Stochastic simulation of near-surface atmospheric forcings for distributed hydrology

Sheng Chen

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Océan, Atmosphère, Hydrologie**

Arrêté ministériel : 7 Août 2006

Présentée par

Sheng CHEN

Thèse dirigée par **Etienne LEBLOIS**
et codirigée par **Sandrine ANQUETIN**

préparée au sein **de l'Unité de Recherche Hydrologie-Hydraulique (HHLY) de l'Institut National en Sciences et Technologies pour l'Environnement et l'Agriculture (Irstea, centre de Lyon-Villeurbanne)**
dans l'Ecole Doctorale **Terre Univers Environnement**

Stochastic simulation of near-surface atmospheric forcings for distributed hydrology

Thèse soutenue publiquement le **01 Février 2018**,
devant le jury composé de :

Mme Anne-Catherine FAVRE
Professeure, Université Grenoble-Alpes, CNRS, IRD, G-INP, IGE (Grenoble), Présidente

M. Alain MAILHOT
Professeur, INRS-Eau, Terre et Environnement (Montréal, Quebec, Canada), Rapporteur

M. András BÁRDOSSY
Professeur, University of Stuttgart (Allemagne), Rapporteur

M. Christophe BOUVIER
Directeur de recherches, IRD, HSM (Montpellier), Examinateur

M. Etienne LEBLOIS
Ingénieur-Chercheur, Irstea, UR HHLY (Villeurbanne), Directeur de thèse

Mme Sandrine ANQUETIN
Directrice de recherches, Université Grenoble-Alpes, CNRS, IRD, G-INP, IGE (Grenoble), Co-Directrice de thèse

ABSTRACT

This PhD work proposes new concepts and tools for stochastic weather simulation activities targeting the specific needs of hydrology. We used, as a demonstration, a climatically contrasted area in the South-East of France, Cévennes-Vivarais , which is highly attractive to hydrological hazards and climate change.

Our perspective is that physical features (soil moisture, discharge) relevant to everyday concerns (water resources assessment and/or hydrological hazard) are directly linked to the atmospheric variability at the basins scale, meaning firstly that relevant time and space scales ranges must be respected in the rainfall simulation technique. Since hydrological purposes are the target, other near-surface variates must be also considered. They may exhibit a less striking variability, but it does exist. To build the multi-variable modeling, co-variability with rainfall is first considered.

The first step of the PhD work is dedicated to take into account the heterogeneity of the precipitation within the rainfall simulator SAMPO [*Leblois and Creutin*, 2013]. We cluster time steps into rainfall types organized in time. Two approaches are tested for simulation: a semi-Markov simulation and a resampling of the historical rainfall types sequence. Thanks to clustering, all kind of rainfall is served by some specific rainfall type. In a larger area, where the assumption of climatic homogeneity is not considered valid, a coordination must be introduced between the rainfall type sequences over delineated sub-areas, forming rainy patterns at the larger scale.

We first investigated a coordination of Markov models, enforcing observed lengths-of-stay by a greedy algorithm. This approach respects long duration aggregates and inter-annual variability, but the high values of rainfall are too low. As contrast, the joint resampling of historically observed sequences is easier to implement and gives a satisfactory behavior for short term variability. However it lacks inter-annual variability. Both approaches suffer from the strict delineation of homogeneous zones and homogeneous rainfall types.

For these reasons, a completely different approach is also considered, where the areal rainfall totals are jointly modeled using a spatio-temporal copula approach, then disaggregated to the user grid using a non-deterministic, geostatistically-based conditional simulation technique. In the copula approach, the well-known problem of rainfall having atom at zero is handled in replacing historical rainfall by an appropriated atmospheric based rain-

fall index having a continuous distribution. Simulated values of this index can be turned to rainfall by quantile-quantile mapping.

Finally, the copula technique is used to link other meteorological variables (i.e. temperature, solar radiation, humidity, wind speed) to rainfall. Since the multivariate simulation aims to be driven by the rainfall simulation, the copula needs to be run in conditional mode. The achieved toolbox has already been used in scientific explorations, it is now available for testing in real-size application. As a data-driven approach, it is also adaptable to other climatic conditions. The presence of atmospheric precursors a large scale values in some key steps may enable the simulation tools to be converted into a climate simulation disaggregation.

CONTENTS

Bibliography

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Water

Water is a vital element for human existence. The importance of water is reflected in all kinds of effects on many issues in the world.

Effects on climate

Water has a regulatory role on the climate. Water vapor is a greenhouse gas that can protect the Earth from cooling, even more effective (60%) at absorbing the thermal radiation from the Earth's surface than carbon dioxide (25%) or ozone (8%) [see *Barkstrom*, 1990; *Karl and Trenberth*, 2003]. Marine and terrestrial water absorb, accumulate, modulate and distribute heat and maintain a livable temperature whatever the season.

The water in the oceans and on the earth's surface evaporates and is then advected in the atmosphere, to form clouds by condensation. A variety of processes redistribute cloud water on the surface as rainfall or snowfall, depending on the altitude and the temperature. Most water evaporates back, the rest often infiltrates in the soil and filtrate at water springs or river beds, sometimes runs overland, eventually flowing into the sea through rivers. This forms the water cycle.

Effects on topography

71% of the Earth's surface is covered by water [*Nace*, 1967]. The Earth is seen as a "Blue Planet" from the sky. Water erodes rock and soil. Water erodes river beds, transports sediments, lowers mountains and reshape alluvial plains. These processes act on surface morphology.

Effects on human society

Water is an important resource for human life. Daily people's needs are around 50 liters by inhabitant [*Gleick*, 1996], covering daily life needs, industrial and agricultural production. In particular, agriculture requires huge amounts of water, either rain feed or through irrigation. The origins of human civilization are mostly in the vicinity of rivers. Early cities were generally established at the water's edge to address irrigation and drinking issues.

With the development of science and technology, water conservancy is being built to fight natural disasters such as floods and waterlogging (which refers to the saturation of soil with water). As a result, a number of water-related activities developed, progressively backed by branches of water science such as hydrology, hydraulics, hydrobiology, etc.

The importance of hydrology science

The term *hydrology* comes from Greek: ὕδωρ [hýdōr] ("water" in English) and λόγος [lógos] ("study" in English). So literally, hydrology is the study of water.

In the Nature scientific journal[1], hydrology science is described as follows.

> Hydrology is the study of the cycling of water through different reservoirs on Earth. [...] Hydrology focuses on the distribution of water in the subsurface, surface and atmosphere, the chemistry of that water, and the effects of climate on the water cycle.

Among numerous published papers and text books, Hydrology: A Science of Nature of *Musy and Higy* [2010] present the various components of water cycle, the catchments or river basins, the factors affecting their hydrological response and hydrological regimes, as well as issues related to measurement and control of hydrological data. Research in hydrology aims at providing methods and tools that are essential for solving concrete problems related to water resources and the associated risks. Management of water resources and related hydrological hazards (e.g., flooding, landslides, mudslides, erosion, drought, etc.) are among the major hydrological concerns. Water resources and hydrological hazards are to be considered in territorial development studies, aiming at structural or non-structural development designs or at financial evaluations (e.g., costs/benefits relations, assurances practice, renewing of hydroelectric concessions). However, operators in charge of territorial assessment face major methodological difficulties in constructing climate scenarios at the regional scale, even where only present climate scenarios are requested, not to speak about taking climate change scenarios into account.

As an example, *Musy et al.* [2014] pointed out four key hydrological issues related to water resources and hydrological hazards which are (1) prediction of hydrological variables, (2) hydrological forecasting, (3) hydrological impacts of human and activities and (4) hydrological impact of climate change. Dealing with such key issues calls for innovating approaches linking observations with modeling tools. Observation networks are first needed to document the system of interest. Obviously, observations can not be provided everywhere, thus modeling approaches may be developed to first represent the missing values at the ungauged places, but are then widely used to forecast hydrological variables in the future (i.e. hours, day, years ahead). Hydrological models then need inputs at the proper targeted spatial and temporal scales.

[1] `https://www.nature.com/subjects/hydrology`, last check on 2017/11/23

The main objectives of the thesis

In order to address the hydrological issues related to water resources and hydrological hazards, the main objective of this PhD work is to provide efficient spatio-temporal stochastic simulations of hydrometeorological variables as inputs for hydrological models.

Two major issues are explored in this research.

1. The first concerns the heterogeneity of the rainfall that needs to be properly reproduced.

2. Being able to supply the required inputs for hydrological modeling constitutes the second issue. Identifying the meteorological inputs and proposing a multivariate system are thus investigated.

The approach should be able to be generalized in order to be applied in any territory under different climatic conditions. The strategy of regionalization needs to respect the resonance in spatial and temporal scales associating atmospheric and hydrological phenomena.

The progress of the thesis

This thesis is organized as follows.

Part I (i.e. Chapter 1-3) introduces the general context dealing with hydrological simulations, the geographical context and the main challenges of this PhD work. Chapter 1 emphasizes the importance of hydrological simulations in particular when water resources and hydrological hazard management are concerned. Chapter 2 introduces the studied area and available data that will be later used. Our main area of interest is located on the Cévennes-Vivarais region which is one of the main regions concerned by flash floods in Europe [*Boudevillain et al.*, 2011]. Two sources of meteorological data are considered in this PhD work. The observatory OHM-CV (Observatoire Hydro-météorologique Méditerranéen Cévennes-Vivarais) is used for its collection of precipitation observations. The ECMWF (European Center for Medium-Range Weather Forecasts) is a major institution providing climate services like atmospheric reanalyses. The ERA-Interim reanalysis of the ECMWF is used for all needed meteorological variables but precipitation. Chapter 3 points out the two main issues. First, given the availability at Irstea of the local stochastic rainfall simulator SAMPO (Simulation of Advected Mesoscale Precipitations and their Occurrence), several methods are proposed to adapt SAMPO to simulate over a heterogeneous rainfall field. Modeling under multivariate framework constitutes the second challenge of this PhD work: to provide proper meteorological inputs required for hydrological models, several meteorological variables such as temperature, wind speed, solar radiation and water vapor pressure must be taken into account.

Part II (i.e. Chapter 4-8) deals with the heterogeneity problem. The spatial variability of rainfall is usually more difficult to be captured than the temporal variability due to the

resolution of precipitation data. One easy but not ideal way to overcome the heterogeneity problem is to partition the whole region into several zones that can be considered homogeneous in term of rainfall regimes. Instead of modeling spatial correlations with large variability, the heterogeneity problem then becomes how to spatially coordinate these homogeneous rainfall zones in space. Chapter 4 reviews some existing stochastic rainfall models that simulate precipitation at more than one location. Chapter 5 proposes two different approaches which are parametric (hierarchical Hidden Markov model) and non-parametric (resampling model) to deal with the coordination problem. Chapter 6 shows the statistical diagnosis of the two approaches. Chapter 7 proposes a contrasting approach inspired by copula technique. Along with copula approach, a disaggregation model is introduced to generate spatially fine simulations which respect large scale values. Chapter 8 makes the conclusions of our approaches concerning to the heterogeneity of the rainfall field problem.

Part III (i.e. Chapter 9) deals with multivariate simulations. The copula technique is a very promising statistical approach to deal with multivariate situations. Combining with auto-regressive process (which is used to model temporal correlations for time-series) and kriging technique (which is used to generate sequentially conditional simulations), a copula based multivariate model is proposed to generate long term simulations of input variables such as precipitation, temperature, wind speed, solar radiation and water vapor pressure for the needs of hydrological models.

Part IV (i.e. Chapter 10-11) makes the conclusions and perspectives of this PhD work.

Part I
General context

CHAPTER 1

SOME MAIN HYDROLOGICAL CONCERNS

[XXX] The title of this chapter should be changed since its topic is not hydrological modelling. How to change the title ?

Water is one of the most important natural resources on our earth. In science, hydrology is the study of water. More precisely, Hydrology is the science that encompasses the occurrence, distribution, movement and properties of the water on earth and its relationship with the environment at each phase of the water cycle. The water cycle, also known as the hydrological cycle, describes the continuous movement of water on, above and below the surface of the Earth.

The main issues in hydrology are to enhance our descriptive capacity and to build the proper synergy between observations and modeling. Practical issues are how to better manage the water resources and wisely assess the hydrological risks. Water cycle refers to the water circulation on different parts of the Earth. Absorbing the energy of the sun, water changes state and moves to other places on earth. [XXX rephrase] For example, the water on the surface turns into water vapor by evaporation as the sun warms. Then, the water vapor is advected in the atmosphere, later forms clouds that lead to precipitation. Energy and water cycles are deeply interconnected. While the state of water in the earth includes solid, liquid and gaseous, the earth's water is present in the atmosphere, in the soil, in the ground, in the lakes, in the rivers and in the oceans [see *Gleick and Howe*, 1995]. The water moves from one place to another through physical processes such as evaporation, precipitation, infiltration, surface flow and underground flow. Water movement can also be induced by biological processes (e.g., evapotranspiration).

The water cycle is an important process that redistributes water and nutrients at the global scale. Change of phase of water (e.g., liquid to vapor) plays also a major role in energy transfer. In doing so, it brings freshwater to people, animals and plants all around the world. The main roles of water cycle can be synthesized as:

1. Water is the medium of all nutrients. The circulation of nutrients and the water cycle are inextricably linked together.

2. Water is a good solvent for substances involved in the energy transfer and essential

for the ecosystems.

3. Water plays a role at the geological scale. The loss of a local mineral element and its deposition at another place are often done through the water cycle.

In the framework of this PhD work, two particular themes in hydrology which are water resource management and hydrological hazards are first introduced.

1.1 Water resources

Water resources are sources of water that are potentially useful. Uses of water include many activities such as agricultural, industrial, household, recreational and environmental activities. All living entities require water to grow and reproduce. Figure 1.1 presents the distribution of earth's water.

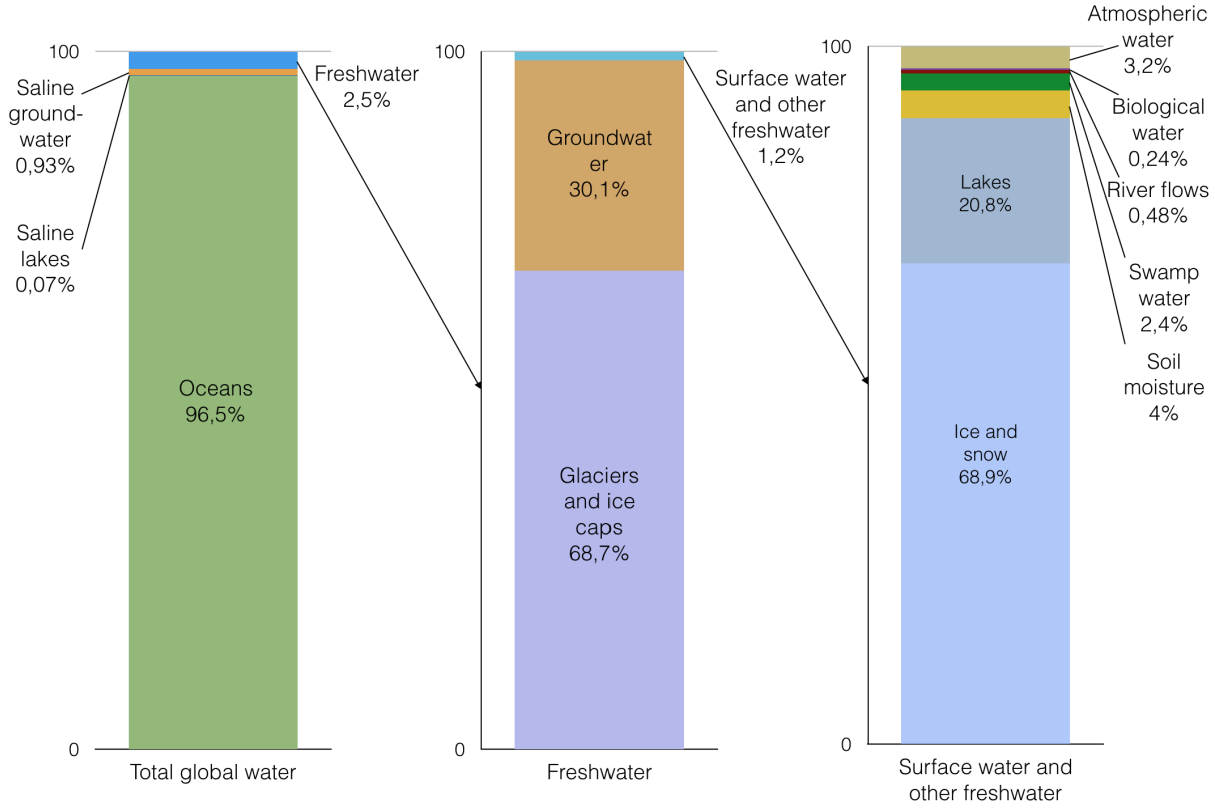


Figure 1.1: Distribution of earth's water. Only 3% *of the Earth's water is fresh water. Most of it is in icecaps and glaciers* (69%) *and groundwater* (30%)*, while all lakes, rivers and swamps combined only account for a small fraction* (0.3%) *of the Earth's total freshwater reserves. Source: [Gleick and Howe, 1995].*

Salt water occupies 97% of the water on the Earth and only 3% is fresh water; around 2/3 of fresh water is frozen in glaciers and polar ice caps. The remaining unfrozen freshwater is found mainly as groundwater, with only a small fraction present above ground or in the atmosphere. *Gleeson et al.* [2012] mentioned that fresh water is a renewable resource, yet

the world's supply of groundwater is steadily decreasing, with depletion occurring most prominently in Asia, South America and North America, although it is still unclear how much natural renewal balances this usage, and whether ecosystems are threatened.

It is essential to understand that the management of water resources is directly related to human life and the future of our earth. Water resources in France are on the order of 200 billion m^3 per year on average, i.e. around 3,300 m^3 per inhabitant per year [see *FAO*, 2003]. The average annual water withdrawal is 33 billion m^3, of which 19% is from groundwater and 81% from surface water. Agriculture accounts for 70% of anthropogenic consumption, domestic use 23% and energy production 7%. Water resources are widely used by human beings in production and living activities, not only widely used in agriculture, industry and life, but also for hydroelectric power generation, water transport, aquatic products, tourism and environmental transformation. Among the various uses, some are water consumption, others are non-expendable or consume very little water. Therefore, the requirements for water quality are different.

[XXX rephrase] Water resources are unevenly distributed among the earth due to natural atmospheric cycles, linked to the solar cycle, and due to the different properties of the ground worldwide. The annual wet and dry seasons and their associated surface and groundwater evolutions are ones of popular knowledge. Locally, such seasonal or monthly trends are partly known. Variability around these "natural" trends is an important aspect of water resource and how to predict and assess this part of the water resources is still an open question. For example, inter-annual variability of a river discharge remains difficult to capture as well as induced changes of the water table. Understanding may be improved by the use of statistical analysis requiring long series of observation data. Such study is thus possible only for very few parts of the world where the data are available.

Another issue associated with the study of the water resource concerns the signature of global warming. Indeed, *Huntington* [2006] showed that the intensification of the water cycle may lead to i) changes in water-resource availability, ii) an increase in the frequency and intensity of tropical storms, floods, and droughts, and iii) an amplification of warming through the water vapor feedback. Therefore, improving our knowledge on the variability of water resources, at different scales, may help to better anticipate such important issues.

1.2 Hydrological hazards

[XXX-NON] This section is confused, jumping from one topic to another without clear guidelines.

Hydrological hazards include floods, storm surges, coastal erosion and droughts. It is important to understand the relationship of hydrological hazards with other hazards (e.g., weather or solid earth (earthquakes, volcano, etc.)) [XXX-NO] what is referring the author ?. For example, extreme rainfall from a thunder and lightning event can cause flooding; winds and low pressure from a tropical cyclone can exacerbate storm surge and coastal erosion.

Floods

Flooding typically results from large-scale weather systems generating prolonged rainfall or on-shore winds. Other causes of flooding include locally intense thunderstorms, snow melt, ice jams, and dam failures. Flash floods, which are characterized by rapid onset and high velocity waters, may carry large amounts of debris. They occur at smaller time and space scales. Floods are capable of undermining buildings and bridges, eroding riverbeds and riverbanks, tearing out trees, washing out access routes, and causing loss of life and injuries, and disruption in economy.

In France, a study lead by *Lang and Coeur* [2014] established lists and descriptions of noteworthy floods. Such reference works serve memory, provide insight to typical or unexpected mechanisms.

Considerable research has been devoted to investigate the most appropriate frequency distribution and fitting method for flood-frequency analyses [*Stedinger*, 1993]. Different frequency distributions and fitting methods have been suggested as superior to the Pearson Type III frequency distribution [*Singh*, 1998], which has been used widely for many years.

1.3 Conclusions

Water provides the possibility of life for the Earth and mankind. Relative to the total amount of water in the Earth, the proportion of available water resources is very small. Although the water cycle helps to recycle water resources, water resources are still limited and unevenly distributed. Large amounts of fresh water are present in polar ice caps and glaciers, but many countries and regions around the world are facing a crisis of water scarcity. How to wisely deal with water resources is still an unavoidable problem.

Water fluxes variability is natural and plays a positive role in a natural environment. It is also a major source of natural disasters such as floods and droughts. The assessment and prevention of hydrological hazard is essential to the protection of human life and property.

As a way of helping water resource management and hydrological hazard assessment, the spatio-temporal variability of hydrological variables must be made explicit and considered. [XXX rephrase] Therefore, hydrological simulation is of great significance. The relevance of hydrological simulation is largely based on the selection of adequate time and space scales.

Water resources are indispensable resources for human survival. But at the same time, hydrological hazards as floods and droughts often bring disasters to people. This is the starting point of this PhD work that aims to serve simulation frameworks in order to enhance our understanding of the evolution of the water resources. The region of interest is the French Mediterranean region, in particular the Cévennes-Vivarais region. This area benefits of many observations [*Boudevillain et al.*, 2011; *Braud et al.*, 2016] and numerous modeling works [*Nuissier et al.*, 2008; *Godart et al.*, 2011; *Adamovic et al.*, 2016] have been already realized to partially study the different parts of the water cycle, such as, processes associated with intense precipitation, contribution of the orographic precipitation to the water resource and identifying hydrological signatures at different basin scales, respectively.

CHAPTER 2

PRESENTATION OF THE STUDIED AREA AND OF DATA

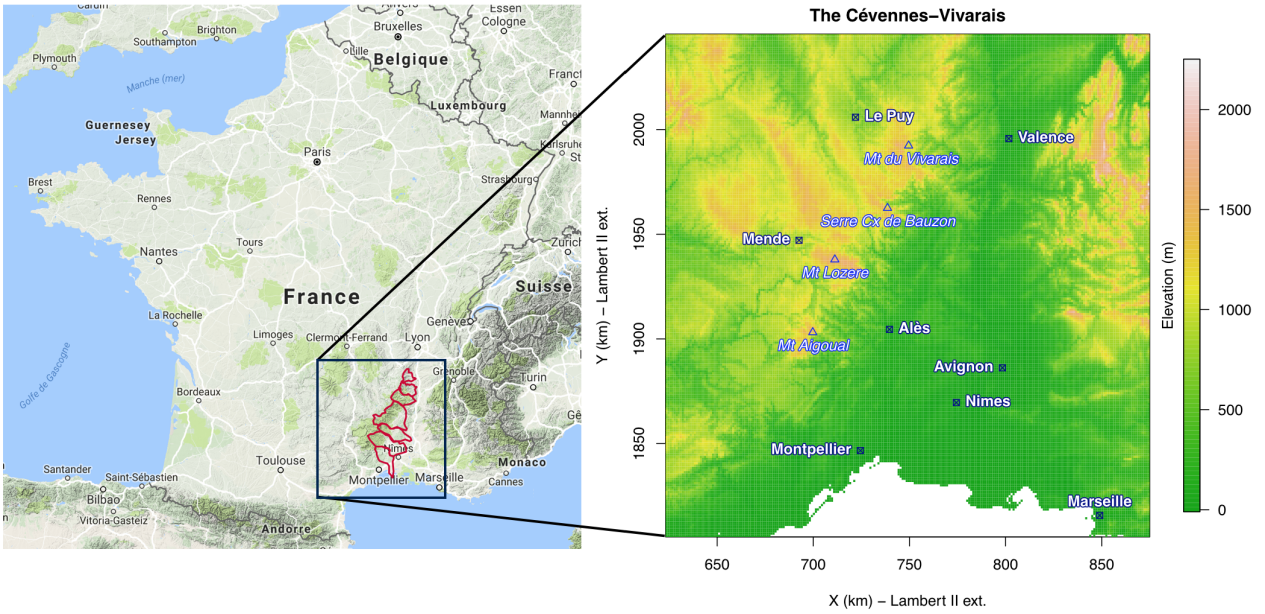
In this chapter, the studied area of the Cévennes-Vivarais region is first introduced and two meteorological databases are then presented for the needs of the hydrological inputs. Due to its main topographic features, the Cévennes-Vivarais region leads to a clear heterogeneous behaviour of the rainfall, which constitutes the main guideline of this chapter.

2.1 Studied area: The Cévennes-Vivarais, France

The Cévennes (Figure 2.1(b)) are a range of mountains in south-central France (Figure 2.1(a)), covering parts of the French administrative Departments of Ardèche, Gard, Hérault and Lozère. The Cévennes are a part of the Massif Central. They run from southwest (le Causse Noir) to northeast (Monts du Vivarais). The highest peak are the Mont Lozère (1702 m) and the Mont Aigoual (1567 m). Among all rivers having their headwaters in the Cévennes, the Loire river and its tributary the Allier river flow northwestward to the Atlantic ocean, whereas the Ardèche, Chassezac and Cèze rivers, the different Gardons, the Vidourle, Hérault and Dourbie rivers flow Southeastward either as tributaries to the Rhone or directly to the Mediterranean Sea.

This mountainous area, 3730 km^2, is prone to flash flood hydrometeorological events, and has been intensively monitored within the framework of the long term natural observatory OHM-CV [*Boudevillain et al.*, 2011]. The OHM-CV began activity in 2000 compiling a research database by gathering, analyzing, and archiving meterological and hydrological data. As mentioned by *Boudevillain et al.* [2011], the Cévennes-Vivarais region is one of the main regions concerned by flash floods in Europe. *Nuissier et al.* [2008] explained that the simultaneous presence of a number of meteorological factors is propitious to extreme precipitation events.

One of the interesting fact is that Cévennes-Vivarais belongs to the Mediterranean region. As *Christensen et al.* [2007] reports, intense precipitation events will likely increase over central Europe in winter, but trends over the Mediterranean regions remain uncertain because of complex interactions at different scales (presence of topography; interactions and feedback among atmosphere-ocean–land processes) that play a predominant role in



(a) *(b)*

Figure 2.1: (a) Location of the Cévennes-Vivarais region in the map of France, the red contours refer to the main catchments of the Cévennes-Vivarais area; (b) zoom on the study area, the triangles refer to the main summits and the squares to the largest cities.

climate and the related ecosystems. So, the Mediterranean region has been identified as a "hot spot" of climatic change, in the sense that climate change is expected to modify hydrological regime in this region, but the direction of change is uncertain, justifying dedicated and dense monitoring. More details about the topological features, climate characteristics and research activities in the Cévennes-Vivarais area can be found, as example, in [*Delrieu*, 2003; *Braud et al.*, 2010, 2016].

2.2 Observation and re-analysis data

Two meteorological databases are considered in this thesis. The first one concerns the precipitation observations. Meteorological variables such as temperature, wind speed or solar radiation are retrieved from the second one. These two databases differ from the origin of the data. Precipitation data are local rain gauge observations, whereas meteorological variables refer to computed values obtained with the integration of radiosonde observations by the mean of physical equations.

OHM-CV Database

A primary objective of the observatory OHM-CV [1] (**O**bservatoire **H**ydrométéorologique **M**éditerranéen **C**évennes-**V**ivarais) is to bring together the skills of meteorologists and hydrologists, modelers and instrumentalists, researchers and practitioners, to improve knowledge and capacity for forecasting the hydrometeorological risk associated with

[1] `http://www.ohmcv.fr/P100_objectifs.php?lang=en`: last check on 2017/11/31

heavy rainfall and flash floods in the Mediterranean region. The OHM-CV observation strategy consists of three complementary lines of effort:

1. detailed, long-lasting, and modern hydrometeorological observation over part of the region of interest, the Cévennes-Vivarais region, dedicated to process studies and to the improvement and assessment of hydrometeorological predictive models;

2. multi-disciplinary post-flood investigations after any event occurring over the entire Mediterranean region to document and analyze the physical and societal processes associated with such extremes;

3. use of historical information available on past floods to better characterize the frequency of extreme hydrometeorological events and the possible trends under a changing climate.

The OHM-CV has been strongly involved in the meta-program MISTRALS (Mediterranean Integrated STudies at Regional And Local Scales) of the Inter-Organizations Environment Committee, and in particular the HyMeX[2] (HYdrological cycle in the Mediterranean EXperiment) project. This latter project is dedicated to the study of the water cycle in the Mediterranean, with a particular interest for the evolution of climate variability and for the genesis and predictability of intense events.

European Center for Medium-Range Weather Forecasts (ECMWF) Database

ECMWF is an independent intergovernmental organization supported by 34 states. ECMWF is both a research institute and a 24h/7d operational service, producing and disseminating numerical weather predictions to its Member States. The data are fully available to the national meteorological services in the Member States. The Center also offers a catalogue of forecast data that can be purchased by businesses worldwide and other commercial customers. The supercomputer facility and associated data archive at ECMWF is one of the largest of its type in Europe; Member States can use 25% of its capacity for their own purposes. The organization was established in 1975 and now employs around 350 staff from more than 30 countries. ECMWF is one of the six members of the Co-ordinated Organizations, which also include the North Atlantic Treaty Organization (NATO), the Council of Europe (CoE), the European Space Agency (ESA), the Organization for Economic Co-operation and Development (OECD), and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT).

In this thesis, data out of the ECMWF ERA-Interim reanalysis [*Dee et al.*, 2011] is used. ERA-Interim is a global atmospheric reanalysis from 1979, continuously updated. The temporal resolution is a 6-hours time step and the spatial resolution is 0.75°.

The data assimilation scheme of ERA-Interim included many ground observations (temperature, pressure, etc.) and precious vertical profiles brought by radio-soundings, and

[2] `https://www.hymex.org`: last check on 2017/11/31

also satellite imagery was considered, bringing in cloud presence and cloud height information (the presence of satellite data is the reason why ERA-Interim database could not be extended before 1979). Locally observed rainfall, however, was not considered in ERA-Interim reanalysis. The reasons for that are the huge gap in scales and the difficulty of entering strongly non-linear precipitation processes schemes into assimilation systems.

2.2.1 Precipitation

Rainfall observations are provided through the OHM-CV database[3]. Hourly and daily precipitation data are available (Table 2.1).

Table 2.1: Precipitation observations available in the OHM-CV database

Type of data	period available	number of rain gauge stations
hourly precipitation	2005 - 2014	146
daily precipitation	1985 - 2014	840

The observations used in this PhD work are the 146 hourly rain gauge stations. The location of these stations in the Cévennes-Vivarais area is illustrated in Fig. 2.2.

Many previous works proposed a climatology of the Cévennes-Vivarais [e.g., *Molinié et al.*, 2012]. Below simple analysis are proposed to introduce the main scales of the precipitation variability. Figure 2.3 depicts the monthly average precipitation of the 146 hourly rain gauge stations from 2005 to 2014. This figure shows that there are more precipitation in fall than in summer over Cévennes-Vivarais region. Monthly variability is more important in February, April, May and November, than for the rest of the year. In November, the monthly average precipitation could exceed 200 mm (222 mm for 2011 and 295 mm for 2014).

Four rain gauges are chosen to illustrate the inter-annual variability of the precipitation in different parts of the study area. Their locations are presented in Fig. 2.4. In Fig. 2.5, the time series of daily precipitation are plotted from 2005 to 2014 at the 4 rain gauges as presented in Fig. 2.4. Clearly, the rain gauge station MONT AIGOUAL, located in Mountain area, records more precipitation, also frequent strong rainfall events.

[3] `http://ohmcv.osug.fr/spip.php?article30`: last check on 2017/11/31

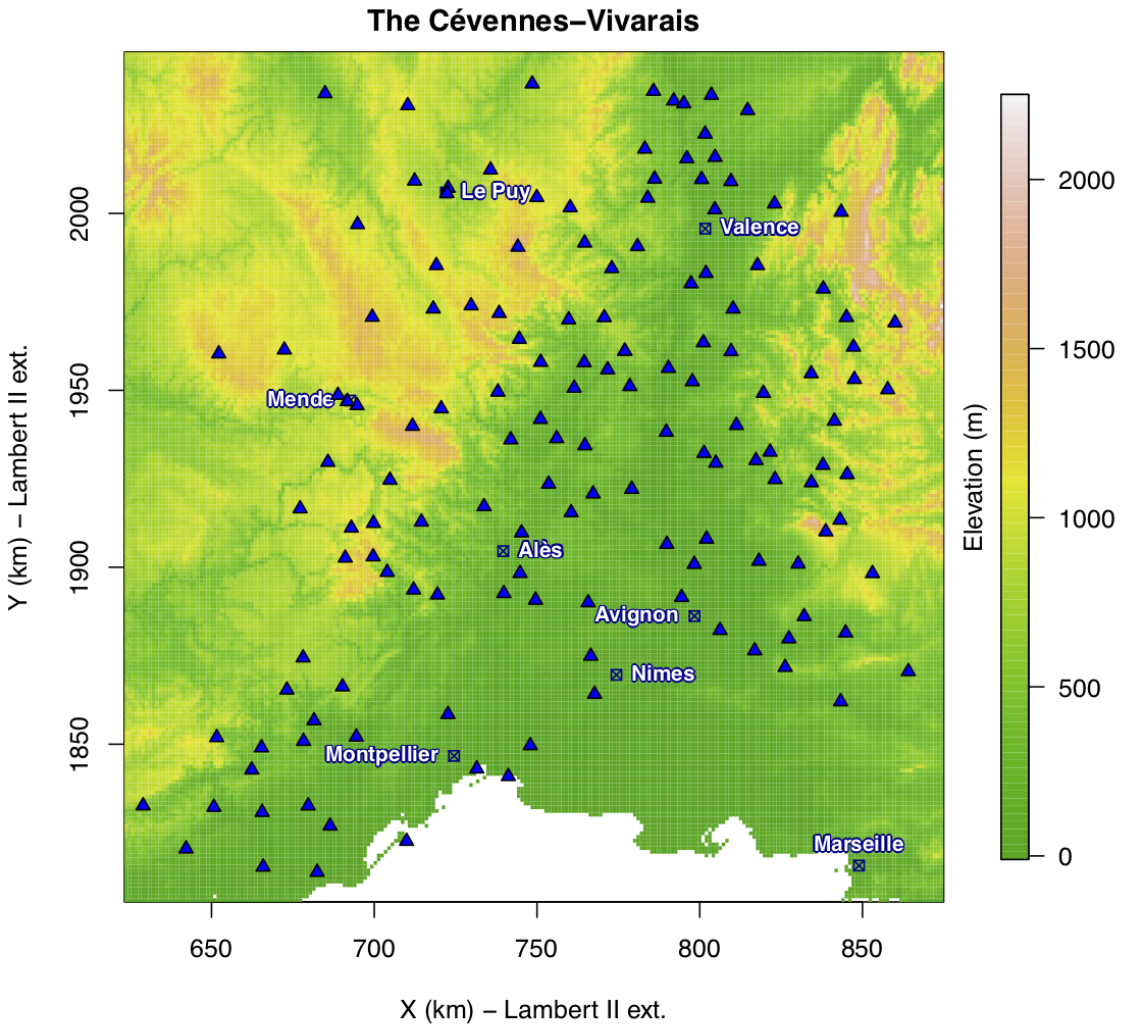


Figure 2.2: Locations of the 146 hourly rain gauge stations in the Cévennes-Vivarais region.

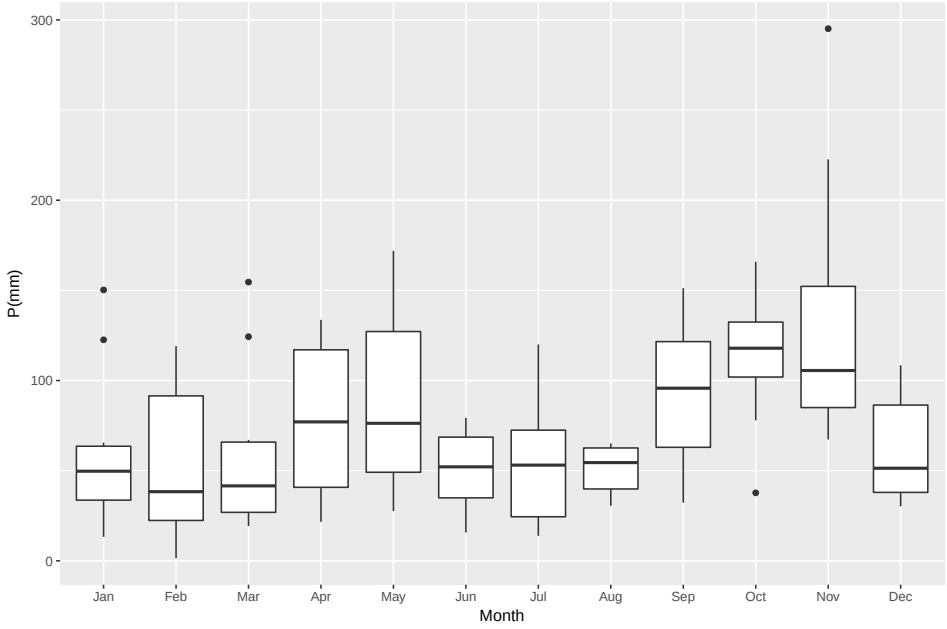


Figure 2.3: Monthly average precipitation of the 146 hourly rain gauge stations from 2005 to 2014.

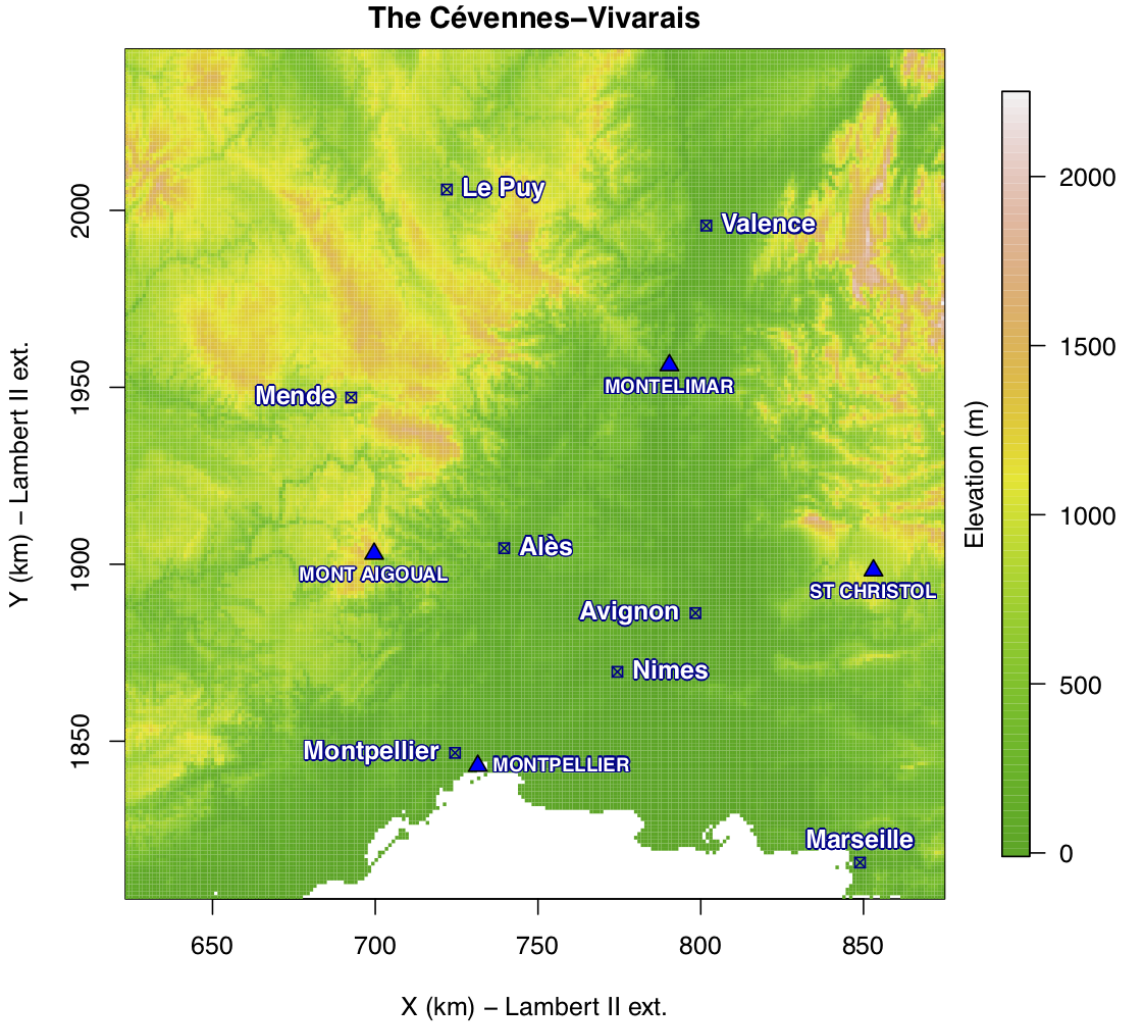


Figure 2.4: Location of the 4 rain gauges in blue triangle (MONTPELLIER, ST CHRISTOL, MONTELIMAR, MONT AIGOUAL) located in different areas.

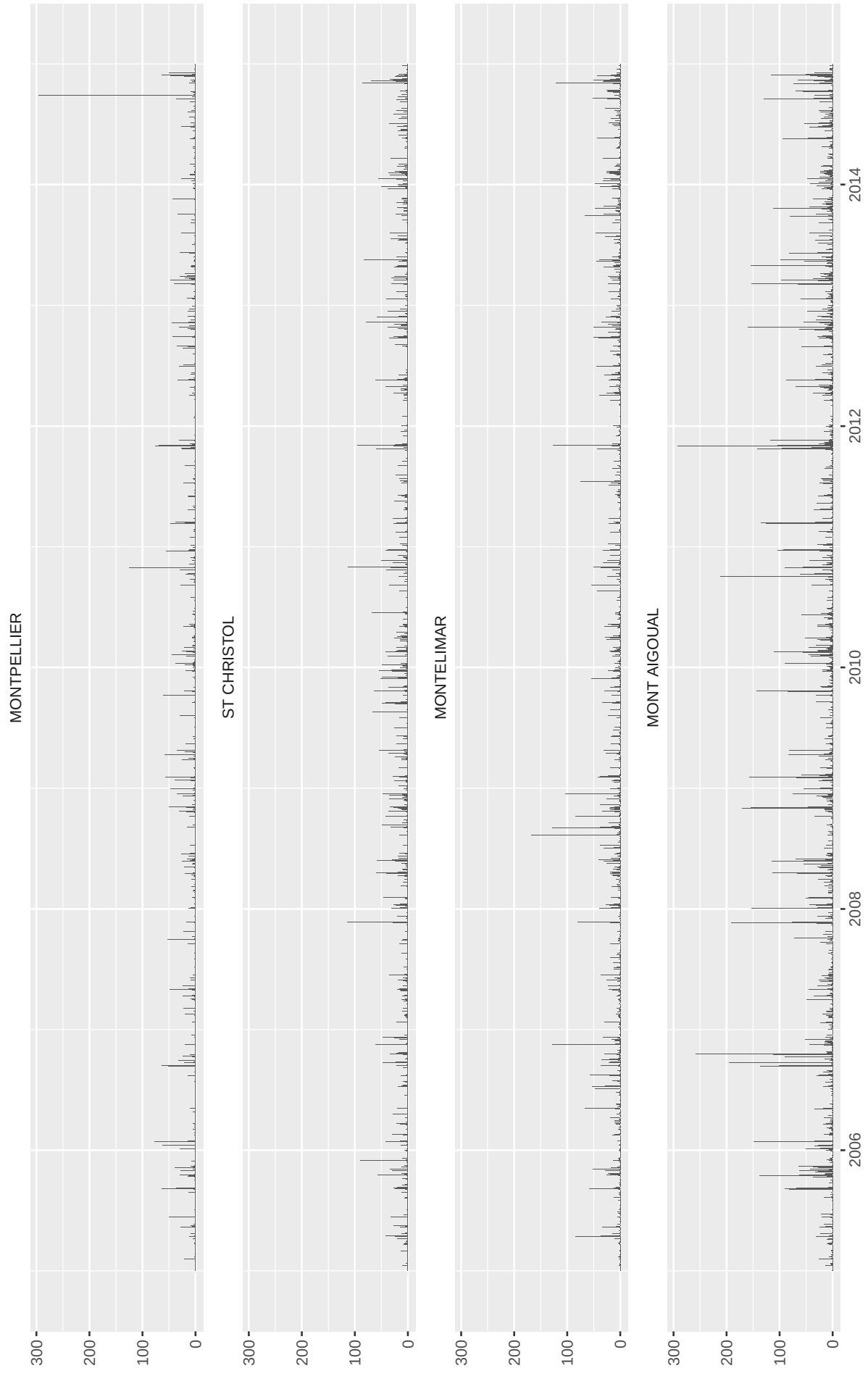


Figure 2.5: Time series of daily precipitation from 2005 to 2014 for the 4 rain gauges in Fig. 2.4.

Precipitation variability takes place within a large range of time and space scales that make this meteorological variable highly complex to observe and to forecast. As far as stochastic rainfall generators are concerned, *Koch and Naveau* [2015] indicate that, most of the stochastic rainfall generators nowadays have a good performance with daily precipitation data, but still present drawbacks when hourly precipitation is concerned. As a matter of fact, hourly precipitation stochastic simulation poses a difficult challenge, given values that appear non-negative, positively skewed, possibly heavy tailed, containing a lot of zeros (dry hours) sometimes organized in dry episodes having long persistence.

Hourly precipitation data at the 4 rain gauges, are illustrated for two given years, for example 2005 and 2014, using histograms in Fig. 2.6 and Fig. 2.7, respectively. They appears that for each rain gauge station in both years, the hourly precipitation distribution is positive, highly skewed to the highest values and contains high values. Mont Aigoual and St Christol also have much frequent and more rain than Montelimar and Montpellier, although high hourly values are likely seen also in Montpellier, as year 2014 gives a clear example.

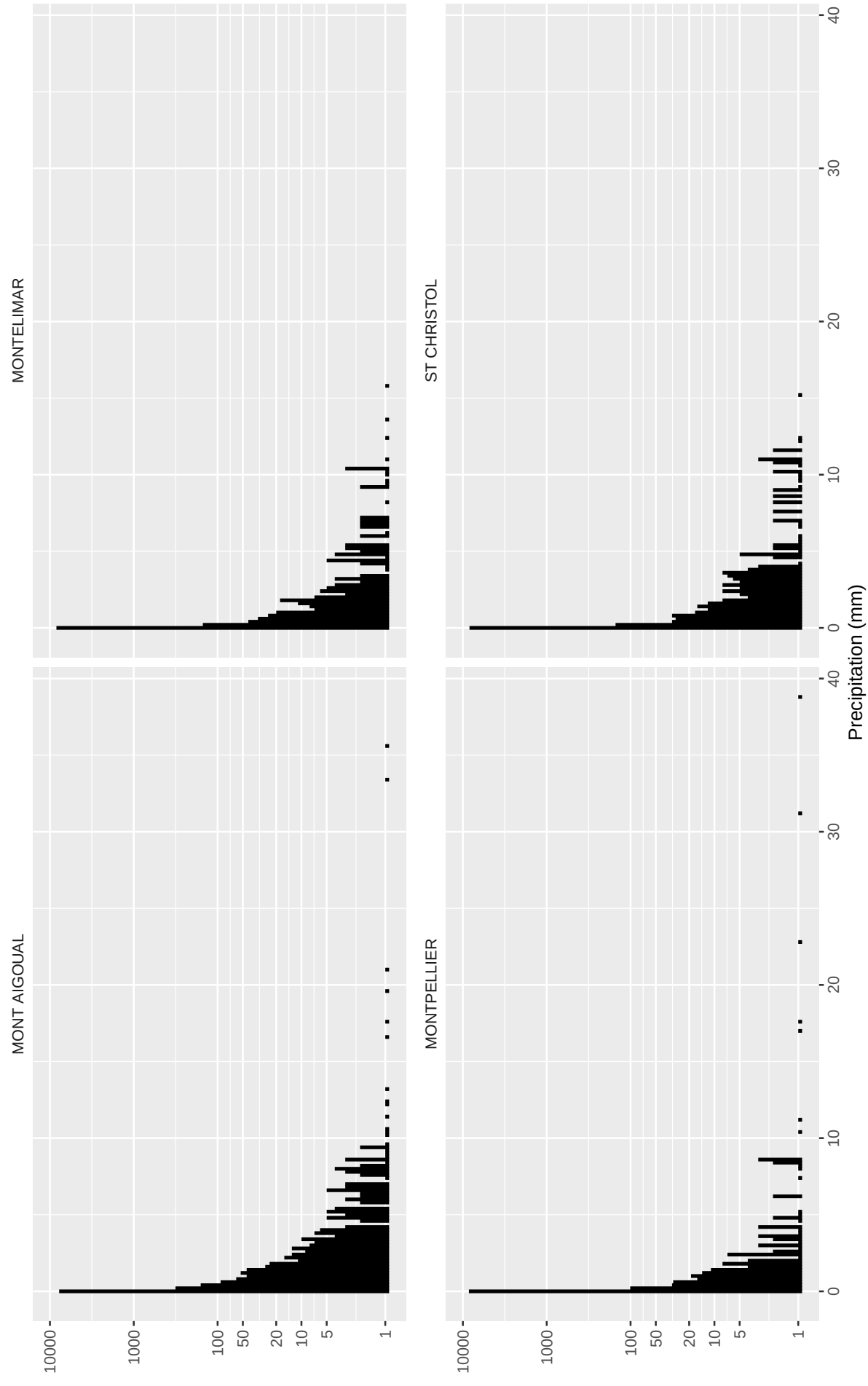


Figure 2.6: Histogram of hourly precipitation in 2005 for the 4 rain gauges.

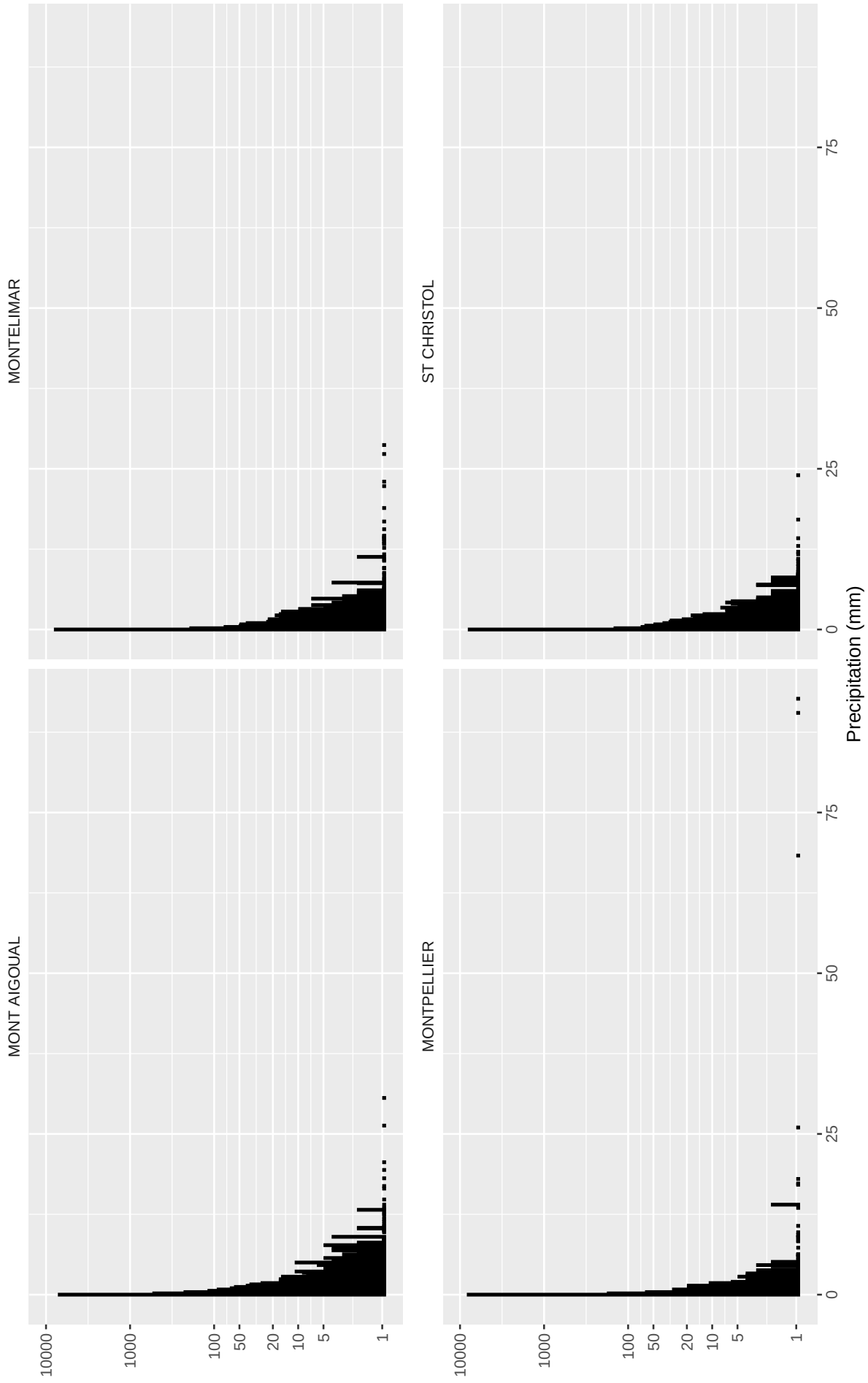


Figure 2.7: Histogram of hourly precipitation in 2014 for the 4 rain gauges.

The Cévennes-Vivarais region: a pool of several relatively homogeneous rainfall zones

As mentioned in Section 2.1, several different topological features form the Cévennes-Vivarais region lead to a strong heterogeneity state of the precipitation. *Guttorp and Sampson* [1994] gave a series of methods for estimating heterogeneous spatial covariance functions for environmental applications. Nevertheless, the parameter estimation associated with spatial covariance functions can be very difficult when the number of considered points/stations/rain gauges increases, running into a curse of dimensionality. To cope with spatial heterogeneity, a much simpler and possibly the simplest way is to divide the whole region into sub-regions or zones considered as "relatively" homogeneous. These zones need to be determined. The identification of a possible set of "k" zones is based on clustering techniques applied to the correlations between each pair of hourly rain gauges. In the Cévennes-Vivarais region, the 146 hourly rain gauge stations are therefore used, and the stations presenting similar vector of correlations to all stations will be gathered in the same homogeneous zone.

The n rain gauge stations $\{S_1, \ldots, S_n\}$ have the same record length of precipitation data. The correlation between any pair of stations (S_i, S_j) is calculated by correlation function. There are three typical correlation coefficients which are:

1. Pearson correlation coefficient [see *Galton*, 1886] which is a measure of the linear correlation between two variables;

2. Spearman correlation coefficient [see *Spearman*, 1904];

3. Kendall correlation coefficient [see *Kendall*, 1938; *Kruskal*, 1958].

The latter two refer as rank correlation coefficients which measure the extent to which, as one variable increases, the other variable tends to increase, but without assuming a linear relation between the two increments. Considering the type of data used in this PhD work, the rank correlation coefficients are more suitable. Thus, the Kendall rank correlation coefficient (called Kendall's tau) is chosen consideration the possible ties ($x_1 = x_2$; $y_1 = y_2$). In most situations, the interpretations of Kendall's tau and Spearman rank correlation coefficient are very similar and lead to the same inferences. Advantages of using Kendall's tau over Spearman rank correlation are as follows:

- The distribution of Kendall's tau has better statistical properties [see *Hamed*, 2009a,b].

- The interpretation of Kendall's tau in terms of the probabilities of observing the concordant or discordant pairs is very direct.

Let (x_1, y_1), (x_2, y_2), ..., (x_m, y_m) be a set of observations of the joint random variables X and Y respectively. One says that (x_i, y_i) and (x_j, y_j) are concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$ (i.e., if $(x_i - x_j)(y_i - y_j) > 0$); and discordant if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$ (i.e., if $(x_i - x_j)(y_i - y_j) < 0$). There are $\binom{m}{2}$ distinct pairs of

observations in the sample, and each pair (barring ties) is either concordant or discordant. The Kendall τ coefficient for the pair (X and Y) is defined as:

$$\tau(X,Y) = \frac{(\text{number of concordant pairs}) - (\text{number of disconcordant pairs})}{m(m-1)/2}. \tag{2.1}$$

□

For any hourly rain gauge station S_i ($i \in [1:146]$), the Kendall correlation coefficients between S_i and all 146 stations consist a $n(=146)$ dimension correlation vector τ_{S_i} which is defined as:

$$\tau_{S_i} = \begin{pmatrix} \tau(S_i,S_1) \\ \tau(S_i,S_2) \\ \vdots \\ \tau(S_i,S_n) \end{pmatrix}, \quad n = 146. \tag{2.2}$$

A matrix M_τ such as

$$M_\tau = \left(\tau_{S_1} \mid \tau_{S_2} \mid \cdots \mid \tau_{S_n}\right) = \left(\begin{array}{c|c|c|c} \tau(S_1,S_1) & \tau(S_2,S_1) & \cdots & \tau(S_n,S_1) \\ \tau(S_1,S_2) & \tau(S_2,S_2) & & \tau(S_n,S_2) \\ \vdots & \vdots & & \vdots \\ \tau(S_1,S_n) & \tau(S_2,S_n) & & \tau(S_n,S_n) \end{array}\right)$$

consists the Kendall coefficients between all pairs of stations. Each column of M_τ is the rank correlation vector of one station which represents the entire relation of the rank correlations between this station and all stations (include itself).

Classification is not only a matter of choosing an algorithm, but ultimately a question of choosing the relevant descriptor for the purpose. [XXX-NON rephrase] In this work, choosing correlations coefficients calculated on time-series as descriptors will favor that the stations in one group enter rainy condition in the same time, in related quantities. So, to obtain the homogeneous areas, it is our choice that the classification technique will be applied on the matrix M_τ. The principle of clustering is the task of grouping a set of objects (here the 146 stations) in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other clusters. There are three major clustering techniques which are (1) Hierarchical clustering, (2) k-means clustering and (3) Density-based clustering. More details for different types of clustering methods could be found in *Rokach and Maimon* [2005]. In this work, we use the k-means clustering technique on the rank correlation matrix M_τ to partition the rain gauge stations into several clusters. The term "k-means" was first used by *MacQueen et al.* [1967], though the idea goes back to *Steinhaus* [1956].

Given a set of observations $(X_1, X_2, \ldots, X_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $G = \{G_1, G_2, \ldots, G_k\}$ so as to minimize the within-cluster sum of squares (i.e. variance). Formally, the objective is to find:

$$\underset{G}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in G_i} \|x - \mu_i\|^2 = \underset{G}{\operatorname{argmin}} \sum_{i=1}^{k} |G_i| \operatorname{Var} G_i, \tag{2.3}$$

where μ_i is the mean of points in G_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\underset{G}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in G_i} \|x - \mu_i\|^2 \equiv \underset{G}{\operatorname{argmin}} \sum_{i=1}^{k} \frac{1}{2|G_i|} \sum_{x,y \in G_i} \|x - y\|^2. \tag{2.4}$$

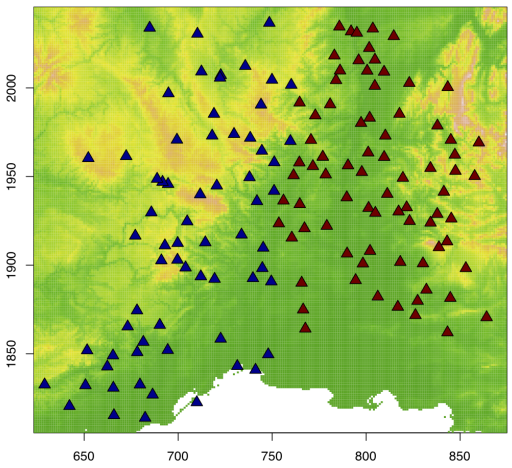
□

In our case, a set of observations $(X_1, X_2, \ldots, X_n)$ is the vectors $(\tau_{S_1}, \tau_{S_2}, \ldots, \tau_{S_n})$ and $n = 146$.

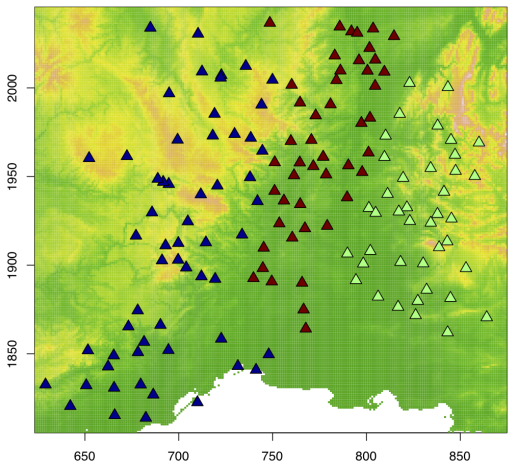
The algorithm of k-means clustering could be found in *MacKay* [2003]. In practice, *kmeans* is a function in the package *stats* of R software.

Number of clusters

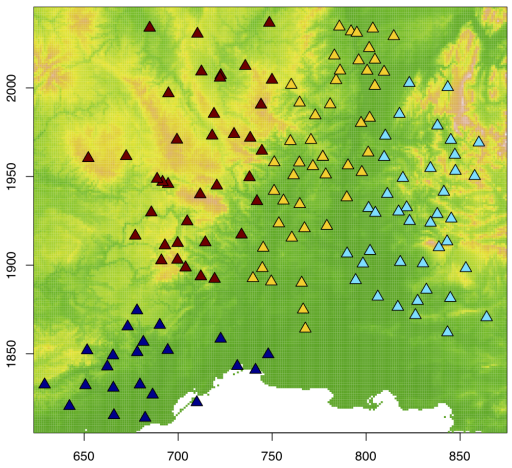
To identify the optimal number of clusters, k-means clustering is thus applied on k = 2 to 7 clusters. The results of these 6 clusterings are presented in Fig. 2.8, where each color refers to one cluster gathering n stations among the 146 rain gauge stations. Table 2.2 gives the number of stations in each cluster for each clustering. The number of rain gauge stations in each cluster needs to be large enough to maintain robust statistics. Therefore, the number of clusters should remain limited, but average intra-class distance needs also to be as minimum as possible showing a certain form of class homogeneity. On the other hand, to choose the number of clusters, one needs to deal with the average distance of inter-class that needs to be as large as possible and with the intra-class variability that must be the lowest as possible. Figure 2.9 illustrates these two quantities for the 6 clusterings.



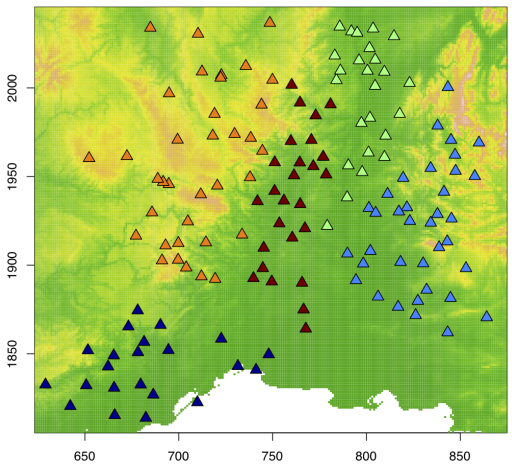
(a) 2 clusters



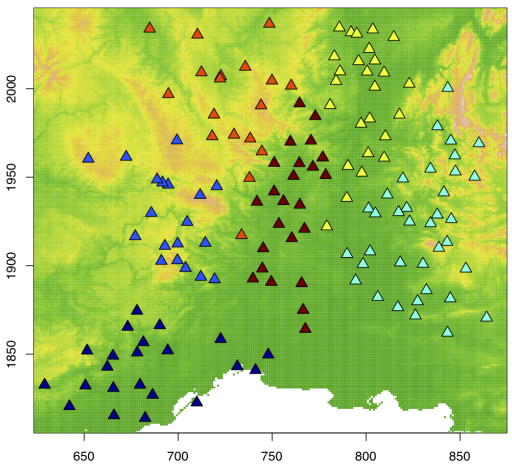
(b) 3 clusters



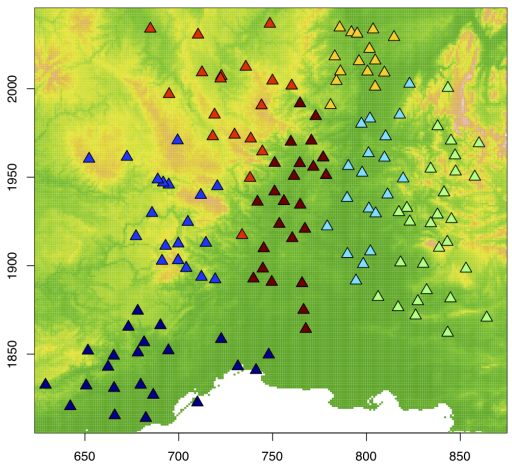
(c) 4 clusters



(d) 5 clusters



(e) 6 clusters



(f) 7 clusters

Figure 2.8: 146 rain gauge stations are partitioned into several clusters by using k-means clustering technique. From 2.8a: 2 clusters to 2.8f: 7 clusters.

Table 2.2: Number of stations in each class, depending on the total number of clusters. The partition refers to the ones given in Fig. 2.8.

2 clusters	67	79					
3 clusters	48	58	40				
4 clusters	22	37	40	47			
5 clusters	22	27	36	25	36		
6 clusters	22	26	18	25	19	36	
7 clusters	22	19	25	28	18	19	15

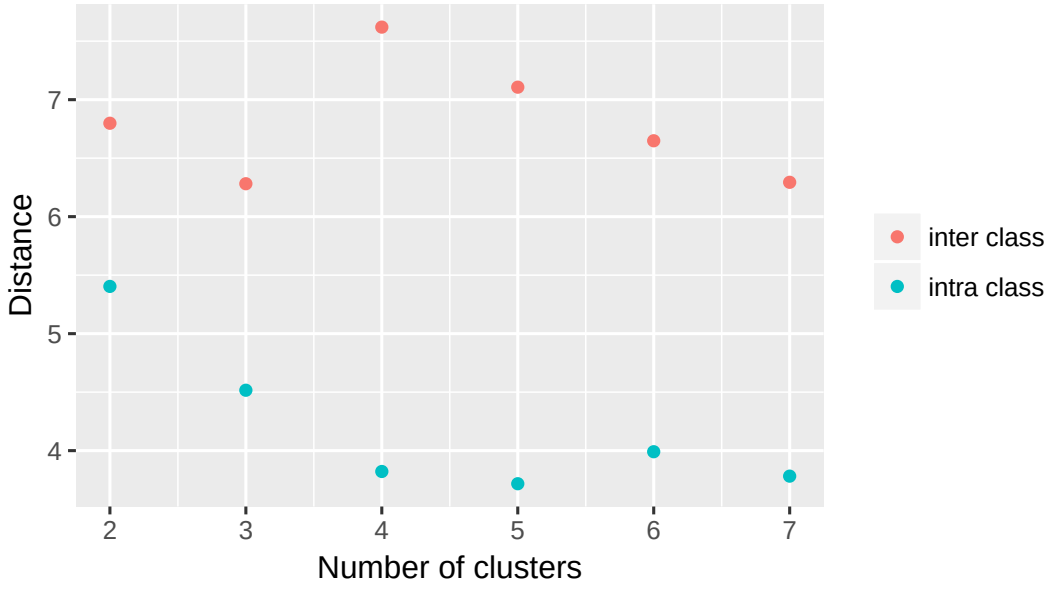


Figure 2.9: Average inter-class (in red) distance and intra-class (in blue) distances as a function of the number of classes.

Based on the results illustrated in Fig. 2.9, we decide to choose 4 "homogeneous" areas (i.e. Fig. 2.10) to represent the Cévennes-Vivarais . These 4 "areas" overlap pretty well with the topographical features of the region, as detailed in Table 2.3, strengthening the idea of a certain form of homogeneity. In Fig. 2.10, the red cluster refers to the Mountain area presenting the highest elevation, whereas the yellow, blue and cyan, refer, respectively, to the Piedmont between two ranges of mountains, to the Mediterranean affected by sea forcing, and to the Northeastern hilly terrain.

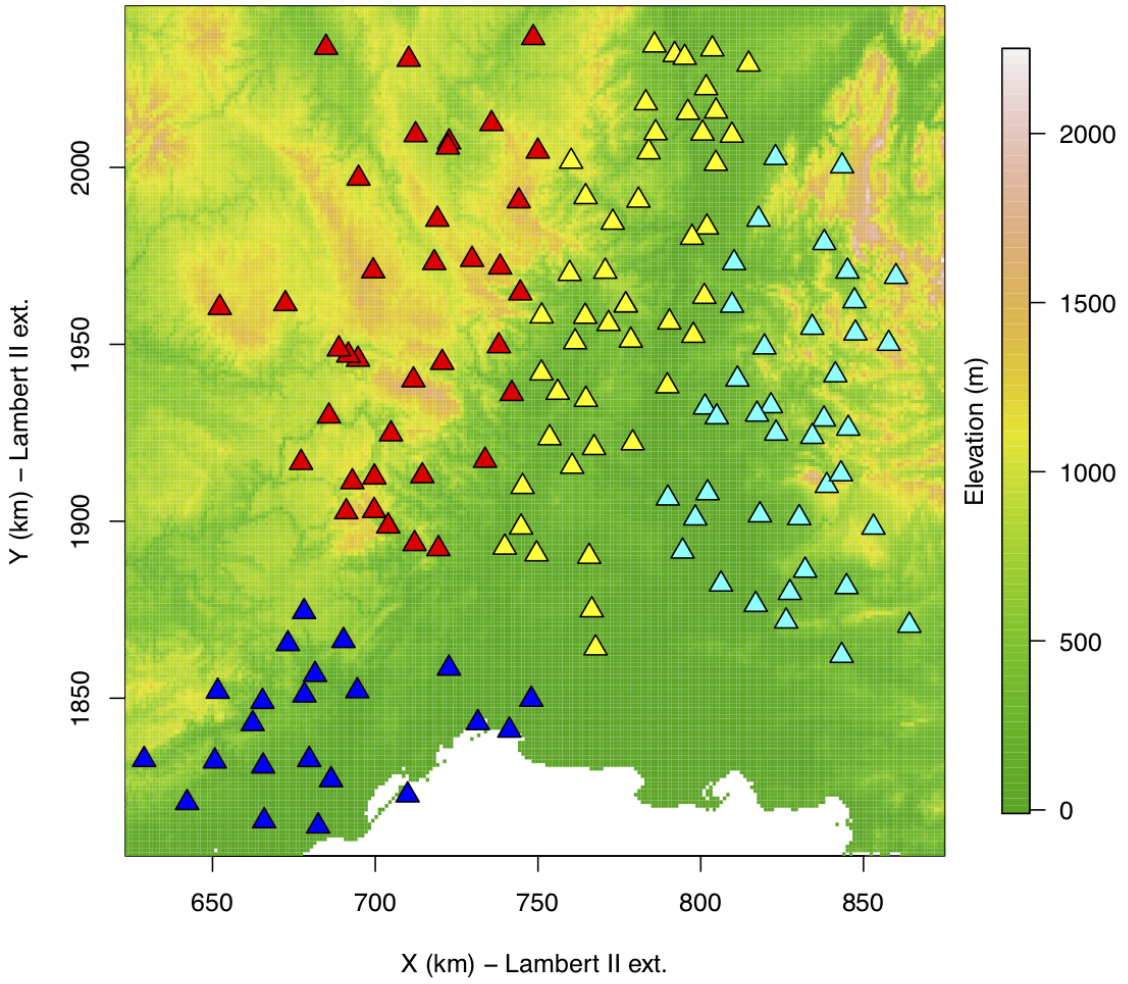


Figure 2.10: 146 rain gauge stations in the Cévennes-Vivarais region are partitioned into 4 clusters which are located in the Mediterranean area (in blue), the hilly area (in cyan), the Piedmont area (in yellow) and the Mountain area (in red).

Table 2.3: Partition of the studied area.

zone	Conventional name	Geographic feature	Location	Number of stations
1	Gard and Hérault	Mediterranean	South West	22
2	Drôme	Hills	North East	37
3	Ardèche	Piedmont	Central North	47
4	Haute-Loire and Lozère	Mountain	North West	40

Statistical analysis

A first climatological analysis is proposed to examine the class features. The monthly average precipitation is given, in Fig. 2.11, for the 4 zones from 2005 to 2014, based on the hourly precipitation data. As usually mentioned, autumn presents the highest amount of precipitation whatever the area [*Nuissier et al.*, 2011; *Molinié et al.*, 2012].

With the Piedmont area (zone 3), the Mountain area (zone 4) presents the highest monthly average precipitation, in November, associated with orographic lifting leading to local precipitation enhancement that may contribute to 40% of the rainfall regime as shown by *Godart et al.* [2011]. The Mediterranean area (zone 1) presents important precipitation variability, associated with synoptic conditions and sea-surface-temperature that strongly modulate moisture content in the atmosphere and thus the resulting precipitation [see *Nuissier et al.*, 2008; *Lebeaupin et al.*, 2006].

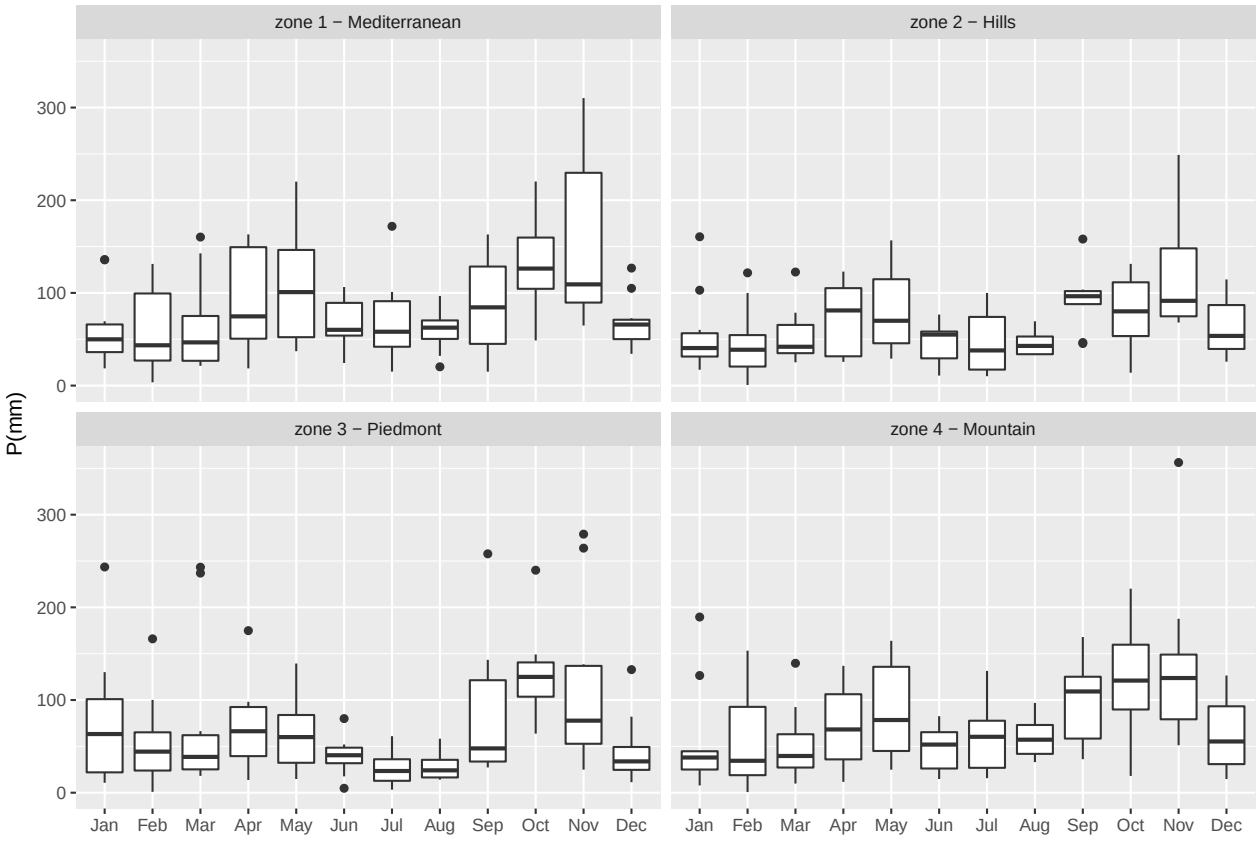


Figure 2.11: Monthly precipitation evolution in the 4 zones from hourly precipitation 2005-2014 data set.

Figure 2.12 presents the comparison of the monthly average precipitation among the 4 zones. During summer, the Mediterranean area presents higher precipitation amounts than other areas, and the Piedmont area has much less rain on average. On the contrary, the Piedmont area accumulates more precipitation in the beginning of the year (such as January, February and March). In October, the Hills area has much less precipitation than other areas because of the topological feature.

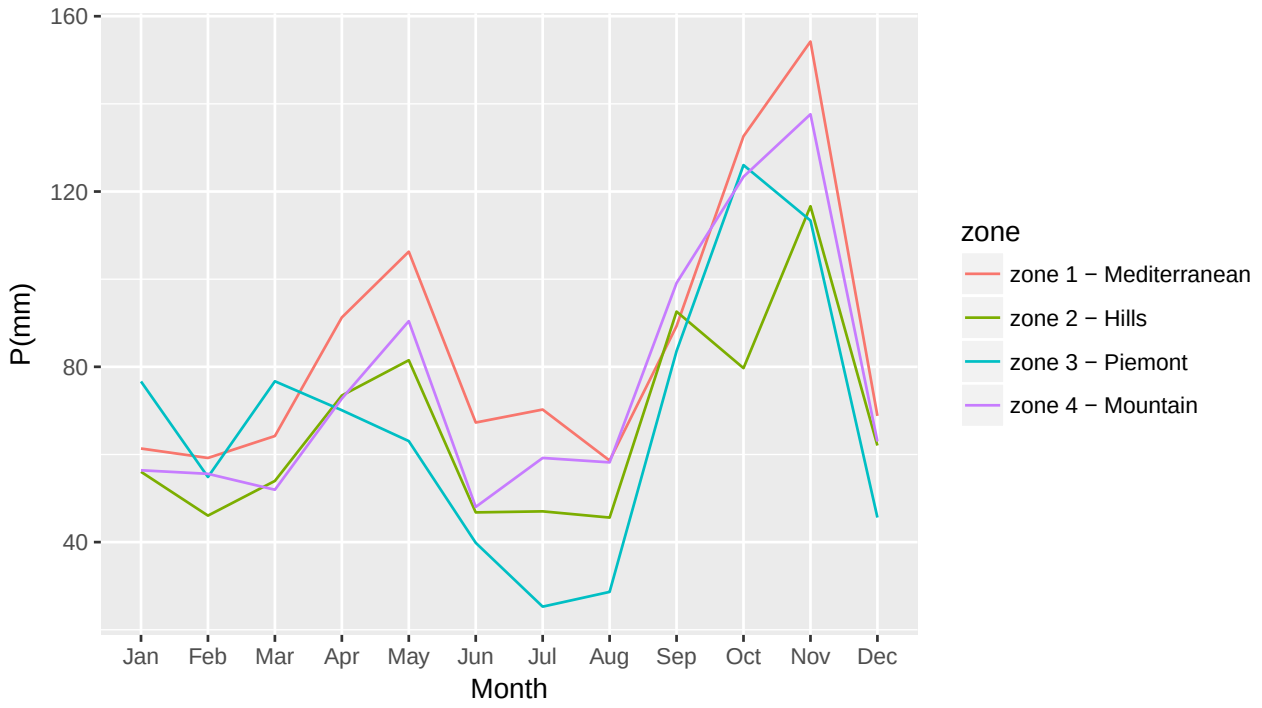


Figure 2.12: Comparison of the monthly average precipitation in the 4 zones from 2005 to 2014 using hourly precipitation data.

Due to the different topological features and climate conditions, the precipitation fields over the Cévennes-Vivarais region can not be considered as a homogeneous rainfall field. Hourly precipitation data from 2005 to 2014 have been selected to analyze rainfall situation in the Cévennes-Vivarais region. The 146 rain gauge stations in the Cévennes-Vivarais region are partitioned into 4 relatively homogeneous rainfall zones, using the k-means technique.

2.2.2 Hydrological model inputs

As mentioned in the introduction, our target is to provide near-surface meteorological variables that will be latter used as inputs of hydrological models. Such models are a simplification of a real-world system that help in the understanding, the predicting and the managing of water resources. The water balance may be seen as the simplest way to model water flows in an hydrological system. Its equation is:

$$P = R + E + \Delta S. \tag{2.5}$$

where P is the precipitation, E is the (actual) evapotranspiration, R is the stream-flow and ΔS is the change in storage (as soil moisture or as groundwater in aquifers). The precipitation data have been described in Section 2.2.1. To solve the water balance, we need an estimation of evaporation.

Evaporation

A most accepted standard formula to estimate it is the Penman equation. *Penman* [1948] combined the energy balance with the mass transfer method and derived an equation to compute the evaporation from an open water surface from standard climatological records of sunshine, temperature, humidity and wind speed. The original equation was developed by Howard Penman at the Rothamsted Experimental Station, Harpenden, UK [see *Penman*, 1948]. The Penman's equation for evaporation is:

$$E_{\text{mass}} = \frac{mR_n + \rho_a c_p \delta_e g_a}{\lambda_v (m + \gamma)} \tag{2.6}$$

where

$m =$ Slope of the saturation vapor pressure curve $(Pa \cdot K^{-1})$
$R_n =$ Net irradiance $(W \cdot m^{-2})$
$\rho_a =$ density of air $(kg \cdot m^{-3})$
$c_p =$ heat capacity of air $(J \cdot kg^{-1} \cdot K^{-1})$
$\delta_e =$ water vapor pressure deficit (Pa)
$g_a =$ momentum surface aerodynamic conductance $(m \cdot s^{-1})$
$\lambda_v =$ latent heat of vaporization $(J \cdot kg^{-1})$
$\gamma =$ psychrometric constant $(Pa \cdot K^{-1})$

which will give the evaporation E_{mass} in $kg/(m^2 \cdot s)$.

In words, evaporation from open water surfaces, like the ocean or lakes, can be simply estimated as proportional to the difference between the saturation vapor pressure at the water surface and the actual vapor pressure above the surface, although the higher the wind speed and the greater the turbulent mixing of the atmosphere, the greater the evaporation. In Equation 2.6, most of parameters are constants or accessible variables such as net irradiation. The water vapor pressure deficit is the difference between the absolute humidity and the absolute humidity at saturation. The absolute humidity at saturation is mostly a temperature dependent quantity. This so-called combination method was further developed by many researchers into many variants, especially suitable extended to cultivated surfaces by introducing resistance factors.

Here, it is enough to say that to serve a water balance model, we need five main near-surface meteorological variables which are *Precipitation*, *Temperature*, *Solar radiation*, *Wind speed*, *Absolute humidity = Water vapor pressure*. Using the same variables, it will also be possible to serve other energy-related issues like estimating the share of snow and rain in precipitation and snow-melt rate. So these five variables are the target of this PhD work. [XXX rephrase] All these near-surface meteorological variables data (except of precipitation data from the OHMCV collection of local raingauges data) are available in ERA-Interim. To illustrate the yearly evolution of the wind speed, the solar radiation, the temperature and the water vapor pressure data in the 4 different zones, the associated $0.75° \times 0.75°$ grid is chosen in each zone (Fig. 2.13). The 6-hours data from 2005 to 2014 in each grid are accumulated into daily data for wind speed, solar radiation, temperature and water vapor pressure.

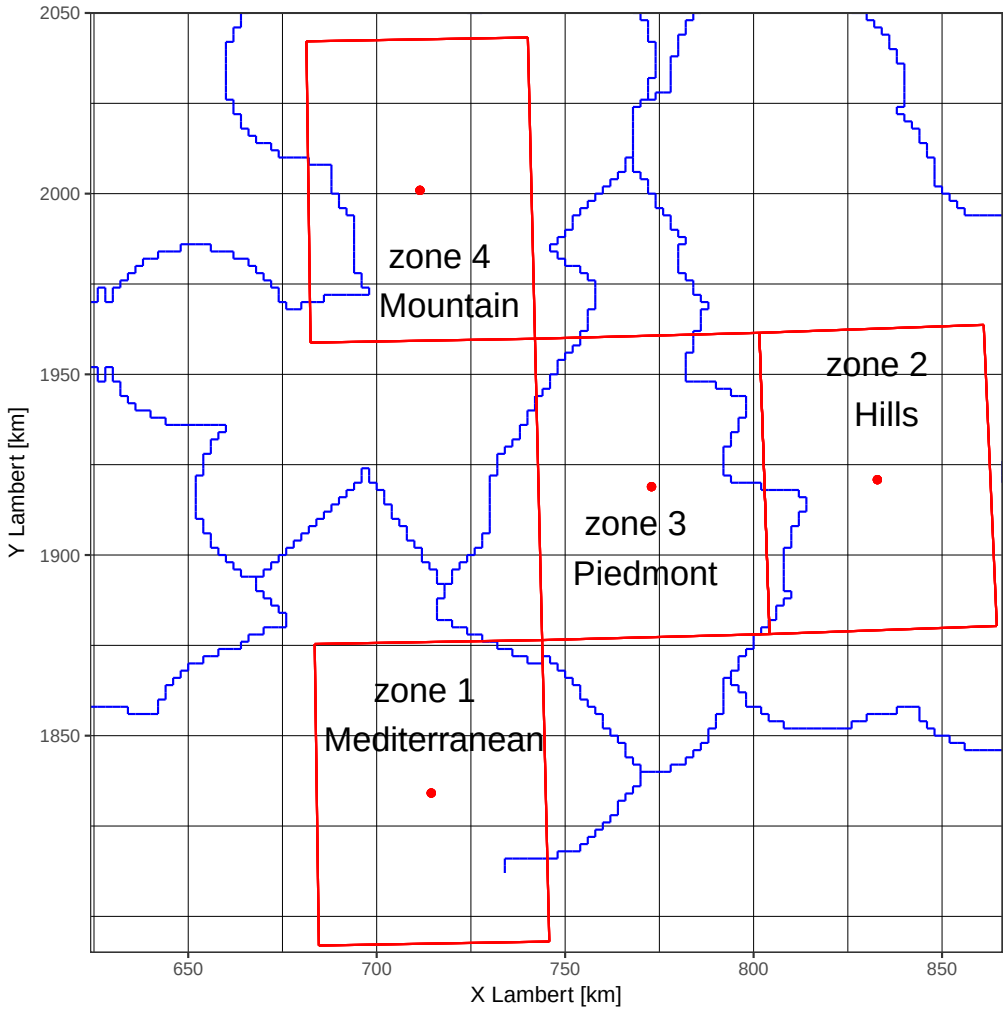


Figure 2.13: Location of ERA-Interim pixel (red rectangle) associated with each zone Rectangle (in red) is the $0.75° \times 0.75°$ *grid associated to each zones. Points (in red) are the centers of the grids. [XXX-NO Need new figure]*

Figure 2.14 - 2.17 present the monthly evolution and its associated variability, of the wind speed, the solar radiation, the temperature and the water vapor pressure, at the 4-zone scale.

Even if globally and due to loose resolution of the ERA-Interim reanalysis, the meteorological variables in the 4 different zones present a similar behaviour as some discrepancies appear, again confirming the necessity to split the whole region in 4 areas. Indeed, the wind speed variability is the highest in the Piemont, probably due to the topographical feature presenting a succession of deep valleys that locally may enhance divergence or convergence of the incoming flow [*Anquetin et al.*, 2003]. As expected, solar radiation, temperature and water vapor pressure show a very strong seasonal evolution during the year. Some discrepancies appear when comparing the 4 zones. In particular, the solar radiation strongly differs within the 4 zones at the period of the year of maximum solar radiation (spring-summer); likewise for the temperature, the 4 zones present the largest differences in winter, when the topography plays an important role of the temperature evolution, and

in summer-fall, the water vapor pressure is slightly different from one zone to the other as the precipitation is.

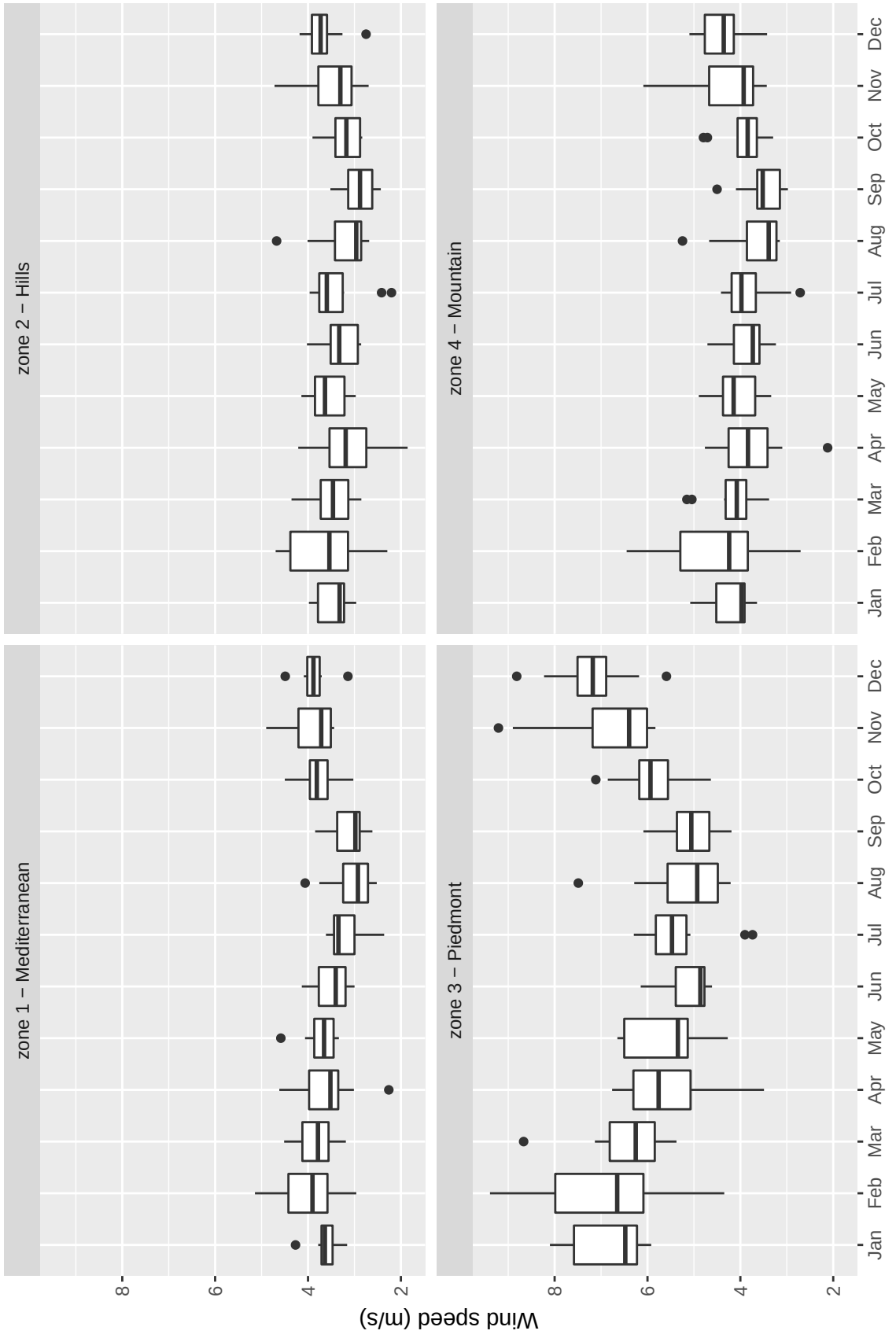


Figure 2.14: Monthly wind speed evolution in the 4 zones, and during the 2005-2014 period.

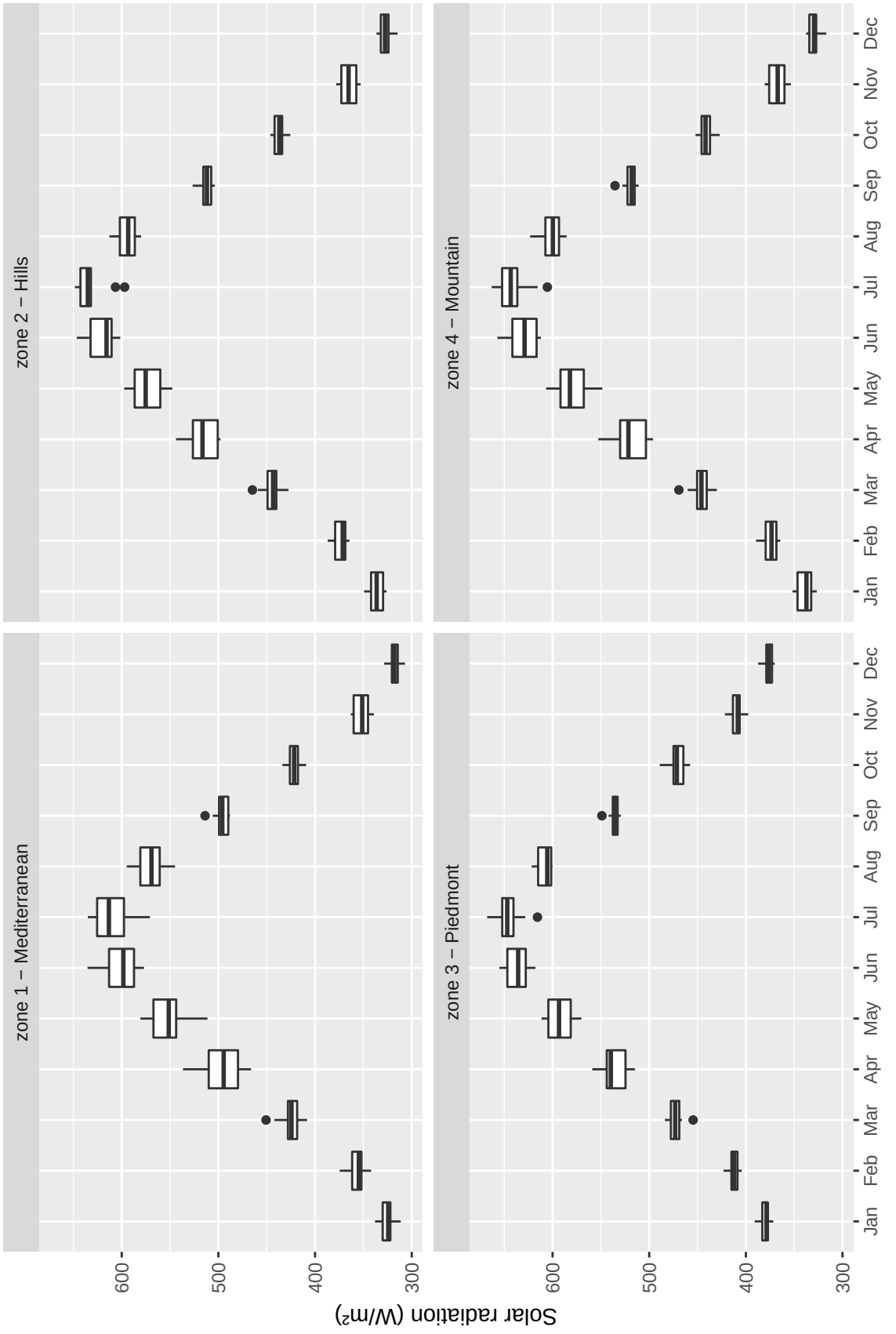


Figure 2.15: Monthly solar radiation evolution in the 4 zones, and during the 2005-2014 period.

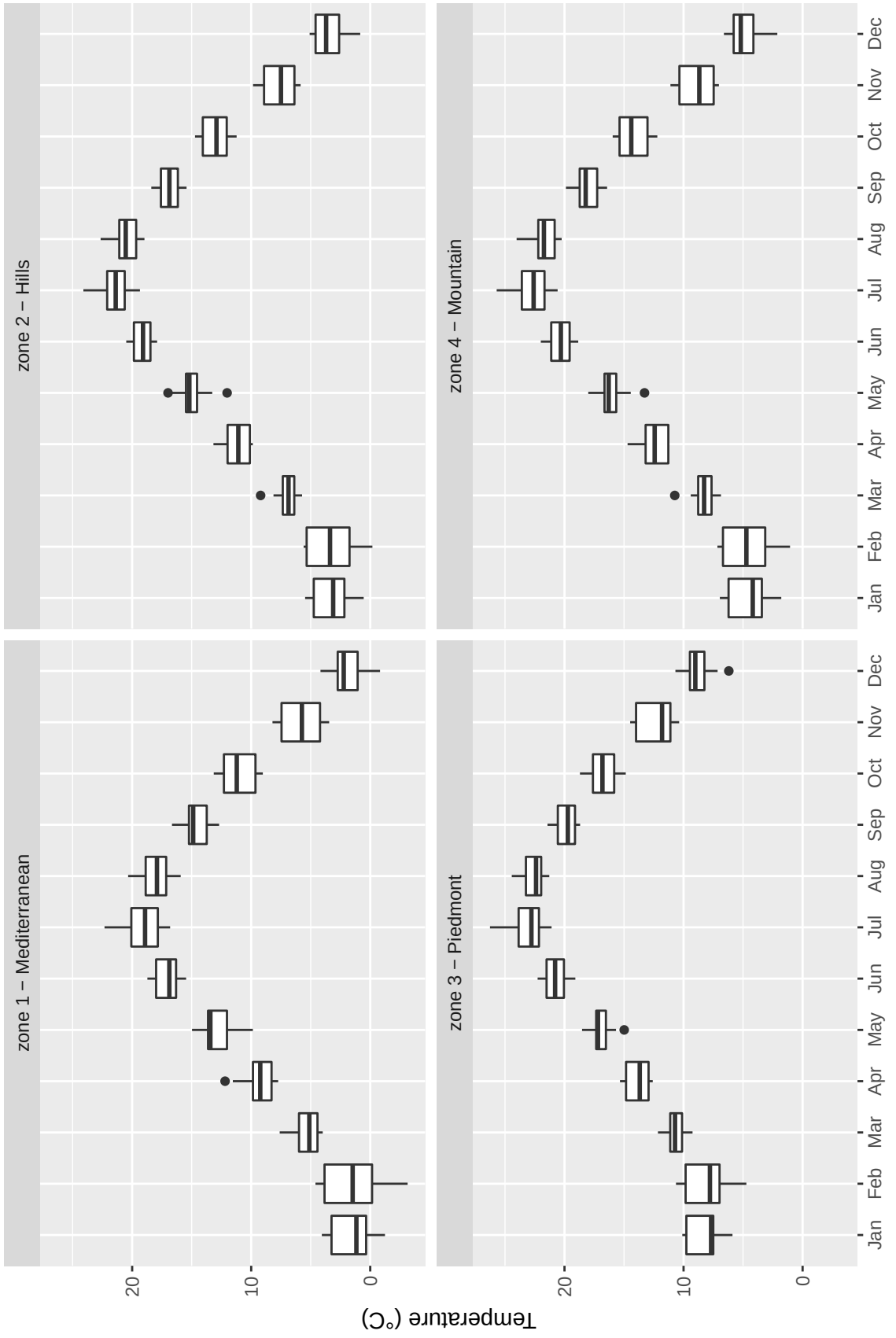


Figure 2.16: Monthly temperature evolution in the 4 zones, and during the 2005-2014 period.

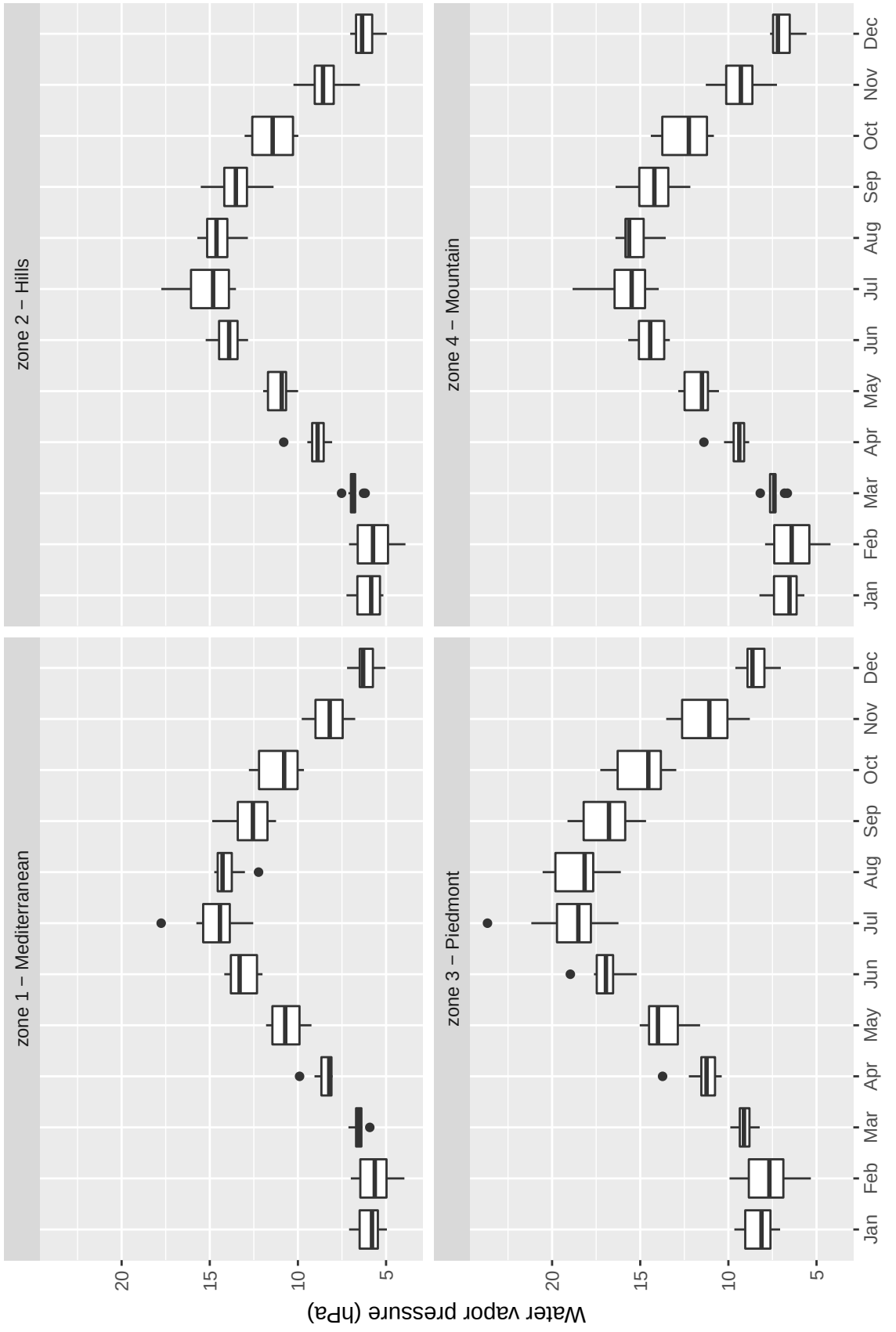


Figure 2.17: Monthly water vapor pressure evolution in the 4 zones, and during the 2005-2014 period.

2.3 Conclusions

The main objective of this work is to produce stochastic simulations of input variables for hydrology, these being first precipitation, then wind speed, solar radiation, temperature and water vapor pressure, all for the estimation of evaporation, precipitation phase and snowmelt. The study area "Cévennes-Vivarais " is located in a well documented mountainous terrain which contains several different climates and topological situations. This region benefits of a long-term and qualified meteorological observation data, as necessary to implement the simulations.

In the next chapter, we shall define the research questions to address.

CHAPTER 3
MAIN CHALLENGES

The hydrological models are used to manage water resources and assess hydrological hazards. Application of hydrological models requires input variables such as precipitation, temperature, solar radiation, water vapor pressure and wind speed which must be modeled together.

Among all these input variables, precipitation is the most important because of its direct link to water. On the basis of SAMPO (**S**imulation of **A**dvected **M**esoscale **P**recipitations and their **O**ccurrence), a stochastic rainfall generator developped at Irstea which can simulate spatio-temporal precipitation over a homogeneous domain, a first main challenge and suggestion made to this PhD work was to expand SAMPO to make it possible to simulate precipitation over a much larger, non-homogeneous domain.

A second main challenge is obviously to build a multivariate model to generate the simulations of all input variables needed for a hydrological model. However, each input variable has its own proper physical behaviour and distinct statistical properties.

In the following Section 3.1, by comparing the different approaches, the stochastic weather generators are considered as the direction of the research modelling. A local stochastic rainfall generator is thus introduced in Section 3.2. In Section 3.3 and 3.4, two major research problems are highlighted that will guide this work.

3.1 Stochastic simulations

Before introducing the stochastic rainfall generator SAMPO, the choice of the stochastic approach rather the physical approach in this PhD work must be justified. There are two different types of climate models for the multivariate simulations:

1. physically based atmospheric models (e.g., several existing French models: Meso-NH [*Lafore et al.*, 1997], LMDZ [*Hourdin et al.*, 2006] and AROME [*Seity et al.*, 2011]);
2. stochastic weather generators (SWG).

The physically based atmospheric models use the physical theory described with mathematical equations, numerically solved. The simulated atmospheric variables are the state

variables (i.e. pressure, temperature, wind) and the variables associated with the water cycle (variables related to all phases of water, would it be water vapor, liquid or solid; for the condensed phases it can be in suspension or precipitating).

On the other hand, the stochastic weather generators are statistical models. By using statistical tools with observed atmospheric data, the weather generators aim at producing simulations with similar statistical properties as observation. Though the history of the developments of the physical models (1920s) is longer than the stochastic models (1950s), there are several disadvantages by using the physical models comparing to the stochastic models:

- The physical models are based on the computation of numerous meteorological variables, describing the whole atmospherical volume above the region of interest. The stochastic models are built only on the concerned variables.

- Spatial and temporal resolutions are still an important problem for the physical models, even nowadays where computer power allows better resolution to address impact issues with more accuracy. On the contrary, the spatial and temporal resolutions of stochastic models are by design the ones of the observed data. Therefore, the stochastic models generate more easily the simulations with finer spatial and temporal resolution (e.g., hourly or even sub-hourly resolution for the time scale and 1 km resolution for the space scale, depending on the observation network).

- The computation time and cost are not negligible for running the physical models; this is still a considerable issue since the structure of the physical models are more and more complex. Generally, the physical models need hours, or even days to complete one simulation ; this may come along with a tendency of communities involved in physical atmospheric modelisation to investigate detailed events more than exploring the behavioral variability of the system over long runs.

For these reasons, the stochastic approach has been chosen in this PhD work.

3.2 SAMPO

This section aims at presenting our local stochastic rainfall generator called **SAMPO** (**S**imulation of **A**dvected **M**esoscale **P**recipitations and their **O**ccurrence) developed by *Leblois and Creutin* [2013]. SAMPO uses the Turning Bands Method (TBM), introduced in its general form by *Matheron* [1973] and popularized for 2-*D* applications in hydrology by *Mantoglou and Wilson* [1982]. *Lepioufle* [2009] and *Lepioufle et al.* [2012] gave the statistical principles to estimate rainfall parameters in a context of spatial rainfall accumulating over time, that the rainfall simulation aims to reproduce. *Renard et al.* [2011] and *Labbas* [2015] used extension of the code for conditional simulation. SAMPO only works with one homogeneous zone.

The idea of considering separately the rainfall occurrences and rainfall amounts and one homogeneous rainfall type at a time is quite general in several models, but usually these

two processes are modeled in very different statistical ways [see *Richardson*, 1981; *Racsko et al.*, 1991; *Bárdossy and Plate*, 1991; *Breinl et al.*, 2015].

SAMPO contributes to the stochastic approach with the particularity of adapting a classical Gaussian random field generator, the Turning Bands Method (TBM), to simulate advected intermittent rainfall fields. The relying assumptions are intimately related to the time and space resolution of the simulated fields. The objective of SAMPO is to simulate rainfall fields at a resolution of typically 10 min and 1 km^2 over domains of several thousands of square kilometers, where intermittency and advection are relevant features.

3.2.1 Turning Bands Method

Here is how *Mantoglou and Wilson* [1982] introduced TBM:

> The Turning Bands Method(TBM) for the simulation of multidimensional random fields is presented. These fields commonly occur in the Monte Carlo simulation of hydrologic processes, particularly groundwater flow and mass transport. The general TBM equations for two- and three-dimensional fields are derived with particular emphasis on the more complicated two-dimensional case. For stationary two-dimensional fields the uni-dimensional line process is generated by a simple spectral method, a technique which can be generally applied to any two-dimensional covariance function and which is easily extended to anisotropic and areal averaged processes. Theoretically and by example the TBM is shown to be ergodic even for a finite number of lines, and it is demonstrated that it rapidly converges to the true statistics of the field. Guide lines are presented for the selection of model parameters which will be helpful in the design of simulation experiments. The TBM is compared to other methods in terms of cost and accuracy, demonstrating that the TBM is as accurate as and much less expensive than multidimensional spectral techniques and more accurate than the most expensive approaches which use matrix inversion, such as the nearest neighbor approach. The uni-dimensional spectral technique presented here permits, for the first time, the inexpensive and accurate TBM simulation of any proper two-dimensional covariance function and should be of some help in the stochastic analysis of hydrologic processes.

SAMPO uses the geostatistical TBM in 3D to generate series of rainy periods with homogeneous statistical properties. Rainfall fields are constructed from the product of two independent fields (Fig. 3.1): (i) a Boolean indicator field representing pixels with zero and non-zero rainfall; and (ii) a field of non-zero precipitation generated from a pre-specified distribution.

The TBM simulation depends on parameters describing the at-site rainfall distribution (e.g., mean and variance of a log-normal distribution) and the spatio-temporal properties of the observed rainfall fields (e.g., the spatio-temporal variogram). SAMPO commonly uses the inverse Gaussian distribution [*Chhikara and Folks*, 1974].

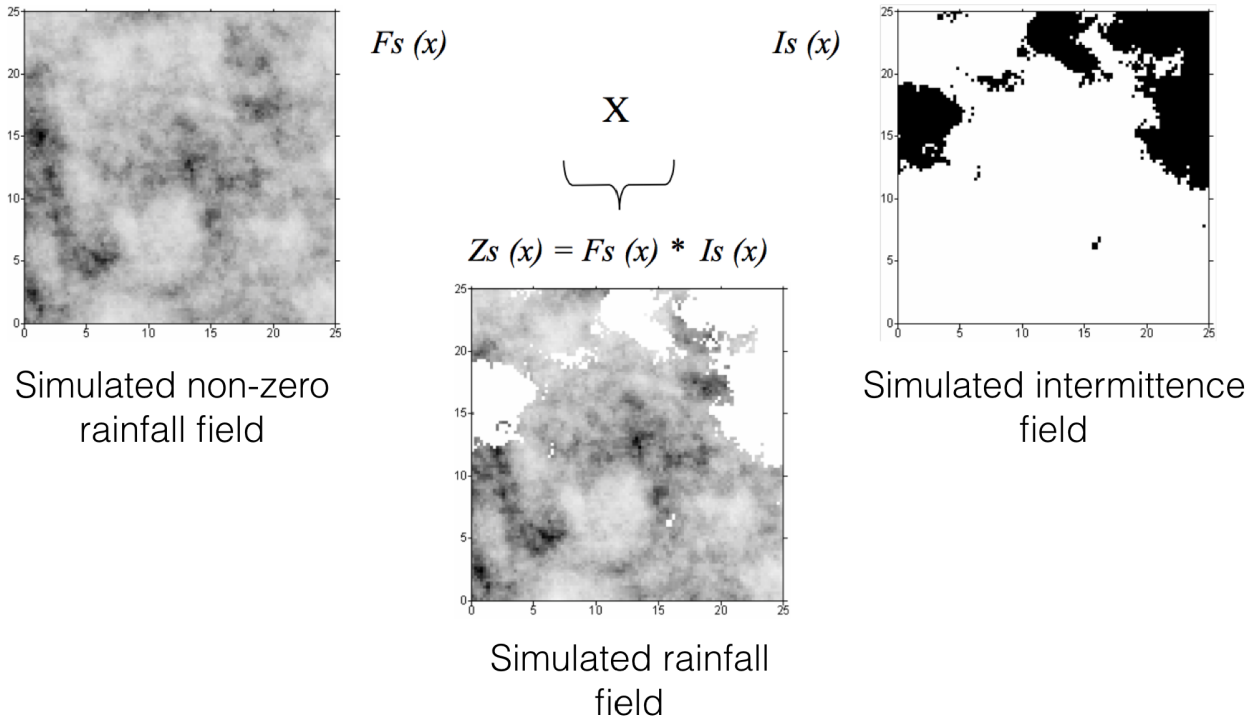


Figure 3.1: Diagram of the simulation of rainfall field which are constructed as the product of two independent fields: $F_s(x)$ - simulated non-zero rainfall field; $I_s(x)$ - simulated intermittence field; $Z_s(x)$ - Final rainfall field obtained. In the simulated intermittence field, the non-rainfall zone is indicated in black. Source: modified from [Domingues-Ramos, 2002].

The general formulation of the intermittent rainfall fields R_I proposed by *Leblois and Creutin* [2013] takes the following form:

$$R_I(\mathbf{x}_E, t) = \varphi(Y_R(\mathbf{x}_L, t.U_R))\mathbf{1}_{Y_I(\mathbf{x}_L, t.U_I)\geq\lambda} \tag{3.1}$$

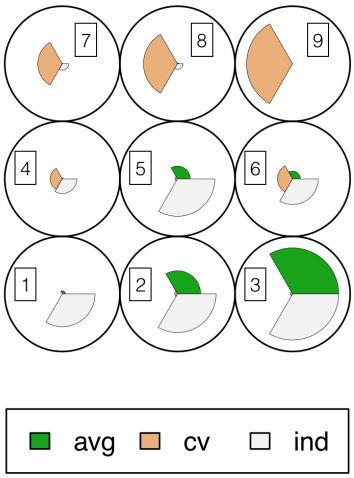
where Y_R and Y_I are the two independent Gaussian functions used to represent nonzero rainfall and intermittency with U_R and U_I featuring their respective dynamics; φ stands for the anamorphosis used to care about the skewed distribution; λ characterizes the fraction of intermittency; the combined use of Lagrangian $\mathbf{x}_L$ and Eulerian $\mathbf{x}_E$ coordinates takes care about advection. The separate specification of nonzero rainfall, intermittency and advection is easy to carry on and gives flexibility to control the properties of the resulting compound field. *Lepioufle* [2009]; *Lepioufle et al.* [2012]; *Creutin et al.* [2015] gave the principal for analyzing raingauge or weather radar data to these models.

3.2.2 Calendar of rainfall types

A second part of SAMPO represents the time variation of the rain field statistical properties from one homogeneous period to the next, including the alternation of rainy and dry periods. As this separation of observed rainfall into homogeneous rainfall chunks is an important starting point for the work in this thesis, and as it was not yet formally published but in French language reports, short notes or academic dissertations (e.g., *Leblois*

[2012]), it will be described below with some details. The obtained sequence of regional rainfall periods directly governs the statistical properties of rainfall accumulation over long time steps (weeks to years). The situation of a homogeneous rainfall field at one time step called **rainfall type** is characterized by several rainfall descriptors such as the average rainfall intensity, the rainfall variation, the average indicator function or the spatial structure. SAMPO simulates the occurrence of rainy periods in time using a Hidden semi-Markov's Model [*Baum and Petrie*, 1966; *Barbu and Limnios*, 2008] applied to rainfall types classified by a Kohonen's Self-Organization Map [*Kohonen*, 2001] (called the Kohonen algorithm or SOM). More details of Self-Organization Map will be shown in Part II and Appendix B.

The Kohonen algorithm is applied to create clusters regrouping *similar* (in the descriptors' sense) time steps. An attractive feature of this algorithm is that the clustering is performed with a *neighboring constraint* between clusters. In other words, clusters are organized so that neighboring clusters are similar, while distant clusters are more markedly different. This organization creates a map where each cell represents a cluster. Figure 3.2 is a simple example of using Self-Organization Map to classify time-steps documented over a set of raingauges by time-series of rainfall precipitations into 9 rainfall types. The rain-



(a)

	avg	cv	ind
code1	0.6473	1.0827	0.7082
code2	4.5814	1.2289	0.8203
code3	9.0694	1.1943	0.9624
code4	0.2032	1.7115	0.3637
code5	2.5378	1.2065	0.7982
code6	1.5537	1.8621	0.5737
code7	0.1262	2.3528	0.2050
code8	0.2228	2.8385	0.1907
code9	0.0335	3.4628	0.0840

(b)

Figure 3.2: (a) Diagram of self-organization map with 9 clusters. Each cluster is represented by three descriptors which are the average rainfall intensity (avg), the coefficient of variation (cv) and the indicator function (ind). (b) The code values of three descriptors for 9 clusters.

fall type 3 represents the time steps when there are high precipitation intensity and rainy everywhere. The rainfall type 6 represents the time steps when there are medium rainfall intensity with small variation at about half the space. The rainfall type 9 represents the time steps when there are no rains almost everywhere but there are still some locations with low precipitation. Classes may seem numerous, but this can be justified on both the demand and offer sides. On the demand side, rainfall types combines several aspects (intensity, variability, coverage, etc.) and even a coarse distinction on each aspect will necessarily generate a number of combinations. On the offer side, the rainfall types are not defined

over one sole time-series observed at one raingauge, but over a set of many raingauges sampling the same rainfall events; this later fact is in favor of a reasonable information / parameter ratio.

Such mapping-with-continuity of rainfall types is suitable to possible further grouping, as is common in SOM approaches. However, both the number of clusters and the topology of the map need to be provided to the algorithm. Such information may be difficult to specify and require additional analyses. In the case of rain, besides absolutely dry time-steps that are not included in the classification, the separation of classes is largely conventional. The coupling of SAMPO and SOM algorithms presented above is quite straightforward. Once the rainfall types are identified through Kohonen's classification and the appropriate transition and emission matrices of a (seasonal) hidden Markov model are identified, the simulated time series of rainfall types called **calendar** will be obtained by considering the weather types. And then we can use the TBM simulator to produce a space-time simulation of rain-fields following the rainfall types prescribed by the calendar.

3.3 Heterogeneity problem of rainfall field

Since the 1970s, stochastic rainfall models are discussed and developed a lot [e.g., *Todorovic and Woolhiser*, 1975; *Bras and Rodríguez-Iturbe*, 1976; *Katz*, 1977]. **Stochastic Weather Generators (SWG)** are statistical models that aim at quickly simulating realistic random sequences of atmospheric variables such as precipitation, temperature and wind. We can find some classic models and different approaches in these review articles [*Wilks and Wilby*, 1999; *Srikanthan and Mcmahon*, 2001; *Ailliot et al.*, 2015].

[XXX-NON rephrase] For large region of interest, the one of most challenging problems in the construction of SWG when we want to simulate in a large surface is the regional heterogeneity, would the origin of this fact be orographic or meteorological. Prior to the 1980s, heterogeneity was most often addressed (when recognized) by partitioning a spatial field into relatively homogeneous regions [see *Guttorp and Sampson*, 1994]. *Hingray* [2003] reviewed the different existing stochastic weather models, in particular, he distinguished the difference between the multi-site models and the spatial models. Multi-site models faithfully reproduce rainfall variability at a discrete set of locations, but don't describe the spatial variations in a continuous way. As a continuous description is very important to the simulation of rainfall which displays one of the largest variability among meteorological variables in time and space, we prefer to model a space-time weather generator that can simulate the variables continuously in time and space, but the challenge is more complicated.

Before solving this heterogeneity problem, the notion of **homogeneity** and **heterogeneity** must be well defined, because with different concepts, the definition can be changed. Several criterion can be considered, for example the different climate, the different topography or the combination of different aspects. On a statistical point of view, we define a relatively homogeneous rainfall field by considering all stations in this field to have similar precipitation distribution, in the sense of rank correlation (see Section 2.2.1). Here, we

insist that the definition of homogeneous rainfall field is **relative**, as using different rainfall duration, or focusing on a specific season, could slightly change the boundaries of homogeneous zones within the rainfall field.

With our choice of definition of relatively homogeneity, a large heterogeneous rainfall field can be partitioned into several relatively homogeneous rainfall zones (e.g., Fig. 2.10). Each zone can be represented with its calendar of rainfall types. Based on the above situation, the main challenge in this PhD work for resolving heterogeneity problem is to provide joint simulated calendars of locally homogeneous rainfall types for use in SAMPO. The question is how to coordinate the calendars of each zone in time and space.

3.4 Multivariate analysis for hydrological input variables

Water resources and hydrological modeling projects typically involve simulating systems made up of many component parts, strongly interrelated, and in some cases, poorly characterized. *Fatichi et al.* [2016] currently reviewed the applications, challenges, and future trends in distributed process-based models in hydrology. In most situations, the hydrological system is driven by stochastic variables (i.e. precipitation, potential evapotranspiration, etc.) and still involves uncertain processes and parameters. Recent articles [e.g., *Devia et al.*, 2015; *Sood and Smakhtin*, 2015] reviewed several types of hydrological models highlighting the important numbers of inputs required for handling simulations as accurate as possible. These inputs concern topography, soil characteristics, vegetation, land surface classification, and meteorological forcings. Runoff observations are used for evaluation.

Thus, it is vital to provide long term simulations of the meteorological input variables for hydrological models. The essential concern of generating well distributed simulations of multivariate context is how to formally describe the relationships between variables and their relevance to the problem being studied. Multivariate statistics is a form of statistics encompassing the simultaneous observation and analysis of more than one outcome variable.

There are two major problems associated with the multivariate modeling. The first one concerns the representation of the observed data distributions as described by the multivariate model. The hydrometeorological data (precipitation, wind speed, cloud cover, relative humidity, etc.) often turn out to be non-Gaussian, which belong to bounded or skewed distributions [*Schoelzel and Friederichs*, 2008]. So, attention needs to be paid on individual distributions and it may be necessary to discuss the dependence of most of these hydrometeorological variables to figure out their relationships. The second difficulty deals with the identification of the multivariate joint distribution. The textbook of *Anderson* [1958] educated a generation of theorists and applied statisticians. In this book, we can find the essential problems and theories for multivariate analysis, especially, it gives us clear notions and methods to deal with standard situations (e.g., the multivariate normal distribution) and simple cases (e.g., the general linear hypothesis).

In hydrology, the development of the multivariate model began with the seminal paper

of *Richardson* [1981] where the author modified the point rainfall generator described in *Katz* [1977] by the introduction of three other meteorological variables, the minimum and maximum temperatures and the solar radiation given at the daily time scale. This was the first stochastic weather generator which took into account simultaneously four meterological variables. Since then, many other works developed with the same idea. Two articles [*Wilks and Wilby*, 1999; *Srikanthan and Mcmahon*, 2001; *Hao and Singh*, 2016] reviewed this long list of multivariate approaches.

Several methods are now available. As examples, multivariate Gaussian mixture models [*Marin et al.*, 2005; *McLachlan and Peel*, 2004] rely on parametric descriptions of the relationship among the variables, or the Bézier distribution [*Wagner and Wilson*, 1995], or the Johnson distribution [*Johnson*, 1949]. Within these approaches, multivariate extensions of the log-normal or gamma distribution are not possible. More generally, it appears that not all distributions are suitable to multi-dimension extensions, and it is matter of fact that variables to be linked usually have different distributions.

Given these limitation, resorting to homogeneous multi-dimensional distributions appears to not be the one and final good solution in general. Instead, the use of so called copulas [e.g., *Nelsen*, 2007] became more and more popular. The main advantage of the copula approach for hydrology relies on the selection of an appropriate model for the dependence among the variables, represented by the copula, that proceed independently from the choice of the marginal distributions. The copula approach can be seen as a simple and straight forward method to find parametric descriptions of multivariate distributions in a context of non-normally distributed random variables, eventually achieving the program that was preliminary addressed by multidimensional approaches. Today, copulas have now a strong record of applications, among others in finance and climatology. They are also used in hydrology [e.g., *Genest and Favre*, 2007; *Bárdossy and Li*, 2008; *Schoelzel and Friederichs*, 2008; *Erhardt et al.*, 2015; *Evin et al.*, 2017]. *Georgakakos and Kavvas* [1987] pointed out an important and another interesting point of view when other meteorological variables are combined with precipitation in stochastic simulations. They reported that doing so, it improved the capability of the models to capture the structure of the precipitation and it also facilitated the assessment of the effect of climate change on the precipitation structure, as including some physical drivers as covariates could help a multivariate simulation to hold over a variety of contexts. Given this general context, this PhD work will develop its own multivariate modeling using the copula technique.

3.5 Conclusions

The main objective of this PhD work is to contribute to build and to evaluate a strategy of climate regionalization targeting the specific needs for hydrology. The approach should be able to be generalized in order to be applied in other territories under different climatic conditions. The strategy of regionalization needs to respect the resonance in spatial and temporal scales associating with atmospheric and hydrological phenomena. A simulation in present climate is first targeted. The possibility of application under climate change must

be kept in mind.

Two different major problems are introduced. The heterogeneity of rainfall field is first discussed and appear to not have a real good solution among the existing models. There are two types of rainfall models for capturing the spatial variability of entire rainfall field which are spatial and multi-site models. Flexible but spatially discontinuous multi-site models have less advantage for hydrological use (e.g., urban drainage risk). On the other hand, spatial models do not have much flexibility to capture all different kinds of spatial situations and variabilities and are often much simplified, and this is also true for SAMPO.

Another problem is how to build a multivariate model to provide the hydrometeorological input variables such as precipitation, wind speed, solar radiation, temperature and water vapor pressure for hydrological models. The inputs must to be statistically consistent, within a large range of time and space scales, with observation data. As is well-known, the multivariate joint distribution remains one of the major difficulties for multivariate modeling.

These two problems will be addressed in the two next parts of the thesis.

Part II
Heterogeneity problem of rainfall field

CHAPTER 4
STATE OF THE ART

Part II deals with the heterogeneity problem of rainfall field pointed out in Section 3.3. This chapter first reviews the state of the art on this issue. Chapter 5 proposes two approaches which are parametric and non-parametric to resolve the problem. Chapter 6 diagnoses the statistical properties of two proposed approaches in time and space. Chapter 7 presents a third approach based on the copula technique and a disaggregation method. Chapter 8 gives some conclusions.

4.1 Motivations

As mentioned in Section 3.1, this PhD work is based on stochastic simulations of rainfall or other meterological variables. Several overview articles [*Wilks and Wilby*, 1999; *Srikanthan and Mcmahon*, 2001; *Ailliot et al.*, 2015] reviewed different types of stochastic weather generators. Stochastic weather generators can be classified into two categories which are rainfall models and multivariate models. Rainfall models can also be classified into three categories which are (1) single-site models, (2) multi-site models and (3) spatio-temporal models. For the heterogeneity problem, we concentrate on the rainfall models. *Srikanthan and Mcmahon* [2001] gave an extensive review of rainfall models, including multisite network modelling. *Hingray* [2003] reviewed different modelling approaches developed over the last decades for the generation of space-time rainfall fields. Table 4.1 presents a overview of available stochastic weather generators of different categories.

In hydrological models, we often need spatially distributed processes, the spatial dependence between the weather inputs at different sites has to be accommodated. Especially, this is very important to the simulation of rainfall which displays the largest variability among meteorological variables in time and space. But we can't ignore the utility of the multisite models and some useful techniques in these models that may also help us in modeling spatial correlations. For example, *Georgakakos and Kavvas* [1987] showed that the multisite precipitation models were essential to allow the characterization of the precipitation process over spatial domains of area similar to medium sized river basins (e.g., hundreds to thousand of km^2).

Table 4.1: Overview of available stochastic weather generators.

Reference	Daily	Hourly	Single site	Multi-site	Spatial	Precipitation	Multivariable	Method
Richardson and Wright [1984]	×		×			×	T_{max}, T_{min}, solar radiation	Two-part process
Cox and Isham [1988]	×	×		×		×		Poisson-cluster processes
Gupta and Waymire [1990]	×				×	×		Scaling property
Racsko et al. [1991]	×		×			×	temperature, solar radiation	Two-part process
Wilson et al. [1992]	×			×		×		Hierarchical model
Hughes et al. [1999]	×			×		×		non-homogeneous HMM
Wilks [1999]	×			×		×	T_{max}, T_{min}, solar radiation	WGEN, spatially correlated random numbers
Wojcik et al. [2000]	×			×		×	×	Resampling
Bouvier et al. [2003]	×	×			×	×		Empirical Orthogonal Function based model
Kilsby et al. [2007]	×		×			×	temperature, humidity, wind, sun shine, derivation of potential evapotranspiration	Neyman Scott Rectangular Pulses model
Lennartsson et al. [2008]	×		×			×		Generalized Pareto distribution
Bárdossy and Pegram [2009]	×			×		×		Copula
Ailliot et al. [2009]	×			×		×		HMM, censored Gaussian distributions
Flecher et al. [2010]	×		×			×	T_{max}, T_{min}, global radiation, wind speed	Multivariate closed skew-normal distributions
Kleiber et al. [2012]	×				×	×		Kriging
Leblois and Creutin [2013]	×	×			×	×		TBM
Oriani et al. [2014]	×				×	×		Resampling
Wilks [2014]	×	×	×				×	Copula
Verdin et al. [2015]	×				×	×	T_{max}, T_{min}	GLM-based method

In the next section, a few multisite and spatial models are shortly introduced.

4.2 Existing solutions

4.2.1 Kriging models

A daily spatio-temporal precipitation model, similar to the approach of *Wilks* [1998, 2009], is proposed by *Kleiber et al.* [2012]. They generated spatially and temporally correlated precipitation in two steps. The first step was dedicated to the generation of the precipitation occurrence with a latent Gaussian process. The simulation of precipitation amounts was then performed, in the second step, by means of a transformed Gaussian process. This is similar to the approach used in SAMPO. At individual locations, Kleiber's model reduced to a Markov chain for precipitation occurrence and a gamma distribution for precipitation intensity, allowing statistical parameters to be included in a generalized linear model (GLM) framework. Statistical parameters were modeled as spatial Gaussian processes, which allowed for interpolation to locations where there are no direct observations via kriging [*Matheron*, 1963; *Cressie*, 1992]. Kriging models, as a spatiotemporal model, minimize and produce an uncertainty estimation at any location.

In the model of *Kleiber et al.* [2012], the estimations of both precipitation occurrence and intensity are obtained by the use of an isotropic correlation function whose scale parameter varies with time. For much more complex terrain or larger domain, alternative anisotropic and non-stationary models [*Baigorria et al.*, 2007] are more preferable. Another problem appears when extreme rainfall events are concerned; in their context the gamma distribution is no more suitable. *Kleiber et al.* [2012] suggested that it would be desirable to combine their model with that of *Buishand et al.* [2008] who described a model for spatially correlated extreme precipitation. And also, their approach requires no extra effort to incorporate the important low-frequency behavior and interannual variability required of precipitation generators. *Verdin et al.* [2015] proposed some extended works and developed a GLM-based spatial weather generator which combined the precipitation and temperature generator. The model can generate sequences at any arbitrary location.

4.2.2 Conditional models

Charles et al. [1999] extended the non-homogeneous hidden-state Markov model of *Hughes et al.* [1999] by incorporating rainfall amounts. The joint distribution of daily rainfall at n sites was evaluated through the specification of n conditional distributions for each weather state $(s = 1, \ldots, N)$. The conditional distributions consist of regressions of transformed amounts at a given site on precipitation occurrence at neighboring sites within a set radius (e.g., d km). An automatic variable selection procedure was used to identify the key neighboring sites. The precipitation model can be expressed as

$$z_s^{(i)} = \theta_{0s}^{(i)} + \sum_{k \in n_i(d)} \theta_{ks}^{(i)} + \varepsilon_s^{(i)} \qquad i = 1, \ldots, n$$

where the $\theta_{ks}^{(i)}$ are regression parameters, $n_i(d)$ denotes the set of indices of the key neighboring sites for site i, $\varepsilon_s^{(i)}$ is an error term modeled stochastically by assuming $\varepsilon_s^{(i)} \sim \mathcal{N}(0, \sigma_s^2(i))$, and

$$z_s^{(i)} = \Phi^{-1}\{F(y_s^{(i)})\}$$

in which F denotes the normal cumulative distribution function and $F(y_s^{(i)})$ is the empirical distribution function of $y_s^{(i)}$, the rainfall amount on days with $r^{(i)} = 1$.

Another conditional model is proposed by *Bárdossy and Plate* [1992], they developed a multi-dimensional stochastic model for the space-time distribution of daily rainfall linked to atmospheric circulation patterns using conditional distributions and conditional spatial covariance functions. The model is a transformed multivariate first-order auto-regressive model with parameters depending on the atmospheric circulation patterns. The negative values are declared as dry days.

4.2.3 Copula-based multisite models

Bárdossy and Pegram [2009] proposed the use of the multivariate copula to relate spatial and temporal observation trends at many sites. As the models previously discussed, the spatial and temporal dependence structure of hydrometeorological data sets are more complex than using conventional correlation of the multivariate normal. *Thomas and Fiering* [1962]; *Matalas* [1967] used the normal score transform for multi-site stochastic simulation in hydrology at the earliest. The consideration of using multivariate copula instead of the classical normal score transform is that the classical one is not rich enough to capture the range of pair-wise correlations being strong at high rainfall values and weak at low rainfall amounts.

Example of the precipitations in two pair stations

Extracted from *Bárdossy and Pegram* [2009], Figure 4.1 shows the locations of the observations used to build the Copula-based multisite model.

Same as Fig. 4.1, extracted from *Bárdossy and Pegram* [2009], Figure 4.2 shows two examples of empirical copulas derived from scatter-plots of pairs of daily recording rain gauges (stations 1 and 23 as shown in Fig. 4.1). The empirical densities are given for two seasons, Fig. 4.2**(a)**, for summer (June, July, and August) and, **(b)**, for winter (December, January, and February). As illustrated by the positions of the horizontal and vertical lines, the summer appears to be drier. These figures exhibit a constant density for the "dry" conditions located in the lower left corner for both seasons. The upper left and lower right quadrants show the one-dimensional marginal densities of the wet gauge given that the other is dry. All the conditional distributions are therefore clearly identified in only one figure.

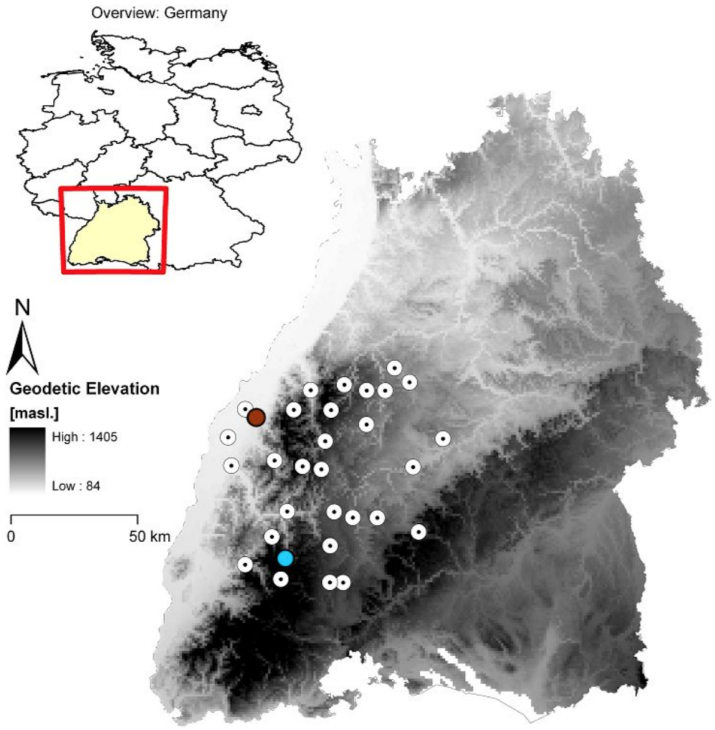


Figure 4.1: Locations of the rain gauge stations indicated by circled dots, around the Black Forest within the German state of Baden-Württemberg used in this study; shading darkens with increasing altitude above sea level. Stations 1 and 23 are colored brown and blue respectively. Source: [Bárdossy and Pegram, 2009]

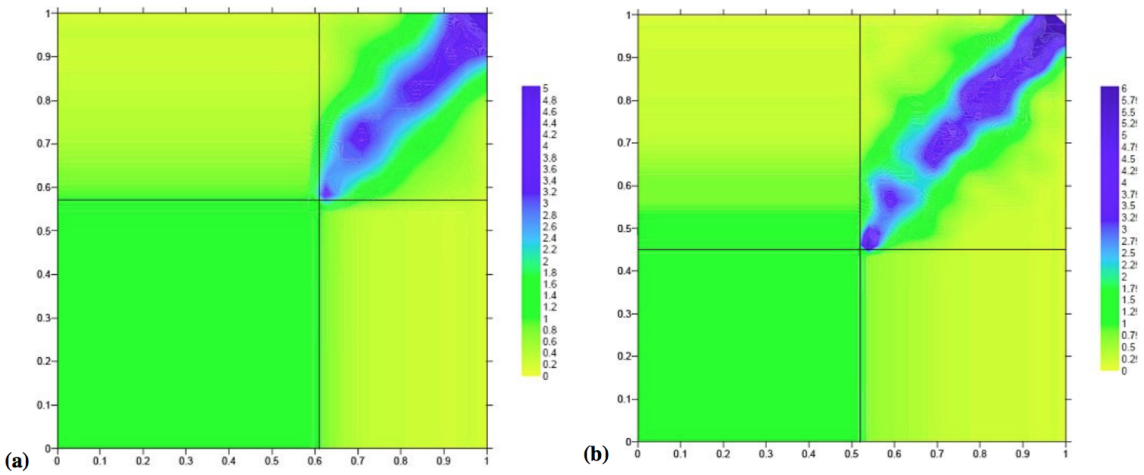


*Figure 4.2: Two sample copulas for station pairs 1 and 23, the left panel **(a)** for Summer (June to August) and the right panel **(b)** for Winter (December to February), more wet. The horizontal and vertical lines indicate the probability limits for the dry/wet boundaries. Source: [Bárdossy and Pegram, 2009]*

4.2.4 Condorcet model

The method was first proposed in a Master thesis [*Ollagnier*, 2013], and then developed and reported in an unpublished scientific rapport [*Leblois*, 2014]. The idea behind this model is to combine two or several non-homogeneous areas into large-scale consensus state by the use of mutual information.

Example: coordinating local atmospheric models together

Similarity of two weather-type sequences considered jointly can be summarized by a contingency matrix that may demonstrate the tendency of given states to occur at the same time, as the contingency matrix identify conditional probabilities of states in one sequence given the state in the other sequence.

The criteria to express the strength of the dependency is the Mutual Information (MI). Formally, the mutual information of two discrete random variables X and Y is defined as:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} \mathbb{P}(x,y) \log \left(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)} \right). \tag{4.1}$$

where $\mathbb{P}$ is the probability function.

Among a set of sequence, maximum mutual information determaines the most similar sequences, to possibly hierarchically aggregate. As an illustration, Figure 4.3 represents a control plot showing the successive linkages established between 5 local atmospheric weather classes systems for five countries: DE (Germany), DK (Denmark), FR (France), NO (Norway) and UK (United Kingdom).

An important by-product of the Condorcet model is a table of historical consensus states at all aggregation levels. As an illustration, Figure 4.4 shows an excerpt of the Condorcet model outputs used to coordinate the weather types over 5 countries ; the weather types were determined based on ERA-Interim on a 4 time-steps in a day basis.

The Condorcet model may thus be used to coordinate the rainfall types, separately identified at each time step for each homogeneous region. The coordination could thus provide the rainfall pattern at large scale, also for a large heterogeneous areas, gathering n homogeneous regions states into one large scale state.

Limitations

The Condorcet model is a good solution to describe the state of a heterogeneous region. But it gives too much priority to the spatial aspect, neglecting to consider temporal coherence ; if we try do simulate the top-level and then local states conditionally to the top-level state, temporal coherence is readily lost by the successive decisions taken. Starting from this evidence, the next chapter tries to find a compromise that would preserve both space and time correlations.

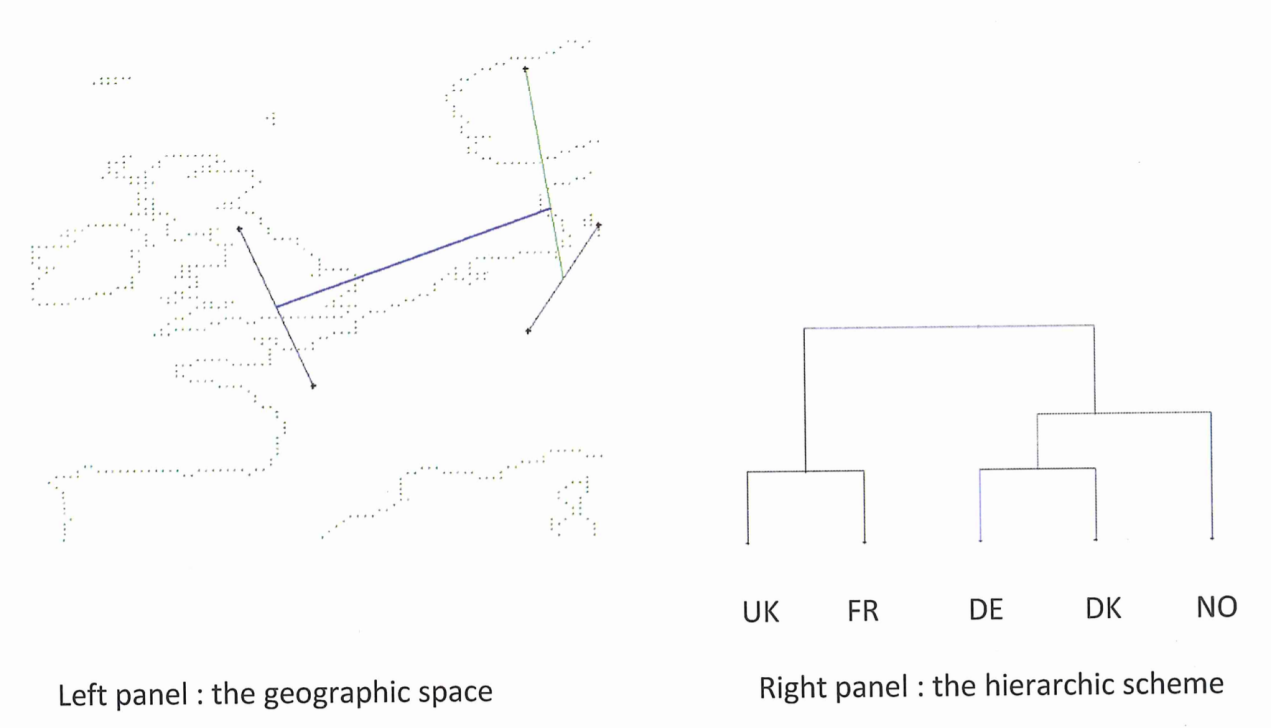


Figure 4.3: Coordinating atmospheric weather classes over five countries by means of Condorcet model. Source: [Leblois, 2014]

Date\Group	#001	#002	#003	#004	#005	#006	#007	#008	#009
[...]									
08/12/2013	34	32	1	28	8	5	7	3	12
08/12/2013	34	26	1	22	1	7	7	6	12
08/12/2013	34	26	1	22	1	7	7	6	12
08/12/2013	34	26	1	22	1	7	7	6	12
09/12/2013	34	6	1	28	1	5	7	9	22
09/12/2013	35	6	12	4	1	5	18	9	22
09/12/2013	10	6	12	4	1	5	18	9	22
09/12/2013	11	29	12	4	31	19	18	19	23
10/12/2013	11	29	12	4	36	23	18	25	19
10/12/2013	11	29	12	4	6	23	18	25	19
10/12/2013	11	22	12	15	3	29	18	33	20
10/12/2013	11	28	12	15	3	29	18	16	20
11/12/2013	11	28	12	2	3	29	18	16	20
11/12/2013	11	22	12	2	10	18	18	13	20
11/12/2013	11	29	12	2	4	25	18	25	19
11/12/2013	11	34	12	2	4	25	18	16	20
12/12/2013	11	34	12	2	10	18	18	13	20
12/12/2013	29	3	12	2	10	18	18	13	20
12/12/2013	29	29	12	33	10	17	18	21	20
12/12/2013	28	35	1	33	11	17	23	21	15
13/12/2013	30	14	1	33	11	17	8	2	14
13/12/2013	28	14	1	33	12	17	23	2	26
13/12/2013	27	28	13	33	12	17	19	16	19
13/12/2013	4	6	2	33	11	17	20	21	9
14/12/2013	4	6	34	34	23	4	8	8	7
14/12/2013	5	24	34	34	12	4	17	8	13
14/12/2013	5	20	7	21	6	4	18	17	20
14/12/2013	33	27	12	27	3	12	18	13	20

5 nation with 36 weather classes
#001: UK
#002: FR
#003: DE
#004: DK
#005: NO

#006: DE + DK
MI = 2.47

#007: UK + FR
MI = 1.98

#008: (DE + DK) + NO
MI = 1.67

#009: (DE + DK + NO) + (UK + FR)
MI = 1.34

Figure 4.4: Excerpt of a table of historical states at all aggregation levels. Each column has 36 weather classes (which are not the same in each column). Source: [Leblois, 2014].

CHAPTER 5

METHODOLOGY I: COORDINATION OF RAINFALL TYPES CALENDARS

A new approach is proposed in this chapter to better take into account the heterogeneity with stochastic rainfall simulator. As mentioned in Section 3.1, SAMPO is a spatio-temporal rainfall simulator which only generates homogeneous rainfall fields. The model input is a time-series of rainfall types (called calendar), provided in our case by means of self-organizing map (SOM). Given a rainfall structure lead by different weather conditions, it is not realistic to apply SAMPO over such rainfall field under the assumption of homogeneity since it will probably demonstrate a clear heterogeneous behaviour. However, by using clustering method, a heterogeneous rainfall field can be partitioned into several homogeneous rainfall zones (e.g., Fig. 2.10). Thus, SAMPO can be applied over each homogeneous rainfall zones. A calendar is thus created for each homogeneous zones, using the precipitation observations located in the zone (Section 3.2.2). These calendars are *observed calendars*. A interesting feature is that the rainfall classes system almost surely differ in each zone to accommodate local rainfall regime. The objective of this chapter is to coordinate these observed calendars in time and space, to later simultaneously generate *simulated calendars* for each homogeneous region, and finally to import the simulated calendars into SAMPO to generate separately the spatio-temporal simulations for each homogeneous rainfall zone. The spatio-temporal rainfall simulation over the whole region is then obtained combining all the simulated homogeneous rainfalls.

One parametric approach (coupled hidden Markov model) and one non-parametric approach (resampling model) are introduced in this chapter to deal with the coordination of rainfall-type calendars. Figure 5.1 presents the complete methodology used for the coordination.

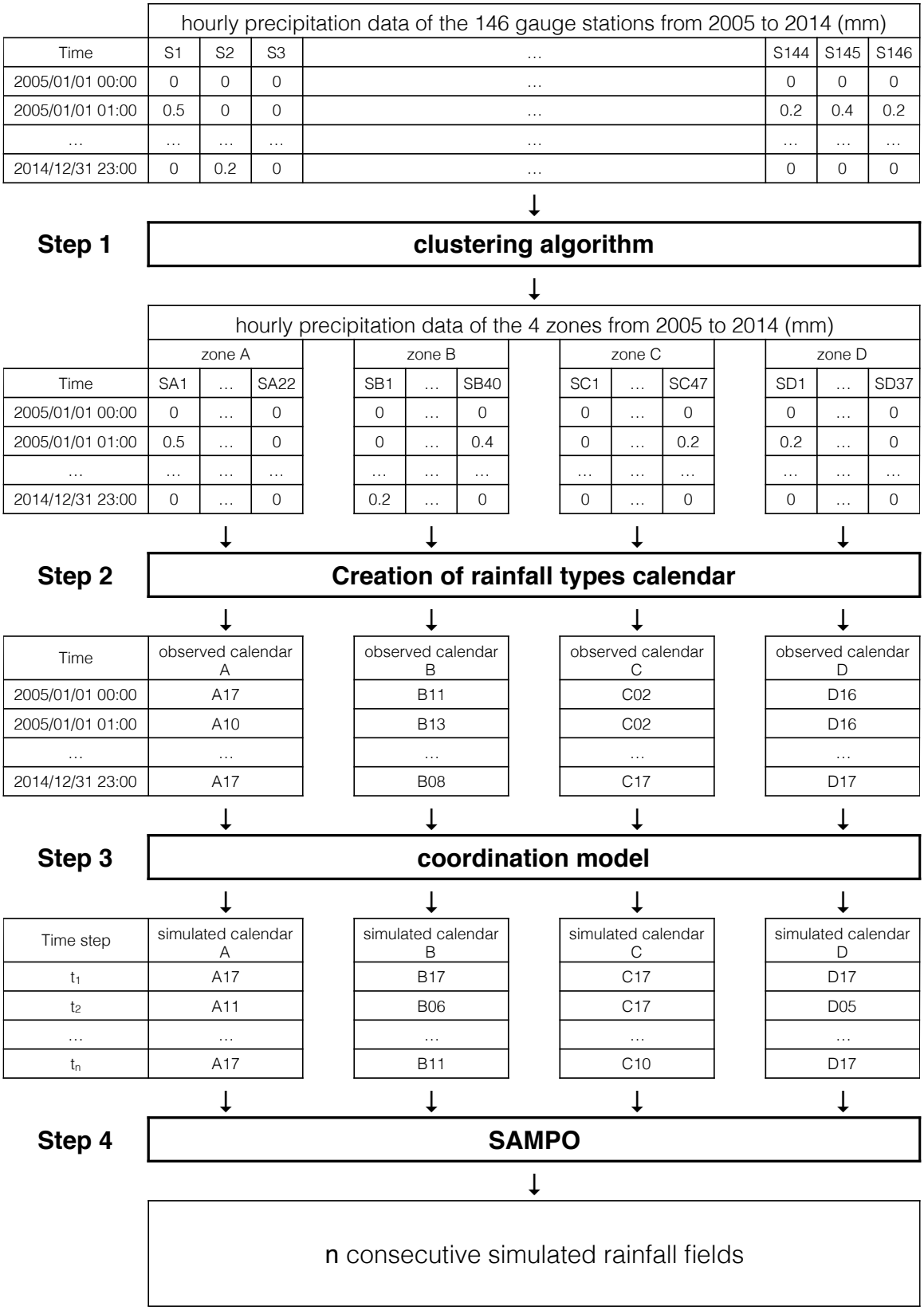


Figure 5.1: Framework of the coordination of rainfall-type calendars.
***Step 1**: Determination of the homogeneous rainfall regions using the clustering algorithm to partition the 146 hourly raingauge stations; **Step 2**: Creation of the observed calendars, one for each individual zone, using the hourly precipitation data; **Step 3**: Coordination of the observed calendars and generation of the simulated calendars, one for each individual zone; **Step 4**: Simulation of the rainfall fields over the whole region.*

As mentioned in the previous chapters, the region of interest is the Cévennes-Vivarais (Fig. 2.1) that benefits of long term observation data. In Fig. 2.2, the locations of the 146 hourly rain gauge stations are shown; this work uses their records from 2005 to 2014. As presented in Section 2.2, the partition of the whole area in several homogeneous regions was done based on the *k*-mean clustering algorithm of these records and lead to 4 homogeneous zones (Fig. 2.10). Figure 5.2 presents the partition of the whole region into the 4 homogeneous rainfall zones on a nearest neighbor basis. The notations, zone A, B, C and D, are preferred here in this chapter, to facilitate mathematical expression and, thus replace zone 1, 2, 3 and 4 (Table 2.3) respectively.

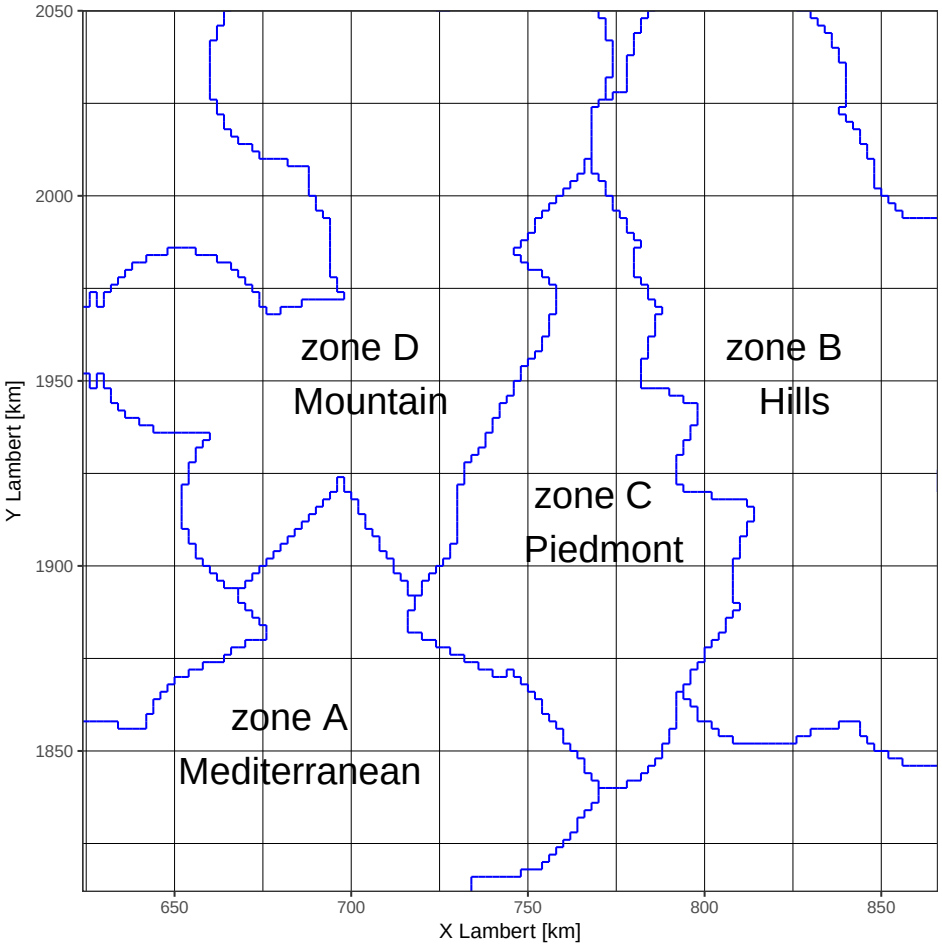


Figure 5.2: Partition of Cévennes-Vivarais region into the 4 homogeneous rainfall zones (blue contours).

5.1 Creation of a rainfall-type calendar for each homogeneous zone

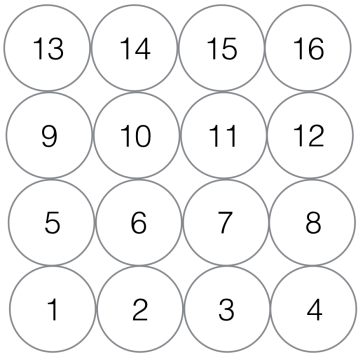
Since each zone is considered as homogeneous, at each time step, the rainfall structure is represented as a *rainfall type* (Section 3.2.2). There are 4 rainfall descriptors used for classification of rainfall type. These descriptors are chosen with their hydrological significance

in mind. They are:

1. the average rainfall intensity over all the gauged stations; it represents the quantity of precipitation over the whole zone.
2. the coefficient of variation of non-zero rainfall intensity; it represents the variability of non-zero rainfall intensity inside the zone.
3. the average indicator function or rainfall intermittency; it represents rainfall frequency over the zone.
4. the Geary's C spatial-correlation coefficient; it represents the spatial structure [*Geary*, 1954; *Khalili et al.*, 2007; *Banerjee et al.*, 2014].

The descriptors are normalized between 0 and 1 to balance their weight in the classification. The classification is made without considering the associated uncertainties. The self-organizing map (Kohonen algorithm) classifies rainy types out of 87648 steps (10 years at hourly time step) into 16 types (or classes). The type 17 is the dry class. Figure 5.3 shows the self-organization maps in each zone with 4 descriptors. Table 5.1 gives the code vectors for the 16 rainfall types with 4 descriptors by using SOM algorithm. The content of each class is relevant to rainfall modeling as in [*Leblois and Creutin*, 2013; *Creutin et al.*, 2015]. However, advection is not considered in the present context; the rainfall is analyzed from a Eulerian perspective (as seen from the ground) and simulated as such.

Table 5.1: Self-organizing map with the 16 rainfall types. **Left**: the order of the 16 rainfall types in self-organizing maps as presented in Fig. 5.3. **Right**: the code vectors for each rainfall type with the 4 descriptors by using Kohonen algorithm, example of zone D. "avg" (in mm) stands for the average rainfall intensity over all the gauged stations; "cv" (without unit) stands for the coefficient of variation of non-zero rainfall intensity; "ind" (without unit) stands for the average indicator function or rainfall intermittency and "GearyC" (without unit) stands for the Geary's C spatial-correlation coefficient. [XXX-NO Units should be indicated]



	avg	cv	ind	GearyC
Type 1	9.3245	1.0768	0.8781	0.6775
Type 2	2.9306	1.0743	0.8236	0.6884
Type 3	1.2771	0.9556	0.7943	0.7525
Type 4	0.4287	1.4303	0.4925	0.8469
Type 5	5.2610	1.1083	0.8517	0.6679
Type 6	1.4608	2.0529	0.4670	0.5035
Type 7	0.2187	2.1798	0.2626	0.6281
Type 8	0.1423	2.9836	0.1550	0.4456
Type 9	0.0536	3.3559	0.0934	4.7428
Type 10	0.0882	3.8550	0.0911	0.3564
Type 11	0.0640	4.9169	0.0579	0.1394
Type 12	0.0103	6.3971	0.0245	1.0000
Type 13	0.0722	2.7793	0.1340	13.5728
Type 14	0.0135	4.6436	0.0455	1.0864
Type 15	0.1169	5.7475	0.0567	0.0642
Type 16	0.0088	6.6673	0.0225	1.0000

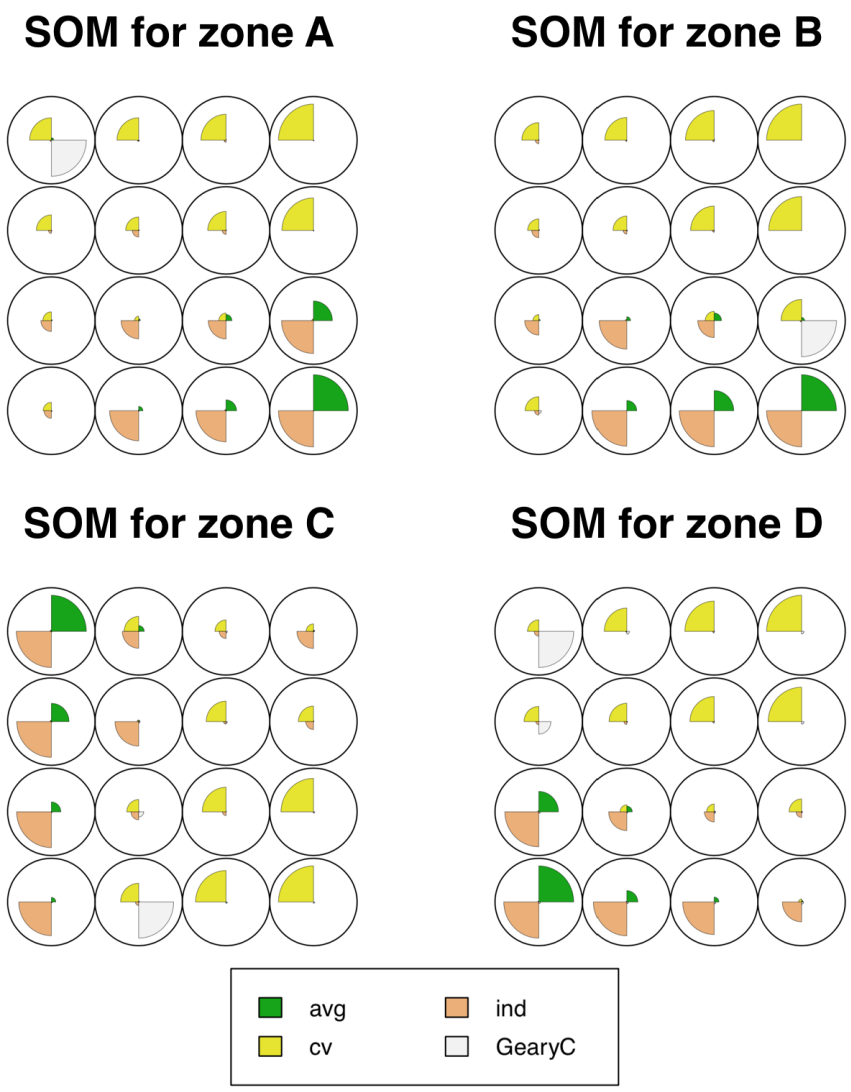


Figure 5.3: Results of the classification, in 16 rainy classes, using Kohonen algorithm for the 4 homogeneous zones. The classification is based on four descriptors: the average total rainfall(avg), the coefficient of variation of non-zero rainfall(cv), the average indicator function(wetness) and the Geary's C spatial-correlation coefficient(gearyC).

The classification thus provides the observed calendars of rainfall types for each zone, as illustrated in Table 5.2. Structurally, the calendar analysis could be conducted for any rainfall duration. In our case, the considered data are:

- hourly observed gauge precipitation at the 146 stations from 01/01/2005 to 31/12/2014 (10 years = 87648 hours);
- 4 homogeneous rainfall zones (A, B, C, D) as a partition of the 146 stations;
- 4 rainfall-type calendars (S_A, S_B, S_C, S_D), one for each of the 4 homogeneous rainfall zones;
- each calendar has its own 17 rainfall types ($\{A1, A2, \ldots, A17\}$ refer to zone A, $\{B1, B2, \ldots, B17\}$ to zone B, $\{C1, C2, \ldots, C17\}$ to zone C and, finally, $\{D1, D2, \ldots, D17\}$ refer to zone D);
- each calendar gathers 87648 time steps.

Table 5.2: Hourly calendar for each of the 4 zones for the 2005-2014 period. There are 17 rainfall types in each zone, the type 17 is the dry class and the other 16 types are the rainy classes.

Time	S_A	S_B	S_C	S_D
2005/01/01 00:00	A17	B11	C02	D16
2005/01/01 01:00	A10	B13	C02	D16
2005/01/01 02:00	A17	B12	C17	D17
2005/01/01 03:00	A17	B04	C02	D03
2005/01/01 04:00	A17	B04	C17	D17
2005/01/01 05:00	A17	B04	C17	D07
2005/01/01 06:00	A11	B13	C16	D07
2005/01/01 07:00	A06	B16	C17	D17
...	...	...	...	...
2014/12/31 23:00	A17	B08	C17	D17
	17 types	17 types	17 types	17 types

At a time step t, the heterogeneous rainfall field at the whole Cévennes-Vivarais scale, is described by (Ai_t, Bi_t, Ci_t, Di_t), $i_t \in [1, 17]$. With 17 different rainfall types in each calendar, there exist $17^4 = 83521$ possible joint rainfall types for (Ai_t, Bi_t, Ci_t, Di_t) at each time step. The 4 calendars can be thus considered as one calendars S_{ABCD} with $17^4 = 83521$ possible joint rainfall types for the entire Cévennes-Vivarais region. Theoretically, S_{ABCD} can be modeled by the Markov model with transition matrix from one time step t to next time step $t + 1$. But the dimension of such transition matrix will be $17^4 \times 17^4$. Not only it will

be difficult to model with actual computer capacity, but anyhow the data don't provide evidence for transition probability.

In the next section, a coordination method is proposed to overcome this dimension problem. It then allows the generation of realistic simulations. The coordination aims at building a model that can generate the simulations corresponding to the observed rainfall-type calendars. That means:

- the spatial synchronicity (Ai_t, Bi_t, Ci_t, Di_t) must be as similar as possible between the simulations and observed rainfall-type calendars;

- the temporal correlation of the simulations in each zone must be as similar as possible to the one of observed rainfall-type calendar in the same zone.

The coordination is built in two phases. In the first one, the spatial synchronicity is modeled for the 4 homogeneous rainfall zones by using hierarchical method and hidden Markov models. In the second phase, a reorganization method is proposed to re-organize the rainfall types sequences which are generated by previous hierarchical hidden Markov models in order to respect the observational evidence for observed length of stay. The reorganization method is applied separately in each homogeneous rainfall zone for optimizing the temporal correlation of rainfall-type simulations. A final check insures that the result is a good approximation to both criteria.

5.2 Hierarchy of homogeneous zones: Coupled Hidden Markov Model

This new approach aims at reducing the bias in temporal correlation by coupling two similar sequences of rainfall-type of different domains while respecting the joint probability matrix. The idea is thus to couple two or several hidden Markov models to obtain one "coupled" hidden Markov model that correctly presents the joint situations.

5.2.1 Hidden Markov Models

This section presents how hidden Markov models (HMMs) can be used in stochastic rainfall modeling. Special cases of (two-state) HMMs for precipitation occurrence were presented by *Foufoula-Georgiou and Lettenmaier* [1987]; *Smith* [1987]. HMMs were later introduced formally as a general mean of modelling single-site and multisite precipitation occurrence data by *Zucchini and Guttorp* [1991]. In their model, precipitation occurrences were assumed to be conditionally independent across the spatial network, given the weather state. Figure 5.4 presents a diagram of hidden Markov model.

Rabiner [1989] is probably the most referenced contribution for hidden Markov models. In this article, the author clearly showed three basic problems of interest that must be solved with such modeling approach, and being useful in real-world applications. These problems are the following:

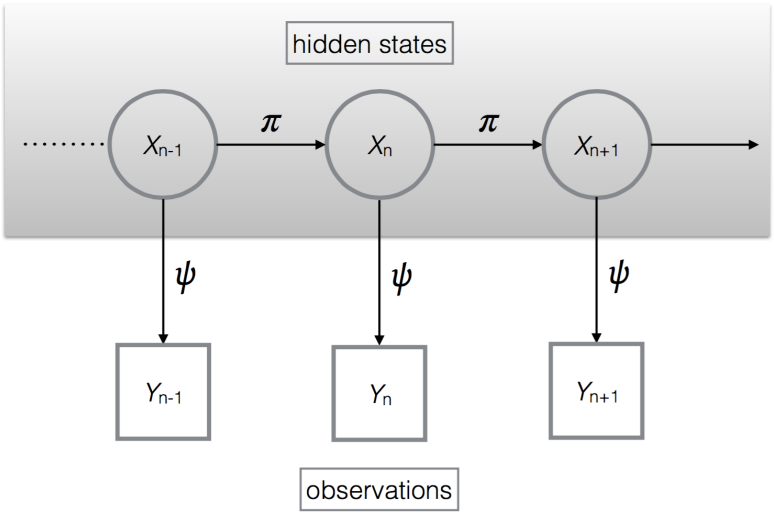


Figure 5.4: Hidden Markov model: $\{X_n\}$ refers to the hidden states, and $\{Y_n\}$ to the observation sequence; π is the matrix of transition probabilities for $\{X_n\}$ and ψ is the matrix of emission probabilities from $\{X_n\}$ to $\{Y_n\}$.

Problem 1: Given the observation sequence $Y = Y_1 Y_2 \dots Y_T$ and a model $\lambda = (\pi, \psi, \theta)$, how do we efficiently compute $P(Y|\lambda)$, the probability of the observation sequence, given the model ?

Problem 2: Given the observation sequence $Y = Y_1 Y_2 \dots Y_T$ and the model λ, how do we choose a corresponding state sequence $X = X_1 X_2 \dots X_T$ which is optimal in some meaningful sense (i.e. best "explains" the observations) ?

Problem 3: How do we adjust the model parameters $\lambda = (\pi, \psi, \theta)$ to maximize $P(Y|\lambda)$?

Rabiner [1989] also discussed explicitly the problems and gave the solution of them. More theoretical contents about HMMs can be found in Appendix A.

- *Problem 1:* Evaluation (Forward algorithm);
- *Problem 2:* Decoding (Viterbi algorithm);
- *Problem 3:* Training (Baum-Welch algorithm).

The idea of coupled hidden Markov models is to model systems of multiple interacting process. The method is firstly mentioned in *Brand* [1997]. In the past 20 years, this method has been very useful in vision and speech applications [e.g., *Natarajan and Nevatia*, 2007; *Nefian et al.*, 2002]. SAMPO uses the single hidden Markov model to simulate rainfall precipitation in one homogeneous field, so the idea is to coupled two (or more) hidden

Markov models of homogeneous rainfall zone to generate a heterogeneous field (Fig. 5.5). The main approach is to reduce the joint situations of two sequences of the hidden states. The modeling part is focused on the hidden states since we want to keep the most informa-

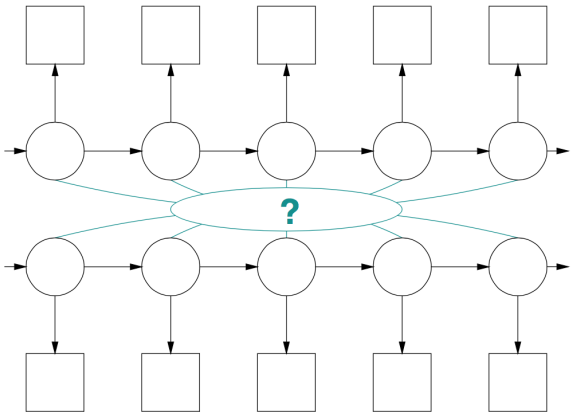


Figure 5.5: How to couple two HMMs ? The circles refer to the hidden states and the squares to the observations. Source: Brand [1997]

tion as possible on the coupled hidden sequence with the least parameters as possible to generate both two hidden sequences. The reason why we want to improve the model with a more complicated model is that the hidden sequence is the Markov chain which gives the temporal property. For example, in Fig. 5.6, there are different types of probabilities between two HMMs, which one will be the most important ? Furthermore, how the temporal correlations remain in these joint probabilities ? We shall also use the singular-value decomposition (SVD) technique to diagnose the performance of this approach (Fig. 5.6).

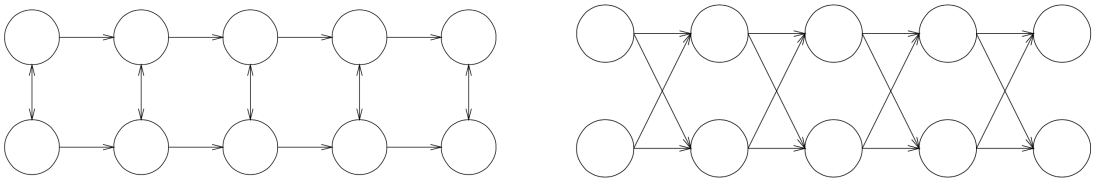


Figure 5.6: ***Left****: the transition probabilities between two HMMs;* ***Right****: the cross probabilities between two HMMs. Source: Brand [1997]*

5.2.2 Coupled Hidden Markov Models

The calendar of a homogeneous zone is represented by the hidden Markov model which contains a transition matrix and a emission matrix. This hidden Markov model also provides an associated hidden state sequence (called Viterbi sequence) with much less classes than the rainfall types in the calendar. Appendix A describes how to create a HMM corresponding to the calendar with an optimized hidden state sequence (Viterbi sequence)

which has less number of choices than rainfall-type calendar, which is a real advantage. That's the main idea to reduce the dimension problem.

In our case, for the 17 rainfall types, the number of hidden states is reduced to 4 (Table 5.3).

Table 5.3: Reducing the sequences of the 17 observed rainfall types into the optimized sequences of the 4 hidden types. **Left**: the 4 calendars in each zone (17 rainfall types in each calendar). **Right**: the 4 optimized hidden type sequences in each zone (4 hidden types in each hidden type sequence).

	Historical calendars from SOM			
Time	S_A	S_B	S_C	S_D
2005/01/01 00:00	A17	B11	C02	D16
2005/01/01 01:00	A10	B13	C02	D16
2005/01/01 02:00	A17	B12	C17	D17
2005/01/01 03:00	A17	B04	C02	D03
2005/01/01 04:00	A17	B04	C17	D17
2005/01/01 05:00	A17	B04	C17	D07
2005/01/01 06:00	A11	B13	C16	D07
2005/01/01 07:00	A06	B16	C17	D17
...	...	...	...	...
2014/12/31 23:00	A17	B08	C17	D17
	17 types	17 types	17 types	17 types

Historical hidden types sequences (Viterbi optimal estimation)			
HS_A	HS_B	HS_C	HS_D
HA 4	HB 2	HC 1	HD 1
HA 2	HB 2	HC 1	HD 1
HA 4	HB 3	HC 4	HD 2
HA 4	HB 3	HC 1	HD 2
HA 4	HB 3	HC 4	HD 2
HA 4	HB 3	HC 4	HD 3
HA 2	HB 2	HC 4	HD 3
HA 2	HB 1	HC 4	HD 3
...	...	...	...
HA 4	HB 3	HC 4	HD 4
4 hidden types	4 hidden types	4 hidden types	4 hidden types

The idea of creating a model to generate several simulated calendars simultaneously is to apply hierarchically the Markov reduction to joint states until all zones will be connected. Figure 5.7 presents how the hierarchical process works in a typical case. The construction of the model is representing from bottom (local ground evidence) to top (large scale rainfall type pattern), whereas the simulation generated afterward by the model is building from top to bottom. In each step of aggregation, two sequences (e.g., calendars or Viterbi sequences) which have the most similarity are chosen to be combined. To assess the similarity, the **Mutual Information** (MI) is used as a natural criterion and is based on the co-occurrence matrix (see Equation 4.1). As it is shown in Fig. 5.7, the sequence B and the sequence C have the maximum mutual information among all the pairs of sequences {A, B, C, D}, so Aggregation 1 combines the sequence B and the sequence C to obtain a sequence (B+C) for the further modeling. With the same process in each aggregation step, the final sequence (B+C+D+A) can be modeled to generate the simulations. This is the main approach called Coupled Hidden Markov Model (CHMM) which applied hierarchically HMMs in each aggregation to construct the coordination model.

The advantage of hierarchical process is not just to make certain order for combining, but also to reduce the dimension of the joint situations of combined sequence in each ag-

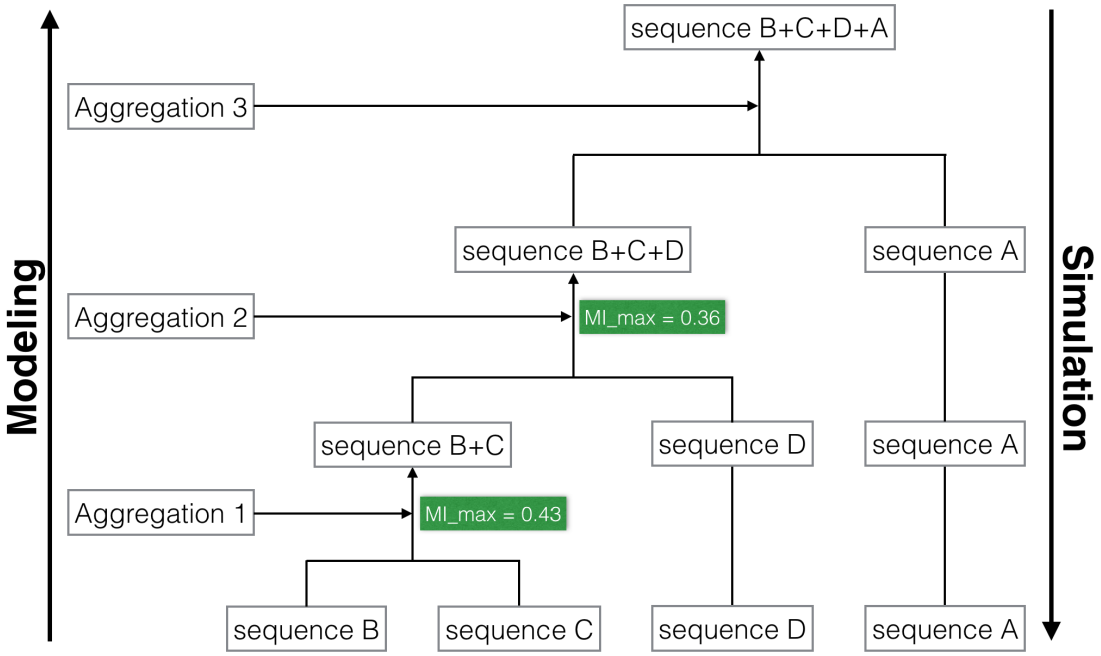


Figure 5.7: Diagram of hierarchy by using CHMM.

gregation. The hierarchical process will begin with the optimized Viterbi sequences with the 4 hidden types instead of the 17 rainfall types (Table 5.3). This choice is also made by reducing the number of classes at the beginning. For Aggregation 1, zone B and zone C have been combined due to the maximum mutual information (MI_max= 0.43) among all pairs. So the hidden type sequence HS_B and HS_C combine as the joint situations sequence for zone (B+C). Then, by applying HMM to the joint situation sequence (with 16 possible joint situations), an optimized hidden type sequence HS_{BC} (with 4 possible hidden types corresponding to 16 joint situations) presents zone (B+C) for the next aggregation. Table 5.4 presents how Aggregation 1 works. Table 5.5 presents the entire hierarchical process (Fig. 5.7) of merged zones (B+C), (B+C+D) and (B+C+D+A).

The CHMM contains all the transition and emission matrix when HMM is applied in each aggregation during the modeling. In our case, the final sequence HS_{BCDA} of (B+C+D+A) is also modeled by HMM, then this HMM associated (B+C+D+A) generates the initial simulation. With the emission matrix, two hidden sequences for both merged (B+C+D) and C zones are generated after the simulations of merged zone (B+C+D+A). The simulated sequences for individual zone A, B, C and D are generated after several times of the same procedure.

Within hierarchical simulation, the spatial co-occurrence is preserved, but temporal continuity is broken by successive draws in emission matrices. So we need a final step to enforce correct temporal distribution of each type within each homogeneous zone. In our case, we propose to guide the simulation by the use of the length of stay within types. The length of stay distributions for wet spell and dry spell have important impact in hydrological application (e.g., heavy rain accumulation, drought, etc.), this is the reason why we

choose this criterion: combined with rainfall types content, it represents the distribution of accumulated rainfall.

Table 5.4: Using HMM to reduce the number of joint situations of a combined hidden type sequence. Here is the example of combining zone B and zone C to obtain a hidden type sequence for zone (B+C).

Time	HS_A	HS_B	HS_C	HS_D	joint situations for zone B+C			HS_{BC}
2005/01/01 00:00	HA 4	HB 2	HC 1	HD 1	21	S1		HBC 2
2005/01/01 01:00	HA 2	HB 2	HC 1	HD 1	21	S1		HBC 2
2005/01/01 02:00	HA 4	HB 3	HC 4	HD 2	34	S2		HBC 3
2005/01/01 03:00	HA 4	HB 3	HC 1	HD 2	31	S3	HMM	HBC 3
2005/01/01 04:00	HA 4	HB 3	HC 4	HD 2	34	S2	⟶	HBC 3
2005/01/01 05:00	HA 4	HB 3	HC 4	HD 3	34	S2		HBC 3
2005/01/01 06:00	HA 2	HB 2	HC 4	HD 3	24	S4		HBC 3
2005/01/01 07:00	HA 2	HB 1	HC 4	HD 3	14	S5		HBC 3
...	...	...	...	...	...	...		...
2014/12/31 23:00	HA 4	HB 3	HC 4	HD 4	34	S2		HBC 3
	4 hidden types	4 hidden types	4 hidden types	4 hidden types	4x4=16			4 hidden joint types

Table 5.5: Aggregations of 4 zones by using CHMM corresponding to the diagram of hierarchy Fig. 5.7.

Time	HS_A	HS_B	HS_C	HS_D	HS_{BC}	HS_{BCD}	HS_{BCDA}
2005/01/01 00:00	HA 4	HB 2	HC 1	HD 1	HBC 2	HBCD 2	HBCDA 2
2005/01/01 01:00	HA 2	HB 2	HC 1	HD 1	HBC 2	HBCD 2	HBCDA 2
2005/01/01 02:00	HA 4	HB 3	HC 4	HD 2	HBC 3	HBCD 3	HBCDA 3
2005/01/01 03:00	HA 4	HB 3	HC 1	HD 2	HBC 3	HBCD 3	HBCDA 3
2005/01/01 04:00	HA 4	HB 3	HC 4	HD 2	HBC 3	HBCD 3	HBCDA 3
2005/01/01 05:00	HA 4	HB 3	HC 4	HD 3	HBC 3	HBCD 3	HBCDA 3
2005/01/01 06:00	HA 2	HB 2	HC 4	HD 3	HBC 3	HBCD 3	HBCDA 3
2005/01/01 07:00	HA 2	HB 1	HC 4	HD 3	HBC 3	HBCD 3	HBCDA 3
...	...	...	...	...	...	...	...
2014/12/31 23:00	HA 4	HB 3	HC 4	HD 4	HBC 3	HBCD 3	HBCDA 3
	4 hidden types	4 hidden types	4 hidden types	4 hidden types	4 hidden joint types	4 hidden joint types	4 hidden joint types

5.2.3 Re-organizing method

The objective of reorganization is to optimize the time-series of simulated sequence in each zone with local observed length of stay distributions. The main idea is to reorganize a simulated sequence to a new (reorganized) sequence which preserves the synchronicity as much as possible with the original simulated sequence, and without changing the number of the types of the simulated sequence. The reorganization aims at following the local observed length of stay distributions, as much as possible.

Given S, a time-series of a simulated sequence given one rainfall type at each time step.

- n is the length of the simulated sequence S;
- m is the number of rainfall types;
- $T = \{T_1, T_2, \ldots, T_m\}$ is the set of rainfall types;
- $S = \{S_1, S_2, \ldots, S_n\}$ (where $S_i \in T$, for $i \in \{1, \ldots, n\}$) is a simulated sequence generated by CHMM.
- $O = \{O_1, O_2, \ldots, O_N\}$ (where $O_t \in T$, for $t \in \{1, \ldots, N\}$) is the observed calendar, O and S therefore consist of the same set of rainfall type T.

In general, the length of simulated sequence S is much longer than the length of observed calendar O, that means $n \gg N$.

The objective is to reorganize S to obtain a new sequence named $C = \{C_1, C_2, \ldots, C_n\}$ (where $C_i \in T$, for $i \in \{1, \ldots, n\}$). The three criteria of the reorganization are:

Criterion 1: $\text{Card}\{T_k \in C\} = \text{Card}\{T_k \in S\}$, for all $k \in \{1, \ldots, m\}$;

Criterion 2: {the length of stay distribution for T_k in C} $\cong$ {the length of stay distribution for T_k in O}, for all $k \in \{1, \ldots, m\}$;

Criterion 3: to maximize the synchronicity between C and S,

$$\max \left(\sum_{i=1}^{n} \mathbb{1}_{C_i = S_i} \right). \tag{5.1}$$

Criterion 1 means that the number of each rainfall type in C is exactly the same as in S. Criterion 2 is the main feature that we want to make sure that the length of stay distributions of the simulated sequences will respect the observed calendars. Criterion 3 is also important because the synchronicity between S and C was granted by CHMM approach. However, while Criterion 1 is maintained by design, some loss must be accepted on Criterion 3 to reach Criterion 2.

The re-organizing method is presented as follows. We define $(S_p, \ldots, S_q)$, for $1 \leq p < q \leq n$, a sub-sequence of S of length $q - p + 1$ as a **fragment** of the rainfall type T_k ($k \in \{1, \ldots, m\}$) when $(S_p, \ldots, S_q)$ consists of $q - p + 1$ consecutive rainfall type T_k, and also

both the time step just before S_p and the time step just after S_q are not the rainfall type T_k. That means

$$S_{p-1} \neq T_k, S_h = T_k, S_{q+1} \neq T_k, \text{for } h \in \{p, \ldots, q\}. \tag{5.2}$$

If $(S_p, \ldots, S_q)$ is a fragment of the rainfall type T_k, then we call the **length of stay** (or **dwell time**) for rainfall type T_k of $(S_p, \ldots, S_q)$ is $(q - p + 1)$.

M_S, the table of lengths of stay for each rainfall type for sequence S is presented by Table 5.6. d_S is the maximum length of stay among all types and $S_{k,l}$ (for $k \in \{1, \ldots, m\}$, $l \in \{1, \ldots, d_S\}$) is the frequency of length of stay l for type k.

Table 5.6: Table of lengths of stay for each rainfall type for sequence S: M_S

	1	2	...	d_S
type 1	$S_{1,1}$	$S_{1,2}$	...	S_{1,d_S}
type 2	$S_{2,1}$	$S_{2,2}$	...	S_{2,d_S}
⋮	⋮	⋮	...	⋮
type m	$S_{m,1}$	$S_{m,2}$	...	S_{m,d_S}

Correspondingly, M_O, the table of lengths of stay for each rainfall type for observed calendar O is presented by Table 5.7.

Table 5.7: Table of lengths of stay for each rainfall type for sequence O: M_O

	1	2	...	d_O
type 1	$O_{1,1}$	$O_{1,2}$	...	O_{1,d_O}
type 2	$O_{2,1}$	$S_{2,2}$	...	O_{2,d_O}
⋮	⋮	⋮	...	⋮
type m	$O_{m,1}$	$O_{m,2}$	...	O_{m,d_O}

In probability theory, Kullback-Leibler divergence [*Kullback and Leibler*, 1951] is a measurement of the differences between two probability distributions. For discrete probability distributions P and Q, the Kullback-Leibler divergence from Q to P is defined in *MacKay* [2003] to be

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{5.3}$$

In our case, for a rainfall type T_k (for $k \in \{1, \ldots, m\}$), the Kullback-Leibler divergence from

$(S_{k,1}, S_{k,2}, \ldots, S_{k,d_S})$ to $(O_{k,1}, O_{k,2}, \ldots, O_{k,d_O})$ is

$$D_{KL} = \sum_{l=1}^{\max(d_S,d_O)} \left(\frac{O_{k,l}}{\sum_{i=1}^{N} \mathbb{1}_{O_i=k}} \right) \log \left(\frac{\frac{O_{k,l}}{\sum_{i=1}^{N} \mathbb{1}_{O_i=k}}}{\frac{S_{k,l}}{\sum_{i=1}^{n} \mathbb{1}_{S_i=k}}} \right) = \sum_{l=1}^{\max(d_S,d_O)} O(k,l) \log \frac{O(k,l)}{S(k,l)} \tag{5.4}$$

which $\frac{O_{k,l}}{\sum_{i=1}^{N} \mathbb{1}_{O_i=k}}$ (noted $O(k,l)$) is the probability of length of stay of l for the rainfall type T_k in observed calendar O and $\frac{S_{k,l}}{\sum_{i=1}^{n} \mathbb{1}_{S_i=k}}$ (noted $S(k,l)$) is the probability of length of stay of l for the rainfall type T_k in the observed calendar S. We can use this measurement to examine whether the simulated length of stay distribution for each type is similar to the observed one. Equation 5.4 shows if $S(k,l)$ equals to $O(k,l)$ (for all $l \in \{1, \ldots, \max(d_S, d_O)\}$), then D_{KL} will be equal to 0, meaning that the simulated sequence S has exactly the same distribution of length of stay for the rainfall type T_k.

So now, the idea is to create a table of lengths of stay M_W (Table 5.8) to follow the next three rules.

Table 5.8: Table of lengths of stay for each rainfall type for sequence W: M_W

	1	2	...	d_W
type 1	$W_{1,1}$	$W_{1,2}$	...	W_{1,d_W}
type 2	$W_{2,1}$	$W_{2,2}$	...	W_{2,d_W}
⋮	⋮	⋮	...	⋮
type m	$W_{m,1}$	$W_{m,2}$	...	W_{m,d_W}

First, the maximum length of stay among all types of M_W is the same as M_O.

$$d_W = d_O. \tag{5.5}$$

Second, M_W contains the same quantity for each type.

$$\sum_{l=1}^{d_W} W_{k,l} = \sum_{l=1}^{d_S} S_{k,l} \qquad \text{for all } k \in \{1, \ldots, m\}. \tag{5.6}$$

Third, M_W has the same length of stay distribution for each type as M_O. But on account of previous rule (Equation 5.6), the consistency of the two length of stay distributions will be approximated.

$$\frac{W(k,1)}{O(k,1)} \cong \frac{W(k,2)}{O(k,2)} \cong \ldots \cong \frac{W(k,d_W)}{O(k,d_O)} \Leftrightarrow \frac{W_{k,1}}{O_{k,1}} \cong \frac{W_{k,2}}{O_{k,2}} \cong \ldots \cong \frac{W_{k,d_W}}{O_{k,d_O}} \qquad \text{for all } k \in \{1, \ldots, m\}. \tag{5.7}$$

With Equation 5.5, 5.6 and 5.7, we obtain a new table of lengths of stay M_W which represents the very similar length of stay distribution as M_O for each rainfall step. Basically, we shift a minimum number of types in the sequence S to make a corrected sequence C (from S) that has the exactly the same table of lengths of stay M_W. Once M_W is fixed, we reorganize S to obtain a corrected sequence C by following the previous three criteria.

Criterion 1 means that the total quantity of each type in C is the same as S;

Criterion 2 means that the table of lengths of stay of C is M_W;

Criterion 3 means to maximize the synchronicity between C and S [XXX-NO rephrase].

Criterion 1 is preserved by design. Criterion 2 is easy to check, so Criterion 3 is our priority. Criterion 3 means to make the minimize changes comparing to S, the question is how to shift a minimum number of types in S to reach the Criterion 2.

The principle of the reorganization approach is similar to the so-called greedy algorithm, that is an algorithm which always makes the choice that seems to be the best at that moment. This means that it makes a locally-optimal choice in the hope that this choice will lead to a globally-optimal or at least usable solution. We thus follow the table M_W and begin with the maximum (most "difficult block") length of stay d_W ($= d_O$). For each length of stay l ($l \in \{1, \ldots, d_W\}$) and for each type T_k ($k \in \{1, \ldots, m\}$), we need to find the $W_{k,l}$ positions in S where we can place the $W_{k,l}$ fragments of l-length for type T_k in C. These positions will be chosen by shifting the minimum number of types in S.

Here is a simple example.

We have a sequence $X = (T_1, T_3, T_1, T_2, T_2, T_1, T_1, T_3, T_1, T_1, T_2, T_1, T_3, T_2, T_3, T_1)$.

time step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
type	T_1	T_3	T_1	T_2	T_2	T_1	T_1	T_3	T_1	T_1	T_2	T_1	T_3	T_2	T_3	T_1

If a fragment (T_1, T_1, T_1, T_1) needs to be placed in X, we shall examine all segments (subsequence) of length 4 in X. In this example, it does not exist a fragment (T_1, T_1, T_1, T_1), so we begin with all segments of length 4 which contain 3 types T_1. The possible candidates are

time step	6	7	8	9
type	T_1	T_1	T_3	T_1

7	8	9	10
T_1	T_3	T_1	T_1

9	10	11	12
T_1	T_1	T_2	T_1

But if first two candidates are chosen, then X will have a fragment $(T_1, T_1, T_1, T_1, T_1)$ of length of 5 instead of length of 4 because of the time steps 6 and 10 are also type T_1. So we have to avoid such mistake to find the right place for the expected fragment in the re-organizing method.

The algorithm of the re-organizing method for obtaining a corrected sequence C is presented as follows:

1. set $C = S$;
2. the initial length of stay $l := d_W (= d_O)$;
3. for k from 1 to m, $W_{k,l}$ fragments of type T_k of length l need to be placed in C by making the minimum changes comparing to S;
4. fix the time steps of placed $W_{k,l}$ fragments in the sequence C (they will not be replaced anymore in further algorithm);
5. then $l := l - 1$ and repeat the step 2 and 3.

In the end of the re-organizing algorithm, the sequence C has

- the same quantity of each rainfall type as the sequence S, that means Criterion 5.2.3 is respected.
- the same table of lengths of stay as M_W which is very similar to the table of lengths of stay M_O, that means Criterion 5.2.3 is respected.

Table 5.9 shows the percentage of preserved time steps between the simulated sequence S and the corrected sequence C in each zone. The percentages reach more than 90% in our case.

Table 5.10 shows the percentage of spatial synchronicity of the 4 zones between the simulated sequences and the corrected sequences. We can see that there are more than 96 percent of time steps when at least 3 zones have the same synchronicity as the sequences before the reorganization. This is important because the spatial rainfall type prescribed by the hierarchical simulation is respected at all these time-steps.

Table 5.9: For each zone, percentage of the total number of time steps when the corrected sequence C and simulated sequence S have the same rainfall types after the reorganization algorithm.

zone 1	zone 2	zone 3	zone 4
92.7%	91.5%	88.1%	90.5%

Figure 5.8 presents the methodology schema of CHMM with re-organizing method from the beginning to the end.

Table 5.10: Percentages of spatial synchronicity between the simulated sequences and the corrected sequences. The first column corresponds the percentage of time steps when the simulated sequences and the corrected sequences have the same rainfall types in all 4 zones. The last column corresponds to the percentage of time steps when the simulated sequences and the corrected sequences do not have any similar rainfall type in each zone.

4	3	2	1	none
81.59%	15.70%	2.48%	0.22%	0.01%

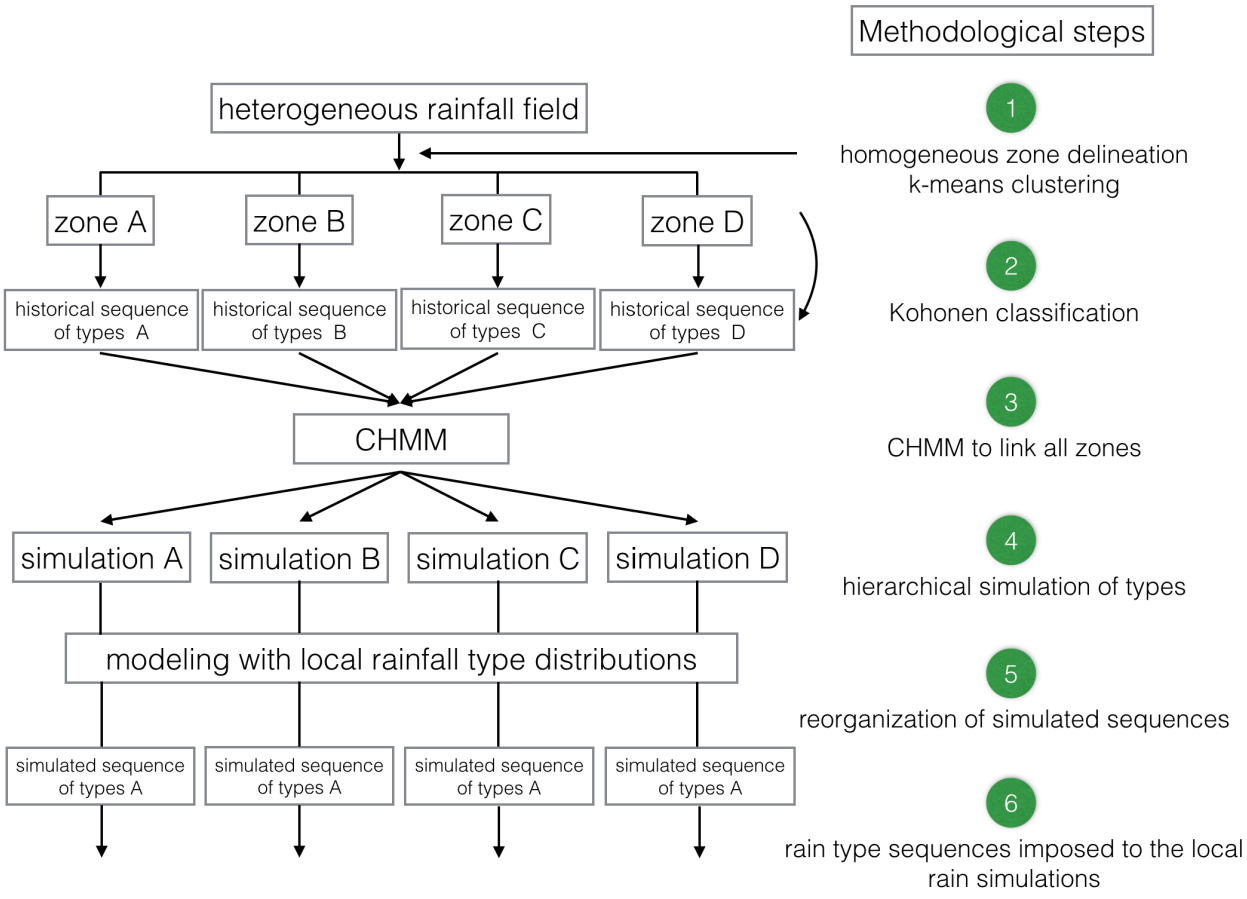


Figure 5.8: Overview of rainfall simulation by coordination of homogeneous zones.

Step by step

To summarize, the following steps (Fig. 5.8) are followed to get the simulation of an heterogeneous rainfall:

1. the partition a heterogeneous rainfall field to several homogeneous zones;
2. the creation of an observed calendar in each time step for each zone by using self-organizing maps (Kohonen classification);
3. the generation of the simulated calendar of each zone by using CHMM;
4. the reorganization of the individual simulated sequence generated previously in each zone with following objectives:
   - the reorganized sequences will have the same the length of stay distributions in each type as the observed calendars;
   - the reorganized sequences will have the same number of each rainfall type as the simulated sequence;
   - the reorganized sequences will keep the maximum number of rainfall types equal to the simulated sequence.

5.3 Non-parametric method: Resampling technique

The resampling model presented in this section is an alternative to the coupled hidden Markov model in Section 5.2. Common resampling techniques include bootstrapping, jackknifing and permutation tests [*Efron*, 1982]. In this PhD work, the resampling method used is bootstrapping, which draws randomly with replacement from available data.

The starting point is that for the objective of coordination of rainfall types calendars, both spatial and temporal aspects are important to be captured. More specifically, with the case of hourly calendars for the 4 zones (Table 5.2), the resampling technique will be applied to preserve the synchronicity of rainfall types of the 4 zones at each time step. To construct one year of rainfall-type simulation corresponding to the historical calendars, the procedure is the following (Fig. 5.9):

(1) defining t as a certain day in the year;

(2) choosing a random number of days n (e.g., between 10 days to 20 days);

(3) choosing a random year Y (e.g., between 2004 and 2015 in hourly case);

(4) targeting the same period of days in the chosen year as simulated year (e.g., the period from date t to next n days in the year Y of observed calendars);

(5) allowing a random shift of several days from the exact same period (e.g., between 0 and 15 days before or after the targeting period).

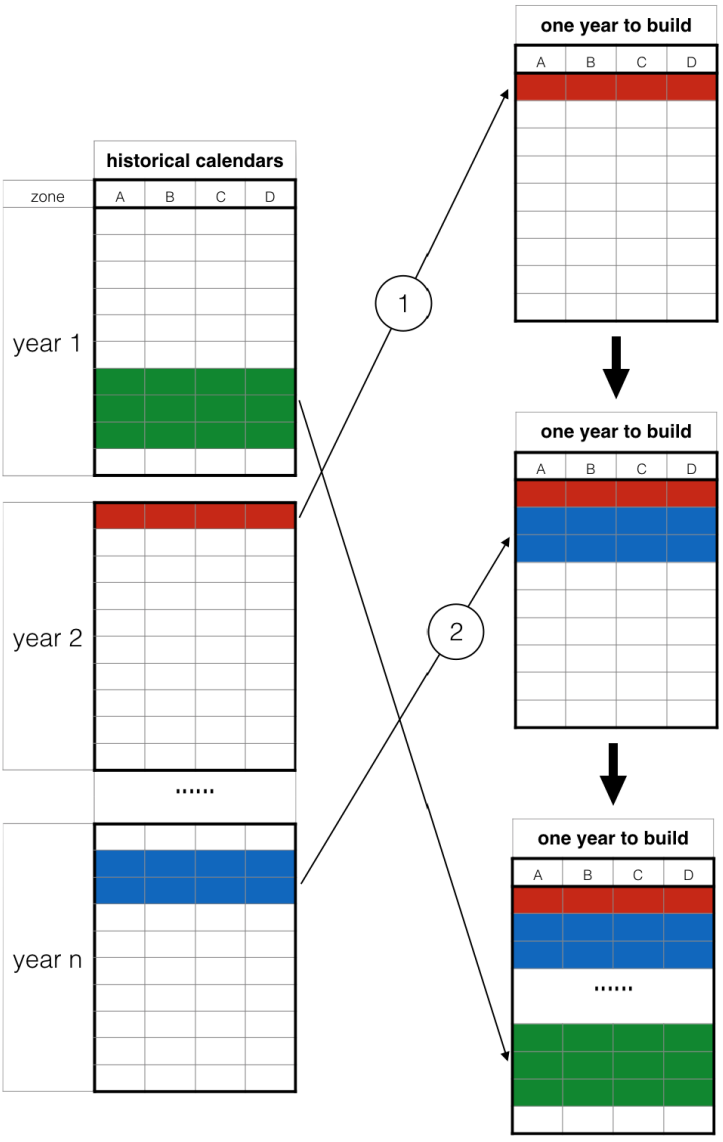


Figure 5.9: Diagram of resampling technique.

The algorithm is easy to implement and the long-term simulation is also simple to generate. A small number of numerical parameters of the method, however, must be chosen. The above-mentioned values want (2) to honor the duration of rainfall/hydrological events in the area and (5) to give some flexibility in the resampling, while keeping seasonality.

Next chapter, we will diagnose the simulations of the coupled hidden Markov model with the reorganization method and the resampling method.

CHAPTER 6

RESULTS AND CONCLUSIONS FOR METHODOLOGY I

In this chapter, the results of four modeling approaches are analyzed, as well as their capacity to reproduce the rainfall field at the Cévennes-Vivarais scale is examined. The first modeling approaches is SAMPO applied over the entire rainfall field (**monobloc**). The 3 other modeling approaches are applied to SAMPO applied over each of the 4 homogeneous zones, where SAMPO uses simulated calendars out of CHMM optimized by reorganization method (**reorg**), simulated calendars out of resampling technique (**resam**), or simulated calendars out of resampling technique optimized by reorganization method (**resamreorg**).

To get rid of possible interference with the peculiarities of the homogeneous simulation technique used, the reference will not be the observed rainfall, but the result of combining SAMPO simulated rainfall with observed/historical calendars as established from hourly observed precipitation data from 2005 to 2014. This reference is denoted **obs** (observation), but this (**obs**) properly only denotes observed calendars.

For all the models including the reference with observed calendars (obs), 50 replications of 10 years hourly simulations are generated and analyzed with box-plot illustrations.

Figure 6.1 presents the domains of simulation. In each homogeneous zone, 5 windows at different scales, $1km \times 1km$, $2km \times 2km$, $4km \times 4km$, $8km \times 8km$ and $16km \times 16km$ respectively, are chosen to analyze the statistical properties and how they evolve with spatial support. The larger $128km \times 128km$ domain intersect all zones, so refers to an actually heterogeneous region.

6.1 Statistical analysis

The simulation where conducted on hourly time step, but statistics are conducted on daily accumulation, stressing on the temporal aggregation of the simulated fields. Fig. 6.2 - 6.9 present, for each homogeneous zone and for the 5 domains, the comparison between the 5 models (referred with a different color) on features like the daily accumulation, its standard deviation, yearly maximum value, average indicator function or wetness, wet spell duration and its the standard deviation, dry spell duration and its standard deviation, respectively.

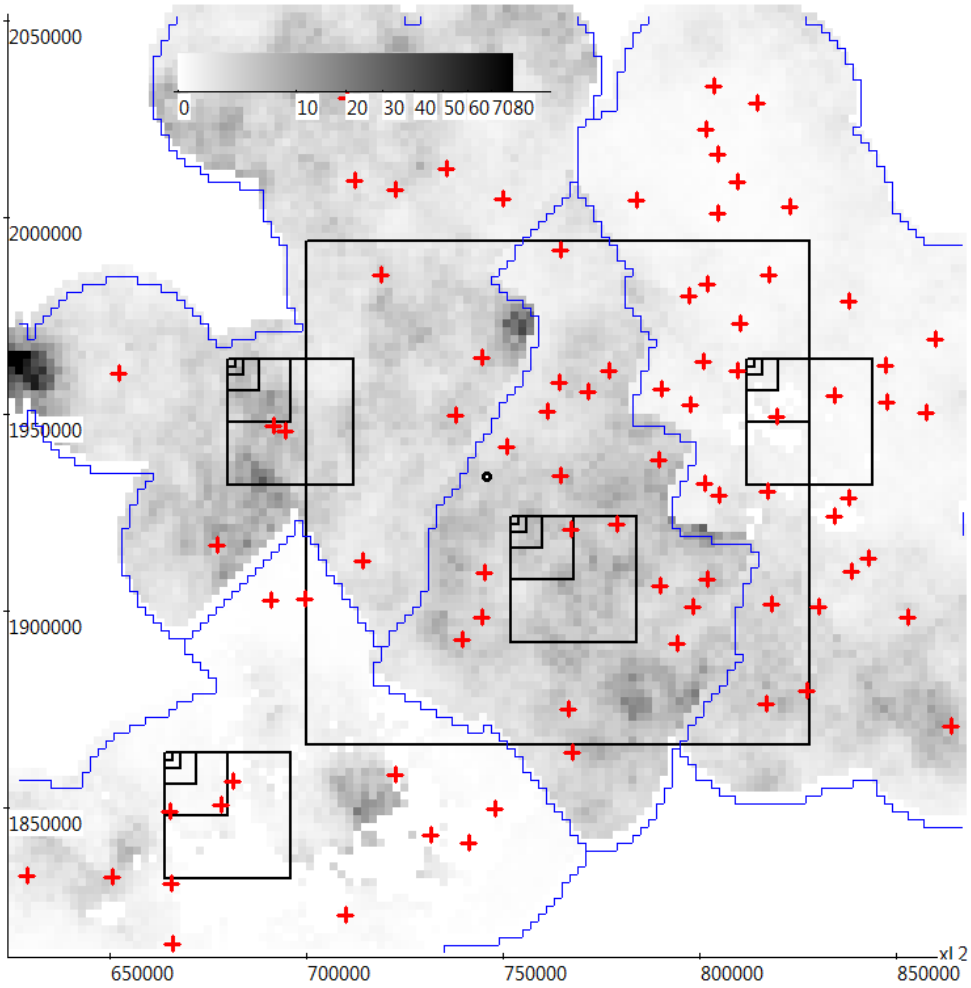


Figure 6.1: The Cévennes-Vivarais region with different domains of simulations. In each homogeneous zone, 5 domains are chosen at different scales, 2 km×2 km, 4 km×4 km, 8 km×8 km, 16 km×16 km and 32 km×32 km. The larger square of 128 km×128 km refers to the heterogeneous region since it overlaps the 4 homogeneous zones. In color, precipitation intensity in mm/h is shown as an example.

As illustrated in these Figures, the resampling technique gives a very good estimation of historical statistics. The reorganization technique does not improve the resampling simulation much better. The Markov parametric method CHMM with reorganization technique performs as not as good as the non-parametric resampling method, but the diagnostics for different statistics are not bad either when comparing to the reference simulations, obs, based on historical rainfall types sequences. As expected, the SAMPO simulations over the whole region give worst results in particular when compared to the "observations". This result clearly justifies the partition over the heterogeneous region into several homogeneous rainfall zones. Comparing the diagnostics of different zones, the CHMM with reorganization technique gives the best results for zone C where there are less stations with smaller surface.

Figure 6.2 shows the statistical fluctuations of the simulations for all models. The boxplots of the different models are close, but they should be in the same level if the proportion

of rainfall types in the simulations converges to the proportion of rainfall types in the observation. Figure 6.3 shows that the simulations of **reorg** lack variability when compared with resam and **resamreorg**. Figure 6.4 shows that the simulation of **reorg** under-estimate extreme values, that are not explicitly considered by the design of the hidden Markov models. The facts that the simulations of **reorg** cover too much rain in space (Fig. 6.5) and are inconsistent with the reference one at dry spell and wet spell (Fig. 6.6-6.9) also is a drawback of HMM. That is because the random draws of simulations by HMM break the transition probabilities of rainfall types in the observations. Non-parametric models like **resam** and **resamreorg** are more capable to reproduce the extreme events.

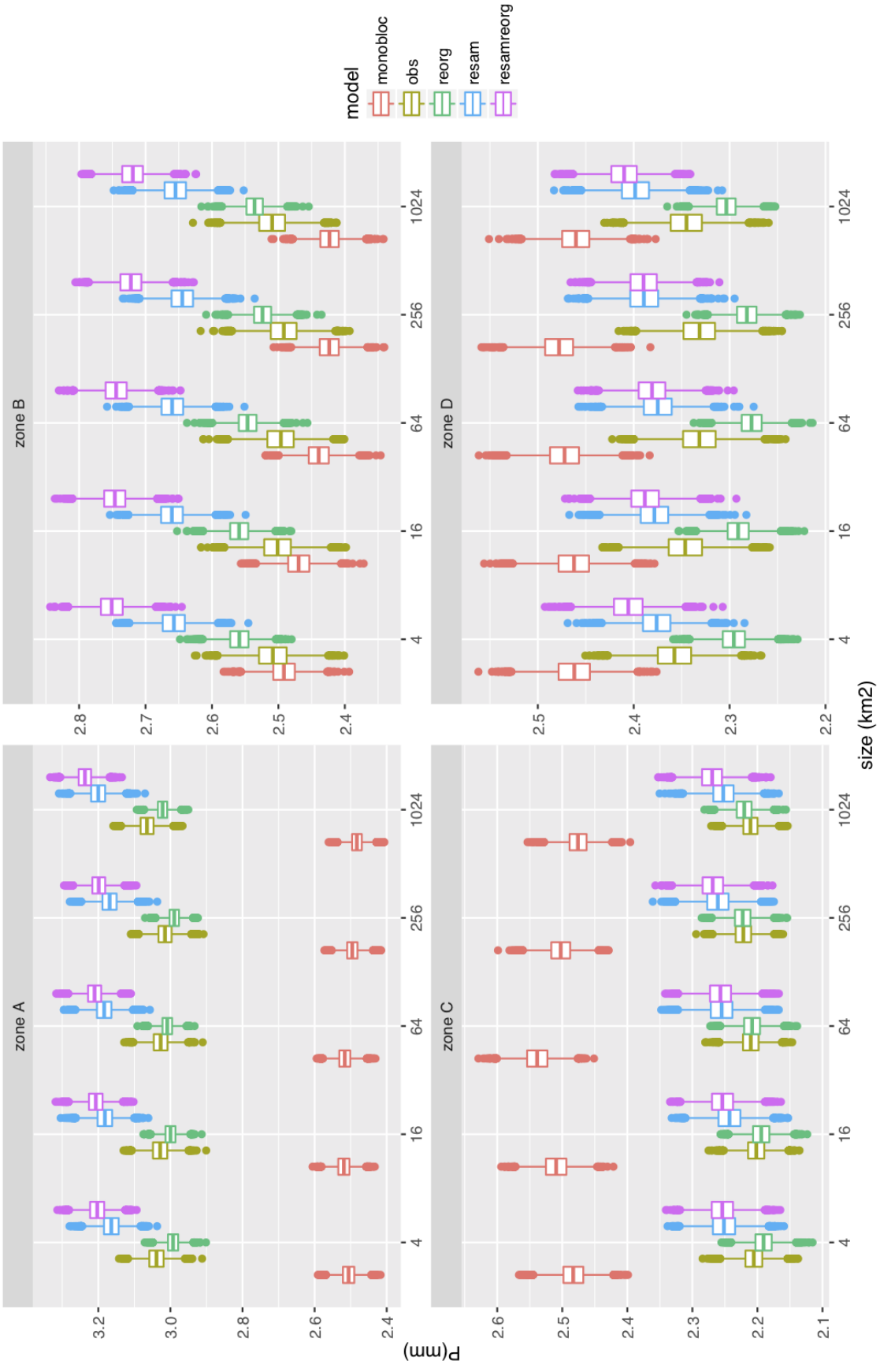


Figure 6.2: Average of daily accumulation of hourly precipitation for 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

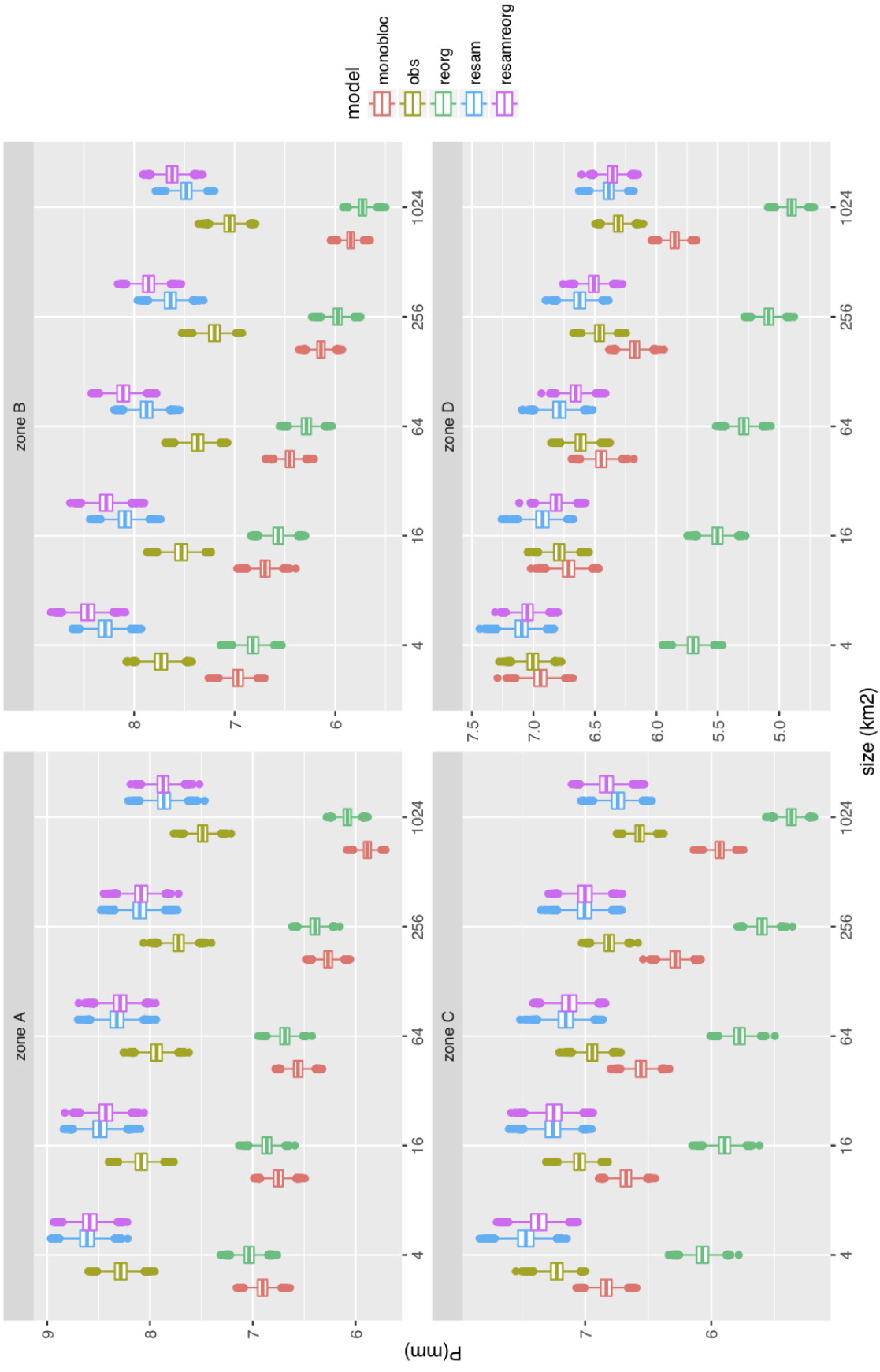


*Figure 6.3: Annual average of standard deviation of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: **monobloc** stands for SAMPO applied over the entire rainfall field, **reorg** for SAMPO with CHMM optimized by reorganization method, **resam** for SAMPO with resampling technique and, **resamreorg** for SAMPO with resampling technique and reorganization method.*

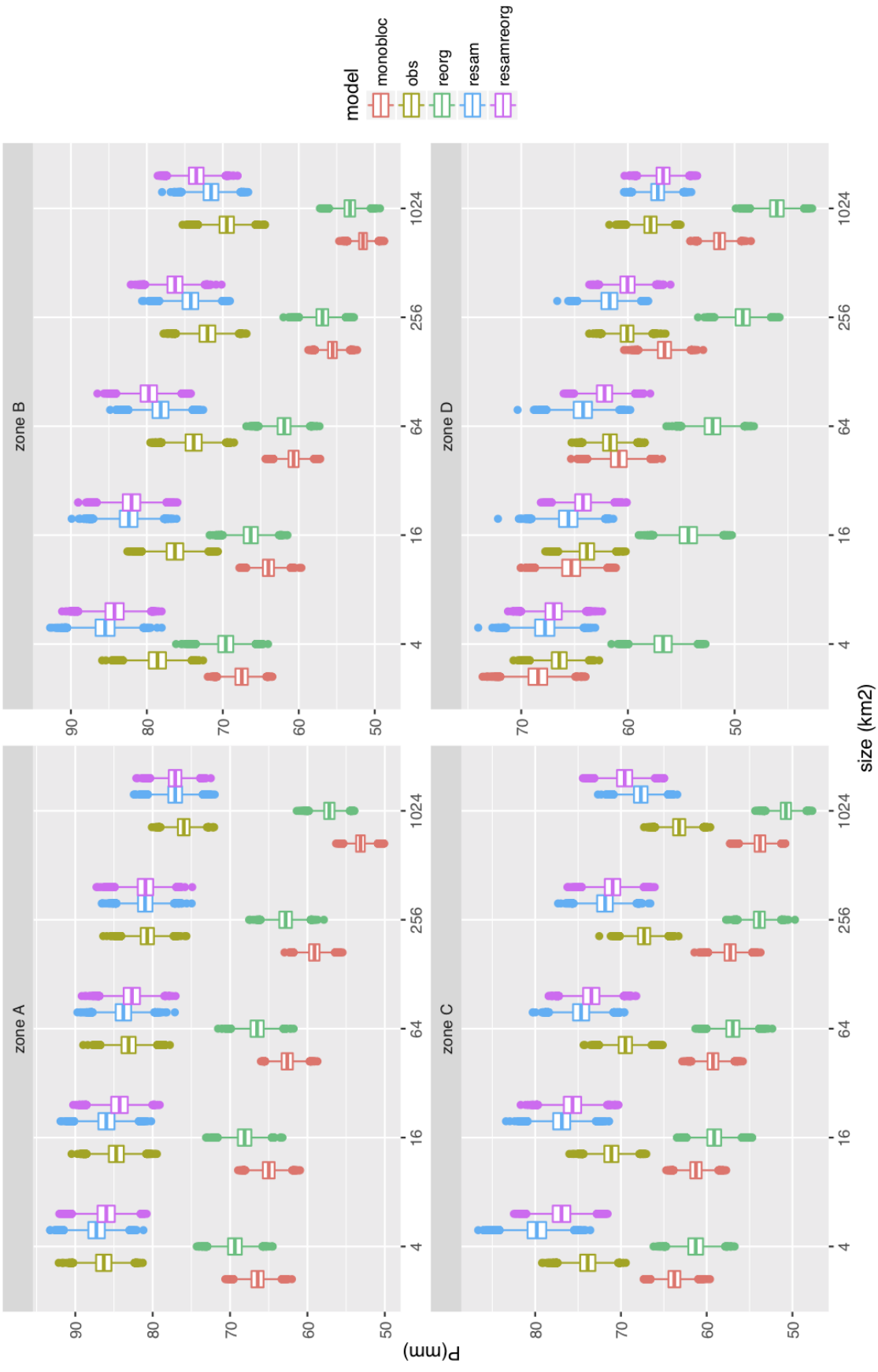


Figure 6.4: Annual average of maximum value of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

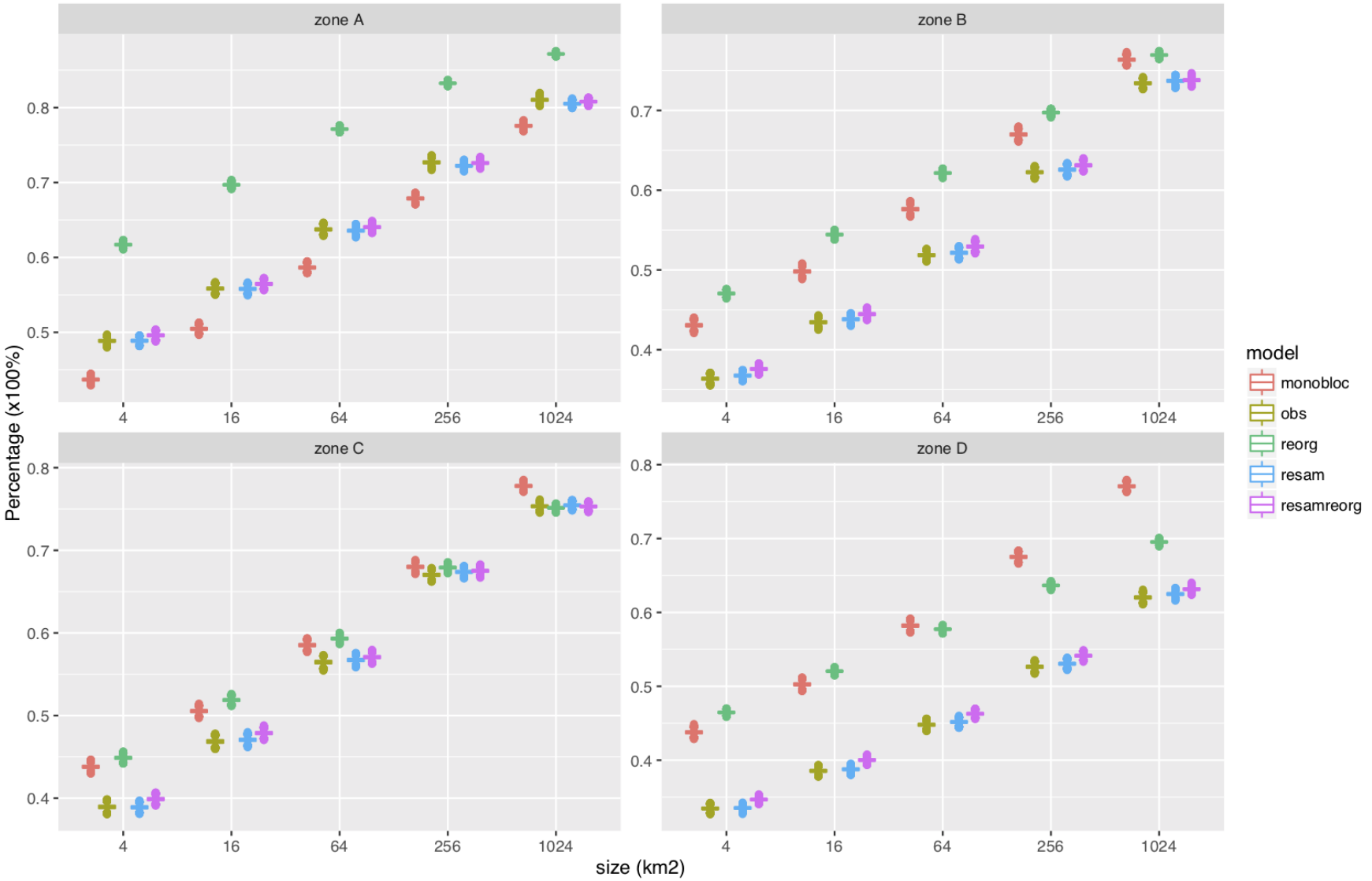


Figure 6.5: Average indicator function of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

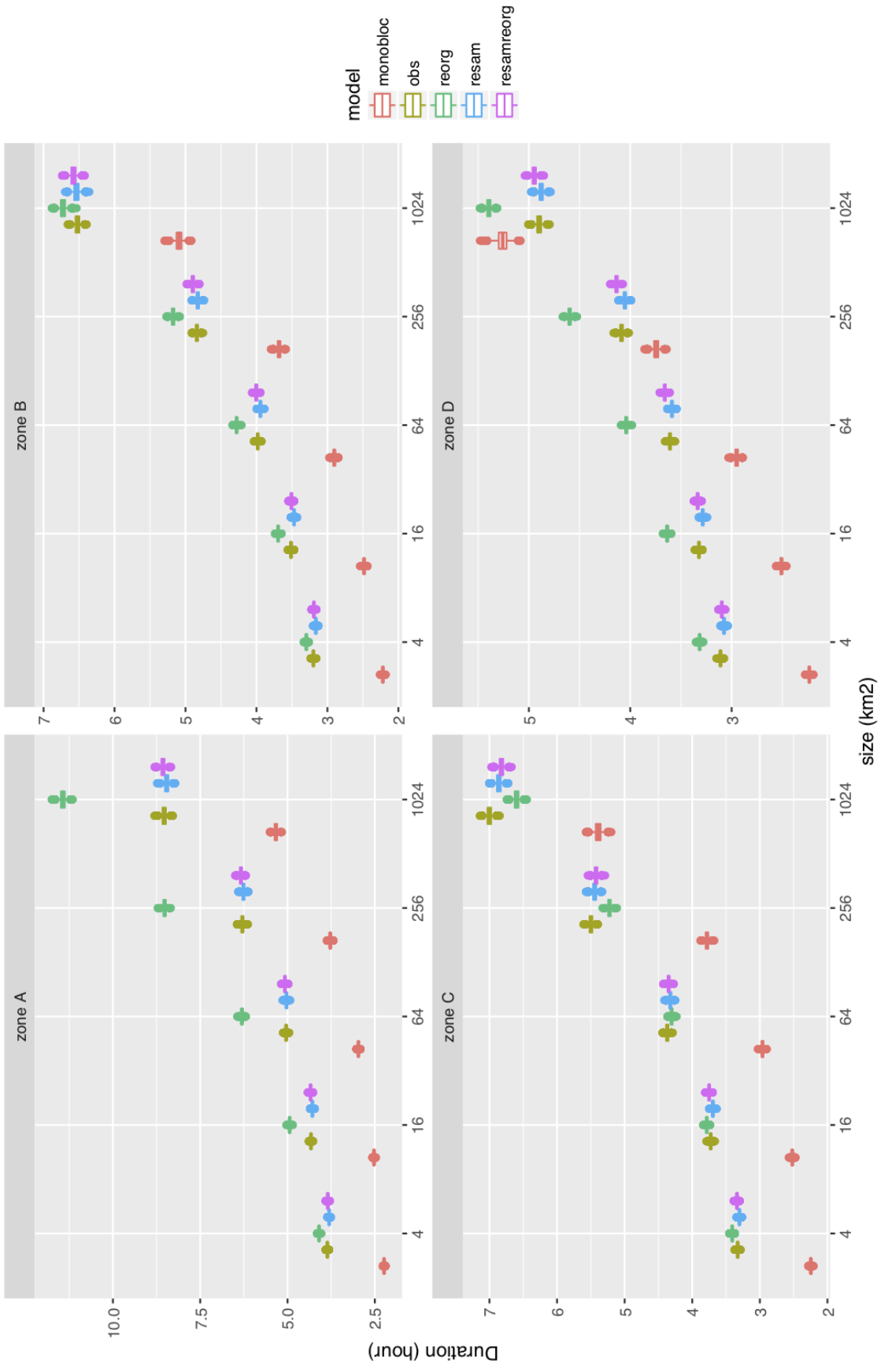


Figure 6.6: Average wet spell length of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

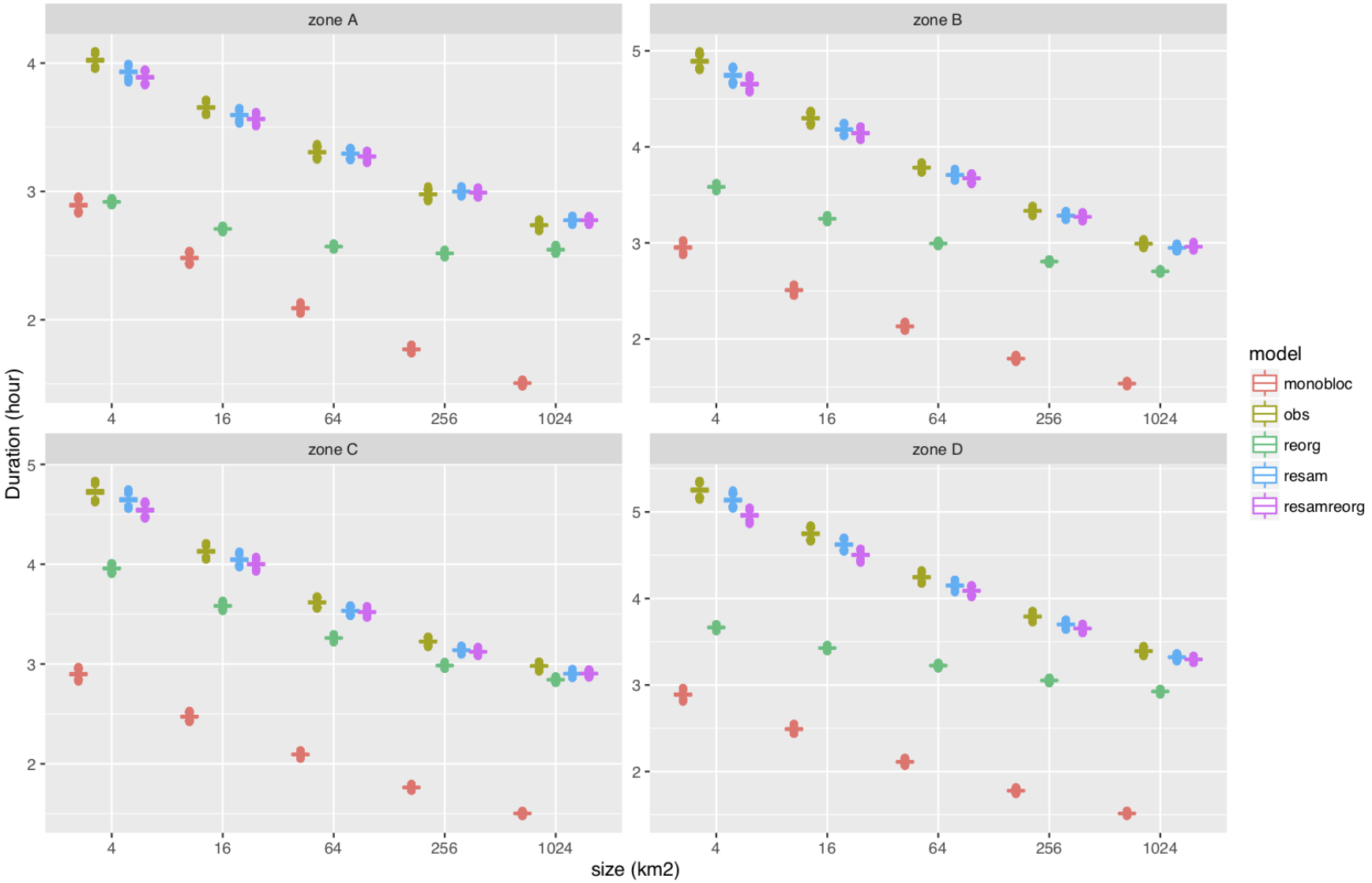


Figure 6.7: Standard deviation of wet spell length of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

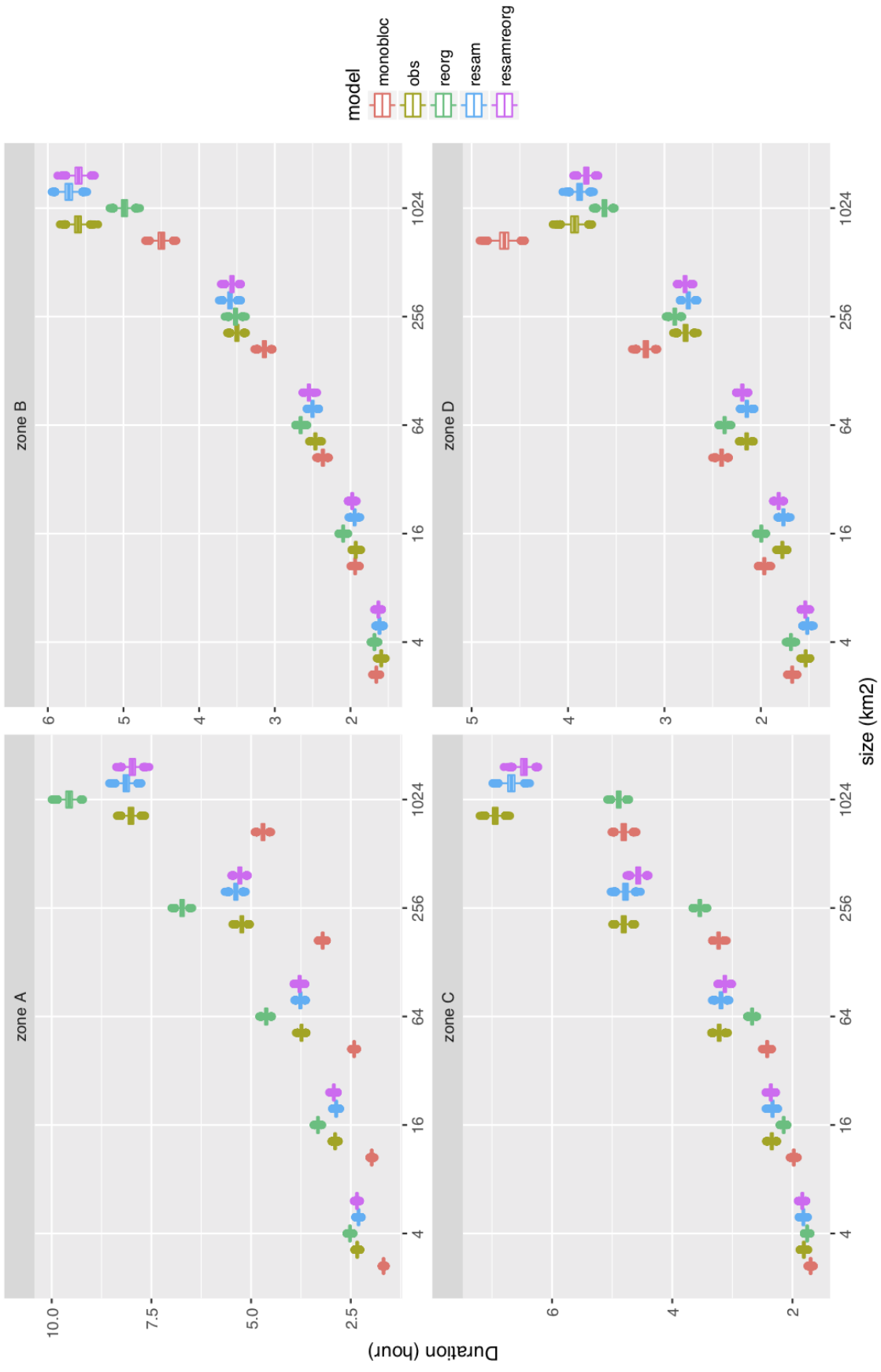


Figure 6.8: Average of dry spell length of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

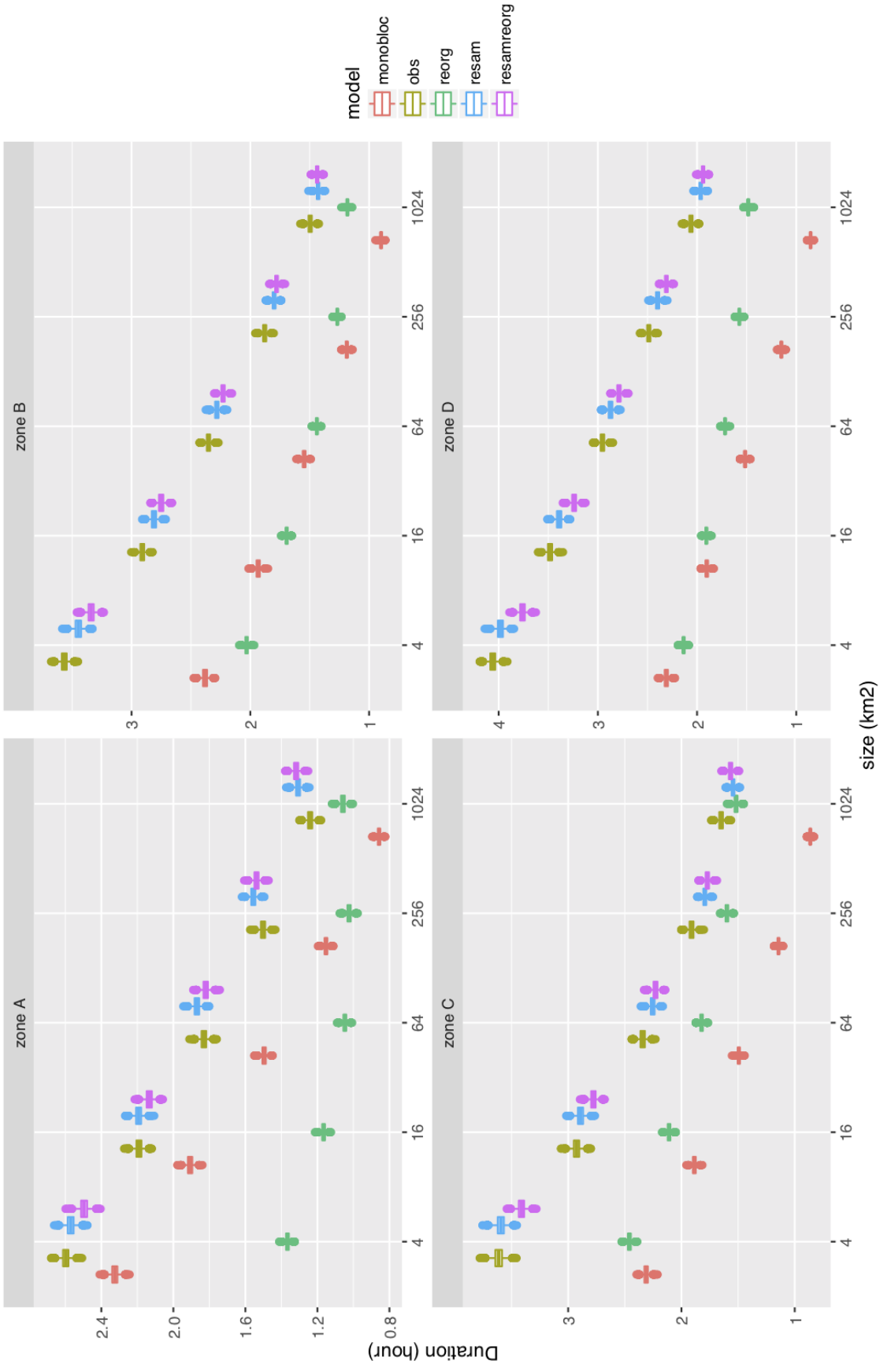


*Figure 6.9: Standard deviation of dry spell length of daily accumulation of hourly precipitation for the 4 zones. In each zone, the 4 different models are used and are referred as: **monobloc** stands for SAMPO applied over the entire rainfall field, **reorg** for SAMPO with CHMM optimized by reorganization method, **resam** for SAMPO with resampling technique and, **resamreorg** for SAMPO with resampling technique and reorganization method.*

6.2 Spatial correlation

In Fig. 6.10 - Fig. 6.12, the inter-stations correlations are presented within each homogeneous zone, between the reference, **obs**, and the computed values with the **monobloc**, **reorg** and **resam** models, respectively. As illustrated in Fig. 6.10, whatever the zones, the inter-station correlations are not too bad. For a given zone, the correlations computed with the simulated values are usually smaller than observed ones.

In Fig. 6.11 (CHMM with the reorganization method) and Fig. 6.12 (resampling model), for a given zone, the inter-station correlations computed with the simulated values, whatever the model, are more accurate than the ones obtained with the monobloc simulation, and this is good. But if we look at two stations located in two different zones, the inter-station correlations is completely lost. It appears that due to the hierarchical simulation containing a partition choice and independent rainfall simulation in each zone, simulated rainfall values pertaining to distinct zones are conditionally independent. The correlation reduces to the part conveyed / preserved by the rainfall typing and appears only zone dependent, not distance dependent unless the zone is the same and the correlation within SAMPO simulated fields appears. This is especially noticeable when we consider two stations in two different zones but located near the border of the zones, they should present a strong correlation, what is not the case in the simulations, as can be seen clearly in Fig. 6.1. This feature is not only unaesthetic, but correspond to a systematic bias.

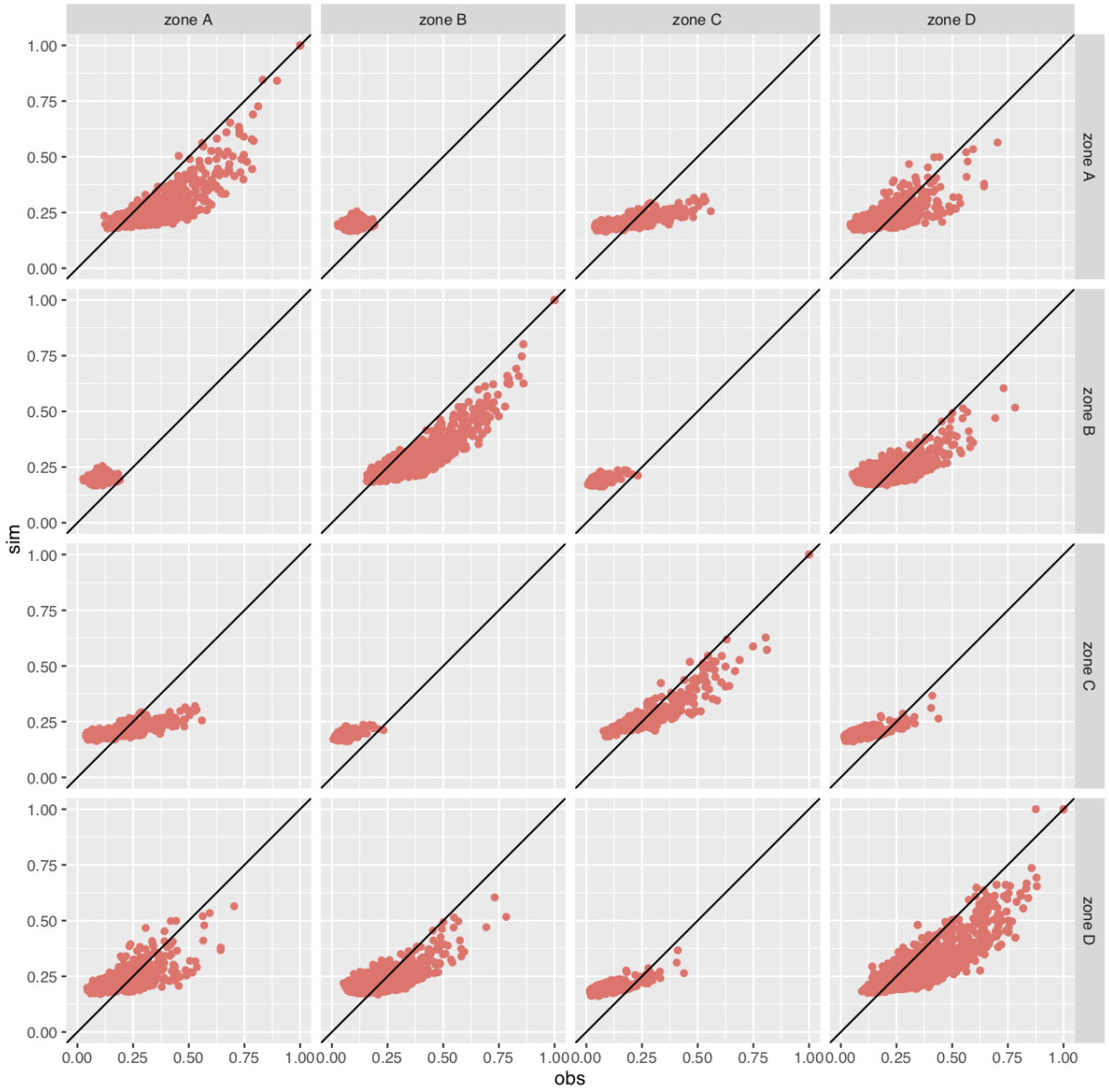


Figure 6.10: Inter-station correlations between hourly observation and monobloc simulation (SAMPO applied to whole surface). The length of hourly simulation is 10*x*10 *years.*

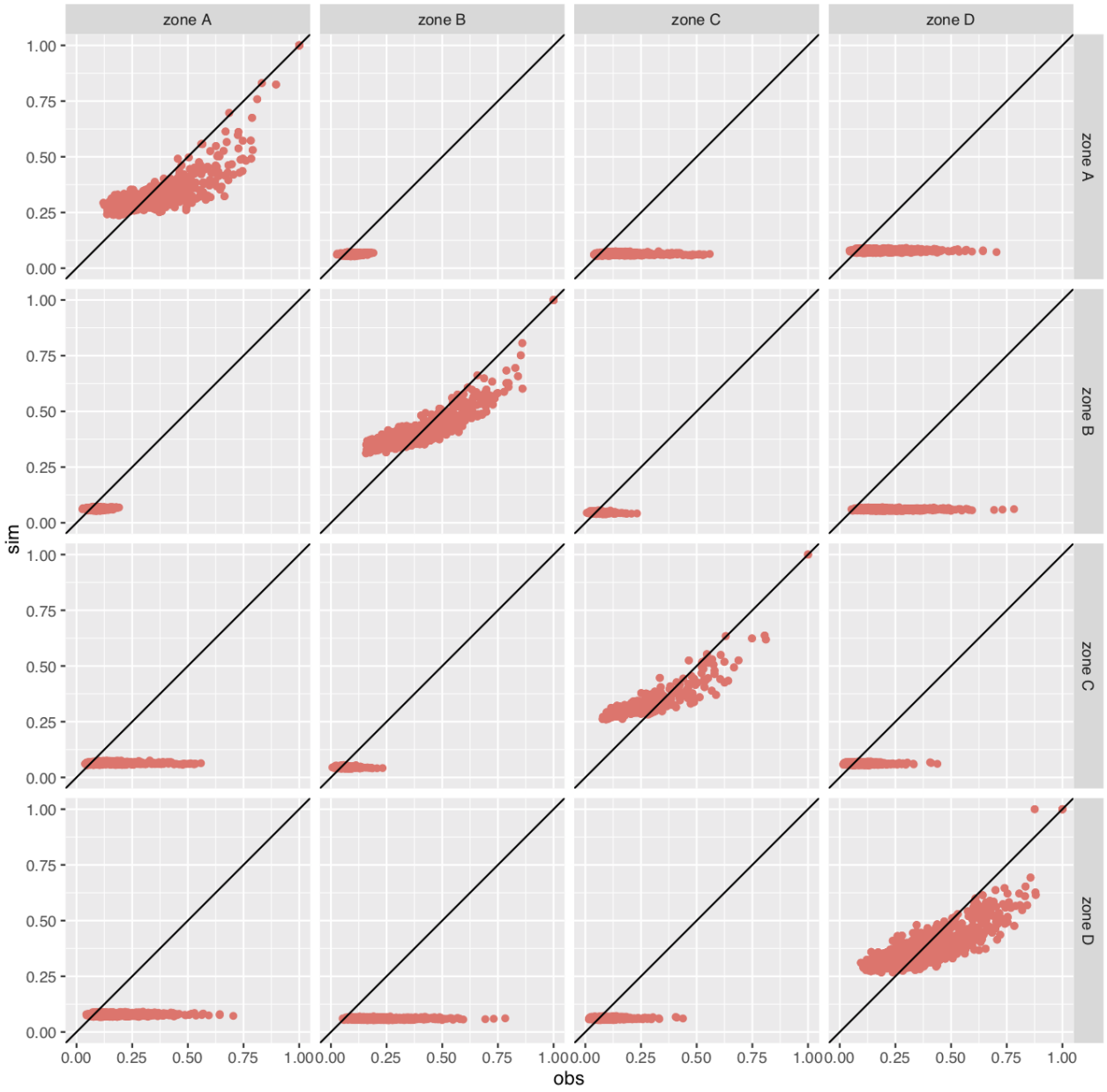


Figure 6.11: Inter-station correlations between hourly observation and CHMM-reorganization simulation. The length of hourly simulation is $10x10$ *years.*

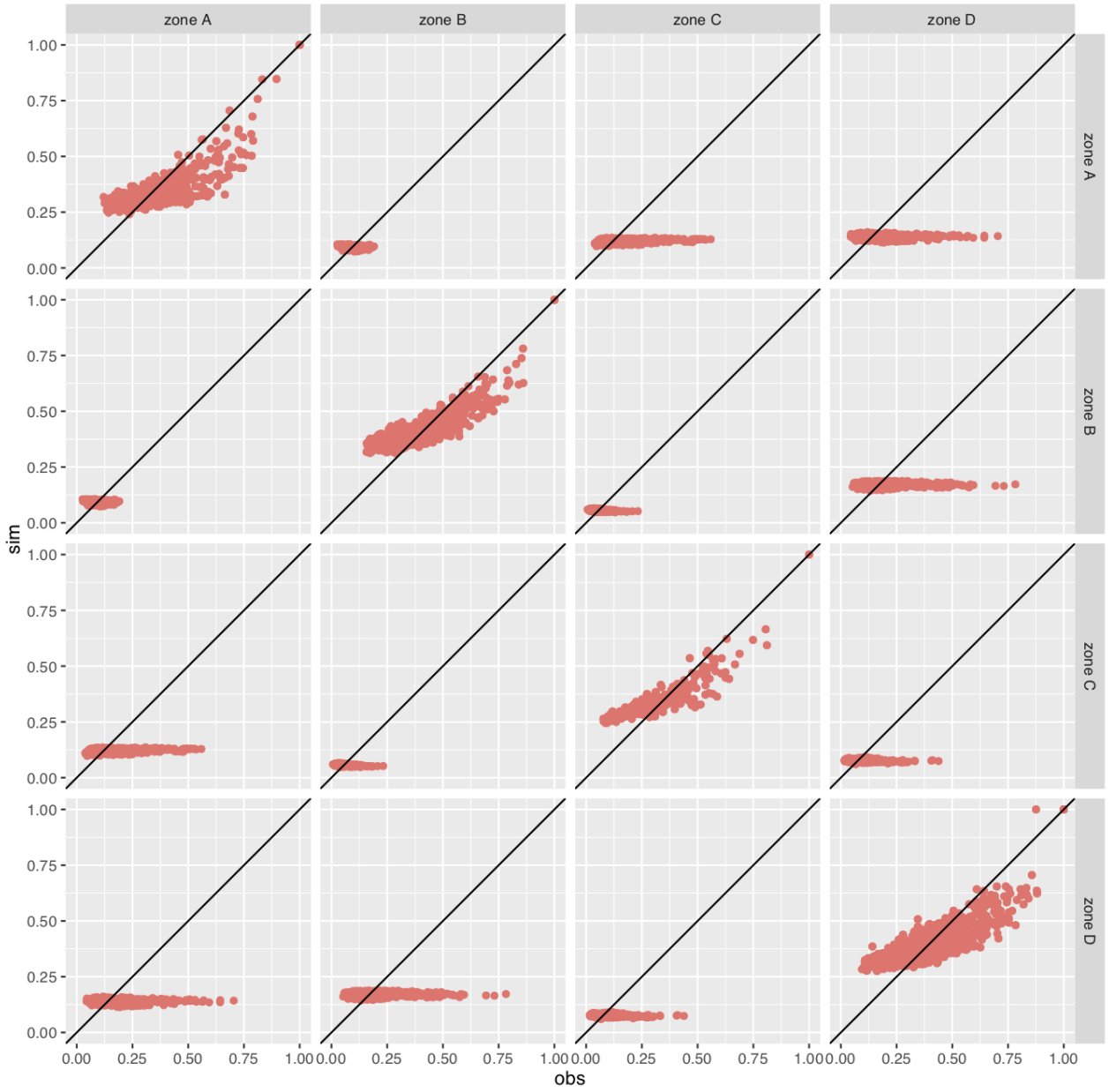


Figure 6.12: Inter-station correlations between hourly observation and resampling simulation. The length of hourly simulation is $10x10$ *years.*

6.3 Temporal correlation

Length of stay distribution

[XXX-YES change "length of stay" to dry/wet spells]

In this Section, the distributions of dry/wet spells are analyzed and discussed, since they express the temporal correlation which is a major issue.

For each time step, if the hourly precipitation values of all rainfall gauge stations in the zone (or the entire rainfall field) are 0, then the zone (or the entire rainfall field) is called dry at the time step, otherwise the zone (or the entire rainfall field) is called wet. Figure 6.13 and Figure 6.14 - Figure 6.15 present the distributions of dry/wet spells which are displayed for each of the 4 models and compared to the reference, for the whole domain and for each of the homogeneous zones, respectively.

As a first comment when looking at Fig. 6.13 - Fig. 6.15, the distributions of wet spell retrieved from the 4 models are over-estimated to the reference one (red line in Fig. 6.13, whatever the area (i.e. whole region versus each of the 4 homogeneous zones). For dry spell, **reorg** performs well comparing with other models, especially for the whole domain and zone A. However, both resampling models are under-estimated with respect to the reference, especially for individual zones. This suggests that the reorganization method prioritized dry spells, possibly because they occupy a majority of time steps in hourly scale. However, the reorganization method is not able to improve the distribution of wet spell, it may be due to the drawback of CHMM which can not produce satisfied simulations.

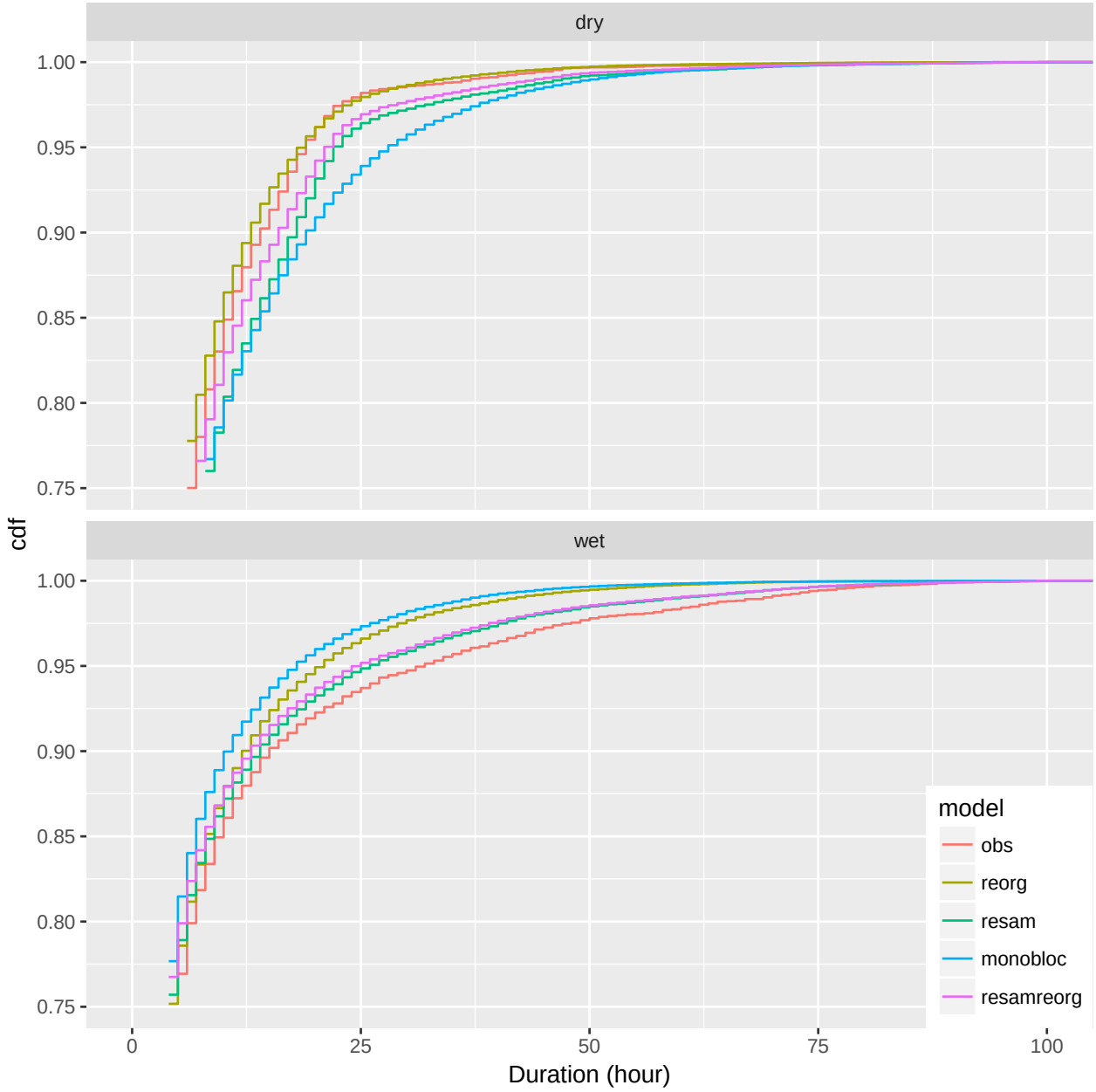


Figure 6.13: Distributions of simulated dry (above) and wet (below) spells, for hourly precipitation. The 4 models are: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method. These results are for the whole region. [XXX-YES]*

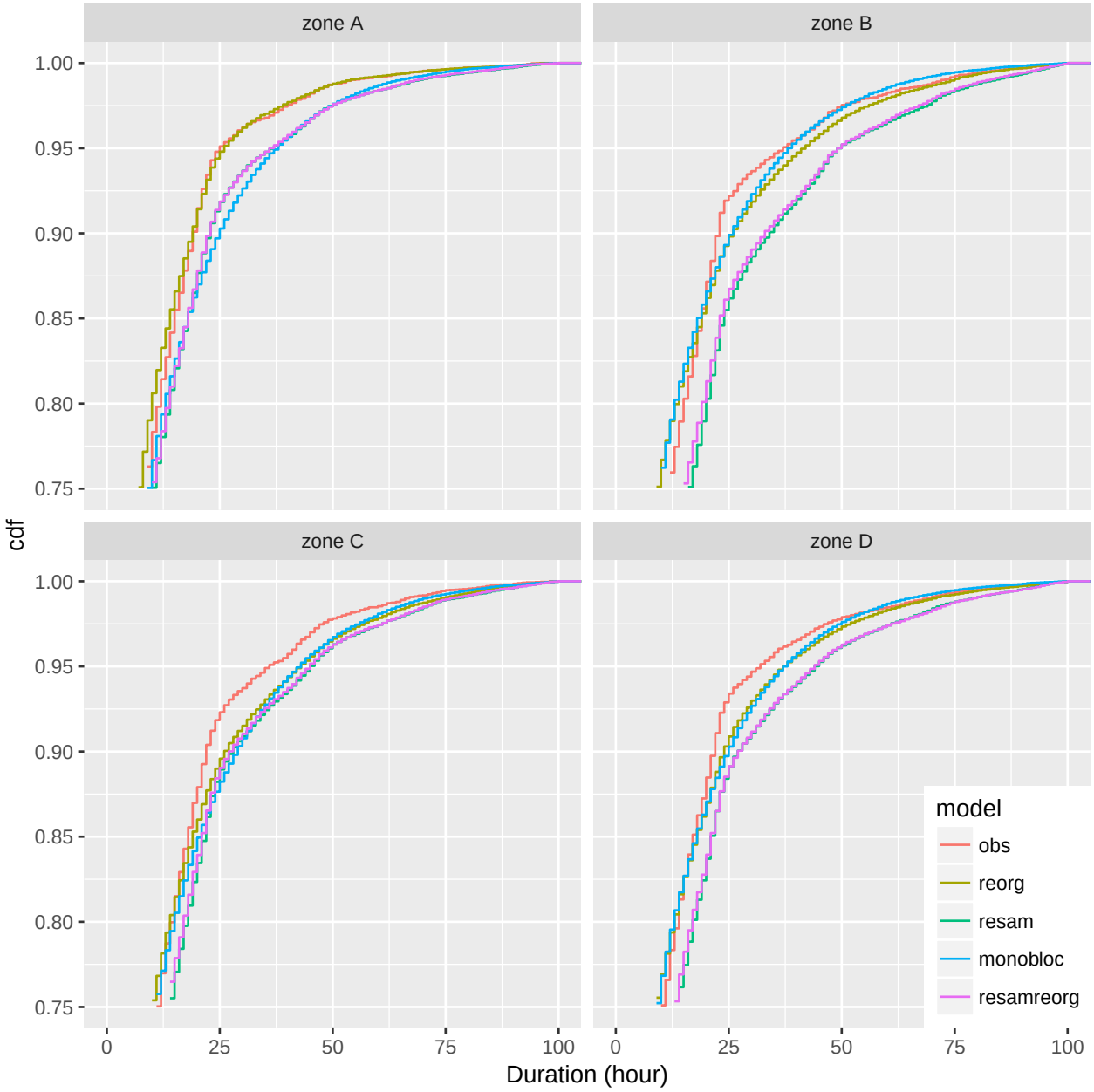


Figure 6.14: Distribution of dry spells with hourly simulation for 4 zones. The 4 models are: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

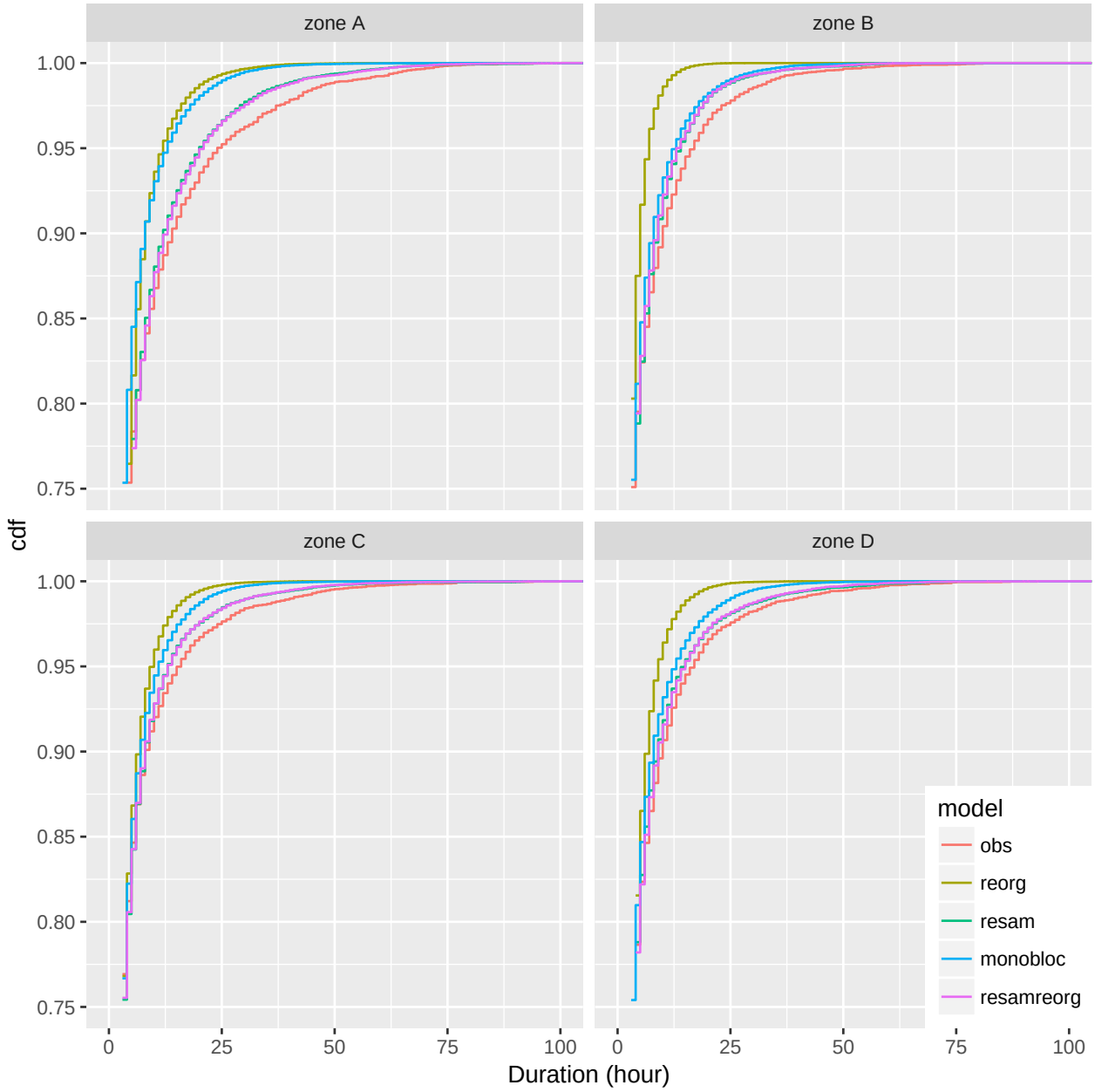


Figure 6.15: Distribution of wet spells with hourly simulation for 4 zones. The 4 models are: ***monobloc*** *stands for SAMPO applied over the entire rainfall field,* ***reorg*** *for SAMPO with CHMM optimized by reorganization method,* ***resam*** *for SAMPO with resampling technique and,* ***resamreorg*** *for SAMPO with resampling technique and reorganization method.*

6.4 Conclusions

This chapter proposes hierarchical techniques to enable a "homogeneous zone" rainfall generator to address rainfall simulation over a large, heterogeneous domain. The development of this new simulator is based on assumed good performance of a rainfall generator over one homogeneous zone. In this PhD work, we use the SAMPO weather generator, but any rainfall-type based generator could be handle in the same way.

To compensate for the priority coordination of hidden Markov models give to spatial description, a reorganization method was introduced to enforce the duration of stay in local rainfall classes. It performs correctly, despite the reorganization does not consider the actual content of qualitative rainfall classes (it could be possible to do slightly better in defining a penalty $W_{i,j}$ of changing rainfall type i to rainfall type j).

On the other hand, the new model with several homogeneous zones presents better scores than when SAMPO is applied on only one large region. Thus, our new approach is useful when spatial scale increases, as spatial variability can be crucial in the applications (when a hydrological model is to be applied for a large river for example).

The new models do not solve entirely the heterogeneity problem, as they still have a bias that is the inter-station correlations for stations located in different zones. This ultimately comes from the very principle of the hierarchical simulation based on homogeneous zones.

CHAPTER 7

METHODOLOGY II: CONTINUOUS TYPE MODEL

In this chapter, a new approach, inspired by copula technique, is proposed as an alternative for the simulation of heterogeneous rainfall fields. The calendars of the rainfall types, as previously defined, are no more used; the proposed approach only uses the rainfall descriptors, and more specifically, for a first exploration, the average rainfall and the rainfall intermittency are kept. Each homogeneous rainfall zone is considered as a large scale (LS) cell for which the average precipitation and the rainfall intermittency are identified. The copula technique is used to jointly simulate the calendars of the average precipitation and the rainfall intermittency of the 4 homogeneous rainfall zones. Then, a geostatistical disaggregation technique is proposed to generate fine scale rainfall fields which respect large scale values.

For each homogeneous rainfall zone, the values of the daily average precipitation and the daily rainfall intermittency are obtained by hourly precipitation data of the 146 rain gauge stations. Thus, a set of 8 calendars (i.e. 2 calendars per zone), as presented in Table 7.1, is built and constitutes the data that will be later used in the copula modeling approach. Within the 2005-2014 period, each calendar may be considered as a time-series of each of the variable, the daily average precipitation in 4 zones (noted **P1**, **P2**, **P3** and **P4**) and the daily rainfall intermittency (noted **I1**, **I2**, **I3** and **I4**), respectively. As an illustration, Table 7.1 presents the data of 12 continuous days, for the 4 homogeneous zones.

7.1 Reconstructed precipitation index

As mentioned in Chapter 2, precipitation is a very complex data in term of its variability within a large range of time and space scales. Precipitation data set presents a large portion of 0 values that can produce two major problems when copula is used. Since, copula uses rank correlation, too many 0 values will hamper the calculation of this correlation. Then, with such data set, it may be difficult to find an appropriate distribution for precipitation data. This problem is obviously not new. Standard approaches used hydrology are either to use a truncated distribution [*Loaiciga et al.*, 1992; *Sharma*, 1997] or to use an independent binary rain indicator $(0/1)$ and non-zero rainfall [*Zucchini and Guttorp*, 1991; *Bárdossy and*

Table 7.1: Daily average precipitation and daily rainfall intermittency for the 4 homogeneous zones. Only 12 continuous days are presented as an illustration. $P1$-$P4$ (in mm) stand for the daily average precipitation in 4 zones. $I1$-$I4$ (without unit) stand for the daily rainfall intermittency in 4 zones. [XXX-YES indicated the units.]

Date	$P1$	$P2$	$P3$	$P4$	$I1$	$I2$	$I3$	$I4$
2005-04-10	0.20	1.28	0.03	0.73	0.29	0.47	0.05	0.53
2005-04-11	0.00	0.02	0.00	0.01	0.00	0.03	0.00	0.07
2005-04-12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2005-04-13	0.03	0.00	0.00	0.00	0.14	0.00	0.00	0.00
2005-04-14	0.01	0.00	0.00	0.00	0.07	0.00	0.00	0.00
2005-04-15	7.50	0.14	5.24	14.25	0.86	0.19	0.74	1.00
2005-04-16	4.54	58.64	59.29	17.74	0.79	1.00	1.00	1.00
2005-04-17	5.52	38.51	7.69	17.85	0.93	1.00	0.95	1.00
2005-04-18	0.13	7.96	0.08	0.27	0.29	0.94	0.21	0.40
2005-04-19	1.23	1.06	1.32	4.44	0.79	0.69	0.74	0.93
2005-04-20	1.34	0.97	0.15	1.83	0.43	0.56	0.16	0.73
2005-04-21	0.73	0.15	0.01	0.25	0.36	0.11	0.05	0.33

Plate, 1991]. Both approaches have advantages and drawbacks.

In this section, a third approach is proposed and takes advantage of neural network. Indeed, the precipitation data is linked to a smooth variate obtained from artificial neural network (ANN) approach. For more theoretical details on artificial neural network, refer to Appendix D.

The method is the following. The inputs of ANN are several meteorological variables, extracted from the ERA-Interim database and suitable to reconstruct precipitation data. The meteorological variables, the location and the atmospheric levels to be considered have been optimized by using ANN with the criterion of the best coefficients of determination.

The target of ANN for training step is just observed precipitation data. Only the non-zero precipitation values are used when tuning the ANN. Precisely, we use normalized observed precipitation data $(\frac{\mathrm{P}}{\max \mathrm{P}})^{0.3}$ (for all days where P>0) for training step. Then we apply the ANN to recover complete time-series of rainfall-related reconstructions, also for all dry days. The reconstructed precipitation is later referred as the **precipitation index**. It appears that reconstructions may generate positive or negative "precipitation values" depending on the estimated ANN, and this is especially true for dry days. This index is interesting as a diagnostic property ; considering all the 0 in the observed precipitation time series it appears that the ANN makes a difference between nearly-rainy and really-dry contexts. Finally precipitation could be quantile-quantile related to the precipitation index, of course with some uncertainty in the reconstruction. Figure 7.1 presents the diagram of ANN for the training step. 28 meteorological variables data are used in the training phase to reconstruct the non-zero precipitation. Three hidden nodes have been considered. The

coefficient of determination between the normalized non-zero precipitation data and the reconstructed precipitation index is generally more than 0.7 (0.702 in this very case) which makes this approach encouraging in its use.

Figure 7.2 presents the comparison between the normalized precipitation data and the reconstructed precipitation index, for the 2012 year. During several periods when the reconstructed precipitation index turned negative, the observed precipitation data showed considerable dry spells. On the contrary, when the observed precipitation data showed a very short dry spell, the precipitation index could remain positive value but very close to zero. Due to the information provided by the other meteorological variables through ANN, the precipitation index can make the different between nearly dry (positive value), dry (0) and very dry (negative value). In the same way, we transform the rainfall intermittency data to the reconstructed intermittency data referred as the **intermittency index**.

Learning from the atmospheric context, the precipitation index is a data-mining reconstruction that has similarities to and possibly justifies the latent variate, included in several stochastic rainfall models [e.g., *Bárdossy and Plate*, 1991; *Vrac et al.*, 2007; *Baxevani and Lennartsson*, 2015].

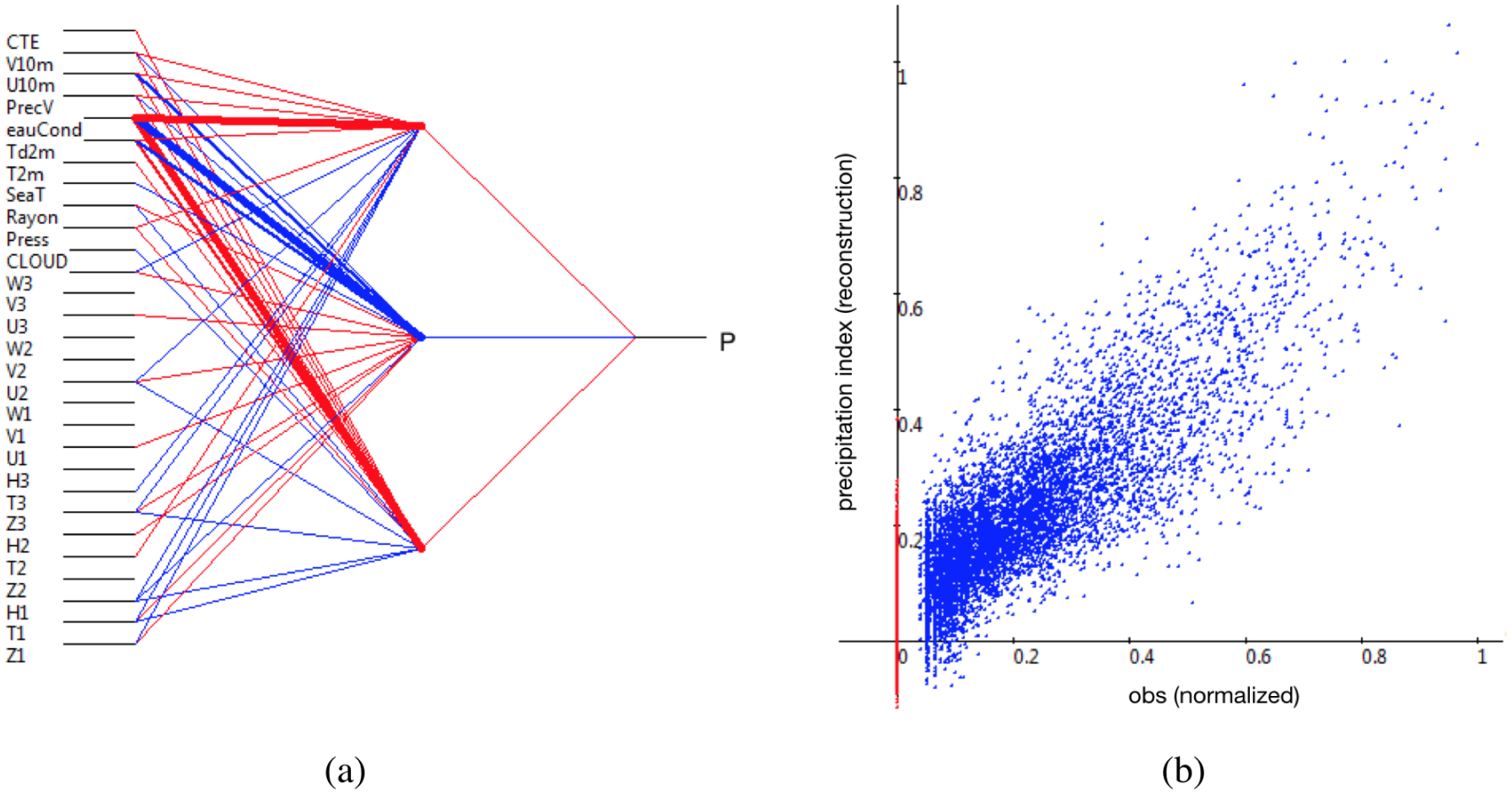


Figure 7.1: (a) Diagram of artificial neural network implement in the training phase where 3 hidden nodes have been chosen. The normalized precipitation (P) is the output node. The input nodes contain 28 meteorological variables and a constant (CTE= 1). Among them, there are 10 *surface variables which are* 10 *metre V wind component (V10m), 10 metre U wind component (U10m), water vapor (PrecV), Condensed water (eauCond), 2 metre dewpoint temperature (Td2m), 2 metre temperature (T2m), sea surface temperature (SeaT), surface solar radiation (Rayon), surface pressure (Press), total cloud cover (CLOUD). Other* 18 *variables are geopotential (Zi), temperature (Ti), relative humidity (Hi), U component of wind (Ui), V component of wind (Vi) and vertical velocity (Wi) at 3 pressure levels which are* 400 *hPa (i =* 1*),* 500 *hPa (i =* 2*) and* 825 *hPa (i =* 3*). All data where taken from ERA-Interim. The red lines indicate the positive correlations whereas the blue ones the negative correlations. The thicker the line is, the stronger the correlation is. (b) Scatter plot between the normalized non-zero precipitation data and the reconstructed precipitation index. The coefficient of determination is* 0.702*.*

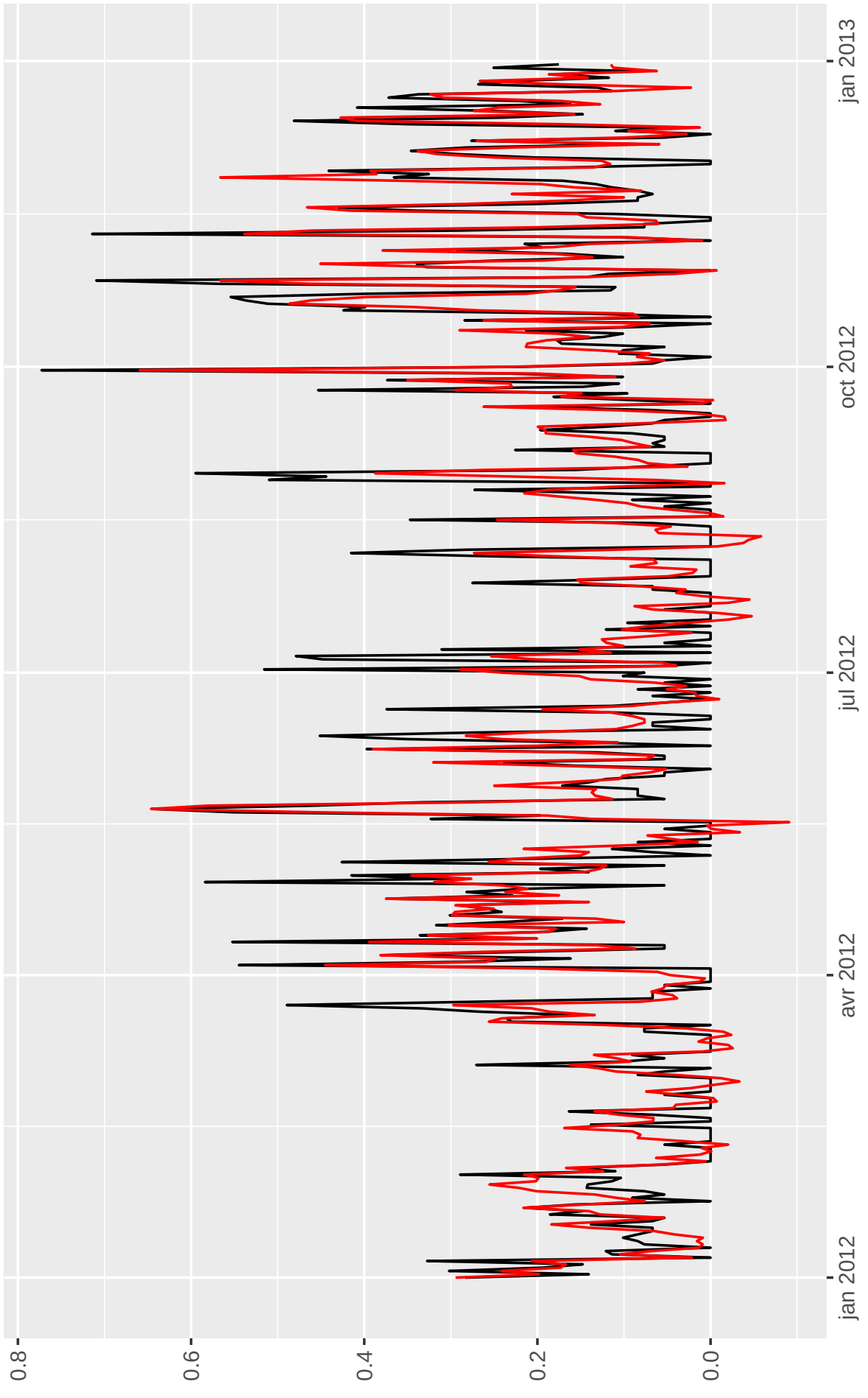


Figure 7.2: Comparison between the normalized daily precipitation data (in black) and the reconstructed precipitation index (in red), for the 2012 year in the Mediterranean homogeneous zone (zone A).

7.2 Copula based parametric model

This section aims at simulating the time-series of daily average precipitation and daily rainfall intermittency for several homogeneous rainfall zones.

Our approach relies on two steps. First, the copula technique is to be implemented in combination with auto-regressive process to capture the joint distribution of multi time-series calendars and temporal correlation of each variable. Then, kriging technique is applied to generate the sequential simulations.

7.2.1 Copula approach

Copulas are functions that link multivariate distribution functions to their constituent univariate marginal distributions [*Nelsen*, 2007]. Basic elements about copula technique can be found in Appendix C.

Sklar's Theorem (C.4) provides the framework where copula plays an essential role between multivariate distribution functions and each marginal distribution. Each marginal distribution is estimated separately. The main issue is then to identify the proper copula function to hold the different margins together into a joint multivariate distribution.

In our case, Gaussian copula is used for its simplicity and, has will be shown, explicit links it has with other statistics tools. The Gaussian copula is a distribution over the unit cube $[0,1]^d$. It is constructed from a multivariate normal distribution over $\mathbb{R}^d$ by using the probability integral transform. For a given correlation matrix $R \in [-1,1]^{d\times d}$, the Gaussian copula, with parameter matrix R, is written as

$$C_R^{\text{Gauss}}(u) = \Phi_R(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_d)), \tag{7.1}$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal and $\Phi(R)$ is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix R. The density of Gaussian copula can be written as

$$c_R^{Gauss}(u) = \frac{1}{\sqrt{\det R}} \left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T \cdot (R^{-1} - I) \cdot \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix} \right) \tag{7.2}$$

where I is the identity matrix.

Therefore, only the correlation matrix of d variables (e.g., d calendars in our context) is needed when applying the Gaussian copula.

7.2.2 Auto-regressive process (AR)

The copula technique makes link between the multivariate joint distribution and each marginal distribution. However, the temporal correlation in each calendar needs to be considered as well. For this, the use of auto-regressive model allows to deal with time-varying

process. Indeed, the auto-regressive model (AR) specifies that the output variable linearly depends on its own previous values and on a stochastic term which has a probability distribution with zero mean and finite variance.

The $AR(p)$ notion indicates an auto-regressive model of order p. The $AR(p)$ model is defined as :

$$X_t = a + \sum_{i=1}^{p} \phi_i X_{t-i} + v_t \tag{7.3}$$

where X_t is the scalar value at time t in case of uni-variable AR models ($p = 1$), $\{X_{t-1}, \ldots, X_{t-p}\}$ are the corresponding values at previous time $\{t-1, \ldots, t-p\}$, $\phi_1, \cdots, \phi_p$ are the parameters of the model, a is a constant, and v_t is a white noise.

In case of multivariate AR models, X_t, X_{t-i}, a, ϕ_i are matrices and v_t a stochastic matrix. The combination of the copula technique with auto-regressive model allows the construction of a multi-time-varying-variables model.

7.2.3 Kriging technique

Kriging is a well known method in geostatistics [*Matheron*, 1963] that aims to predict the value of a function at a given point by computing a weighted average of the known values of the function in its neighborhood.

As a reminder, simple kriging is a linear estimation method. A value from location x_1 (notation of a any set of geographic or temporal coordinates) is interpreted as a realization $z(x_1)$ of the random variable $Z(x_1)$. In the space A, where the set of samples is dispersed, there are N realizations of the random variables $Z(x_1), Z(x_2), \cdots, Z(x_N)$, correlated between themselves. Spatial estimation of a quantity $Z : \mathbb{R}^n \to \mathbb{R}$, at an unobserved location x_0, is obtained from a linear combination of the observed values $z_i = Z(x_i)$ and weights $w_i(x_0), i = 1, \ldots, n$:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} w_i(x_0) \times Z(x_i). \tag{7.4}$$

If kriging is often used as a *clever* interpolation technique, this method can be used, more generally, in conditional simulation situation. In this case, the variable $Z(x_i)$ is thus simplified to z_i. After estimating the kriging weights, the simulation is generated by the use of Equation 7.4 with an added random residual. In our case, only simple kriging is used. Simple kriging is mathematically the simplest, but the least general [*Olea*, 2012]. It assumes the expectation of the random field to be known, and relies on a covariance function. The kriging weights of simple kriging have no unbiasedness condition and are given by the simple kriging equation system:

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} v(x_1, x_1) & \cdots & v(x_1, x_n) \\ \vdots & \ddots & \vdots \\ v(x_n, x_1) & \cdots & v(x_n, x_n) \end{pmatrix}^{-1} \begin{pmatrix} v(x_1, x_0) \\ \vdots \\ v(x_n, x_0) \end{pmatrix} \tag{7.5}$$

where $v(x, y) = \mathrm{Cov}(Z(x), Z(y))$. In this case, the estimation of variance is:

$$\sigma^2 = v(x_0, x_0) - \sum_{i}^{n} v(x_i, x_0) w_i \tag{7.6}$$

In the approach developed in this section, the conditional simulation is based on sequential kriging with uncertainty. Within the Gaussian framework chosen from the beginning, a drawn value is thus obtained from conditional distribution by the use of $\hat{Z}(x_0)$ (the expected value) and σ^2 (the prediction variance) before proceeding the estimation to the next step.

7.2.4 Diagnosis and results

Several statistical diagnosis are presented in this section. The used data for modeling are the 8 time-series of daily data from 2005 to 2014 noted (P1, P2, P3, P4, I1, I2, I3, I4) which Pi is the daily average precipitation of zone i and Ii is the daily rainfall intermittency of zone i. The simulations are 50 replications of the 10-years period as the same length as the 2005-2014 period.

The observed data used to model are the reconstructed precipitation index and the intermittency index (Section 7.1). The simulated values are thus a simulated precipitation index and a simulated intermittency index. These two simulated values are then transformed to "real" values by using quantile-quantile plot of observed data.

The comparisons between the observed and the simulated values are presented in Fig. 7.3 and 7.4, for respectively, the average daily precipitation and rainfall intermittency and their associated standard deviations, given in the 4 homogeneous zones. The general agreement is good.

The observed 2005-2014 average values (in black) are inside of the boxes (in red) given the statistics of the 50 replications of the 10-year simulated period. Furthermore, the results for the daily rainfall intermittency are better than those of the daily average precipitation, since the mean simulated values (horizontal red line) are very closed to the mean observed values (black dots), whatever the homogeneous zone. This may be due to the variability of the average precipitation higher than the rainfall intermittency.

Figure 7.5 and 7.6 show the comparison between the observed and the simulated marginal distributions. The general results are satisfactory, but in all cases, the comparison fails for high extreme values.

Moreover, Figure 7.7, 7.8 and 7.9 present the bivariate diagnosis between the observed and the simulated data. For this analysis, the period length must be the same for the observed and the simulated data. Therefore, just one replication among the 50 is randomly chosen for each variable. Figure 7.7 compares the daily average precipitation in the 4 homogeneous zones. The simulations seem more dispersive than the observations as illustrated in the upper right part of the figure. But surprisingly, the bivariate distributions between the simulated and the observed data, showed the lower left part of the figure, are quite good. Figure 7.8 compares the daily rainfall intermittency in the 4 homogeneous zones. The upper right part of the figure gives less information due to the structure of the daily rainfall intermittency. However, the bivariate distributions of simulated values are quite similar to the observations. It's difficult to show all results in the same figure, so Fig. 7.9 is selected as an illustration of the comparison between zones 1 and 2. Focusing the 4 plots

(P1, I1), (P1, I2), (P2, I1) and (I1, I2) in the upper right part of the figure, the simulated values of the daily average precipitation are generally slightly overestimated, especially when the values of the average rainfall intermittency are close to 1. Nevertheless, the bivariate distributions between the simulations and observed data show good consistency.

The auto-correlation function (ACF) is finally used to examine the temporal correlations for each variable. The results shown in Fig. 7.10 are generally satisfactory. Even for 20 days in lag, the ACF values are preserved in the simulated values as compared in the observed data. I1 is a very good example.

These different statistical diagnosis between the simulated and the observed data present very encouraging and promising results. The diagnosis of both marginal distributions and joint distributions show the power of copula technique when dealing with the multivariate framework. The auto-regressive process ensures that the temporal correlations of the variables can be preserved. At this point, the simulations generated by this copula based parametric model are usable for the further disaggregation approach.

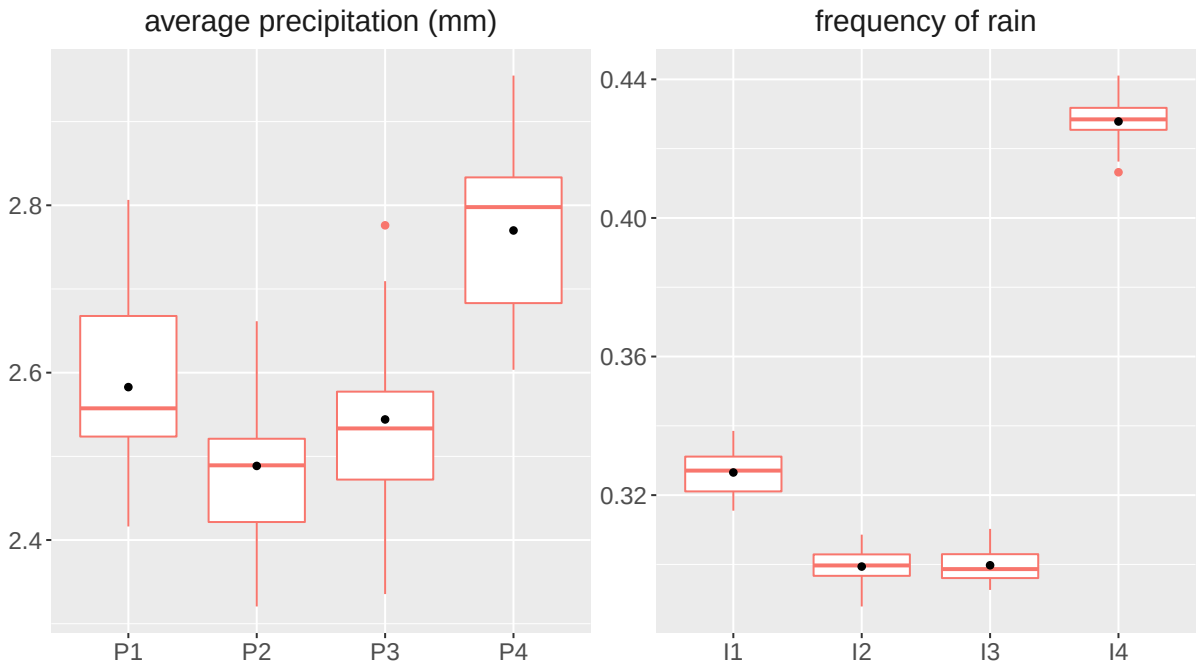


Figure 7.3: Statistics, illustrated with boxplots, of the average values for the daily average precipitation and the daily rainfall intermittency in the 4 homogeneous over the 10-years period. Black points refer to the observed average value. The statistics of the simulated values, in red, rely on the 50 replications of the 10-years simulations.

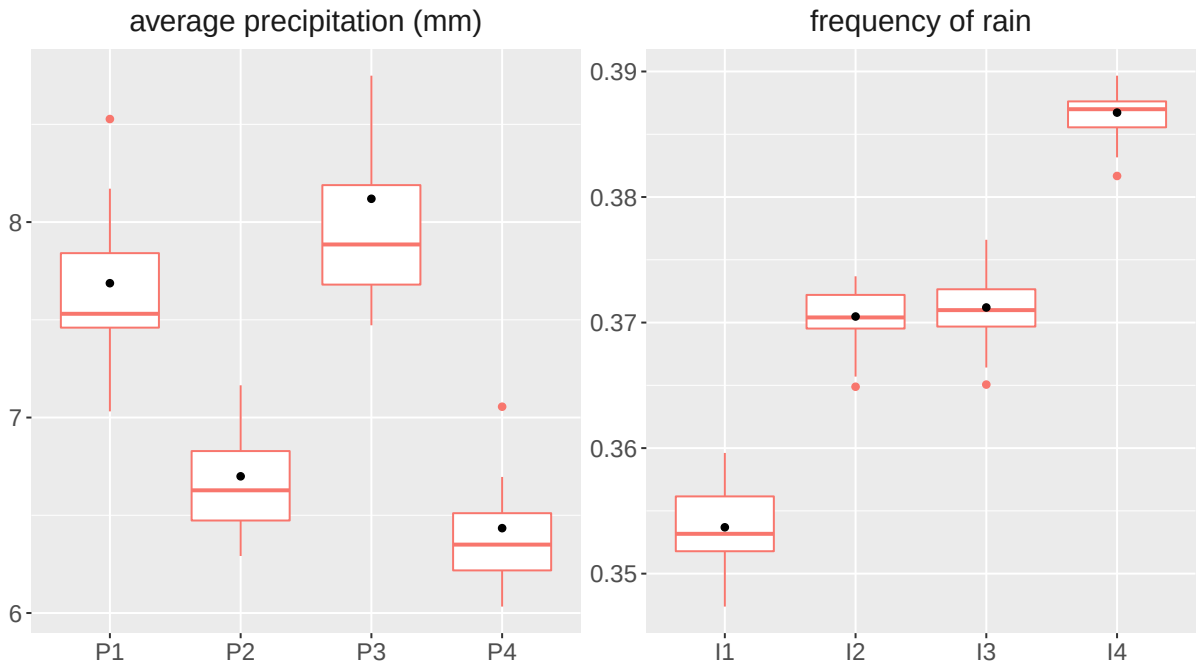


Figure 7.4: Statistics, illustrated with boxplots, of the standard deviation values for the daily average precipitation and the daily rainfall intermittency in the 4 homogeneous zones over the 10-years period. Black points refer to the observed average value. The statistics of the simulated values, in red, rely on the 50 replications of the 10-years simulations.

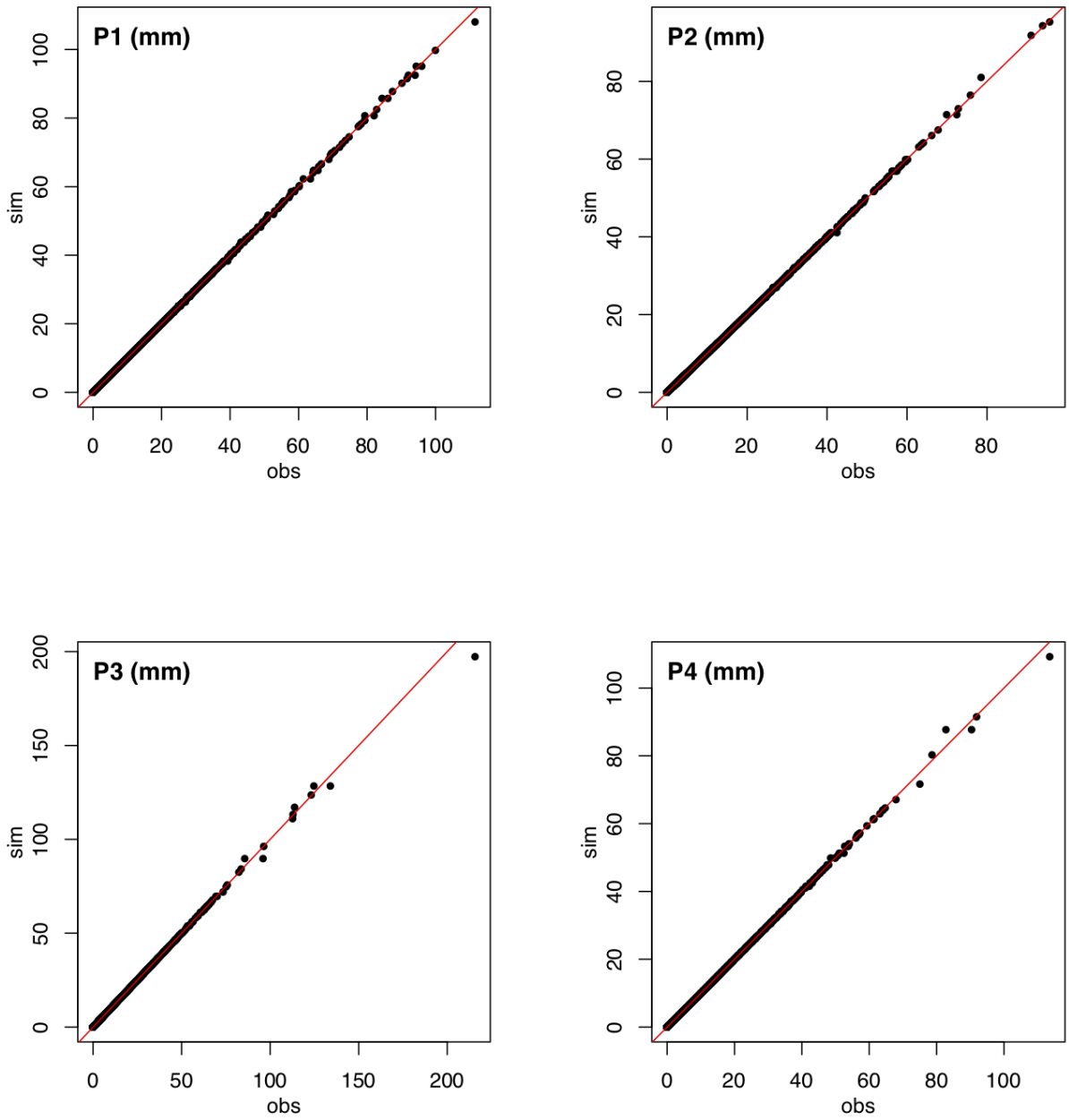


Figure 7.5: QQ-plots for the daily average precipitation by comparing the marginal distributions bewteen the observed and the simulated data in 4 zones.

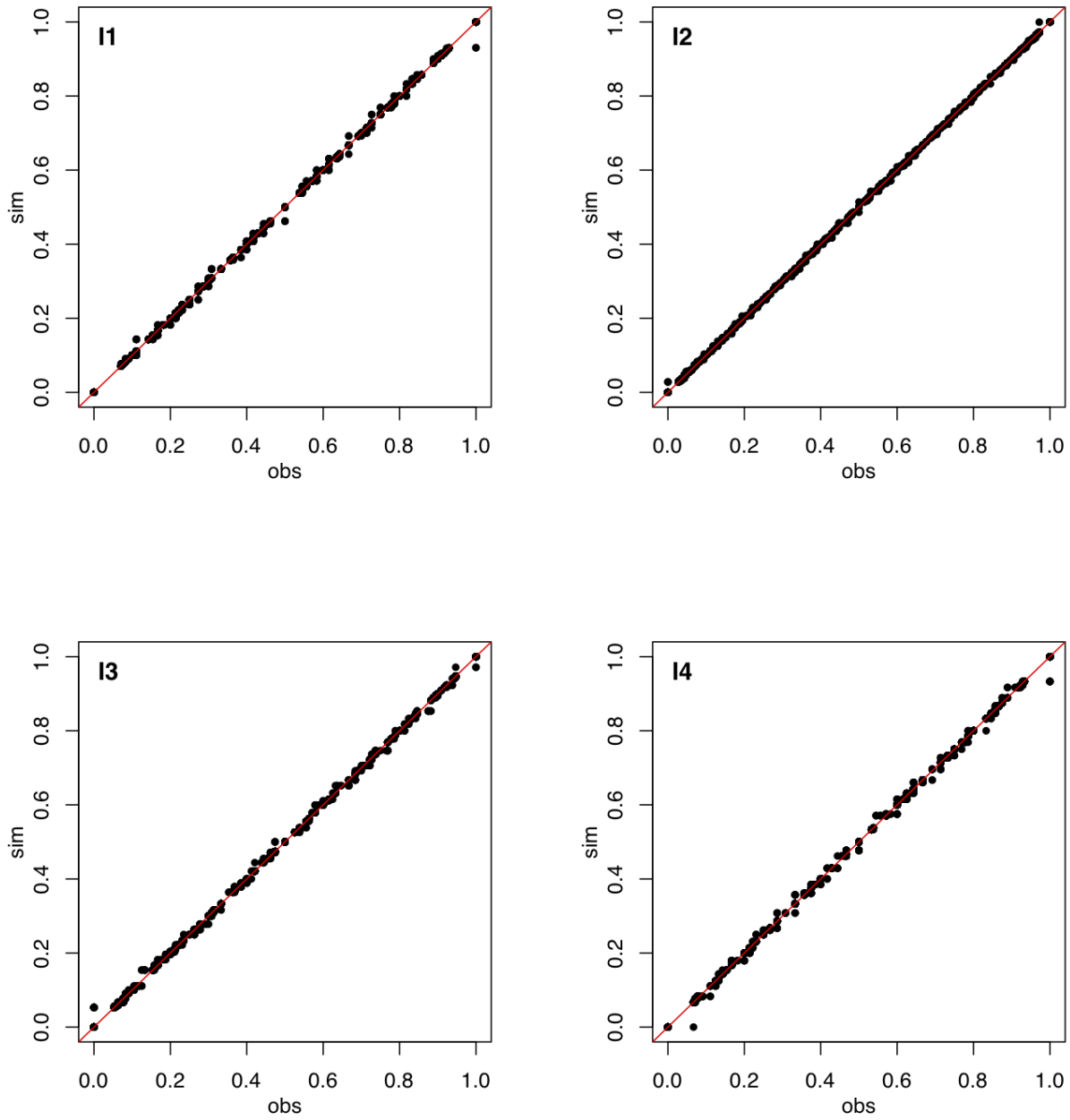


Figure 7.6: QQ-plots for the daily rainfall intermittency by comparing the marginal distributions bewteen the observed and the simulated data in 4 zones.

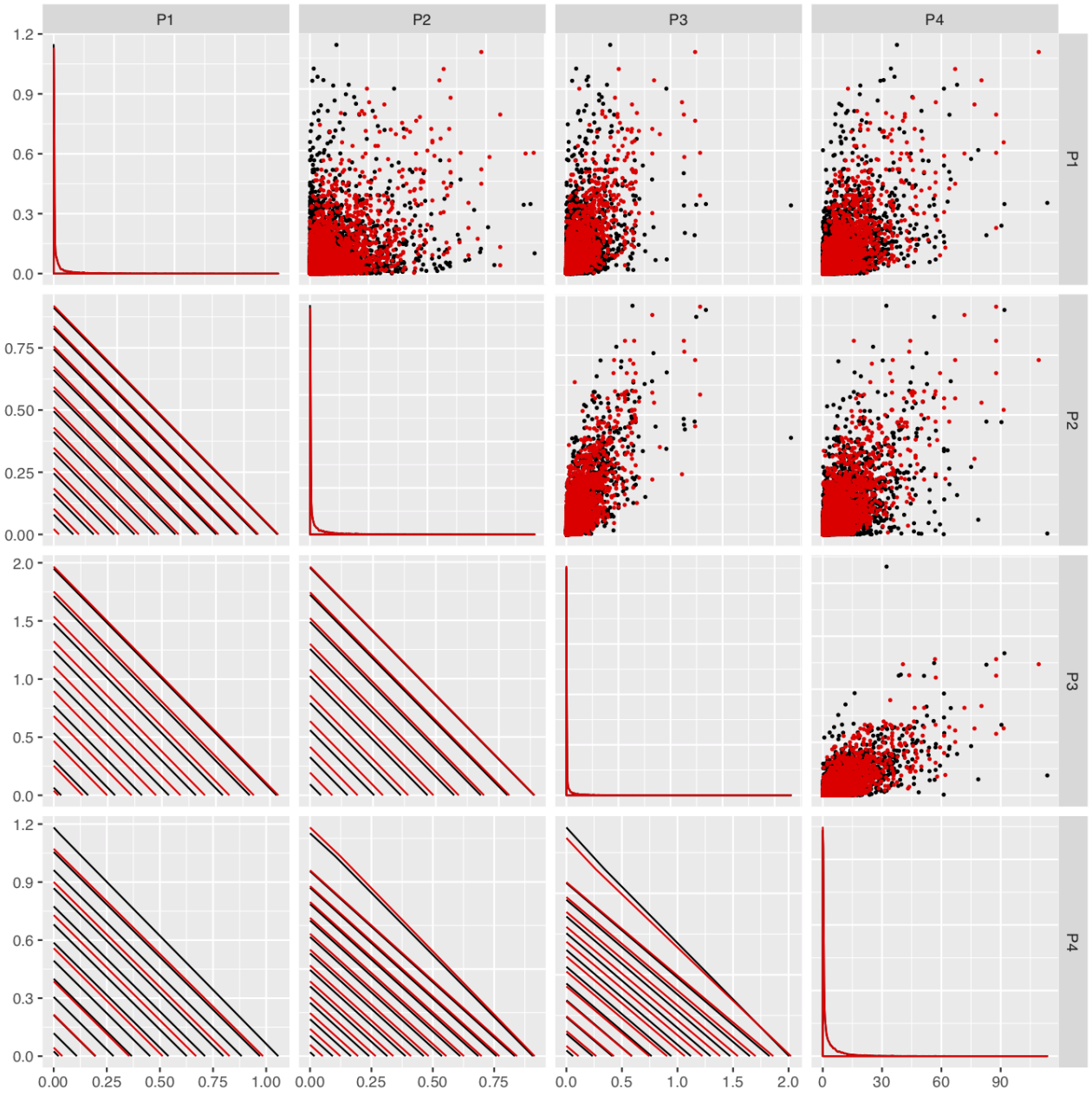


Figure 7.7: Bivariate distributions of the observed (in black) data within the 2005-2014 period and the 10-years simulations (in red) for the daily average precipitation. The simulated contours refer to the mean of the 50 replications of the 10-years simulation. Upper right, the point-by-point representations of all pairs of variables; lower left, the bivariate distributions of all pairs of variables; in the diagonal, the comparison between the observed and the simulated data distribution for each variable.

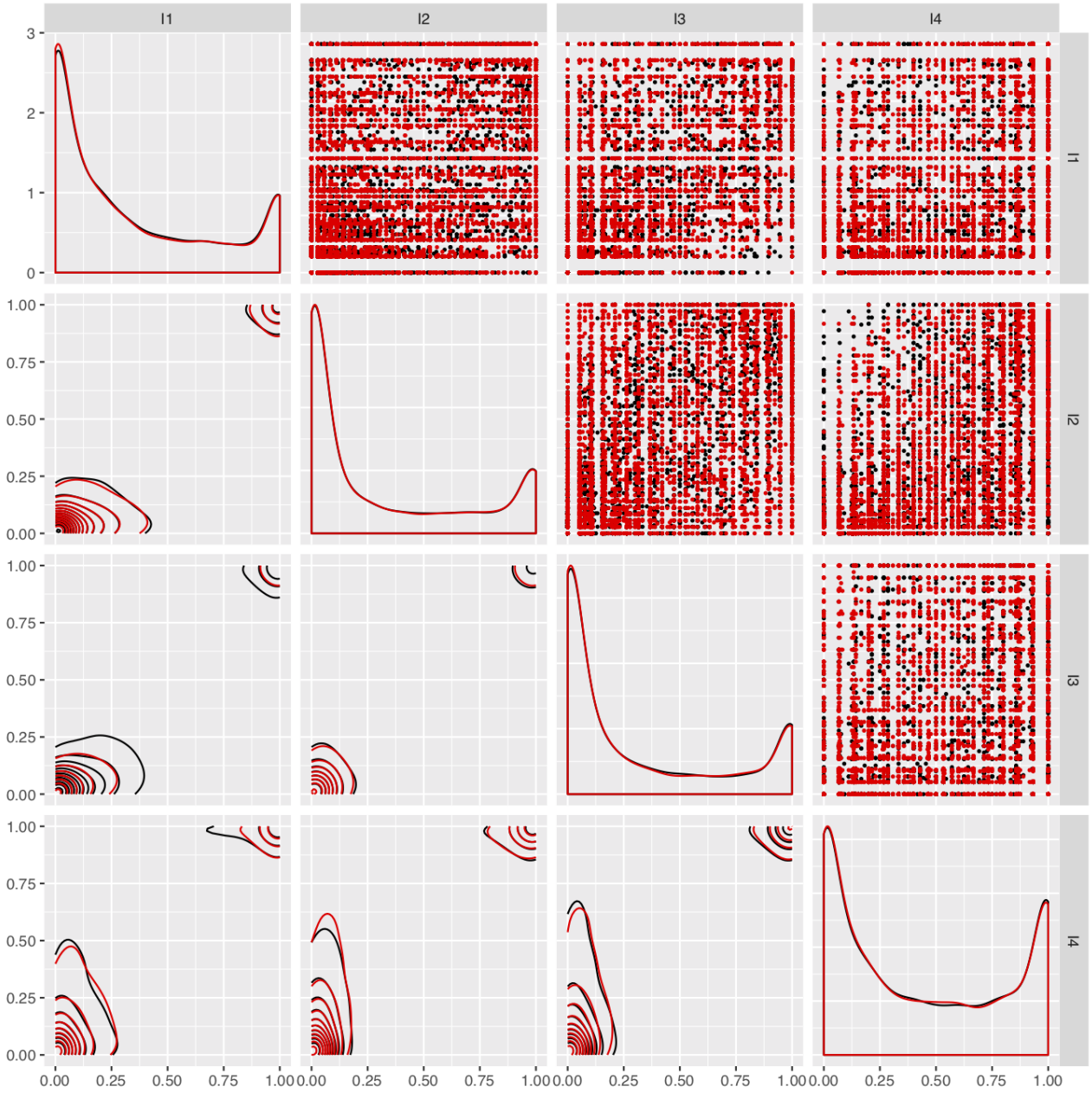


Figure 7.8: Bivariate distributions of the observed (in black) data within the 2005-2014 period and the 10-years simulations (in red) for the daily rainfall intermittency. The simulated contours refer to the mean of the 50 replications of the 10-years simulation. Upper right, the point-by-point representations of all pairs of variables; lower left, the bivariate distributions of all pairs of variables; in the diagonal, the comparison between the observed and the simulated data distribution for each variable.

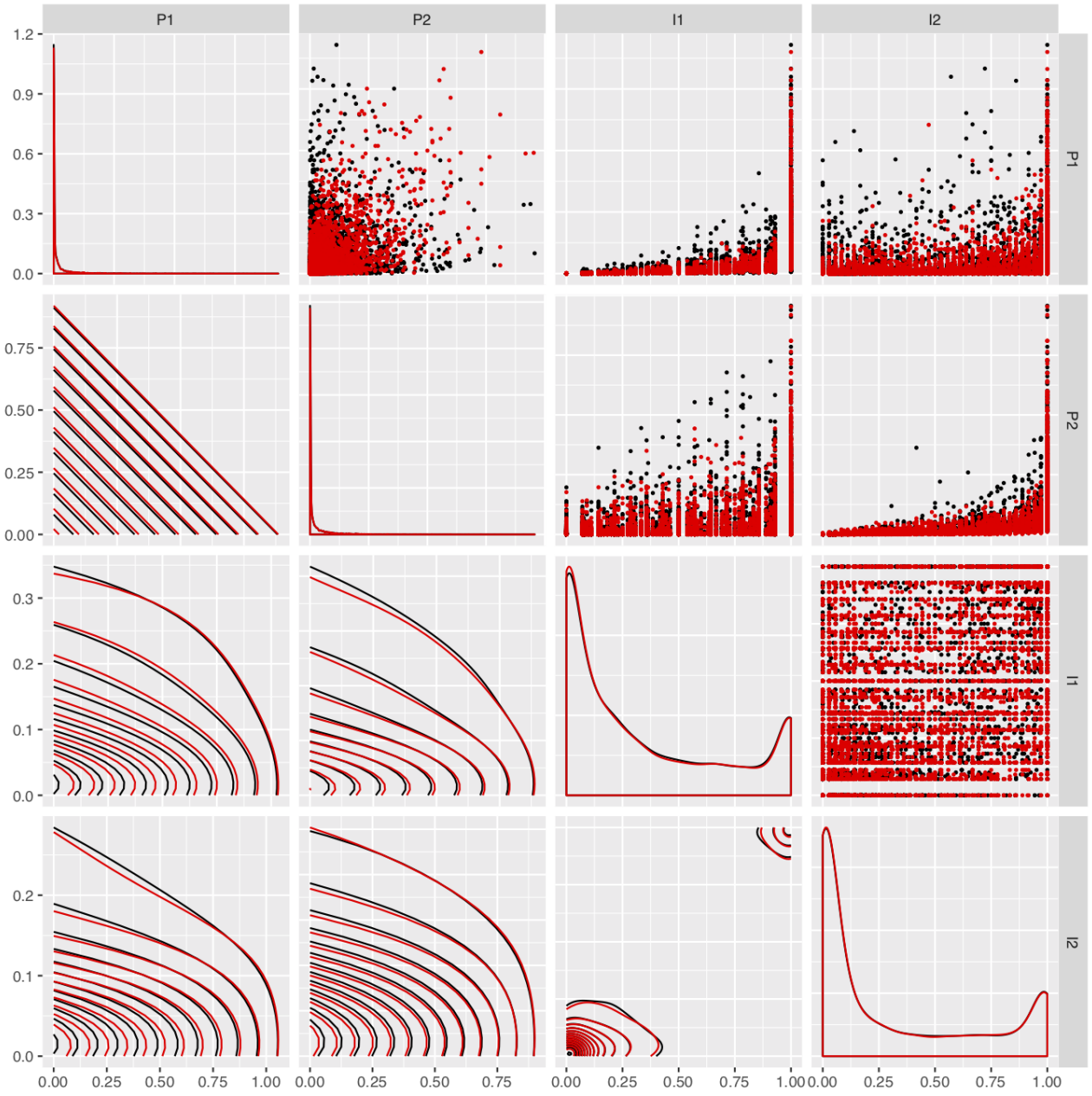


Figure 7.9: Bivariate distributions of the observed (in black) data within the 2005-2014 period and the 10-years simulations (in red) for zone 1 and zone 2. The simulated contours refer to the mean of the 50 replications of the 10-years simulation. Upper right, the point-by-point representations of all pairs of variables; lower left, the bivariate distributions of all pairs of variables; in the diagonal, the comparison between the observed and the simulated data distribution for each variable.

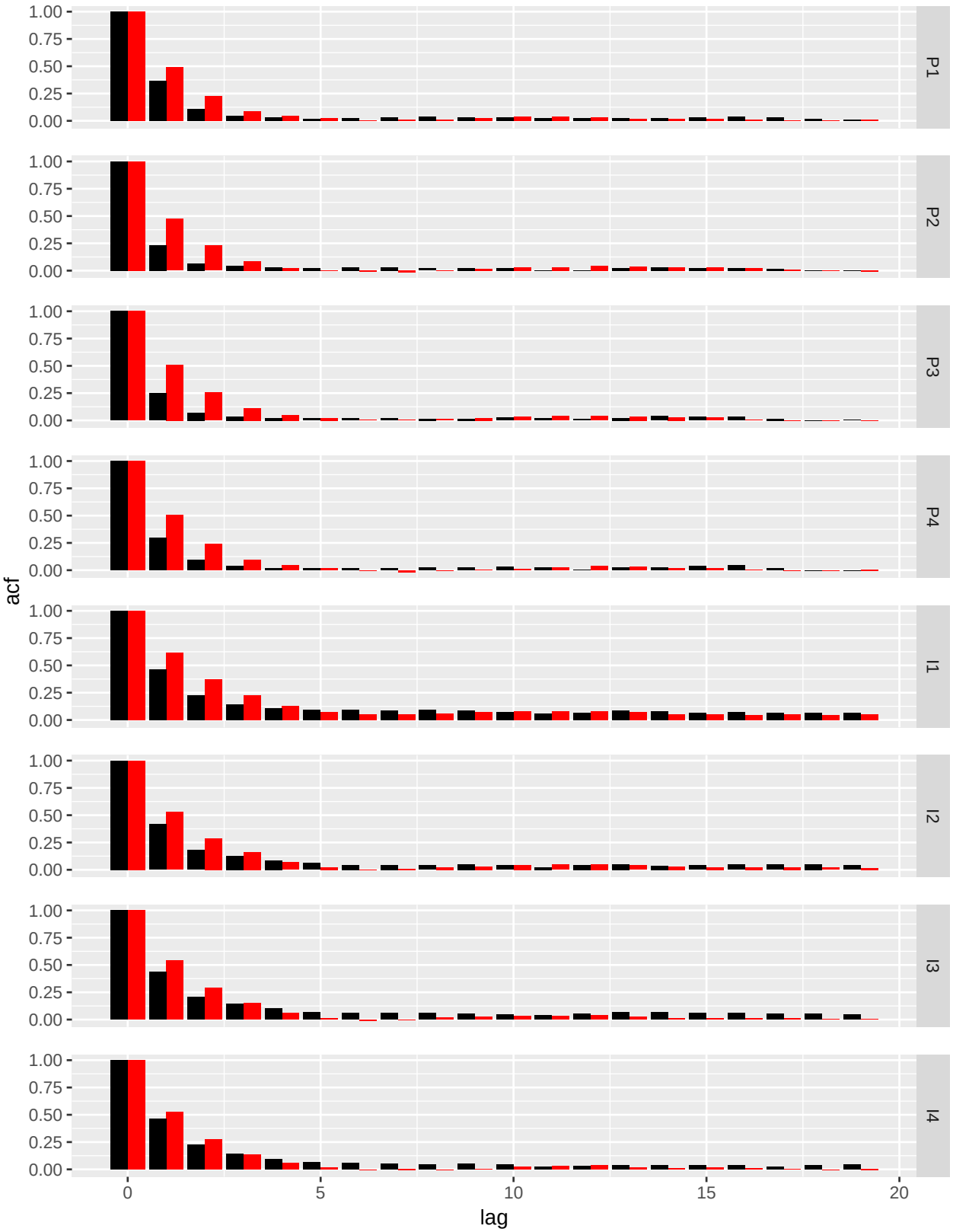


Figure 7.10: Auto-correlation functions of the observed (in black) data within the 2005-2014 period and the 10-years simulations (in red) for the daily average precipitation and the daily rainfall intermittency in 4 zones.

7.3 Geostatistical disaggregation of a rainfall field

At any time step, each homogeneous zone is described by two rainfall descriptors (i.e. the average precipitation of the zone, the rainfall intermittency) as the large scale (LS) values. As seen in Section 7.2, the simulated large-scale values calendars are generated by the use of copula model. In this section, a new disaggregation method is proposed to generate a fine-resolution rainfall field which respects large scale values. In our case, at a given time step, each rainfall zone of a fine resolution simulation must be preserved the average daily precipitation and the rainfall intermittency, provided by the copula based parametric simulations.

7.3.1 Method algorithm

Two steps are achieved in the disaggregation method. At a given time step, the small-scale intermittency field is first generated and needs to respect the LS value. Then, the small-scale non-zero rainfall field is generated and needs to respect the LS non-zero precipitation value. The product of the intermittency field with the non-zero rainfall field constitutes the final simulated rainfall field. The quantity to disaggregate is total rainfall, but the fractional wet area (the fraction of the area with positive rainfall) can also be preserved.

More precisely, the intermittency step is presented as follows and is later illustrated in the next section:

1. generate a Gaussian field for the small-scale grid having the expected spatio-temporal structure suitable for thresholding-based simulation of intermittency.

2. check how the thresholded field compares with the expected large-scale wetness values

3. where the recovered wetness is too much, gently lower the Gaussian field ; where the recovered wetness is not enough, gently higher the Gaussian field.Adjust as necessary to recover every prescribed large-scale wetness value.

Then, the non-zero rainfall step is presented as follows:

1. generate a Gaussian field for the small-scale grid having the spatio-temporal structure suitable for anamorphosis-based simulation of non-zero rainfall.

2. check how the anamorphosed field compares with the expected large-scale rainfall values (keep only the average on wet cells).

3. where the recovered precipitation amount is too much, gently lower the Gaussian field; where the recovered precipitation amount is not enough, gently higher the Gaussian field. Adjust as necessary to recover every prescribed large-scale wetness value.

How to "gently change" a Gaussian field ?

The basic idea is to choose one scalar shift value per control zone (LS cell); using block-to-point kriging [*Kerry et al.*, 2012], these shifts are distributed (interpolated) into a small-scale grid and this small-scale distributed shift is added to the Gaussian grid. (A warning about block-to-point kriging: what is easily found about block kriging usually refers to point-to-block kriging, where the kriging matrix is between data points and the right hand vector is average covariance between data and a target domain of finite non point size, so a block. Here we really mean block-data to points kriging, where the kriging matrix is build on block covariance between data blocks and the right vector is covariance between data blocks and target point.) As the kriging variance is not needed, using dual-formulation of kriging [*Royer and Vieira*, 1984] is possible.

How to respect the exact values of large scale ?

The previous paragraph has explained briefly how we "gently change " a Gaussian field. In order to enable each control zone to respect the large scale values, a technique called dichotomy method is introduced. The dichotomy method is a root-finding method that repeatedly bisects an interval and then selects a sub-interval in which a root must lie for further processing. The technique was inspired by solutions for solving inverse problems in hydrogeology given by *Certes and de Marsily* [1991]; *de Marsily et al.* [1999] under the name of pilot points approach. Their motivation was then aquifers reconstruction given observed macro-scale properties. The pilot points where points of arbitrary value set to condition the Gaussian field underlying the aquifer simulation. In changing their values, the aquifer properties could be tuned as wished.

In our context, the dichotomy method will allow each control zone of a arbitrary Gaussian field to come close to prescribed large-scale value by gently changing the Gaussian field in each iteration, until the large-scale values are respected for all control zones.

7.3.2 Application: the Cévennes-Vivarais region

As mentioned previously, the 4 homogeneous rainfall zones (Fig. 2.10) are the large scale area control zones (LS cells). Figure 7.11 presents the control zones (LS cells) delineation and the small-scale 2km resolution grid. The LS cells (i.e. the homogeneous rainfall zone) are described by the two known values, the average daily precipitation (i.e. average rain over the whole LS cell) and the daily rainfall intermittency (i.e. fraction of grid cells in the LS cell presenting a non-zero rain). As an illustration, Table 7.2 presents 12 continuous daily simulated values, obtained with the copula based parametric model, as described in Section 7.2.4, that are used as input in the disaggregation model.

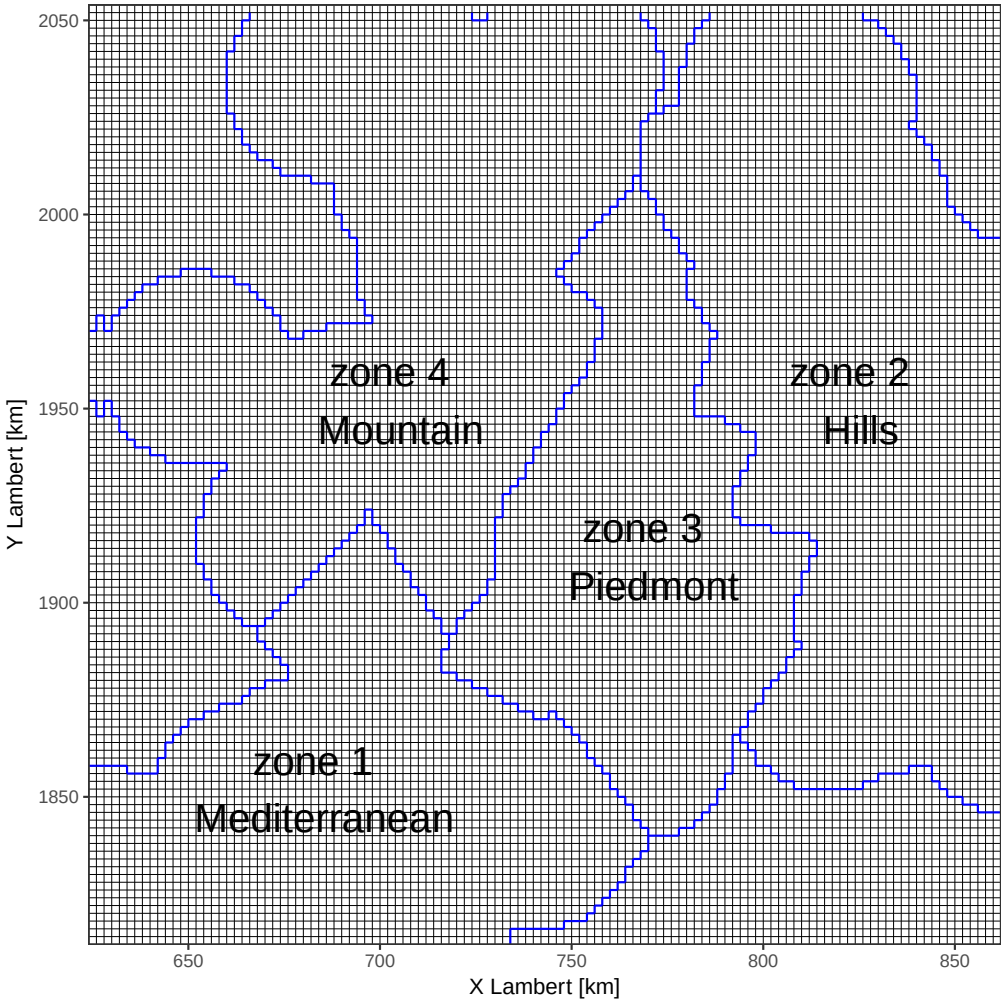


Figure 7.11: Gridded Cévennes-Vivarais region. Two resolutions are shown: large-scale resolution (blue lines) based on the 4 homogeneous rainfall zone, small-scale regular resolution (black squares) given at 2km *resolution.*

Figure 7.12 to 7.20 present 12 continuous daily simulated rainfall fields at each step of the proposed disaggregation model. Let's introduce nz, the number of homogeneous rainfall zones, here $nz = 4$ and nt, the number of time steps, here $nt = 12$.

We begin with the first step which is to generate the intermittency fields. The nt free Gaussian fields having the expected spatio-temporal structure representing the rainfall in-

Table 7.2: Twelve continuous daily simulated values of the average precipitation and the rainfall intermittency in the 4 homogeneous zones.

t	P1	P2	P3	P4	I1	I2	I3	I4
1	1.76	0.01	0.01	3.01	0.64	0.17	0.05	1.00
2	0.00	0.02	0.01	1.53	0.00	0.14	0.05	1.00
3	0.07	0.01	0.28	4.37	0.36	0.64	0.33	1.00
4	0.52	0.55	0.03	6.00	0.43	0.97	0.10	1.00
5	0.41	0.21	0.03	1.44	0.64	0.19	0.63	0.89
6	2.18	1.16	1.68	5.97	0.82	0.86	0.97	1.00
7	7.46	20.32	24.81	11.19	1.00	1.00	1.00	1.00
8	3.33	5.63	3.54	5.39	0.86	0.97	1.00	1.00
9	0.28	0.01	0.03	0.55	0.36	0.14	0.47	0.57
10	1.14	3.63	1.93	10.20	0.64	1.00	0.83	1.00
11	11.33	5.35	25.21	4.92	1.00	0.95	1.00	0.93
12	9.28	2.66	17.88	1.53	1.00	0.81	1.00	0.54

termittency fields are generated (see Fig. 7.12).

Given a threshold value p, each generated Gaussian field (in Fig. 7.12) can be truncated at this threshold to get a binary mask for dry-wet areas (called the thresholded field). By comparing the thresholded field with the LS intermittency values Ii ($i \in [1..4]$), we understand that the generated Gaussian fields need to be adjusted to match the 4 Ii. That means, at each time step and for each homogeneous zone, we need to add a new field (called the interpolated pilot values field) to the free Gaussian field so that the thresholded field of the sum of the two fields has the same intermittency value as the expected Ii.

Before modeling the interpolated pilot values fields, several details must be clarified.

- The interpolated pilot values fields are generated by using block-to-point kriging so the resulting spatio-temporal structure remains the same as the free Gaussian fields. In this case, the block-to-point kriging is used in a three-dimensional context (space and time; later called 3D domain).

- The interpolated pilot values are obtained by using the dichotomy method.

The kriging matrix relevant to the intermittency for block-to-point kriging needs first to be calculated. Below is the case of ordinary kriging with exponential covariance.

Let B_i^k the block of zone k at time step i, B_j^h the block of zone h at time step j, the exponential covariance function from the block B_i^k to the block B_j^h is

$$\mathrm{C}(B_i^k, B_j^h) = \frac{\sum_{x \in B_i^k, y \in B_j^h} \exp(-d(x,y))}{\sum_{x \in B_i^k, y \in B_j^h} \mathbb{1}_{(x,y) \in B_i^k \times B_j^h}} \qquad \text{for} \quad (i,j) \in [1:nt]; (k,h) \in [1:nz] \tag{7.7}$$

where d is the distance function.

Thus, the kriging matrix relevant to the intermittency M is

$$\mathrm{MK} = \left[\begin{array}{ccc|c} & \left(\mathrm{C}(B_i^k, B_j^h)\right)_{(i,j)\in[1..nt];(k,h)\in[1:nz]} & & \begin{matrix}1\\ \vdots \\ 1\end{matrix} \\ \hline 1 & \cdots & 1 & 0 \end{array}\right] \tag{7.8}$$

The dimension of the kriging matrix MK is $(nt \times nz + 1) \times (nt \times nz + 1) = 49 \times 49$.

For each cell x (including cells outside the 4 zones) in the 3D domain, the exponential covariance function from x to the block B_j^h of zone j at time step h is

$$\mathrm{C}(x, B_j^h) = \sum_{y \in B_j^h} \exp(-d(x,y)) \qquad \text{for} \quad j \in [1:nt]; h \in [1:nz] \tag{7.9}$$

Thus, the vector of our target cell x linked to all blocks of the 3D domain is

$$V = \left[\begin{array}{c} \mathrm{C}(x, B_1^1) \\ \vdots \\ \mathrm{C}(x, B_j^h) \\ \vdots \\ \mathrm{C}(x, B_{nt}^{nz}) \\ 1 \end{array}\right] \tag{7.10}$$

where the dimension of V is 49 ($= nt \times nz + 1$).

According to Equation 7.4, the kriging weights of V for point-to-block kriging is

$$\lambda_x = \mathrm{MK}^{-1} \times V \tag{7.11}$$

Let PV a vector of the pilot values, the dimension of the PV is $nt \times nz + 1$ where the first $nt \times nz$ values correspond to the pilot values of the nz zones for all nt time steps and the last one is 1 (non-bias parameter). The estimated value of the target cell x associated with the pilot values vector PV and the kriging weights vector λ_x is

$$x_{\text{pilot values}} = < \lambda_x, PV > \tag{7.12}$$

where $< u, v >$ is the scalar product of the vector u and the vector v.

Thus, a 3D field (called the pilot values fields) corresponding to a given pilot values vector PV can be generated by calculating each cell x in the 3D domain through Equation 7.12.

With the same threshold p, the threshold field of the sum of the free Gaussian field and the pilot values fields present an adjusted intermittency field. Comparing the rainfall intermittency value of the threshold field to Ii in each zone i, we can decide whether the pilot values should be increased or diminished so that the prescribed intermittency value Ii is recovered.

The dichotomy method is used to find the interpolated pilot values.

Since the interpolated pilot values have been found in all zones for all time steps, the interpolated pilot values field for the intermittency field (Fig. 7.13) can be generated.

Figure 7.14 presents the sum of Fig. 7.12 and Fig. 7.13.

The simulated intermittency field (Fig. 7.15) is obtained by a binary mask for Fig. 7.13 with the threshold value p. That means the cell in Fig. 7.15 equals to 1 (i.e. the cell is wet) if the same cell in Fig. 7.13 is greater than p, otherwise the cell in Fig. 7.15 equals to 0 (i.e. the cell is dry).

A very similar procedure is used to generate the simulated non-zero rainfall field.

1. The nt free Gaussian fields having the expected spatio-temporal structure are generated to be used for modeling the non-zero rainfall field (see Fig. 7.16).
2. The kriging matrix relevant to the non-zero rainfall is calculated through 7.7 and 7.8.
3. The interpolated piloted values field (Fig. 7.17) can be generated by using the dichotomy method.
4. Figure 7.18 is the sum of Fig. 7.16 and Fig. 7.17
5. Two parameters (m and σ) of a log-normal distribution are introduced for the non-zero rainfall. The value r of a cell in Fig. 7.18 is transformed to $\exp(m + \sigma \times r)$.
6. The simulated non-zero rainfall field is generated in Fig. 7.19 after the transformation.

The final simulated rainfall (Fig. 7.20) is obtained by the product of the cell of Fig. 7.15 with the cell of Fig. 7.19 in the same location.

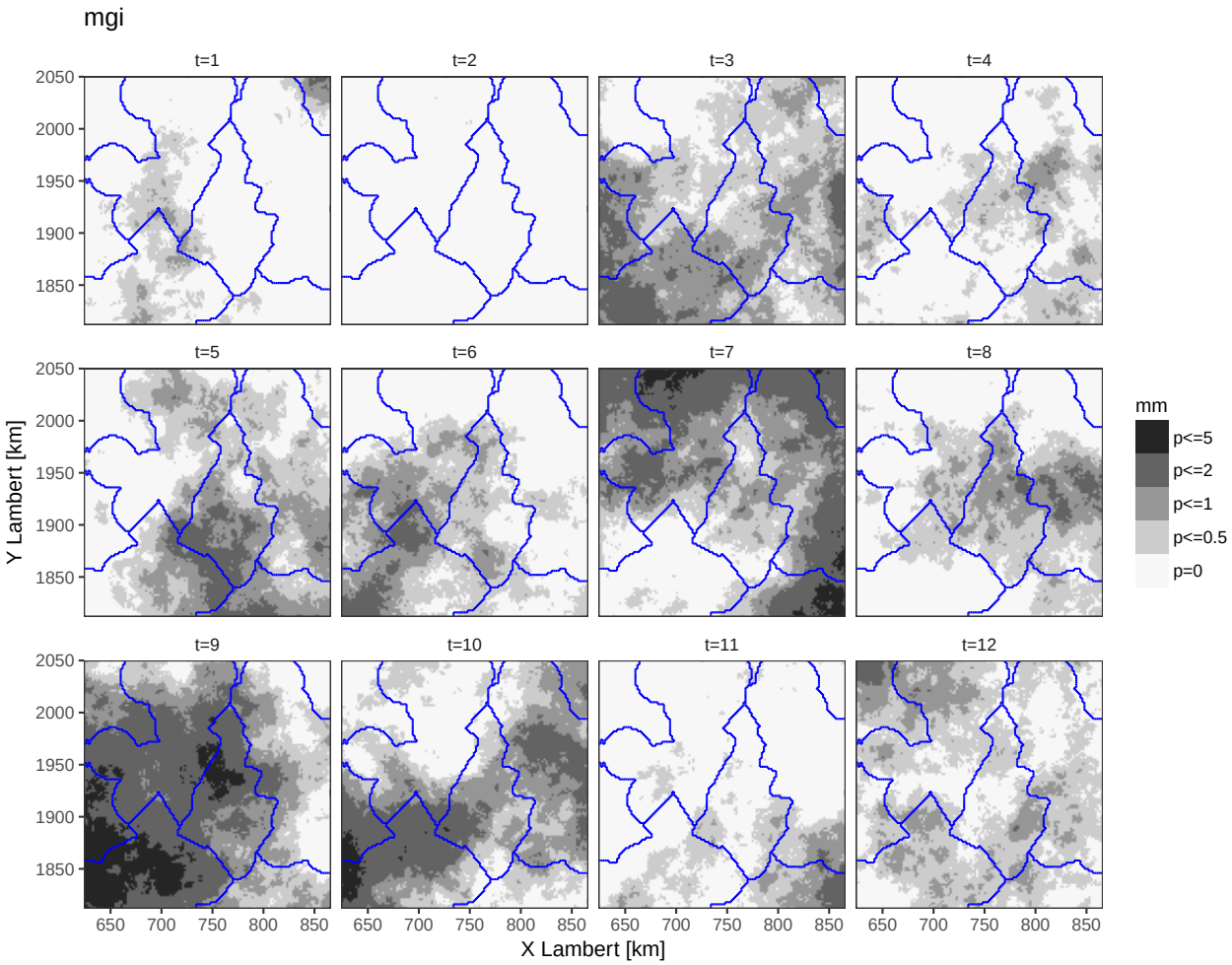


Figure 7.12: The 12 continuous free Gaussian fields, used for intermittency fields.

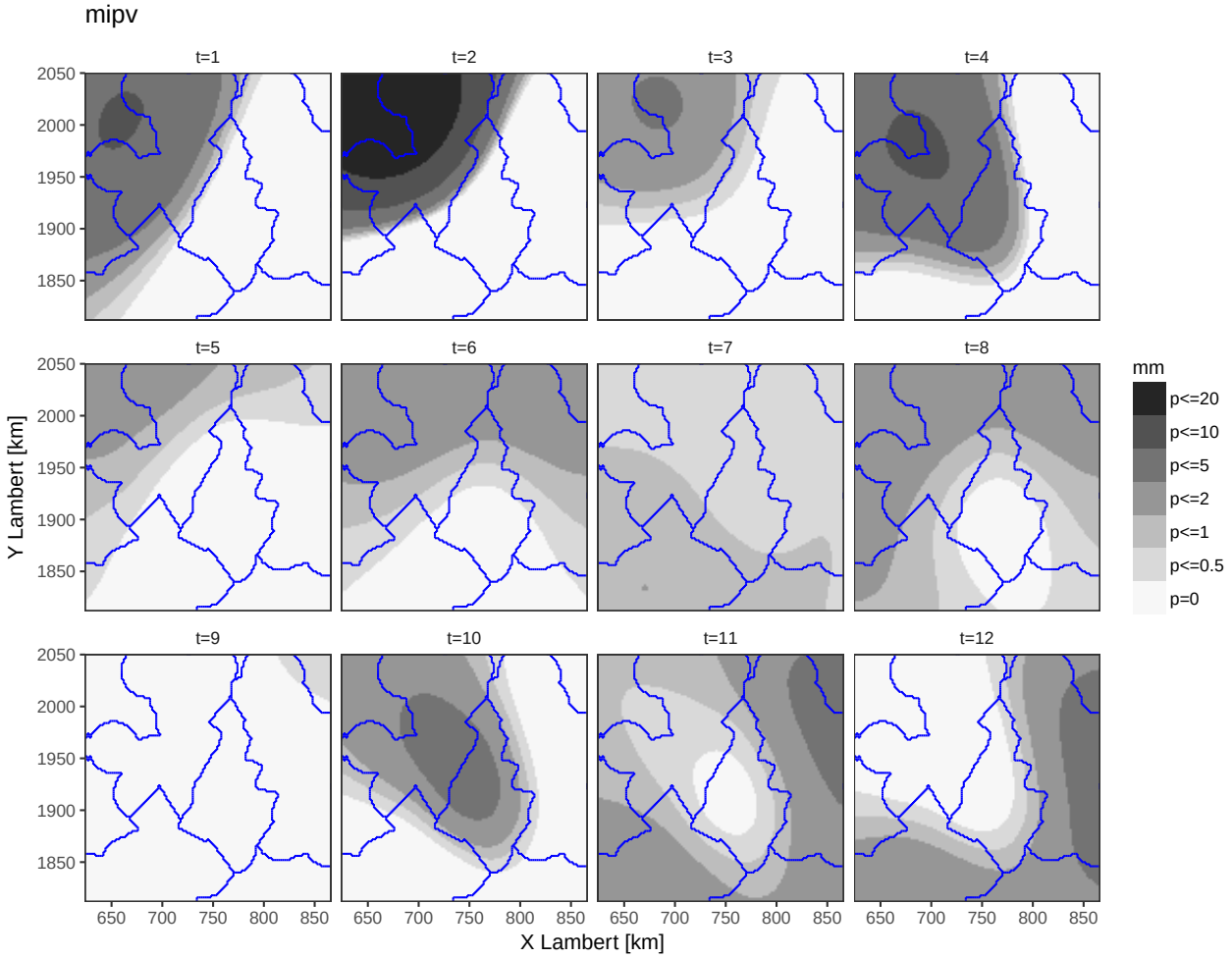


Figure 7.13: The 12 continuous fields of the interpolated pilot values, then added to the Gaussian fields.

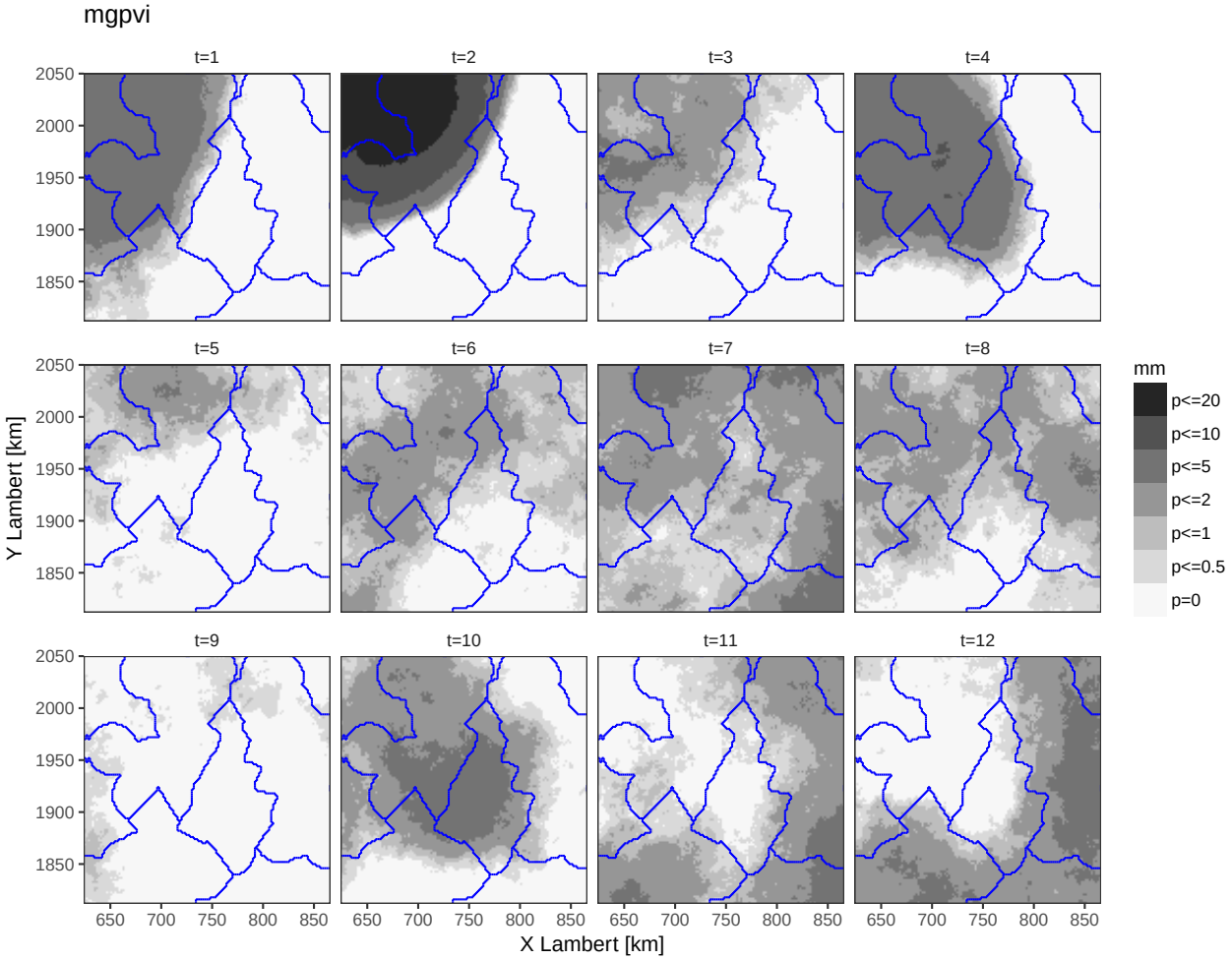


Figure 7.14: The 12 continuous sum of a priori Gaussian plus shift, used for intermittency fields.

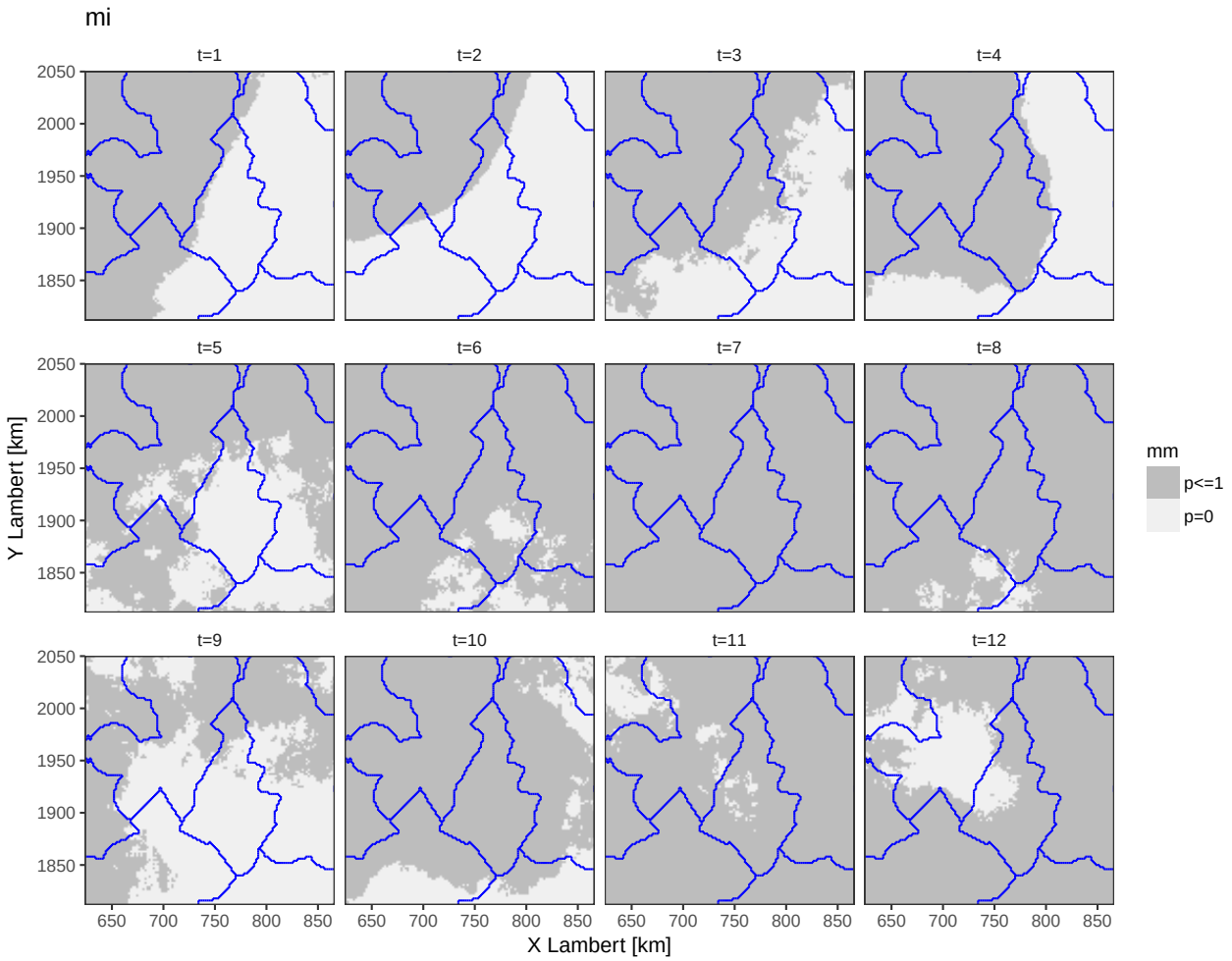


Figure 7.15: The 12 continuous simulated intermittency fields.

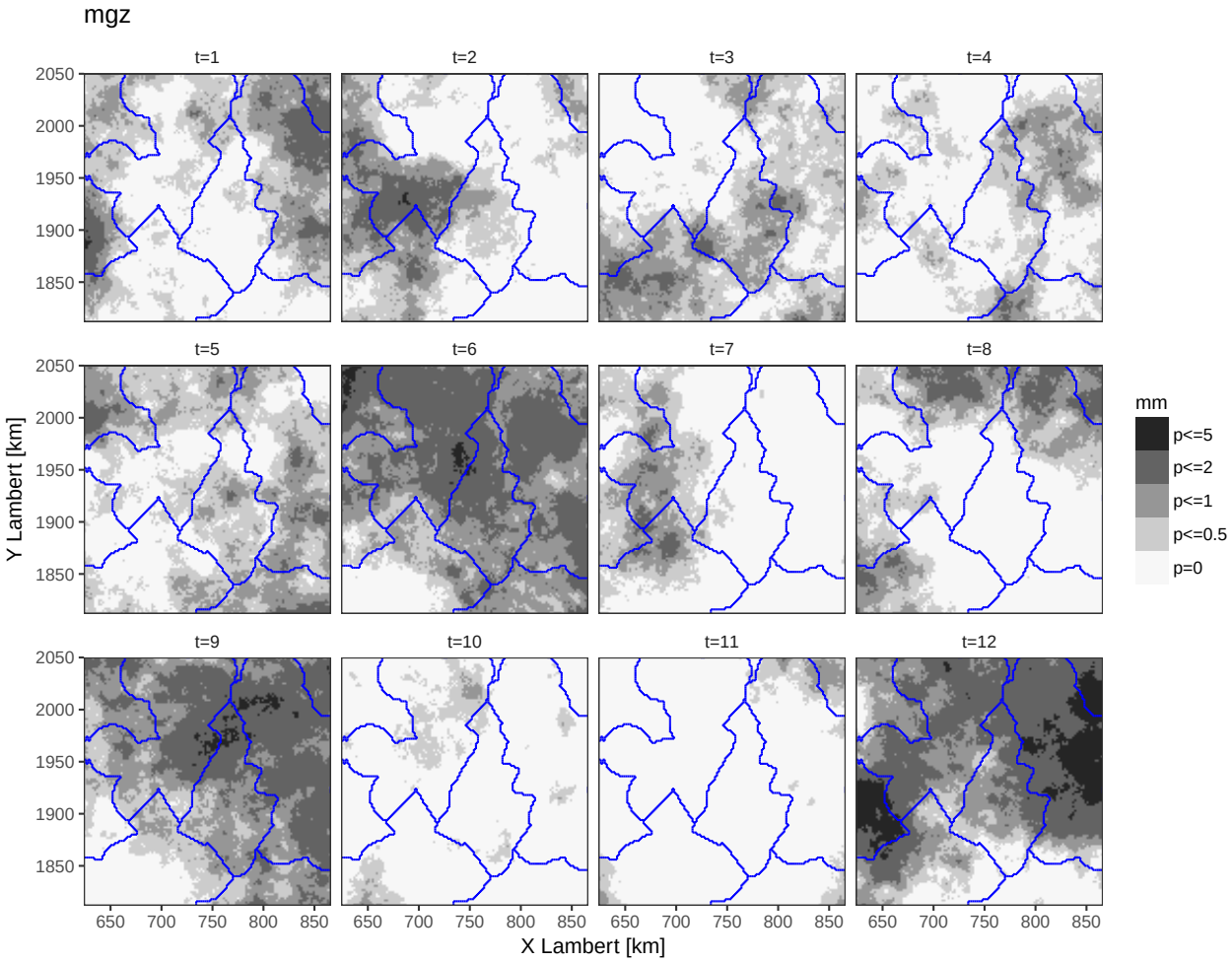


Figure 7.16: The 12 continuous free Gaussian fields, used for non-zero rainfall fields.

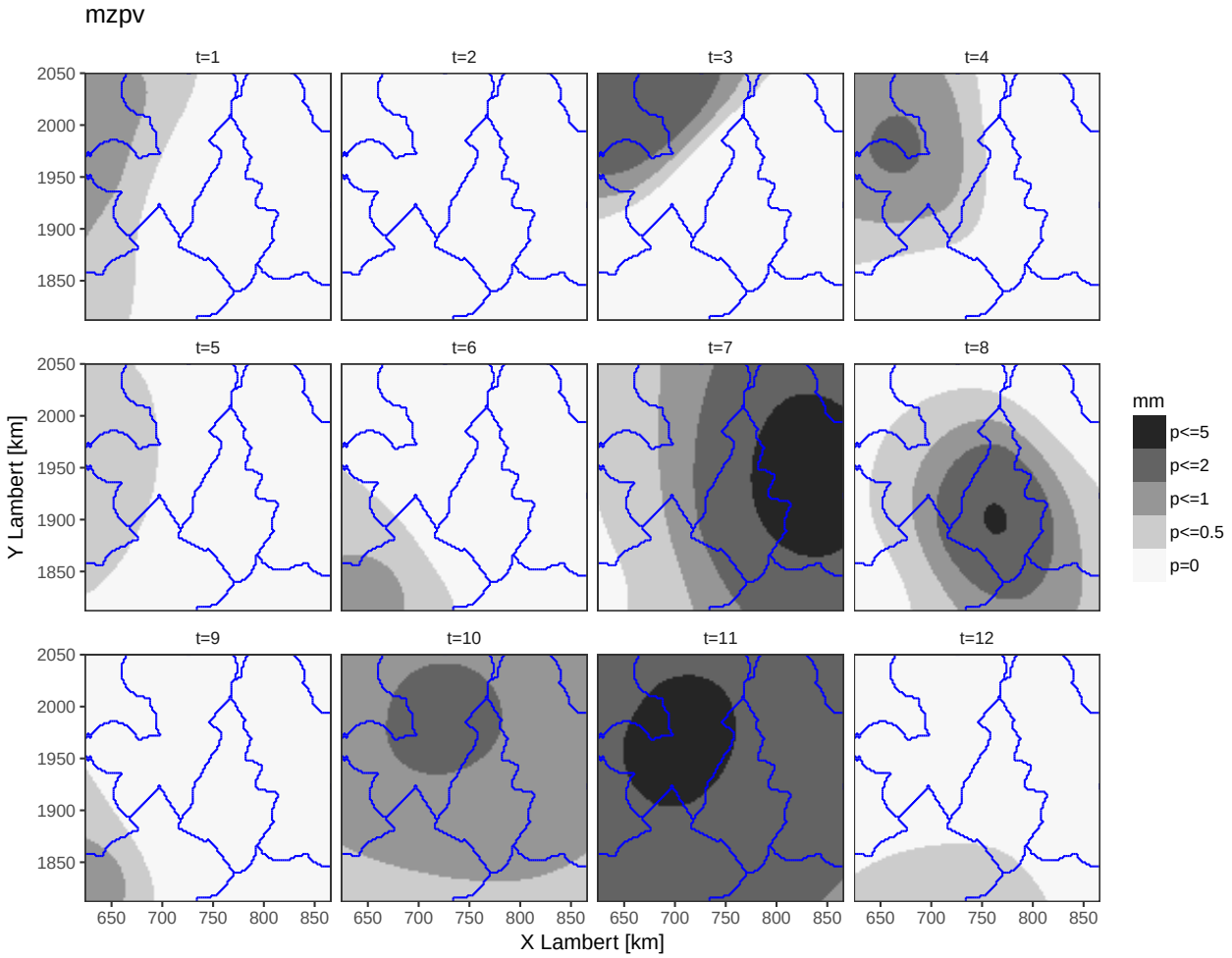


Figure 7.17: The 12 continuous fields of interpolated pilot values, later added to the Gaussian fields.

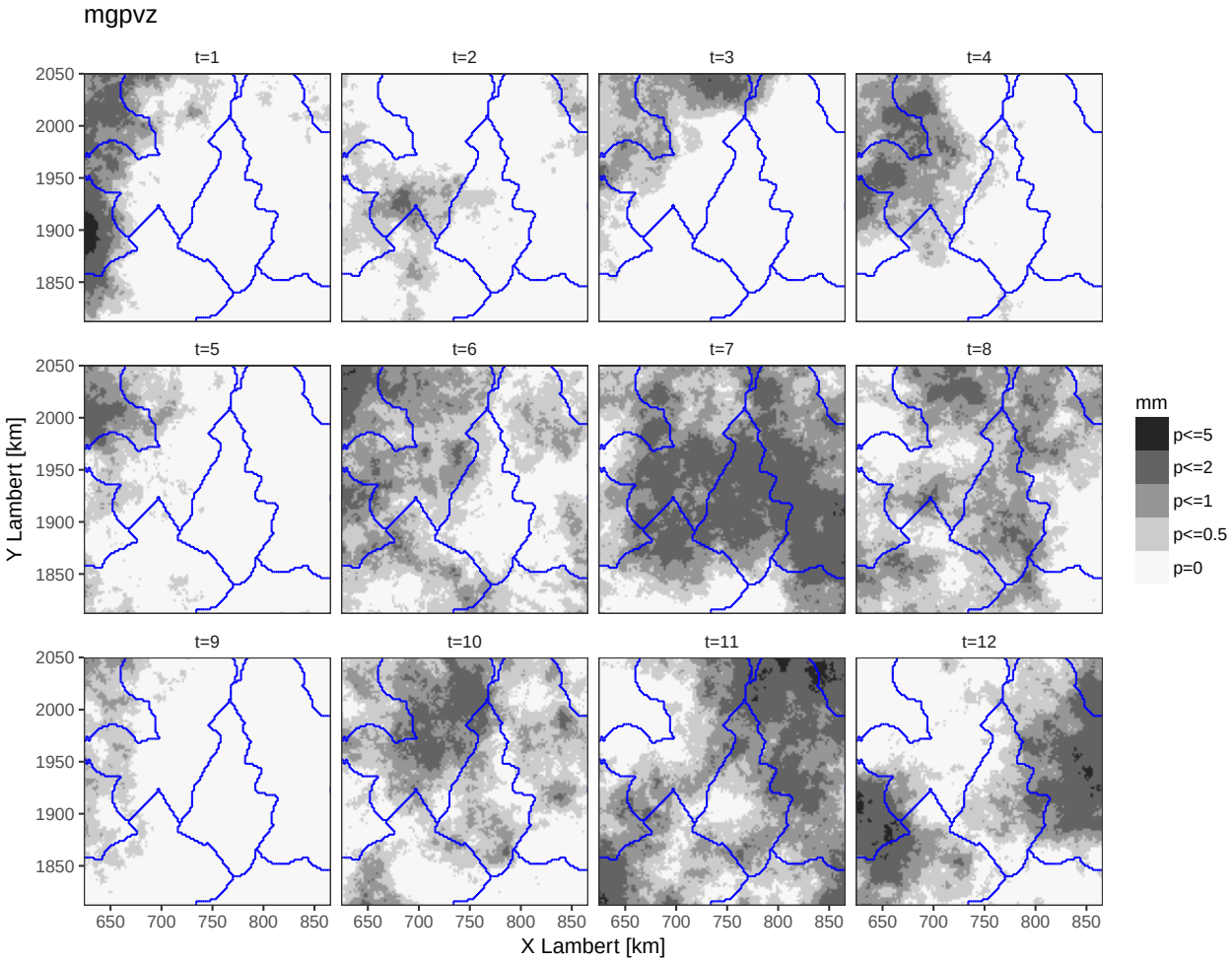


Figure 7.18: The 12 continuous sum of a priori Gaussian plus shift, used for non-zero rainfall fields.

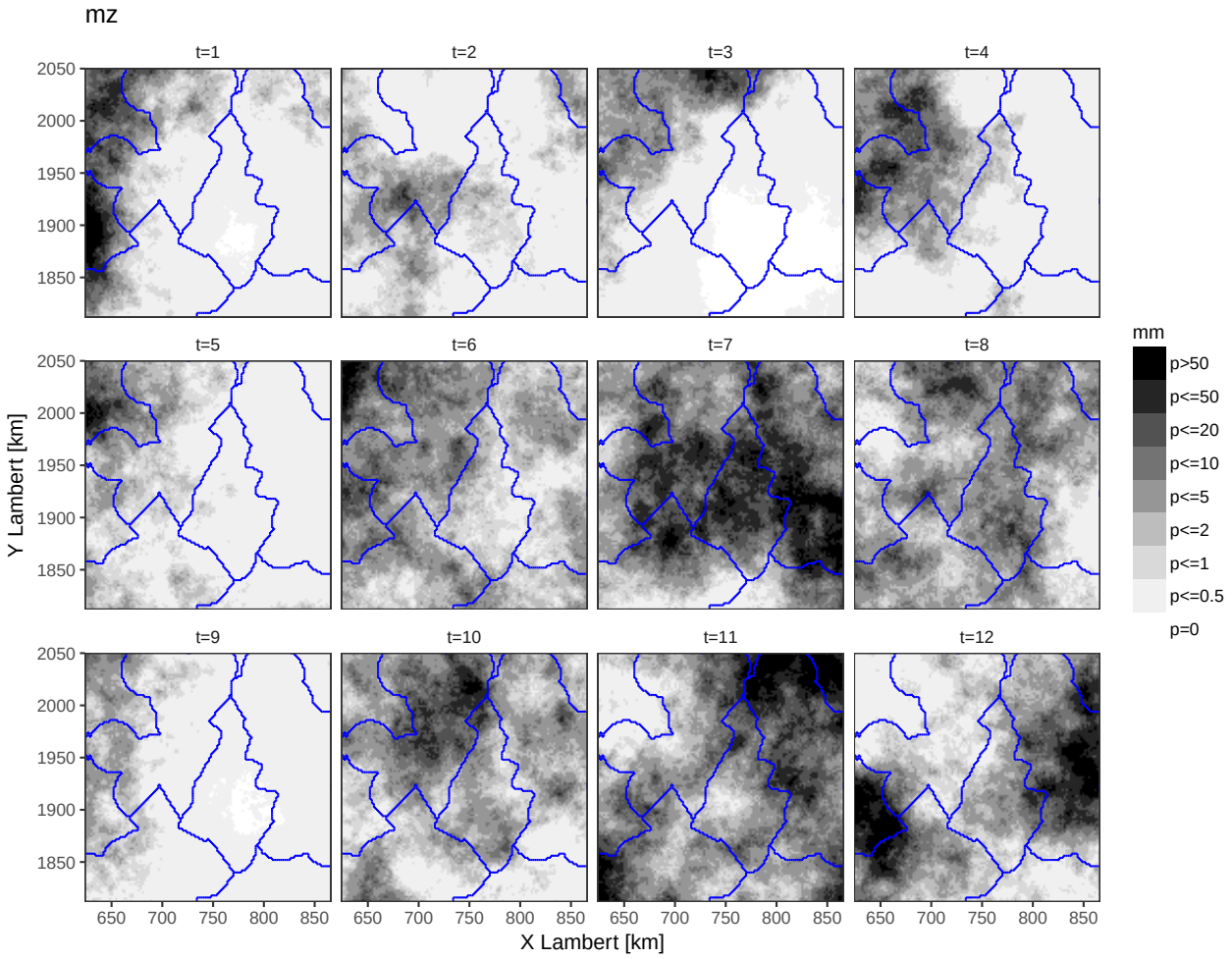


Figure 7.19: The 12 continuous simulated non-zero rainfall fields.

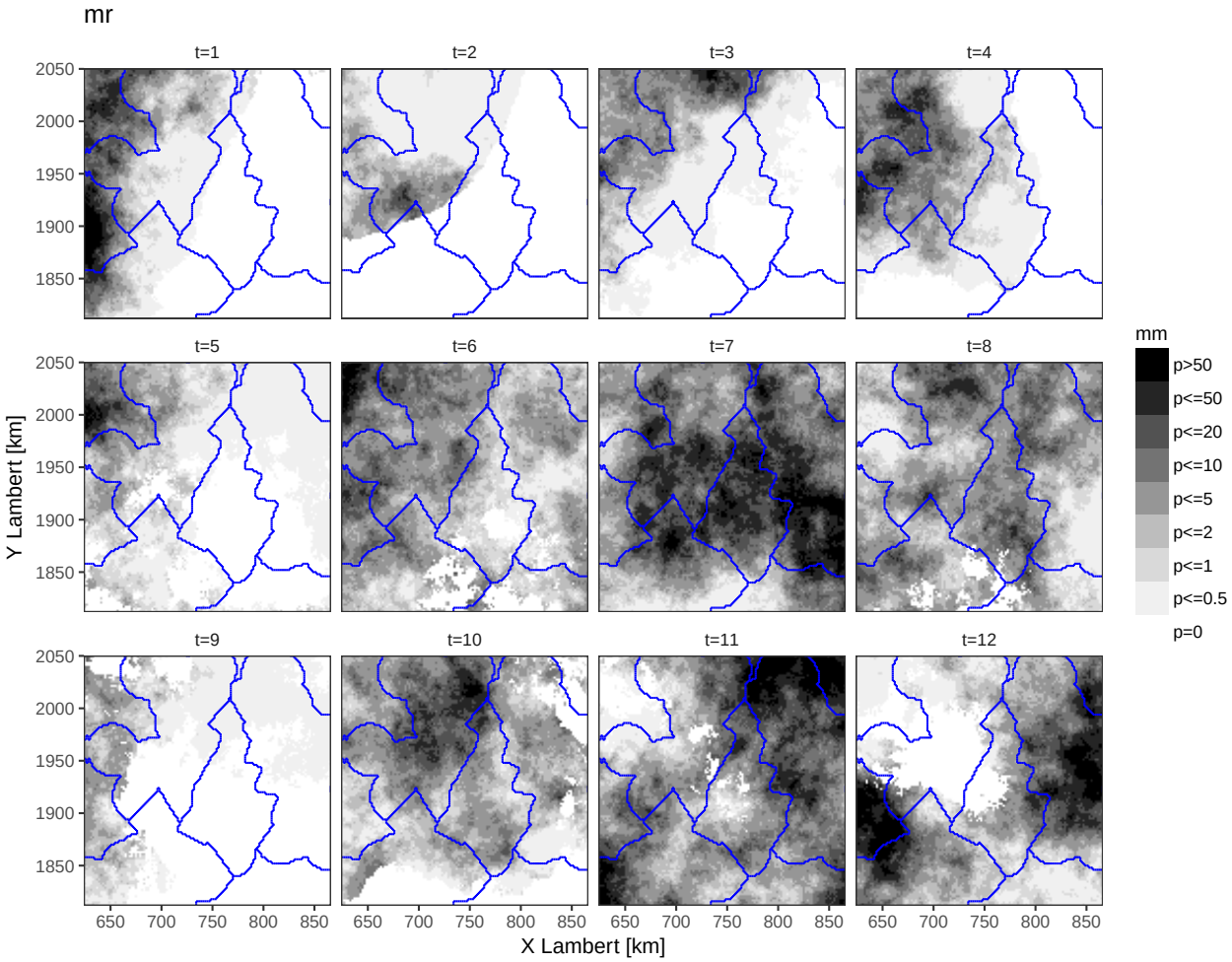


Figure 7.20: Final composite rainfall fields for 12 consecutive days.

7.3.3 Details on the algorithm

1. We assume a stationary covariance structure; the simplest is an exponential decay. If several time steps are jointly disaggregated the suggestion is to use an adimensional equivalent distance

$$d_{\text{eq}} = \sqrt{\left(\frac{\mathrm{d}x}{L}\right)^2 + \left(\frac{\mathrm{d}t}{D}\right)^2} \tag{7.13}$$

   where L and D are correlation parameters tuning variability in space and time.

   In geostatistical terms, a Gaussian field presenting a covariance only dependent on the separation of two points (i.e. a stationnary covariance structure), is called SRF-2 (stationary random function of order 2) [*Chilès and Delfiner*, 2008a; *Christakos*, 2012]. However the technique introduced could be implemented as well in the context of an IRF-0 (intrinsic random function of order 0) [*Chilès and Delfiner*, 2008b] structure - the geostatistical equivalent of a Brownian motion, where stationarity is not on the data but on the increments.

2. The shifts have initial zero-values, they are likely to evolve between a minimum (say -10) and a maximum (say +10).

3. For mathematical consistency, kriging must be done using the same stationarity model and the same structure that has been used for the Gaussian field simulation. This can be shown to give an unbiased sampling of the parental Gaussian field conditionally to the zonal average.

4. Kriging used here is the block-to-point ordinary kriging with unknown mean, a usual and tolerant variant. According to the theoretical point recalled above, a stricter alternative would be to use simple kriging with known (0) mean. These variants may deserve later study.

5. Kriging is an exact estimator and a linear estimator when the data is fixed (the sum of kriging estimates is the kriging of the sum); The control zone (LS cell) average of the kriged shift will by construction respect the given pilot value. So the average over any control zone (LS cell) of the kriged field is the data average plus the shift value for the given control zone (LS cell).

6. The operation, an anamorphosis, turns the gridded field into a field of the quantity of interest to the user. In all cases but for linear transformations, the sum of the transform, however, is the transform of the sum; this is why we can't explicitly state what is the right pilot value to use: the pilot value has to be tuned.

   It is known that anamorphosis induces a decay in correlation, that can be formally described in the case of transforming unconditional Gaussian fields using Hermite decomposition of the anamorphosis function (e.g., *Leblois and Creutin* [2013]); here the phenomenon also exists, but we transform conditioned fields, not free fields; we are

not aware whether a formulation for the covariance decay exists in this case; for now we just do not consider the problem; a possibility later would be to use the general formulation as an approximation.

7. Starting from one given underlying Gaussian field, the zonal average is expected to be in monotonous relation to the shift - an increase of the shift will result in an increase in the target value. So tuning the shifts is an obvious way to drive the simulation to respect the wished LS values.

   The simplest way to tune a value to get a prescribed derived value in a monotonous context is a dichotomy search. Here several shift values must be tuned together. It was found efficient to include all shifts in the same dichotomic loop.

   However, we may expect that if the parameters for the underlying process are inconsistent to the forcing averages, forcing averages may be conflicting. So caution was taken to not miss the solution by a too quick convergence.

   Let us consider the iterative one-step feasible range reduction factor α of the dichotomy, commonly 0.5 (at each step, the legal domain [0-1] is reduced to [0-0.5] or [0.5-1]).

   We choose to take an α reduction of less than 0.5 so that at each step, the legal domain [0-1] is reduced to either [0 to (1-α)] or [α to 1]. As compared to the ordinary dichotomy, this allows the domain [α; (1- α)] around the tested value, to be maintained in the convergence procedure.

   A check was done about smooth convergence: we traced where each final value was located within the limits of the nested intervals out of the dichotomy search, so long these intervals still have a significant width. The guess is that, if there is no conflict between LS values given the assumed small scale variability, the final value should not have been on the limit of the feasible domain while tuning.

   We tested alphas from 0.01 to 0.50 on synthetic examples and found the strategy correct. However, there is a huge difference between NZR (Non-Zero Rainfall) and IND (INDcator function) simulation. [XXX-YES NZR? IND?]

7.3.4 Specific details for rainfall

Rainfall has an enormous atom at 0 that makes possible to consider it a bivariate quantity, an occurrence of rain (0/1), and a quantity of rain (in R+*) that is only observable where rainy. Most authors make the design choice of perfect dependence (rainfall can be simulated using a one step simulation, in thresholded Gaussian style), other prefer independence (occurrence of rain independent of the quantity).

(Certainly a better choice would be to consider both aspects of the rainfall phenomena being linked with a explicit bivariate structure that may depend of the space and time resolution, but this idea is still place for research).

In this contribution, we shall consider both phenomena to be independent, with a benefit of a much simpler solution. As a consequence, the disaggregation algorithm described above will be applied two times.

1. First, we disaggregate the wetness. The user transformation is thresholding to the threshold that corresponds to the local rainfall percentile. If the shifted Gaussian field is above that value, the location will be declared rainy, if it is below the location will be declared dry. Counting the fraction of wet cells over a control zone (LS cell) makes the assessment of the zonal wetness, to be compared with target value.

   A detail is that the number of cells in the control zone (LS cell) makes only discrete values for wetness reachable, and the zonal predicted wetness must be rounded to such a feasible value – excluding zero if we know the total rainfall over the control zone (LS cell) to be non-zero.

   IND simulation is much more sensitive, because of the threshold effect making a limited set of fractional intermittency feasible. A low alpha value (around 0.02) and numerous iterations make the job. A termination peculiarity is that if the intermittency is OK for a given LS cell, no reduction is taken. When this is true for all LS domains, the tuning is complete.

   The result of this step is a wet/dry pattern over the area that fulfills the LS wetness constraint up to the resolution of the grid; the scale size of the pattern is dominated by the structure parameters of the underlying Gaussian if the LS simulation is really loose versus the size of the grid cell size.

2. Second, we disaggregate the rain total. The user transformation is there just the anamorphosis between the underlying Gaussian and the local rainfall percentile (QQ-plot). Tuning the shifts relies on comparing the zonal average (over all cells, including dry ones) rainy cells to the LS rainfall total – this is why the wetness pattern was decided first. The dichotomy can be stopped when the marginal changes in user space are below any practical meaning (say 0.001 mm). For NZR simulation, a α value of 0.30 to 0.35 was found able to maintaining a satisfactory convergence rate and complain only in the case of ill balanced LS data (suggesting to review the choice of parameters).

We see the disaggregated field keeps a noticeable conditional variability and this makes sense if one considers disaggregation as a scale-bridging technique.

Elaborated in this PhD , this geostatistical disaggregation technique could be useable in many disaggregation issues. It has been suggested to our colleagues at Sintef[1] (Norway) for use in their own SWG prototype, and we intent a paper with our colleague Sara MARTINO (SINTEF & NTNU).

[1] `https://www.sintef.no/en/`: last check on 2017/11/27

CHAPTER 8

CONCLUSION FOR HETEROGENEITY PROBLEM

In this part, two different approaches are proposed to solve the problem of the simulation of the heterogeneous rainfall field.

Chapter 4 reviews some existing rainfall simulation methods which deal with rainfall simulation at more than one location, including kriging models, conditional models and copula based multisite models. Several references on multisite models [e.g., *Bárdossy and Pegram*, 2009; *Chen et al.*, 2015; *Evin et al.*, 2017] demonstrates the potential of the copula technique, which will eventually be used in this PhD work.

In this PhD work, we aim at generating spatial-temporal rainfall simulations, not just multisite rainfall models. We first tried to use SAMPO [*Lepioufle et al.*, 2012; *Leblois and Creutin*, 2013], the local rainfall simulation tool, which relies on the idea that a spatio-temporal rainfall field can be considered as an instance of a homogeneous rainfall type. Ongoing work at Irstea suggests that local rainfall can be summarized as a alternation of rainfall fields of several rainfall types. So in Chapter 5, we proposed two models to coordinate parallel calendars of qualitative rainfall types, which are a parametric model based on coupled hidden Markov model (CHMM), and a non-parametric model based on a resampling technique. Parametric models are sometimes difficult to model because of the complexity of the observations. The CHMM is modeled by using hidden Markov models. In hidden Markov models, the statistical properties of the qualitative (types) calendars are captured with a transition matrix and a emission matrix. These matrices are estimated with the well known Baum-Welch algorithm (see Appendix A). This established technique is a very important part of how the model is fitted to the observations. But a problem appears in the generation of simulations from the estimated transition matrix and emission matrix. The successive random draws in a emission matrix generate correct frequencies but incorrect distribution of length of stay in the observed states. This effect is worse in coupled hierarchical Markov model where the simulation chain includes several conditional draws for any single time step. We defined a re-organization method to adjust the length of stay distributions; the statistical results show a clear but limited improvement. As an other way to make the compromise, we introduce a non-parametric model based on the resampling

technique. Due to the non-parametric nature of this resampling, such a model can not be used to generate all different kinds of scenarios, especially could not be run in a climate change context, and simulation demonstrated a slightly eroded interannual variability as each simulated year is by design a mixture from the sampled years. But the resampling technique is easily applied in our case and the statistical diagnosis of the simulations are generally satisfying.

As a main result, both models were formally capable to generate the calendars to drive SAMPO simulations over a large heterogeneous domain. However, both models also have difficulties to fully capture the spatial correlations between the different zones; the root of the evil comes from the fact that the achieved/preserved correlation is only the part of the correlation conveyed by the rainfall types system between successive distinct rainfall types in one zone and between two zones. This comes evident in the statistical analysis of both models, and we eventually came to recognize that the real problem is associated with the very concept of homogeneous rainfall type and the strict delineation of homogeneous rainfall zones.

Given the above diagnostic, a completely different approach was designed, where the areal rainfall totals and intermittency want to be kept in their original distribution, jointly modeled using the copula technique. The idea of partitioning a heterogeneous rainfall field into several rainfall zones still remains, but we target the average precipitation and the rainfall intermittency values for each homogeneous rainfall zone, not just a qualitative type of homogeneous rainfall. As it is well known, atoms are a problem in copula modeling, and specially the atom at 0 is a well known problem in rainfall modeling. We introduced a detour through data mining the historical atmospheric context as documented by ERA-Interim. This provided a rainfall index and intermittency index, having no atom and suitable for use as surrogates to the targeted quantities in the copula approach. Results are quantile-quantile transferred to the targeted average precipitation and wetness values, that finally have similar statistical properties as the observations. Final step is to use these large-scale simulations as input to a disaggregation model, to generate fine resolution rainfall fields. The disaggregation is handled using a geostatistically-based conditional simulation technique combining block to point kriging and optimization. The technique first generates the rainfall intermittency fields, then generate the non-zero rainfall fields taking into account the dry/wet pattern previously simulated. The simulation of 12 continuous days offers a realistic spatial connection between the zones. This simulation technique still needs to be diagnosed statistically, but shows great potential.

In the next part, a multivariate model is proposed to deal with several hydrometeorological variables in time.

Part III
Multivariate modeling for hydrological inputs

CHAPTER 9

DRIVING VARIABLES FOR WATER RESOURCES MODELING: COPULA BASED MULTIVARIATE APPROACH

A multivariate model generating time-series simulations for the hydro-meteorological variables is proposed in this chapter.

This multivariate model is similar to the copula based parametric model introduced for rainfall in Section 7.2 of Part II, and uses several techniques that appears to be closely related:

- The Gaussian copula, used to capture the multivariate joint distribution;
- The multi-variate auto-regression, used to explicit temporal correlations in the time-series of the variables;
- The sequential kriging perspective, used to generate complete simulations given possibly already known variables.

The target, here, is the joint simulation of five different hydro-meteorological variables (precipitation, temperature, solar radiation, water vapor pressure and wind speed). The inter-variability among these hydro-meteorological variables can be complex, and each variable has its own, possibly very different distribution. In addition, we intend to make precipitation the primary variable, possibly simulated by a different model component, so that other considered variables must be simulated conditionally to the precipitation. The sequential kriging technique is the natural technique to make such conditional simulations.

The chapter is presented in the form of a scientific paper which is hopefully to be submitted in the future.

Abstract [XXX-YES the abstract should end with one or two sentences presenting the main conclusions of the paper]

For water resources modeling, several meteorological inputs such as precipitation, temperature, solar radiation, wind speed and water vapor pressure are required in hydrological models. In statistical approach, multivariate models are to provide such simulations. In this paper, Gaussian copula based multivariate approach is proposed to simulate sequentially these five variables. The multivariate auto-regression is introduced to explicitly use the temporal correlation for the variables. The simulations of multivariable can be performed in any desired order by using sequential kriging technique. The studied area is in Cévennes-Vivarais area, a mountainous region in the south of France well documented in local rainfall data-bases and as any other place by meteorological reanalysis. The seasonality is considered in this paper by standard 3-months seasons. The diagnosis of the statistical analysis for this copula based multivariate approach are generally quite promising.

9.1 Introduction

Water resources, essential for lifebeing, are widely shared by different users and economical sectors, such as agriculture, industry, household, recreational and environmental activities. Often, sharing raises conflicts between specific users and very generally between the sustainable environment and economical purposes like hydro-power, irrigation for agriculture and domestic and industry water supply, where total flows are diverted without releasing water for ecological conservation [*Tessema*, 2011]. Such conflicts need a balanced assessment of water resources and this remains a difficult task given their high variability in space, time, and also the presence of water in different physical compartments (i.e. atmosphere for rain water; soil and hydrographic network for surface water; aquifers for groundwater). *Vörösmarty et al.* [2000] mentioned that the future adequacy of freshwater resources is also difficult to assess, owing to a complex and rapidly changing geography of water use and regulating schemes.

When evaluating water supply and resource systems, the challenge is to build an approach that can incorporate all the knowledge available for planners and water managers into a quantitative framework that can be used to simulate and predict the outcome of alternative approaches and policies. To do so, hydrological models are thus required [e.g., *Maidment et al.*, 1993; *Legesse et al.*, 2003; *Musy et al.*, 2014; *Fatichi et al.*, 2016].

Water resources and hydrological modeling projects typically involve simulating systems made up of many parts, strongly interrelated, and in some cases, poorly characterized. *Fatichi et al.* [2016] currently reviewed the applications, the challenges, and the future trends in distributed process-based models in hydrology. In most situations, the hydrological system is driven by physical variables (i.e. precipitation, potential evapotranspiration, etc.) and still involves uncertain processes and parameters. Recent articles [e.g., *Devia et al.*, 2015; *Sood and Smakhtin*, 2015] reviewed several types of hydrological models highlighting the important numbers of inputs required for handling simulations as accurate as possible. These inputs concern topography, soil characteristics, vegetation, land surface

classification, and meteorological forcings. Runoff observations are used for evaluation.

This paper aims at providing relevant meterological forcings for the use of hydrological models. To close the water balance at the catchment scale, it is necessary to have available precipitation and potential evapotranspiration, both forcings associated with meteorological processes. This paper takes part of a long term scientific dynamics where a local stochastic rainfall generator, named SAMPO [*Lepioufle et al.*, 2012; *Leblois and Creutin*, 2013], has been developed and provides simulated stochastic rainfall fields for a given homogeneous region. Evaporation is under control of potential evapotranspiration ; the stochastic simulation of the potential evapotranspiration is more complicated due to the large number of relevant parameters involved [e.g., *Penman*, 1948; *Zotarelli et al.*, 2010]. To obtain the evapotranspiration, mainly four meteorological variables are needed: temperature, solar radiation, wind speed, and water vapor pressure. These other variables, by the way, are not only useful for potential evapotranspiration estimation, but also rule some water demands uses like irrigation and connected topics like assessment of renewable energy potential. Since the target is to feed an hydrological model, all the meteorological forcings must be available at the same time and space scales. It is thus necessary to develop a multi-variate model to reach this objective.

Georgakakos and Kavvas [1987] give another interesting point of view about the interest of other meteorological variables to be combined with precipitation in stochastic simulations. They reported that doing so, it improves the capability of the models to capture the structure of the precipitation and it also facilitates the assessment of the effect of climate change on the precipitation structure. Including some physical drivers as covariates may help a multivariate simulation to hold over a variety of contexts.

The principal concern of multivariate statistics is to formally describe the relationships between variables and their relevance to the problem being studied. There are two major problems associated with the multivariate modeling. The first one concerns the representation of the observed data distributions as described by the multivariate model. Meteorological or climatological data (precipitation, wind speed, cloud cover, relative humidity) often turn out to be non-Gaussian, having bounded or skewed distributions [*Schoelzel and Friederichs*, 2008]. So, attention needs to be payed on individual distributions and it may be necessary to discuss the dependence of most of these meterological variables to figure out their relationship.

The second difficulty deals with the identification of the multivariate joint distribution. In hydrology, the development of the multivariate model began with the seminal paper of *Richardson* [1981] where the author modified the rainfall generator described in *Katz* [1977] introducing three other meteorological variables, the minimum and maximum temperatures and the solar radiation given at the daily time scale. This was the first stochastic weather generator to simultaneously consider four meterological variables. Many later works developed the same idea, and already [*Wilks and Wilby*, 1999; *Srikanthan and Mcmahon*, 2001] reviewed a long list of multivariate approaches.

Several methods are now available. Multivariate distributions had some success, as an example multivariate Gaussian used in mixture models [*Marin et al.*, 2005; *McLachlan and*

Peel, 2004]. Such multivariate distributions rely on one parametric description to accommodate all relationship among all variables. However, not all distributions are suitable to multi-dimension extensions, and variables of interest often follow different distributions. This why the so called copulas [e.g., *Nelsen*, 2007] became more and more popular. Copulas relies on the selection of an appropriate model for the dependence among the variables, represented by the copula that is understood independently from the choice of the marginal distributions, what is very convenient in many applications. Today, copulas have now a strong record of applications, among others in finance and climatology. They are also used in hydrology [e.g., *Genest and Favre*, 2007; *Bárdossy and Li*, 2008; *Schoelzel and Friederichs*, 2008; *Erhardt et al.*, 2015; *Evin et al.*, 2017].

Within this general context, this paper aims at developing a multivariate models using the copula technique. In our case the precipitation is the priority because of the complexity of its spatio-temporal variability. Assuming a specialized component (either SAMPO or any other simulator) already generates reasonable precipitation simulations, there is a need for a capacity of partial simulation model driven by already available simulated precipitation. So we may have to use the copula in conditional mode. Interestingly, if using the Gaussian copula, we find a practical situation that is strictly equivalent to sequential kriging or auto-regression techniques. Sequential kriging is a well-know method in geostatistics to operate conditional simulations [*Zimmerman et al.*, 1998; *Bayraktar and Turalioglu*, 2005].

The multivariate model is used in a context of water resources in the French Mediterranean region, the Cévennes-Vivarais region, where many observations [*Boudevillain et al.*, 2011; *Braud et al.*, 2016] and modeling [*Nuissier et al.*, 2008; *Godart et al.*, 2011; *Adamovic et al.*, 2016] works have been already realized. These previous works pointed out the need for long-term stochastic meteorological forcings to assess the evolution of the water resources in a region recognized [*Intergovernmental Panel on Climate Change*, 2014] as a hot-spot of global change.

The article is organized as follows. The overall methodology including the copula method seen in its connection to sequential kriging and auto-regression techniques is first described in Section 9.2. Section 9.3 details its application to daily hydrological multivariate model in the Cévennes region. Section 9.4 compares the statistics of observations and simulations and synthesizes the results. Section 9.5 offers conclusions and perspectives.

9.2 Methods

The three techniques (copula, auto-regressive process, kriging) used in this copula based multivariate model were respectively described in the previous Section 7.2.1, 7.2.2 and 7.2.3 (of this PhD). In this section, we present more details of the mathematical formulations about how Gaussian copula, auto-regression and sequential kriging work together in our context.

Equivalence between Gaussian copula, auto-regressive process and simple kriging

Let $\{V_1, V_2, \ldots, V_n\}$ n be time-series variables, all of them supposedly already normalized. In this part, we concentrate in $AR(1)$ model for auto-regression, but there could be a higher lag. So the concerned variables will extend to $\{V_1, V_2, \ldots, V_n, V_1^1, V_2^1, \ldots, V_n^1\}$ (also noted as $\{X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}\}$) with

$$V_i^1(t) = V_i(t+1) \quad \text{for } i \in [1:n]. \tag{9.1}$$

The associated covariance matrix of the variables $\{V_1, V_2, \ldots, V_n, V_1^1, V_2^1, \ldots, V_n^1\}$ is

$$M_{\text{cov}} = \left[\begin{array}{c|c} \text{cov}(V_i, V_j) & \text{cov}(V_i, V_j^1) \\ \hline \text{cov}(V_i^1, V_j) & \text{cov}(V_i^1, V_j^1) \end{array}\right] = (a_{i,j}).$$

As a symmetric real matrix, the covariance matrix M_{cov} admits the unique Cholesky decomposition that has the form

$$M_{\text{cov}} = M_{\text{Chol}} M_{\text{Chol}}^{\intercal}$$

which M_{Chol} is a lower triangular matrix with having only real entry and positive values (usually strictly positive) on the diagonal

$$M_{\text{Chol}} = \begin{bmatrix} b_{1,1} & 0 & \cdots & 0 \\ b_{2,1} & b_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{2n,1} & b_{2n,2} & \cdots & b_{2n,2n} \end{bmatrix}.$$

Let us decide we want to make the simulation sequentially. This means that a value of X_i is to be obtained conditionally on previously simulated values such as $\{X_1, \ldots, X_{i-1}\}$. We also use an AR process to formulate X_i.

$$X_i = \sum_{j=1}^{i-1} \alpha_{i,j} X_j + \epsilon_i. \tag{9.2}$$

So we can write the vector $X = [X_1, \ldots, X_{2n}]^{\intercal}$ under the form,

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{2n} \end{bmatrix} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,2n} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{2n,1} & \alpha_{2n,2} & \cdots & \alpha_{2n,2n} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{2n} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{2n} \end{bmatrix}.$$

In fact, the vector of innovations $[\epsilon_1, \ldots, \epsilon_{2n}]^{\intercal}$ equals to the diagonal of M_{Chol},

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{2n} \end{bmatrix} = \begin{bmatrix} b_{1,1} \\ b_{2,2} \\ \vdots \\ b_{2n,2n} \end{bmatrix}. \tag{9.3}$$

With the structure of Equation 9.2, $[\alpha_{i,j}]$ is a lower triangular matrix. Therefore, each variable of $\{X_{n+1}, \ldots, X_{2n}\}$ $(= \{V_1^1, \ldots, V_n^1\})$ can be expressed as the linear combination of variables $\{X_1, \ldots, X_n\}$ $(= \{V_1, \ldots, V_n\})$ only. This means that a multivariate auto-regressive model can be built consistent to the Gaussian copula based model. A multivariate auto-regressive model of order p (noted $MAR(p)$) predicts the next value in a d-dimensional time series, y_n as a linear combination of the m previous vector values

$$y_n = \sum_{i=1}^{m} y_{n-i} A(i) + e_n \tag{9.4}$$

where $y_n = [y_n(1), y_n(2), \ldots, y_n(d)]$ is the nth sample of a d-dimensional time series, each $A(i)$ is a d-by-d matrix of coefficients and $e_n = [e_n(1), e_n(2), \ldots, e_n(d)]$ is additive Gaussian noise with zero mean and covariance.

Since we have concentrated in $AR(1)$ model, we should build a $MAR(1)$ for n variables $\{V_1, \ldots, V_n\}$.

$$\begin{bmatrix} V_1^1 \\ V_2^1 \\ \vdots \\ V_n^1 \end{bmatrix} = A \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} + e \tag{9.5}$$

where A is a n-by-n matrix and $e = [e_1, e_2, \ldots, e_n]$ is a n-dimension Gaussian noise.

Actually, the parameters A and e could be easily found by using associated Cholesky matrix M_{Chol}. A series of sub-Cholesky-matrix $\{M_1, M_2, \ldots, M_{2n}\}$ of M_{Chol} is introduced

$$M_i = \left[\begin{array}{c|c} M_{\text{Chol}}^i & 0_{i\times(2n-i)} \\ \hline 0_{(2n-i)\times i} & I_{2n-i} \end{array}\right]$$

where

$$M_{\text{Chol}}^i = \begin{bmatrix} b_{1,1} & 0 & \cdots & 0 \\ b_{2,1} & b_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ b_{i,1} & b_{i,2} & \cdots & b_{i,i} \end{bmatrix}.$$

The inverse of M_i has the same form of M_i such as:

$$M_i^{-1} = \left[\begin{array}{c|c} (M_{\text{Chol}}^i)^{-1} & 0_{i\times(2n-i)} \\ \hline 0_{(2n-i)\times i} & I_{2n-i} \end{array}\right]$$

where I_d stands for identity matrix of dimensions $d \times d$ and $0_{p\times q}$ stands for zero matrix of dimensions $p \times q$.

As we mentioned before, we use kriging technique for conditional simulation. When we multiply M_{2n} and M_n^{-1}, we can obtain

$$M_{2n}M_n^{-1} = M_{\text{Chol}} \left[\begin{array}{c|c} (M_{\text{Chol}}^n)^{-1} & 0_{n\times n} \\ \hline 0_{n\times n} & I_n \end{array}\right]$$

$$= \left[\begin{array}{c|c} M_{\text{Chol}}^n & 0_{n\times n} \\ \hline B_1 & B_2 \end{array}\right] \left[\begin{array}{c|c} (M_{\text{Chol}}^n)^{-1} & 0_{n\times n} \\ \hline 0_{n\times n} & I_n \end{array}\right]$$

where

$$B_1 = \begin{bmatrix} b_{n+1,1} & b_{n+1,2} & \cdots & b_{n+1,n} \\ b_{n+2,1} & b_{n+2,2} & \cdots & b_{n+2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{2n,1} & b_{2n,2} & \cdots & b_{2n,n} \end{bmatrix}, \; B_2 = \begin{bmatrix} b_{n+1,n+1} & 0 & \cdots & 0 \\ b_{n+2,n+1} & b_{n+2,n+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{2n,n+1} & b_{2n,n+2} & \cdots & b_{2n,2n} \end{bmatrix}. \quad (9.6)$$

Thus, we can obtain

$$M_{2n}M_n^{-1} = \left[\begin{array}{c|c} M^n_{\text{Chol}}(M^n_{\text{Chol}})^{-1} = I_n & 0_{n\times n} \\ \hline B_1(M^n_{\text{Chol}})^{-1} & B_2 \end{array}\right]$$

A and e can be expressed as follows.

$$A = B_1(M^n_{\text{Chol}})^{-1}. \quad (9.7)$$

and

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = B_2 \begin{bmatrix} b_{1,1} \\ b_{2,2} \\ \vdots \\ b_{n,n} \end{bmatrix}. \quad (9.8)$$

With Equation 9.7 and 9.8, we build the relationship between the parameters of multivariate auto-regressive model of order 1 and the Cholesky matrix for the variable $\{V_1, \ldots, V_n\}$.

Equation 7.4 provides the kriging weights for simple kriging which are the parameters for the simulation. In fact, the kriging weights can be found in the series of sub-matrix of Cholesky. For example, the kriging weights vector $[w_1^i, \ldots, w_{i-1}^i]^\intercal$ for X_i which is simulated conditionally on $\{X_1, \ldots, X_{i-1}\}$ can be calculated as

$$\begin{bmatrix} w_1^i \\ \vdots \\ w_{i-1}^i \end{bmatrix} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,i-1} \\ \vdots & \ddots & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,i-1} \end{bmatrix}^{-1} \begin{bmatrix} a_{i,1} \\ \vdots \\ a_{i,i} \end{bmatrix}.$$

However, we can also find the same kriging weights vector $[w_1^i, \ldots, w_{i-1}^i]^\intercal$ by multiply-

ing M_i and M_{i-1}^{-1}.

$$M_i M_{i-1}^{-1} = \left[\begin{array}{c|c} M_{\text{Chol}}^{i} & 0_{i\times(2n-i)} \\ \hline 0_{(2n-i)\times i} & I_{2n-i} \end{array}\right] \left[\begin{array}{c|c} (M_{\text{Chol}}^{i-1})^{-1} & 0_{(i-1)\times(2n-i+1)} \\ \hline 0_{(2n-i+1)\times(i-1)} & I_{2n-i+1} \end{array}\right]$$

$$= \left[\begin{array}{ccc|c|c} & M_{\text{Chol}}^{i-1} & & 0 & \\ & & & \vdots & 0_{i\times(2n-i)} \\ & & & 0 & \\ \hline b_{i,1} & \cdots & b_{i,i-1} & b_{i,i} & \\ \hline \multicolumn{3}{c|}{0_{(2n-i)\times i}} & & I_{2n-i} \end{array}\right] \left[\begin{array}{ccc|c|c} & (M_{\text{Chol}}^{i-1})^{-1} & & 0 & \\ & & & \vdots & 0_{i\times(2n-i)} \\ & & & 0 & \\ \hline 0 & \cdots & 0 & 1 & \\ \hline \multicolumn{3}{c|}{0_{(2n-i)\times i}} & & I_{2n-i} \end{array}\right]$$

$$= \left[\begin{array}{c|c|c} I_{i-1} & 0_{(i-1)\times 1} & 0_{i\times(2n-i)} \\ \hline [b_{i,1}\cdots b_{i,i-1}]\times(M_{\text{Chol}}^{i-1})^{-1} & b_{i,i} & \\ \hline \multicolumn{2}{c|}{0_{(2n-i)\times i}} & I_{2n-i} \end{array}\right]$$

Thus, we have

$$[w_1^i \cdots w_{i-1}^i] = [b_{i,1}\cdots b_{i,i-1}] \times (M_{\text{Chol}}^{i-1})^{-1}. \tag{9.9}$$

Equation 9.9 shows the relationship between the kriging weights and the Cholesky matrix.

Our approach is to use both auto-regressive process and kriging technique in Gaussian copula framework to model a multivariate situation. The auto-regressive process makes sure that the model has certain temporal correlation for each variable. The kriging technique has been used to simulate sequentially on previous variables.

In Gaussian copula framework, the Cholesky matrix associated with the covariance matrix is vital. Our approach is similar to model a multivariate auto-regressive model (Equation 9.5). Both weight matrix A and white noisy vector e are accessible by using M_{Chol} through Equations 9.7 and 9.8.

On the other hand, we need kriging weights and innovation vector to generate sequential conditional simulations. The same conclusion as above, the Cholesky matrix are used to obtain kriging weights in each simulation step with Equation 9.9 and the innovation vector with Equation 9.3.

Let us recall the matrix called Schur complement that is interesting to consider in the context of this kind of conditional simulations.

Suppose P, R, S and Q are respectively $p \times p$, $p \times q$, $q \times p$ and $q \times q$ matrices. In addition, Q is considered as invertible. Let

$$M = \left[\begin{array}{c|c} P & R \\ \hline S & Q \end{array}\right]$$

where M is a $(p+q)\times(p+q)$ matrix.

Then the Schur complement of the block Q of matrix M, denoted as M/Q, is the $p \times p$ matrix.

$$M/Q := P - RQ^{-1}S.$$

In the applications to probability theory and statistics, suppose the random vectors Y and Z belong to $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, and the vector (Y, Z) in $\mathbb{R}^{p+q}$ has a multivariate normal distribution whose covariance is the symmetric positive-definite matrix

$$\Sigma = \begin{bmatrix} P & R \\ R^\intercal & Q \end{bmatrix}$$

where $P \in \mathbb{R}^{p\times p}$ is the covariance matrix of Y, $Q \in \mathbb{R}^{q\times q}$ is the covariance matrix of Z and $R \in \mathbb{R}^{p\times q}$ is the covariance matrix between Y and Z.

Then the conditional covariance of Z given Y is the Schur complement of P in Σ:

$$\mathrm{Cov}(Z|Y) = Q - RP^{-1}R^\intercal \tag{9.10}$$

In our context,

$$\begin{aligned} p &= n; \\ q &= m; \\ Y &= [V_1, \ldots, V_n]; \\ Z &= [V_1^1, \ldots, V_n^1]. \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathrm{Cov}(Y, Z) &= M_{\mathrm{cov}} \\ &= \left[\begin{array}{c|c} \mathrm{cov}(V_i, V_j) & \mathrm{cov}(V_i, V_j^1) \\ \hline \mathrm{cov}(V_i^1, V_j) & \mathrm{cov}(V_i^1, V_j^1) \end{array}\right] \end{aligned}$$

where $P = [\mathrm{cov}(V_i, V_j)]_{1\le i,j\le n}$, $Q = [\mathrm{cov}(V_i^1, V_j^1)]_{1\le i,j\le n}$ and $R = [\mathrm{cov}(V_i, V_j^1)]_{1\le i,j\le n}$.

Because of Equation 9.10, we obtain the conditional covariance of $\{V_1^1, \ldots, V_n^1\}$ given $\{V_1, \ldots, V_n\}$ as the Schur complement of P in M_{cov}

$$\begin{aligned} \mathrm{Cov}(V_1^1, \ldots, V_n^1|V_1, \ldots, V_n) &= Q - R^\intercal P^{-1} R \\ &= [\mathrm{cov}(V_i^1, V_j^1)] - [\mathrm{cov}(V_i^1, V_j)][\mathrm{cov}(V_i, V_j)]^{-1}[\mathrm{cov}(V_i, V_j^1)]. \end{aligned}$$

Thus, we can also find the Schur complement of $P = [\mathrm{cov}(V_i, V_j)]_{1\le i,j\le n}$ with B_2 in Equation 9.6:

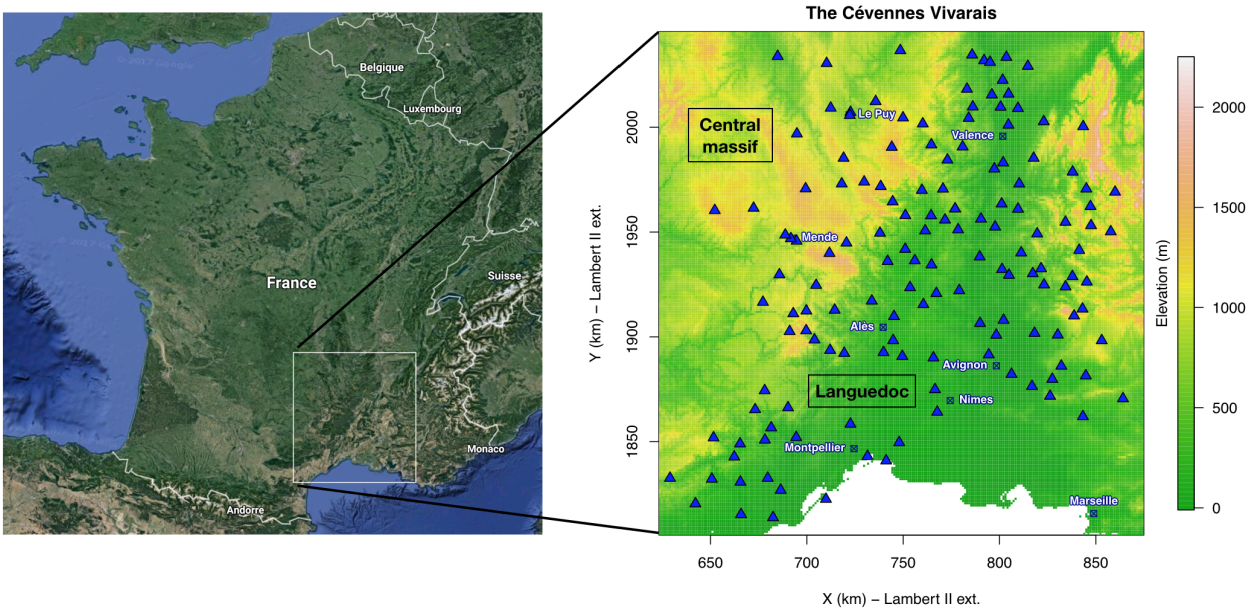
$$\mathrm{Cov}(V_1^1, \ldots, V_n^1|V_1, \ldots, V_n) = B_2 B_2^\intercal.$$

All the above can be summarized in saying that sequential kriging, auto-regression and the Gaussian dependence part of the Gaussian copula techniques are different aspects of the same system of dependence. So one can express things resorting to one or the other aspect according to convenience, and links are explicit to switch from one perspective to another.

9.3 A multivariate model for hydrological purposes

9.3.1 Studied area and data

The studied area is located in the Cévennes which are a mountain range in south-central France (Fig. 9.1).



(a) (b)

Figure 9.1: (a) Location of the Cévennes. (b) 146 daily rain gauge stations available over years 1989-2013.

The Cévennes form a mountainous region at the interface of the Massif Central and the plains of Languedoc. The Cévennes are subjected to Mediterranean influences and its situation as first heighs facing south makes the area sensitive to the so-called *épisode cévenol* meteorological phenomenon, where convective thunderstorms discharge considerable amounts of water in few hours. These events occur at fall when south wind blows moist lukewarm air from the Mediterranean, while upper air is colder, yielding a convective context [see *Tudurí and Ramis*, 1997].

Figure 9.1 shows different topological features in Cévennes area. By applying k-means clustering method on correlation matrix of daily precipitation data, the 146 rainfall gauge stations can be partitioned into 4 clusters, as illustrated in Fig. 9.2.

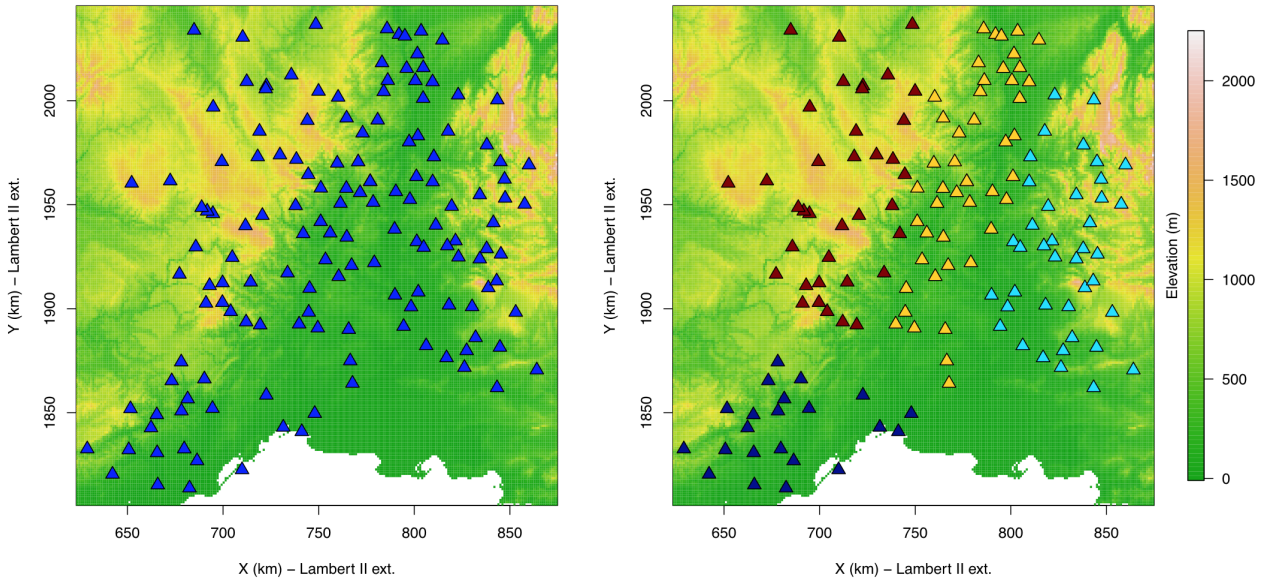


Figure 9.2: (a) The 146 daily rainfall gauge stations in Cévennes Vivarais within the 1989-2013 period. (b) The 146 gauges stations are partitioned into 4 clusters by using the k-means clustering. The clusters are located in the Mediterranean area (in dark blue), the Hill area (in cyan), the Piedmont area (in yellow) and the Mountain ares (in red).

Table 9.1 gives more details about the partition.

Table 9.1: Partition of the study area.

zone	Conventional name	Geographic feature	Location	Number of stations
1	Gard and Hérault	Mediterranean	South West	22
2	Drôme	Hills	North East	37
3	Ardéche	Piemont	Central North	47
4	Haute-Loire and Lozère	Mountain	North West	40

As already mentioned, a multivariate model aims at providing simulation of several variables relevant to hydrologic modeling, used to improve our understanding and our capacity of prediction and management of water resources.

Starting from the beginning, the water balance at the catchment scale may represent the simplest description of the flow of water in and out of a system. Its general equation is:

$$P = R + E + \Delta S. \tag{9.11}$$

where P is precipitation, E is evapotranspiration, R is runoff or streamflow and ΔS is the change in storage (in soil or the bedrock / ground water). E is a response to an atmospheric potential evaporation (PET) that summarizes, from a physically based point of view, available energy. A typical estimation of PET originates in the Penman formula that indicates most relevant atmospheric factors. To serve relevant inputs to a hydrologic model, one finally comes to five main near-surface meteorological variables driving the water balance: Precipitation, Temperature, Solar radiation, Water vapor pressure and Wind speed.

The objective of the multivariate model is thus to make these five variables available to the simulations.

Rain gauge observation precipitation data is extracted from the OHM-CV (Observatoire Hydrométéorologique Méditerranéen Cévennes-Vivarais) database [*Boudevillain et al.*, 2011]. Cévennes-Vivarais region covers an area of 160 km $\times$ 210 km. In this case, the average of daily precipitation data of 22 gauge stations in the Mediterranean area (see Table 9.1) from 1989 to 2013 is considered as precipitation data for modeling. As seen in Section 7.1 and Section 7.2.4, the precipitation index is used as the precipitation variable.

Temperature, solar radiation, water vapor pressure and wind speed data are extracted from the ERA-Interim database[1] [*Dee et al.*, 2011]. ERA-Interim is a global atmospheric reanalysis starting in 1979, continuously updated. The temporal resolution is a 6-hours time step and the spatial resolution is 0.75°. Figure 9.3 presents the location of the Mediterranean area where the extracted temperature, solar radiation, water vapor pressure and wind speed data are joined to the precipitation data as inputs of the multivariate modeling.

Table 9.2 gives a short sample of the type of data used in the multivariate model. The 6-hours ERA-Interim data are averaged to obtain a daily observation.

[1] `http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/`: last check on 2017/11/23

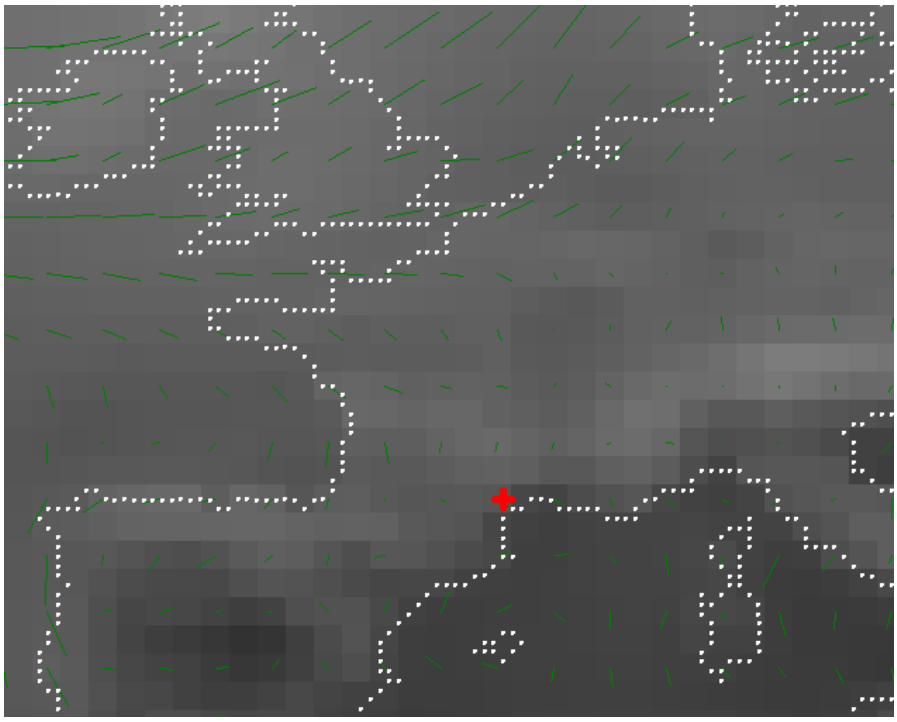


Figure 9.3: Location of the ERA-Interim cell used for providing atmospheric data for zone 1. The cell center is located at (43°30'N, 3°00'E) and indicated by a red cross. The allusive background is a ERA-Interim temperature field, with spatial resolution 0.75°. Green lines are surface wind vectors.

Table 9.2: Short sample of the daily average data used in the multivariate model.

date	Wind speed (m/s)	Solar radiation (W/m²)	Temperature (°C)	Water Vapor pressure (hPa)	Precipitation index
1989/01/01	7.69	363.07	5.40	6.81	0.08
1989/01/02	13.51	369.32	7.72	7.49	0.29
1989/01/03	10.85	363.15	4.99	5.72	0.18
1989/01/04	10.26	360.50	3.92	5.21	0.06
1989/01/05	8.87	367.86	6.22	6.44	0.00
1989/01/06	3.32	375.45	6.81	7.31	-0.02
1989/01/07	1.09	390.14	6.81	8.89	-0.03
.....	.....	.....	.....	.....	.....
2013/12/31	4.32	379.52	8.07	8.04	0.06

9.3.2 Modeling

This section presents the different steps of the modeling process.

In Fig. 9.4, the daily average observations of the 4 variables extracted from the ERA-Interim database (i.e. temperature, solar radiation, water vapor pressure,wind speed) as well as the precipitation index are plotted for the 1989-2013 period. As expected, three among these 5 variables, temperature, water vapor pressure and solar radiation, present a strong seasonality.

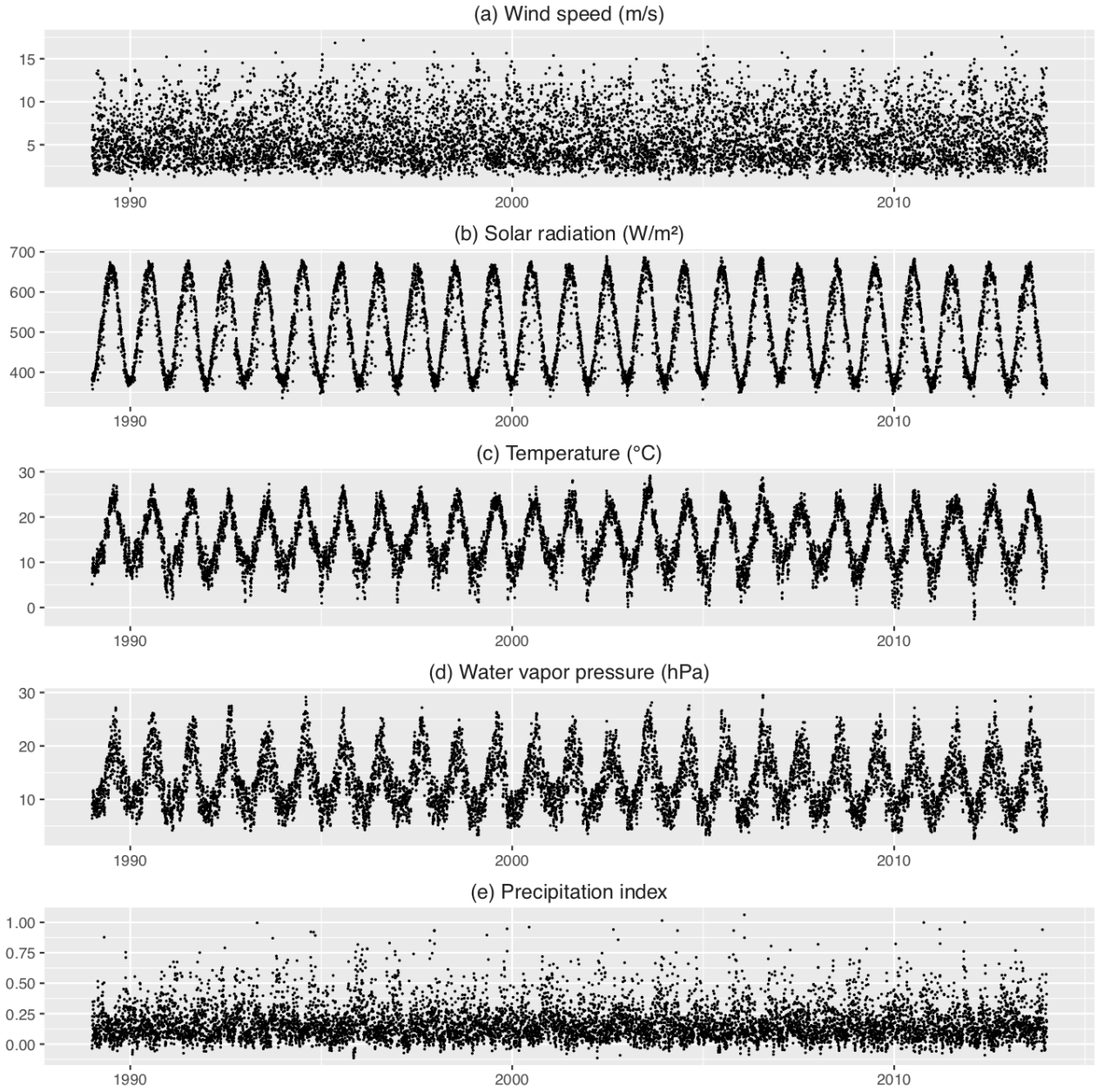


Figure 9.4: The time-series of daily average observation of 5 variables from 1989 to 2013.

To accomodate seasonal variability, the modeling process was repeated for each of the 4 seasons commonly used in climatology : Spring runs from March to May, Summer from Jun to August, Autumn from September to November, and Winter runs from December to

February. The modeling process is thereafter only illustrated for the Summer.

Figures 9.5 - 9.9 present the time series of the five variables and their corresponding fitted distributions.

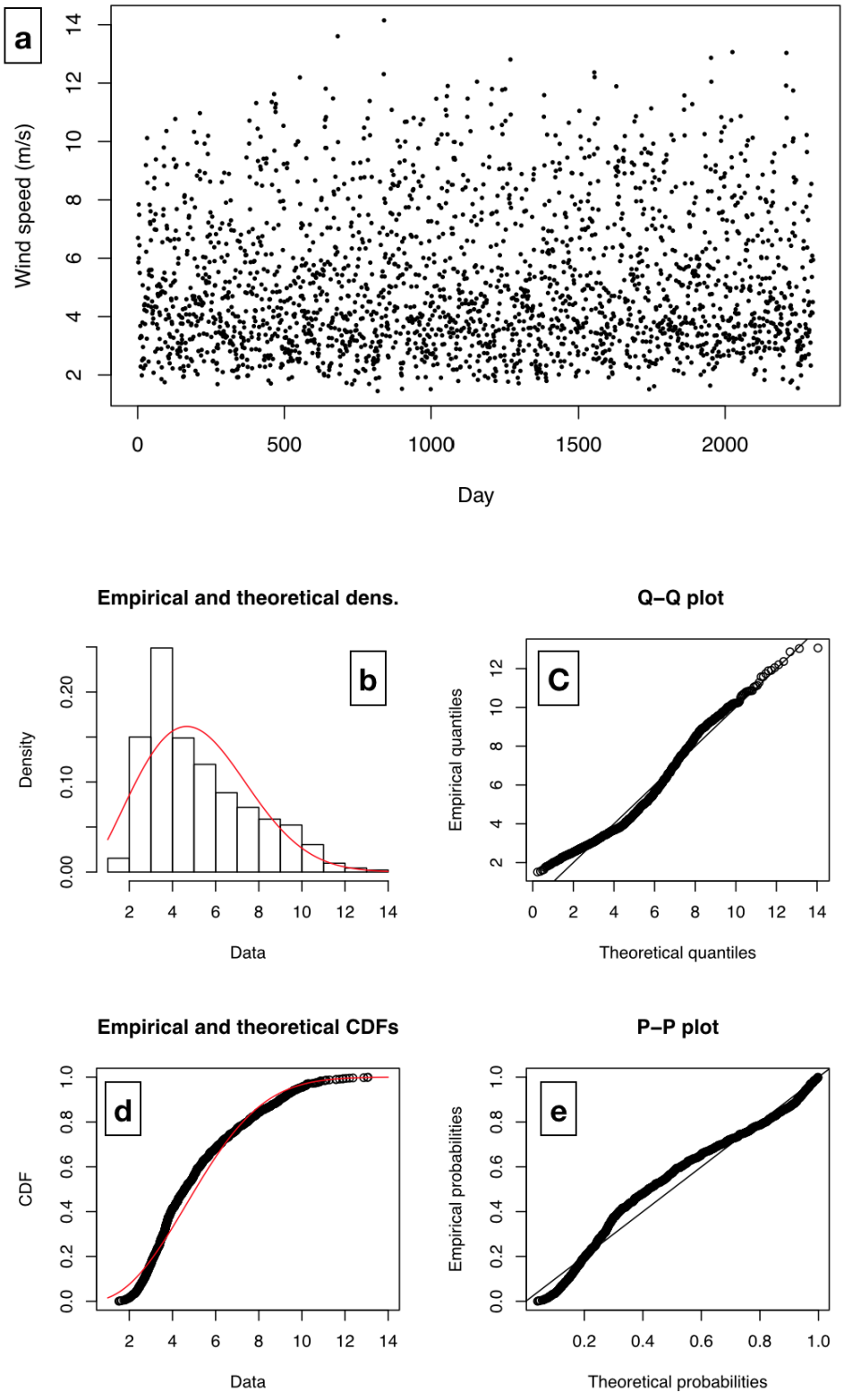


Figure 9.5: (a) the time series of average daily wind speed (m/s) in summer periods from 1989 to 2013. (b) the histogram of data and the fitted Weibull distribution (red line) with parameters (shape, scale) = (2.336, 5.922). (c) the Q-Q (quantile-quantile) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis). (d) the comparison of cumulative distribution function between the data (in black) and theoretical Weibull distribution (in red). (e) the P-P (probability–probability) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis).

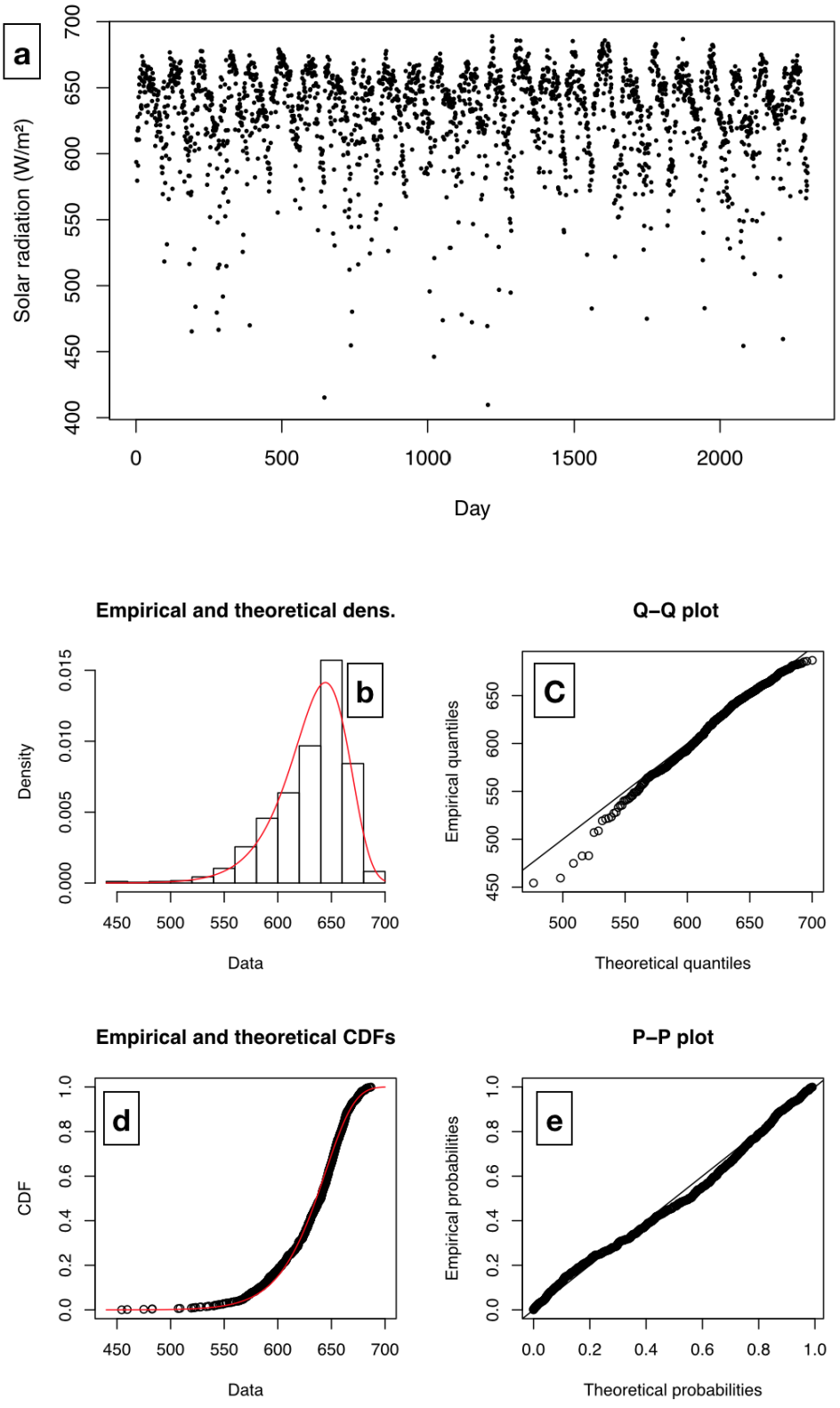


Figure 9.6: (a) the time series of average daily solar radiation (W/m^2) in summer periods from 1989 to 2013. (b) the histogram of data and the fitted Weibull distribution with parameters (shape, scale) = (24.78, 645.4). (c) the Q-Q (quantile-quantile) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis). (d) the comparison of cumulative distribution function between the data (in black) and theoretical Weibull distribution (in red). (e) the P-P (probability–probability) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis).

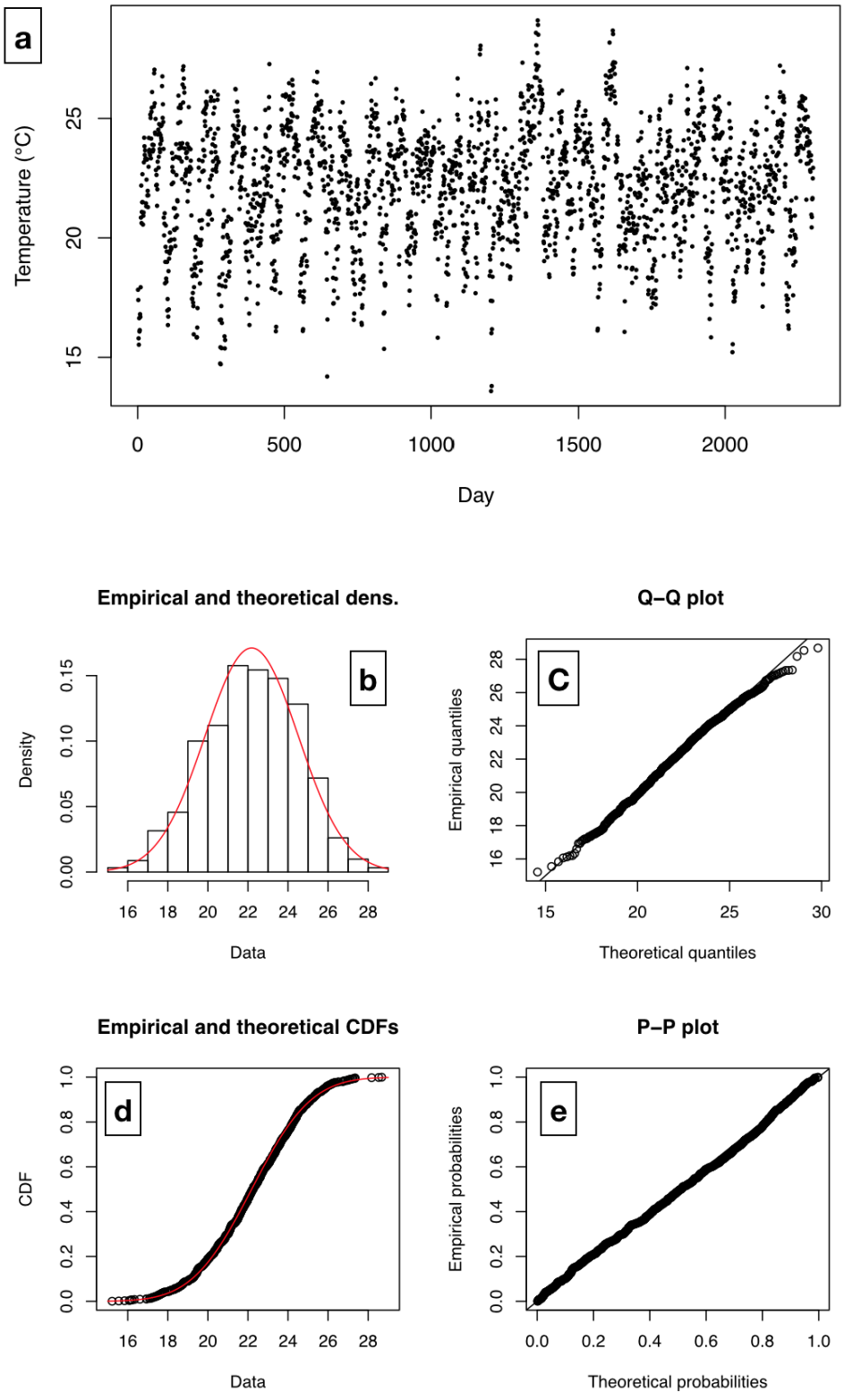


Figure 9.7: (a) the time series of average daily temperature (Celsius) in summer periods from 1989 to 2013. (b) the histogram of data and the fitted Normal distribution with parameters (mean, sd) = (22.18, 2.332). (c) the Q-Q (quantile-quantile) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis). (d) the comparison of cumulative distribution function between the data (in black) and theoretical Normal distribution (in red). (e) the P-P (probability–probability) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis).

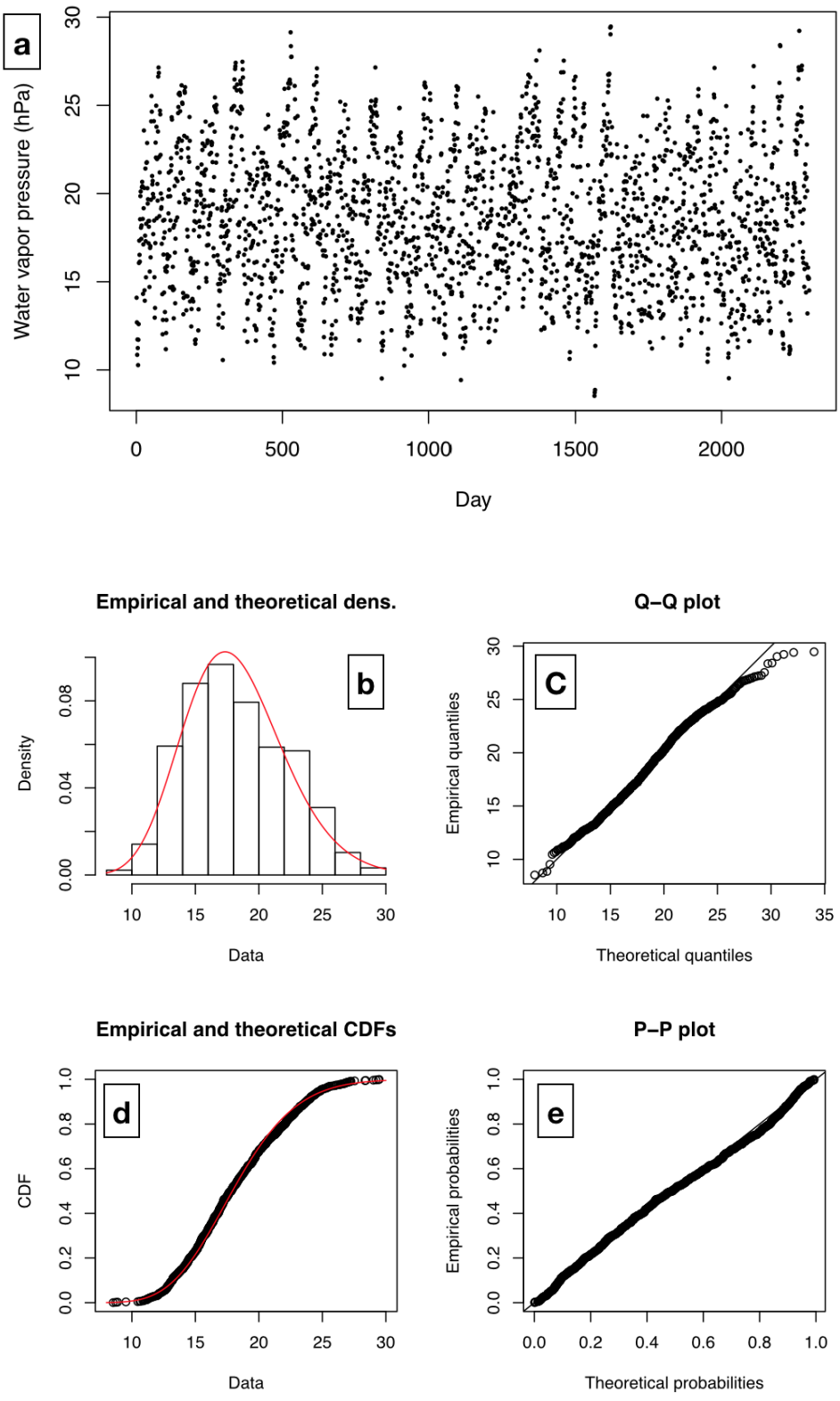


Figure 9.8: (a) the time series of average daily water vapor pressure (hPa) in summer periods from 1989 to 2013. (b) the histogram of data and the fitted Gamma distribution with parameters (shape, rate) = (21.02, 1.155). (c) the Q-Q (quantile-quantile) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis). (d) the comparison of cumulative distribution function between the data (in black) and theoretical Gamma distribution (in red). (e) the P-P (probability–probability) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis).

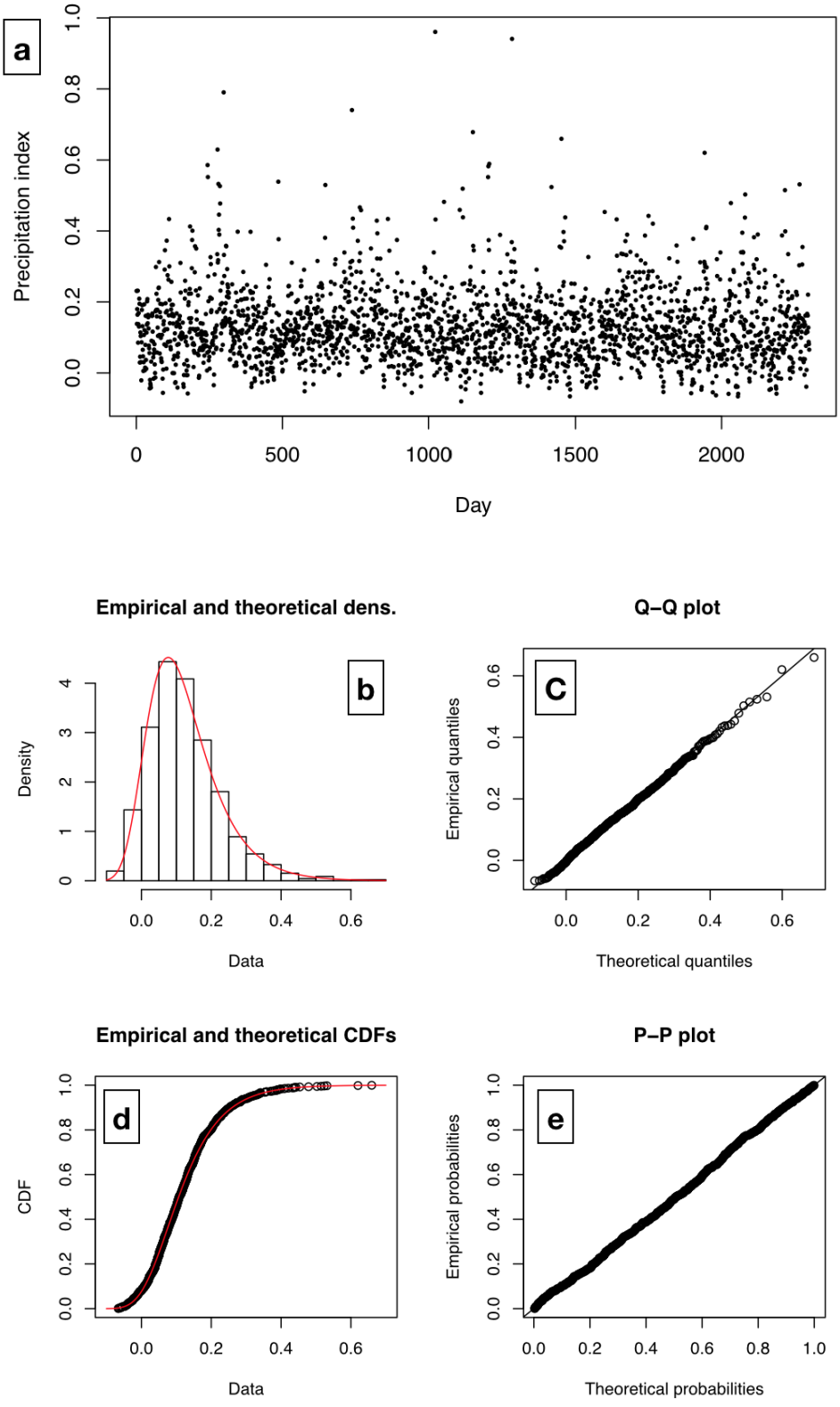


Figure 9.9: (a) the time series of daily precipitation index in summer periods from 1989 to 2013. (b) the histogram of data and the fitted Gumbel distribution with parameters (location, scale) = (0.0768, 0.0814). (c) the Q-Q (quantile-quantile) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis). (d) the comparison of cumulative distribution function between the data (in black) and theoretical Gumbel distribution (in red). (e) the P-P (probability–probability) plot between empirical quantile of data (y-axis) and theoretical quantiles (x-axis).

For each variable, the distribution is fitted among the general common uni-distributions such as Normal distribution, Log-normal distribution, Exponential distribution, Gamma distribution, Weibull distribution and Gumbel distribution.

As mentioned in Section 7.2.2, the auto-regressive process is included in the model and is illustrated in Table 9.3 for the Summer season. In Table 9.3, the data of next day is then added to each line of the original data. Indeed for the 5 variables W (FF[2] [XXX-YES need explain FF], R,T, V, P referred as Wind speed, Solar radiation, Temperature, Water vapor pressure and precipitation index respectively), the temporal relation between a day and its following is preserved by adding their shifted variant $AR(1)$ process ϕ^1 (i.e., FF^1, R^1, T^1, V^1, P^1).

$$\phi^1(t) = \phi(t + dt), \quad \text{for } \phi \in \{\text{FF, R, T, V, P}\}. \tag{9.12}$$

where dt is one time-step (or one day).

These data are then used to build the multivariate $AR(1)$ model for summer period from 1989 to 2013.

Table 9.3: Daily average data in summer period from 1989 to 2013. Columns from 2 to 5 refer to the original data. (FF, R, T, V, P) is the abbreviation of (Wind speed (m/s), Solar radiation (W/m^2), Temperature (°C), Water vapor pressure (hPa), Precipitation index). Columns from 6 to 9 refer to the data one day in advance in comparison to the original data.

date	FF	R	T	V	P	FF (d+1)	R (d+1)	T (d+1)	V (d+1)	P (d+1)
1989/06/01	6.65	588.96	19.85	16.50	0.16	10.85	641.13	19.78	14.28	0.20
1989/06/02	10.85	641.13	19.78	14.28	0.20	11.59	651.48	19.03	12.27	0.01
1989/06/03	11.59	651.48	19.02	12.27	0.01	9.18	649.31	20.11	13.54	-0.01
1989/06/04	9.18	649.31	20.11	13.54	-0.01	data of date 1989/06/05				
.....										
1989/08/31	7.83	653.98	20.88	14.87	0.02	5.35	651.33	20.46	15.13	0.04
1990/06/01	5.35	651.33	20.46	15.13	0.04	data of date 1990/06/02				
.....										
2013/08/30	3.01	649.95	19.49	16.93	0.04	data of date 2013/08/31				

The copula model, based on the 10 variables (i.e., FF, R, T, V, P, FF^1, R^1, T^1, V^1, P^1), is

[2]FF comes from the SYNOP (surface synoptic observations) code - `https://donneespubliques.meteofrance.fr/client/document/doc_parametres_synop_168.pdf`: last check on 2017/11/31.

implemented by the use of the Gaussian copula, defined by its Equation 7.1.

The kriging is then used to simulate the variables sequentially, meaning that in this application, the multi-variable (FF, R, T, V, P, FF^1, R^1, T^1, V^1, P^1) is operated in this specific order. At each time-step, the value of R is simulated conditionally to the variable FF, the value of T is simulated conditionally to the variables R and FF, and so on, finally, the value of P^1 is simulated conditionally to the variables FF, R, T, V, P, FF^1, R^1, T^1, V^1. The way how a simulation is done, whether all at a time or as chained conditional simulation, does not make any difference about the final result, if everything is to be simulated and all copula parameters are supposed exactly known. The flexibility of conditional simulation is useful when some data are already known (like rainfall) and all other variates are simulated conditionally to it as statistically dependent.

Due to the sequential conditional characteristics of the simulations, the construction of the model is also sequential. To do so, the multivariate auto-regressive model needs three factors to use kriging for the simulation :

(a) the correlation matrix (as for the copula parameters);

(b) the parameters of white noise in auto-regressive process;

(c) the kriging weight parameters.

The first one (a) is easily calculated with the data. Table 9.4 gives the example of the correlation matrix of the variables obtained with the Summer data.

Table 9.4: The correlation matrix of multivariables (FF, R, T, V, P, FF^1, R^1, T^1, V^1, P^1)

	FF	R	T	V	P	FF1	R1	T1	V1	P1
FF	1.00	-0.10	-0.30	-0.60	0.03	0.56	-0.09	-0.33	-0.62	-0.30
R	-0.10	1.00	0.44	0.18	-0.34	-0.19	0.61	0.45	0.31	-0.39
T	-0.30	0.44	1.00	0.63	-0.20	-0.17	0.34	0.90	0.57	-0.14
V	-0.60	0.18	0.63	1.00	0.18	-0.25	0.16	0.59	0.74	0.45
P	0.03	-0.34	-0.20	0.18	1.00	0.17	-0.13	-0.20	-0.10	0.53
FF1	0.56	-0.19	-0.17	-0.25	0.17	1.00	-0.10	-0.28	-0.60	0.03
R1	-0.09	0.61	0.34	0.16	-0.13	-0.10	1.00	0.43	0.18	-0.32
T1	-0.33	0.45	0.90	0.59	-0.20	-0.28	0.43	1.00	0.62	-0.18
V1	-0.62	0.31	0.57	0.74	-0.10	-0.60	0.18	0.62	1.00	0.20
P1	-0.30	-0.39	-0.14	0.45	0.53	0.03	-0.32	-0.18	0.20	1.00

From these numerical values, we can see there is no general rule weather the inter-variate or inter-time correlation will be stronger. This agnostic perspective of the technique is a strength that makes it very flexible. In distinct context and use cases, especially in distinct season or time scales, different links may appear, that can be understood or used, even understood and used in the best case.

Equation 7.3 and 7.4 give access to the parameters (b) and (c), respectively, reported in Table 9.5 for the summer season.

Table 9.5: The sequential system of simple kriging for multivariables (FF, R, T, V, P, FF^1, R^1, T^1, V^1, P^1). The first 9 columns contain the kriging weight parameters for the sequential system. The simulated value of the variable $\phi \in \{R, T, V, P, FF^1, R^1, T^1, V^1, P^1\}$ at time t can be obtained by using Equation 7.4, the weight parameters can be found in the row corresponding to the variable ϕ. The last column is the estimation of variance for simple kriging (see Equation 7.6).

	Kriging weight parameters									
	FF	R	T	R	P	FF^1	R^1	T^1	R^1	σ
R	-0.13									0.99
T	-0.22	0.44								0.86
V	-0.45	-0.07	0.54							0.64
P	0.30	-0.17	-0.47	0.76						0.84
FF^1	0.54	-0.14	0.04	0.04	0.09					0.83
R^1	-0.10	0.67	0.12	-0.10	0.11	0.04				0.72
T^1	-0.01	-0.04	0.82	0.00	0.00	-0.12	0.17			0.41
V^1	-0.01	0.12	-0.18	0.57	-0.09	-0.35	-0.11	0.33		0.49
P^1	-0.10	-0.08	-0.22	0.57	0.31	0.17	-0.12	-0.29	0.24	0.58

9.4 Results and discussion

This section presents the results of the simulations generated with the copula-based multivariate model and the comparison between the data from ERA-Interim and the simulated data. Since the daily observations are available from 1989 to 2013, so a 25 years data-set, the simulation is considered in chunks of 25 years. 50 replicates are generated (1250 simulated years all together).

The simulations include the seasonality by generating the model separately for the 4 different seasons. Thus, for each season, a specific model is provided. Several statistical are checked to evaluate the simulations:

1. a general statistical analysis (e.g., the average and the standard deviation) ;
2. the marginal distribution of each variable ;
3. the joint distribution of several variables ;
4. the temporal correlation.

The Section is thus structured in order to discuss on these 4 items.

9.4.1 General statistical analysis

Basic statistic properties such as the mean and standard deviations of the 25-years simulations are first diagnosed. These simple statistics are essential as they provide a preliminary validation of the model before exploring more complex indicators. Figures 9.10 and 9.11 present the statistics of the mean and standard deviations of the simulated period. For each variable, there are 50 replications of 25-years simulations. The black points refer to the mean and standard deviations of 25-years observation for each variable. By its construction, the simulations present a large variability, but the comparisons between the average observed and simulated data show a good agreement for almost all the variables.

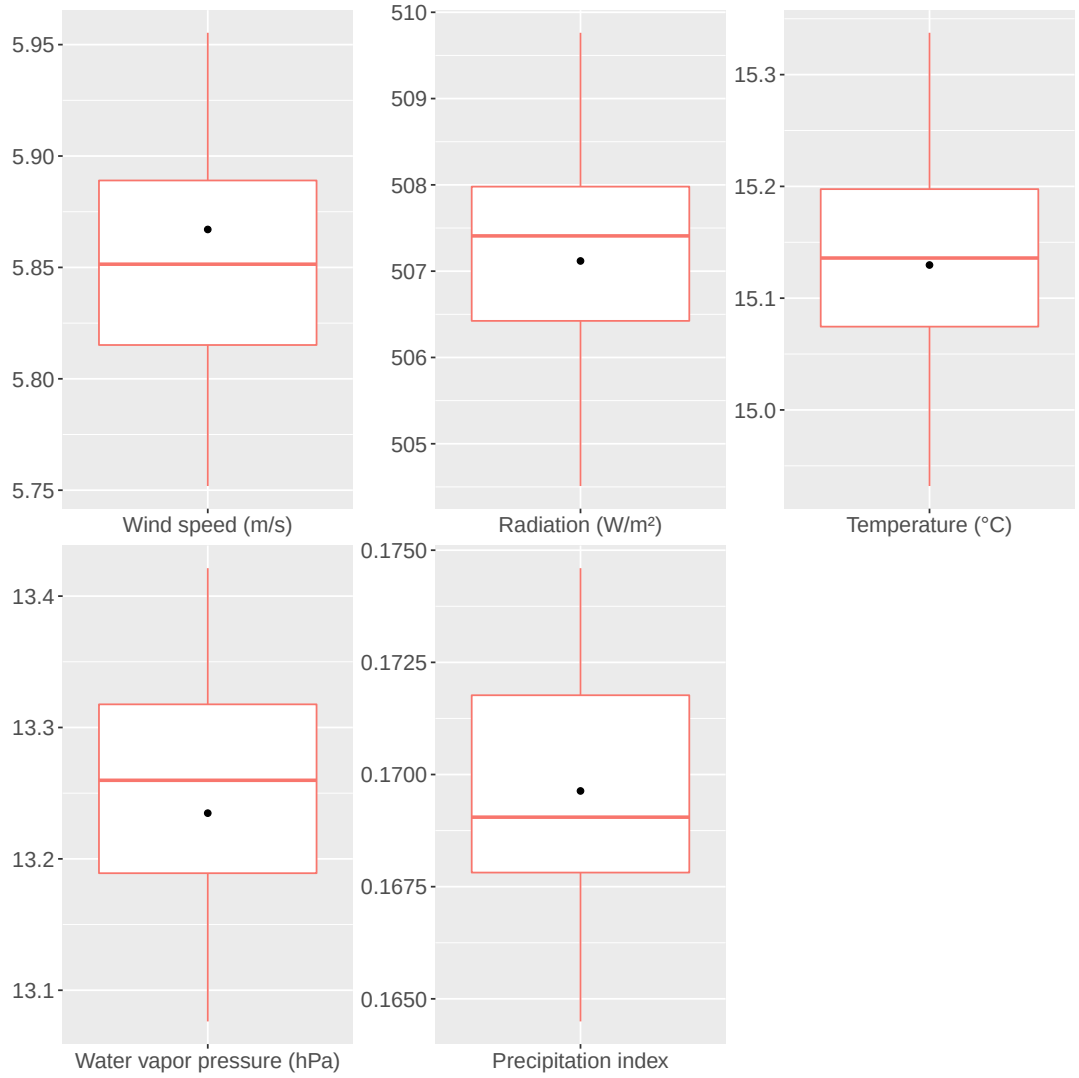


Figure 9.10: Statistics, illustrated with boxplots, of the average values for each variable over the 25-years period. Black points refer to the observed average value. The statistics of the simulated values, in red, rely on the 50 replications of 25-years simulations.

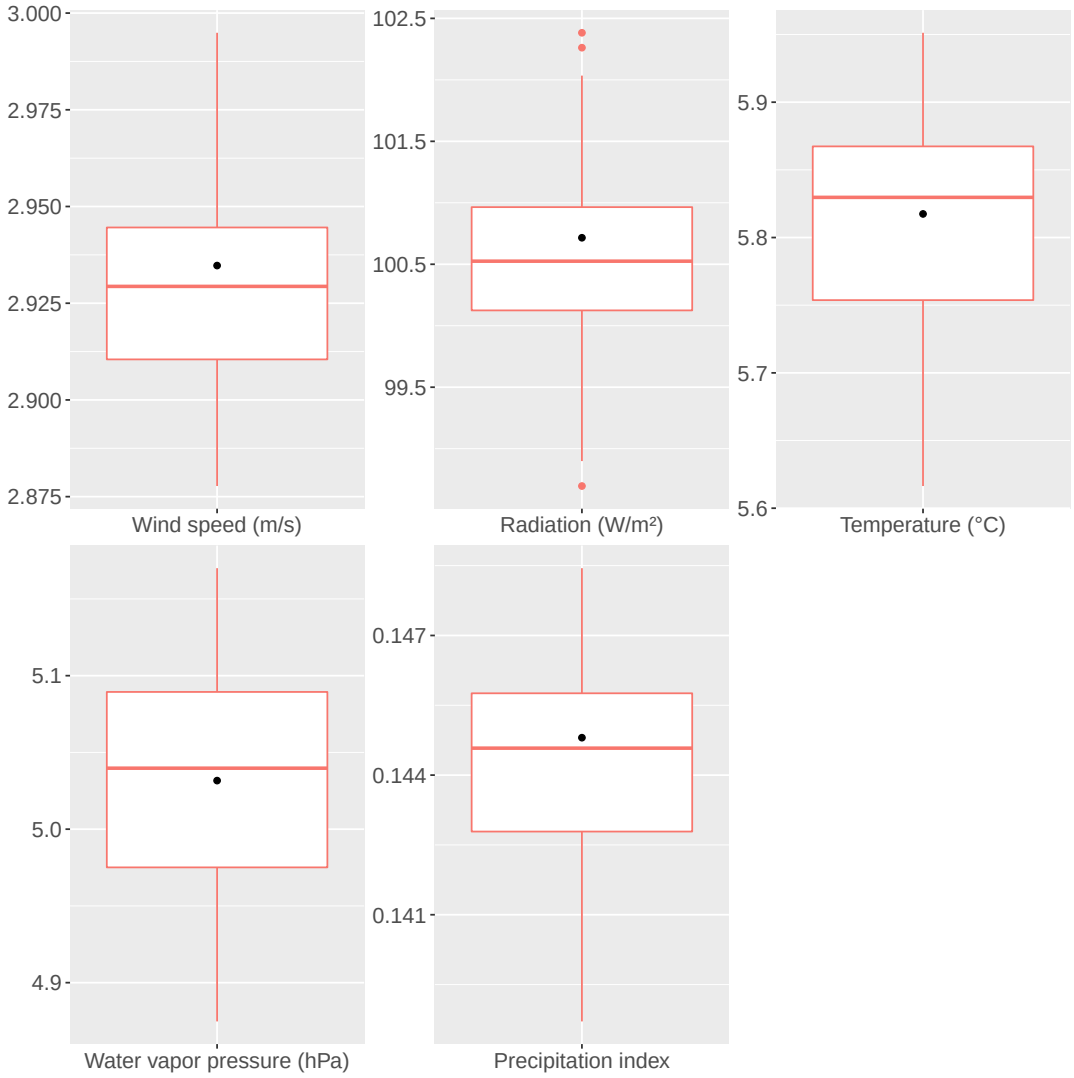


Figure 9.11: Statistics, illustrated with boxplots, of the standard deviation for each variable over the 25-years period. Black points refer to the observed standard deviation. The statistics of the simulated values, in red, rely on the 50 replications of 25-years simulations.

The capacity of the model to reproduce the seasonality is illustrated in Fig. 9.12. As expected, the variability within each season is strongly reduced compared to the one for the whole year, and the comparison with the observation presents better agreement.

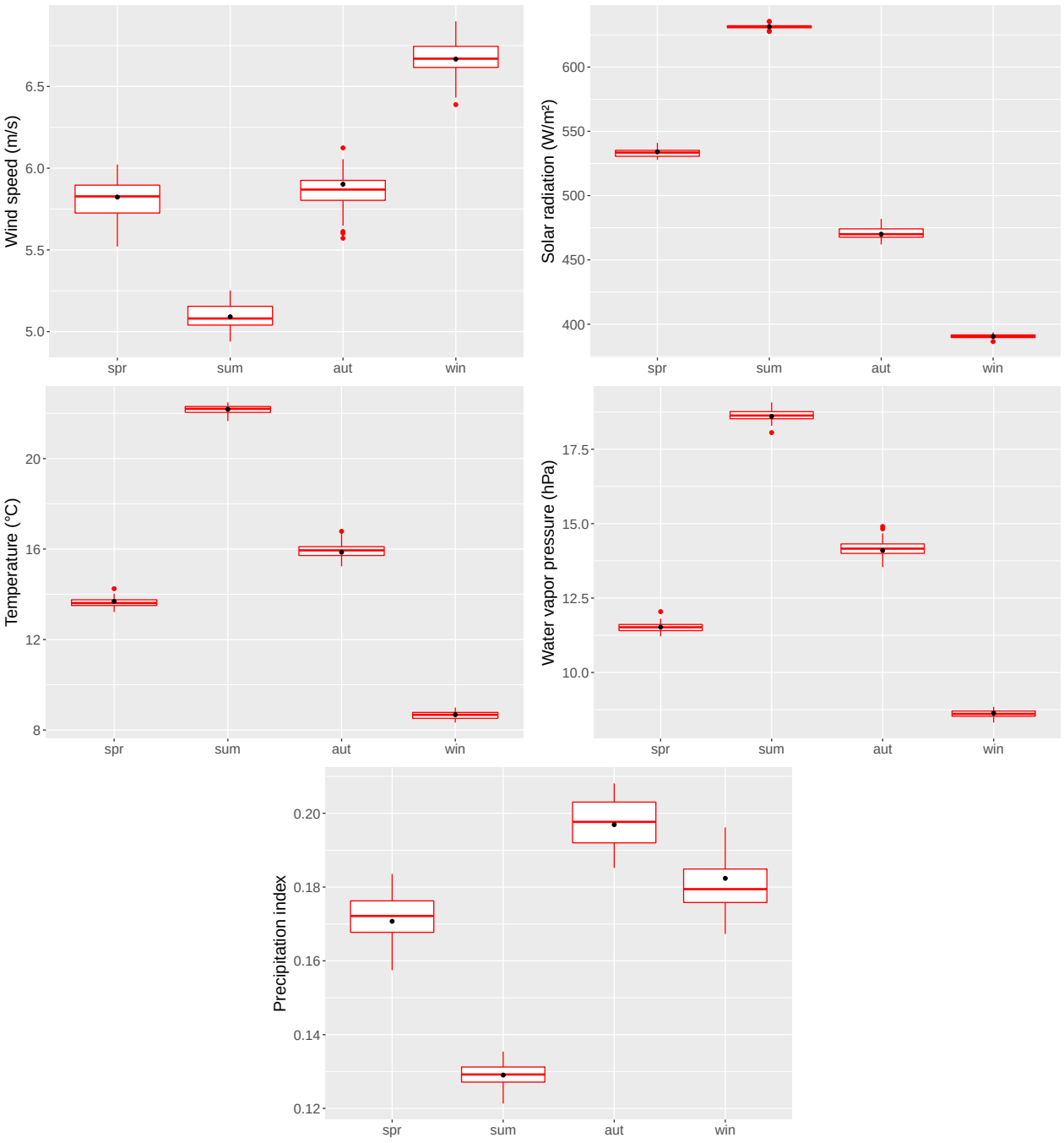


Figure 9.12: Statistics, illustrated with boxplots for each season, of the average value for each variable over the 25-years period. Black points refer to the observed average value. The statistics of the simulated values, in red, rely on the 50 replications of 25-years simulations.

This result is confirmed with the QQ-plot graphics, introduced in Fig. 9.13, where the comparison between the observed and the simulated probability distributions presents a fairly good agreement. The points approximately remain on the line $y = x$.

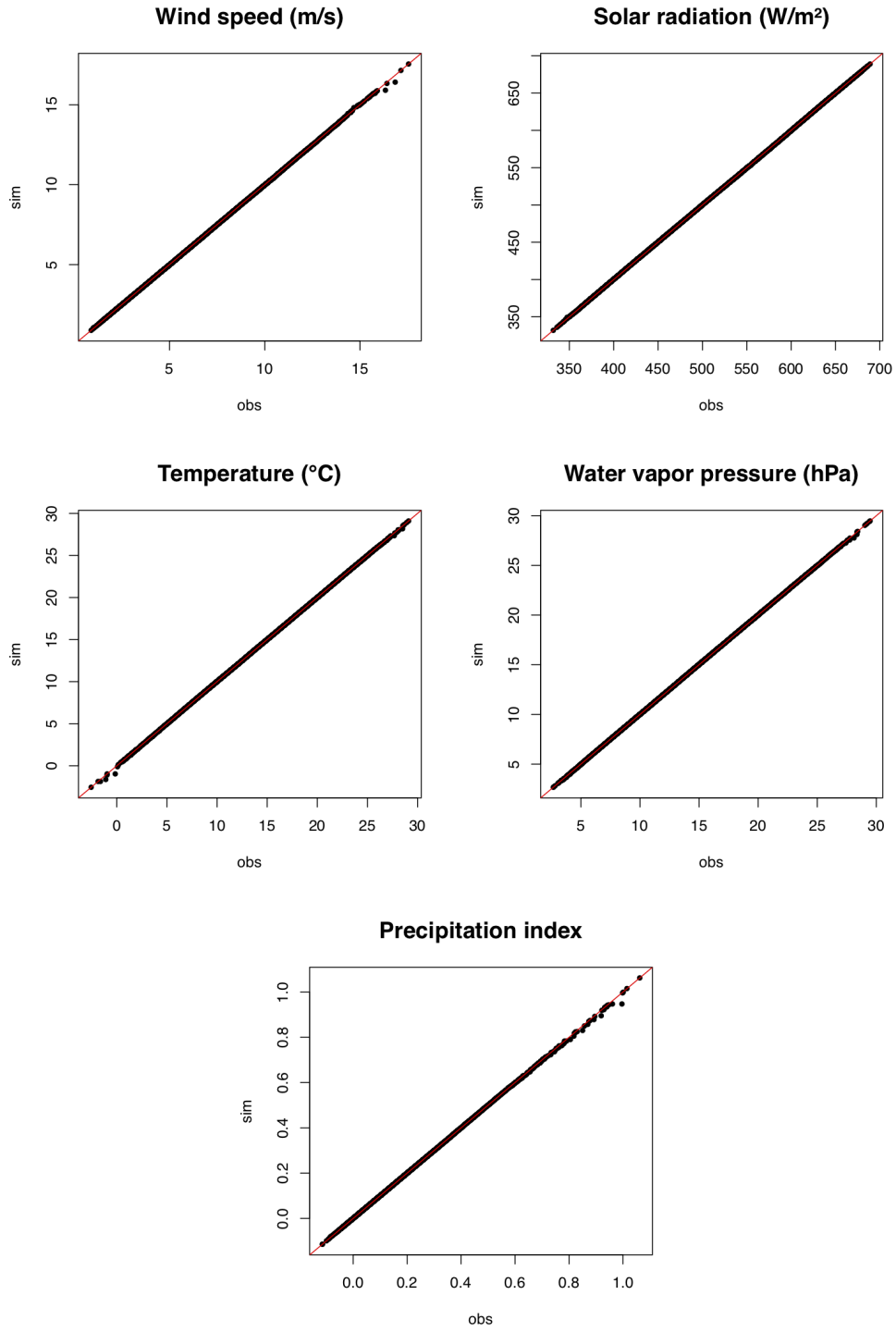


Figure 9.13: QQ-plots for each variable by comparing the marginal distributions between 25-years observed and one 25-years simulation.

Some discrepancies still exist for the high values, in particular for water vapor pressure and precipitation index. This result is due to approximate adjustment of their marginal

distributions. As Sklar's Theorem C.4 mentioned, the joint distribution of multivariable can be treated separately with a copula function and all marginal distributions. If also means that if the marginal distributions could not be fitted correctly to observed data, the copulas will not solve the problem. The theorem makes sure that the marginal distributions of variables are not affected by the nature of copulas. This is one of the main advantage to use copula technique to deal with multivariate problem. In this paper, the empirical distribution function is chosen to fit marginal distribution of the variables.

9.4.2 Bi-variate distribution

Bivariate distributions of all pairs of variables are presented in Fig. 9.15 for the summer season. The comparison between the observed (in black) and the simulated (in red) focus on the 25-years simulated period.

The Figures may be split in two parts. The upper right part presents the point-by-point comparison between the observed and simulated data. As illustrated in the Fig. 9.15, the red and the black clusters occupy more or less the same area. Few discrepancies appear when analyzing the highest precipitation index values. In the lower part, the plots show the bivariate distributions of all pairs of variables. The simulated contours (in red) refer to one replication of 25-years simulations. Even if the two colored contours do not coincide exactly, both bivariate distributions present more or less the same shape, which is an encouraging result. Looking in more details (Figure 9.18), it is almost impossible to identify the observed bivariate distribution among to the simulated ones, and this is as expected.

The results are encouraging even though the model is based on simple Gaussian copula.

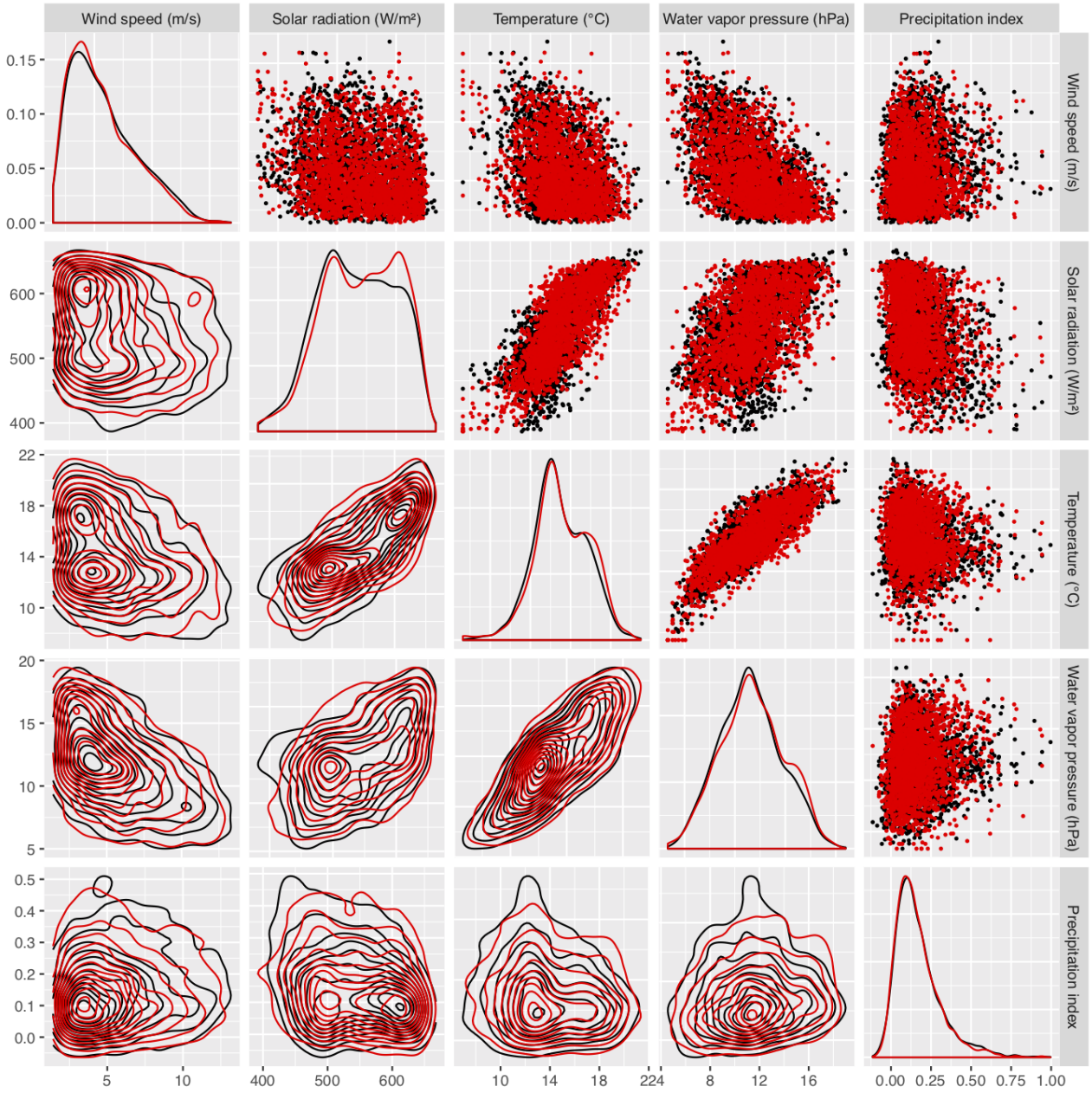


Figure 9.14: Bivariate distributions of the observed (in black) and the simulated (in red) data for the spring season within the 1989-2013 period. The simulated contours refer to the mean of the 50 replications of the 25-years simulation. Upper right part of the figure) the point-by-point representations of all pairs of variables; lower left part of the figure) the bivariate distributions of all pairs of variables; the diagonal) the comparison between the observed and the simulated data distribution for each variable.

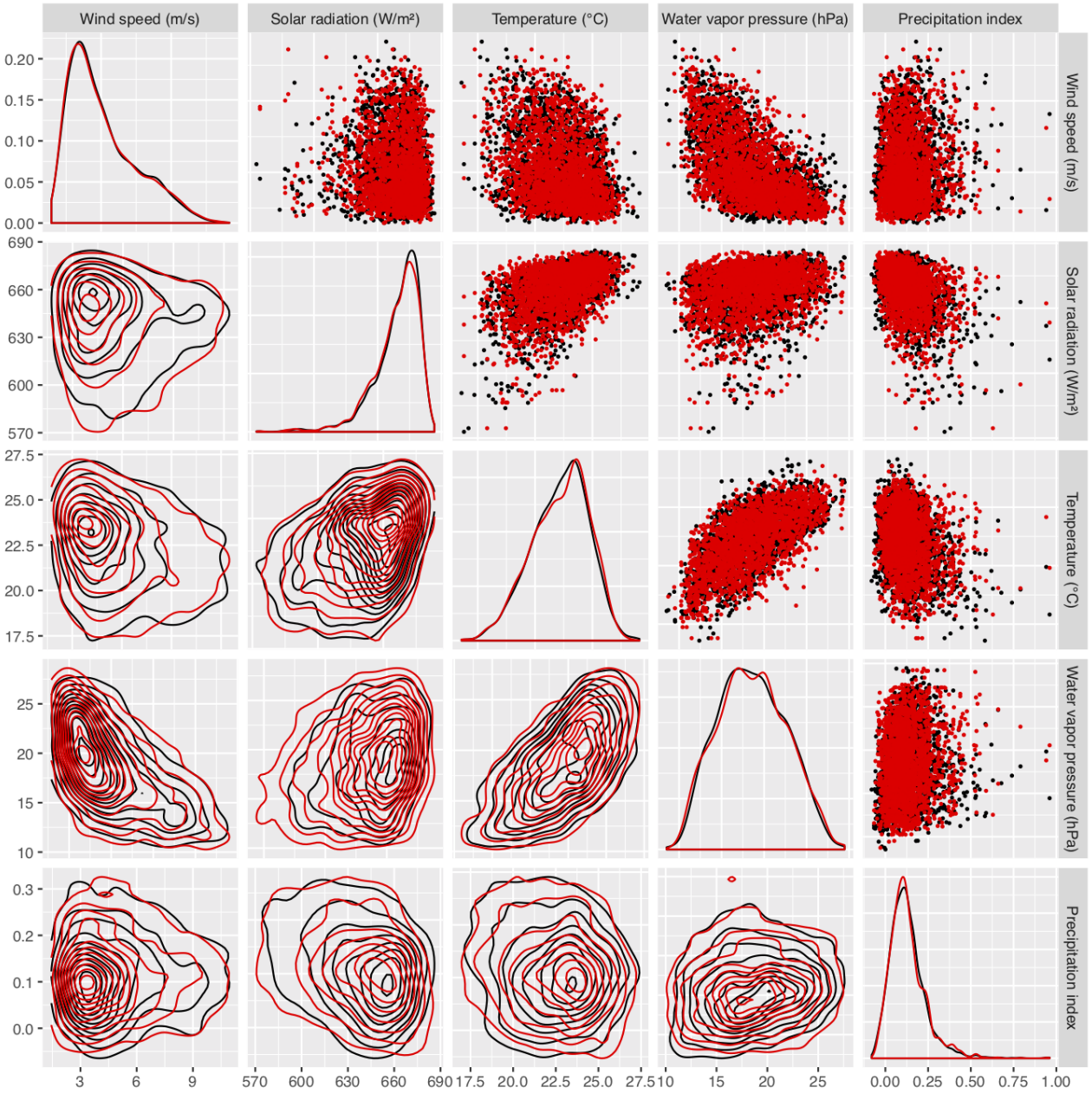


Figure 9.15: Bivariate distributions of the observed (in black) and the simulated (in red) data for the summer season within the 1989-2013 period. The simulated contours refer to the mean of the 50 replications of the 25-years simulation. Upper right part of the figure) the point-by-point representations of all pairs of variables; lower left part of the figure) the bivariate distributions of all pairs of variables; the diagonal) the comparison between the observed and the simulated data distribution for each variable.

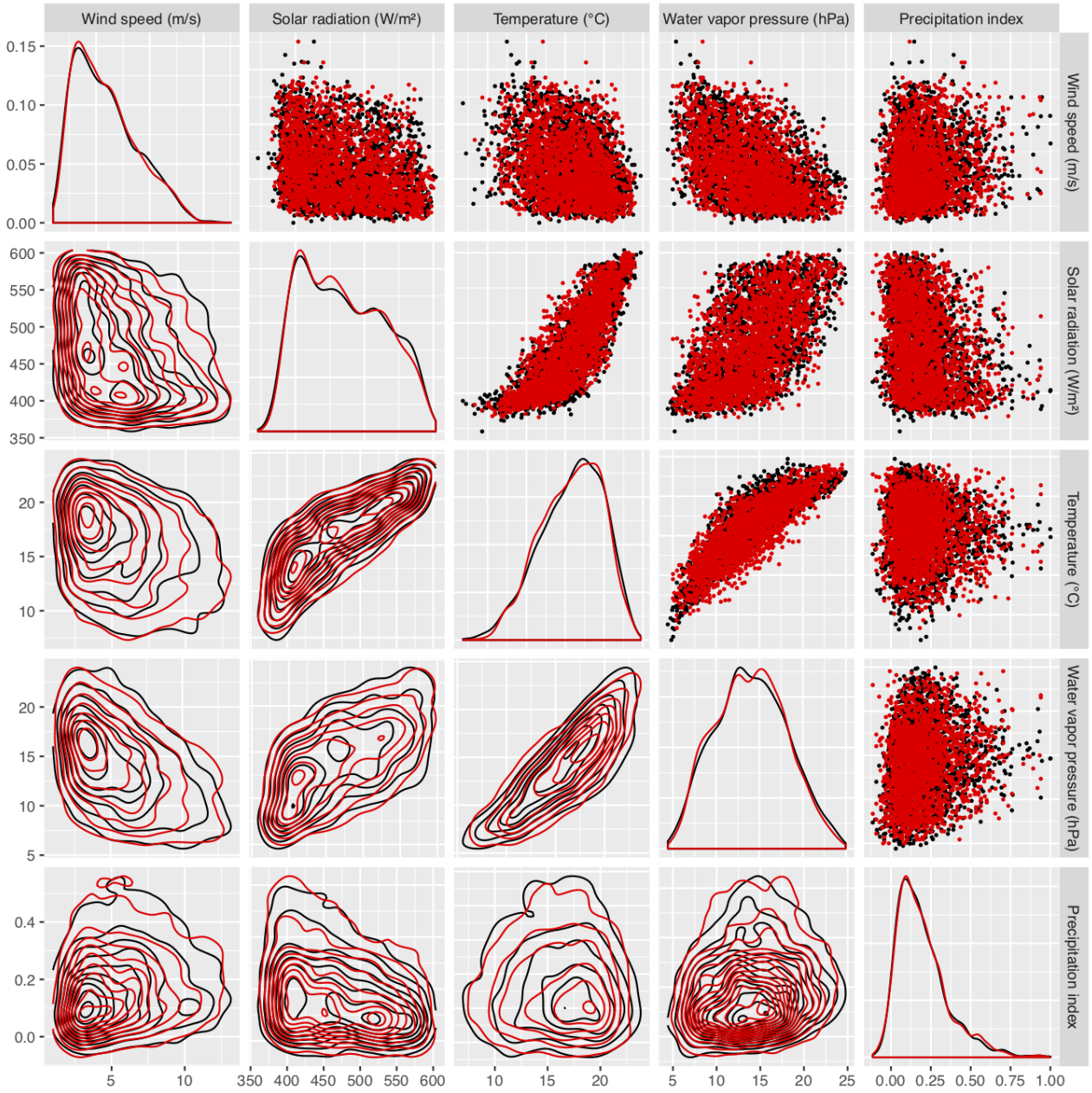


Figure 9.16: Bivariate distributions of the observed (in black) and the simulated (in red) data for the autumn season within the 1989-2013 period. The simulated contours refer to the mean of the 50 replications of the 25-years simulation. Upper right part of the figure) the point-by-point representations of all pairs of variables; lower left part of the figure) the bivariate distributions of all pairs of variables; the diagonal) the comparison between the observed and the simulated data distribution for each variable.

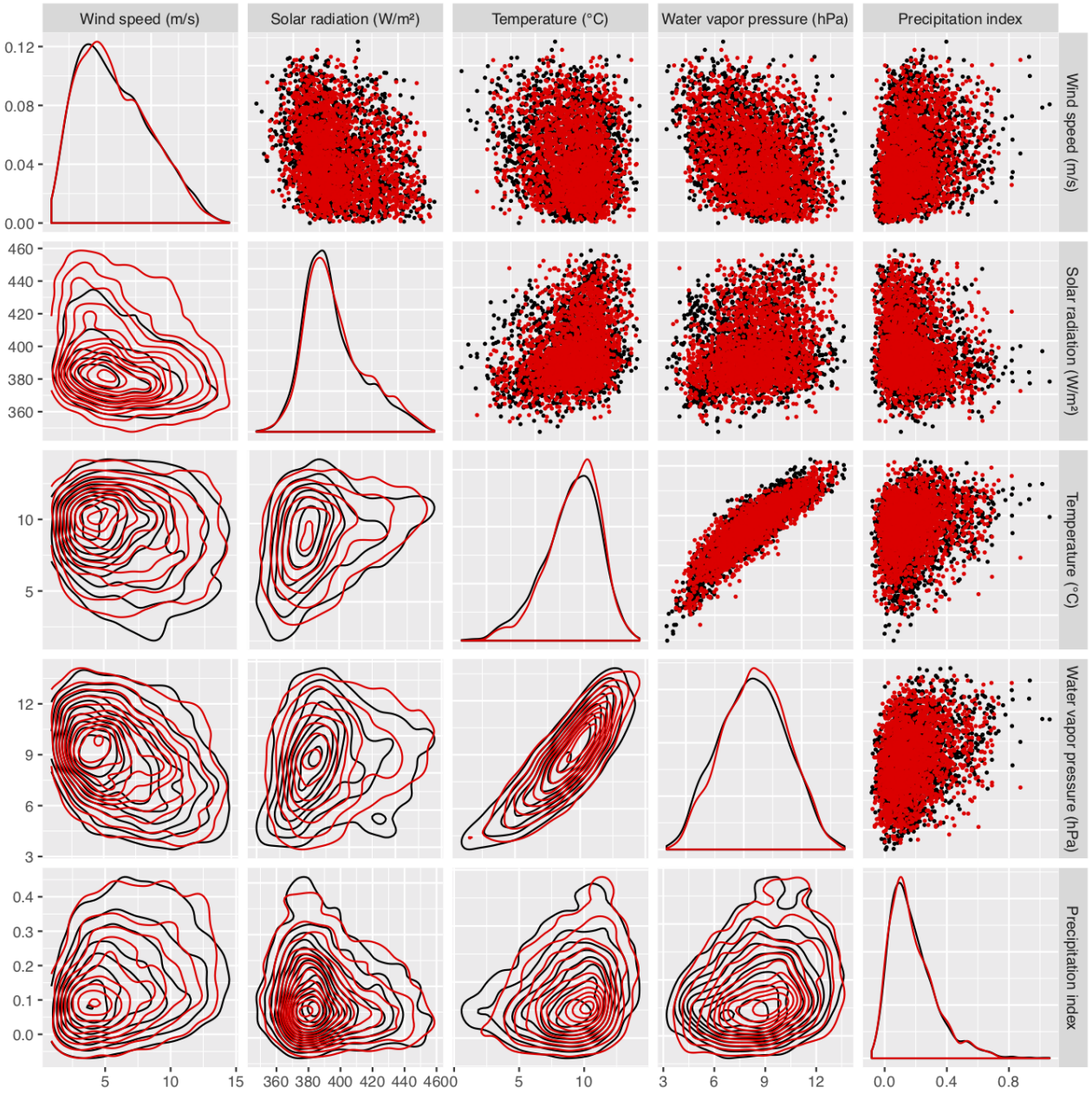


Figure 9.17: Bivariate distributions of the observed (in black) and the simulated (in red) data for the winter season within the 1989-2013 period. The simulated contours refer to the mean of the 50 replications of the 25-years simulation. Upper right part of the figure) the point-by-point representations of all pairs of variables; lower left part of the figure) the bivariate distributions of all pairs of variables; the diagonal) the comparison between the observed and the simulated data distribution for each variable.

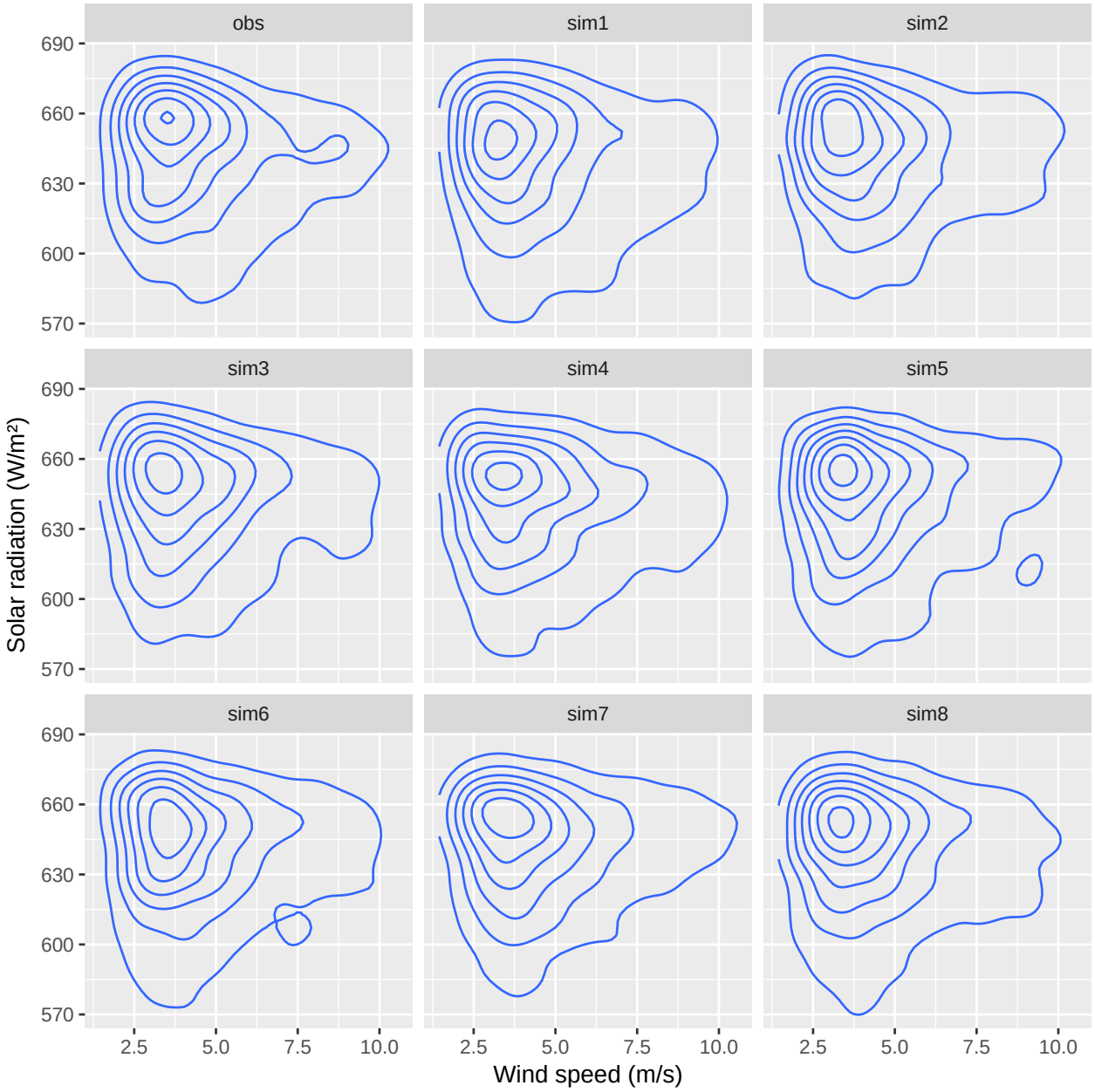


Figure 9.18: The comparison of bivariate distributions of wind speed and solar radiation between the observation data and simulation data. There are one observation plot and 8 simulation plots. The span of each data is 25-years for summer periods. This displays the variability of 25-years chunks, and suggest there is no obvious difference between any of them and the observed.

9.4.3 Temporal correlation

Temporal correlation is an important issue since the objective of this multivariate model is to provide relevant simulated meteorological variables that will be further used as inputs of hydrological models in order to evaluate water resource at the regional scale. In this context, the auto-correlation functions (ACF) are used to evaluate the capacity of the model to proper reproduce the observed temporal correlation. They are drawn in Fig. 9.19 - Fig. 9.22 for each variable, for the spring, summer, autumn and winter seasons, respectively.

Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. ACF represents the serial dependence for a time series of random variable.

Overall, Fig. 9.19 - Fig. 9.22 present a fairly good agreement between the observed and simulated ACF for all seasons. For the variables which do not present any strong periodic change in time, like wind speed and precipitation index, the performance of the model is really good. The observed and the simulated ACF present very weak discrepancies. Nevertheless, for the three other variables, in particular the solar radiation, the multivariate model is not able to properly reproduce its periodic feature as the one observed in the observation data. In spring, the model correctly reproduce the ACF of the simulated solar radiation up to 15 days, but is not able to capture the change of the sign.

This result is associated with the choice of the 3-month block to define the season, and consequently to implement the multivariate model.

But the existence of non-stationarity (periodic trend) [XXX-NON inside for ?] inside of certain variables (such as temperature or solar radiation) for 3-month block could be an issue for temporal modeling. One possible solution is to reduce 3-month block into a shorter interval (e.g., one-month interval). Another way would be to fit a trend where obvious and physically appropriate and remove it from stochastic simulation.

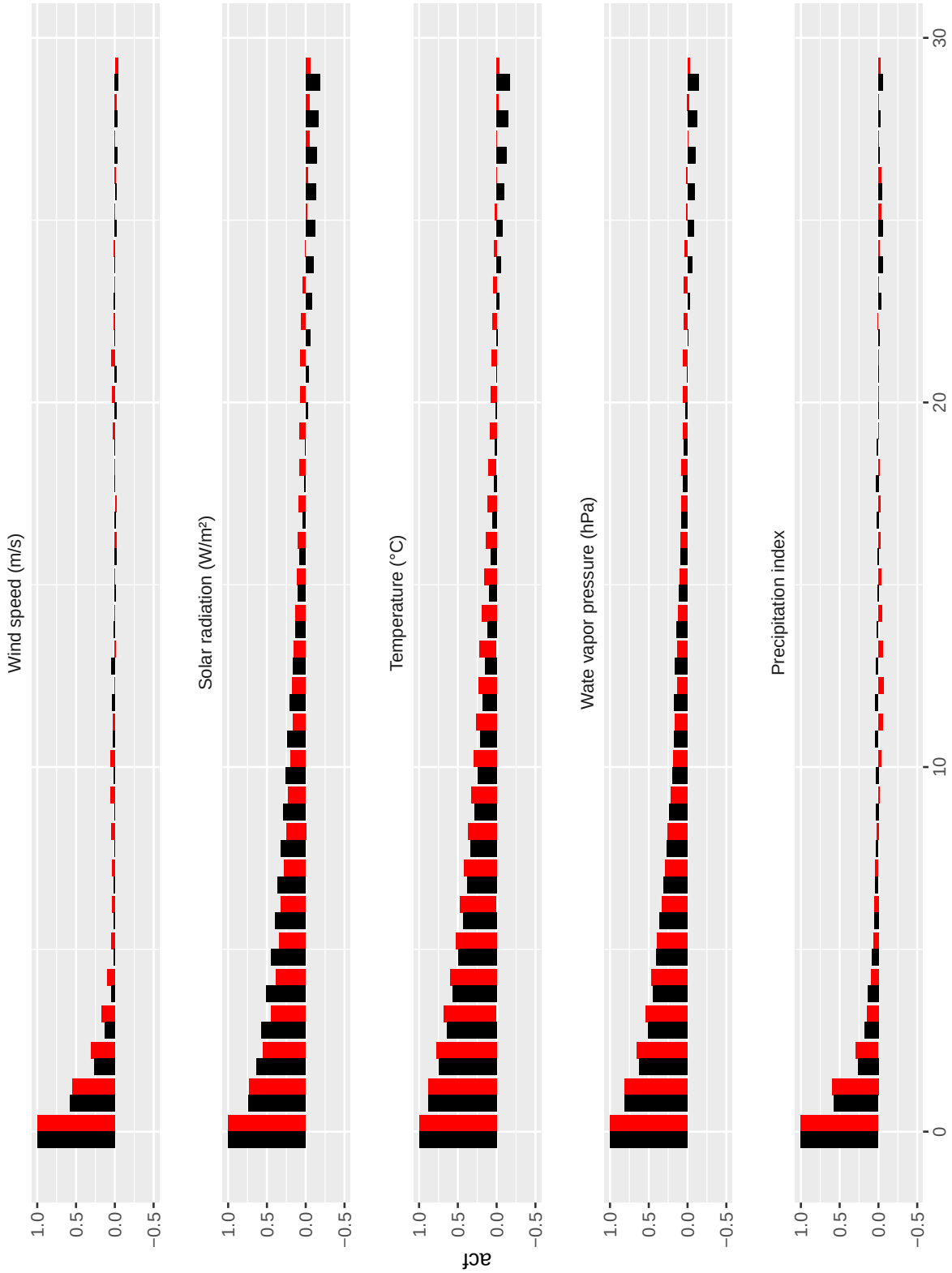


Figure 9.19: Auto-correlation functions (ACF) of observed (in black) and simulated (in red) data for the spring season within the 1989-2013 period.

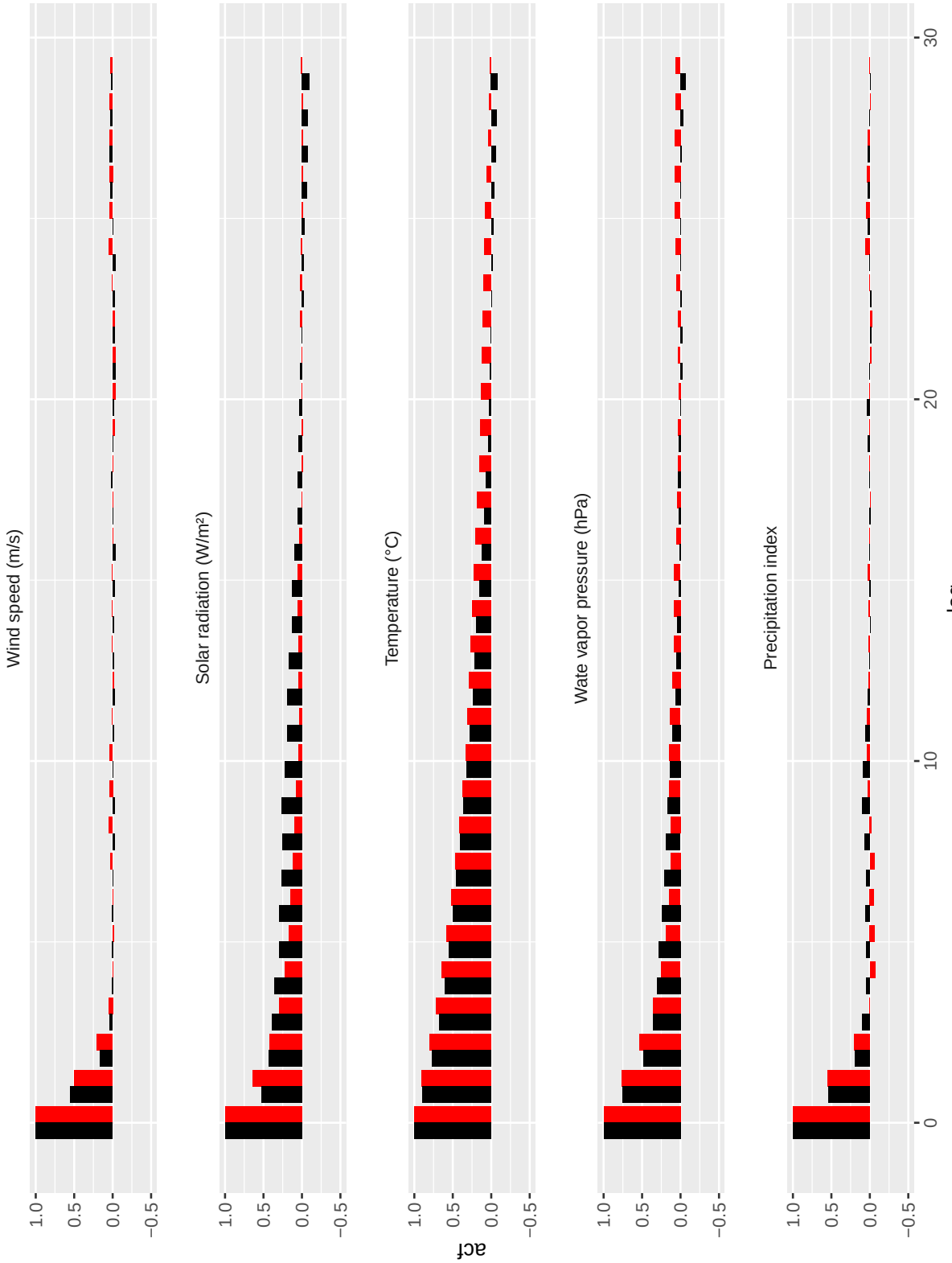


Figure 9.20: Auto-correlation functions (ACF) of observed (in black) and simulated (in red) data for the summer season within the 1989-2013 period.

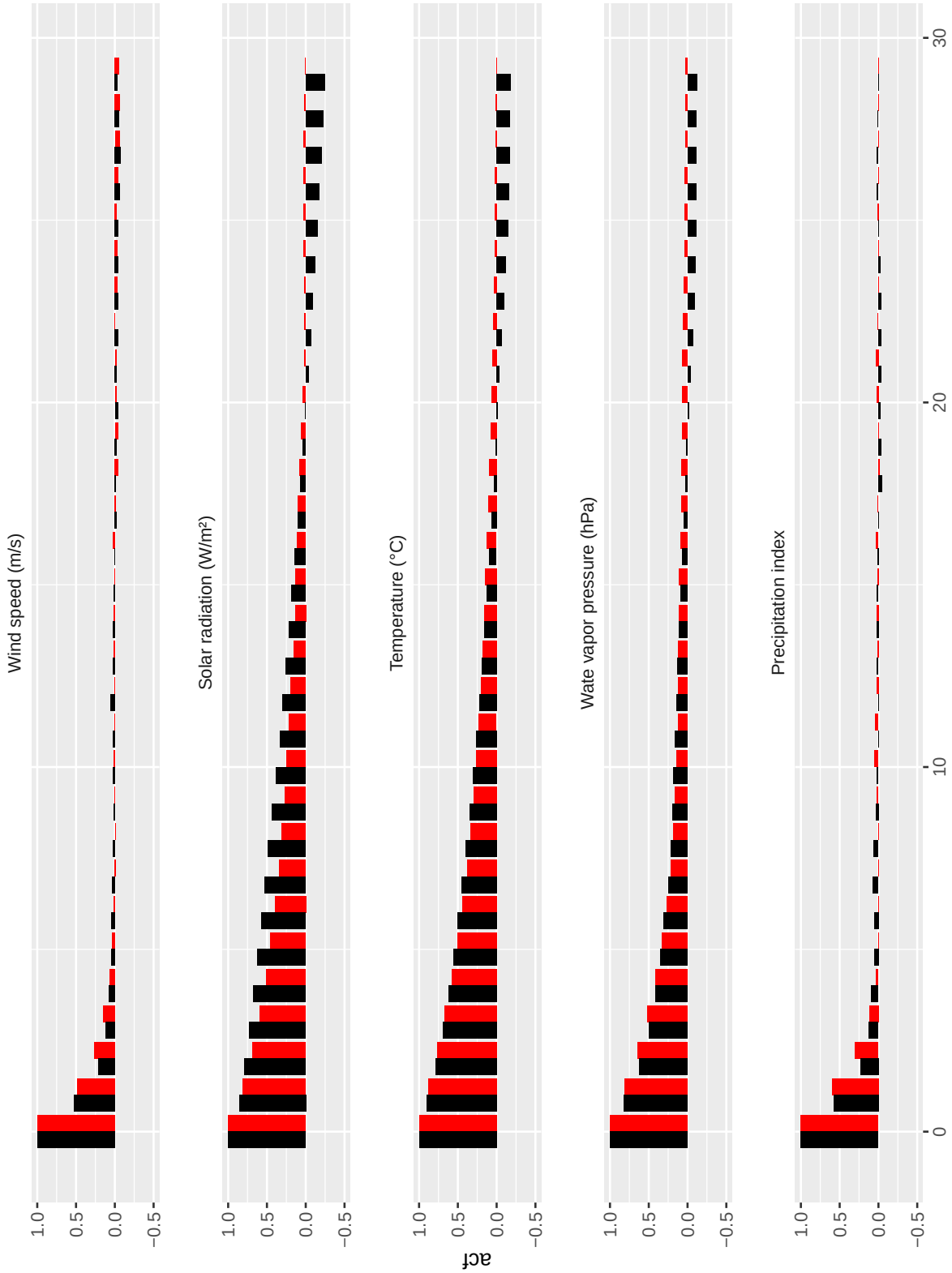


Figure 9.21: Auto-correlation functions (ACF) of observed (in black) and simulated (in red) data for the autumn season within the 1989-2013 period.

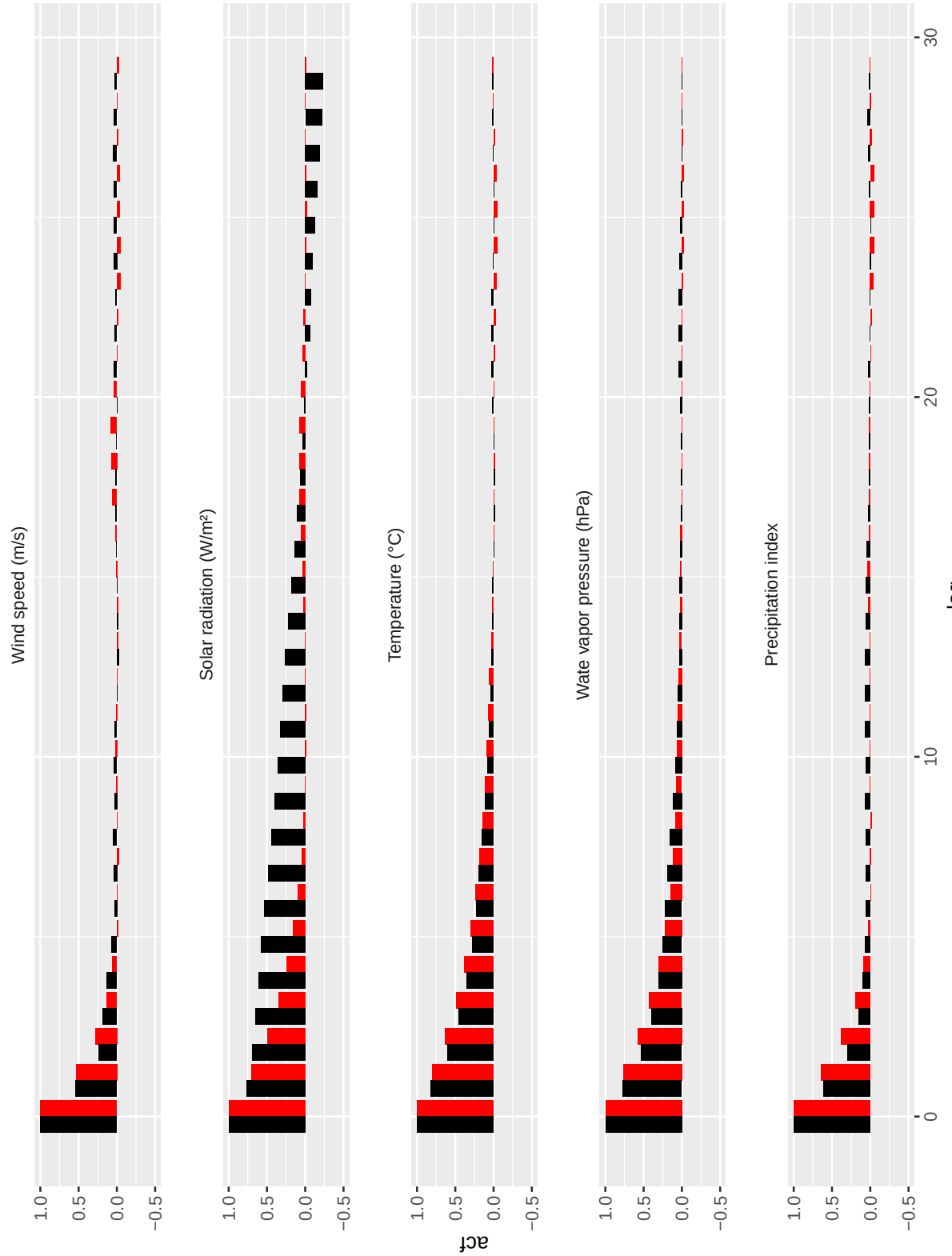


Figure 9.22: Auto-correlation functions (ACF) of observed (in black) and simulated (in red) data for the winter season within the 1989-2013 period.

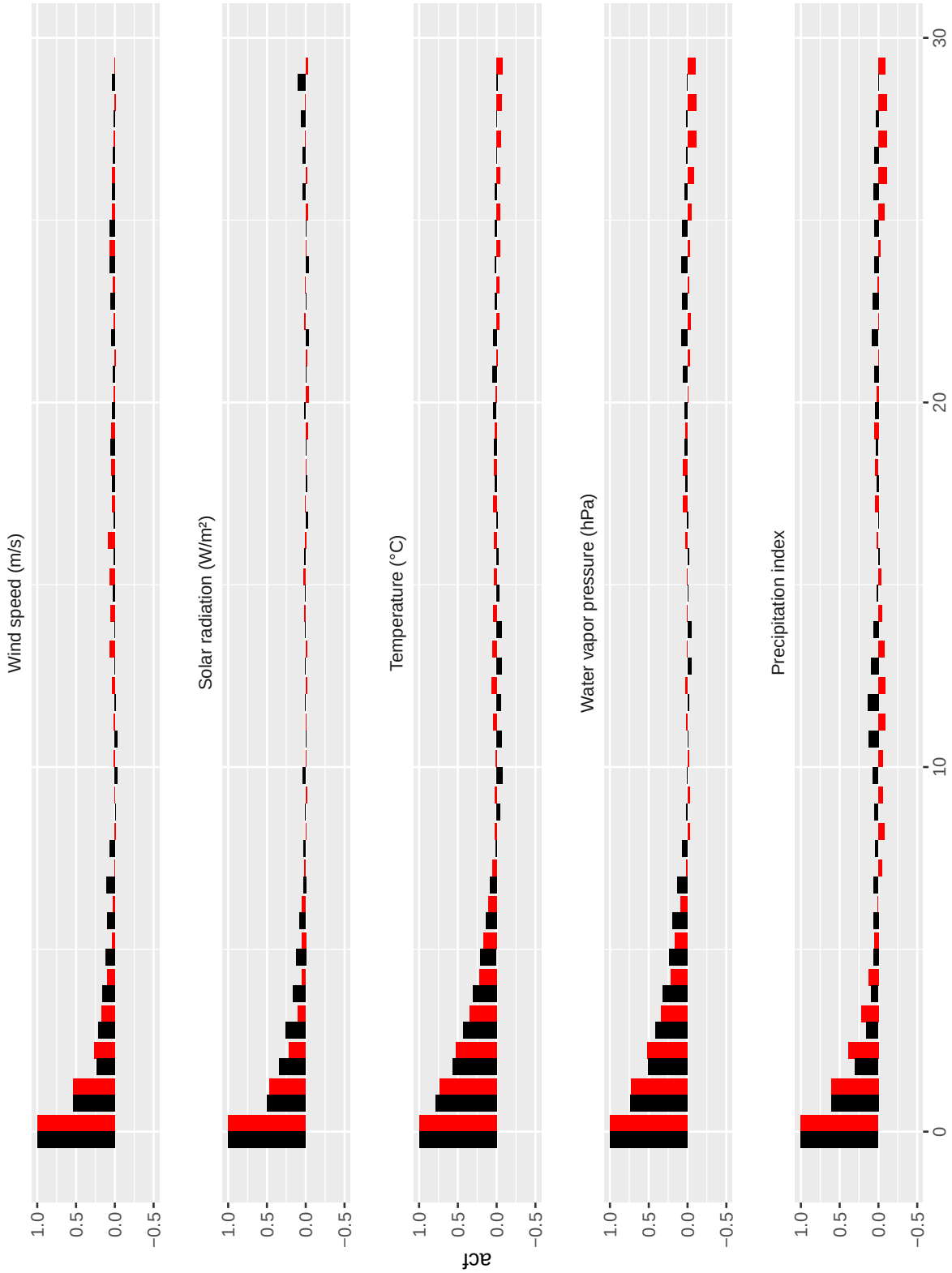


Figure 9.23: Representations of the auto-correlation function of observed data and simulated data for each variable in January periods. Black represents the observed data and red represents the simulated data.

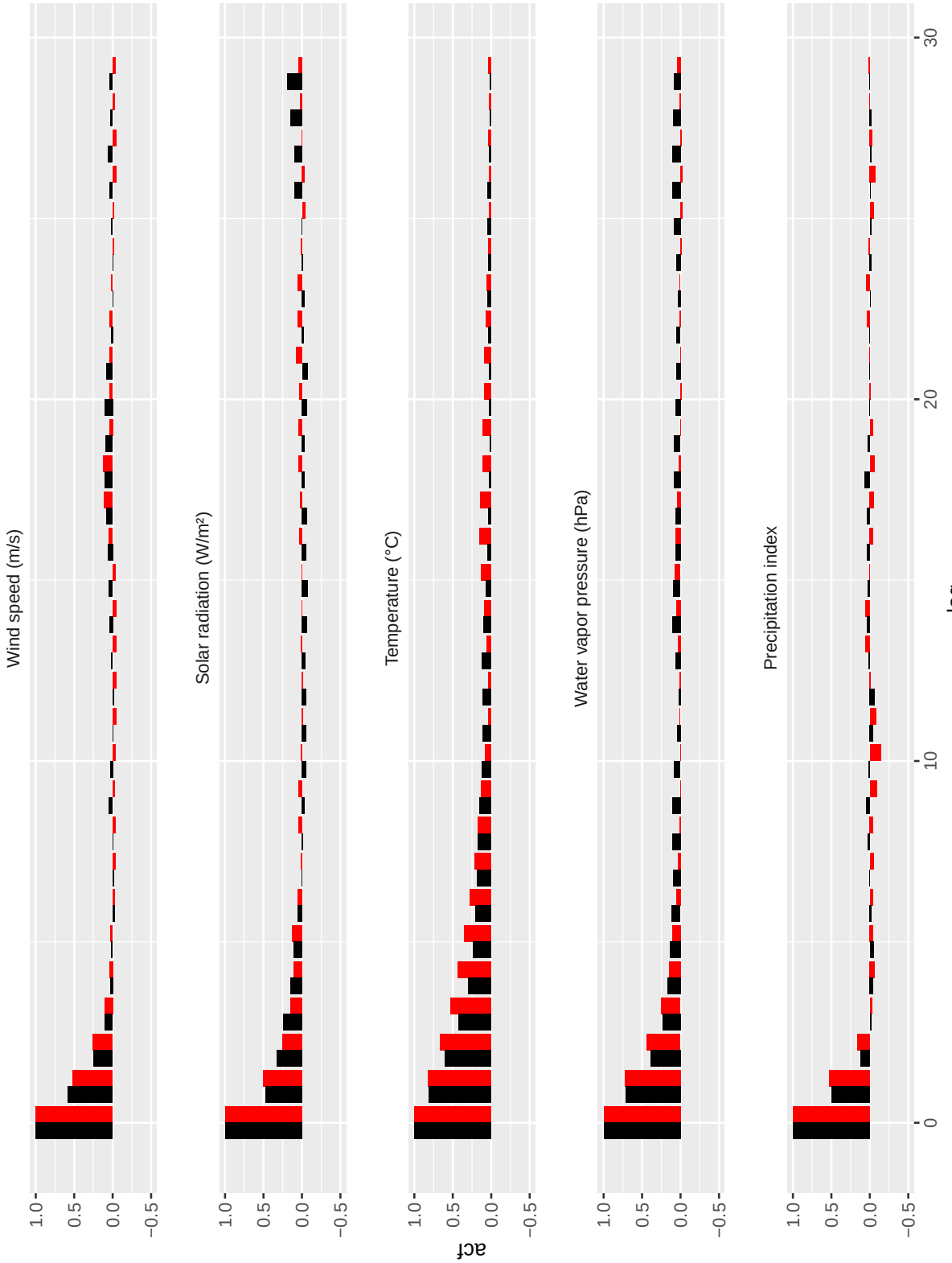


Figure 9.24: Representations of the auto-correlation function of observed data and simulated data for each variable in August periods. Black represents the observed data and red represents the simulated data.

9.4.4 Seasonality

The definition of season and its associated months is based on usual meteorological conventions. However, the temporal correlation diagnosis (Section 9.4.3) shows the importance of such seasonality choice. In this section, the signature of the definition of the season not only on the performance, but also on the multivariate dependence structure is shown. One copulas model per month is built, and the impact of the definition of the season (i.e. 4 seasons per year versus 12 seasons per year [XXX-NON monthly values ?]) in the multivariate simulation is illustrated in Figure 9.25 and Figure 9.26.

Figure 9.25 presents the Kendall rank correlation coefficients between temperature and precipitation index within each individual season. The overall evolution of the correlation coefficients presents the same pattern over the year, whatever the number of seasons in a year (4 or 12). Temperature has the positive correlation with precipitation index during the winter and autumn periods, but a negative correlation during the spring and summer periods. This clearly demonstrates a change in the inter-variable structure along the year, that we know has climatological relevance and can not be ignored in practice.

On the contrary, Figure 9.26 presents the Kendall rank correlation coefficients between solar radiation and precipitation index within each individual season. Just in December, the precipitation index has a positive correlation with solar radiation, but this is not very strong and disappear when only 4 seasons per year are considered.

The segmentation of the year in seasons of homogeneous dependence structure, or the relevant parametrisation of significant gradual changes, could be interesting as such. It is certainly a next step in SWG development.

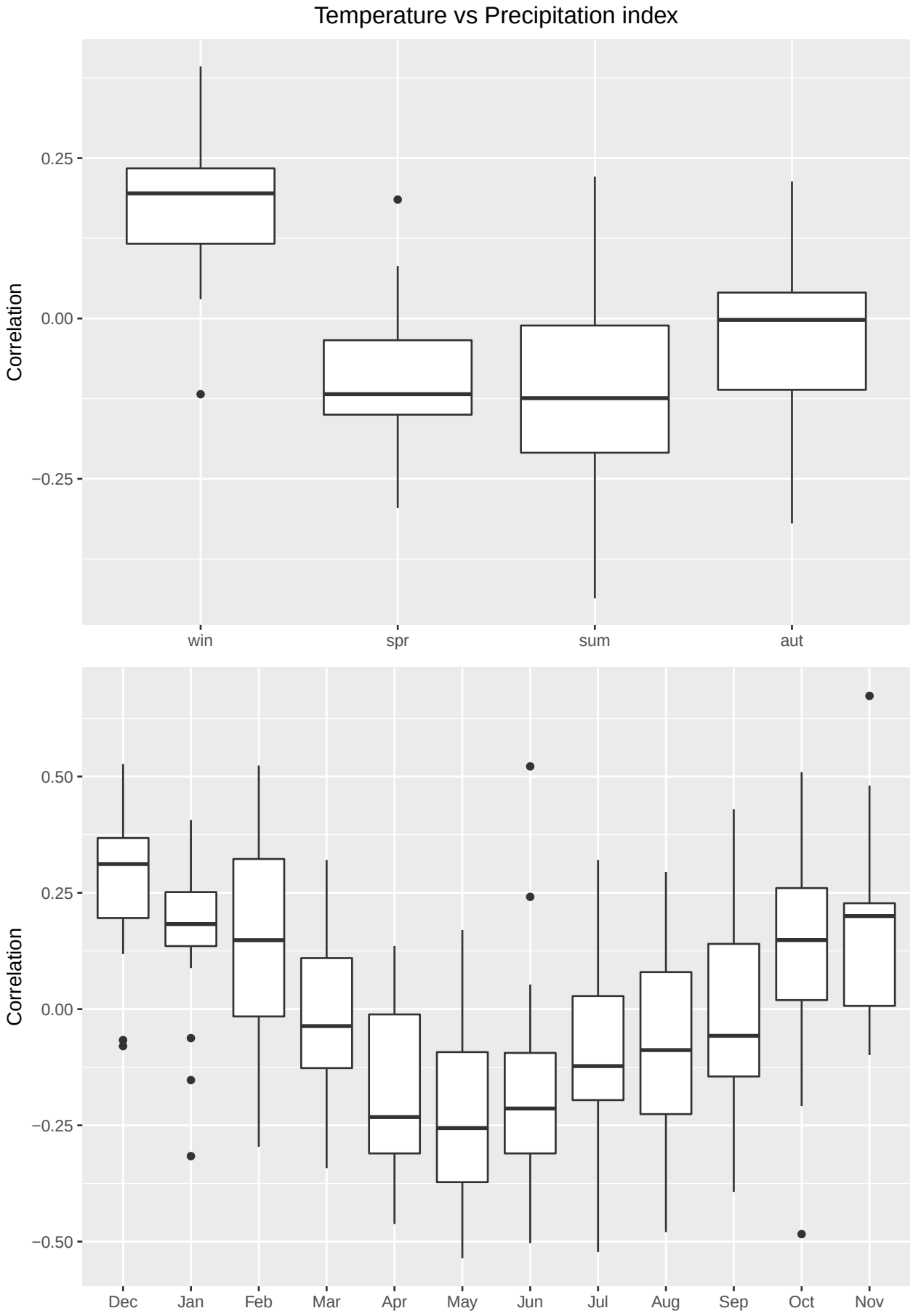


Figure 9.25: Kendall rank correlation coefficients between temperature and precipitation index within each individual seasons. Each boxplot presents 10 yearly correlation coefficients within the 1989-2013 period. Above: 4 seasons per year; Below: 12 seasons per year.

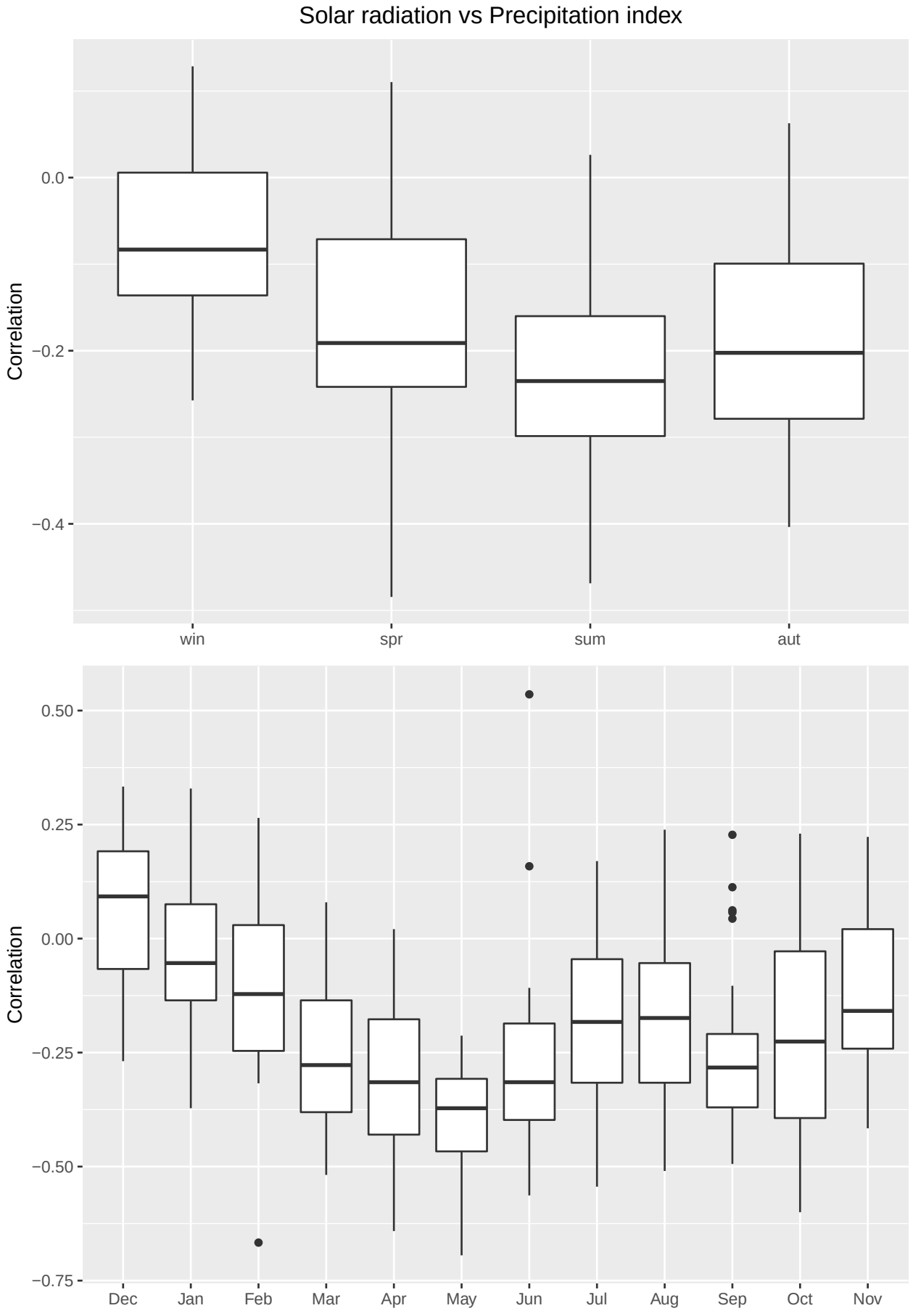


Figure 9.26: Kendall rank correlation coefficients between solar radiation and precipitation index within each individual different seasons. Each boxplot presents 10 yearly correlation coefficients within the 1989-2013 period. Above: 4 seasons per year; Below: 12 seasons per year.

9.5 Conclusions and discussions

A multivariate model is necessary to provide relevant meteorological variables required as inputs of hydrological models. The inputs must to be statistically consistent, within a range of time and space scales determined by the intended use of the simulation, with observation data.

As it is well known, the multivariate joint distribution remains one of the major difficulties for multivariate modelling. Within the framework proposed by *Sklar* [1959] in its theorem (Equation C.4), a continuous multivariate joint distribution may be expressed as the product of the marginal distributions of each single variable and a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. The importance of this theorem is to establish the relationship between the multivariate joint distribution and each univariate marginal distribution, by means of a copula function that describes the dependence structure between the uniform transformed variables. Therefore, the copula technique has become one more and more attractive statistical tool to analyze multivariate model. As one of the elliptical copula families, the Gaussian copula is most popular because of the simplicity of their formula and the easy links with other statistical tools under Gaussian framework (such as geostatistics). Certainly, the Gaussian copula is frequently inadequate, because it can not model tail dependence, making it unsuitable for the situations where tail dependence exists and must be considered [see *Mazur and Piterbarg*, 2015; *Serinaldi et al.*, 2015; *Furman et al.*, 2016; *Hao and Singh*, 2016]. A possible improvement would be to use multivariate t-distribution which resembles multivariate normal distribution with one extra parameter in t-distribution that can give more flexibility. *Evin et al.* [2017] has used t-copula to generate multi-site daily precipitation for the assessment of extreme floods in Switzerland.

Another major issue is the capacity of the multivariate model to reproduce a sound temporal correlation of each variable. Here the auto-regression aspect was explicitly introduced to check how easily the temporal dependence can be handled in the copula framework. In this paper, only $AR(1)$ process has been applied, but the development done do not suggest any lack of generality. If the results are globally encouraging, a main difficulty remains associated with the important seasonality of some variables. Introducing seasonality by 3-month blocks seems to be not enough because seasonal trends in some variates like temperature and solar radiation make a *stationary* model ineffective.

In this paper, only Gaussian copula has been considered, the general results are quite promising. On one hand, because of the limitations of the Gaussian framework itself, it is clear that extreme values will still be an issue. On the other hand, as demonstrated, the Gaussian copula comprises known results of auto-regression and sequential kriging techniques. This suggest that the copula technique is a very flexible and no regret approach for building stochastic weather generator, even if the parallel between copula and other efficient techniques may be more obscure or only approximate if other copulas are used. [XXX-NON clarify please]

The studied area in this paper is in Cévennes-Vivarais area which is mountainous region

with several well documented data-bases for the meteorological variables. Five hydrological variables such as Wind speed, Solar radiation, Temperature, Water vapor pressure and precipitation have been chosen to build a multivariate model. We know that the atom at zero in precipitation data usually cause difficulties in copula approach. In this paper, using the meteorological context data is efficient to scrutinize dry-days and suggest and adjust a **precipitation index** that is much more continuous and easily enters the copula approach.

The diagnosis of the statistical analysis for this copula based multivariate approach are generally quite promising. The simulations have the similar average values and variability of each variable as observed data in both yearly and seasonal span. The marginal distributions of the variables have also been reproduced in the simulations. In a more complex bivariate diagnosis, the joint distributions of all pairs of the variable have been nicely represented. Even though, the choice of 3-month block for separating the seasons is not good enough in the temporal correlation diagnosis for certain variables such as temperature or solar radiation, the modification to monthly block might solve the problem, and additionally suggest further investigation of the seasonal changes in the dependence structure among variables under study.

Part IV
Conclusions and perspectives

CHAPTER 10

CONCLUSIONS

This work has been dedicated to propose new concepts and tools for stochastic weather simulation activities targeting the specific needs of hydrology. Two main contributions are:

1. to generate spatio-temporal rainfall simulations over a large heterogeneous domain;

2. to generate multivariate simulations related to precipitation, temperature, solar radiation, water vapor pressure and wind speed in a flexible way, using conditional mode on already known elements to extend the simulation in time or across variates.

Chapter 1 pointed out two main hydrological concerns which were water resources management and hydrological hazards assessment that can be served with hydrological simulations. [XXX-NON rephrase] Water resources are indispensable for human survival. But at the same time, hydrological hazards like floods often bring disaster to people. This was the starting point of this PhD work that aimed to propose a simulation framework in order to make explicit a picture of the variability of water fluxes at the entrance to river basins.

Chapter 2 introduced the Cévennes-Vivarais region as the studied region in this PhD work. The Cévennes-Vivarais region is located in a mountainous terrain which contains several different climates and topological situations. This well-documented region benefits of a long-term and qualified meteorological observation data, necessary to implement the simulations. The OHM-CV database provided precipitation data. There are 146 available rain gauge stations that record the hourly precipitation. Data from 2005 to 2014 were used to analyze rainfall situations in the Cévennes-Vivarais region. Due to the different topological features and climate conditions, the rainfall field at the Cévennes-Vivarais region can not be considered as homogeneous. By using k-means clustering method applied on hourly data, we classified the whole Cévennes-Vivarais region into relatively homogeneous rainfall zones. A classification of 4 zones was selected and preserved during the PhD work be representative of the heterogeneity problem of a rainfall field. We identified precipitation, temperature, solar radiation, water vapor pressure and wind speed as the key hydrological model inputs to serve the water balance models, so hydrologic distributed models in gen-

eral. The latter 4 hydro-meteorological variables data were selected from the ERA-Interim database. The variabilities inside the zone were not considered.

Chapter 3 identified the main challenges in this PhD work. The main objective is to enable hydrological simulations to realize various long term hydrological strategies on different spatial and temporal scales. There were two steps to achieve our objective, corresponding to two main challenges. The first main challenge to overcome was the heterogeneity of a rainfall field mentioned in Chapter 2. The spatio-temporal rainfall simulator SAMPO available at Irstea was introduced. This simulator is built under the assumption of the homogeneity of the rainfall field, while the temporal sequence is built as a succession of different rainfall types. Hence, our direct approach was to adapt SAMPO to simulate over a non-homogeneous rainfall field considered as a set of several homogeneous rainfall zones. The second main challenge was the multivariate modeling. Five hydro-meteorological variables chosen in Chapter 2 needed to be modeled simultaneously for the spatio-temporal simulations for the hydrological uses to be consistent.

Heterogeneity problem

Chapter 4 reviewed existing approaches relevant to the problem of the heterogeneity of a rainfall field. An overview table (Table 4.1) on available stochastic weather generators (SWG) was presented to distinguished SWG into different categories. Several existing methods dealing with spatial correlations of a rainfall field and their limitations were reviewed. But they were not suitable for our approach of adapting SAMPO, as SAMPO is based on homogeneous rainfall types. Therefore, approaches dealing with coordination of rainfall types calendars were considered.

Chapter 5 proposed two different approaches based on rainfall types calendars to enable a "homogeneous" rainfall generator to address rainfall simulations over a large, heterogeneous field. The first one was a parametric approach based on coupled hidden Markov model (CHMM). The CHMM was built by using the well known hidden Markov model in a hierarchical way. As a parametric model, the modeling process needs to estimate a number of parameters from Baum-Welch training algorithm and Viterbi decoding algorithm (Appendix A) for hidden Markov models. This condense the information provided by the data in a reduced number of explicit parameters, that are necessary and sufficient to simulate other instances of their observed process. The other approach was a non-parametric approach based on resampling of historical calendars of homogeneous rainfall types. Both models were able to generate a set of simulated calendars of rainfall types for the homogeneous rainfall zones.

Chapter 6 showed the statistical results of the simulations generated by SAMPO through the two approaches based on rainfall types calendars, comparing to a reference simulation built on the historical sequences of rainfall types. Simulations (called monobloc) generated by SAMPO over the whole rainfall field without the separation into several homogeneous rainfall zones were also included in the comparison of the results. The results with coordination methods were much better than monobloc simulations, as expected, confirming

that the assumption of homogeneity for the whole domain would not be realistic. The simulations of the resampling based method represented better results than those of the CHMM with the re-organization method, when comparing to simulation based on the observed calendars in term of the general statistical values (the average, the standard deviation or the maximum, etc) and the temporal correlations (dwell time for wet and dry period). However, parametric methods have a more pronounced interannual variabilities and the potential to be driven by the large scale. So both approaches have their relevance. However, the new methods share a common drawback that the inter-station correlations for stations in different zones are not correct. Given the partition of a large surface into several homogeneous zones, the spatial correlation between two zones is only conveyed by the rainfall types system. Ultimately, the real problem came from the very principle of delineating homogeneous zones and homogeneous rainfall types.

Chapter 7 proposed a new approach taking the opposite perspective to the previous two methods to overcome the heterogeneity problem. This new approach uses a copula based coordination method to generate the simulated time-series with the continuous values, and a disaggregation method to simulate small-scale rainfall field while respecting the large scale values provided by the simulated continuous type time-series. The modeling of the copula based coordination approach is made under the Gaussian framework. This is a bit restrictive choice as a copula, but allows different techniques such as the geostatistical tools, the auto-regressive process and the Gaussian copula to work together. The simulated time-series of the average daily precipitation and the daily rainfall intermittency seems correct. The diagnostic of the marginal distributions and the joint distributions showed the benefits of the copula technique when dealing with the multivariate framework. The auto-regressive process ensures the preservation of temporal correlations inside of the simulated time-series. Thereafter, reasonably simulated time-series of average daily precipitation and daily rainfall intermittency considered as large scale values could be submitted to a disaggregation model. By using a combination of block-to-point kriging and optimization search inspired by inverse modeling in hydrogeology, the rainfall fields were simulated in a fine scale. Especially using an adapted dichotomy search enable the rainfall field at small scale to respect daily average precipitation and the daily rainfall intermittency prescribed at large scale. The evident improvement of this approach is that the field is now smooth at boundaries between zones. The new approach makes the simulations much more realistic in space, while correlation in time is also preserved, if it exists. However, lack of statistical diagnosis on this new approach suggests further assessment of this new technique.

Multivariate modeling

Part III proposed a copula based multivariate model to provide relevant hydro-meteorological variables required as inputs of hydrological models. Five hydro-meteorological variables (precipitation, temperature, solar radiation, water vapor pressure and wind speed) were chosen for multivariate modeling. The daily precipitation data and the daily data of the other meteorological variables from 1989 to 2013 were selected from

the OHM-CV and ERA-Interim database, respectively. The main focus of multivariate modeling was on the inter-variable link and on the time scale, the spatial disaggregation of other variates than precipitation was not considered. Similar to the model proposed in Chapter 7 to simulate several time series of large-scale rainfall, we used the copula technique combining with the auto-regressive process and kriging technique to deal with the time series of these five hydro-meteorological variables under the Gaussian framework.

Several issues have been overcome in this part. Firstly, the precipitation index was introduced to deal with the problem that the distribution of daily rainfall total has a large proportion of days with zero rainfall. By using an artificial neural network applied on context meteorological variables, the non-zero precipitation data being used as target, the zeros in daily precipitation data could be transformed into non-zero values. In our case, the coefficient of determination between the precipitation data and the reconstructed precipitation index was more than 0.7 which was considered as a reasonable result to allow us using the precipitation index in the modeling. Secondly, since the spatio-temporal rainfall simulation could be generated as in Chapter 5 and 7, the kriging technique was introduced to simulate sequentially the other hydro-meteorological variables conditionally to the precipitation simulation. The equivalence between the Gaussian copula technique, the auto-regressive process and the simple kriging method under the Gaussian framework makes the multivariate model easy to implement.

The general statistical analysis showed the good consistency between simulated and observed datasets. The bivariate diagnostic validated the choice of the Gaussian copula in our case. If the results are globally encouraging, a main difficulty remains associated with the strong dependency of the some meteorological variables to the season. Introducing standard seasonality by 3-months blocs seems to be not enough because of the non-stationarity of the temperature and the solar radiation. The results of auto-correlation functions were improved when the interval of one season reduced to one month. So the choice of seasons is a key factor to in the time-series modeling. Actually, the statistical analysis of seasonality in inter-variable dependence must be first investigated, and also has an interest for itself.

With this, we are moving to the perspective section.

CHAPTER 11

PERSPECTIVES

This PhD work (1) simulated large, non-homogeneous rainfall field with two different methodologies, the rainfall types calendars based coordination approach and the copula-disaggregation approach; (2) generated multivariate simulations of the hydrological inputs conditioned by precipitation simulations.

Possible further developments follow.

Coupled hidden Markov model

In Chapter 10, we mentionned drawbacks of the coupled hidden Markov model (CHMM) linked to the very concept of homogeneous rainfall types. But the hidden Markov model still allows us to simulate rainfall on homogeneous rainfall zone, having key statistical properties (e.g., rainfall probabilities, dry/wet spell lengths) of the simulated rainfall that do match those of the observed rainfall records. This can be useful for generating large numbers of synthetic realizations of rainfall for input into statistical analysis. Also, the unobserved hidden states introduced by the hidden Markov model have different rainfall distributions associated with them. More, the hidden states can be coupled with the states of the atmosphere; hence it is certainly possible to drive a HMM of rainfall from classified weather types out of atmospheric runs for climate change.

Disaggregation model

The disaggregation approach was developed by the end of the PhD and certainly needs to be diagnosed more sharply using different statistical analysis.

The current model shows that the average precipitation and the rainfall intermittency of small scale respect the large scale values, as expected by design. However, spatial changes in distribution parameters could be considered as they influence the fine-scale rainfall field achieved. This is needed to better respect the local climatology.

Prescribed small scale variability has a clear impact on the disaggregation results. Despite it can be first "expert-prescribed", this suggests a detailed estimation of the small scale

variability to be conducted, as customary in any simulation; it can be based on raingauge data but could also be supplemented by rainfall radar images as these are gaining enough record to serve in climatological studies.

Multivariate model

Even though the results are generally good with the copula based multivariate model, there are still several points which need to be investigated. Firstly, in the absence of tail dependence of Gaussian copula, it is important to see whether Gaussian copula is suitable for concerned variables or should be replaced by more adequate copulas [*Furman et al.*, 2016], provided these copulas can also conveniently run in conditional and autoregressive mode. Secondly, other copula families should be tested as well to compare the differences. This kind of comparison will provide more information about the "real" multivariate distribution. Thirdly, the investigate of seasonality must be more precise. In the case of this PhD work, the 3-month interval for one season was taken for granted. Though a reduced one-month interval for one season showed a better result, it still needs to understand how to classify different seasons where the variables can be considered as stationary. Finally, in the aspect of auto-regressive process, only $AR(1)$ has been applied in this PhD work, and trying AR with higher order may be another step forward for further development.

Long term development

Looking forward, the ultimate goal will be to provide a complete procedure to generate the spatio-temporal simulations of concerned hydrometeorological variables over any reasonable large-scale studied area (e.g., large catchments).

Precipitation is generally considered as primary variable for stochastic weather generators in the statistical approach, but which variables conduct rainfall events in the real world ? The investigations on the physical interactions among the meterological variables can be useful in the multivariate models, especially when the multivariate models operate in conditional mode. Opening the set of variables concerned to other variables that the ones strictly needed, possibly including obviously physically relevant background variables, may help respect a minimum of physical sense and better prepare the stochastic models to run as a disaggregator.

For other hydro-meteorological variables, their simulations should be generated in space and in time with the same spatio-temporal resolution as the precipitation simulations. The spatio-temporal simulations of all concerned hydrometeorological variables finally could be an option, but it may be not the most urgent, if we consider the dynamic of hydrology is mostly sensitive to rainfall distribution and other factors mostly act through by their accumulated effect on evaporation.

Eventually, the simulations should be capable to produce different hydrological scenarios for the different hydrological uses with no noticeable bias when comparing to using historical data.

Part V
Appendix

APPENDIX A

HIDDEN MARKOV MODELS

Theory

Let

$$\begin{aligned}
T &= \text{length of the observation sequence;}\\
N &= \text{number of states in the model;}\\
M &= \text{number of observation symbols;}\\
S &= \{S_0, S_1, \ldots, S_{N-1}\} = \text{distinct states of the Markov process;}\\
V &= \{v_0, v_1, \ldots, v_{M-1}\} = \text{set of possible observations;}\\
\pi &= \text{state transition probabilities;}\\
\psi &= \text{observation probability matrix;}\\
\mu &= \text{initial state distribution;}\\
\mathcal{O} &= (\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1}) \text{ observation sequence;}\\
Q &= (q_0, q_1, \ldots, q_{T-1}) \text{ one possible hidden state sequence.}
\end{aligned}$$

The state transition probabilities $\pi = \{a_{ij}\}$ where

$$a_{ij} = \mathbb{P}(q_{t+1} = S_j | q_t = S_i), \quad 1 \leq i, j \leq N-1. \tag{A.1}$$

and the observation probability matrix in state j, $\psi = \{b_j(k)\}$, where

$$b_j(k) = \mathbb{P}(v_k \text{ at } t | q_t = S_j), \quad 1 \leq j \leq N-1,\ 1 \leq k \leq M-1. \tag{A.2}$$

A generic hidden Markov model is illustrated in Fig. A.1, where the $\{X_i\}$ presents the hidden state sequence and all other notations are as given above. The Markov process which is hidden behind the dashed line is determined by the current state and the transition matrix π . We are only able to observe the $\{\mathcal{O}_i\}$, which are related to the (hidden) states of the Markov process by the emission matrix ψ.

Three fundamental problems

There are three fundamental problems that we can solve using HMMs.

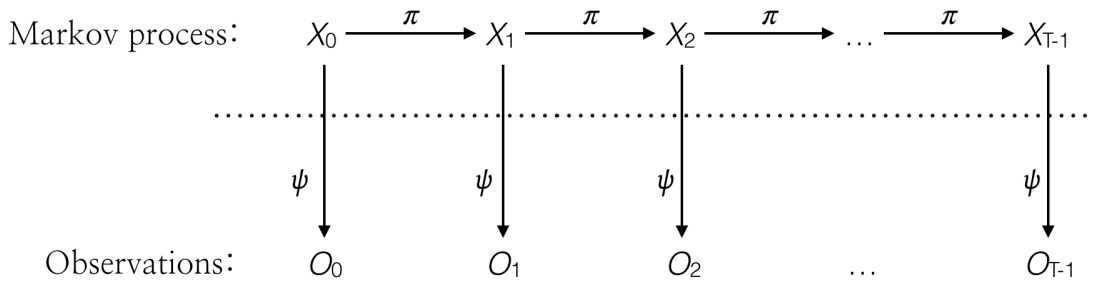


Figure A.1: Hidden Markov Model.

- **Problem 1: Evaluation**

  Given a model $\lambda = (\pi, \psi, \mu)$ and a sequence of observations $\mathcal{O}$, find $\mathbb{P}(\mathcal{O}|\lambda)$. Here, we want to determine the likelihood of the observed sequence $\mathcal{O}$, given the model.

  We can use the **forward algorithm** [*Baum et al.*, 1967, 1968] to resolve the evaluation problem. Problem 1 allows us to choose the model which best matches the observations.

- **Problem 2: Decoding**

  Given $\lambda = (\pi, \psi, \mu)$ and an observation sequence $\mathcal{O}$, find an optimal state sequence for the underlying Markov process. In other words, we want to uncover the hidden part of the Hidden Markov Model.

  We can use the **Viterbi algorithm** [*Viterbi*, 1967; *Forney*, 1973] to resolve the decoding problem. The Viterbi algorithm is a dynamic programming algorithm for finding most likely sequence of hide states. The principle of the Viterbi algorithm is to find the single best state sequence, $Q = \{q_0, q_1, \ldots, q_T\}$, for the given observation sequence $\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_T)$. For this purpose, we need to define the quantity

  $$\delta_t(i) = \max_{q_0, q_1, \cdots, q_{t-1}} \mathbb{P}(q_0, q_1, \ldots, q_t = i, \mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_t|\lambda) \tag{A.3}$$

  i.e. $\delta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state S_i. Thus, we have

  $$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \times b_j(\mathcal{O}_{t+1}). \tag{A.4}$$

- **Problem 3: Training**

  Given an observation sequence $\mathcal{O}$ and the dimensions N and M, find the model $\lambda = (\pi, \psi, \mu)$ that maximizes the probability of $\mathcal{O}$. This can be viewed as training a model to best fit the observed data. Alternatively, we can view this as a (discrete) hill climb on the parameter space represented by π, ψ and μ.

  We can use the **Baum-Welch algorithm** [*Dempster et al.*, 1977] to resolve the training problem. The Baum-Welch algorithm is used to find the unknown parameters

of a hidden Markov model. It uses the well known **EM algorithm** to find the maximum likelihood estimate of the parameters of a hidden Markov model given a set of observed feature vectors. The Baum-Welch algorithm finds a local maximum for $\lambda^* = \text{argmax}_\lambda \mathbb{P}(\mathcal{O}|\lambda)$ (i.e. the HMM parameters λ that maximize the probability of the observation).

A simple case: use HMM to generate one sequence of rainfall types

For example, the below table presents a sequence of rainfall types.

Observations	$\mathcal{O}_1$	$\mathcal{O}_2$	$\mathcal{O}_3$	...	$\mathcal{O}_T$
Type of rainfall	3	3	5	...	1

For a hidden Markov model, we need three things which are the number of the hidden states, a initial state transition probabilities π_{ini} and a initial observation probability matrix ψ_{ini}.
Once we get a initial hidden Markov model, we use **Baum Welch algorithm** to improve the model. The Baum Welch algorithm optimizes the model parameters so as to best describe how a given observation sequence comes about. It allows us to optimally adapt model parameters to observed training data i.e. to create best models for real phenomena. Then, another important step is to find the "correct" state sequence. This is the decoding problem which we mentioned before. The **Viterbi algorithm** can solve this problem as best as possible. But we must point out here that the **Viterbi algorithm** is not the only solution for the decoding problem. After the training step for finding the best model with given observation and the fixed number of states, and the decoding step for finding the correct state sequence with given observation and given model, we can generate observation sequence with a hidden Markov model and get a correct state sequence.

If the number of observation symbols is M and the number of state is N, we can obtain a transition matrix with the dimension of $N \times N$ for the states, a emission matrix with the dimension of $N \times M$ between the states and observations, and also a state sequence of the same length of observation sequence. These are three important elements when we construct the coupled hidden Markov model in Section 5.2.2

APPENDIX B

SELF-ORGANIZING MAP

Definition

The Self-Organizing Map (SOM), commonly known as Kohonen network introduced by the Finnish professor *Kohonen* [1982] is a computational method for the visualization and analysis of high-dimensional data. The self-organizing map is a type of artificial neural network (Appendix D) that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as back-propagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space. There are two modes when operating SOM.

- Training mode is to build the map using input examples (a competitive process, also called vector quantization).

- Mapping mode is to automatically classify a new input vector.

Same as artificial neural network, a self-organizing map consists of components called nodes or neurons. Associated with each node are a weight vector of the same dimension as the input data vectors, and a position in the map space. The usual arrangement of nodes can be defined to be rectangular, hexagonal or even irregular; hexagonal is effective for visual display. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

Learning Algorithm

SOM mapping steps starts from initializing the weight vectors. From there a sample vector is selected randomly and the map of weight vectors is searched to find which weight best represents that sample. Each weight vector has neighboring weights that are close to

it. The weight that is chosen is rewarded by being able to become more like that randomly selected sample vector. The neighbors of that weight are also rewarded by being able to become more like the chosen sample vector. From this step the number of neighbors and how much each weight can learn decreases over time. The whole process is repeated a large number of times. The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The neuron whose weight vector is most similar to the input is called the Best Matching Unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance (within the lattice) from the BMU. Kohonen learning uses a neighborhood function φ, whose value $\varphi(i,k)$ represents the strength of the coupling between unit (or neuron) i and the BMU k during the training process. A simple choice is defining $\varphi(i,k) = 1$ for all units i in a neighborhood of radius r of unit k and $\varphi(i,k) = 0$ for all other units. Regardless of the functional form, the neighborhood function shrinks with time. At the beginning when the neighborhood is broad, the self-organizing takes place on the global scale. When the neighborhood has shrunk to just a couple of neurons, the weights are converging to local estimates.

Algorithm

start : The n-dimensional weight vectors $w_1, w_2, \ldots, w_m$ of the m computing units are selected at random. An initial radius r, a learning constant η, and a neighborhood function φ are selected.

step 1 : Select an input vector ξ using the desired probability distribution over the input space.

step 2 : The unit k with the maximum excitation is selected (that is, for which the distance between w_i and ξ is minimal, for $i = 1, \ldots, m$).

step 3 : The weight vectors are updated using the neighborhood function and the update rule

$$w_i \leftarrow w_i + \eta\varphi(i,k)(\xi - w_i), \quad i = 1, \ldots, m. \tag{B.1}$$

step 4 : Stop if the maximum number of iterations has been reached; otherwise modify η and φ as scheduled and continue with step 1.

In sum, the training mode occurs in several steps and over many iterations.

1. Each unit's weights are initialized.
2. A vector is chosen at random from the set of training data.
3. Every unit is examined to calculate which one's weights are most like the input vector. The winning unit is selected and called BMU.

4. Then the neighborhood of the BMU is calculated. The amount of neighbors decreases over time.

5. The winning weight is rewarded with becoming more like the sample vector. The neighbors also become more like the sample vector. The closer a unit is to the BMU, the more its weights get altered and the farther away the neighbor is from the BMU, the less it learns.

6. Repeat step 2 for N iterations.

APPENDIX C

COPULA

In probability theory and statistics, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables. They are named for their resemblance to grammatical copulas in linguistics.

Definition

A definition of copula can be found in *Nelsen* [2007] which is a main reference. A copula is defined as a distribution function on the n-dimensional unit cube. All marginal distributions are uniform:

$$\begin{aligned} &C : [0,1]^n \longrightarrow [0,1] \\ &C(u) = u_i \qquad \text{if the vector } u = (1,\ldots,1,u_i,1,\ldots,1). \end{aligned} \tag{C.1}$$

For every n dimensional hyper-cube within the unit hyper-cube, the corresponding probability has to be non-negative. Copulas and multivariate distributions are linked to each other. Each multivariate distribution $F(x_1,\ldots,x_n)$ can be represented with the help of a copula:

$$F(x_1,\ldots,x_n) = C(F_{x_1}(x_1),\ldots,F_{x_n}(x_n)). \tag{C.2}$$

where $F_{x_i}(x_i)$ represents the i-th one-dimensional marginal distribution of the multivariate distribution. If the distribution is continuous then the copula C is unique. Copulas can be constructed from distribution functions:

$$C(u) = C(u_1,\ldots,u_n) = F(F_{x_1}^{-1}(x_1),\ldots,F_{x_n}^{-1}(x_n)) \tag{C.3}$$

Theory

Abe Sklar's theorem [*Sklar*, 1959], provides the theoretical foundation for the application of copulas.

$$f_X(x_1,\ldots,x_n) = c_X(F_{X_1}(x_1),\ldots,F_{X_n}(x_n)) \cdot f_1(x_1)\cdots f_n(x_n), \tag{C.4}$$

where c_X is the density of the copula for X.

Sklar's Theorem states that any multivariate joint distribution can be written in terms of uni-variate marginal distribution functions and a copula which describes the dependence structure between the variables.

The advantages that can be described to using copula densities to represent interdependence between variables include the following properties:

- The empirical copulas (probability density scatter-plots in many dimensions) are independent of their corresponding marginal distributions, so that copulas display interdependence between variables in its purest or essential form;
- Empirical copulas are easily computed from data;
- Differences in types of association between variables are readily identified by copula shape;
- A suite of theoretical copula density functions has been developed to model these attributes.

The copula technique is used to preserve the rank-correlations for each pair variables. Because of the Sklar theorem, the construction of copula will not affect the marginal distribution of each variable. Yet, the copula approach doesn't deal with the seasonality problem. The copula technique does not presume where these variables are, why they are different, and if they are different by nature, by measurement technique, by support, or by place in space or time.

APPENDIX D

ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is a computational model that is inspired by the way biological neural networks in the human brain process information. Artificial Neural Networks have generated a lot of excitement in Machine Learning research and industry, thanks to many breakthrough results in speech recognition, computer vision and text processing. [*French et al.*, 1992] uses a neural network to forecast rainfall in space and time. Artificial neural network has been the core of data learning and deep learning science in recent years. The advantages of artificial neural network are which

1. ANN is nonlinear model that is easy to use and understand compared to statistical methods.

2. ANN is non-parametric model that is not concerned with dimensionality.

3. ANN with back propagation learning algorithm is widely used in solving various classification and forecasting problems.

However, ANN is black box learning approach, can not interpret the relationship between input and output and can not deal with uncertainties.

A single Neuron

The basic unit of computation in a neural network is the neuron, often called a node or unit. It receives input from some other nodes, or from an external source and computes an output. Each input has an associated weight (w), which is assigned on the basis of its relative importance to other inputs. The node applies a function f (defined below) to the weighted sum of its inputs as shown in Fig. D.1.

The above network takes numerical inputs $X1$ and $X2$ and has weights $w1$ and $w2$ associated with those inputs. Additionally, there is another input 1 with weight b (called the Bias) associated with it. The output Y from the neuron is computed as shown in the Fig. D.1.

$$Y = f(w_1 \times x_1 + w_2 \times x_2 + b) \tag{D.1}$$

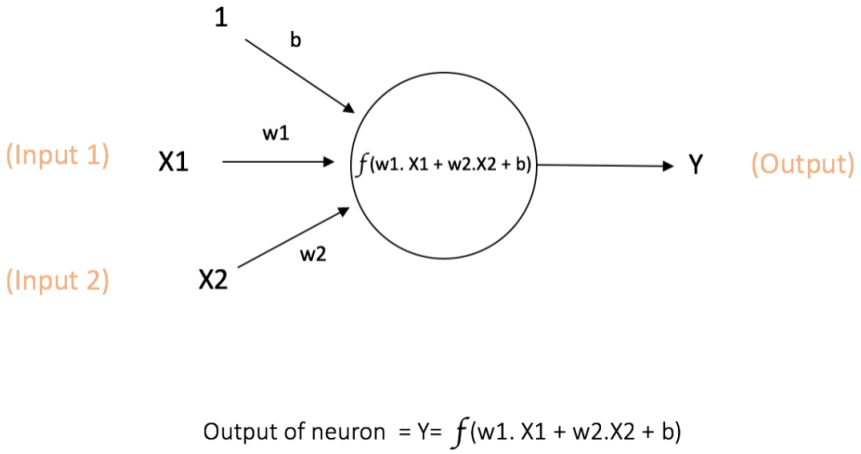


Figure D.1: A single neuron.

where the function f is non-linear and is called the **Activation Function**. The purpose of the activation function is to introduce non-linearity into the output of a neuron. This is important because most real world data is non linear and we want neurons to learn these non linear representations. There are several activation functions that one may use in practice.

- **Sigmoid**: takes a real-valued input and squashes it to range between 0 and 1

$$\sigma(x) = 1/(1 + \exp(-x)) \tag{D.2}$$

- **tanh**: takes a real-valued input and squashes it to the range $[-1, 1]$

$$\tanh(x) = 2\sigma(2x) - 1 \tag{D.3}$$

Figure D.2 shows each of the above activation functions.

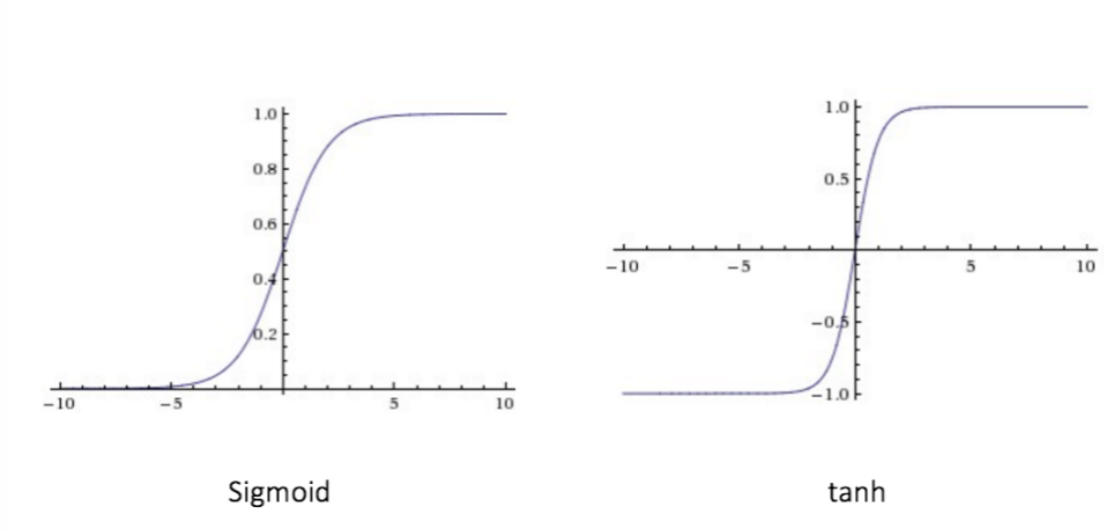


Figure D.2: Different activation functions.

Importance of Bias: The main function of Bias is to provide every node with a trainable constant value (in addition to the normal inputs that the node receives).

Feedforward Neural Network

The feedforward neural network was the first and simplest type of artificial neural network devised. It contains multiple neurons (nodes) arranged in layers. Nodes from adjacent layers have connections or edges between them. All these connections have weights associated with them. An example of a feedforward neural network is shown in Fig. D.3.

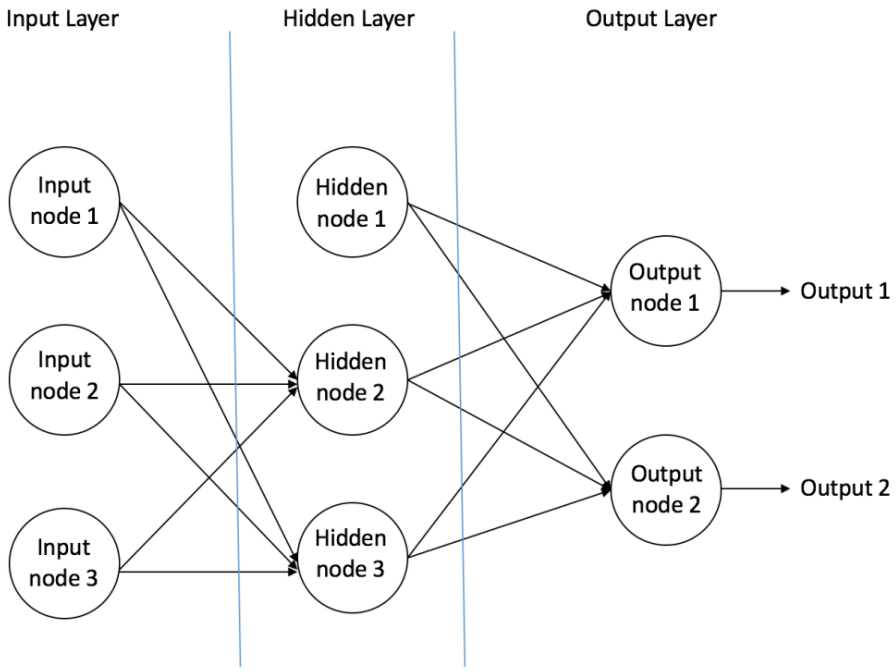


Figure D.3: An example of articificial neural network.

A feedforward neural network can consist of three types of nodes:

1. **Input Nodes** - The Input nodes provide information from the outside world to the network and are together referred to as the "Input Layer". No computation is performed in any of the Input nodes - they just pass on the information to the hidden nodes.

2. **Hidden Nodes** - The Hidden nodes have no direct connection with the outside world (hence the name "hidden"). They perform computations and transfer information from the input nodes to the output nodes. A collection of hidden nodes forms a "Hidden Layer". While a feedforward network will only have a single input layer and a single output layer, it can have zero or multiple Hidden Layers.

3. **Output Nodes** - The Output nodes are collectively referred to as the "Output Layer" and are responsible for computations and transferring information from the network to the outside world.

In a feedforward network, the information moves in only one direction - forward - from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network (this property of feed forward networks is different from Recurrent Neural Networks in which the connections between the nodes form a cycle).

BIBLIOGRAPHY

Adamovic, M., F. Branger, I. Braud, and S. Kralisch (2016), Development of a data-driven semi-distributed hydrological model for regional scale catchments prone to Mediterranean flash floods, *Journal of Hydrology*, *541*(Part A), 173–189, doi:10.1016/j.jhydrol.2016.03.032.

Ailliot, P., C. Thompson, and P. Thomson (2009), Space–time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *58*(3), 405–426, doi:10.1111/j.1467-9876.2008.00654.x.

Ailliot, P., D. Allard, V. Monbet, and P. Naveau (2015), Stochastic weather generators: an overview of weather type models, *Journal de la Société Française de Statistique*, *156*(1), 101–113.

Anderson, T. W. (1958), *An introduction to multivariate statistical analysis*, vol. 2, Wiley New York.

Anquetin, S., F. Minsicloux, J.-D. Creutin, and S. Cosma (2003), Numerical simulation of orographic rainbands, *Journal of Geophysical Research: Atmospheres*, *108*(D8), doi:10.1029/2002JD001593, 8386.

Baigorria, G. A., J. W. Jones, and J. J. O'Brien (2007), Understanding rainfall spatial variability in southeast USA at different timescales, *International Journal of Climatology*, *27*(6), 749–760, doi:10.1002/joc.1435.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014), *Hierarchical modeling and analysis for spatial data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.

Barbu, V., and N. Limnios (2008), Hidden Semi-Markov Model and Estimation, in *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*, *Lecture Notes in Statistics*, vol. 191, pp. 1–48, Springer New York, doi:10.1007/978-0-387-73173-5_6.

Bárdossy, A., and J. Li (2008), Geostatistical interpolation using copulas, *Water Resources Research*, *44*(7), doi:10.1029/2007WR006115.

Bárdossy, A., and G. G. S. Pegram (2009), Copula based multisite model for daily precipitation simulation, *Hydrology and Earth System Sciences*, *13*(12), 2299–2314, doi:10.5194/hess-13-2299-2009.

Bárdossy, A., and E. J. Plate (1991), Modeling daily rainfall using a semi-Markov representation of circulation pattern occurrence, *Journal of Hydrology*, *122*(1), 33–47, doi:10.1016/0022-1694(91)90170-M.

Bárdossy, A., and E. J. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resources Research*, *28*(5), 1247–1259, doi:10.1029/91WR02589.

Barkstrom, B. R. (1990), Earth radiation budget measurements: pre-ERBE, ERBE, and CERES, in *SPIE 1299, Long-Term Monitoring of the Earth's Radiation Budget*, vol. 1299, pp. 52–60, doi:10.1117/12.21364.

Baum, L. E., and T. Petrie (1966), Statistical Inference for Probabilistic Functions of Finite State Markov Chains, *The Annals of Mathematical Statistics*, *37*(6), 1554–1563.

Baum, L. E., J. A. Eagon, et al. (1967), An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, *Bull. Amer. Math. Soc*, *73*(3), 360–363, doi:10.1090/S0002-9904-1967-11751-8.

Baum, L. E., G. R. Sell, et al. (1968), Growth transformations for functions on manifolds, *Pacific Journal of Mathematics*, *27*(2), 211–227.

Baxevani, A., and J. Lennartsson (2015), A spatiotemporal precipitation generator based on a censored latent Gaussian field, *Water Resources Research*, *51*(6), 4338–4358, doi:10.1002/2014WR016455.

Bayraktar, H., and F. S. Turalioglu (2005), A Kriging-based approach for locating a sampling site—in the assessment of air quality, *Stochastic Environmental Research and Risk Assessment*, *19*(4), 301–305, doi:10.1007/s00477-005-0234-8.

Boudevillain, B., G. Delrieu, B. Galabertier, L. Bonnifait, L. Bouilloud, P.-E. Kirstetter, and M.-L. Mosini (2011), The Cévennes-Vivarais Mediterranean Hydrometeorological Observatory database, *Water Resources Research*, *47*(7), doi:10.1029/2010WR010353.

Bouvier, C., L. Cisneros, R. Dominguez, J.-P. Laborde, and T. Lebel (2003), Generating rainfall fields using principal components (PC) decomposition of the covariance matrix: a case study in Mexico City, *Journal of Hydrology*, *278*(1), 107 – 120, doi:10.1016/S0022-1694(03)00122-7.

Brand, M. (1997), Coupled hidden Markov models for modeling interacting processes, *Tech. rep.*, The Media Lab, MIT.

Bras, R. L., and I. Rodríguez-Iturbe (1976), Rainfall generation: A nonstationary time-varying multidimensional model, *Water Resources Research*, *12*(3), 450–456, doi:10.1029/WR012i003p00450.

Braud, I., H. Roux, S. Anquetin, M.-M. Maubourguet, C. Manus, P. Viallet, and D. Dartus (2010), The use of distributed hydrological models for the Gard 2002 flash flood event: Analysis of associated hydrological processes, *Journal of Hydrology*, *394*(1), 162–181, doi:10.1016/j.jhydrol.2010.03.033, flash Floods: Observations and Analysis of Hydrometeorological Controls.

Braud, I., P. Ayral, C. Bouvier, F. Branger, G. Delrieu, G. Dramais, J. Le Coz, E. Leblois, G. Nord, and J. Vandervaere (2016), Advances in flash floods understanding and modelling derived from the FloodScale project in south-east France, in *3rd European Conference on Flood Risk Management, Innovation, Implementation, Integration (FLOODrisk 2016)*, vol. 7, p. 04005, Lyon, France, doi:10.1051/e3sconf/20160704005.

Breinl, K., T. Turkington, and M. Stowasser (2015), Simulating daily precipitation and temperature: a weather generation framework for assessing hydrometeorological hazards, *Meteorological Applications*, *22*(3), 334–347, doi:10.1002/met.1459.

Buishand, T. A., L. de Haan, and C. Zhou (2008), On spatial extremes: with application to a rainfall problem, *The Annals of Applied Statistics*, *2*(2), 624–642.

Certes, C., and G. de Marsily (1991), Application of the pilot point method to the identification of aquifer transmissivities, *Advances in Water Resources*, *14*(5), 284–300, doi:10.1016/0309-1708(91)90040-U.

Charles, S. P., B. C. Bates, and J. P. Hughes (1999), A spatiotemporal model for downscaling precipitation occurrence and amounts, *Journal of Geophysical Research: Atmospheres*, *104*(D24), 31,657–31,669, doi:10.1029/1999JD900119.

Chen, L., V. P. Singh, S. Guo, J. Zhou, and J. Zhang (2015), Copula-based method for multisite monthly and daily streamflow simulation, *Journal of Hydrology*, *528*(Supplement C), 369–384, doi:10.1016/j.jhydrol.2015.05.018.

Chhikara, R. S., and J. L. Folks (1974), Estimation of the Inverse Gaussian Distribution Function, *Journal of the American Statistical Association*, *69*(345), 250–254, doi:10.1080/01621459.1974.10480165.

Chilès, J.-P., and P. Delfiner (2008a), *Geostatistics: Modeling Spatial Uncertainty*, vol. 497, John Wiley & Sons, doi:10.1002/9780470316993.

Chilès, J.-P., and P. Delfiner (2008b), *Intrinsic Model of Order k*, chap. 4, pp. 231–291, John Wiley & Sons, Inc., doi:10.1002/9780470316993.ch4.

Christakos, G. (2012), *Random field models in earth sciences*, Courier Corporation.

Christensen, J. H., B. Hewitson, A. Busuioc, A. Chen, X. Gao, R. Held, R. Jones, R. K. Kolli, W. Kwon, R. Laprise, et al. (2007), Regional climate projections, in *Climate Change, 2007: The Physical Science Basis. Contribution of Working group I to the Fourth Assessment Report of*

the Intergovernmental Panel on Climate Change, University Press, Cambridge, Chapter 11, pp. 847–940.

Cox, D. R., and V. Isham (1988), A Simple Spatial-Temporal Model of Rainfall, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *415*(1849), 317–328, doi:10.1098/rspa.1988.0016.

Cressie, N. (1992), Statistics for spatial data, *Terra Nova*, *4*(5), 613–617, doi:10.1111/j.1365-3121.1992.tb00605.x.

Creutin, J.-D., E. Leblois, and J.-M. Lepioufle (2015), Unfreezing Taylor's hypothesis for precipitation, *Journal of Hydrometeorology*, *16*(6), 2443–2462, doi:10.1175/JHM-D-14-0120.1.

de Marsily, G., J.-P. Delhomme, F. Delay, and A. Buoro (1999), Regards sur 40 ans de problèmes inverses en hydrogéologie, *Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science*, *329*(2), 73–87, doi:10.1016/S1251-8050(99)80208-0.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597, doi:10.1002/qj.828.

Delrieu, G. (2003), L'Observatoire Hydro-météorologique Méditerranéen Cévennes-Vivarais, *La Houille Blanche*, (6), 83–88, doi:10.1051/lhb/2003116.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Devia, G. K., B. Ganasri, and G. Dwarakish (2015), A Review on Hydrological Models, *Aquatic Procedia*, *4*(Supplement C), 1001–1007, doi:10.1016/j.aqpro.2015.02.126.

Domingues-Ramos, M. H. (2002), Analyse de la pluviométrie sous des systèmes nuages convectifs: Etude des cas sur des données de la ville de Marseille et de la méthode ISIS de Météo-France, Ph.D. thesis, Université Joseph Fourier.

Efron, B. (1982), *The jackknife, the bootstrap and other resampling plans*, vol. 38, SIAM, doi:10.1137/1.9781611970319.fm.

Erhardt, T. M., C. Czado, and U. Schepsmeier (2015), R-vine models for spatial time series with an application to daily mean temperature, *Biometrics*, *71*(2), 323–332, doi:10.1111/biom.12279.

Evin, G., A.-C. Favre, and B. Hingray (2017), Stochastic generation of multi-site daily precipitation for the assessment of extreme floods in Switzerland, doi:10.5194/hess-2017-226.

FAO (2003), *Review of World Water Resources by Country*, no. 23 in Water Report, Rome: Food and Agriculture Organization of the United Nations.

Fatichi, S., E. R. Vivoni, F. L. Ogden, V. Y. Ivanov, B. Mirus, D. Gochis, C. W. Downer, M. Camporese, J. H. Davison, B. Ebel, N. Jones, J. Kim, G. Mascaro, R. Niswonger, P. Restrepo, R. Rigon, C. Shen, M. Sulis, and D. Tarboton (2016), An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, *Journal of Hydrology*, *537*, 45–60, doi:10.1016/j.jhydrol.2016.03.026.

Flecher, C., P. Naveau, D. Allard, and N. Brisson (2010), A stochastic daily weather generator for skewed data, *Water Resources Research*, *46*(7), doi:10.1029/2009WR008098.

Forney, G. D., Jr. (1973), The Viterbi algorithm, *Proceedings of the IEEE*, *61*(3), 268–278, doi:10.1109/PROC.1973.9030.

Foufoula-Georgiou, E., and D. P. Lettenmaier (1987), A Markov Renewal Model for rainfall occurrences, *Water Resources Research*, *23*(5), 875–884, doi:10.1029/WR023i005p00875.

French, M. N., W. F. Krajewski, and R. R. Cuykendall (1992), Rainfall forecasting in space and time using a neural network, *Journal of Hydrology*, *137*(1–4), 1–31, doi:10.1016/0022-1694(92)90046-X.

Furman, E., A. Kuznetsov, J. Su, and R. Zitikis (2016), Tail dependence of the Gaussian copula revisited, *Insurance: Mathematics and Economics*, *69*(Supplement C), 97–103, doi:10.1016/j.insmatheco.2016.04.009.

Galton, F. (1886), Regression Towards Mediocrity in Hereditary Stature., *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263, doi:10.2307/2841583.

Geary, R. C. (1954), The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, *5*(3), 115–146, doi:10.2307/2986645.

Genest, C., and A.-C. Favre (2007), Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask, *Journal of Hydrologic Engineering*, *12*(4), 347–368, doi:10.1061/(ASCE)1084-0699(2007)12:4(347).

Georgakakos, K. P., and M. L. Kavvas (1987), Precipitation analysis, modeling, and prediction in hydrology, *Reviews of Geophysics*, *25*(2), 163–178, doi:10.1029/RG025i002p00163.

Gleeson, T., Y. Wada, M. F. Bierkens, and L. P. van Beek (2012), Water balance of global aquifers revealed by groundwater footprint, *Nature*, *488*(7410), 197–200, doi:10.1038/nature11295.

Gleick, P. H. (1996), Basic Water Requirements for Human Activities: Meeting Basic Needs, *Water International*, *21*(2), 83–92, doi:10.1080/02508069608686494.

Gleick, P. H., and C. W. Howe (1995), Water in Crisis: A Guide to the World's Fresh Water Resources, *Climatic Change*, *31*(1), 119–122.

Godart, A., S. Anquetin, E. Leblois, and J.-D. Creutin (2011), The contribution of orographically driven banded precipitation to the rainfall climatology of a Mediterranean region, *Journal of Applied Meteorology and Climatology*, *50*(11), 2235–2246, doi:10.1175/JAMC-D-10-05016.1.

Gupta, V. K., and E. Waymire (1990), Multiscaling properties of spatial rainfall and river flow distributions, *Journal of Geophysical Research: Atmospheres*, *95*(D3), 1999–2009, doi:10.1029/JD095iD03p01999.

Guttorp, P., and P. D. Sampson (1994), 20 Methods for estimating heterogeneous spatial covariance functions with environmental applications, in *Environmental Statistics*, *Handbook of Statistics*, vol. 12, edited by G. Patil and C. Rao, pp. 661–689, Elsevier, doi:10.1016/S0169-7161(05)80022-7.

Hamed, K. (2009a), Exact distribution of the Mann–Kendall trend test statistic for persistent data, *Journal of Hydrology*, *365*(1), 86–94, doi:10.1016/j.jhydrol.2008.11.024.

Hamed, K. H. (2009b), Effect of persistence on the significance of Kendall's tau as a measure of correlation between natural time series, *The European Physical Journal Special Topics*, *174*(1), 65–79, doi:10.1140/epjst/e2009-01090-x.

Hao, Z., and V. P. Singh (2016), Review of dependence modeling in hydrology and water resources, *Progress in Physical Geography*, *40*(4), 549–578, doi:10.1177/0309133316632460.

Hingray, B. (2003), Multisite and space-time rainfall models: a review, *Tech. Rep. 50*, Laboratory of Hydrology and Land Improvement, EPFL, Lausanne.

Hourdin, F., I. Musat, S. Bony, P. Braconnot, F. Codron, J.-L. Dufresne, L. Fairhead, M.-A. Filiberti, P. Friedlingstein, J.-Y. Grandpeix, G. Krinner, P. LeVan, Z.-X. Li, and F. Lott (2006), The LMDZ4 general circulation model: climate performance and sensitivity to parametrized physics with emphasis on tropical convection, *Climate Dynamics*, *27*(7), 787–813, doi:10.1007/s00382-006-0158-0.

Hughes, J. P., P. Guttorp, and S. P. Charles (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(1), 15–30, doi:10.1111/1467-9876.00136.

Huntington, T. G. (2006), Evidence for intensification of the global water cycle: Review and synthesis, *Journal of Hydrology*, *319*(1), 83–95, doi:10.1016/j.jhydrol.2005.07.003.

Intergovernmental Panel on Climate Change (2014), *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, doi:10.1017/CBO9781107415324.

Johnson, N. L. (1949), Bivariate Distributions Based on Simple Translation Systems, *Biometrika*, *36*(3–4), 297–304, doi:10.2307/2332669.

Karl, T. R., and K. E. Trenberth (2003), Modern Global Climate Change, *Science*, *302*(5651), 1719–1723, doi:10.1126/science.1090228.

Katz, R. W. (1977), Precipitation as a chain-dependent process, *Journal of Applied Meteorology*, *16*(7), 671–676, doi:10.1175/1520-0450(1977)016<0671:PAACDP>2.0.CO;2.

Kendall, M. G. (1938), A New Measure of Rank Correlation, *Biometrika*, *30*(1/2), 81–93, doi:10.2307/2332226.

Kerry, R., P. Goovaerts, B. G. Rawlins, and B. P. Marchant (2012), Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale, *Geoderma*, *170*(Supplement C), 347–358, doi:10.1016/j.geoderma.2011.10.007.

Khalili, M., R. Leconte, and F. Brissette (2007), Stochastic multisite generation of daily precipitation data using spatial autocorrelation, *Journal of hydrometeorology*, *8*(3), 396–412, doi:10.1175/JHM588.1.

Kilsby, C., P. Jones, A. Burton, A. Ford, H. Fowler, C. Harpham, P. James, A. Smith, and R. Wilby (2007), A daily weather generator for use in climate change studies, *Environmental Modelling & Software*, *22*(12), 1705–1719, doi:10.1016/j.envsoft.2007.02.005.

Kleiber, W., R. W. Katz, and B. Rajagopalan (2012), Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes, *Water Resources Research*, *48*(1), doi:10.1029/2011WR011105, w01523.

Koch, E., and P. Naveau (2015), A frailty-contagion model for multi-site hourly precipitation driven by atmospheric covariates, *Advances in Water Resources*, *78*, 145–154, doi:10.1016/j.advwatres.2015.01.001.

Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, *43*(1), 59–69, doi:10.1007/BF00337288.

Kohonen, T. (2001), *Self-Organizing Maps*, *Springer Series in Information Sciences*, vol. 30, 3rd ed., Springer Berlin Heidelberg, doi:10.1007/978-3-642-56927-2.

Kruskal, W. H. (1958), Ordinal Measures of Association, *Journal of the American Statistical Association*, *53*(284), 814–861, doi:10.2307/2281954.

Kullback, S., and R. A. Leibler (1951), On Information and Sufficiency, *The Annals of Mathematical Statistics*, *22*(1), 79–86, doi:10.1214/aoms/1177729694.

Labbas, M. (2015), Modélisation hydrologique de bassins versants périurbains et influence de l'évolution de l'occupation du sol et de la gestion des eaux pluviales-Application au bassin de l'Yzeron (130 km2), Ph.D. thesis, Irstea Lyon-Villeurbanne.

Lafore, J. P., J. Stein, N. Asencio, P. Bougeault, V. Ducrocq, J. Duron, C. Fischer, P. Héreil, P. Mascart, V. Masson, J. P. Pinty, J. L. Redelsperger, E. Richard, and J. V.-G. de Arellano (1997), The Meso-NH Atmospheric Simulation System. Part I: adiabatic formulation and control simulations, *Annales Geophysicae*, *16*(1), 90–109, doi:10.1007/s00585-997-0090-6.

Lang, M., and D. Coeur (2014), *Les inondations remarquables au XXe siècle: Inventaire 2011 pour la directive Inondation*, Editions Quae.

Lebeaupin, C., V. Ducrocq, and H. Giordani (2006), Sensitivity of torrential rain events to the sea surface temperature based on high-resolution numerical forecasts, *Journal of Geophysical Research: Atmospheres*, *111*(D12), doi:10.1029/2005JD006541, d12110.

Leblois, E. (2012), Le bassin versant, système spatialement structuré et soumis au climat, HDR Dissertation.

Leblois, E. (2014), SWG - Irstea's activity report for 2014, *Tech. rep.*, SINTEF, Norway.

Leblois, E., and J.-D. Creutin (2013), Space-time simulation of intermittent rainfall with prescribed advection field: Adaptation of the turning band method, *Water Resources Research*, *49*(6), 3375–3387, doi:10.1002/wrcr.20190.

Legesse, D., C. Vallet-Coulomb, and F. Gasse (2003), Hydrological response of a catchment to climate and land use changes in Tropical Africa: case study South Central Ethiopia, *Journal of Hydrology*, *275*(1), 67–85, doi:10.1016/S0022-1694(03)00019-2.

Lennartsson, J., A. Baxevani, and D. Chen (2008), Modelling precipitation in Sweden using multiple step markov chains and a composite model, *Journal of Hydrology*, *363*(1–4), 42–59, doi:10.1016/j.jhydrol.2008.10.003.

Lepioufle, J.-M. (2009), Modélisation spatio-temporelle d'un champ de pluie: application aux pluies journalières du bassin versant de la Loire, Ph.D. thesis, Grenoble, INPG.

Lepioufle, J.-M., E. Leblois, and J.-D. Creutin (2012), Variography of rainfall accumulation in presence of advection, *Journal of Hydrology*, *464–465*, 494–504, doi:10.1016/j.jhydrol.2012.07.041.

Loaiciga, H. A., J. Michaelsen, and P. F. Hudak (1992), Truncated distributions in hydrologic analysis, *JAWRA Journal of the American Water Resources Association*, *28*(5), 853–863, doi:10.1111/j.1752-1688.1992.tb03187.x.

MacKay, D. J. (2003), *Information theory, inference and learning algorithms*, Cambridge university press.

MacQueen, J., et al. (1967), Some methods for classification and analysis of multivariate observations, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA.

Maidment, D. R., et al. (1993), *Handbook of hydrology*, vol. 1, McGraw-Hill New York.

Mantoglou, A., and J. L. Wilson (1982), The Turning Bands Method for simulation of random fields using line generation by a spectral method, *Water Resources Research*, *18*(5), 1379–1394, doi:10.1029/WR018i005p01379.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005), Bayesian Modelling and Inference on Mixtures of Distributions, in *Bayesian Thinking Modeling and Computation*, *Handbook of Statistics*, vol. 25, edited by D. Dey and C. Rao, pp. 459–507, Elsevier, doi:10.1016/S0169-7161(05)25016-2.

Matalas, N. C. (1967), Mathematical assessment of synthetic hydrology, *Water Resources Research*, *3*(4), 937–945, doi:10.1029/WR003i004p00937.

Matheron, G. (1963), Principles of geostatistics, *Economic geology*, *58*(8), 1246–1266.

Matheron, G. (1973), The intrinsic random functions and their applications, *Advances in applied probability*, *5*(3), 439–468.

Mazur, A. E., and V. I. Piterbarg (2015), Gaussian copula time series with heavy tails and strong time dependence, *Moscow University Mathematics Bulletin*, *70*(5), 197–201, doi:10.3103/S0027132215050010.

McLachlan, G., and D. Peel (2004), *Finite mixture models*, John Wiley & Sons.

Molinié, G., D. Ceresetti, S. Anquetin, J. D. Creutin, and B. Boudevillain (2012), Rainfall Regime of a Mountainous Mediterranean Region: Statistical Analysis at Short Time Steps, *Journal of Applied Meteorology and Climatology*, *51*(3), 429–448, doi:10.1175/2011JAMC2691.1.

Musy, A., and C. Higy (2010), *Hydrology: A Science of Nature*, CRC Press.

Musy, A., B. Hingray, and C. Picouet (2014), *Hydrology: a science for engineers*, CRC Press.

Nace, R. L. (1967), Are we running out of water?, *Tech. rep.*, US Geological Survey.

Natarajan, P., and R. Nevatia (2007), Coupled Hidden Semi Markov Models for Activity Recognition, in *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, p. 10, doi:10.1109/WMVC.2007.12.

Nefian, A., L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy (2002), A coupled HMM for audio-visual speech recognition, in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, pp. II–2013–II–2016, doi:10.1109/ICASSP.2002.5745027.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Nuissier, O., V. Ducrocq, D. Ricard, C. Lebeaupin, and S. Anquetin (2008), A numerical study of three catastrophic precipitating events over southern France. I: Numerical framework and synoptic ingredients, *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 111–130, doi:10.1002/qj.200.

Nuissier, O., B. Joly, A. Joly, V. Ducrocq, and P. Arbogast (2011), A statistical downscaling to identify the large-scale circulation patterns associated with heavy precipitation events over southern France, *Quarterly Journal of the Royal Meteorological Society*, *137*(660), 1812–1827, doi:10.1002/qj.866.

Olea, R. A. (2012), *Geostatistics for engineers and earth scientists*, Springer Science & Business Media.

Ollagnier, M. (2013), Climatologie des pluies en région Cévennes-Vivarais : caractérisation des situations pluvieuses à l'échelle régionale, Master's thesis, Université Joseph Fourier, LTHE.

Oriani, F., J. Straubhaar, P. Renard, and G. Mariethoz (2014), Simulation of rainfall time series from different climatic regions using the direct sampling technique, *Hydrology and Earth System Sciences*, *18*(8), 3015–3031, doi:10.5194/hess-18-3015-2014.

Penman, H. L. (1948), Natural evaporation from open water, bare soil and grass, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *193*(1032), 120–145, doi:10.1098/rspa.1948.0037.

Rabiner, L. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, *77*(2), 257–286, doi:10.1109/5.18626.

Racsko, P., L. Szeidl, and M. Semenov (1991), A serial approach to local stochastic weather models, *Ecological Modelling*, *57*(1–2), 27–41, doi:10.1016/0304-3800(91)90053-4.

Renard, B., E. Leblois, J. Le Coz, L. Bonnifait, F. Branger, M.-H. Ramos, T. Van Pham, C. Perrin, V. Andressian, M. Thyer, D. Kavetski, G. Kuczera, and G. Evin (2011), Quantifying uncertainties in hydrologic prediction, with application to flood forecasting, *Tech. rep.*, Irstea Lyon, Iresta Antony, University of Adelaide and University of Newcastle.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resources Research*, *17*(1), 182–190, doi:10.1029/WR017i001p00182.

Richardson, C. W., and D. A. Wright (1984), *WGEN: A model for generating daily weather variables*, no. 8 in Report (United States. Agricultural Research Service), US Department of Agriculture, Agricultural Research Service Washington, DC, USA.

Rokach, L., and O. Maimon (2005), Clustering methods, in *Data mining and knowledge discovery handbook*, pp. 321–352, Springer, doi:10.1007/0-387-25465-X_15.

Royer, J., and P. Vieira (1984), Dual formalism of kriging, *Geostatistics for natural resources characterization, part*, *2*, 691–702, doi:10.1007/978-94-009-3701-7_8.

Schoelzel, C., and P. Friederichs (2008), Multivariate non-normally distributed random variables in climate research - introduction to the copula approach, *Nonlin. Processes Geophys.*, *15*(5), 761–772.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. BÃ©nard, F. Bouttier, C. Lac, and V. Masson (2011), The AROME-France Convective-Scale Operational Model, *Monthly Weather Review*, *139*(3), 976–991, doi:10.1175/2010MWR3425.1.

Serinaldi, F., A. Bárdossy, and C. G. Kilsby (2015), Upper tail dependence in rainfall extremes: would we know it if we saw it?, *Stochastic Environmental Research and Risk Assessment*, *29*(4), 1211–1233, doi:10.1007/s00477-014-0946-8.

Sharma, T. C. (1997), Estimation of Drought Severity on Independent and Dependent Hydrologic Series, *Water Resources Management*, *11*(1), 35–49, doi:10.1023/A:1007904718057.

Singh, V. P. (1998), *Pearson Type III Distribution*, pp. 231–251, Springer Netherlands, Dordrecht, doi:10.1007/978-94-017-1431-0_14.

Sklar, M. (1959), *Fonctions de répartition à n dimensions et leurs marges*, Université Paris 8.

Smith, J. A. (1987), Statistical modeling of daily rainfall occurrences, *Water Resources Research*, *23*(5), 885–893, doi:10.1029/WR023i005p00885.

Sood, A., and V. Smakhtin (2015), Global hydrological models: a review, *Hydrological Sciences Journal*, *60*(4), 549–565, doi:10.1080/02626667.2014.950580.

Spearman, C. (1904), The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, *15*(1), 72–101, doi:10.2307/1412159.

Srikanthan, R., and T. A. Mcmahon (2001), Stochastic generation of annual, monthly and daily climate data: A review, *Hydrology and Earth System Sciences Discussions*, *5*(4), 653–670.

Stedinger, J. R. (1993), Frequency analysis of extreme events, *Handbook of hydrology*, *18*.

Steinhaus, H. (1956), Sur la division des corp materiels en parties, *Bull. Acad. Polon. Sci*, *IV*(12), 801–804.

Tessema, S. M. (2011), Hydrological modeling as a tool for sustainable water resources management: a case study of the Awash River Basin, Master's thesis, KTH Royal Institute of Technology.

Thomas, H. A., and M. B. Fiering (1962), Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, *Design of water resource systems*, pp. 459–493.

Todorovic, P., and D. A. Woolhiser (1975), A stochastic model of n-day precipitation, *Journal of Applied Meteorology*, *14*(1), 17–24, doi:10.1175/1520-0450(1975)014<0017:ASMODP>2.0.CO;2.

Tudurí, E., and C. Ramis (1997), The Environments of Significant Convective Events in the Western Mediterranean, *Weather and Forecasting*, *12*(2), 294–306, doi:10.1175/1520-0434(1997)012<0294:TEOSCE>2.0.CO;2.

Verdin, A., B. Rajagopalan, W. Kleiber, and R. Katz (2015), Coupled stochastic weather generation using spatial and generalized linear models, *Stochastic Environmental Research and Risk Assessment*, *29*(2), 347–356, doi:10.1007/s00477-014-0911-6.

Viterbi, A. J. (1967), Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *Information Theory, IEEE Transactions on*, *13*(2), 260–269, doi:10.1109/TIT.1967.1054010.

Vörösmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers (2000), Global Water Resources: Vulnerability from Climate Change and Population Growth, *Science*, *289*(5477), 284–288, doi:10.1126/science.289.5477.284.

Vrac, M., M. Stein, and K. Hayhoe (2007), Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, *Climate Research*, *34*(3), 169–184, doi:10.3354/cr00696.

Wagner, M. A. F., and J. R. Wilson (1995), Graphical interactive simulation input modeling with bivariate Bézier distributions, *ACM Trans. Model. Comput. Simul.*, *5*(3), 163–189, doi:10.1145/217853.217854.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *Journal of Hydrology*, *210*(1–4), 178–191, doi:10.1016/S0022-1694(98)00186-3.

Wilks, D. S. (1999), 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, doi:10.1016/S0168-1923(99)00037-4.

Wilks, D. S. (2009), A gridded multisite weather generator and synchronization to observed weather data, *Water Resources Research*, *45*(10), doi:10.1029/2009WR007902, w10419.

Wilks, D. S. (2014), Multivariate ensemble Model Output Statistics using empirical copulas, *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 945–952, doi:10.1002/qj.2414.

Wilks, D. S., and R. L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Progress in Physical Geography*, *23*(3), 329–357, doi:10.1177/030913339902300302.

Wilson, L. L., D. P. Lettenmaier, and E. Skyllingstad (1992), A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation, *Journal of Geophysical Research: Atmospheres*, *97*(D3), 2791–2809, doi:10.1029/91JD02155.

Wojcik, R., J. J. Beersma, and T. A. Buishand (2000), Rainfall generator for the Rhine basin: Multi-site generation of weather variables for the entire drainage area, *KNMI Publications*.

Zimmerman, D. A., G. de Marsily, C. A. Gotway, M. G. Marietta, C. L. Axness, R. L. Beauheim, R. L. Bras, J. Carrera, G. Dagan, P. B. Davies, D. P. Gallegos, A. Galli, J. Gómez-Hernández, P. Grindrod, A. L. Gutjahr, P. K. Kitanidis, A. M. Lavenue, D. McLaughlin, S. P. Neuman, B. S. RamaRao, C. Ravenne, and Y. Rubin (1998), A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resources Research*, *34*(6), 1373–1413, doi:10.1029/98WR00003.

Zotarelli, L., M. D. Dukes, C. C. Romero, K. W. Migliaccio, and K. T. Morgan (2010), Step by step calculation of the Penman-Monteith Evapotranspiration (FAO-56 Method), *Institute of Food and Agricultural Sciences. University of Florida*.

Zucchini, W., and P. Guttorp (1991), A Hidden Markov Model for Space-Time Precipitation, *Water Resources Research*, *27*(8), 1917–1923, doi:10.1029/91WR01403.

Abstract

This PhD work proposes new concepts and tools for stochastic weather simulation activities targeting the specific needs of hydrology. We used, as a demonstration, a climatically contrasted area in the South-East of France, Cévennes-Vivarais , which is highly attractive to hydrological hazards and climate change.

Our perspective is that physical features (soil moisture, discharge) relevant to everyday concerns (water resources assessment and/or hydrological hazard) are directly linked to the atmospheric variability at the basins scale, meaning firstly that relevant time and space scales ranges must be respected in the rainfall simulation technique. Since hydrological purposes are the target, other near-surface variates must be also considered. They may exhibit a less striking variability, but it does exist. To build the multi-variable modeling, co-variability with rainfall is first considered.

The first step of the PhD work is dedicated to take into account the heterogeneity of the precipitation within the rainfall simulator SAMPO [*Leblois and Creutin*, 2013]. We cluster time steps into rainfall types organized in time. Two approaches are tested for simulation: a semi-Markov simulation and a resampling of the historical rainfall types sequence. Thanks to clustering, all kind of rainfall is served by some specific rainfall type. In a larger area, where the assumption of climatic homogeneity is not considered valid, a coordination must be introduced between the rainfall type sequences over delineated sub-areas, forming rainy patterns at the larger scale.

We first investigated a coordination of Markov models, enforcing observed lengths-of-stay by a greedy algorithm. This approach respects long duration aggregates and inter-annual variability, but the high values of rainfall are too low. As contrast, the joint resampling of historically observed sequences is easier to implement and gives a satisfactory behavior for short term variability. However it lacks inter-annual variability. Both approaches suffer from the strict delineation of homogeneous zones and homogeneous rainfall types.

For these reasons, a completely different approach is also considered, where the areal rainfall totals are jointly modeled using a spatio-temporal copula approach, then disaggregated to the user grid using a non-deterministic, geostatistically-based conditional simulation technique. In the copula approach, the well-known problem of rainfall having atom at zero is handled in replacing historical rainfall by an appropriated atmospheric based rainfall index having a continuous distribution. Simulated values of this index can be turned to rainfall by quantile-quantile mapping.

Finally, the copula technique is used to link other meteorological variables (i.e. temperature, solar radiation, humidity, wind speed) to rainfall. Since the multivariate simulation aims to be driven by the rainfall simulation, the copula needs to be run in conditional mode. The achieved toolbox has already been used in scientific explorations, it is now available for testing in real-size application. As a data-driven approach, it is also adaptable to other climatic conditions. The presence of atmospheric precursors a large scale values in some key steps may enable the simulation tools to be converted into a climate simulation disaggregation.

Keywords: stochastic weather generator; hydrology; multivariate; heterogeneity.