dubrovsky. Simulation of extreme temperature events by a stochastic weather generator: Effects of interdiurnal and interannual variability reproduction. *International Journal of Climatology*, 25(2):251–269, Feb. 2005.

[108] H. Kyznarova and P. Novak. CELLTRACK - convective cell tracking algorithm and its use for deriving life cycle characteristics. *Atmospheric Research*, 93(1-3):317–327, July 2009.

[109] U. Lall, B. Rajagopalan, and D. G. Tarboton. A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resources Research*, 32(9):2803–2823, Sept. 1996.

[110] U. Lall and A. Sharma. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, 32(3):679–693, Mar. 1996.

[111] B. Lamrini, E.-K. Lakhal, M.-V. Le Lann, and L. Wehenkel. Data validation and missing data reconstruction using self-organizing map for water treatment. *Neural Computing & Applications*, 20(4):575–588, June 2011. WOS:000290319100013.

[112] L. G. Lanza. A conditional simulation model of intermittent rain fields. *Hydrology and Earth System Sciences*, 4(1):173–183, Mar. 2000.

[113] E. Leblois and J.-D. Creutin. Space-time simulation of intermittent rainfall with prescribed advection field: Adaptation of the turning band method. *Water Resources Research*, 49(6):3375–3387, June 2013.

[114] T. Lee, T. B. M. J. Ouarda, and C. Jeong. Nonparametric multivariate weather generator and an extreme value theory for bandwidth selection. *Journal of Hydrology*, 452:161–171, July 2012.

[115] D. R. Legates and G. J. McCabe. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1):233–241, Jan. 1999. WOS:000078123800021.

[116] Y. Li, W. Cai, and E. P. Campbell. Statistical modeling of extreme rainfall in southwest western australia. *Journal of Climate*, 18(6):852–863, Mar. 2005.

[117] L. Liao, R. Meneghini, and T. Iguchi. Simulations of mirror image returns of air/space-borne radars in rain and their applications in estimating path attenuation. *Ieee Transactions On Geoscience and Remote Sensing*, 37(2):1107–1121, Mar. 1999.

[118] S. Lovejoy and B. B. Mandelbrot. Fractal properties of rain, and a fractal model. *Tellus Series A-dynamic Meteorology and Oceanography*, 37(3):209–232, 1985.

[119] S. Lovejoy and D. Schertzer. Generalized scale-invariance in the atmosphere and fractal models of rain. *Water Resources Research*, 21(8):1233–1250, 1985.

[120] G. Mariethoz, M. F. McCabe, and P. Renard. Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach. *Water Resources Research*, 48:W10507, Oct. 2012. WOS:000309608800003.

[121] G. Mariethoz and P. Renard. Reconstruction of incomplete data sets or images using direct sampling. *Mathematical Geosciences*, 42(3):245–268, Apr. 2010.

[122] G. Mariethoz, P. Renard, and J. Straubhaar. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46:W11536, Nov. 2010.

[123] D. Marsan, D. N. Schertzer, and S. Lovejoy. Causal space-time multifractal processes: Predictability and forecasting of rain fields. *Journal of Geophysical Research-atmospheres*, 101(D21):26333–26346, Nov. 1996.

[124] E. Meerschman, G. Pirot, G. Mariethoz, J. Straubhaar, M. Van Meirvenne, and P. Renard. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. *Computers & Geosciences*, 52:307–324, Mar. 2013.

[125] R. Mehrotra and A. Sharma. A nonparametric stochastic downscaling framework for daily rainfall at multiple locations. *Journal of Geophysical Research-atmospheres*, 111(D15):D15101, Aug. 2006.

[126] R. Mehrotra and A. Sharma. Preserving low-frequency variability in generated daily rainfall sequences. *Journal of Hydrology*, 345(1-2):102–120, Oct. 2007.

[127] R. Mehrotra and A. Sharma. A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low-frequency variability. *Journal of Hydrology*, 335(1-2):180–193, Mar. 2007.

[128] R. Mehrotra, R. Srikanthan, and A. Sharma. A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology*, 331(1-2):280–292, Nov. 2006.

[129] M. Menabde, D. Harris, A. Seed, G. Austin, and D. Stow. Multiscaling properties of rainfall and bounded random cascades. *Water Resources Research*, 33(12):2823–2830, Dec. 1997.

[130] M. Menabde, A. Seed, D. Harris, and G. Austin. Self-similar random fields and rainfall simulation. *Journal of Geophysical Research-atmospheres*, 102(D12):13509–13515, June 1997.

[131] H. Millan, J. Rodriguez, B. Ghanbarian-Alavijeh, R. Biondi, and G. Llerena. Temporal complexity of daily precipitation records from different atmospheric environments: Chaotic and levy stable parameters. *Atmospheric Research*, 101(4):879–892, Sept. 2011.

[132] A. M. Moffat, D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H. Braswell, G. Churkina, A. R. Desai, E. Falge, J. H. Gove, M. Heimann, D. Hui, A. J. Jarvis, J. Kattge, A. Noormets, and V. J. Stauch. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3-4):209–232, Dec. 2007. WOS:000251469900009.

[133] F. J. Moral. Comparison of different geostatistical approaches to map climate variables: application to precipitation. *International Journal of Climatology*, 30(4):620–631, Mar. 2010.

[134] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the Asabe*, 50(3):885–900, June 2007. WOS:000248036800021.

[135] D. Munoz-Diaz and F. S. Rodrigo. Effects of the north atlantic oscillation on the probability for climatic categories of local monthly rainfall in southern spain. *International Journal of Climatology*, 23(4):381–397, Mar. 2003.

[136] J. Ndiritu. A variable-length block bootstrap method for multi-site synthetic streamflow generation. *Hydrological Sciences Journal*, 56(3):362–379, 2011.

[137] T. R. Nkuna and J. O. Odiyo. Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks. *Physics and Chemistry of the Earth*, 36(14-15):830–835, 2011. WOS:000296306100013.

[138] P. Northrop. *Modelling and statistical analysis of spatialtemporal rainfall fields.* PhD thesis, Department of Statistical Science, University College London, 1996.

[139] V. Nourani, A. H. Baghanam, and M. Gebremichael. Investigating the Ability of Artificial Neural Network (ANN) Models to Estimate Missing Rain-gauge Data. *Journal of Environmental Informatics*, 19(1):38–50, Mar. 2012. WOS:000302846300005.

[140] A. Ochoa, L. Pineda, P. Crespo, and P. Willems. Evaluation of trmm 3b42 precipitation estimates and wrf retrospective precipitation simulation over the pacific-andean region of ecuador and peru. *Hydrology and Earth System Sciences*, 18(8):3179–3193, 2014.

[141] F. Oriani, J. Straubhaar, G. Mariethoz, and P. Renard. Spatial rainfall simulation: Trading time for space with multiple point statistics. In *AGU Fall Meeting*, 2014.

[142] F. Oriani, J. Straubhaar, P. Renard, and G. Mariethoz. Modeling future climatic time-series according to climate change scenarios. In *The Water Cycle in a Changing Climate*, 2013.

[143] 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.

[144] T. M. Over and V. K. Gupta. Statistical-analysis of mesoscale rainfall - dependence of a random cascade generator on large-scale forcing. *Journal of Applied Meteorology*, 33(12):Amer Meteorol Soc; Amer Geophys Union, Dec. 1994.

[145] T. M. Over and V. K. Gupta. A space-time theory of mesoscale rainfall using random cascades. *Journal of Geophysical Research-atmospheres*, 101(D21):26319–26331, Nov. 1996.

[146] A. Paschalis, P. Molnar, S. Fatichi, and P. Burlando. A stochastic model for high-resolution space-time precipitation simulation. *Water Resources Research*, 49(12):8400–8417, Dec. 2013.

[147] A. Pathirana and S. Herath. Multifractal modelling and simulation of rain fields exhibiting spatial heterogeneity. *Hydrology and Earth System Sciences*, 6(4):695–708, Aug. 2002.

[148] G. G. S. Pegram and A. N. Clothier. High resolution space-time modelling of rainfall: the "string of beads" model. *Journal of Hydrology*, 241(1-2):26–41, Jan. 2001.

[149] N. Peleg and E. Morin. Stochastic convective rain-field simulation using a high-resolution synoptically conditioned weather generator (hires-wg). *Water Resources Research*, 50(3):2124–2139, Mar. 2014.

[150] B. Peng, J. Shi, W. Ni-Meister, T. Zhao, and D. Ji. Evaluation of trmm multisatellite precipitation analysis (tmpa) products and their potential hydrological application at an arid and semiarid basin in china. *Ieee Journal of Selected Topics In Applied Earth Observations and Remote Sensing*, 7(9):3915–3930, Sept. 2014.

[151] J. R. Prairie, B. Rajagopalan, T. J. Fulp, and E. A. Zagona. Modified k-nn model for stochastic streamflow simulation. *Journal of Hydrologic Engineering*, 11(4):371–378, July 2006.

[152] J. L. Proud, K. K. Droegemeier, V. T. Wood, and R. A. Brown. Sampling strategies for tornado and mesocyclone detection using dynamically adaptive doppler radars: A simulation study. *Journal of Atmospheric and Oceanic Technology*, 26(3):492–507, Mar. 2009.

[153] A. Putthividhya and K. Tanaka. Optimal rain gauge network design and spatial precipitation mapping based on geostatistical analysis from co-located elevation and humidity data. *Chiang Mai Journal of Science*, 40(2):187–197, Apr. 2013.

[154] Y. Qin, Z. Chen, Y. Shen, S. Zhang, and R. Shi. Evaluation of satellite rainfall estimates over the chinese mainland. *Remote Sensing*, 6(11):11649–11672, Nov. 2014.

[155] B. Rajagopalan and U. Lall. A k-nearest-neighhor simulator for daily precipitation and other weather variables. *Water Resources Research*, 35(10):3089–3101, Oct. 1999.

[156] A. Randrianasolo, M. H. Ramos, G. Thirel, V. Andreassian, and E. Martin. Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters*, 11(2):100–107, Apr. 2010.

[157] U. F. A. Rauf and P. Zeephongsekul. Analysis of rainfall severity and duration in victoria, australia using non-parametric copulas and marginal distributions. *Water Resources Management*, 28(13):4835–4856, Oct. 2014.

[158] C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar-radiation. *Water Resources Research*, 17(1):182–190, 1981.

[159] T. Rientjes, A. T. Haile, and A. A. Fenta. Diurnal rainfall variability over the upper blue nile basin: A remote sensing based approach. *International Journal of Applied Earth Observation and Geoinformation*, 21:311–325, Apr. 2013.

[160] I. Rodriguez-iturbe, D. R. Cox, and P. S. Eagleson. Spatial modeling of total storm rainfall. *Proceedings of the Royal Society of London Series A-mathematical and Physical Sciences*, 403(1824):27–50, Jan. 1986.

[161] D. E. Rupp, P. Licznar, W. Adamowski, and M. Lesniewski. Multiplicative cascade models for fine spatial downscaling of rainfall: parameterization with rain gauge data. *Hydrology and Earth System Sciences*, 16(3):671–684, 2012.

[162] F. Russo, F. Lombardo, F. Napolitano, and E. Gorgucci. Rainfall stochastic modeling for runoff forecasting. *Physics and Chemistry of the Earth*, 31(18):European Geosci Union, 2006.

[163] J. F. Sanchez-Moreno, C. M. Mannaerts, and V. Jetten. Influence of topography on rainfall variability in santiago island, cape verde. *International Journal of Climatology*, 34(4):1081–1097, Mar. 2014.

[164] A. Sarangi, C. A. Cox, and C. A. Madramootoo. Geostatistical methods for prediction of spatial variability of rainfall in a mountainous region. *Transactions of the Asae*, 48(3):943–954, May 2005.

[165] D. Schertzer, I. Tchiguirinskaia, S. Lovejoy, P. Hubert, H. Bendjoudi, and M. Larcheveque. Discussion of "evidence of chaos in the rainfall-runoff process" - which chaos in the rainfall-runoff process? *Hydrological Sciences Journal-journal Des Sciences Hydrologiques*, 47(1):139–148, Feb. 2002.

[166] M. Schleiss, S. Chamoun, and A. Berne. Nonstationarity in intermittent rainfall: The "dry drift". *Journal of Hydrometeorology*, 15(3):1189–1204, June 2014.

[167] M. Schleiss, S. Chamoun, and A. Berne. Stochastic simulation of intermittent rainfall using the concept of "dry drift". *Water Resources Research*, 50(3):2329–2349, Mar. 2014.

[168] D. H. Schoellhamer. Singular spectrum analysis for time series with missing data. *Geophysical Research Letters*, 28(16):3187–3190, Aug. 2001. WOS:000170348100033.

[169] W. Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, 1992.

[170] M.-L. Segond, H. S. Wheater, and C. Onof. The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the lee catchment, uk. *Journal of Hydrology*, 347(1-2):116–131, Dec. 2007.

[171] D. J. Seo. Real-time estimation of rainfall fields using radar rainfall and rain gage data. *Journal of Hydrology*, 208(1-2):37–52, July 1998.

[172] D. J. Seo and J. P. Breidenbach. Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *Journal of Hydrometeorology*, 3(2):93–111, Apr. 2002.

[173] H. O. Sharif, F. L. Ogden, W. F. Krajewski, and M. Xue. Numerical simulations of radar rainfall error propagation. *Water Resources Research*, 38(8):1140, Aug. 2002.

[174] M. Sharif and D. H. Burn. Improved k-nearest neighbor weather generating model. *Journal of Hydrologic Engineering*, 12(1):42–51, Jan. 2007.

[175] A. Sharma and R. Mehrotra. Rainfall generation. In *Rainfall: state of the science*, number 191 in Geophysical Monograph Series, pages 215–246. AGU, Washington, D. C., 2010.

[176] A. Sharma, D. Tarboton, and U. Lall. Streamflow simulation: A nonparametric approach. *Water Resources Research*, 33(2):291–308, Feb 1997.

[177] I. V. Sideris, M. Gabella, R. Erdin, and U. Germann. Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of switzerland. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1097–1111, Apr. 2014.

[178] J. Simpson, R. F. Adler, and G. R. North. A proposed tropical rainfall measuring mission (trmm) satellite. *Bulletin of the American Meteorological Society*, 69(3):278–295, Mar. 1988.

[179] S. Sinclair and G. Pegram. Combining radar and rain gauge rainfall estimates using conditional merging. *Atmospheric Science Letters*, 6(1):19–22, Jan. 2005.

[180] N. Singhrattna, B. Rajagopalan, M. Clark, and K. K. Kumar. Seasonal forecasting of thailand summer monsoon rainfall. *International Journal of Climatology*, 25(5):649–664, Apr. 2005.

[181] B. Sivakumar, R. Berndtsson, J. Olsson, and K. Jinno. Evidence of chaos in the rainfall-runoff process. *Hydrological Sciences Journal-journal Des Sciences Hydrologiques*, 46(1):131–145, Feb. 2001.

[182] B. Sivakumar, S. Liong, and C. Liaw. Evidence of chotic behaviour in singapore rainfall. *JAWRA Journal of the American Water Resources Association*, 34(2):301–310, 1998.

[183] B. Sivakumar, F. M. Woldemeskel, and C. E. Puente. Nonlinear analysis of rainfall variability in australia. *Stochastic Environmental Research and Risk Assessment*, 28(1):17–27, Jan. 2014.

[184] W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, M. Duda, X. Huang, W. Wang, and J. Powers. A description of the advanced research wrf version 3. Technical report, NCAR, 2008.

[185] B.-J. So, H.-H. Kwon, D. Kim, and S. O. Lee. Modeling of daily rainfall sequence and extremes based on a semiparametric Pareto tail approach at multiple locations. Journal of Hydrology, In press.

[186] F. A. Souza and U. Lall. Seasonal to interannual ensemble streamflow forecasts for ceara, brazil: Applications of a multivariate, semiparametric algorithm. *Water Resources Research*, 39(11):1307, Nov. 2003.

[187] R. Srikanthan. Stochastic generation of daily rainfall data using a nested model. In *57th Canadian Water Resources Association Annual Congress*, pages 16–18, 2004.

[188] R. Srikanthan. Stochastic generation of daily rainfall data using a nested transition probability matrix model. In *29th Hydrology and Water Resources Symposium: Water Capital, 20-23 February 2005, Rydges Lakeside, Canberra*, page 26. Engineers Australia, 2005.

[189] R. Srikanthan and G. G. S. Pegram. A nested multisite daily rainfall stochastic generation model. *Journal of Hydrology*, 371(1-4):142–153, June 2009.

[190] R. Srikanthan, A. Sharma, and T. A. McMahon. Comparison of two nonparametric alternatives for stochastic generation of monthly rainfall. *Journal of Hydrologic Engineering*, 11(3):222–229, May 2006.

[191] V. V. Srinivas and K. Srinivasan. Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows. *Journal of Hydrology*, 302(1-4):307–330, Feb. 2005.

[192] S. Steinschneider and C. Brown. A semiparametric multivariate, multisite weather generator with low-frequency variability for use in climate risk assessments. *Water Resources Research*, 49(11):7205–7220, Nov. 2013.

[193] J. Straubhaar. *MPDS technical reference guide*. Centre d'Hydrogeologie et de Geothermie, University of Neuchâtel, September 2011.

[194] J. Straubhaar, P. Renard, and G. Mariethoz. Conditioning multiple point statistics simulations to block data using direct sampling. In *Geoenv 2014, Paris*, 2014.

[195] J. Straubhaar, P. Renard, G. Mariethoz, R. Froidevaux, and O. Besson. An improved parallel multiple-point algorithm using a list approach. *Mathematical Geosciences*, 43(3):305–328, 2011.

[196] S. Strebelle. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34(1):1–21, 2002.

[197] K. H. Syed, D. C. Goodrich, D. E. Myers, and S. Sorooshian. Spatial characteristics of thunderstorm rainfall fields and their relation to runoff. *Journal of Hydrology*, 271(1-4):PII S0022–1694(02)00311–6, Feb. 2003.

[198] P. Tahmasebi, A. Hezarkhani, and M. Sahimi. Multiple-point geostatistical modeling based on the cross-correlation functions. *Computational Geosciences*, 16(3):779–797, 2012.

[199] L. Telesca, V. Lapenna, E. Scalcione, and D. Summa. Searching for time-scaling features in rainfall sequences. *Chaos Solitons & Fractals*, 32(1):35–41, Apr. 2007.

[200] C.-K. Teo and D. I. F. Grimes. Stochastic modelling of rainfall from satellite data. *Journal of Hydrology*, 346(1-2):33–50, Nov. 2007.

[201] C. S. Thompson, P. J. Thomson, and X. Zheng. Fitting a multisite daily rainfall model to new zealand data. *Journal of Hydrology*, 340(1-2):25–39, June 2007.

[202] E. Todini. A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences*, 5(2):187–199, 1999.

[203] R. Trigo, J. L. Zezere, M. L. Rodrigues, and I. F. Trigo. The influence of the north atlantic oscillation on rainfall triggering of landslides near lisbon. *Natural Hazards*, 36(3):331–354, Nov. 2005.

[204] T. Vansteenkiste, M. Tavakoli, N. Van Steenbergen, F. De Smedt, O. Batelaan, F. Pereira, and P. Willems. Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. *Journal of Hydrology*, 511:335–349, Apr. 2014.

[205] C. A. Velasco-Forero, D. Sempere-Torres, E. F. Cassiraga, and J. Jaime Gomez-Hernandez. A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Advances In Water Resources*, 32(7):986–1002, July 2009.

[206] T. Vischel, T. Lebel, S. Massuel, and B. Cappelaere. Conditional simulation schemes of rain fields and their application to rainfall-runoff modeling studies in the sahel. *Journal of Hydrology*, 375(1-2):273–286, Aug. 2009.

[207] R. M. Vogel and A. L. Shallcross. The moving blocks bootstrap versus parametric time series models. *Water Resources Research*, 32(6):1875–1882, June 1996.

[208] T. W. R. Wallis and J. F. Griffiths. Simulated meteorological input for agricultural models. *Agricultural and Forest Meteorology*, 88(1-4):241–258, Dec. 1997.

[209] Q. J. Wang and R. J. Nathan. A daily and monthly mixed algorithm for stochastic generation of rainfall time series. In *Water Challenge: Balancing the Risks: Hydrology and Water Resources Symposium 2002*, page 698. Institution of Engineers, Australia, 2002.

[210] S. H. Wang. Application of self-organising maps for data mining with incomplete data sets. *Neural Computing & Applications*, 12(1):42–48, Sept. 2003. WOS:000186039600006.

[211] E. Waymire, V. K. Gupta, and I. Rodriguez-iturbe. A spectral theory of rainfall intensity at the meso-beta scale. *Water Resources Research*, 20(10):1453–1465, 1984.

[212] R. L. Wilby. Modelling low-frequency rainfall events using airflow indices, weather patterns and frontal frequencies. *Journal of Hydrology*, 212(1-4):380–392, Dec. 1998.

[213] D. S. Wilks. Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research*, 25(6):1429–1439, June 1989.

[214] D. S. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1-4):178–191, Sept. 1998.

[215] D. S. Wilks. Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, 96(1-3):85–101, Aug. 1999.

[216] C. B. Wilson, J. B. Valdes, and I. Rodriguez-iturbe. Influence of the spatial-distribution of rainfall on storm runoff. *Water Resources Research*, 15(2):321–328, 1979.

[217] R. Wojcik and T. A. Buishand. Simulation of 6-hourly rainfall and temperature by two resampling schemes. *Journal of Hydrology*, 273(1-4):69–80, Mar. 2003.

[218] R. Wojcik, D. McLaughlin, A. Konings, and D. Entekhabi. Conditioning stochastic rainfall replicates on remote sensing data. *IEEE T. Geosci. Remote*, 47(8):2436–2449, 2009.

[219] D. A. Woolhiser and D. C. Goodrich. Effect of storm rainfall intensity patterns on surface runoff. *Journal of Hydrology*, 102(1-4):335–354, Sept. 1988.

[220] D. A. Woolhiser, T. O. Keefer, and K. T. Redmond. Southern oscillation effects on daily precipitation in the southwestern united-states. *Water Resources Research*, 29(4):1287–1295, Apr. 1993.

[221] J. Yan and M. Gebremichael. Estimating actual rainfall from satellite rainfall products. *Atmospheric Research*, 92(4):481–488, June 2009.

[222] S.-S. Yoon and D.-H. Bae. Optimal rainfall estimation by considering elevation in the han river basin, south korea. *Journal of Applied Meteorology and Climatology*, 52(4):802–818, Apr. 2013.

[223] K. C. Young. A multivariate chain model for simulating climatic parameters from daily data. *Journal of Applied Meteorology*, 33(6):661–671, June 1994.

[224] M. P. Young, C. J. R. Williams, J. C. Chiu, R. I. Maidment, and S.-H. Chen. Investigation of discrepancies in satellite rainfall estimates over ethiopia. *Journal of Hydrometeorology*, 15(6):2347–2369, Dec. 2014.

[225] D. Zanchettin, S. W. Franks, P. Traverso, and M. Tomasino. On enso impacts on european wintertime rainfalls and their modulation by the nao and the pacific multi-decadal variability described through the pdo index. *International Journal of Climatology*, 28(8):995–1006, June 2008.

[226] I. I. Zawadzki. Statistical Properties of Precipitation Patterns. *Journal of Applied Meteorology*, 12(3):459–472, Apr. 1973.

[227] C. Zhang, Y. Wang, A. Lauer, and K. Hamilton. Configuration and evaluation of the wrf model for the study of hawaiian regional climate. *Monthly Weather Review*, 140(10):3259–3277, Oct. 2012.

[228] T. Zhang, P. Switzer, and A. Journel. Filter-based classification of training image patterns for spatial simulation. *Math. Geol.*, 38(1):63–80, 2006.