SAPIENZA
UNIVERSITÀ DI ROMA

Statistical methods for the analysis of climatic phenomena

Department of Statistical Sciences-University of Rome La Sapienza

Dottorato di Ricerca in Statistica Metodologica – XXIV Ciclo

Candidate

Edmondo Di Giuseppe
ID number 951655

Thesis Advisor

Prof. Giovanna Jona Lasinio

Co-Advisors

Dr. Stanislao Esposito
Dr. Massimiliano Pasqui

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Methodological Statistics

December 2013

Thesis defended on 13 December 2013
in front of a Board of Examiners composed by:

Prof. Mariaclelia Stefania Di Serio (chairman)

Prof. Fabio Divino

Prof. Pier Paolo Brutti

Statistical methods for the analysis of climatic phenomena
Ph.D. thesis. Sapienza – University of Rome

© 2013 Edmondo Di Giuseppe. All rights reserved

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: edidigiu@gmail.com

Dedicated to Piera, my beloved wife

Contents

Prelude

Statistical Climatology investigates the application of statistics to atmospheric and climate science. This Thesis is intended to contribute some developments in this field. The Thesis is composed of two distinct applied problems: the first presents a functional clustering procedure applied to meteorological time series to obtain homogeneous climate zones; the second builds a hierarchical Bayesian model aiming at the prediction of 15- and 30-minutes cumulated precipitation at unknown locations and time using information on lightnings in the same area.

Part I: Functional Clustering for climate zone determination

The first part of the Thesis presents a functional clustering procedure applied to meteorological time series. Our proposal combines time series interpolation with smoothing penalized B-spline and the Partitioning Around Medoids (PAM) clustering algorithm. We compare this approach to standard methods based on a combination of Principal Component Analysis and Cluster Analysis (CA) and we discuss it in relation to other functional clustering approaches based on Fourier analysis and CA. We show that a functional approach is simpler than standard methods from a methodological and interpretability point of view. Indeed it becomes natural to find a clear connection between mathematical results and physical variability mechanisms. We discuss how the choice of the basis expansion (splines, Fourier) affects the analysis and propose some comments on their use. An assessment based on climatic patterns is presented to prove the consistency of the clustering and a comparison of results obtained with different methods is used to judge the functional data approach. The basis for classification is formed by monthly values of temperature and precipitation recorded during the period 1971-2000 over 95 and 94 Italian monitoring stations respectively. This work has been presented at *11th International Meeting on Statistical Climatology* and *Gfkl-Cladag Joint Meeting 2010* and published in *Theoretical and Applied Climatology* journal.

Part II: Predicting rainfall fields from lightnings records: a hierarchical Bayesian approach

The general purpose of the second part of the Thesis is to build a spatio-temporal model of lightnings records with the final aim of improving rainfall fields predictions.

Our proposal is a Bayesian space-time hierarchical model. Basically we adopt a mixed model approach to the representation of precipitation as function of lightnings counts during storms. The first issue to be addressed is the identification of single storms (event) among several severe meteorological events. Then our first contribution in this study, is a simple and effective scan statistics procedure to separate 'events'. These events are *convective storms* from which either precipitation or lightnings might be generated. Then we present the Mixed Models approach in which precipitations are modeled as function of lightning counts (fixed effects) and space time variation is handled using specific random effect. The space-time random effect is modeled as separable, with a Conditional Autoregressive model (CAR) to model the spatial random component and a simple AR(1) model to represent time variation. We show that our modelling approach has a good capability in identifying peaks whilst it is likely to underestimate rainfall quantity, though the mean predicting error is not large. Moreover, great part of zero precipitation cases are well identified and predictive intervals have empirical coverage closer to the nominal values attesting to the accuracy of predictions. The area of study is located in Central Italy and the period of analysis is 2003-2006. The database is composed of lightnings records (instant-point fields), satellite precipitation records (hourly-10×10 km interpolated fields) and the weather stations precipitation records (sub hourly-point fields). This work has been presented at *Bayesian Young Statisticians Meeting 2013* and *IX Conference on Geostatistics for Environmental Applications GeoENV2012*.

Part I

Functional Clustering

Introduction to Part I

A key issue in meteorological fields analysis is played by the study of their spatio-temporal variability. It exists a structural variability which describes the nature of a phenomenon both to intra-annual (*seasonality*) and long term variability (*climate trend*) and it is relevant to be able to analyse them over homogeneous climate areas. A set of different methods are used for climate zones determination, typically a combination of Principal Component Analysis (PCA) and Cluster Analysis (CA). Guidelines on the use of PCA in meteorology and climatology have been set in the work of Preisendorfer and Mobley (1988) [Preisendorfer et al. 1988]. A theoretical and applied framework of the principal component analyses of climate-related fields is given in Chapter 13 of Von Storch and Zwiers (1999) [Von Storch and Zwiers 1999]. The spatial domain PCA (S-mode) is a reduction of the information related to the temporal patterns of the locations [Ehrendorfer 1987]. Thus, each component generates a mapping of mixed physical features. On the other hand the temporal domain PCA (T-mode) by reducing the information seen from the time series point of view, attempt to describe climate regime [Richman 1986]. Finally, the R-mode approach points at locales similarity in mean and variances of meteorological fields across a fixed time by means of CA [Fovell and Fovell 1993]. The main drawback of PCA based techniques is that the reduced space they return as output does not have an immediate connection with the physical one.

In this work a combination of Functional Data Analysis (FDA) and Partioning Around Medoids (PAM) clustering technique is applied in Italy to monthly surface temperature and precipitation fields in order to delineate locale climate zones. FDA is a collection of techniques to model data from dynamic systems in terms of some set of basis functions, which are a linear combination of known functions. FDA consists of converting observations gathered at discrete time into functional data. The choice of the basis to implement this conversion is crucial. The functional data approach is typically used in genetic [Kim et al. 2008] and pollution's diffusion analysis [Ignaccolo et al. 2008] and only very recently in climate studies [Laguardia 2011]. Kim et al. 2008 use functional data approach for modeling the time-dependent expression value of genes in the genome of yeast and they find that the features of those genes are properly modeled by a 3-orders Fourier series approximation. Ignaccolo et al. 2008 fit the functional data to pollutant ceoncentrations time series using B-splines system of basis, with a fixed number of knots. Then, they produce a zonal index of pollutant's concentration in Northern Italy based on a clustering of estimated coefficients. In Laguardia 2011, a Fourier basis expansion is adopted to model a very large amount of precipitation data (2043 rain gauges). The clustering is performed using a k-means clustering algorithm. Our approach differs from his first of all for the choice of the clustering algorithm, and secondly as in our setting penalized B-splines are preferred to Fourier basis. Our choices are discussed below in details. We also note that in our work a smaller amount of data than in Laguardia is considered, nevertheless returning very coherent results.

Temperature and precipitation time series can be considered as realizations of continuous processes recorded in discrete time. Thus, they are converted into functional data through the estimation of spline coefficients and the latter used for the final classification as each time series is representative of location climate variability.

Here a penalized B-spline basis system is adopted to map observations gathered at discrete time into functional data. Our proposal is named *Bsplines30 model* and reproduces data intra-annual variability by means of B-splines basis system over a 30-years period (1971-2000). A fixed number of knots guarantees a comparability of responses from the 95 and 94 time series, which constitute the data-set for the analysis of temperature and precipitation, respectively. On the contrary, a system with a free number of knots would leads each series to be smoothed according to different scale of variability and, *de facto*, the delineation of homogeneous zones would not be done. Finally the estimated coefficients are partitioned by PAM classification technique and average silhouette width method is used to determine the number of climate zones [Rousseeuw 1987].

The main advantage of a functional approach to this type of data is dimensional reduction, as the information on monthly temporal pattern given by a large number of observations (time series) is summarized by a small number of coefficients that describe the basis spanning the functions [Ramsay and Silverman 1997]. Furthermore the proposed approach overcomes the problem of connecting the reduced space to the physical one. Indeed the fitting of B-splines allows to define in a clear way which type of variability is considered.

This part of the Thesis is organized as follows. In Chapter 1 the available data and an exploratory analysis of temperature and precipitation spatial patterns are presented. In Chapter 2, we illustrate functional clustering in general terms and then we move to illustrate our proposal: Section 2.1 is devoted to the presentation of penalized B-spline; in Section 2.2 a description of the k-medoids clustering procedure is reported and Section 2.3 details our proposal. In Chapter 3, we prove and discuss the validity of the method and we compare the final grouping with the same results obtained by means of PCA method in T-Mode. Relations of our proposal to Fourier analysis approach is discussed in the same section. Finally, in the last section some concluding remarks are presented.

Chapter 1

Dataset for functional clustering

In this work we define "Regime" the signal obtained by averaging monthly values over the years in each station. The dataset is composed of daily precipitation and daily minimum and maximum temperature data collected by CRA-CMA Research Unit for Climatology and Meteorology Applied to Agriculture for the period 1971-2000 from 98 Italian stations.

1.1 Data description: missings, outliers and imputation

The total number of stations is 98 (Fig. 1.1), but only 92 are in common with Temperature and Precipitation, then we have 96 stations for Temperature and 94 for Precipitation. This dataset is composed of climatic time series with a relatively small amount of missing values over the considered time window. This fact in addition to the continuity of these time series is a considerable advantage over other more numerous, in terms of monitoring stations, Italian dataset, such as SIMN (SIMN is affected by a severe missing data problem and, moreover, its collection ends around 1989 when it was dismissed).

Among the 98 stations 7 are located above 1500 meters and it is natural to expect that, due to correlation between temperature and altitude, they may form a cluster. For those stations, especially during winter season, observed values of precipitation might be due to snow events which amount is usually transformed into equivalent precipitation quantity. Nevertheless this fact do not affect the analysis provided that all the mountains' stations were grouped in a unique cluster. One station time series (Pian Rosà) has been removed as the station is located at 3480 meters and becomes an outlier with respect to the other stations (see Fig. 1.2).

Minimum and maximum temperatures were averaged to obtain a rough estimate of daily medium temperatures. An outliers detection is performed eliminating each element outside the range $(x - 4var; x + 4var)$ for daily medium temperatures, where var is the variance and greater than $99th$ percentile for daily precipitation. Then Monthly Mean of Medium Temperature (Tmed-MM) and Monthly Cumulated Rainfall (Prec-MC) were calculated provided that at least 21 daily data in a month were registered. If not, the correspondent monthly value is set to NA (Not Available), i.e. missing value. Besides, a non parametric test for outliers detection of these monthly values is performed. This test is based on Median Absolute deviation (MAD)

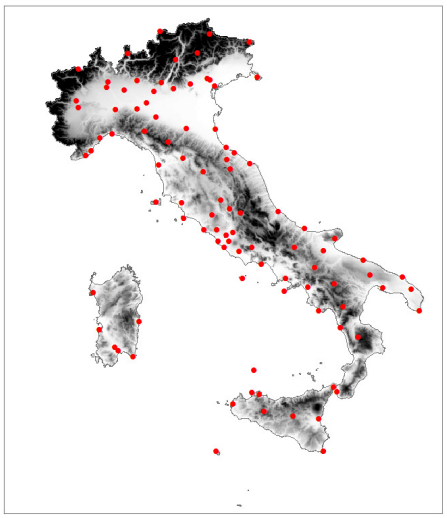


Figure 1.1. Location of weather stations.

and is suggested in Sprent 1998 as "simple and reasonably robust test" [Sprent 1998]. In fact, MAD is itself a robust estimator of the spread of a univariate data series. More specifically, let x_i be the element of a data series with $i = 1, \ldots, n$ and x_{Med} the median of the series, then MAD is the median of the absolute deviation from the median:

$$MAD = Median(|x_i - x_{Med}|) \tag{1.1}$$

and x_i is detected as outlier if:

$$\frac{|x_i - x_{Med}|}{MAD} > M \tag{1.2}$$

where $M = 5$ following Sprent and Smeeton, 2001 [Sprent and Smeeton 2001]. They suggest this rule of thumb because of the approximate relation $5MAD = 3Sd$, with Sd denoting standard deviation. The cross stations outliers detected and, successively removed, by the MAD-based test applied over the period 1971-2000 are 0 for Tmed-MM and 150 for Prec-MC. The latter is not significant with respect to the 33840 overall number of data (360 monthly values x 94 stations). Summary statistics of the eliminated cases for Prec-MC are in Appendix 3.2 3.4 whilst general summary statistics for Tmed-MM and Prec-MC are reported in Table 1.1 together with the overall number of missing data. Moreover, the number and percentage of monthly missings data of the stations relatively to the period 1971-2000 is listed in Appendix (Table 3.2).

An *imputation* of missing monthly data has been performed accounting for seasonal variability and 3-years climate cycle, since the completeness of the series makes the application of FDA method easier from a computational point of view and it eases the output interpretation. In particular to estimate spline coefficients from a series completeness is necessary, but the values of the curve - giving rise

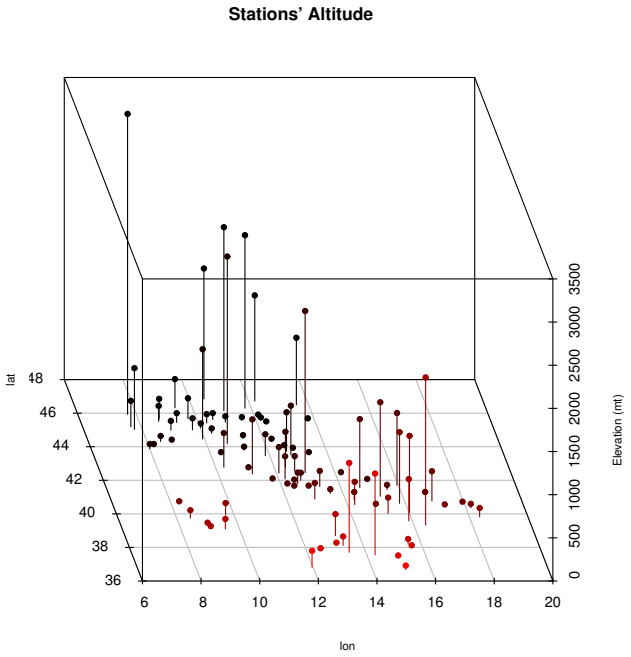


Figure 1.2. Altitude of weather stations.

Variables	Min	1st Qu	Median	Mean	3rd Qu	Max	% of missings data
Tmed-MM	-14.2	8.5	13.5	13.7	20.0	30.2	4.36
Prec-MC	0	19.6	46.7	60.1	86.3	393.6	5.36

Table 1.1. Tmed-MM and Prec-MC summary statistics and percentage of monthly missing data calculated across the overall locations.

to that series - can be observed on an irregular grid. Therefore, if the amount of missing data is small, it is possible to omit NA's and estimate spline coefficients even when time series are not complete. Nevertheless, the completeness of the series is fundamental for establishing the announced connection to physical variability. Briefly, the missing data $\tilde{y}_{ij}$ in year i and month j, is imputed as:

$$\tilde{y}_{ij} = 1/3(\overline{y}_j + [1/2 \cdot (y_{i-1,j} + y_{i+1,j})] + [1/2 \cdot (y_{i,j-1} + y_{i,j+1})]) \tag{1.3}$$

where $\overline{y}_j$ is the 30-years average corresponding to the jth month value. Whenever contiguous missing data are found, they are directly imputed with the 30-years average. Being the number of the monthly missing values small, we decided to not adopt a complex statistical model (such as ARIMA or VARMAX) for imputation. We use the above described procedure that takes into account the general features of monthly regimes and, it is conservative in terms of variability since we use climatological levels. We experimented with other techniques, such as spline

imputation as implemented in the `na.spline()` function in R (package "zoo") and other versions of our approach, observing that the proposed functional clustering approach is robust to missing values imputation, i.e. stations are classified in the same way regardless of the chosen imputation technique. This result is most likely due to the reduced number of missing values present in the data. Besides, because of extremely high variability of precipitation, a Cox-Box transformation with coefficient $\lambda = 0.5$ has been performed on Monthly Precipitation data [Box and Cox 1964]. This transformation corresponds to a square root of the initial data and determines tighter high scale data and looser low scale data (Fig. 1.3). Finally, our dataset is composed by 95 and 94 time series of 360 monthly values of Tmed-MM and square root Prec-MC, respectively, since the removed station of Pian Rosà was originally included only in Temperature dataset. In the following, we mention Prec-MC always referring to square root of Prec-MC whereas the levels expressed in millimeters are back-transformed to the original scale.

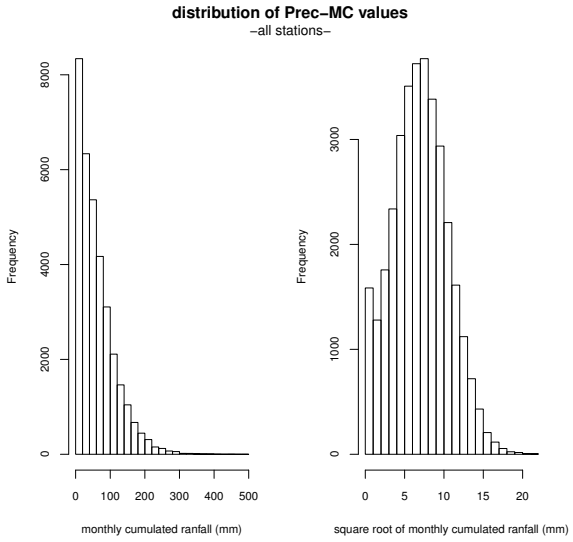
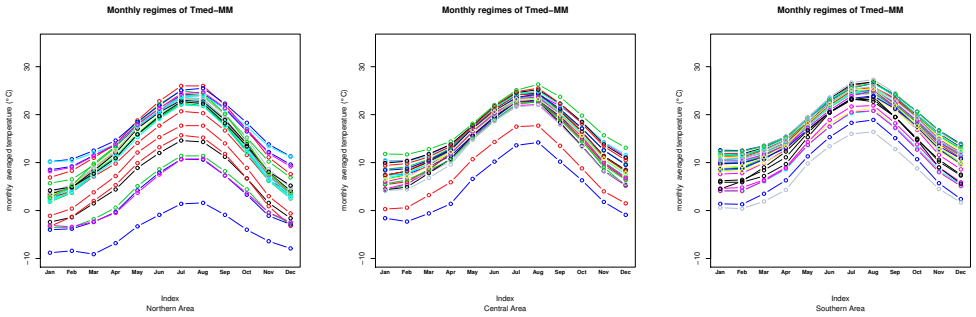


Figure 1.3. The effect of square root trasformation of monthly precipitation data.

1.2 Exploratory data analysis

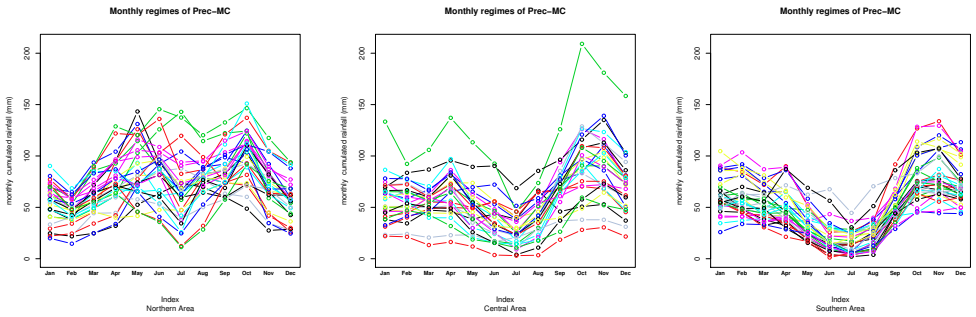
A major result of the exploratory data analysis is the strong influence of local factors in determining both temperature and precipitation spatial patterns. In figures 1.4 and 1.5 monthly regimes of Tmed-MM and Prec-MC grouped according to latitude classification of Italian territory are shown. From these representations it appears clear the strong influence of latitude in determining different climate features of grouped locations. The longitude also plays a key role in this building due to the presence of Appenini mountain chain that extends throughout the Italian peninsula (not shown here). Furthermore, the complex topography of Italian peninsula along with the strong influence of sea over air masses flow generate a large amount of small

scale atmospheric variability which is able to modulate not only the temperature field, but also the precipitation one [Trigo et al. 2006]. These local variability make the building of climatic homogeneous classification particularly challenging.



(a) (b) (c)

Figure 1.4. Tmed-MM: 30-years monthly average of 96 stations grouped into Northern, Central and Southern area.



(a) (b) (c)

Figure 1.5. Prec-MC: 30-years monthly average of 94 stations grouped into Northern, Central and Southern area.

Chapter 2

Method

Functional clustering combines the functional representation through a given basis expansion of a time series with a cluster algorithm with the aim of finding observed units homogeneous groups. The choice of a basis implies the type of features of the series that are to be enhanced or hidden in the representation [Ramsay and Silverman 1997] and then become relevant in the classification building. The two most commonly chosen basis are Fourier and B-splines. The first one is mostly adopted when data are assumed to have an important periodic component; the second one is particularly suitable when no periodicity is anticipated in the data or periodicity is affected by some type of changing component. A B-splines smoothing is able to incorporate the shifts in the mean level of the time series caused by a breakpoint into the estimates of the coefficients. This may constitute an advantage especially in the study of climate variables and obviously, depending on the scope of the analysis. For instance, an *ad hoc* analysis can be conducted combining two B-splines systems of basis: a first smoothing placing a knot every year in order to model the *trend* component and, in case, the breakpoints; a second smoothing placing a knot every 2-, 3-, or 4-months for modelling the *seasonal* component (see chapter 7 of Ramsay and Silverman, 2002 [Ramsay and Silverman 2002]). In particular, the penalized version of B-splines, which we adopt here becomes useful when the interest is in representing smooth functions without completely removing local behaviour in time, such as changes in the time series level that persist for a limited time [Ramsay and Silverman 1997]. Furthermore this basis allows to capture specific variability patterns with an appropriate choice of knots localization.

The second element of functional clustering is the Clustering Algorithm. In the literature k-means algorithm has already been used in application to precipitation data [Laguardia 2011]. Here we prefer a Partitioning Around Medoids Algorithm [Kaufman and Rousseeuw 1990]. The k-medoids algorithm is a clustering algorithm related to the k-means algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars), making easier to identify groups features. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean

distances [Kaufman and Rousseeuw 1990].

In what follows we report a brief description of our main tools: penalized B-spline basis and Partitioning Around Medoids Algorithm.

2.1 Functional data smoothing

FDA transforms discrete data y_j in a functional form using a system of basis. B-splines basis are piece-wise polynomials of degree d joined at $k+1$ fixed points named *knots*. Two adjacent polynomials are required to have matching $d-1$ (continuous) derivatives. The order of the polynomial B-splines is $d+1$ and the free parameters are $k+d+1$. The degree of smoothing is determined both by the location and the number of knots. Thus a function $x(t)$ can be represented as a linear combination of k known basis functions ϕ_j:

$$x(t) = \sum_{j=1}^{k+d+1} c_j \phi_j(t) \tag{2.1}$$

the coefficients of the expansion c_j are determined by minimizing a least square criterion. In the penalized B-splines basis a penalization term is added to ensure control over local variability and to reduce outliers influence on the least squares estimates. The penalization term involves a smoothing parameter λ and a linear differential operator $PEN(x)$ which is a measure of the function roughness (it is the value of an approximate integral over the x range of the square of the $d-1$ derivative of the curve, which quantifies the total curvature of the function). The penalized least square criterion adopted for coefficients estimation is:

$$PENSSE_\lambda(x \mid \mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]' \mathbf{W} [\mathbf{y} - x(\mathbf{t})] + \lambda PEN(x) \tag{2.2}$$

where $\mathbf{W}$ is a symmetric, positive definite weights matrix. The smoothing parameter λ is chosen by *generalized cross validation* criterion:

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{PENSSE}{n - df(\lambda)}\right) \tag{2.3}$$

where *df* are the degrees of freedom in the smoothing curve and its value depends on the number of knots and the spline degree that will be specified in Chapter 3. The best choice of λ is associated to the minimum value of GCV(λ). For large values of λ, the curve approaches the standard linear regression. A penalized B-splines smoothing with a small number of coefficients is able to capture the shape of the curve and accommodate for local features. Indeed, by using penalized B-spline we obtain that outliers in the data do not affect coefficients estimates. We run our method with and without outliers in the data and the interesting feature is that conclusions are not affected by the presence of outliers. However in our data we have only few *anomalies* then, as a good practice, we suggest to remove outliers identified by MAD. Notice that this identification method finds very extreme values (approximately larger than 3 times the standard deviation) and no outliers are found for Tmed-MM while 150 values are identified for Prec-MC. Thus, the estimate of the coefficients we use in the clustering method, is robust. Simple polynomial regression

does not have this kind of robustness and small changes in the data can dramatically affect the coefficients estimates [Abraham et al. 2003].
In practice the construction of the "best" penalized B-spline representation proceed by iterating two steps: (i) fix the number of parameters (knots and polynomial degree) and choose λ by GCV; (ii) compute RMSE then change the number of parameters and go back to (i); repeat this two steps until no more sensible reduction in RMSE is obtained. Finally the combination of λ and parameters number that returns the smallest RMSE is chosen. In general, this sequence of steps can be carried on automatically or a data-driven choice of parameters can be performed. In our case study we choose the latter as we want the final clustering to have a physical meaning and, at the same time, we want to minimize the number of estimated parameters (details are given in Section 2.3).

2.2 Partitioning Around Medoids classification method

K-medoids algorithm is based on the object called *medoid* (most centrally located point in the cluster) instead of the centroid of k-means algorithm (average of objects coordinates in the cluster). This has two advantages: firstly, the medoid is a real object and it is representative of group features; secondly, there is no need to calculate distances at each iteration since the reference is the distance matrix between objects. The steps of k-medoids algorithm can be summarized as follows:

1. choose randomly k objects of the n data points to be the initial cluster medoids;
2. assign objects to the cluster with the closest medoid;
3. recalculate the k medoids of clusters as formed at step 2;
4. repeat steps 2 and 3 until the medoids do not change.

Step 3 is performed by finding the object i which minimizes

$$\sum_{j \in C_i} d(i; j) \tag{2.4}$$

where C_i is the *cluster* including i and *d(i; j)* is any measure of dissimilarity (common choices are Euclidean and Manhattan norms) between observations i and j.

Among k-medoids algorithm, the most used and powerful is Partitioning Around Medoids (PAM) algorithm proposed by Kaufman and Rousseeuw [Kaufman and Rousseeuw 1990]. This algorithm is characterized by an efficient procedure for determining the set of medoids, which can be described in two phases: the "build" and the "swap". The gain in the algorithm efficiency introduced with PAM is described in [Reynolds et al. 1992]. In the build phase the algorithm looks for a good initial set of medoids. Then, in the swap phase it calculates the loss in the objective function determined by changing medoid. More specifically, consider the effect of removing object i from the set of medoids and re-placing it with object h. The total cost of the change is given by the sum of the cost associated to each object j that move from other clusters to the new cluster h determined by the change. In particular, there are 3 cases:

a) the cost is zero since object j does not move;

b) object j is closer to the initial medoid i than any other medoid before the swap, then the cost associated to the moving is $d(j,h) - d(j,i)$;

c) j is further from i than from some other medoid, then the cost of the moving is $d(j,h) - D_j$, where D_j is the distance of object j from the closest medoid (if the closest is h then the cost is zero).

If the total cost is negative, then the move gives an improvement in the clustering. The whole neighbourhood is evaluated in each iteration of the algorithm. Here Kaufman and Rousseeuw [Kaufman and Rousseeuw 1990] suggest that calculating the change in cost rather than the total cost at each iteration is less operational demanding.

Once the medoids have been fixed, clustering quality indexes can be calculated. Let *a(i)* be the average dissimilarity between i and all objects in cluster C_i and let $d(i;C)$ be the average dissimilarity of i to all objects in C, with $C \neq C_i$. Denote with *b(i)* the smallest distance $d(i;C)$ found among all clusters $C \neq C_i$, then C is the neighbour cluster of i. An evaluation of how well the object i is classified in C_i or in the *neighbour cluster* is given by the *silhouette width index*:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.5}$$

Observations with a $s(i)$ value close to 1 are very well clustered, a small value of $s(i)$ means that the observation can be assigned to two clusters, and observations with a negative $s(i)$ are misplaced. The number of clusters can be determined by the *average silhouette width*, which is the mean of *s(i)* over all objects of any possible clustering [Rousseeuw 1987].

2.3 Proposed functional clustering

In Section 2.1 we describe a general functional smoothing for one time series. We now consider the multiple time series framework that is proper to climatological studies. The first point that requires attention is how apply the protocol of Section 2.1 to all series in order to obtain comparable results. We propose to use the same penalized B-splines for all time series i.e. we modify, following [Ramsay and Silverman 1997], steps (i) and (ii) as follows: (i.a) we fix the same number of knots (or their position) and polynomial degree for all time series and we choose a unique smoothing parameter λ by GCV. First for each time series the GCV corresponding to a given value of λ is computed and then the the average of these GCV values is associated to the specific λ; (ii.a) we compute RMSE for each time series and then the average RMSE; we repeat (i.a) and (ii.a) until no sensible changes are obtained in the average RMSE. Finally we choose the combination of λ and parameters number that returns a meaningful clustering and, simultaneously, a small average RMSE. In fact, a key point in our procedure is the choice of the number of knots and their positions, sometimes it is necessary to compromise between a small average RMSE value and a set of knots that returns a meaningful representation of the time series. For instance

in our study, by placing a knot every 4 months we capture intra-annual variability with considerable accuracy; with a knot every 3 months we obtain a smaller average RMSE but the series representation becomes more sensitive to outliers and, if no outliers are present, it is identical to the 4 months one with a considerable increase in the number of parameters to be estimated from the data. Remark that the introduction of a large number of parameters not necessarily helps the understanding of climatic features as not only information is thus added but also noise (variability).

Once the representation of the time series is obtained, the coefficients of the functional smoothing become input of the clustering algorithm with the aim of obtaining climate zones delineation. Here we use the k-medoids algorithm PAM as implemented in the R `cluster` library, [R Development Core Team 2011] and illustrated above (Section 2.2). The number of clusters is chosen by average silhouette and climatological considerations. In other words if the largest silhouette value is given by a very small number of clusters, say 2, that does not have a climatological meaningful interpretation, we look for the second best or the third best and so on. Besides, the choice of the proper number of clusters is done taking into account also the information associated to the PAM algorithm, as isolation, diameter of clusters, separation and silhouette width of each group.

In the present study we are going to call our procedure *Bspline30* as here we eventually adopt a penalized B-splines basis with a knot every 4 month over a 30-years period (1971-2000), which is a period commonly used as Climatic Normals [WMO 1989]. We fix the two knots corresponding to the edges of the smoothing interval respectively on January 1971 and on December 2000 whereas the position of the interior knots, the degree of the polynomial and the smoothing parameter are determined as illustrated in Section 2.1 and above, using `fda` library [Ramsay et al. 2011], implemented in R [R Development Core Team 2011]. The clustering is performed using PAM implemented in the R library `cluster`. Details of the results are given in the next Chapter.

Chapter 3

Results of Functional Clustering

This work is motivated by the need of finding a segmentation procedure of the available time series leading to homogeneous classes. Most of the analysis regarding the determination of homogeneous climatic regions are based on the monthly time scale. Then, we adopt the monthly scale using monthly averages. This time space truncation is commonly adopted in order not to include the synoptic and sub-synoptic variability signals in the atmosphere variability. The climate framework of Italian peninsula is made complex by both the sea mitigation effect on temperature and the presence of Alps in the North as well as Appennino along the latitude extension which affect precipitation distribution. In fact, some studies based on standard clustering techniques classify Italian climate in 7/8 homogeneous sub-regions ([Laguardia 2011], [Toreti et al. 2009], [Brunetti et al. 2006]). On the other hand, the Mennella's basic work of 1972 ([Mennella 1972]) describes at least 20 climate micro-regions using both observations and physical features. From a phenomenological point of view, the main advantage of functional clustering is a clear identification of variability mechanisms whereas standard methods need to find a relation between selected Principal components (Pcs) and climate patterns. Recall that, with the S-mode of PCA, we look for the most significant Pcs of the information matrix over the stations, then we map the elements of the corresponding eigenvectors (*loadings*) which are associated to each station [Ehrendorfer 1987]. On the other hand, with the T-mode, we look for the most significant Pcs of information matrix over time, then mapping the *scores* [Richman 1986]. In this study, we focus on *Intra-annual variability* by placing penalized B-splines knots every 3, 4 and 6 months. which let us to capture intra-annual variation with scale of variability larger equal than 3 months. The functional smoothing performed in this way preserves the bell shaped temperature monthly distribution typical of Italian peninsula and the largest intra-annual precipitation pick. As an abbreviation we use the term *4-monthly* (or 3-monthly or 6-monthly) to recall the variability scale and the placement of penalized B-splines knots. Following the approach proposed in Section 2.3 for the functional smoothing, the most interesting models among all those investigated are reported in Table 3.1 where the average RMSE is reported together with the penalization coefficient (λ), the number of total knots of the B-splines and the degree of the piece-wise polynomials used. Notice that with a knot every 3 months the average RMSE is a little smaller than the one obtained with a knot every 4 months,

but the total number of parameters to be estimated from the data considerably increases, without any advantage in the subsequent classification (details to support this statement are given in Section 3.1 and 3.2).

Tmed-MM Functional Model	**Degree**	**Knots**	**Lambda**	**Rmse (°C)**
Bsplines30 6-monthly	3	60	0.06	1.97
Bsplines30 6-monthly	5	60	1	1.89
Bsplines30 4-monthly	3	90	0.16	1.58
Bsplines30 4-monthly	5	90	3.98	1.44
Bsplines30 3-monthly	3	120	0.25	1.35
Bsplines30 3-monthly	5	120	0.63	1.35
Sqrt Prec-MC Functional Model				**Rmse (mm)**
Bsplines30 6-monthly	3	60	3.98	8.78
Bsplines30 6-monthly	5	60	15.85	8.70
Bsplines30 4-monthly	3	90	6.31	8.44
Bsplines30 4-monthly	5	90	63.1	8.39
Bsplines30 3-monthly	3	120	3.98	7.79
Bsplines30 3-monthly	5	120	63.1	8.29

Table 3.1. Tmed-MM and Sqrt Prec-MC model selection for functional data transformation with penalized B-splines piece-wise polynomials degree, number of knots, penalty coefficient (lambda) and averaged across stations RMSE.

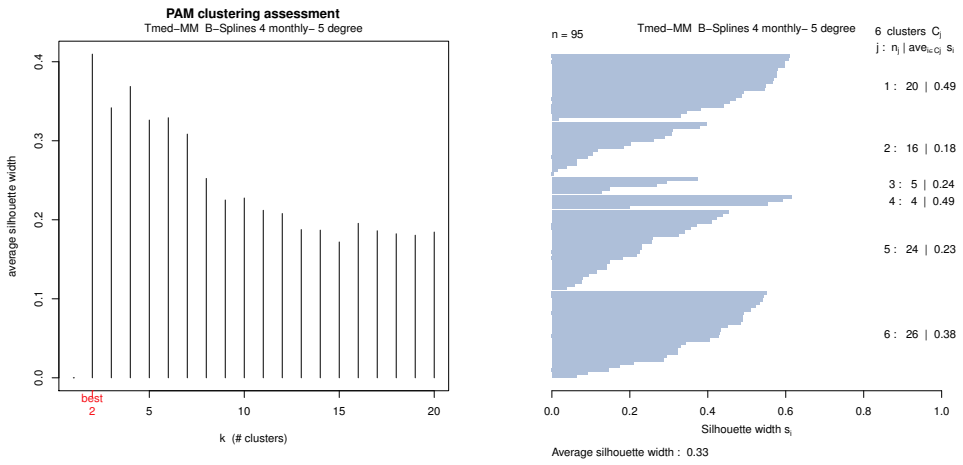
The assessment for determining the proper number of clusters and the corresponding index to evaluate the quality of the chosen clustering are visualized in panel (a) and panel (b) of Figs. 3.1 and 3.5, for the temperature and precipitation respectively. Medoids locations are representative of the climate features of all stations belonging to the corresponding cluster, and are enhanced in the classification maps of Figs. 3.2 and 3.6. Besides, the functional smoothing of medoids' time series over the period 1971-2000 are represented in Fig. 3.3 and 3.7. The maps of classification obtained by Bsplines30 model are reported in Figs. 3.2a (Tmed-MM) and 3.6a (Prec-MC). In the comparison procedure we adopted different ways of summarizing time series features: PCA in T-Mode and Fouries basis functional smoothing. The latter includes, following [Laguardia 2011], 12- and 6-monthly harmonics that should be enough to capture monthly regimes. The final classifications have always been obtained using PAM as in Section 2.3. Classification maps of temperature and precipitation clusters obtained trough PCA standard method are reported in Figs. 3.2b and 3.6b, finally the Fourier functional smoothing clusters are mapped in panels (c) of the same figures to facilitate comparison.

3.1 Results for Monthly Mean of Temperature (Tmed-MM)

The chosen model for Tmed-MM is *Bsplines30 4-monthly 5-degree* with 90 fixed knots (a knot every four months) and 5 degree piece-wise polynomials which corresponds to functional smoothing of order 6 (see Table 3.1). This choice produces a good

smoothing with an average RMSE value of 1.44 °C although the 1.35 °C minimum value of RMSE is achieved with Bsplines30 3-monthly 3-degree. Nevertheless, as mentioned above, the gain in the smoothing is not enough to justify the increase in the number of parameters to be estimated (from 94 to 122) as the bell shape of monthly temperature distribution typical of Italian peninsula is well reproduced by the 4-monthly scale of variability and, moreover, it is not adding any useful information for the final classification. In fact, the "best" number of groups obtained from Bsplines30 3-monthly 3-degree model is 5. This choice is done taking into account all clustering indexes and climate patterns. The maximum value of average silhouette width index corresponds to 3 groups clustering that has no climatic meaning. The 5 groups clustering has an average silhouette width of 0.36 with 1 misplaced unit and returns equivalent results to our 4-monthly 5-degree model except for the northern mountain region. There a single cluster is found by the 3-monthly model, while two clusters (cluster 1 and 2 in our classification mapped in Fig. 3.2 panel (a)) are given by the 4-monthly model the latter being more meaningful from a climatic point of view.

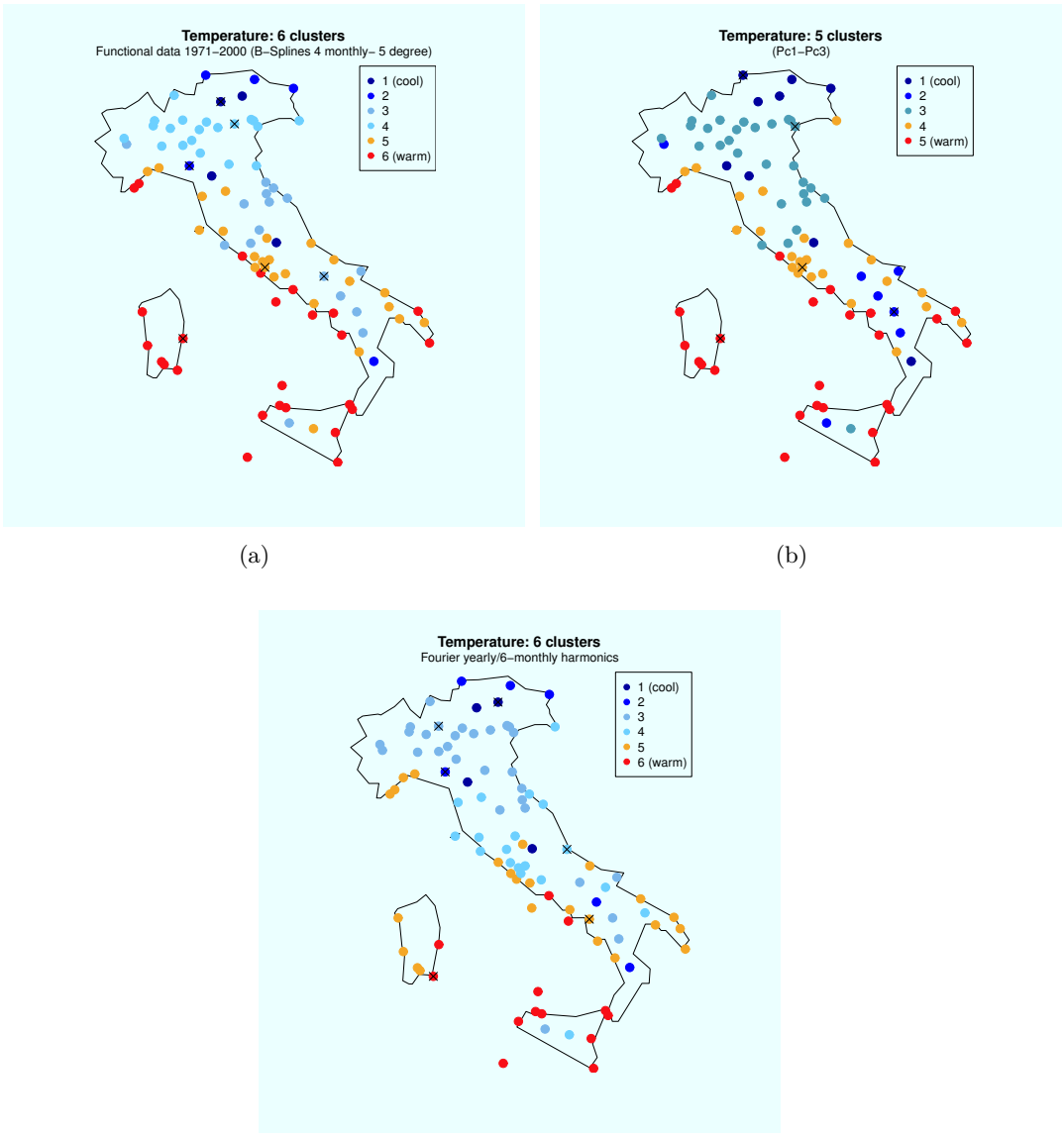
The average silhouette width index reported in panel (a) of Fig. 3.1 is our tools to choose the number of clusters, and we report its value from 2 to 20 groups derived from our chosen model. The best value is obtained with 2 groups which is not very meaningful from a climatological point of view. As mentioned in Section 2.3, we take into account climate features in the choice of the optimal number of clusters and we select the 6 groups partition as a good compromise between average silhouette width value and description of climate features. Moreover the silhouette width values of single groups shown in panel (b) of Fig. 3.1 suggest an appropriate classification with no misclassified units (recall that with misclassified units a negative value of silhouette width index is obtained).



(a) (b)

Figure 3.1. Cluster algorithm assessment of Tmed-MM for 4-monthly variability functional data: (a) Silhouette Average Width for determining the number of clusters; (b) Silhouette width index for each group and for each unit included in the correspondent 6 groups clustering.

Renaming clusters from colder (1) to warmer (6) we obtain the map in Fig. 3.2a) where also the medoids are indicated. The mapping of Bsplines30 functional clustering highlights the following: there are two coldest clusters of mountain stations in the North (cluster 1 and 2); cluster 3 covers a part of Central area mainly close to the Adriatic Sea and some mountain stations in the South which are not included in the northern mountain stations' clusters because of latitude's mitigation effect; cluster 4 represents cold stations of the Northern area; clusters 5 and 6 correspond to warm Southern region near the Adriatic Sea, and nearly the whole of stations located along the Tyrrhenian coast and both the islands of Sicily and Sardinia. By visualizing the smoothed time series of medoids in Fig. 3.3, it is also worth noticing the peculiarity of "hot winters" occurred in mountain regions from 1988 to 1992 (cluster 1 and 2); how cluster 4 differs from cluster 3 for hotter summer temperatures whilst cluster 5 differs from cluster 6 for colder winter temperatures. This cluster analysis can be immediately related to the exposure to the main atmospheric circulations of the different regions. Warm clusters (cluster 5 and 6) location, in the South and along the Tyrrhenian coast, are linked to the south-western flows forced by both cyclonic and anticyclonic circulation over western Mediterranean basin. In this area only mountain stations such as those over the Appennini ridge and Mount Argentario belongs to other clusters. The locations of cold clusters are linked to north-eastern flows driven by cyclonic circulation over East Europe and blocking condition over central Europe that bring cold air masses into the Mediterranean Basin. A detailed summary of monthly and seasonal Tmed-MM 30 years averaged values of each group is given in Fig. 3.4. The Bsplines30 approach leads to results similar to the benchmark PCA in T-mode using 3 Pcs with respect to highlighted climatological features. The PCA based classification returns a unique coldest cluster of mountain stations in the North whereas in Bsplines30, this cluster is more correctly divided into two separated groups. Examining Fig. 3.4, where the general features of cluster are depicted, it appears that the monthly levels of cluster 1 and cluster 2 are clearly different. In panel (c) the Fourier based map is reported. There we choose 6 groups as for the Bsplines30, clusters are very similar however some relevant differences are there: Sardinia is divided into two clusters and several stations around Rome are in a colder cluster with respect to the Bspline30, furthermore the two Rome stations are classified into two different clusters. The general classification has several unclear aspects from a phenomenological point of view. In terms of the best clustering quality Bspline30 with 6 groups reports an average silhouette width of 0.33 with no misclassified stations, PCA with 5 groups is obtained with average silhouette width equal to 0.42 and 2 misplaced units and Fourier with 6 groups reports an average silhouette width of 0.41 and 1 misplaced unit. Say k the number of groups of each cluster, in the case of PCA, the best value of silhouette average width corresponds to k=2, k=3 is the second best and our choice k=5 is the third best. In the Fourier case, the best value of silhouette average width is found for k=3, k=2 is the second best, k=4 is the third and our choice k=6 is the fourth best.



(a) (b)

(c)

Figure 3.2. Cluster maps of Tmed-MM for 4-monthly variability functional data (a), PCA T-Mode method using 3 Principal components (b) and Fourier with 5 basis of 12- and 6-monthly harmonics. Crosses in the maps indicate the location of cluster's medoids.

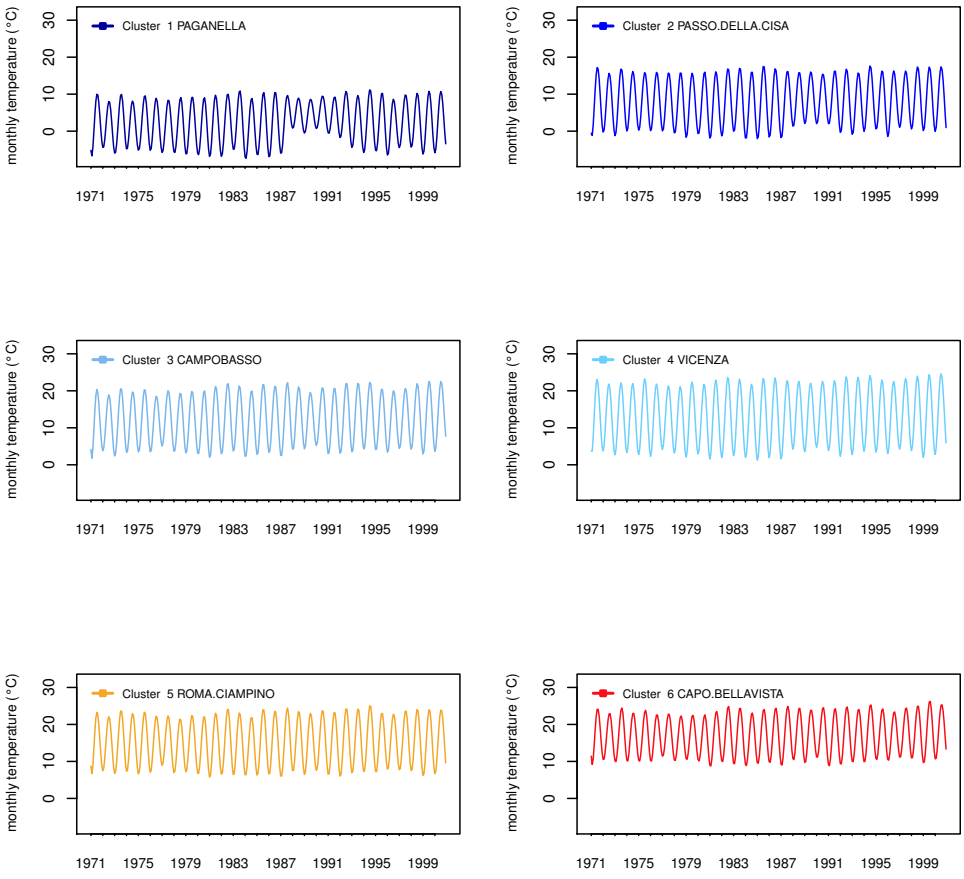
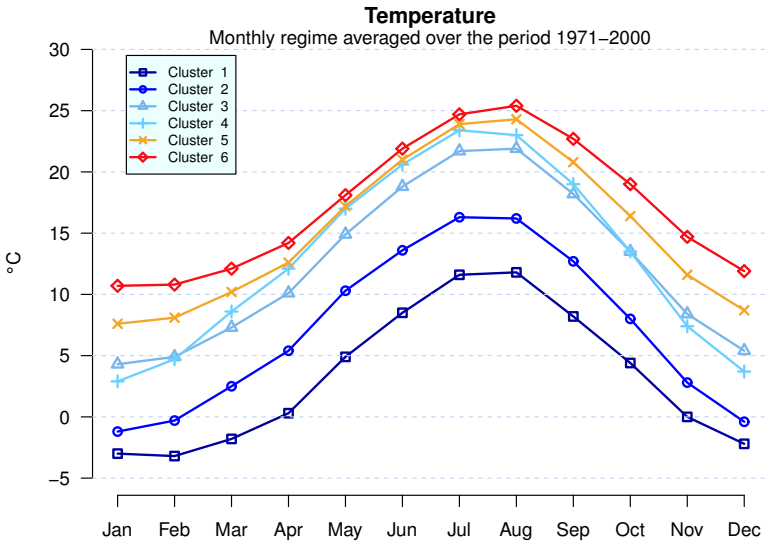


Figure 3.3. Functional smoothing of the 6 medoids of temperature time series 1971-2000 (B-Splines 4-monthly 5-degree).



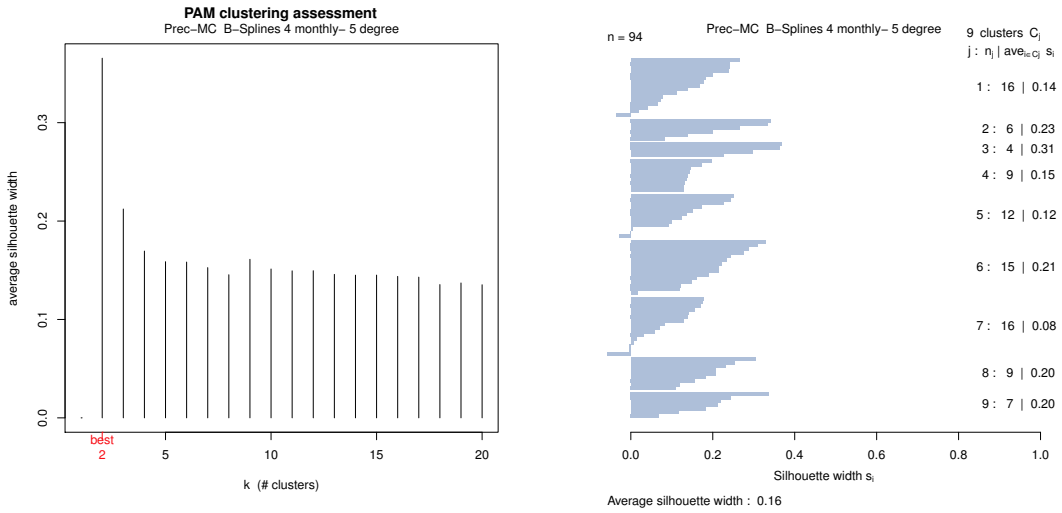
	Units (#)	DJF (°C)	MAM (°C)	JJA (°C)	SON (°C)
Cluster 1	4	−2.8	1.1	10.6	4.2
Cluster 2	5	−0.6	6	15.4	7.8
Cluster 3	16	4.9	10.8	20.8	13.4
Cluster 4	20	3.8	12.6	22.3	13.3
Cluster 5	24	8.1	13.3	23.1	16.3
Cluster 6	26	11.2	14.8	24	18.8

Figure 3.4. Monthly and Seasonal values of Tmed-MM averaged over 1971-2000 for 6 areas delineated by 4-monthly variability functional data (*DJF*: December, January, February; *MAM*: March, April, May; *JJA*: June, July, August; *SON*: September, October, November).

3.2 Results for Monthly Cumulated Precipitation (Prec-MC)

The more appropriate Bsplines30 model for Prec-MC is *Bsplines30 4-monthly 5-degree* with 90 fixed knots and 5 degree piece-wise polynomials. The value of average RMSE is 8.39 millimeters. Similar comments as in Section 3.1 on the model choice apply. Again the smallest average RMSE is obtained with Bsplines30 3-monthly 3-degree but the increase in parameters number (from 94 to 122 parameters to be estimated) (see Table 3.1) and the final classification do not justify the choice of the 3-monthly model. For Prec-MC B-Splines30 3-monthly 3-degree the chosen number of groups is 7 with an average silhouette width of 0.15 which is the fourth best value and the first one with a climatic meaningful interpretation; there are 12 misplaced units according to the silhouette index and the classification is quite consistent with climate patterns. A comparison with our chosen classification reveals that several locations are wrongly classified into cluster 1 (stations along Po river) and two areas are not isolated in single clusters as it should be (stations near the Ligurian sea and Sardinian stations). The Sardinian stations are correctly grouped if we consider the 8 groups clustering, which is the sixth best choice in terms of average silhouette width (0.13) and counts 11 misplaced units. On the contrary, using the chosen model we obtain an acceptable compromise between climatic interpretation of groups and statistical clustering quality indexes. This statement is corroborated by the following results. As far as the number of groups to be chosen, for precipitation data the choice is less straightforward than with temperature data. Thus, very similar values of the average silhouette width index are obtained with 4 up to 20 clusters partitions (see Fig. 3.5a). Nevertheless, in spite of an average silhouette width value of 0.16 and 3 misplaced units (Fig. 3.5b), the 9 clusters partition is "the best" if we take into account all the information associated to the PAM algorithm, as isolation, diameter of clusters, separation and silhouette width of each group. Besides, this clustering returns a representation of climate features of precipitation which is consistent with well known patterns of this variable for the Italian peninsula.

The smoothed time series of 9 medoids represented in Fig. 3.7 reveal the high variability of precipitation and also highlight significant differences between groups. A detailed summary of yearly and seasonal Prec-MC 30 years averaged values of each group is given in Fig. 3.8, where we use line chart instead of bar chart to make a clearer graphical representation of precipitation regime. In the following, we refer to those values for the ordination of the groups from the rainiest to the driest and for a further description of the groups. As it comes out from panel (a) of Fig. 3.6, main patterns of variability are well reproduced and their identification improved with respect to the benchmark in PCA T-mode (panel (b)). In particular, it is worth evaluating the separation of the stations near the Ligurian Sea (cluster 5) and continental stations in the North-West (cluster 1 and 6) into different regions; the clear identification of two precipitation patterns in the northern and southern stations along the Po river (cluster 1 and 6). A central area extending from the Tyrrhenian to the Adriatic Sea (cluster 2) which is the second most rainy region (858 mm of total annual precipitation) behind the northern Po river area (966 mm). A coastal region along Tyrrhenian Sea (cluster 4) is also delineated which is fourth



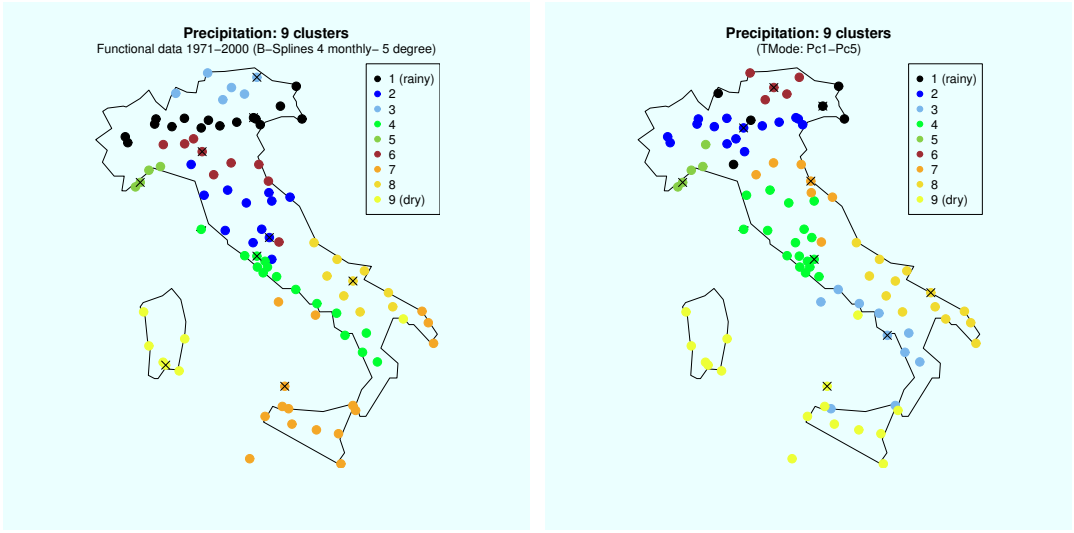
(a) (b)

Figure 3.5. Cluster algorithm assessment of square root Prec-MC for 4-monthly variability functional data: (a) Silhouette Average Width for determining the number of clusters; (b) Silhouette width index for each group and for each unit included in the correspondent 9 groups clustering.

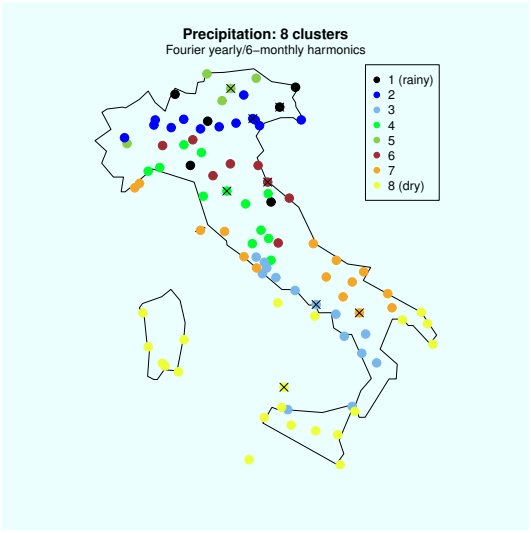
in the rainy ordered classification (807 mm). Regarding the South of the peninsula, Puglia is longitudinally divided into two areas (cluster 7 and 8) because of drier summer regime registered in the southern part (39 mm against 93 mm), which is similar to Sicily precipitation features. This is an improvement of the PCA T-Mode classification. Finally, Sardinia is correctly classified as a unique cluster with driest annual precipitation volume (402 mm) whereas the PCA-based classification proposes a unique cluster of Sicily and Sardinia. Besides, the clustering of stations reflects atmospheric patterns responsible for different precipitation regimes both at large scale and local scale. In particular, the Atlantic storm track determines the grouping of western areas (cluster 2, 4 and 5), of which clusters 2 and 4 are characterized by a prevalence of frontal precipitation and convective events, whereas cluster 5 precipitation signal is due to a more cyclogenetic and convective type of events [Harnik and Chang 2003]. The continental and Alpine regions are characterized by large amount of precipitation due to a orographic enhancement mechanism driven by the presence of mountain ridges (clusters 1 and 3) and a distinct area (cluster 6) in the east side of Appennini lee ridge, which is dryer than other northern clusters since it is not directly exposed to the moist westerly atmospheric flows. Similarly to the case of temperature we perform functional clustering using Fourier basis as well. Following [Laguardia 2011] we adopt 12- and 6-monthly harmonics. The classification map shows noticeable differences with respect to Bsplines30: the locations of the rainiest cluster are far from each other and, moreover, this spatial dispersion does not seem to have a physical motivation; cluster 2 is similar to cluster 1 of Bsplines30; locations by the Ligurian sea do not have a clear identification as it is for our proposal and, finally, the two major Italian islands Sicily and Sardinia

are aggregated in a unique cluster (cluster 8), which is questionable as it is clear by looking at the yearly volume of precipitation of those groups when separated (Fig. 3.8). In terms of best clustering quality Bspline30 reports an average silhouette width of 0.16 with 3 misclassified units, PCA T-mode using 5 Pcs with 9 groups returns a value of 0.31 average silhouette width and 5 misclassified units, while the Fourier based analysis with 8 groups has 0.33 average silhouette width with 7 misplaced units. Say k the number of groups of each clustering, in the case of PCA, the best value of silhouette average width is obtained for k=2, k=11,12 and 13 have the same value of the index which corresponds to the second best and our choice k=9 is the third best. In the Fourier case, the best value of silhouette average width corresponds to k=2, k=3 is the second best, k=4 the third and our choice k=8 is the fourth best.

Finally we replicate, as far as possible, the procedure proposed in [Laguardia 2011] by adopting the same basis for the functional smoothing, i.e. the Fourier basis with 12- and 6-monthly harmonics and the k-means clustering method. Appendix 3.2 contains the map and other information about this part. However in Laguardia's paper the clustering algorithm is not entirely specified and than we choose the default Hartigan and Wong algorithm in the "stats" library of the R software with 25 random starts for the k-means clustering [Hartigan and Wong 1979]. The method for selecting the optimal number of clusters is not specified in Laguardia, than we choose the same number of clusters proposed by Laguardia, i.e. 6 for clustering the 94 stations of our data set. The predicted values of monthly regime obtained from our data set replicating the method proposed in Laguardia are consistent with the climatology of the clusters location (Fig. 3.9). Furthermore, the cross stations RMSE values calculated for each month and reported in Table 3.5 are very similar with respect to the ones in Laguardia. However with the 6 clusters several features captured by the Bsplines30 are not highlighted, and when a larger number of clusters has been tried, the same confusions seen with PAM classification are obtained.



(a) (b)



(c)

Figure 3.6. Cluster maps of precipitation for 4-monthly variability functional data (a), PCA in T-Mode method using 5 Principal components (b) and Fourier with 5 basis of 12- and 6-monthly harmonics (c). Crosses in the maps indicate the location of cluster 's medoids.

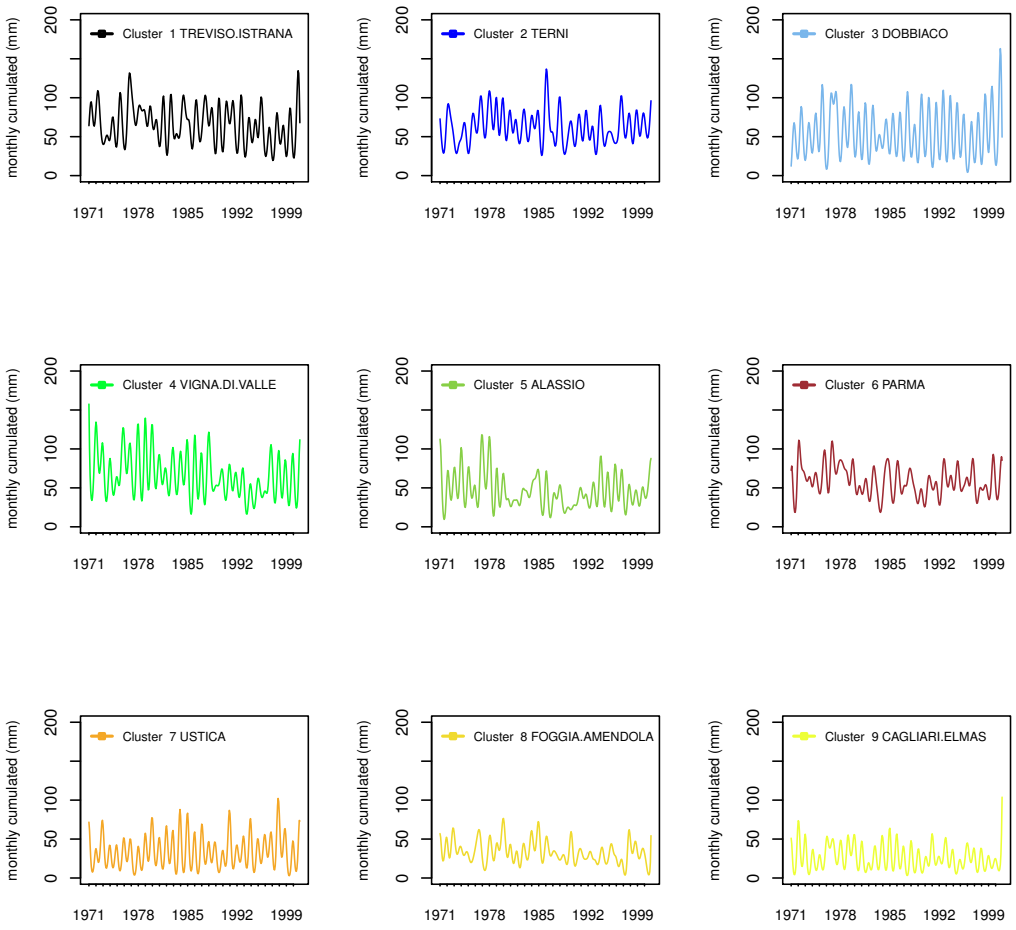
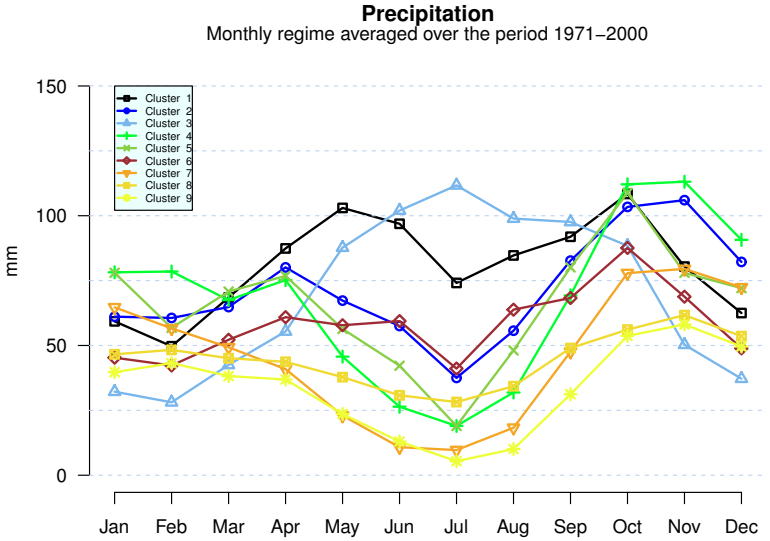


Figure 3.7. Functional smoothing of the 9 medoids of precipitation time series 1971-2000 (B-Splines 4-monthly 5-degree).



	Units (#)	DJF (mm)	MAM (mm)	JJA (mm)	SON (mm)	YEAR (mm)
Cluster 1	16	171.5	258.9	255.6	280.7	966.7
Cluster 2	12	203.8	212.1	150.7	292	858.6
Cluster 3	6	97.6	185.6	312.6	236.3	832.1
Cluster 4	15	247.4	188.4	77.3	294.5	807.6
Cluster 5	4	206.9	203.9	109.2	266.8	786.8
Cluster 6	9	136.5	170.9	164.3	224.8	696.5
Cluster 7	16	193.8	113.3	38.8	205	550.9
Cluster 8	9	148.4	126.7	93.2	166.8	535.1
Cluster 9	7	132.5	98.8	28.5	143	402.8

Figure 3.8. Monthly and Seasonal values of precipitation averaged over 1971-2000 for 9 areas delineated by 4-monthly variability functional data (Seasonal precipitation values are obtained by cumulating monthly values- DJF: December, January, February; MAM: March, April, May; JJA: June, July, August; SON: September, October, November).

Conclusion of Part I

This part of the Thesis has presented a new protocol, based on functional clustering for determining homogeneous climate zones. We showed that by functional clustering, information on temporal pattern relative to the monthly time scale is summarized by a small number of coefficients and those coefficients determine a clear identification of variability mechanisms. The proposed method achieves this goal with a parametrization of function using penalized B-splines basis that returned a clear description of Intra-annual variability. Description of the current distribution of local precipitation is made difficult by the high spatial and temporal variability of this parameter. Nevertheless, the regional distributions obtained not only correspond fairly well to the large, well-known physical regions of Italy, but also go further, improving the classification determined by standard methods. In fact, to identify climate regions using PCAs based methods requires a long and complex analysis of the reduced space to connect it to the physical world. In our approach this is easily achieved by the choice of knots number and locations. Changing place and/or number of interior knots, allows us to investigate different patterns of variability: Long-term variability or trend (yearly variability over at least 30-years interpolation period), Intra-annual variability (bimonthly, quarterly, four monthly or six monthly variability across time series period).

Further development of this approach are possible and have been investigated in our research to some extent (not reported in the present work). For instance a decomposition in trend and short term component of the time series is easily achieved by fitting a B-splines with yearly knots (trend) and a second B-splines with more knots to capture short term features or a Fourier expansion with few harmonics to capture long term cycles. Some caution must be used when using Fourier basis with relatively small number of station such as in our study. Indeed the Fourier expansion reveals a tendency to over-smoothing (not shown in the paper) that influences classification results that may not be very clear, particularly with highly variable quantities such as precipitation. Thus, the Fourier smoothing seems to refer to a numerical smoothing rather than to a physical framework and this drawback might be due to the loss of local element in the time domain. In fact, unlike Bsplines30, the smoothing of Fourier with 12- and 6-months harmonics attempts to reproduce the average features of monthly distribution of the time series smoothing out small and short term changes. Moreover, the reproduction of Fourier predictive regime reveals that the adding of one supplementary harmonics does not let us to catch local element in time domain. However, if a very large number of monitoring stations is available as in [Laguardia 2011], the strong smoothing effect of Fourier basis expansion may mitigate problems deriving from the large variability that is

proper of large datasets.

In general terms our proposal, as described above, creates a very flexible framework in which analysis of climatological features can be carried out. In particular the functional smoothing can be modified including for example both Fourier and penalized B-splines basis, the first to describe periodic components (regime) and the second to describe the trend; this combination of basis is especially effective when the periodicity in the data it is not subject to large changes in the considered time window. Other basis can be considered such as wavelets, or combinations of B-splines with different number of parameters depending always on the aim of the study and type of available data.

In conclusion we believe that the presented functional clustering approach is definitely much more flexible and easier to implement than the current PCAs based methods, regardless the chosen basis representation.

Acknowledgement of Part I

The temperature and rainfall datasets used in this Thesis are from weather stations network of National Centre of Aeronautical Meteorology and Climatology of Italian Aeronautic Army and Agricultural Climatology and Meteorology Research Unit of CRA-Agricultural Research Council.

Appendix of Part I

List of weather stations and monthly missings data

	Weather Station	Tmed-MM missings	%	Prec-MC missings	%
1	ALASSIO	8	2.2	15	4.2
2	ALGHERO	5	1.4	24	6.7
3	AREZZO	1	0.3	5	1.4
4	BARI.PALESE	4	1.1	21	5.8
5	BATTIPAGLIA	2	0.6	5	1.4
6	BOLOGNA.B..PANIGALE	6	1.7	28	7.8
7	BONIFATI	3	0.8	11	3.1
8	BRESCIA.GHEDI	15	4.2	15	4.2
9	BRINDISI	1	0.3	3	0.8
10	CAGLIARI.ELMAS	1	0.3	4	1.1
11	CAMPOBASSO	1	0.3	6	1.7
12	CAPO.BELLAVISTA	7	1.9	24	6.7
13	CAPO.CARBONARA	71	19.7	110	30.6
14	CAPO.FRASCA	4	1.1	10	2.8
15	CAPO.PALINURO	2	0.6	13	3.6
16	CAPRI	109	30.3	116	32.2
17	CATANIA.FONTANAROSSA	13	3.6	34	9.4
18	CHIAVENNA	6	1.7	8	2.2
19	CIVITAVECCHIA	12	3.3	16	4.4
20	COZZO.SPADARO	2	0.6	19	5.3
21	CREMONA	1	0.3	1	0.3
22	DECIMOMANNU	10	2.8	8	2.2
23	DOBBIACO	3	0.8	4	1.1
24	ELBA	49	13.6	56	15.6
25	ENNA	148	41.1	6	1.7
26	FALCONARA	12	3.3	26	7.2
27	FIRENZE.PERETOLA	2	0.6	13	3.6
28	FOGGIA.AMENDOLA	1	0.3	2	0.6
29	FRONTONE	5	1.4	22	6.1
30	FROSINONE	1	0.3		
31	GAETA	25	6.9	27	7.5
32	GENOVA.SESTRI	2	0.6	16	4.4
33	GIOIA.DEL.COLLE	2	0.6	3	0.8
34	GROSSETO	3	0.8	7	1.9
35	GUIDONIA	8	2.2	9	2.5
36	IMPERIA	0	0.0	8	2.2
37	LATINA	2	0.6	2	0.6
38	LATRONICO	4	1.1	12	3.3
39	LECCE	1	0.3	9	2.5
40	MARINA.DI.RAVENNA	10	2.8	23	6.4
41	MESSINA	1	0.3	2	0.6
42	MILANO.LINATE	3	0.8	27	7.5
43	MILANO.MALPENSA	8	2.2	11	3.1
44	MONTE.ARGENTARIO	1	0.3		
45	MONTE.CIMONE	3	0.8	12	3.3
46	MONTE.SANT.ANGELO	2	0.6	7	1.9
47	MONTE.SCURO	2	0.6	3	0.8
48	MONTE.TERMINILLO	15	4.2	36	10.0
49	NAPOLI.CAPODICHINO	14	3.9	18	5.0

Table 3.2. Number and percentage of Tmed-MM and Prec-MC monthly missings data for each station.

	Weather Station	Tmed-MM missings	%	Prec-MC missings	%
50	NOVARA.CAMERI	8	2.2	10	2.8
51	ORIO.AL.SERIO	7	1.9	25	6.9
52	PAGANELLA	2	0.6	5	1.4
53	PALERMO.BOCCADIF.	162	45.0	165	45.8
54	PALERMO.PUNTA.RAISI	5	1.4	29	8.1
55	PANTELLERIA	1	0.3	18	5.0
56	PARMA	8	2.2	10	2.8
57	PASSO.DELLA.CISA	49	13.6	61	16.9
58	PASSO.ROLLE	25	6.9	28	7.8
59	PESARO	108	30.0		
60	PESCARA	2	0.6	13	3.6
61	PIACENZA	3	0.8	4	1.1
62	PISA.SAN.GIUSTO	1	0.3	6	1.7
63	PONZA	2	0.6	7	1.9
64	POTENZA	26	7.2	38	10.6
65	PRATICA.DI.MARE	8	2.2	15	4.2
66	PRIZZI	4	1.1	12	3.3
67	REGGIO.CALABRIA	8	2.2	4	1.1
68	RIMINI	2	0.6	5	1.4
69	ROMA.CIAMPINO	1	0.3	6	1.7
70	ROMA.FIUMICINO	3	0.8	9	2.5
71	ROMA.URBE	14	3.9	21	5.8
72	SALO.	1	0.3	2	0.6
73	SAVONA	28	7.8	40	11.1
74	S..MARIA.DI.LEUCA	2	0.6	7	1.9
75	S..VALENTINO.ALLA.MUTA	21	5.8	27	7.5
76	TARANTO	9	2.5	14	3.9
77	TARVISIO	3	0.8	10	2.8
78	TERMOLI	15	4.2	18	5.0
79	TERNI	0	0.0	8	2.2
80	TODI	0	0.0	16	4.4
81	TORINO.BRIC.CROCE	2	0.6	6	1.7
82	TORINO.CASELLE	9	2.5	32	8.9
83	TRAPANI.BIRGI	1	0.3	3	0.8
84	TREVICO	13	3.6	23	6.4
85	TREVISO.ISTRANA	4	1.1	5	1.4
86	TREVISO.SANT.ANGELO	5	1.4	5	1.4
87	TRIESTE	1	0.3	15	4.2
88	URBINO	2	0.6	2	0.6
89	USTICA	2	0.6	5	1.4
90	VENEZIA.TESSERA	4	1.1	22	6.1
91	VERONA.VILLAFRANCA	1	0.3	4	1.1
92	VICENZA	3	0.8	6	1.7
93	VIGNA.DI.VALLE	1	0.3	5	1.4
94	VITERBO	2	0.6	12	3.3
95	VOGHERA	0	0.0	2	0.6
96	BOLZANO			38	10.6
97	UDINE.RIVOLTO			2	0.6

Table 3.3. Number and percentage of Tmed-MM and Prec-MC monthly missings data for each station.

Summary statistics	Prec-Mc outlier values (mm)
Min.	138.4
1st Qu.	217.0
Median	263.2
Mean	276.9
3rd Qu.	321.0
Max.	525.3
Number of deleted	150

Table 3.4. Prec-MC: summary statistics of outliers detected by MAD-based test in the monthly series of data.

Comparison with method proposed in Laguardia's paper

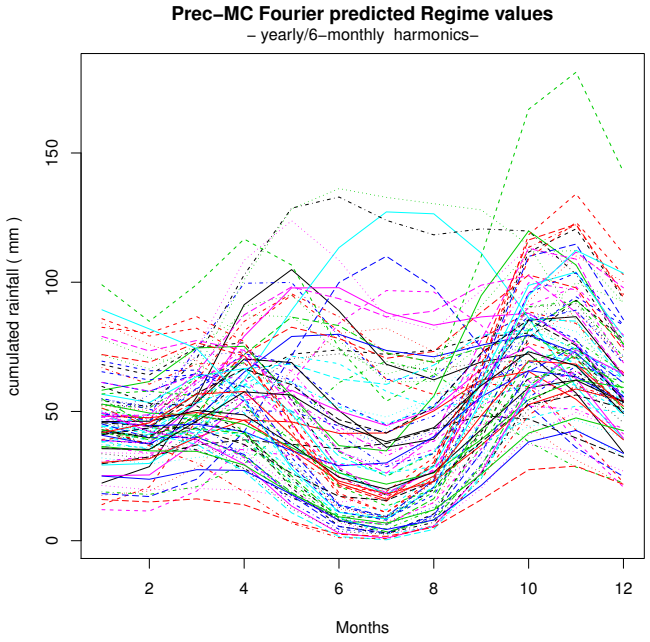


Figure 3.9. Monthly regimes predicted by Fourier 12- and 6-monthly harmonics functional smoothing for the overall stations.

Month	RMSE (mm)
Jan	18.50
Feb	9.82
Mar	13.61
Apr	7.20
May	10.05
Jun	15.62
Jul	21.03
Aug	14.79
Sep	10.72
Oct	16.28
Nov	20.13
Dec	5.11

Table 3.5. RMSE of monthly regime predictions obtained by Fourier 12- and 6-monthly harmonics functional smoothing.

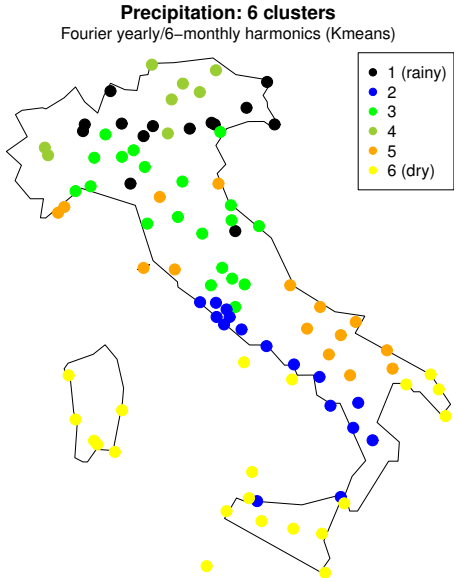


Figure 3.10. Clustering obtained replicating the method proposed in Laguardia's paper: Fourier 12- and 6-monthly harmonics functional smoothing and kmeans clustering algorithm with 6 groups.

Bibliography

[Abraham et al. 2003] Abraham C, Cornillon PA, Matzner-Loeber E and Molinari N (2003) Unsupervised Curve Clustering using B-Splines. Scand J Stat 30:581-595.

[Box and Cox 1964] Box GEP and Cox DR (1964) An analysis of transformations. J Roy Stat Soc B 26:211-246.

[Brunetti et al. 2006] Brunetti M, Maugeri M, Monti F, Nanni T (2006) Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. Int J Climatol 26:345-381.

[Ehrendorfer 1987] Ehrendorfer M (1987) A regionalization of Austria's precipitation climate using principal component analysis. Int J Climatol 7,1:71-89.

[Fovell and Fovell 1993] Fovell RG and Fovell MYC (1993) Climate Zones of the Conterminous United States Defined Using Cluster Analysis. *J Climate* 6:2103-2135.

[Harnik and Chang 2003] Harnik N and Chang EKM (2003) Storm Track Variations As Seen in Radiosonde Observations and Reanalysis Data. J Climate 16:480-495.

[Hartigan and Wong 1979] Hartigan JA and Wong MA (1979) A K-means clustering algorithm. Applied Statistics 28: 100-108.

[Kaufman and Rousseeuw 1990] Kaufman L and Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

[Kim et al. 2008] Kim BR, Zhang L, Berg A, Fan J, Wu R (2008) A Computational Approach to the Functional Clustering of Periodic Gene-Expression Profiles. Genetics 180:821-834.

[Ignaccolo et al. 2008] Ignaccolo R, Ghigo S and Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. Environmetrics 19:672-686.

[Laguardia 2011] Laguardia G (2011) Representing the precipitation regime by means of Fourier series. Int J Climatol 31,9:1398-1407.

[Mennella 1972] Mennella C (1972) Il clima d'Italia nelle sue caratteristiche e varieta' quale fattore dinamico del paesaggi. Vol. II. Fratelli Conte Editore, Napoli.

[Preisendorfer et al. 1988] Preisendorfer RW, Mobley CD (1988) Principal Component Analysis in Meteorology and Oceanography. Elsevier, Amsterdam.

[R Development Core Team 2011] R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[Ramsay and Silverman 1997] Ramsay JO and Silverman BW (1997) Functional Data Analysis. Springer-Verlag, New York.

[Ramsay and Silverman 2002] Ramsay JO and Silverman BW (2002) Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag, New York.

[Ramsay et al. 2011] Ramsay JO, Wickham H, Graves S and Hooker G (2011) Fda: Functional Data Analysis. R package version 2.2.7 http://CRAN.R-project.org/package=fda.

[Reynolds et al. 1992] Reynolds A, Richards G, De La Iglesia B and Rayward-Smith V (1992) Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5: 475-504 http://dx.doi.org/10.1007/s10852-005-9022-1.

[Richman 1986] Richman MB (1986) Rotation of principal components (review article). J Climatol 6:293-335.

[Rousseeuw 1987] Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53-65.

[Sprent 1998] Sprent P (1998) Data driven statistical methods. Chapman & Hall, London.

[Sprent and Smeeton 2001] Sprent P and Smeeton NC (2001) Applied Nonparametric Statistical Methods (3rd ed.). Chapman & Hall/CRC, London.

[Toreti et al. 2009] Toreti A, Fioravanti G, Perconti W and Desiato F (2009) Annual and seasonal precipitation over Italy from 1961 to 2006. Int J Climatol 29,13:1976-1987.

[Trigo et al. 2006] Trigo RM and Coauthors (2006) Chapter 3 Relations between variability in the Mediterranean region and mid-latitude variability in Mediterranean Climate Variability. Lionello P, Malanotte-Rizzoli P and Boscolo R Eds., Developments in Earth and Environmental Sciences, Vol. 4, Elsevier, 179-226.

[Von Storch and Zwiers 1999] Von Storch H and Zwiers FW (1999) Statistical analysis in Climate Research. Cambridge University Press, Cambridge.

[WMO 1989] WMO (1989) Calculation of Monthly and Annual 30-Year Standard Normals. WCDP n.10, WMO-TD/N.341, Geneva.

Part II

Scan statistic and Bayesian Spatio-temporal models

Introduction to Part II

In the second part of the Thesis we build a hierarchical Bayesian model aiming at the prediction of 15-minutes and 30-minutes accumulated precipitation at unknown locations and time using information on lightnings in the same area. More generally, this work is motivated by the poorness of satellite precipitation datasets since rainfall fields estimates derived from satellite sensors are affected by several limitations. Firstly, they do not allow to differentiate between *convective* and *strati-form* rainfall events and, subsequently, they induce systematic errors in the estimates of hourly/sub-hourly rain rates. In fact, convective rainfall events typically can extend over a relatively small area and have a brief lifetime during which large quantity of precipitation is generated whilst strati-form rainfall events can extend over a large area generating large quantity of precipitation as well but during a larger amount of time. Consequently, the rain rates produced by the two types of rainfall systems are quite different and their effects substantially differ from one to another. For instance, soil erosion is often due to high rainfall rates. Secondly, they underestimate the total amount of precipitation during extreme events, occasionally hiding the real causes of landslide or floods, for example. Finally, they are often poor in mountain areas or areas far from satellite axis. As a consequence of these considerations, our final goal is to improve satellite rainfall fields estimate using the information content in the lightning events. Lightnings can be generated either in convective or in strati-form systems although a well delineated spatial and temporal propagation of lightnings has to be associated to convective events, exclusively. Thus, lightning information can enhance our knowledge for discriminating convective rain areas within a cloud system. Convective rainfall events can be self independent, like Thermal Convective System (TCS) or included within strati-form systems. Our attention is devoted to those convective events generated within *Mesoscale Convective Systems* (MCS). Finally, it is worth noticing that the most of rainy clouds in MCS are lightning free [Palmeira et al.] [Morales and Anagnostou 2003]. The work-flow implementation can be described in four steps:

1. Identification of rainfall "convective events";
2. Estimate of Rainfall Lightning Ratio (RLR);
3. Predicting rainfall from lightnings records;
4. Calibration model for satellite rainfall fields estimate.

The first issue to be addressed is the identification of single storms (events) among several severe meteorological events. Then our first contribution in this study, is a simple and effective scan statistic procedure to separate 'events'. These events are *convective storms* from which either precipitation or lightnings might be generated. From identified convective events we estimate the Rainfall Lightning Ratio which determines the mass of rain associated to each flash. Then, we introduce the basic Tapia-Smith-Dixon model used to estimate rainfall fields from lightnings. Finally, we present the Mixed Models approach in which precipitations are modeled as function of lightnings counts (fixed effects) and space time variation is handled using

specific random effect. The space-time random effect is modeled as separable, with a Conditional Autoregressive model (CAR) to model the spatial random component and a simple AR(1) model to represent time variation. The area of study is located in Central Italy and the study events regards the storms of 5th of August 2004 and 9th of May 2006. The database is composed of lightnings records (instant-point fields) and the weather stations precipitation records (sub hourly-point fields). The fourth step of the work is not a part of this Thesis and will be developed in next future.

Chapter 4

A non-physical approach to identify convective events by means of lightning records: scan statistic

4.1 Description and phenomenology of lightnings activity

Lightning activity is closely related to atmospheric instability and is due to the transport of latent heat inside clouds systems such as storm convection. More specifically, lightnings are generated whenever it exists a polarization of positive and negative electric charges between two object such as storm cloud, air or ground. A lightning is the transfer of massive electrostatic charge between any two of those objects. Thus, lightning discharges can occur Intra-Cloud (IC), Cloud-to-Cloud (CC), Cloud-to-Air (CA) and Cloud-to-Ground (CG). The latter have the primary interest of researchers. CG-lightnings are detected by means of at least three sensors located on the Earth's surface which record the electromagnetic signal emitted by the lightning return stroke. The intensity and polarity of the signal are registered for each lightning along with the time of event and the impact point (Latitude and Longitude).

In relation to rainfall fields estimate, the most common use of lightnings data is now-casting, particularly to forecast the volume of precipitation expected over the impact area of storms, tornadoes, etc [Tapia et al. 1998]. An implementation in this field could be done using a point process approach, even though several works use this technique for estimating rainfall fields from satellite data but not from lightnings data (see for example the seminal work of [Rodriguez-Iturbe et al. 1987] or [Wheater et al. 1999]). On the other hand, lightnings data are also used as forcing factors in physical models to improve the forecast performance [Fierro et al. 2012].

On the contrary, the final goal of this work is the the building of tools for *ex post* correction of rainfall fields derived from satellite data using the count and space-time propagation of CG-lightnings (Cloud to Grounds) and their relation to rainfall convective events. Following Morel et al. 2002 [Morel and Senesi, 2002], we refer to the population of rainfall convective events as part of larger *Mesoscale Convective Systems* (MCS). MCS produce a significant fraction of the warm season rainfall, lightning activity and severe weather in the mid latitudes of Northern hemisphere. Within each MCS, the **smallest rainfall convective systems** produce a contiguous precipitation area of about 100 km in one horizontal direction and about 3 hr of duration whilst the **largest convective systems** can extend over about 500 km in one horizontal direction and persist for about 20 hours. A physical description of the convective systems development within MCS systems is given in [Parker and Johnson 2000], integrated in a subsequent paper by Parker et al. 2001 [Parker et al. 2001]. The scheme of Figure 4.1 reports three typical shapes of convective storms inside an MCS.

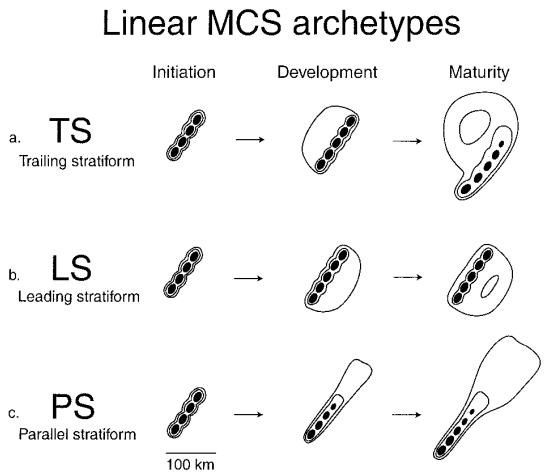


Figure 4.1. The scheme of life cycles for three linear MCS archetypes (from Parker and Johnson 2000 [Parker and Johnson 2000]): (a) leading line trailing stratiform (TS), (b) convective line with leading stratiform (LS), (c) convective line with parallel stratiform (PS).

4.2 Scan statistic procedure

The first step is devoted to use **time-sequence** of lightnings for detecting "rainfall convective events". This complex operation is usually done by analyzing the atmospheric circulation by means of satellite images or others instruments. To isolate a single convective event is fundamental for any type of rainfall prediction. Here, we propose to isolate rainfall convective events using the lightnings data by means of a scan statistic procedure. Firstly, we individuate a daily significant lightnings aggregation using the marginal distribution of hourly lightnings counts over the gridded study region (see panel (b) of figures 4.3 and 4.5). Secondly, in order to capture events taking place around the boundary between two subsequent days our time-window extends from 6 pm of the previous day to 6 am of the next, allowing for a 6 hours overlap between adjacent days. The scan statistic procedure is performed in order to identify the beginning and the end of each convective event.

In particular, we use the **scan statistic procedure** proposed by Ester et al. in 1996 [Ester et al. 1996]. This procedure consists of an iterative algorithm with 2 parameters: the **radius** of a circle drawn around each record and the **minimum number of records** inside the circle which determines the initial record to be clustered or not. In addition to the UTM coordinates (Universal Transverse Mercator system) which uniquely identify the location of each lightning in the $\Re^2$ spatial domain, we consider the instant of lightning hit since the temporal sequence of lightnings determines the clustering of the convective events. Thus, the algorithm scans each item (lightning) in a sphere of the $\Re^3$ spatial-temporal domain. Because of the different scale of measure, we standardize the 3 variables and, subsequently, we find a value of the sphere radius equals to 0.3 by means of K-dist criterion having set the minimum number of items at 10. The graphical representation of the procedure reported in Fig. 4.2 is taken from the paper of [Ester et al. 1996].

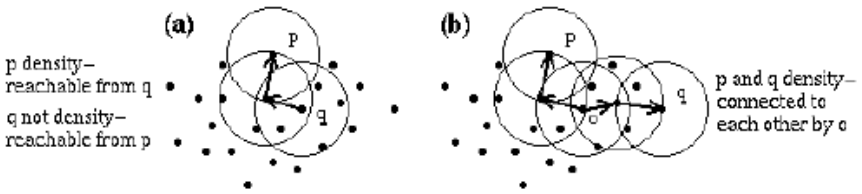


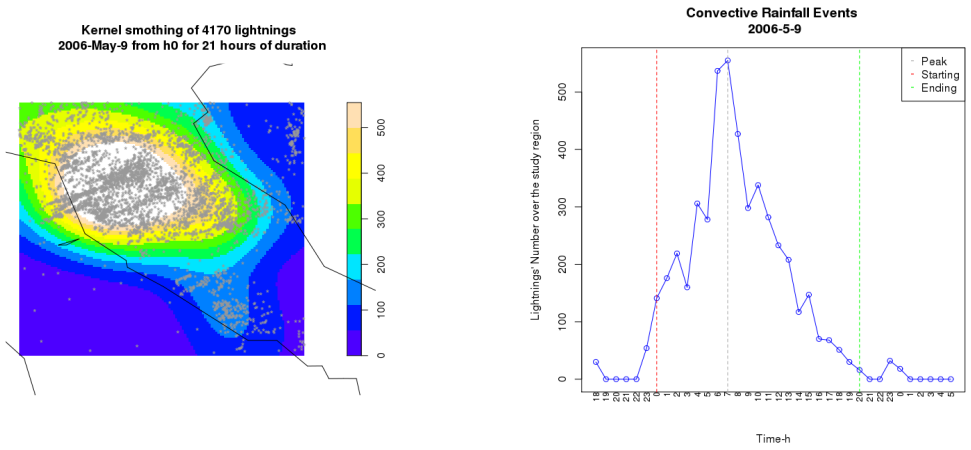
Figure 4.2. The density-based scan clustering as represented in [Ester et al. 1996].

The algorithm is included in R package *fpc* [Henning, 2010]. For the purpose of clustering different convective events which are eventually occurring during the same day, we also experimented the *spatgraphs* R-library [Rajala 2012], which compute a general adjacency of a given point pattern. We set a geometric adjacency determining a spatial clustering of lightnings by means of a connection radius. However, the clustering obtained by applying this alternative method present some limitations for several tested convective events.

In the next two paragraphs the results of scan statistic procedure applied to the storms on the 9th of May 2006 and the 5th of August 2004 are reported. The maps in

Figure 4.4 and 4.6 are done using the `rgeos package` of R [Bivand and Rundel 2012].

Convective events isolated during the storm of 9th May 2006 The scan statistic procedure has identified 4 convective events during the storm of May 9, 2006.



(a) (b)

Figure 4.3. Identification of 9th-May-2006 convective event by means of marginal distribution of lightnings spatial patterns.

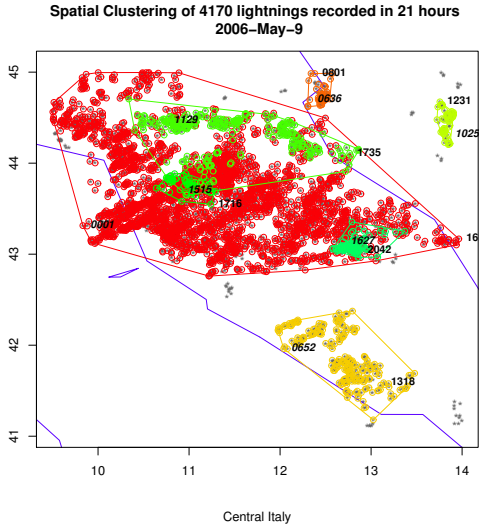
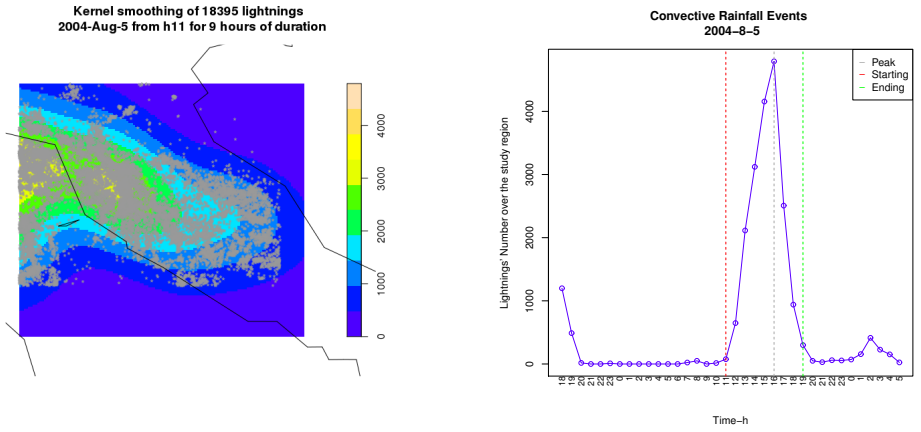


Figure 4.4. Convective events clustered during the storm of 9th May 2006 by scan statistic procedure.

Convective Events	Event A (red)	Event B (orange)
Starting Time	00.01	06.36
Ending Time	16.14	08.01
Duration (min)	974	86
Number of CG-Lightnings	3163	50
Convective Events	**Event C (yellow)**	**Event D (green)**
Starting Time	06.52	11.29
Ending Time	13.18	17.35
Duration (min)	387	367
Number of CG-Lightnings	247	327

Table 4.1. List and details of convective events isolated by scan statistic procedure during the storm of 9th May 2006.

Convective events isolated during the storm of 5th Aug 2004 The scan statistic procedure has identified 2 convective events during the storm of Aug 5, 2004.



(a) (b)

Figure 4.5. Identification of 5th-Aug-2004 convective event by means of scan statistic procedure.

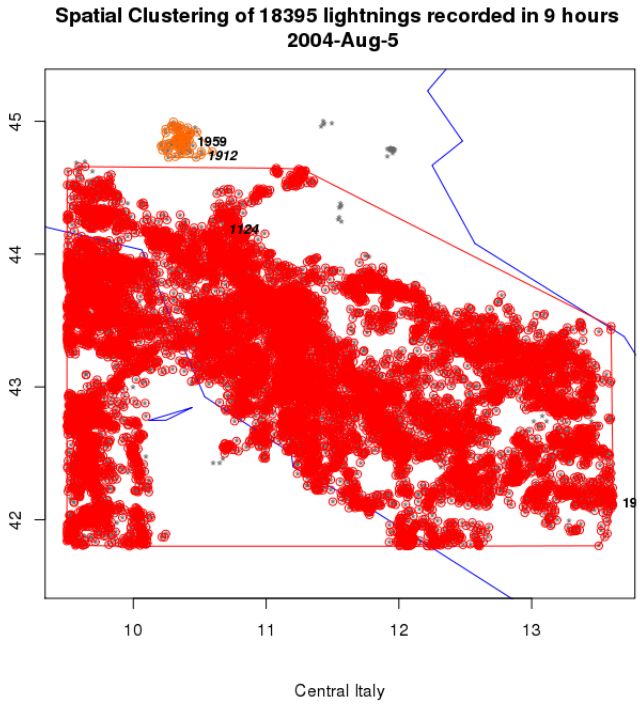


Figure 4.6. Convective events clustered during the storm of 5th August 2004 by scan statistic procedure.

Convective Events	Event A (red)	Event B (orange)
Starting Time	11.24	19.12
Ending Time	19.56	19.59
Duration (min)	513	47
Number of CG-Lightnings	18140	50

Table 4.2. List and details of convective events isolated by scan statistic procedure during the storm of 5th August 2004.

4.3 Analysis of convective events in Central Italy

The scan statistic procedure presented in Section 4.2 is able to identify a great number of convective events. The analysis of spatial and temporal distribution of these events lets eventually help in climatic studies of rainfall storms occurring over the study region. Those types of analysis are beyond the scope of this Thesis. Nevertheless, we present here a part which is fundamental for the building of the model since let us to estimate the Rainfall Lightning Ratio (see further in Section 5.3.2). The events with at least 50 CG-lightnings identified by scan statistic procedure are 767 and are shown in Figure 4.7, where the bubble's area is proportional to the event's total number of lightnings whilst the bubble's center corresponds to the centroid of the event. The largest event counts 33364 lightnings.

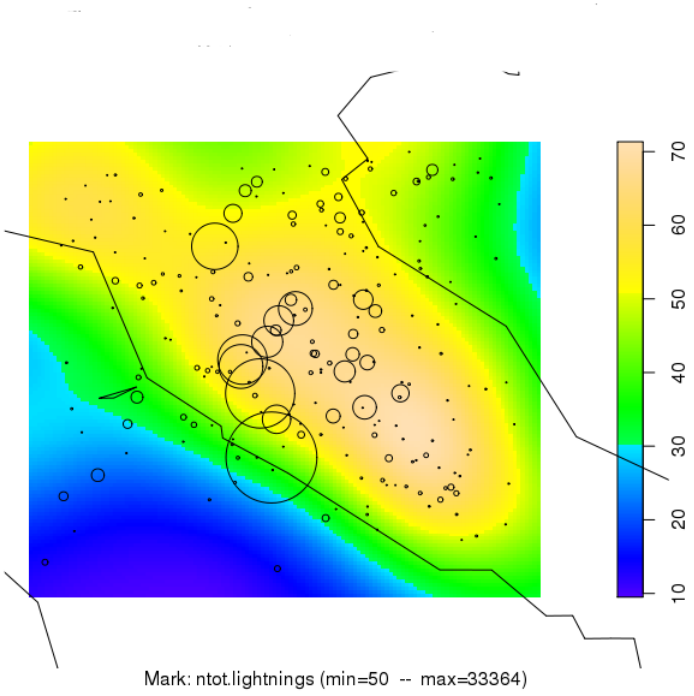


Figure 4.7. Spatial distribution of convective events in Central Italy

We define four categories of convective events based on the total number of lightnings: *Small*, *Medium*, *Large* and *Very Large* events. These categories are described in Table 4.3, where also the number of cases in each class is shown. The choice of thresholds is done using the information from the frequency distribution of events per classes of dimension, i.e. number of lightnings (Fig. 4.8). In this Figure, the dotted green lines indicate the thresholds of 170 and 900 lightnings whereas the dotted red line is drawn at 8000 lightnings. Moreover, thresholds for each category are chosen in order to guarantee an adequate number of cases in each class since those cases are taken as reference for calculating the Rainfall Lightning

Ratio. For this scope, we merge the *Large* and *Very Large* categories re-defining the *Large* category with number of lightnings greater than 900. However, the Very Large definition is used ahead in this paragraph for description purposes, exclusively.

Dimension	Number of Lightnings	Number of cases
Small	≤ 170	403
Medium	$(170, 900]$	270
Large	$(900, 8000]$	84
Very Large	> 8000	10

Table 4.3. Classes of dimensionality of convective events defined on the basis of lightnings number.

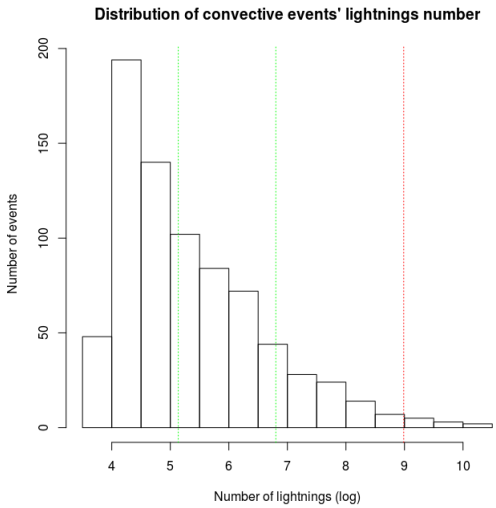


Figure 4.8. Number of events per classes of total lightning's number generated within each one. The dotted green lines indicate the thresholds of 170 and 900 lightnings whereas the dotted red line is drawn at 8000 lightnings.

The monthly percentage of cases in each class of event dimension for the period March to September is reported in Table 4.4 whereas the monthly distribution of convective events as well as a spatial distribution per month are presented in Figure 4.9.

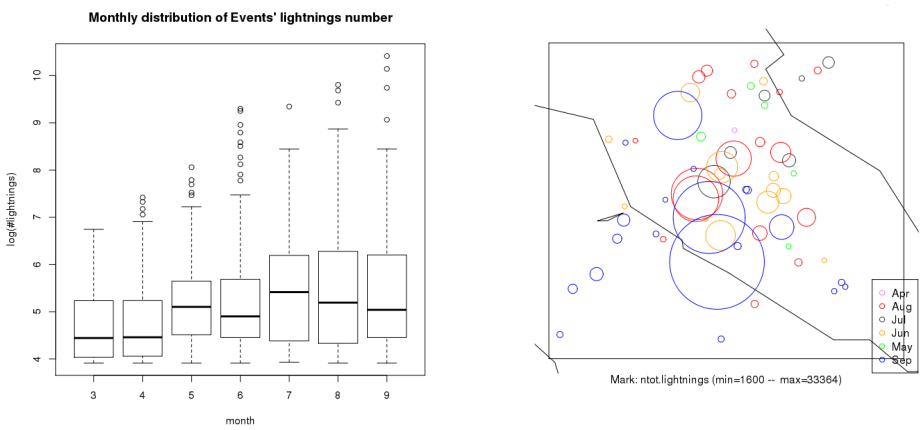
From both Table 4.4 and Figure 4.9, it is worth evaluating the following phenomenological features:

- **largest events** are grouped in the central part of the study area entering from Thyrrenian sea (east) and being blocked by the Appenini's mountain (Fig. 4.7);
- the **average monthly values** of the count of total lightnings is smaller in March-April than in July-August (Fig. 4.9 panel (a));

	Classes of lightnings number			
	$[50, 170]$	$(170, 900]$	$(900, 8000]$	$(8000, 33364]$
March	66.7	33.3	0.0	0.0
April	66.1	25.8	8.1	0.0
May	52.0	39.2	8.8	0.0
June	56.6	30.1	11.8	1.5
July	44.9	46.5	7.9	0.8
August	48.0	35.8	14.5	1.7
September	52.9	30.7	13.6	2.9
Mar-Sept	52.5	35.2	11.0	1.3

Table 4.4. Monthly percentage of cases in each class of event dimension.

- **large events** are more numerous in June, August and September, however, **very large events** are mostly concentrated in September (Tab. 4.4 and Fig. 4.9 panel (b));
- **delineated monthly patterns are**: April along Appenini's mountain, July-August along Appenini and north-east side and September by Tyhrrenian sea coast (Fig. 4.9 panel (b)).



(a) (b)

Figure 4.9. Monthly distribution of convective events identified by means of scan statistic procedure: a) monthly boxplot; b) spatial visualization.

Chapter 5

Lightnings-rainfall relation

5.1 The dataset

The study area is located in Central Italy and is identified by the geographical coordinates 41°-45°LatN, 9.5°-14°LonE. The dataset is composed of three databases covering the time span March-September 2003-2006: the lightnings instantaneous records, the satellite hourly precipitation fields on a 10x10 km regular grid and the weather stations hourly and sub-hourly precipitation records:

Type 1 Lightnings data report the locations and dates of all registered cases of Cloud to Ground (CG) lightnings within the area of study. The CG-lightnings are instantaneously recorded by several sensor located on the earth surface, which detect the electromagnetic field emitted by any cloud-ground lightning (CESI-Sirf [CESI-Sirf] database acquired by Consorzio Lamma (Regione Toscana Cnr-Ibimet) [Consorzio Lamma]).

Type 2 Global satellite precipitation data are distributed by the project GSMaP by the Earth Observation Research Center, Japan Aerospace Exploration Agency (JAXA). The project GSMaP is sponsored by JST-CREST and promoted by the JAXA Precipitation Measuring Mission (PMM) Science Team ([Okamoto et al. 2005];[Kubota et al. 2007];[Aonashi et al. 2009]).
The $GSMaP_MVK+$ dataset has a grid resolution of 0.1 degree lat/lon and a temporal resolution of 1 hour. The estimates of the surface rainfall rates are obtained as a combination of the infrared brightness temperature by the GEO-IR satellites and microwave radiometer estimates, by means of Kalman filter techinique [Ushio et al. 2009].

Type 3 Point precipitation data are composed of hourly and sub-hourly observations time series coming from 181 weather stations located within the study area [Consorzio Lamma].

It is worth noticing that the registration's time of rainfall records from rain gauges follows a different codification with respect to satellite derived records and CG-flashes. In fact, the gauges rainfall recorded at time t indicates the rain accumulated during the time interval $(t-1, t]$ whilst the satellite rainfall record as well as the number of flashes at time t represents the rain accumulated and the whole flashes recorded during the time interval $[t, t+1)$, respectively. Consequently, the 3 databases require spatial and temporal alignement to be used jointly.

5.2 Analysis of lightnings activity and rainfall in Central Italy during the period 2003-2006

In this section an analysis of the three databases of lightnings, satellite rainfall and weather stations rainfall described in Section 5.1 is presented. The curves in Figure 5.1 let us to compare the monthly distribution and the diurnal cycle of lightnings and rainfall which have been recorded over the study area from 2003 to 2006. The graphical comparison is possible since both rainfall and lightnings are drawn on the percentage scale, i.e. monthly levels on total or hourly levels on total. The panel a) of the Figure 5.1 reports the monthly distribution from January to December whilst the diurnal cycle represented in panel b) is calculated during the period March-September, which is the period under analysis. In fact, Mesoscale Convective Systems, which generate the events we are attempting to model in this work, mainly occur during this time window (see Section 4.1 for details).

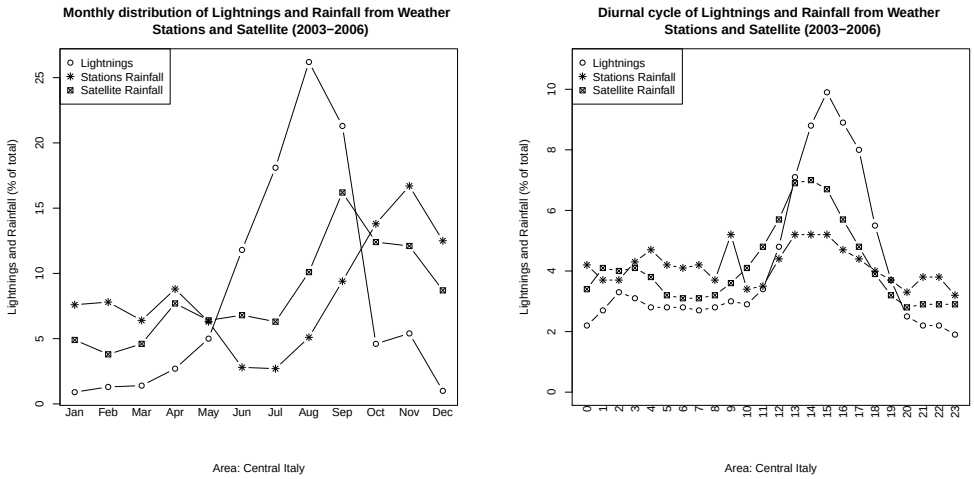
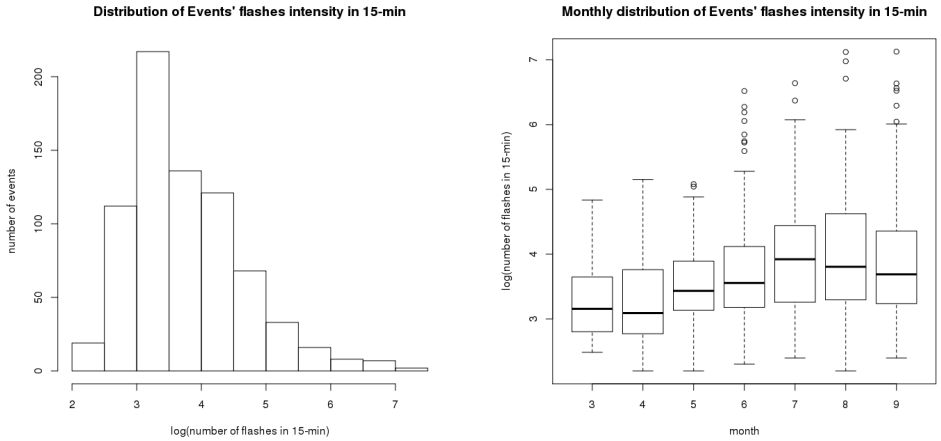


Figure 5.1. CG-lightnings, weather stations and satellite-derived rainfall in Central Italy: a) Monthly distribution; b) Diurnal cycle during the warm season March-September.

The analysis of monthly distribution depicted in panel a) of Figure 5.1 shows that the lightning activity in Central Italy is mostly concentrated during the period May-September with a peak in August which counts roughly 25% of yearly total of CG-lightnings. On the other hand, the rain activity is characterised by two peaks in Spring and Autumn and a dry period during Summer. This bimodal form of the rain distribution is typical at Mid Latitudes. Notice that satellite and rain gauges precipitation data differ from each other, being satellite data summaries in contrast with the well known features of monthly Italian rain activity [Di Giuseppe et al. 2013]. In particular, rainfall levels of summer months are higher in satellite data than in rain gauges data and also September is the most rainy according to satellite data whilst the peak of rain occurs in November in the case of rain gauges data.

The diurnal cycle reproduced in panel b) of Figure 5.1 shows that lightnings activity is more intense during the afternoon as well as the rain activity. However, rain peak typically occurs around 2 pm whilst lightning peak is one hour later. Here, satellite data and rain gauges data coincide with each other except for a stronger activity in the early morning recorded by rain gauges.

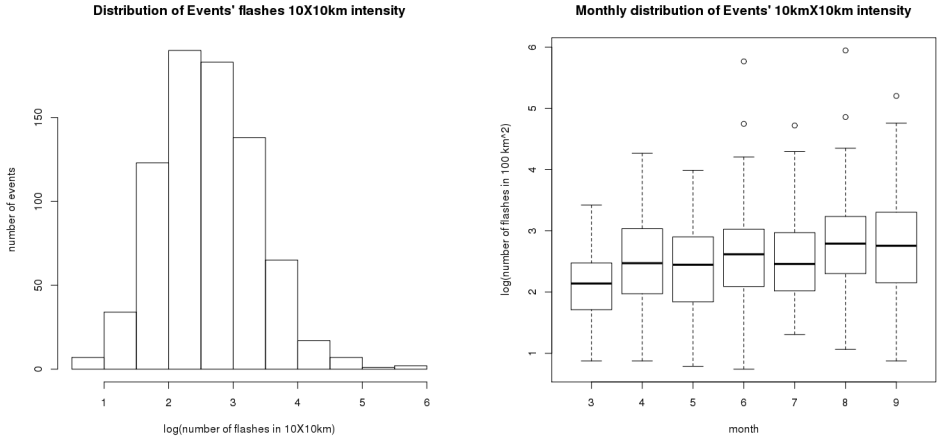
We also make an analysis of the intensity of the convective events in terms of the average number of lightnings in 15-minutes time interval and average number of lightnings at $10 \times 10\,km$ cell. We call the former *temporal intensity* and the latter *spatial intensity*. Notice that this analysis is performed exclusively using the 767 convective events identified by means of the scan statistics procedure which is described in Section 4.2. The most of the events have an average temporal intensity that ranges from 20 to 54 lightnings in 15-minutes, however events in the right tail of the distribution can reach 400 lightnings in 15-minutes (see Fig. 5.2a). Moreover, the analysis of boxplot in Figure 5.2b reveals that the highest averaged 15-minutes temporal intensity is in July whilst events with the higher 15-minutes temporal intensity occur in August and September.



(a) (b)

Figure 5.2. Averaged intensity of lightnings in 15-minutes time interval during the warm season March-September: a) Frequency distribution; b) Monthly distribution.

The analysis of spatial intensity reveals that roughly 650 out of 767 events have an average number of CG-lightnings recorded in $10 \times 10\,km$ cell which ranges from 5 to 33. The most extreme value of spatial intensity is around 400 CG-lightnings. Thus, the distribution of spatial intensity along the 767 convective events is more homogeneous with respect to temporal intensity and also the boxplot depicted in panel b) of Figure 5.3 shows a similar distribution of spatial intensity between months from March to September.



(a) (b)

Figure 5.3. Averaged intensity of lightnings in $10 \times 10\,km$ cell during the warm season March-September: a) Frequency distribution; b) Monthly distribution.

5.3 A spatial-temporal technique to model lightnings-rainfall relation

The Tapia-Smith-Dixon model [Tapia et al. 1998] is widely used to estimate the spatio-temporal propagation of rainfall using instantaneous lightnings records. A key issue of this model is the estimation of Rainfall Lightning Ratio (RLR). Following the approach of Tapia-Smith-Dixon model, we use lightnings records in order to delineate the spatial propagation of rainfall but firstly, we apply a different method for the estimation of the RLR; secondly we build on Tapia-Smith-Dixon proposal by introducing different representations of spatial and spatio-temporal variation in their lightnings-rainfall conversion equation. Eventually, in Chapter 6, we propose a statistical model for rainfall estimation based on lightnings-rainfall relation. Here, we present the basic of Tapia-Smith-Dixon model and we illustrate the estimation phase of Rainfall Lightning Ratio that is crucial when applying our model, as well. Finally, we present the performance of our RLR estimate when incorporated in a simple deterministic model which we use to reconstruct rainfall fields from lightnings records.

5.3.1 Tapia-Smith-Dixon model

The Tapia-Smith-Dixon model [Tapia et al. 1998] is basically a spatio-temporal prediction of rainfall rate based on CG-lightnings patterns and Rainfall Lightnings Ratio estimate. The Tapia-Smith-Dixon model is described as follows:

$$R(t,x) = C \sum_{i=1}^{N_t} Z f(t,T_i) g(x,X_i) \tag{5.1}$$

where

- $R(t,x)$ is the rainfall rate (mm/h) at time t and spatial location x
- C is the units conversion factor
- t denotes forecast time
- i is the ordinal number of flash
- N_t is the number of flashes until time $t + \Delta t/2$
- Z denotes Rainfall-Lightning Ratio (RLR), the convective rainfall mass per flash (total convective rainfall mass divided by the total number of flashes)
- T_i denotes time of ith flash
- X_i denotes location of ith flash
- $f(t,T_i)$ specifies the rainfall flux at time t determined by a lightning flash at time T_i

- $g(x, X_i)$ specifies the rainfall flux at location x having a lightning flash at location X_i.

Either temporal or spatial function are taken to be *uniform* over the interval (-5 min, +5 min) of time t and within a circle of 5 km radius around location x, respectively. These assumptions are declared by the authors as simple and an early stage to build on.

5.3.2 Rainfall Lightning Ratio estimation

The Rainfall Lightning Ratio Z determines the **mass of rain associated to each flash**. This mass is expressed in $kg\ m^{-2}$ whereas the precipitation recorded by rain gauges or estimated from radar as well as satellite data is expressed in $mm\ m^{-3}$. Consequently, we need to transform a mass into a volume applying a conversion factor such that named C in Equation 5.1, which is $C = 10^6 A^{-1}$ where A is the interested area in square kilometers.

In general, the RLR depends very much on the thunderstorm type and region. Quantitative estimations of RLR have been proposed in several studies [Soula and Chauzy 2001] and references therein:

- Battan 1965 determined an RLR of **30** $10^3\ m^3$ per CG-flash that ranges from 3 to 300 in Arizona
- Kinzer 1974 reported **20** $10^3\ m^3$ from *1* thunderstorm in Oklahoma
- Maier et al. 1978 reported **100** $10^3\ m^3$ from *22* thunderstorms in Florida
- Piepgrass et al. 1982 reported **20** $10^3\ m^3$ from *2* cases in Florida
- Buechler et al. 1990 reported **40** $10^3\ m^3$ from *21* cases in Tennessee
- Buechler and Goodman 1991 reported **180** $10^3\ m^3$ from *2* cases in Florida
- Williams et al. 1992 reported **500** $10^3\ m^3$ from *43* cases in Australia
- Soula et al. 1998 reported **30** $10^3\ m^3$ for *1* long lasting and stationary system in Spain
- Tapia et al. 1998 determined a average volume of **43** $10^3\ m^3$ per flash that ranges from 24 to 365, in *22* storm cases in Florida

The simplest estimator of RLR is an average of the set of $\tilde{Z}_e$ calculated for each convective event e:

$$\hat{Z} = \frac{1}{E_n} \sum_{e=1}^{E_n} \tilde{Z}_e$$

where E_n is the total number of convective events identified by the *scan statistic procedure* (see Sec. 4.2). Then, we can estimate RLR accounting for some convective event's features as:

1. total number of CG-lightnings (d-dimension)
2. the size of the covered area (A-area)
3. number of CG-lightnings during a 15-minutes peak (TI-temporal intensity)
4. mean number of CG-lightnings in a pixel $10 \times 10\,km$ (SI-spatial intensity)

Thus the RLR estimator becomes a function of the above mentioned dimensions. In what follows we propose RLR estimates function of the event size considering three categories Small, Medium and Large as identified in Table 4.3 where *Large* and *Very Large* are merged together since cases in *Very Large* category are few. Thus we are going to adopt the following estimators:

$$\hat{Z}^S = \frac{1}{E_n^S} \sum_{e=1}^{E_n^S} \tilde{Z}_e^S \tag{5.2}$$

$$\hat{Z}^M = \frac{1}{E_n^M} \sum_{e=1}^{E_n^M} \tilde{Z}_e^M \tag{5.3}$$

$$\hat{Z}^L = \frac{1}{E_n^L} \sum_{e=1}^{E_n^L} \tilde{Z}_e^L \tag{5.4}$$

where E_n^S, E_n^M and E_n^L are the number of convective events counted in each dimensionality group and, correspondingly, $\tilde{Z}_e^S$, $\tilde{Z}_e^M$ and $\tilde{Z}_e^L$ are the RLR estimates of a single event in each group.

Furthermore, we propose a $\hat{Z}$ estimate that is based on satellite data rather than rain gauges data. In fact, the RLR for a single convective event is calculated dividing the total volume of precipitation registered in the area of interest by the total number of lightnings registered in the same area, and data from sparse dislocated weather stations do not let an exact computation of the numerator. However, we use weather stations data to apply two correction factors since satellite records generally underestimate rainfall quantities and overestimate the number of null precipitation cases. In the case of underestimation, the correction is necessary since the total volume of precipitation is in the numerator of RLR. On the other hand, the correction for overestimation reflects the fact that is frequent to have cases with positive value of lightnings' number and a null precipitation determining a more numerous total number of lightnings in the denominator of RLR when using satellite data instead of rain gauges data. Before giving details on correction factors, let us formalize the estimator. Let us denominate *Poly* a polygon that represents the spatial patterns of lightnings associated to a convective event. *Poly* is determined as a convex hull of the whole finite set of lightnings. A graphical representation of two polygons relative to May 9, 2006 and August 5, 2004 convective events is reported in Figure 4.4 and 4.6, respectively[1]. Moreover, remind that the satellite database is composed of records from cells $10 \times 10\,km$ side of a regular grid. Thus, for instance, the Large event estimate of Z is:

[1]The convex hull is computed using the *sp* R-package [Bivand et al. 2013].

$$\hat{Z}_e^L = \frac{\sum_{h=1}^{T_h} \sum_{p \in Poly} r_h^{SAT}(p)}{\sum_{t=1}^{T_h} \sum_{p \in Poly} L(t,p)} \tag{5.5}$$

where $L(t,p)$ is the number of lightnings recorded at time t and cell p and $r_h^{SAT}(p)$ is the satellite precipitation accumulated in cell p during the 1-hour h time interval. Notice that the time interval is 1-hour since this is the time scale of satellite database, with $h = 1, \cdots, T_h$ being T_h the duration of the event in hours.

Once RLR for the three classes of dimensionality have been calculated, two correction factors are applied. More specifically, the correction analysis is done comparing rainfall data collected from stations and the correspondent satellite data recorded in the cells where stations are located. In particular, let $r_h^{STAT}(p)$ be the stations precipitation accumulated at cell p during the 1-hour time interval h, with spatial domain $D = \{p = 1, \cdots, N\}$, being N the total number of cells where at least one weather station is located. Firstly, the average difference between precipitation volume recorded from stations and that obtained from satellite data in the cells where rain gauges are present is calculated. The resulting quantity is used to augment the satellite rainfall values in every cell where satellite precipitation is not null. More precisely, the correction values is given by:

$$f_1 = \frac{1}{T_h \times N} \sum_{h=1}^{T_h} \sum_{p \in Poly \cap D} r_h^{STAT}(p) - r_h^{SAT}(p) \quad h = 1, \cdots, T_h \quad p = 1, \cdots, N \tag{5.6}$$

Lastly, we count the cases when a null precipitation is encountered and we compute the factor of correction dividing the probability of having a null precipitation in stations data by the same probability in satellite data, such that:

$$f_2 = \frac{Pr\{r_h^{STAT}(p) = 0\}}{Pr\{r_h^{SAT}(p) = 0\}}, \quad h = 1, \cdots, T_h \quad p = 1, \cdots, N \tag{5.7}$$

Correction factor f_2 is used to counter the excessive weight of cases with zero precipitation in the calculation of RLR. Thus, the correction consists of two adjustments:

1. we adjust the satellite precipitation using an average difference with rain gauges levels;
2. we eliminate the excessive cases of null precipitation estimated from satellite data equalizing the frequency to that of rain gauges records.

Finally, RLR estimates obtained from our data-set are reported in Table 5.1 where RLR values are expressed in 10^3 m^3 per CG-flash.

5.3.3 Reconstruction of rainfall field by means of lightnings data

We introduce here a very simple deterministic model for predicting rainfall from lightnings. In particular, we built on the approach of Tapia-Smith-Dixon [Tapia et al. 1998] by introducing estimates of the RLR that depend on the size of the convective event

RLR (10^3 m^3)	**Small events**	**Medium events**	**Large events**	**Entire**
Min.	0	0	0	0
1st Qu.	0	0.2	2.1	0.1
Median	0.1	0.6	8.6	0.2
Mean	0.4	2.6	**24.1**	4.0
3rd Qu.	0.3	2.2	27.4	1.5
Max.	10.4	131.2	131.8	131.8

Table 5.1. Corrected RLR estimates for 3 classes of dimensionality: Small, Medium and Large convective events.

and by removing the temporal component $f(t, T_i)$ from Equation 5.1. In fact, this model reconstructs the precipitation at every cell of the convective event spatial domain for the entire duration of the event. However, it is worth drawing the attention that this approach is fully deterministic and, subsequently, it does not allow for a correct assessment of estimates uncertainty. The model presented further in Chapter 6 overcomes this drawback by introducing a latent variable with a space-time random process. Nevertheless, the deterministic model we present here is a valid test to evaluate the efficiency of RLR estimates done in the previous section.

Recall from Section 5.3.2 that T_h is the duration time in hours of the convective event, the satellite precipitation at cell p in the time interval of the duration of the event is

$$R^{SAT}(p) = \sum_{h=1}^{T_h} r_h^{SAT}(p) \qquad\qquad h = 1, 2, \cdots, T_h.$$

Then, the lightnings-derived precipitation at each cell p for the entire duration T_h becomes:

$$R^{LIG}(p) = (10^6 A_p^{-1})\Big(\sum_{h=1}^{T_h} L_h(p)\Big) * \hat{Z}(d) \qquad\qquad h = 1, 2, \cdots, T_h \qquad (5.8)$$

where

- $r_h^{SAT}(p)$ is the hourly satellite rainfall at cell p;
- $L_h(p)$ is the number of lightnings recorded at cell p during hour h;
- $10^6 A_p^{-1}$ is a factor for converting from $10^6 kg\ m^{-2}$ (mass) to $mm\ m^{-3}$ (volume) where A_p is the area of a cell in square meters;
- $\hat{Z}(d)$ is RLR for class of dimensionality d=Small, Medium, Large events.

Notice that the model presented here is based on the assumption that the total rainfall volume derived from lightnings at cell p is uniformly distributed for the duration of the event.

The map in Figure 5.4 is an example of reconstructed satellite precipitation obtained by applying the Lightnings-Rainfall model described above. The map is referred to the convective event of August 7, 2003 started at 1:35 pm and ended at 18:56 pm. This event is classified as Large event since it counted 1096 total lightnings in 322 minutes and, consequently, we use a RLR equals to 24.1 10^3 m^3 per CG-flash (see Table 5.1).

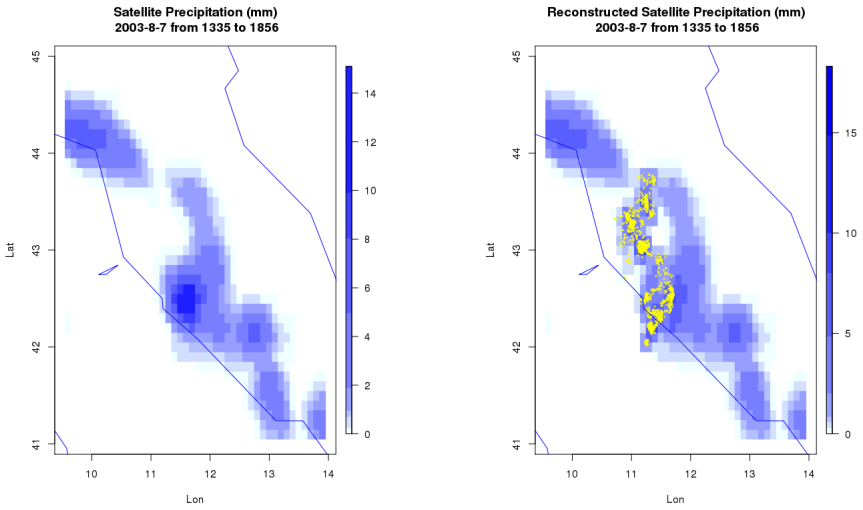


Figure 5.4. Map of the August 7, 2003 convective event: satellite precipitation data (left panel); reconstructed precipitation by adopting a deterministic model that use lightning records (right panel).

The evaluation of the reconstruction model is done using the information from rain gauges. More specifically, for every 767 convective events identified by scan statistics procedure (see Section 4.2) we select those cells where at least one rain gauge is settled in, then we compare the amount of rainfall deriving from satellite data $R^{SAT}(p)$ with the corresponding value calculated from rain gauges records $R^{STAT}(p) = \sum_{h=1}^{T_h} r_h^{STAT}(p)$. When two or more rain gauges are present in the cell, we take the average amount of precipitation. Under the hypothesis that the rainfall volume recorded by a rain gauge is representative of the cell $10 \times 10\,km$ which contains it, we obtain the RMSE values reported in Table 5.2, depending on the dimensionality of the event. In this table, occurrences are the number of cells with a valid record coming from rain gauges.

The analysis of RMSE values reveals that the adoption of our $\hat{Z}(d)$ estimate let us to obtain an **improvement** of rainfall satellite estimate for Large convective events. Furthermore, the probability of hitting rainy cases, i.e. with rainfall greater than $0.2\,mm$ and Probability of False Alarm (POFA)[2] reported in Table 5.3 confirm that our reconstruction method improves the performance of satellite data.

[2]Details on the meanings of POD and FAR are given further on Section 6.7.4. Recall that Probability of Detection (POD) is the number of cases correctly predicted in class *Rain* on the total cases observed in the same class whereas Probability of False Alarm (POFA) is the percentage of cases predicted as rainy but observed as no rainy with respect to the whole rainy predicted cases.

RMSE(mm)	RLR=0.439	RLR=2.585	RLR=24.108
precsat	13.0	22.8	**20.7**
reconstructed	15.3	24.9	**15.9**
occurences	535	1248	2780

Table 5.2. Root Mean Square Error of satellite precipitation (precsat) and reconstructed values compared to rain gauges values.

RLR Large events		
	POD	POFA
reconstructed	81.3	15.8
precsat	76.1	15

Table 5.3. Comparison of hits and false alarms on total of Large events reconstruction.

Chapter 6

Predicting rainfall fields from lightnings records

In this Chapter we present a model to predict the 15-minutes and 30-minutes accumulated precipitation at unknown locations given lightnings counting. In particular, we assume that the accumulated precipitation at time t in cell p of a 10×10 km regular grid is generated by a fixed component related to lightnings and a random $\mathbf{W}$ term structured in space and time. Recall that we refer to convective events as defined in Chapter 4 and that the study area is located in Central Italy. We use lightnings records in the fixed component of the model. The Chapter is organized as follows: in Section 6.1 an analysis of the two case study is presented; the complete model structure is illustrated from Section 6.2 to 6.5; the predictive process is described in Section 6.6. Finally the Section 6.7 includes an illustration of parameters estimation (Sec. 6.7.3) and an evaluation of the model performance (Sec. 6.7.4).

6.1 The case study

The model we propose is implemented on two convective events occurred during the storms of May 9, 2006 and August 5, 2004, particularly the convective event started at 00.01 pm and ended at 16.14 on the 9th of May 2006 and the convective event started at 11.24 am and ended at 19.56 pm on the 5th of August 2004. Events are identified as described in Chapter 4.

Main features of the two events are reported in Table 6.1. They differ from each other mainly because of the total number of lightnings generated. The mean volume of precipitation calculated as the fraction of total rainfall estimated from satellite data in the area on number of cells, is also different, though the difference is not as remarkable as for lightnings counting. According to the definition given in Section 4.3, the convective event of May 9, 2006 can be considered as a *Large event* whereas that of August 5, 2004 belongs to the category of *Very Large event* (see Table 4.3). In fact, the first registered 3163 CG-flashes in 16 hours of duration and the latter 18140 CG-flashes in 9 hours of duration. Thus, the duration of the event differs substantially from one to another as well as the covered area, which is of 54131 km^2 and of 93147 km^2 for May 9, 2006 and August 5, 2004 convective event, respectively. However the location of the two events is very similar: the May 9, 2006 centroid is located at 11.4 degrees Longitude East and 43.8 degrees Latitude North, whereas the August 5, 2004 centroid is located at 11.4 degrees Longitude East and 43.1 degrees Latitude North. Finally, the lightnings' peak of May 9, 2006 at 15 minutes aggregation has been registered during the first quarter of 20 pm with 136 flashes whilst that of August 5, 2004 has been registered during the third quarter of 4 pm with 1236 flashes.

Event	**#Ligh.**	**Duration (hours)**	**Area(km^2)**	**Max # ligh. in 15min**	**Hourly rainfall intensity (mm)**
May 9, 2006	3163	16	54131	136	1.3
Aug 5, 2004	18140	9	93147	1236	1.86

Table 6.1. Study event main features (hourly rainfall intensity is the fraction of total rainfall recorded by rain gauges in the area on number of rain gauges).

Furthermore, we adopt two scales of time aggregation: the 15-minutes aggregation which well suits the temporal evolution of lightnings within a storm but including some noise in the model processing and the 30-minutes aggregation which, on the other hand, is capable of eliminating the noise but might miss some relevant features observable at larger time scales. Since we adopt two time aggregations, the description of the database and the corresponding data summaries vary with the time scale.

A description of space-time support of the study events is reported in Table 6.2. The weather stations located inside the polygons which represent the area covered by the two events under analysis are 316 for May 9, 2006 and 201 for August 5, 2004. Most of those weather stations have been excluded by the analysis because of the great quantity of missing data in the corresponding time series. Then, the two dataset reduce to 179 (181) and 159 (171) locations where the time series of observations are complete for the entire duration of the events[1]. Moreover, the data from weather stations are aggregated in order to have a single rainfall value per each cell of a regular grid. In fact, the support of the model is a regular grid of 10×10 km cell sides. Thus, when two or more rain gauges belong to the same grid cell we take their median over the cell. Consequently, the following summaries always refer to cells rainfall amount calculated on the basis of rain gauges records. Finally, the space-time support of May 9, 2006 15-min (30-min) event is composed of 111(112) cells and 68(34) time units whereas the support of August 5, 2004 15-min (30-min) event is 100(104) cells and 36(18) time units. Thus, the database for the former event is composed of lightnings records (instant-point fields) accumulated in 7548(3808) space-time units and 179(181) time series of rain observations and for the latter in 3600(1872) lightnings records and 159(171) rainfall time series.

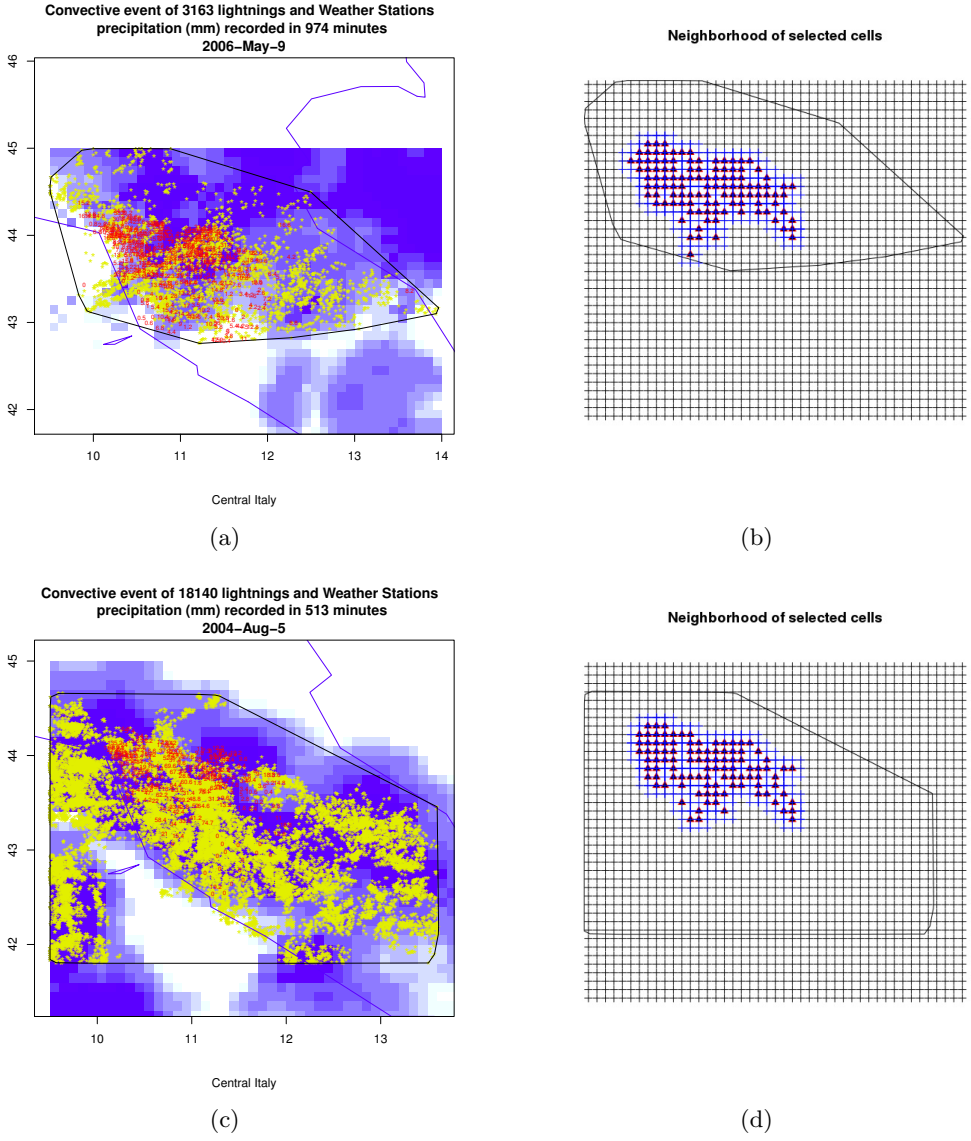
Event	#Rain gauges	#Cells	T	Units
May 9, 2006 15-min	179	111	68	7548
30-min	181	112	34	3808
August 5, 2004 15-min	159	100	36	3600
30-min	171	104	18	1872

Table 6.2. Description of study events at 15- and 30-minutes time aggregation: number of rain gauges, cells of 10×10 km side, time intervals (T) and units ($T \times \#Cells$).

A map of the two events is reported in Figure 6.1 panel a) and c) where it is shown a polygon (black line) that delimits the area of interest of the convective event, the lightnings spatial propagation (yellow), the satellite precipitation (blue palette) and the rain gauges observations (red values). In panel b) and d) of the same figure the cells where at least one complete time series is available and the neighborhood is mapped.

Rainfall data are affected by several problems, on one hand a very large number of zero values is recorded, on the other hand the rain gauges precision (about 0.2 mm) implies an almost discrete measurement of accumulated rain as shown in Table 6.3.

[1]The values in brackets are the available time series for the 30-min time aggregation. The number of weather stations is often greater than that of the 15-min aggregation due to the presence of several rain gauges which register data every 30 minutes.



(a) (b)

(c) (d)

Figure 6.1. The area covered by convective events of May 9, 2006 (a-b) and August 5, 2004 (c-d): a-c) lightnings (yellow), satellite precipitation (blue palette) and rain gauges observations (red values); b-d) cells with at least one rain gauge.

Besides, a comparison of basic summaries of positive precipitation shown in Table 6.4 confirms the different nature of the two events: the maximum value registered at 15-min time steps is 15 mm for the May 9, 2006 event and 34.6 mm for the August 5, 2004 event; the correspondent levels at 30-min time steps are 17.6 mm and 58.8 mm. Moreover, the difference between the two events are more evident in two extremes as it can be seen in the values of 3rd quartile. This is due to the presence of higher levels of precipitation in the 4th quartile of August 5, 2004 rainfall distribution, i.e higher rain rates for extreme cases.

Rain classes (mm)	$[\mathbf{0}, \mathbf{0.2})$	$[\mathbf{0.2}, \mathbf{0.4})$	$[\mathbf{0.4}, \mathbf{0.6})$	$[\mathbf{0.6}, \mathbf{0.8})$	$[\mathbf{0.8}, \mathbf{1})$	$[\mathbf{1}, \mathbf{5})$	$\geq \mathbf{5}$
Cases (9May2006-15min)	5949	451	202	134	125	625	62
%	78.8	6	2.7	1.8	1.7	8.3	0.8
Cases (9May2006-30min)	2746	254	120	69	54	446	119
%	72.1	6.7	3.2	1.8	1.4	11.7	3.1
Cases (5Aug2004-15min)	2384	380	202	136	102	323	73
%	66.2	10.6	5.6	3.8	2.8	9	2
Cases (5Aug2004-30min)	1079	155	98	89	54	312	85
%	57.6	8.3	5.2	4.8	2.9	16.7	4.5

Table 6.3. Frequency distribution of rainfall values observed on 9th of May 2006 and 5th of August 2004 convective events at 15-min and 30-min time aggregation.

Event	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
May 9, 2006 15-min	0.2	0.2	0.6	1.299	1.6	15
30-min	0.2	0.2	1	1.971	2.6	17.6
August 5, 2004 15-min	0.2	0.2	0.6	1.386	1.2	34.6
30-min	0.2	0.4	0.8	2.194	2	58.8

Table 6.4. Summaries of positive precipitation cases observed during the convective event of May 9, 2006 at 15-min and 30-min time aggregation.

6.2 Modeling approach

In Section 5.3, a very simple deterministic model for predicting rainfall from lightnings is illustrated. We have already built on the approach of Tapia-Smith-Dixon [Tapia et al. 1998] by introducing estimates of the RLR that depend on the size of the convective event. In particular, we were able to better estimate precipitation for largest events (see Table 5.2). However the Tapia-Smith-Dixon approach being fully deterministic does not allow for a correct assessment of estimates uncertainty. Furthermore, while the Tapia-Smith-Dixon model is used for now-casting of point estimates, we are interested in areal estimation over a $10 \times 10\,km$ grid, predictions to be in a future used to correct satellite values. We adopt a stochastic approach as we are interested not only in predicted values but also in the uncertainty affecting such predictions. Among other possible stochastic approaches, such as Germ-Grains model ([Neyman and Scott 1958]; [Rodriguez-Iturbe et al. 1987]), we propose a Linear Mixed model, i.e. a linear model with *fixed* and *random* effects. In particular, the fixed component is based on an new version of the Tapia-Smith-Dixon model, where a more complex approach is taken to describe space-time propagation of CG-lightnings inside a convective event and we include the modified RLR estimation described in Section 5.3.2 when computing the precipitation volume in each cell. The random component is associated to the spatial distribution of precipitation among closer cells and to the temporal evolution of the precipitation phenomenon in each cell. The adding of this space-time component is crucial for a corrected prediction, since a prediction based exclusively on lightnings information might mask those cases where there is no rainfall although the presence of lightnings. In fact, it is frequent to have either lightnings events with no rainfall [Parker and Johnson 2000] or viceversa.

6.2.1 The model definition

Mixed Models (linear and non linear) belongs to a class of models in which some of the effects are *fixed* and some are *random*, formalization of these models is easily achieved in a hierarchical Bayesian framework. Here we propose a space-time mixed model to link rain measures and lightning counts in a given area of Central Italy. We envision a model in which the rain at cell p of a regular grid, at time t is described by a latent random process. The rainfall intensity changes rapidly across time and space, mostly depending on intensity and direction of the wind. Since the wind is also a forcing factor of the spatial propagation of lightnings, we make inference about the process directly relating the latent variable to them. In particular, by the use of LMM, the latent variable is generated both by a fixed component connected to spatial patterns of lightnings events and a spatial random component. Our final goal is to built a predictor for the underlying spatial surface of the latent variable.

Let $X(t,p)$ be the latent rainfall field at cell p and time t. $L_{t,p}$ denotes the corresponding number of lightnings. Given the partially discrete nature of the dataset and the zero-inflated distribution (see Table 6.3), we discretize the latent process $X(t,p)$ below 1 mm assuming that there exists five values $\lambda_i, i = 0, \ldots, 4$ described in Table 6.5, that occurs with positive probability whenever $X(t,p)$ belongs to one of the interval reported in the same table. Here, the quantity 0.1 is the

typical measurement error level of rain gauge instrument. This simple method proposed by [Sahu et al. 2005] and [Jona Lasinio et al. 2007] let us to obtain a good performance of the model in predicting zero rainfall events. Other methods can be adopted to treat zero inflated rainfall distributions, such as in [Berrocal et al. 2008], [Fuentes et al. 2008] or [Schmidt and Migon 2009]. Berrocal et al. (2008) specify a spatial model that includes two spatial Gaussian processes driving precipitation occurrence and accumulation, respectively. A spatial-temporal model for rain gauges and reflectivity radar data is developed in Fuentes et al. (2008), where a latent process corresponding to the true rain amount drives the probability of precipitation occurrence and the rainfall accumulation. On the other hand, Smith and Migon (2009) treat observations from rain gauges as generated by a latent process which realizations are a mixture between a Bernoulli distribution that specifies the probability of having positive precipitation and a probability density function for the rainfall accumulation, typically an exponential, a gamma or a log-normal distribution. We choose the above described method at this stage as it is easy to implement and we are on a finer time scale than the cited works that mostly work with hourly rainfall data using radar measurements to improve the rainfall field estimation or with weekly data as in Smith and Migon (2009).

Rain Classes (mm)	Discretization values
$[\mathbf{0}, \mathbf{0.2})$	$\lambda_0 = log(0.1 + 1)$
$[\mathbf{0.2}, \mathbf{0.4})$	$\lambda_1 = log(0.3 + 1)$
$[\mathbf{0.4}, \mathbf{0.6})$	$\lambda_2 = log(0.5 + 1)$
$[\mathbf{0.6}, \mathbf{0.8})$	$\lambda_3 = log(0.7 + 1)$
$[\mathbf{0.8}, \mathbf{1})$	$\lambda_4 = log(0.9 + 1)$

Table 6.5. Discretization values for the latent rainfall field X.

The rainfall latent variable is log-transformed as $Y(t,p) = \log(X(t,p)+1)$ where we add 1 to account for zero values. This transformation is chosen mostly to smooth the impact of strong rainfall intensities [Lee and Zawadzki 2005] and to allow a more sensible adoption of Gaussian representation. Then, the transformed latent variable becomes:

$$Y(t,p) = \log(X(t,p)+1)$$

with elements $\mathbf{Y} = \left(y(t_1,p_1), \cdots, y(t_T,p_N)\right)^T$, at 15- or 30-minutes intervals $t = 1, \cdots, T$ and cell p of a regular grid $n = n_1 \times n_2$. Then, the latent rainfall field on the log scale $Y(t,p)$ is modeled as the sum of a fixed effect and a space-time random effect $\boldsymbol{W}$:

$$y(t,p) = \mu(t,p) + w(t,p) + \epsilon(t,p) \tag{6.1}$$

where $\mu(t,p)$ is as in Eq. 6.9, $w(t,p)$ is the (t,p) element of $\mathbf{W}$ a separable space-time random field such that $w(t,p) = T(t) + S(p)$ with $T(t) = \alpha T(t-1) + \eta(t)$, $\eta(t) \sim N(0, \sigma_\eta^2)$ and

$$\mathbf{S} \sim MN(\mathbf{0}, \sigma_s^2(\mathbf{I} - \rho_s\mathbf{B})^{-1}) \tag{6.2}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{B}$ an adjacency matrix describing a spatial neighborhood structure. Finally, $\epsilon(t,p) \sim N(0,\tau^2)$ are independent, identically distributed random variables. In practice, the random component of the process is divided into two parts: the first is a spatial-temporal process, the second is a measurement error term.

6.3 The fixed effect

The fixed component of the model relates precipitations and lightnings starting from the well known Tapia-Smith-Dixon relation [Tapia et al. 1998] which is reported in Eq. 5.1. By applying the modification of RLR estimation proposed in Section 5.3.2, the estimator of rainfall field at cell p and time t becomes:

$$\hat{r}^{LIG}(t,p) = (10^6 A_p^{-1}) * \hat{Z}(d) * \sum_{i=1}^{T} \sum_{s \in N_p} L_{i,p} * f(t, T_i; V) * g^d(p, P_s; V) \qquad i = 1, 2, \cdots, T \tag{6.3}$$

where

- $\hat{r}^{LIG}(t,p)$ is the rainfall prediction field at cell p and time t;
- P_s is the observed cell-location;
- T_i is the observed time;
- $L_{i,p}$ is the number of lightnings cumulated at the end of a given time (in our case study 15 and 30 minutes) interval i at cell p;
- N_p is the neighbourhood of p;
- $\hat{Z}(d)$ is the estimated Rainfall Lightning Ratio according to the dimensional factor d=Small, Medium, Large as defined in Table 5.3.2;
- $10^6 A_p^{-1}$ is a conversion factor from $10^6 kg\ m^{-2}$ (mass) to $mm\ m^{-3}$ (volume) with A_p being the area of any cell in square meters (see Section 5.3.2 for details);
- V is the velocity of propagation of the convective event;
- $f(t, T_i; V)$ is a time weights function;
- $g^d(p, P_s; V)$ is a spatial weights function.

The life of lightnings pattern inside a rainfall convective event is composed of 3 stages: *Charging phase* (Ch), *Mature state* (Ma) and *Dissipating phase* (Dis). Then, the event duration interval can be partitioned into $[t_0, T_{Ch})$, $[T_{Ch}, T_{Ma})$ and $[T_{Ma}, T]$. We build a time weight function that take into account this feature. In space we assume that the number of lightnings in cell p depends on the number of lightnings occurring in neighboring cells. Then we define a neighborhood structure to handle such dependency: for instance we can adopt a chess queen neighboring structure (see Fig. 6.3) [Cressie 1993].

Time weight function

In Fig. 6.2 the temporal evolution of both lightnings and associated precipitation during a convective event is described. The temporal function is built from Fig. 6.2 such that:

$$f(t,T_i;V) = \begin{cases} f_{Ch}(t) & \text{if} \quad t \,\&\, T_i \in Ch \\ f_{Ma,Dis}(t) & \text{if} \quad t \,\&\, T_i \in Ma \bigcup Dis \end{cases} \tag{6.4}$$

In practice, we assume that the *Mature* and *Dissipating* phases are equivalent in term of the lightnings temporal evolution ending with just two stages $[t_0, T_{Ch})$ and $[T_{Ch}, T)$, such that:

$$f_{Ch}(t) = \exp\left\{ -\frac{(a+bV)}{A_p^{1/2}} |t-T_i| \right\} \qquad t_0 < t \,\&\, T_i < T_{Ch} \tag{6.5}$$

$$f_{Ma,Dis}(t) = \exp\left\{ -\frac{(a+bV)}{A_p^{1/2}} |t-T_i|^2 \right\} \qquad T_{Ch} \le t \,\&\, T_i < T \tag{6.6}$$

where T_{Ch} indicates the end of the *Charging phase* and T is the duration of the entire event, V is the velocity of propagation.

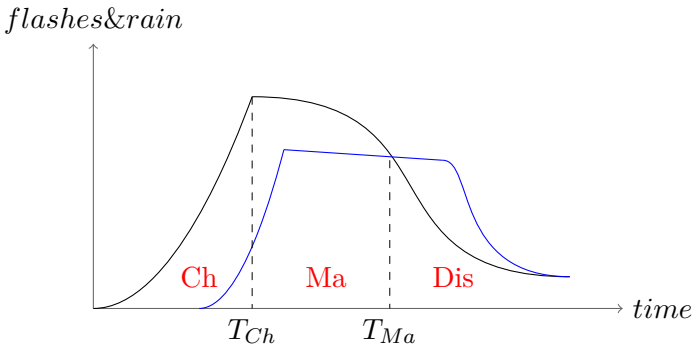


Figure 6.2. Temporal evolution of lightnings (black) and associated rain (blue) within a convective event. The 3 stages of evolution are also indicated: *Charging phase* (Ch), *Mature state* (Ma) and *Dissipating phase* (Dis).

At the stage T_{Ch} is obtained from the data as an exogenous quantity. In Table 6.7 are reported values related to our case studies. Notice that if the predicting time t and observed time T_i are in different phases of event's lifetime, that is both the events $\left\{ \{t \in Ch\} \cap \{T_i \in Ma \cup Dis\} \right\}$ and $\left\{ \{t \in Ma \cup Dis\} \cap \{T_i \in Ch\} \right\}$, then they receive a zero weight, i.e. $f_{Ch,Ma,Dis}(t) = 0$. This assumption is basically done to avoid a strong correlation between model predictions that are quite distant in

time from each other. However, the assumption becomes too strong when treating with instants of observation and prediction that are both near to the T_{Ch} time.

Space weight function

We define $N(p)$ as the second order neighbourhood of cell p (see Fig. 6.3), then

$$g(p, P_s) = \begin{cases} \omega_{i,s} & \text{if } P_s \in N(p) \\ 0 & \text{if } P_s \notin N(p) \end{cases} \tag{6.7}$$

where $\omega_{i,s} = 1/8\, L_{i,P_s}$, L_{i,P_s} is the number of lightnings recorded in observed cell P_s at time i and the proportional factor 8 represents the maximum number of neighbours of cell p for the queen's neighborhood structure.

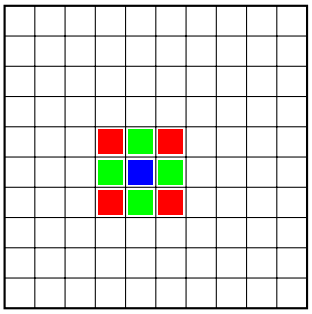


Figure 6.3. Neighbourhood of a generic cell p (blue): first order structure is composed of green pixels whilst second order is red in addition to green pixels (queen structure).

In this formulation, the spatial dependence between neighboring cells is completely determined according to the following equation:

$$\omega_{i,p} = \frac{L_{i,p}}{L_p} + \frac{1}{8} * \frac{L_{i,N_p}}{L_p} \tag{6.8}$$

where $L_{i,p}$ is the number of lightnings at predicting cell p and time i, L_{i,N_p} is the summation of lightnings over the neighborhood cells of p at each time i (i.e. $\sum_{P_s \in N_p} L_{i,P_s}$) and

$$L_p = \sum_{i=1}^{T} \left(L_{i,p} + L_{i,N_p} \right)$$

is the total number of lightnings hitting cell p in addition to the neighbourhood of p for the entire duration of the event.

Possible modification of this spatial weights function could include to take into account the direction and shape of the event. Moreover, the spatial function could vary with the event size (Large, Medium, Small).

Finally, the fixed effect $\mu(t,p)$ is composed of the two parts expressed in Eq. 6.5 and 6.6 and the spatial weight specified in Eq. 6.8. To simplify the notation, let us write $(10^6 A_p^{-1}) * \hat{Z}(d)$ as a constant C since $\hat{Z}(d)$ is estimated apart from the model (see Section 5.3.2), then we have:

$$\begin{aligned}\mu(t,p) &= \log\left(C * \sum_{i=1}^{T} L_{i,p} * \left(\exp\left\{-\frac{(a+bV)}{A_p^{1/2}}|t-T_i|^2\right\} I_{[T_{Ch},T]}(t)\right.\right. \\ &+ \left.\left.\exp\left\{-\frac{(a+bV)}{A_p^{1/2}}|t-T_i|\right\} I_{[0,T_{Ch}]}(t)\right) + C * \sum_{i=1}^{T} \omega_{i,p} + 1\right) \quad (6.9)\end{aligned}$$

where V is again the velocity of propagation of lightnings within a convective event's pattern and A_p is the area of a single cell. Here, $I_{[t',t'']}(\cdot)$ is the indicator function of the time interval $[t', t'']$ which is 1 if either the predicted time t or the observed time T_i are in the same phase of event's lifetime and 0 otherwise.

6.4 The space-time random effect

The $w(t,p)$ component explains the residual spatial variation of rainfall after accounting for that due to $L_{t,p}$ explanatory variable included in the fixed component. In practice, the $\mathbf{W}$ component determines a random increase/decrease of the intercept and it affects the overall variance structure. Recall from Section 6.2.1 that $w(t,p)$ is the (t,p) element of $\mathbf{W}$ a separable space-time random field such that $w(t,p) = T(t) + S(p)$. Thus, the temporal part is specified as $T(t) = \alpha T(t-1) + \eta(t)$ a simple autoregressive model of order 1 with $\eta(t) \sim N(0, \sigma_\eta^2)$ whereas the spatial part is a Conditional Autoregressive model. In the next section, we enter into details of the latter.

6.4.1 Conditional Autoregressive modeling of spatial random effect

The spatial random component S is modeled using a Conditional Autoregressive model (CAR) [Besag 1974]. The CAR model is characterized by a clear link between the conditional and the joint probability distributions. Let $D = \{1, \cdots, p, \cdots, n\}$ be the spatial domain as defined in Section 6.7.1 where n are the 10×10 km square cells that have been hit by the storm event and, simultaneously, have one or more settled in rain gauges[2]. These cells are regularly spaced, then the spatial domain is regular (Fig. 6.4 and 6.5 panel (a)-(c)). Again, let N_p indicate the neighborhood of cell p such that

$$N_p \equiv \{p^* \in D : p^* \neq p \text{ is a neighbor of } p\} \quad p, p^* \in D$$

and $\mathbf{S}^t = (S_1, \cdots, S_p, \cdots, S_n)$ the space random field at each time t. For the sake of simplicity, let us omit the t notation. Under very general conditions, the CAR model is a *Markov Random Field* (MRF) then the knowledge of the set of conditional distributions identifies the joint distribution, furthermore conditional distributions depends only on the neighborhood structure such that:

$$p(s_p | \mathbf{s}_{p'},\, p' \neq p) = p(s_p | \mathbf{s}_{p^*},\, p^* \in N_p) \tag{6.10}$$

and, consequently, we can use the local information to make inference on the random field $\mathbf{S}$ (details can be found for example in [Besag 1974], [Cressie 1993] or [Banerjee et al. 2004]). Furthermore, the *Hammersley-Clifford Theorem* states that the joint distribution deriving from a MRF is a Gibbs distribution, i.e. the joint distribution can be expressed as potentials on cliques (see par. 6.4.1 of [Cressie 1993] for details). On the other hand Geman and Geman [Geman and Geman 1984] demonstrate that a MRF can be sampled from its associated Gibbs distribution (Gibbs sampler). Because of its link between the conditional and the joint probability distributions, the CAR model is particularly useful under a hierarchical Bayesian framework.

[2] The value of n depends on the event under study (see Table 6.7).

The Gaussian case (Autonormal)

Let $\mathbf{B} = (b_{pp'})$ be the matrix for measuring the spatial dependence between grid cells such that, for $p = 1, \cdots, n$, $b_{pp'} > 0$ if p and p' are dependent and 0 otherwise. If it is also assumed that a *pairwise-only dependence* exists (i.e. $b_{pp'} = b_{p'p}$), then the conditional spatial regression has mean and variance:

$$\begin{aligned} E(S_p|\mathbf{s}_{-p}) &= \mu_{S_p} + \sum_{p'=1}^{n} b_{pp'}(s_{p'} - \mu_{S_{p'}}) \qquad p = 1, \cdots, n \\ Var(S_p|\mathbf{s}_{-p}) &= 1/\tau^2_{S_p} \end{aligned} \tag{6.11}$$

where s_{-p} indicates every cell other than p [Cressie 1993].

Now, let us suppose that the random field $\mathbf{S} = (S_1, \cdots, S_p, \cdots, S_n)$ has a multivariate normal distribution. Then, the CAR model is known as *Autonormal* or *Gaussian Markov Random Field* (GMRF) model. Using the Brook's Lemma[3], it has likelihood:

$$p(s_1, \cdots, s_p, \cdots, s_n) \propto exp\left(-\frac{1}{2}\, \mathbf{s^T}\, \mathbf{V}^{-1}(\mathbf{I} - \mathbf{B})\, \mathbf{s} \right) \tag{6.12}$$

where $\mathbf{V}$ is a diagonal matrix with $\mathbf{V}_{pp} = \tau^2_{S_p}$ and $\mathbf{I}$ is the $n \times n$ identity matrix. This form suggests a multivariate normal distribution with $\mu_{S_1} = 0, \cdots, \mu_{S_p} = 0, \cdots \mu_{S_n} = 0$ and covariance matrix $\mathbf{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{V}$. We consider its parametrization by a precision matrix $\mathbf{Q} = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{B})$ with $\mathbf{Q}$ that has zero-pattern structure $Q_{tp,t'p'} = 0$ unless cells (t, p) and (t', p') are neighbors. Consequently, the MRF $\mathbf{S}$ has mean $\mathbf{0}$ and variance $\mathbf{Q^{-1}}$, such that the generic element at time t depends only on the neighborhood p^* and has distribution:

$$S_p|s_{p^*} \sim N\left(\sum_{p^*} b_{pp^*}\, s_{p^*}\, ,\, \tau^2_{S_p} \right) \qquad p = 1, \cdots, n. \tag{6.13}$$

Without loss of generality, it is also worth assuming a common variance $1/\tau^2_S = 1/\tau^2_{S_p}$ for every $p = 1, \cdots, n$ elements of the random field. Thus, the spatial random process has distribution:

$$\mathbf{S} \sim MN\left(\mathbf{0},\ \frac{1}{\tau^2_S}(\mathbf{I} - \mathbf{B})^{-1} \right) \tag{6.14}$$

provided that $(\mathbf{I} - \mathbf{B})$ is invertible and $(\mathbf{I} - \mathbf{B})^{-1}$ is positive-definite. The choice of $\mathbf{B}$ as an adjacency matrix of elements $b_{pp'} = 1$ if p and p' are neighbors and 0 if they are not respect the *pairwise-only dependence* condition and lead us to the previously anticipated Eq. 6.2

$$\mathbf{S} \sim MN(\mathbf{0}, \sigma^2_s(\mathbf{I} - \rho_s \mathbf{B})^{-1})$$

where the symmetry is respected since $b_{pp'}/\tau^2_S = b_{p'p}/\tau^2_S$ for all p, p' and the positive-definite condition is guaranteed by adding a correlation parameter ρ_S. The space

[3]The Brook's Lemma proves that from a set of full conditional distributions one can retrieve the unique joint distribution, provided that the full conditionals one uses are *compatible* [Banerjee et al. 2004].

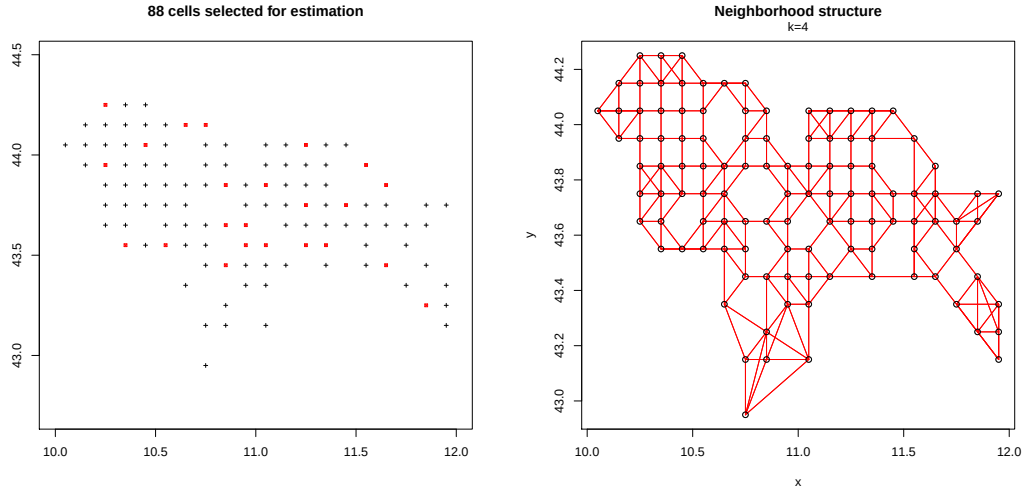
parameter of ρ_S is fixed in $(1/m, 1)$ with m the maximum number of neighbors rather than $(0, 1)$ to make $(\mathbf{I} - \rho_s \mathbf{B})$ non-singular. A further advantage of introducing the correlation parameter in this Gaussian framework is that $\rho_S = 0$ attests conditional independence.

Spatial dependence structure of the model

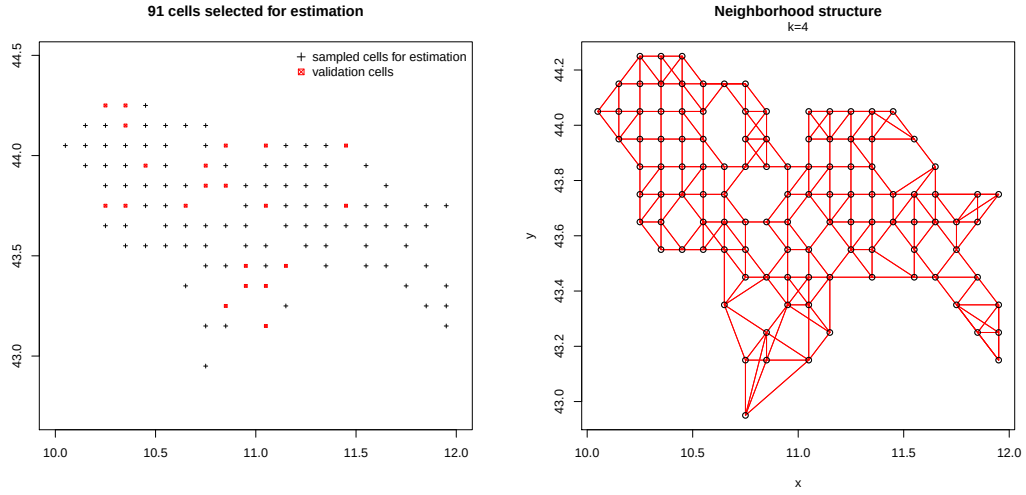
Our choice of neighborhood structure is a second order nearest-neighbor structure where at least $k = 4$ neighbors are selected. This means that we consider as neighbors of cell p, the first 4 cells that are immediately horizontally, vertically or diagonally adjacent (see Fig. 6.3). The resulting structure of dependence between cells is mapped in panel (b)-(d) of Fig. 6.4 and 6.5 for May 9, 2006 and August 5, 2004 convective events, respectively. In the applications we use part of the available observations for estimation and part for validation as reported in Table 6.6 together with the maximum number of neighbors of each neighborhood system. The latter being useful for priors specification.

Model case	**n**	**Est(#cells)**	**Val(#cells)**	**m**
9May2006-15min	111	88	23	7
9May2006-30min	112	91	21	7
5Aug2004-15min	100	86	14	8
5Aug2004-30min	104	91	13	8

Table 6.6. Number of estimation cells, number of validation cells and maximum number of neighbors in the dependence structure of the 4 model cases.

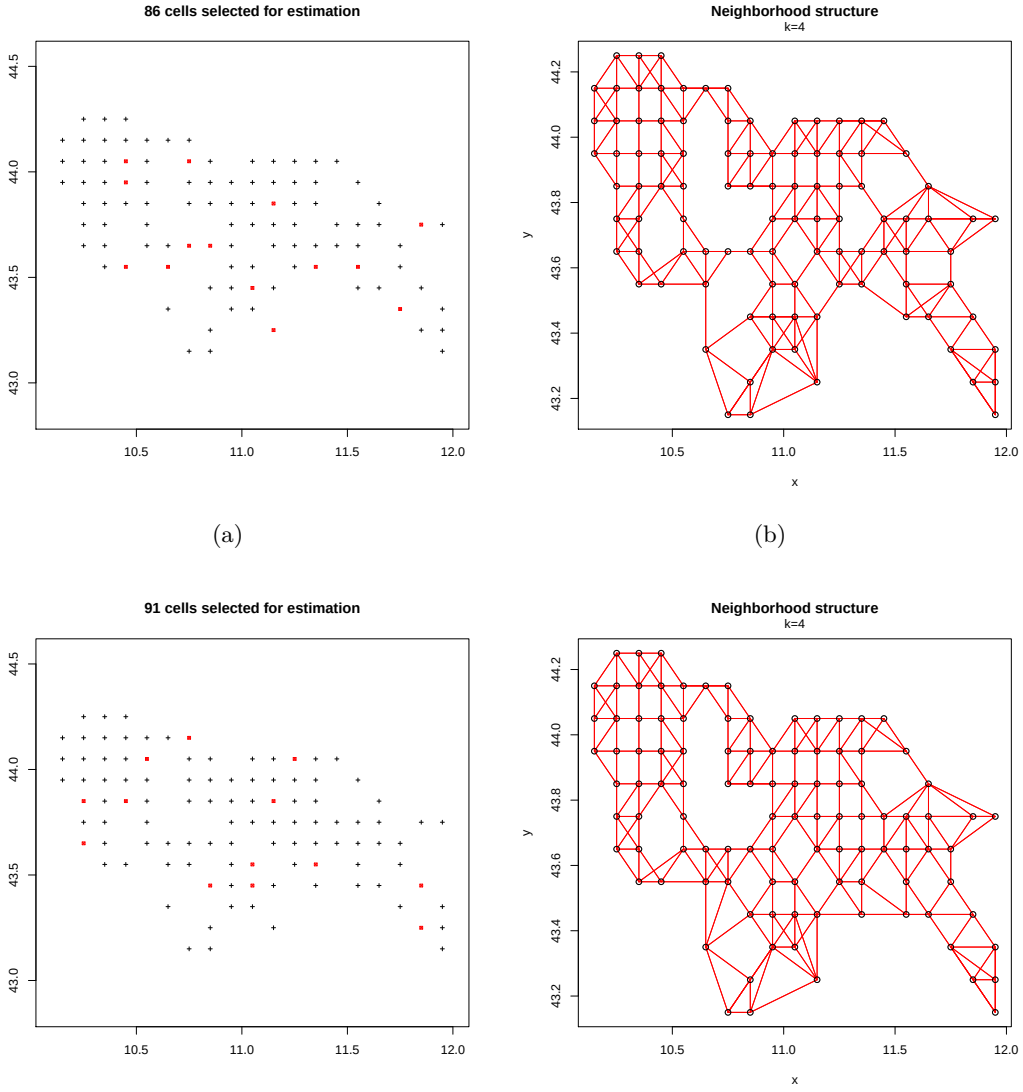


(a) (b)



(c) (d)

Figure 6.4. Spatial domain and neighborhood: (a)-(b) 9-May-2006 15min; (c)-(d) 9-May-2006 30min. In panel (a)-(c) the selected cells for estimation (black crosses) and validation cells (red crosses) are mapped. In panel (b)-(d) the dependence's structure between the whole cells is mapped.



(a) (b)

(c) (d)

Figure 6.5. Spatial domain and neighborhood: (a)-(b) 5-Aug-2004 15min; (c)-(d) 5-Aug-2004 30min. In panel (a)-(c) the selected cells for estimation (black crosses) and validation cells (red crosses) are mapped. In panel (b)-(d) the dependence's structure between the whole cells is mapped.

6.5 The hierarchical Bayesian approach

In a typical Bayesian case the interest is in parameters $\theta = \{\theta_1, \cdots, \theta_m\}$ characterizing the likelihood. Before the sample of data is obtained we have a prior knowledge of the value of the parameters that we model by appropriate probability distributions. Then, once the data have been obtained we *update* our prior beliefs by means of the information coming from the data given the parameters. Firstly, we need to set a *parameters space* Θ as well as the support for the data $\mathbf{Y} = \{y_1, \cdots, y_n\}$. Secondly, we set a prior density $\pi(\theta)$. However, the relationship between parameters θ_i may depend on some parameter ψ which is called *hyper-parameter*. This hyper-parameter controls the *structure* of the common distribution.

The hierarchical approach make possible to define the joint structure of a spatio-temporal process as the product of some simpler conditional distributions. Thus, the joint probability distribution and the full likelihood can be determined using simpler conditional sub-models at each hierarchical stage. Following Berliner 1996 [Berliner 1996] and [Banerjee et al. 2004], the hierarchical modeling structure can be described according to three stages:

Level 1 data|*parameters*, *process*

Level 2 process|*parameters*

Level 3 parameters (hyper)

6.5.1 The hierarchical structure of the model

We start from Eq. 6.1 $y(t,p) = \mu(t,p) + w(t,p) + \epsilon(t,p)$, given the space-time process W we have that:

Level 1 $Y(t,p)|\theta, \mathbf{W} \sim N\Big(\mu(t,p) + w(t,p), \tau^2\Big)$

Level 2 $\mu(t,p)$ see Eq. 6.9
$w(t,p) = T(t) + S(p)$

$T(t)|\alpha, \tau_\eta^2 \sim N(\alpha T(t-1), \frac{\sigma_\eta^2}{1-\alpha^2})$
$\mathbf{S}|\tau_S^2, \rho_S \sim MN(\mathbf{0}, \sigma_s^2(\mathbf{I} - \rho_s\mathbf{B})^{-1})$

Level 3 $a \sim \Gamma(a_0, b_0)$, $b \sim \Gamma(a_1, b_1)$;
$\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\tau_\eta^2 \sim Inv\Gamma(a_\eta, b_\eta)$;
$\tau_S^2 \sim Inv\Gamma(a_S, b_S)$, $\rho_S \sim N(0, \sigma_\rho^2)I_{(0,1/m)}$;
$\tau^2 \sim Inv\Gamma(a_\tau, b_\tau)$

Finally, the complete set of parameters of our model is $\theta = \{a, b, \alpha, \tau_\eta^2, \tau_S^2, \rho_S, \tau^2\}$ where $\sigma_\eta^2 = 1/\tau_\eta^2$ and $\sigma_S^2 = 1/\tau_S^2$.

6.5.2 Priors

The parameters of the fixed component $\mu(t,p)$ are a, b for which we adopt a prior structure based on the *Gamma* distribution: $a, b \sim \Gamma(0.001, 0.001)$. The parameters of the random component $w(t,p)$ are $\{\alpha,\ \tau_\eta^2,\ \tau_S^2,\ \rho_S\}$. For the random element τ_η^2 and τ_S^2 we use an *Inverse-Gamma* distribution whilst for the coefficient α of the temporal auto-regressive model as well as for the correlation parameter ρ_S of the CAR model we fix a Gaussian distribution $\alpha \sim N(0.5,\ 100)$ and $\rho_S \sim N(0, 100)I_{(0,1/m)}$. In the latter, m is the maximum number of neighbors (m) such that, as rule of thumb, its reciprocal is used as limit for the truncation of diffuse uniform distribution. Finally, we adopt an *Inverse-Gamma* distribution also for the common variance of the model. We set independent priors such that $\pi(\theta) = \pi(a)\pi(b)\pi(\alpha)\pi(\tau_\eta^2)\pi(\tau_S^2)\pi(\rho_S)\pi(\tau^2)$.

6.6 Model inference: posterior predictive distribution

We are interested in predictive distribution in order to obtain samples of Y values at q unknown cells $Y_t^0 = \{y_{t,1}^0, y_{t,2}^0, \cdots, y_{t,q}^0\}$ at time t. Let us denote with Y^0 the vector of latent variables at unknown cells and at every time-step of the temporal domain $t = 1, \cdots, T$. In order to obtain predictive values Y^0, we need to generate samples from the conditional distribution of (Y^0, Y), i.e. from the distribution $(Y^0|Y, L, \theta)$ of the predictive model. We know from Section 6.5.1 that the distribution of (Y^0, Y) is multivariate normal. This derives from the assumption that the $Y_{t,p}$ are conditionally independent Gaussian random variable given $\mathbf{W}$ and, furthermore,

$$(Y, Y^0)|\theta, W \sim MN(\mu + \mathbf{w}, \tau^2 I).$$

In this hierarchical framework, we can formulate the posterior predictive distribution conditionally on spatial process $\mathbf{W}$ such that:

$$\pi(Y^0|Y, L, L_t^0) = \int \pi(Y^0|Y, \theta, L^0)\pi(\theta|Y, L)d\theta \tag{6.15}$$

where L^0 is the vector of co-variates at unknown cells, that is the number of lightnings recorded at time t at the cells where precipitation data are not available. In practice, the posterior distribution of the model becomes the prior distribution for updating the likelihood of the predictive model. Notice that also $\pi(Y^0|Y, \theta, L^0)$ has a conditional normal distribution arising from multivariate normal distribution of (Y_t^0, Y). Moreover, since our model implies that Y^0 is conditionally independent of both θ and Y, given W, it follows that:

$$Y^0|Y, \theta \sim MN(\mu^0 + w^0,\ \tau^2 I) \tag{6.16}$$

Thus, although the integration in 6.16 does not have an analytical solution, we can obtain approximations through Monte Carlo methods. In the MCMC algorithm we draw a sample of θ from the posterior $\pi(\theta|Y, L)$ and successively we generate realizations of latent spatial variable (Y^0, Y). The simulations of the posterior distribution and the predictive posterior distribution are implemented in the same step of the algorithm since their mean and variance are the same. This procedure let us to obtain predictive values of rainfall at unknown cells of the spatial domain at each time-step of the temporal domain for every iterations.

6.7 Results

6.7.1 Predictors, exogenous variables and space-time domain

We consider the **15-minutes** and **30-minutes** time-aggregation. Recall from Section 6.1 that the space-time support of May 9, 2006 15-min (30-min) event has 111(112) cells and 68(34) time units whereas the support of August 5, 2004 15-min (30-min) event has 100(104) cells and 36(18) time units. Then, the spatial domain is $D = \{1, \cdots, p, \cdots, n\}$ where p is the 10×10 km square cell of a regular grid and $n = 111$ or 112 as reported in Table 6.7. T_{Ch} is the peak time of the convective event identified on the basis of maximum number of lightnings. T_{Ch} is an exogenous information plugged in Equation 6.9.

Model case	n	T	T_{Ch}	Units
9May2006-15min	111	68	27	7548
9May2006-30min	112	34	13	3808
5Aug2004-15min	100	36	23	3600
5Aug2004-30min	104	18	12	1872

Table 6.7. Space-time domain of the 4 models.

The count of lightnings $L_{t,p}$ generated by the convective system at cell p in $t = 15$, or 30 minutes is the **predictor** of the model. The other exogenous information are represented by the **velocity of propagation** $\boldsymbol{V}$ of lightnings within the convective event and the **RLR** and they are estimated outside of the model (see Section 5.3.2). The velocity of propagation V depends on the intensity of upper tropospheric winds[4]. V influences the temporal and spatial evolution of the event. We incorporate this influence in the time weights matrix, assuming that the correlation at a fixed cell p between number of lightnings at predicting time t and at observed time T_i decreases as the velocity of propagation V increases. Here, we assume that the velocity is 16.1 m/s as suggested in [Levizzani et al. 2010] for convective events spanned within 1000 km and up to 20 hours of duration. Finally, it is worth drawing the attention to the two basic assumptions of this model: 1) the rainfall mass derived from lightnings is uniformly distributed over 10×10 km cell; 2) the rainfall mass derived from lightnings is uniformly distributed for a quarter(half) of hour.

6.7.2 Analysis of data for convective event of 9th of May 2006 and 5th of August 2004

In Figure 6.6, the cross-station mean of rainfall over each time step is plotted versus the correspondent variance. Those figures show a strong variability due to variations in time, that is not entirely removed by log-transforming the observed data.

[4]More precisely, the largest the velocity V, the smallest the mass of precipitation at a single cell p. In fact, V represent the mean velocity referred to the cloud body, and subsequently to the precipitation footprint such that the atmospheric system responsible for convective event spends a small amount of time over the specific cell p.

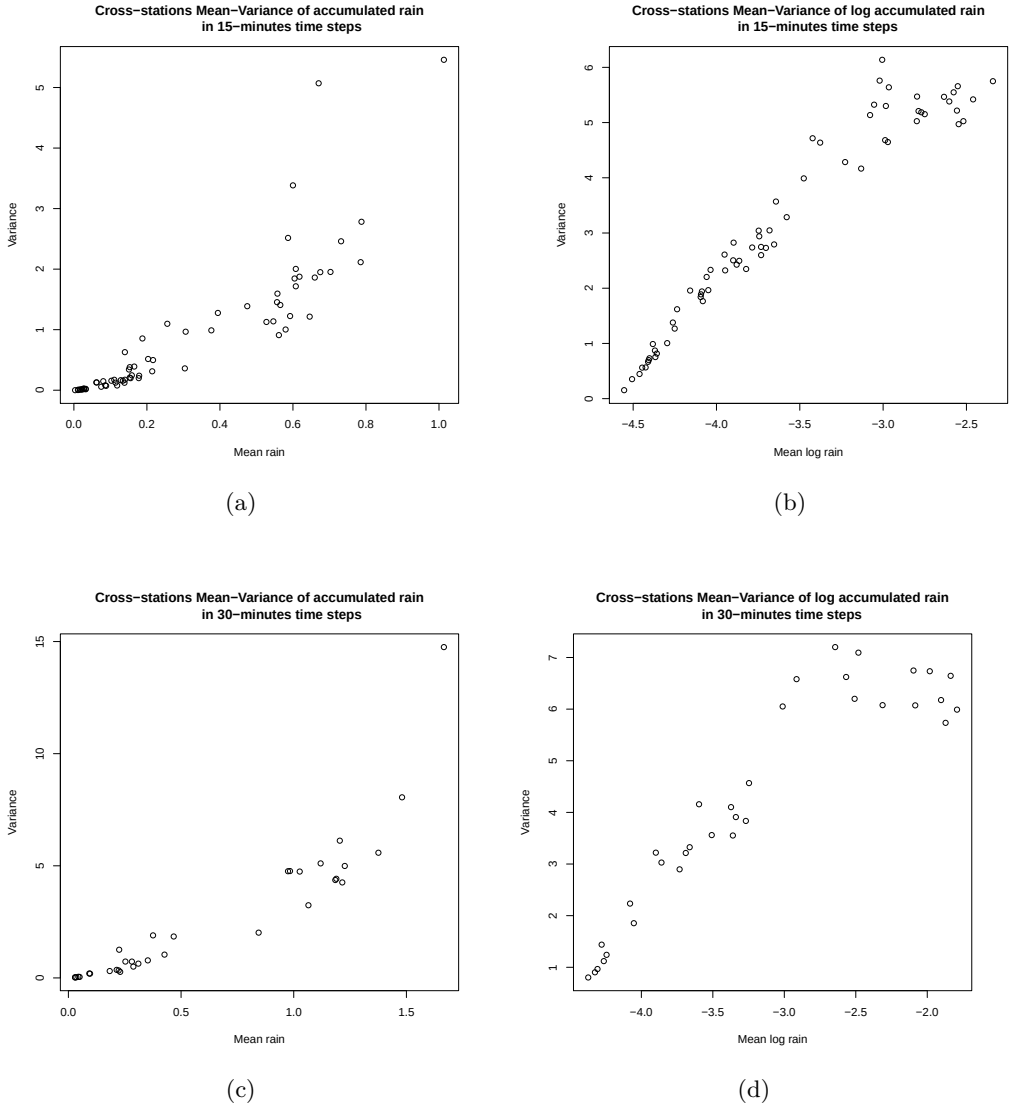
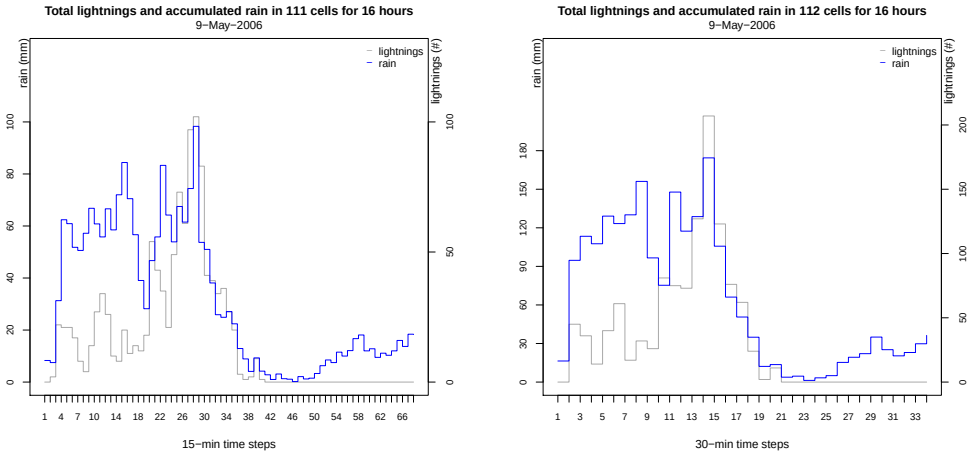


Figure 6.6. Mean-Variance relation of precipitation observed in 68 and 34 time steps during the convective event of May 9, 2006 at 15-min and 30-min time aggregation, respectively. The four Figures correspond to: a) raw scale at 15-min; b) log scale at 15-min; c) raw scale at 30-min and d) log scale at 30-min.

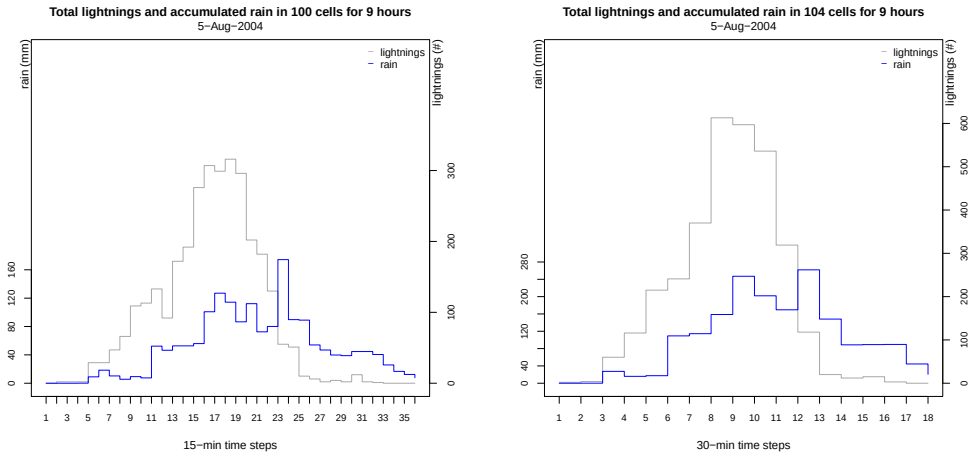
The time series plots of Figure 6.7 reveal that the lightnings activity and rainfall amount follow the same temporal evolution. Here, the counting of lightnings as well as the accumulation of rain per each time step is computed over the 111 (15-min aggregation) and 112 (30-min aggregation) cells.



(a) (b)

Figure 6.7. Total lightnings and rainfall amount observed for the entire duration of the May 9, 2006 convective event: a) 15-min; b) 30-min.

The two time series of the number of lightnings and rainfall amount observed during the 9 hours of duration of the August 5, 2004 convective event at 15-min time aggregation suggest a temporal lag of 75 minutes between their peaks. Nevertheless, the shape of the two curves are very similar.



(a) (b)

Figure 6.8. Total lightnings and rainfall amount observed for the entire duration of the August 5, 2004 convective event: a) 15-min; b) 30-min.

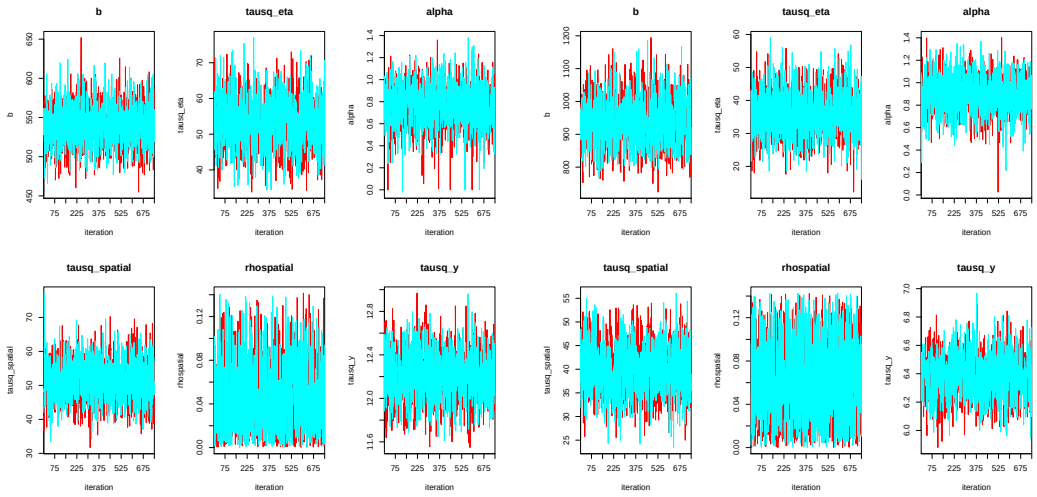
6.7.3 Estimation of parameters

Recall from Section 6.5.1 that the complete set of parameters of our model is $\theta = \{a, b, \alpha, \tau_\eta^2, \tau_S^2, \rho_S, \tau^2\}$ where $\sigma_\eta^2 = 1/\tau_\eta^2$ and $\sigma_S^2 = 1/\tau_S^2$. Recall also that a and b are the polynomial coefficients associated to the velocity of propagation V and that V is exogenous and T_{Ch} is plugged-in the model by finding the time when the maximum number of flashes in 15-, 30-minutes is registered. Moreover, α and τ_η^2 are the parameters of the temporal random component whilst τ_S^2 and ρ_S are the parameters of the CAR model. Finally, τ^2 is the precision assigned to each field. Notice that the inclusion of a δ lightnings-rain delay could be eventually inserted in the model.

A sample of cells is drawn from the entire set of cells of the spatial domain $D = \{1, \cdots, p, \cdots, n\}$ for estimating the parameters. The remaining cells are settled for validating the model. A summary of estimating and validating cells for the four models is reported in Table 6.6.

The model is implemented in JAGS [Plummer 2003] using the package *R2jags* [Yu-Sung and Masanao 2012] to run the simulation within R. We run two chains with dispersed starting points for 20000 iterations, with a burn-in of 5000 and we retain the last 1000 iterations of each chain for estimation. Convergence was inspected both graphically and from several statistics. Simulations summaries for the four models are reported in Table 6.8 whilst the trace plots of all the parameters except a which is equal to zero are presented in Fig. 6.9 and 6.10.

The results obtained show that the MCMC converges rapidly for all the four-cases and also the Potential Scale Reduction Factor $\hat{R}$ proposed by Gelman and Rubin [Gelman and Rubin 1992] confirms the good performance in the parameters estimation phase, being equal to 1 for all parameters except a (see Table 6.8). In fact, the $\hat{R}$ is based on a weighted average of within Var^W and between Var^B chain



(a) (b)

Figure 6.9. Model case 9-May-2006 trace plots of 6 out of 7 parameters obtained with 20000 iterations and 2 chains after a burn-in of 5000: a) 15-min; b) 30-min.

variance, such that:

$$\hat{R} = \sqrt{\frac{\hat{Var}(\theta)}{Var^W}} \tag{6.17}$$

where the $\hat{Var}(\theta)$ of a single parameter is

$$\hat{Var}(\theta) = (1 - \frac{1}{n})\, Var^W + \frac{1}{n}\, Var^B$$

with, given m the number of simulation chains,

$$Var^W = \frac{1}{m} \sum_{j=1}^{m} \{ \frac{1}{n-1} \sum_{i=1}^{n} (\theta_{ij} - \overline{\theta}_j)^2 \}$$

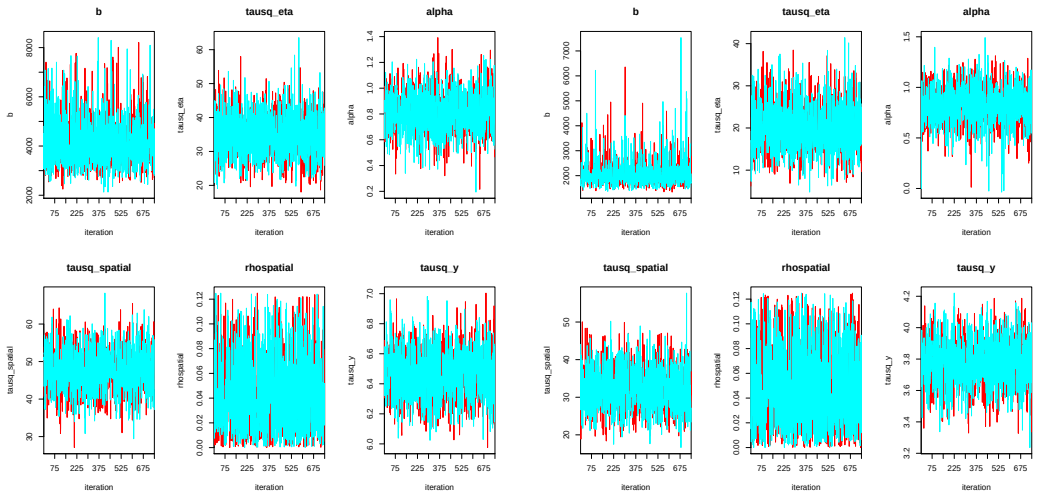
and

$$Var^B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\theta}_j - \widetilde{\theta})^2)$$

where $\widetilde{\theta} = 1/m \sum_{j=1}^{m} \overline{\theta}_j$.

9-May-2006-15min	mean	sd	2.5%	25%	50%	75%	97.5%	$\hat{R}$
a	1.11	7.49	0.00	0.00	0.00	0.00	9.33	2.68
b	538.25	27.44	486.77	519.28	537.18	556.66	596.67	1.00
α	0.76	0.20	0.35	0.63	0.77	0.91	1.13	1.01
τ_η^2	53.21	7.12	39.67	48.28	53.21	57.99	67.29	1.00
τ_S^2	51.16	6.18	39.58	46.75	51.18	55.33	63.33	1.01
ρ_S	0.05	0.03	0.00	0.02	0.04	0.07	0.12	1.00
τ^2	12.20	0.22	11.76	12.05	12.21	12.35	12.64	1.00
9-May-2006-30min								
a	0.02	0.21	0.00	0.00	0.00	0.00	0.00	3.13
b	936.13	71.88	804.86	887.09	932.55	981.03	1087.17	1.00
α	0.89	0.17	0.55	0.79	0.89	1.01	1.20	1.00
τ_η^2	36.07	6.85	23.25	31.26	35.92	40.56	49.73	1.00
τ_S^2	39.92	5.35	29.79	36.19	39.86	43.37	50.78	1.00
ρ_S	0.06	0.04	0.00	0.03	0.06	0.09	0.13	1.00
τ^2	6.37	0.16	6.08	6.26	6.37	6.48	6.69	1.01
5-Aug-2004-15min								
a	0.25	2.47	0.00	0.00	0.00	0.00	0.37	1.31
b	4079.60	979.65	2724.60	3375.50	3910.60	4548.61	6663.88	1.00
α	0.82	0.16	0.50	0.72	0.82	0.93	1.12	1.00
τ_η^2	34.80	6.49	22.52	30.16	34.64	39.22	47.70	1.00
τ_S^2	47.47	5.85	36.23	43.31	47.41	51.26	58.73	1.01
ρ_S	0.04	0.03	0.00	0.02	0.04	0.06	0.11	1.00
τ^2	6.46	0.17	6.14	6.35	6.46	6.58	6.79	1.00
5-Aug-2004-30min								
a	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.51
b	2072.29	535.64	1480.96	1753.67	1949.35	2231.40	3484.40	1.00
α	0.82	0.19	0.45	0.71	0.83	0.94	1.18	1.00
τ_η^2	20.29	5.72	9.92	16.19	19.90	23.94	32.41	1.00
τ_S^2	32.39	5.65	21.80	28.50	32.23	36.22	43.69	1.00
ρ_S	0.05	0.03	0.00	0.02	0.04	0.07	0.12	1.00
τ^2	3.77	0.15	3.48	3.67	3.77	3.87	4.07	1.00

Table 6.8. MCMC Posterior Inference of the four model cases 9-May-2006-15min, 9-May-2006-30min, 5-Aug-2004-15min and 5-Aug-2004-30min.



(a) (b)

Figure 6.10. Model case 5-Aug-2004 trace plots of 6 out of 7 parameters obtained with 20000 iterations and 2 chains after a burn-in of 5000: a) 15-min; b) 30-min.

6.7.4 Evaluation of rainfall fields prediction

The evaluation of the model performance in predicting rainfall fields is done analyzing the predicted rainfall values obtained at validation cells in comparison with the observed values recorded at the same cells. This evaluation phase is performed by means of both a graphical analysis and summary indexes such as RMSE and probability of detection. The number of validation cells of each model is reported in Table 6.6. Recall from Section 6.7.3 that we run two chains for 20000 iterations, retaining the last 1000 of each chain after a burn-in of 5000 iterations, then we take the median of the retained 2000 values drawn from the posterior predictive distribution described in Section 6.6. Those values are considered as predicted values and compared to the corresponding observed values, which are themselves a median of the observations recorded by the whole rain gauges settled in each cell. The core of the study is represented by the four basic cases described along the previous sections, which are the two convective events of the 9th of May 2006 and the 5th of August 2004 in the two time aggregations of 15- and 30- minutes. The $k = 4$ minimum number of neighbors is adopted in these four cases. Nevertheless, model's options such as $k = 2$ and a delay of one time-step in the lightnings-precipitation relation are tested and discussed here although not shown. To simplify the notation, we adopt the scheme in Table 6.9, where the *date* indicates the convective event; *15min* or *30min* is the time aggregation; *k2* or *k4* stands for the minimum number of neighbors in the spatial dependence structure and, finally, *lag0* or *lag1* indicates the framework in which the accumulated precipitation at time t depends on the amount of lightnings at time t or at time $t-1$. Notice that Table 6.9 reports only the list of tested cases. Furthermore, for the evaluation we consider two thresholds of rainfall levels: $0.2\,mm$ and $1\,mm$. The former is usually adopted to discriminate rainy against no rainy records whilst the latter is arbitrary chosen taking into account

both the proportion of cases with rain larger than 1 mm shown in Table 6.3 and the rain rates reported in Table 6.4. Firstly, we discuss some aspects deriving from the graphical analysis; then, we present the evaluation based on summary indexes.

Model case	Short name
9May2006-15min-k4-lag0	**A1**
9May2006-15min-k2-lag0	*A1-k2*
9May2006-15min-k2-lag1	*A1-k2-lag1*
9May2006-30min-k4-lag0	**A2**
9May2006-30min-k2-lag0	*A2-k2*
9May2006-30min-k4-lag1	*A2-lag1*
5Aug2004-15min-k4-lag0	**B1**
5Aug2004-30min-k4-lag0	**B2**
5Aug2004-30min-k4-lag1	*B2-lag1*

Table 6.9. Short name of tested model cases (in bold the four main cases).

Graphical evaluation. The set of graphs of time series predictions at each validation location is presented in appendix 6.7.4. As an example, we report here the prediction graphs of model A1. The scatter plots of Figure 6.11 depict predicted versus observed values at validating cells for the entire set of observations and for the two subsets with rainfall values greater or equals than 0.2 mm and 1 mm. Moreover, each point is classified into Charging (red) or Mature-Dissipating (blue) phase. Notice that Mature and Dissipating phases are always intended to be a unique phase, named Mature. This fact derives from the assumption we have done in Section 6.3 when formulating the fixed effect. In the graph of Figure 6.12 observations time series (blue rings) along with correspondent predicted values (black crosses) are shown. Furthermore, limits of credibility interval of predictions calculated as 0.025 and 0.975 quantiles are drawn(red lines) and correlation between observed and predicted is visualized.

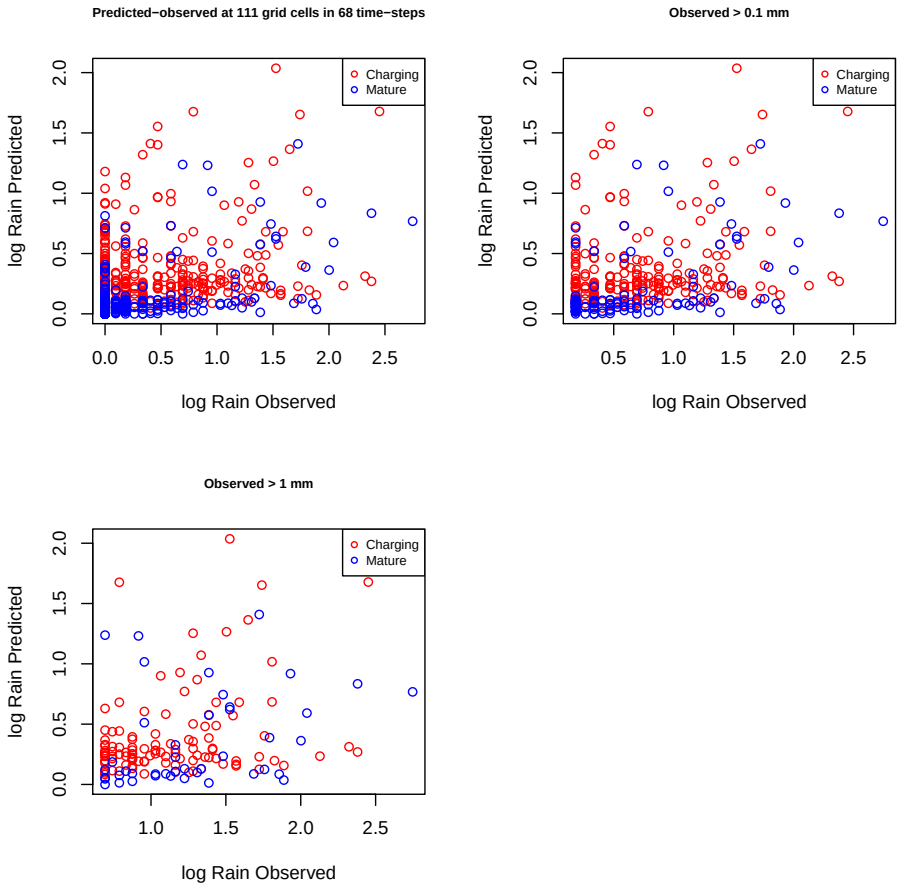


Figure 6.11. Predicted versus observed values at validating cells for model case A1 divided into Charging and Mature-Dissipating phases for three levels of precipitation quantity.

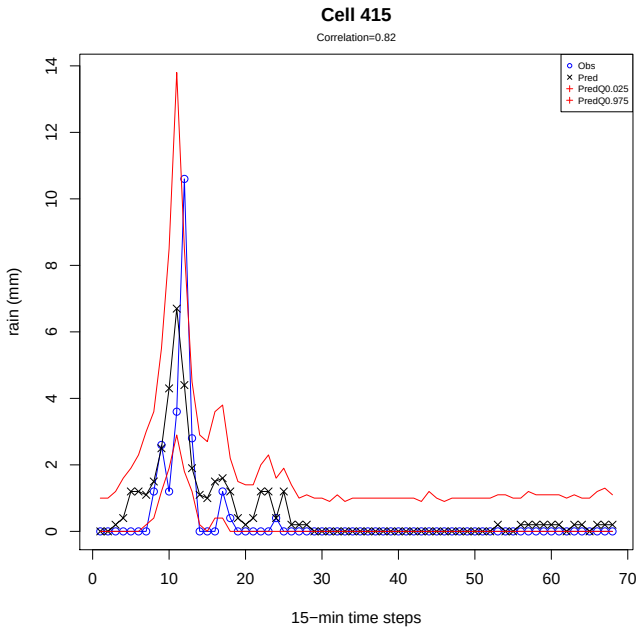


Figure 6.12. Time series of rainfall predicted values at validating cell 415 for 9-May-2006-15min-k4 case: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

Scatter plots of Figure 6.13 and 6.17 point out that both **Model A1** and **Model A2** have an acceptable performance tough are not fully able to correctly predict cases with observed rainfall equals to zero. In fact, several wrong predictions are present either with positive values of precipitation instead of zero or with zero precipitation instead of positive one, the latter being associated mainly with the Mature phase. By visualizing time series plots of Figure 6.14, 6.15 and 6.16 referred to Model A1 as well as Figure 6.18, 6.19 and 6.20 of Model A2 emerge that main time dynamic is captured. Nevertheless, Model A1 reveals a good capability in predicting rain quantity associated to peaks, being errors connected basically to low rain rates whereas Model A2 generally fails in it. The option *k2* instead of *k4* improves the prediction performance both for positive rainfall levels, i.e. greater or equals than $0.2\,mm$ and for rainfall levels greater or equals to $1\,mm$ but it worsens the global predictive capability, that is when also the zero rainfall cases are included. On the other hand, the option *lag1* seems to be uninfluential on the predictive performance.

The analysis of graphs relative to **Model B1** and **Model B2** reveals that the predictive performance is more satisfying than that of Model A1 and A2. Like for Model A1 and A2, the issue of predicting positive rainfall when observed equals to zero persists, however the opposite case is improved (see Fig. 6.21 and 6.24). Time series plots of Model B1 predictions (Figure 6.22 and 6.23) show the same features observed for Model A2. The same argumentation are valid for Model B2, as well (see Fig. 6.25 and 6.26). However, Model B2 shows a very satisfying performance in predicting zero rainfall cases. The option *lag1* applied on Model B2 seems to slightly improves the performance in predicting positive rainfall when observed equals to zero.

Finally, we use two indexes to resume the graphical analysis: the cross correlation and the empirical coverage. Cross correlation is calculated as an average of correlation between predicted and observed values obtained in the validation cells. Empirical coverage is calculated for the entire spatial domain on the 90% credible interval. In practice, based on observations $Y(t,p) = y_{t,p}$, we need empirical coverage for each prediction $\tilde{y}_{t,p}$ to be as follows:

$$Pr(l(\tilde{y}_{t,p}) < y_{t,p} < u(\tilde{y}_{t,p})) = 0.90 \tag{6.18}$$

where $l(\tilde{y}_{t,p})$ and $u(\tilde{y}_{t,p})$ are the 5th and 95th percentiles derived from the 2000 simulations from the posterior predictive distribution. Cross correlation and empirical coverage for each model case are reported in Table 6.10. Notice that cross correlation value of Model B2 is markedly larger than other cases because of the capability in predicting zero rainfall. Besides, predictive intervals have empirical coverage closer to the nominal values of 90% for all the model cases attesting to the accuracy of predictions of our modelling approach.

Summary indexes evaluation. We use two types of indexes to evaluate the predicting performance of our modelling approach: the Root Mean Square Errors (RMSE) and the probability of detecting rainy events. RMSE is computed for three classes of observed quantity of precipitation: any quantity, rain greater or equals to $0.2\,mm$ and rain greater or equals to $1\,mm$. The other is a set of indexes based on the prediction of two complementary events. In particular, we define these two

Model cases	Cross correlation	Empirical Coverage (%)	DIC
A1	0.46	90.7	2158.421
A2	0.49	91.2	3167.798
B1	0.48	92.8	3115.878
B2	0.70	92.4	2035.139

Table 6.10. Cross correlation, Empirical coverage and DIC.

complementary events as *Rain* and *No Rain* although we refer to two different thresholds adopted also for RMSE such that we consider probability of detection of Rain ($\geq 0.2\,mm$) against No Rain ($< 0.2\,mm$), and Rain ($\geq 1\,mm$) against No Rain ($< 1\,mm$). The set of indexes is composed of Probability of Hits on Total (POHT), which is the percentage of cases correctly predicted in both classes *Rain* and *No Rain*; Probability of Detection (POD), which is the number of cases correctly predicted in class *Rain* on the total cases observed in the same class; Probability of False Detection (POFD), which is the percentage of cases predicted as rainy but observed as no rainy with respect to no rainy observed cases; finally, Probability of False Alarm (POFA), which is the percentage of cases predicted as rainy but observed as no rainy with respect to the whole rainy predicted cases, i.e. wrong cases when the model predicts rain. As suggested in [Barnes et al. 2009], we define these four probability using the notation in Table 6.11, then we can explicit them as in Table 6.12.

		Observed		
		Rain	*No Rain*	
Predicted	*Rain*	a	b	a+b
	No Rain	c	d	c+d
		a+c	b+d	n

Table 6.11. Cases for calculating Probability of Hits on Total (POHT), Probability of Detection (POD), Probability of False Detection (POFD) and Probability of False Alarm (POFA) used as evaluation indexes.

Indexes	Acronym	Formula
Probability of Hits on Total	POHT	$(a+d)/n$
Probability of Detection	POD	$a/(a+c)$
Probability of False Detection	POFD	$b/(b+d)$
Probability of False Alarm	POFA	$b/(a+b)$

Table 6.12. Notation and formula of indexes used for the evaluation.

The POHT, POD, POFD and POFA are extensively used for evaluating climatic forecasting models as well as climatic fields reconstruction models[5]. It is worth

[5]The definitions of POHT, POD, POFD and POFA given here are taken from the Forecast

considering POHT when predictions of *Rain* and *No Rain* events have the same importance, as in our case. On the other hand, POD, POFD and POFA are focused on the evaluation of predictions in *Rain* class. A jointly evaluation of these indexes is generally recommended.

The RMSE values obtained from the four cases A1, A2, B1 and B2 are shown in Table 6.13 whereas the probability indexes are reported in Table 6.14.

RMSE	ModelA1 (mm)	ModelA2 (mm)	ModelB1 (mm)	ModelB2 (mm)
global	0.410	0.541	0.541	0.524
observed$\geq 0.2\,mm$	0.954	1.019	0.725	0.710
observed$\geq 1\,mm$	1.679	1.742	1.493	1.085
Relative RMSE	ModelA1 (%)	ModelA2 (%)	ModelB1 (%)	ModelB2 (%)
global	2.7	3.1	1.6	0.9
observed$\geq 0.2\,mm$	6.3	5.8	2.1	1.2
observed$\geq 1\,mm$	11.2	9.9	4.3	1.8

Table 6.13. RMSE of predicted against observed for 3 classes of rain: rain and no rain (global), positive precipitation (rain$\geq 0.2mm$) and rain$\geq 1mm$ (the relative RMSE is calculated with respect to the maximum quantity of observed precipitation).

Recall from Table 6.4 that the observed mean (maximum) quantity of precipitation excluding zero precipitation for the events associated to the cases A1, A2, B1 and B2 are millimeters 1.299(15), 1.971(17.6), 1.386(34.6) and 2.194(58.8), respectively. RMSE values calculated on the whole data (global) are nearly 1/3 of the observed mean (0.410/1.299 and 0.541/1.386) for A1 and A2 whereas they are approximately 1/4 (0.541/1.971 and 0.524/2.194) for B1 and B2. RMSE are in the range of observed mean for A1 and A2 and nearly 1/3 for B1 and B2, when referring only to positive precipitation. Nevertheless, it is worth considering the large variability of precipitation and the high values of extreme precipitation: the relative RMSE calculated with respect to the maximum quantity of observed precipitation is also reported in Table 6.13. On the basis of all the above considerations we can affirm that the model performance is satisfactory in predicting precipitation quantity and that the 30-minutes cases have a better capability than 15-minutes. Besides, B1 and B2 concerning the more intense event of 5th of August 2004 show a better performance than A1 and A2. Furthermore, the options *k2* for Model A1 let us to obtain an appreciable improvement in RMSE values, which are 0.388 for global, 0.865 for observed cases with rainfall greater or equals to $0.2\,mm$ and 1.380 for $1\,mm$ threshold. On the converse *lag1* option is uninfluential on the predictive performance and in some cases, such as Model B2 it worsen the prediction quality (not shown here).

The analysis of probability indexes reported in Table 6.14 reveals a satisfactory performance of our modelling approach with some differences between A and B as well as between $0.2\,mm$ and $1\,mm$ thresholds. In fact, POHT indicates that the correctly predicted cases in both classes *Rain* and *No Rain* ranges from 69.2% to 77.8% with $0.2\,mm$ threshold and from 83.8% to 89.7% with $1\,mm$ threshold. POD

Verification Glossary of NOAA-Space Weather Prediction Center at `http://www.swpc.noaa.gov/forecast_verification/Glossary.html` and from [Barnes et al. 2009].

values confirm a good performance in predicting cases in *Rain* class, particularly, for B1 and B2 that are able to predict correctly 83.2% and 86.7% of total cases with rain greater or equals than $0.2\,mm$, respectively. On the converse, the capability of predicting rainy cases above $1\,mm$ is in agreement with our comments in the previous paragraph about limitations of the model in predicting rainfall quantity, i.e. underestimation of higher levels.

	ModelA1 (%)	ModelA2 (%)	ModelB1 (%)	ModelB2 (%)
Threshold=0.2 mm				
Prob of hits on total	75.1	69.2	75.4	77.8
Prob of detection	56.6	66.5	83.2	86.7
Prob of false detection	19.8	29.6	29.2	28.7
Prob of false alarm	55.9	50.3	37.7	31.5
Threshold=1 mm				
Prob of hits on total	89.7	86	84.9	83.8
Prob of detection	14.2	18.3	25.4	50
Prob of false detection	2	1.8	6.6	8.9
Prob of false alarm	56	35.5	64.4	44.7

Table 6.14. Comparison of model performance indexes calculated on partitioned Rain/No Rain predictions, where Rain is defined either with precipitation quantity $\geq 0.2mm$ or $\geq 1mm$.

POFD is generally ranging from 2% for Model A1 and rain amount$\geq 1mm$ to 29.6% for model A2 and rain $\geq 0.2mm$ whilst the POFA is particularly large. In fact, the number of cases wrongly predicted as rainy on the whole case predicted as rainy are acceptable when B1 and B2 attempt to predict cases with rainfall greater or equals to $0.2\,mm$ for which predictions are wrong in 37.7% and 31.5% of cases, respectively whilst the same percentages in cases with rainfall greater or equals to $1\,mm$ are discouraging. The POFA is too high also for Model A1 and Model A2: 55.9% and 50.3% of cases with $0.2\,mm$ threshold, and 56% and 35.5% of cases with $1\,mm$. In conclusion, our proposed model in its four main cases show a good performance when the evaluation is done subdividing observations with a threshold equals to $0.2\,mm$, i.e. cases in class $No\,Rain < 0.2\,mm$ and cases in class $Rain \geq 0.2\,mm$. In this framework, Model B2 has the best performance. On the converse, when observations are partitioned in $No\,Rain < 1\,mm$ and $Rain \geq 1\,mm$, the model reveals some problems, in particular the underestimation of rainfall quantity. This fact is somehow due to the physical features of rainfall convective event which can have different rain rates during its development. In particular, the larger are the rain rates in a cell, the smaller is the propagation in surrounding cells. Our spatial weight function takes into account eight cells around the predicting cell for a total square area of 30×30 km, that maybe too extended for the cases described above and, especially, when the probability of encountering zero precipitation cells is significantly high, as it is for our two study events (see Table 6.3). Consequently, the rainfall quantity in similar cases, i.e. cases with larger rain rates, is underestimated. Furthermore, it is worth noticing from the same Table that the percentage of zero precipitation cases of the event associated to Model B2 is relatively smaller in comparison to other

events. This fact might have caused Model B2 to better predicting rainfall cases greater than $1\,mm$ with a POD of 50%. In any case, a revision of spatial weight function has to be considered. Furthermore, the options *k2* or *lag1* do not improve the prediction performance. For instance, Model B2-lag1 lead us to obtain the same results as Model B2. On the other hand, from the graphical analysis and the overall RMSE values we can see that the model well captures the events dynamic behaviour. Then some adjustments are necessary mostly in the definition of the mean function.

Conclusion of Part II

The main goal of this Part of the Thesis is to build a model to predict the accumulated precipitation at unknown locations using lightnings counting. The model we propose is based on a stochastic approach as we are interested not only in predicted values but also in the uncertainty affecting such predictions. Nevertheless, the first contribution is a simple and effective scan statistic procedure used for the identification of single storms among several severe meteorological events. This procedure let us to identify 767 convective events during the period March-September of 2003-2006 from which we estimate the Rainfall Lightnings Ratio. As a consequence of the identification process, we are also able to delineate the three life-time phases of a convective event: Charging, Mature and Dissipating using the beginning, peak and ending times of lightnings temporal evolution to separate them. This aspect represents an advantage of our approach since it let us to incorporate into the model the different features of rainfall propagation in time assuming different weight structures in the equation which converts lightning into rain (Eq. 6.5). Furthermore, as we have a large number of identified events we can build a reliable classification distinguishing among Small, Medium and Large events in terms of number of lightnings. This allows us to estimate a different Rainfall Lightning Ratio for each event size. However, the validation of the scan statistics procedure is fully been carried out only for 7 events and we believe it requires further investigation using the in-clouds temperature inversion as validation tool.

The two convective events of May 9, 2006 and August 5, 2004 chosen for testing our model are both classified in the category of *Large events*. This fact have determined the adoption of the same value of Rainfall Lightnings Ratio in the estimation of rainfall predictions. However, the phenomenological features of the two convective events are substantially different from one another both in lightning intensity and in rainfall rates. In particular, the rain rates of the 5th of August 2004 are larger than those of the 9th of May 2006 as well as the lightning intensity. The model predictive performance may depend on these event characteristics.

We estimate our model in four different data situation: Model A1 and A2 for May 9, 2006 event and Model B1 and B2 for August 5, 2004 event, where version 1 and 2 are associated to 15-and 30-minutes time aggregation, respectively. Predictive intervals have empirical coverage closer to the nominal values of 90% for all the model cases confirming the accuracy of predictions of our modelling approach.

In general it appears that the proposed model is able to capture the time dynamic in all considered situations. In particular Model B2 shows a better performance in fitting the data than the others (see Table 6.10 for the DIC and Table 6.13 for the RMSE evaluation). Model A1, the second best in terms of DIC, predict the events

peaks better than the others. In terms of RMSE we can see that predictions are more reliable for the more intense event of August 5, 2004 reinforcing the intuition that RLR based estimate works better for intense events.

The evaluation done by using Probability of Hits on Total, Detection, False Detection and False Alarm makes clear that our modelling approach show a good predictive performance when the evaluation is done partitioning cases with a threshold equals to $0.2\,mm$ whilst reveals some problems with a threshold equals to $1\,mm$.

Finally, we briefly summarize the results obtained varying the neighbourhood dependence structure or allowing for one time step delay in the lightning-rainfall relation although at the moment this analysis is not complete. For instance, the adoption of *k2* option, that is a minimum number of 2 neighbours instead of 4 lead to improve Model A1 in terms of RMSE and Cross Correlation but it is uninfluential for Model A2 (Model B1-k2 and B2-k2 have not been tested). On the other hand, the use of *lag1* option, which is based on the assumption that rainfall at time t is caused by lightnings at time $t-1$ generally worsens the predictive performance of the model or it is uninfluential.

On the whole, our modelling approach shows a good capability in capturing time dynamic although it underestimates substantially the rainfall level with a large performance variability depending on event type and time aggregation. This fact is most likely due to the erratic physical features of rainfall convective event but also to some limitations in the specification of the space and time weight functions in the fixed part of the model. There, we model temporal evolution of the convective event weighting each predictive instant t with the number of lightnings recorded during both previous and following times provided that they belong to the same phase of the evolution: Charging or Mature-Dissipating phase, and spatial propagation using a system of weights that takes into account the whole lightnings recorded in the eight surrounding cells of cell p for the entire duration of the convective event. As a matter of fact, rainfall convective events can have different rain rates during their development such that the larger are the rain rates in a cell, the smaller is the propagation in surrounding cells. In these cases, our spatial weight function incorporate an area that is too extended since it takes into account eight cells around the predicting cell for a total square area of 30×30 km. In this sense, the assumption of constant neighborhood structure for the entire duration of the event is also too strong. This fact clearly emerges when adopting the *lag1* option from which one would expect to have an improvement of predictive performance since a delay between lightning and rainfall is conform to the empirical evidence. Another possible modification of the fixed component can be done implementing the calculation of velocity of propagation directly into the model instead of using an external value. This implementation would allow to specify the propagation patterns of each single event. Notice that these modifications can be easily achieved in our modelling approach. On the other hand, a modification of random component can be adopted to treat zero inflated rainfall distributions, such as mixture between a distribution that specifies the probability of having positive precipitation and a probability density function for the rainfall accumulation. This alternative modeling might improve the prediction of zero rainfall.

As further developments we envision three main points: 1) we need to consider different ways of aggregating point rain records in the grid superimposed to the

study area for instance use rain density computed as cumulated rain/cell's area; 2) as above we would like to investigate the use of lightnings density instead of their counts; and finally 3) we want to develop a data fusion model including satellite rain based on the best model we can obtain once 1 and 2 have been investigated.

Acknowledgement of Part II

The satellite rainfall dataset used in this Thesis is from GSMaP Project. The GSMaP Project was sponsored by JST-CREST and is promoted by the JAXA Precipitation Measuring Mission (PMM) Science Team, and the GSMaP products were distributed by the Earth Observation Research Center, Japan Aerospace Exploration Agency. The lightning records dataset is from CESI-Sirf acquired by Consorzio Lamma (Regione Toscana and Cnr-Ibimet). Finally, the weather stations rainfall dataset is from Biometeorology Institute of CNR-National Research Council.

Appendix of Part II

Predictions graphs

In this appendix a graphical evaluation of time series predictions at each validation location for the four model cases A1, A2, B1 and B2 is presented. In these graphs observations time series (blue rings) along with correspondent predicted values (black crosses) are shown. Furthermore, the limits of credibility interval of predictions calculated as 0.025 and 0.975 quantiles are drawn(red lines) and the correlation between observed and predicted are visualized. Any other explanation is given in Section 6.7.4.

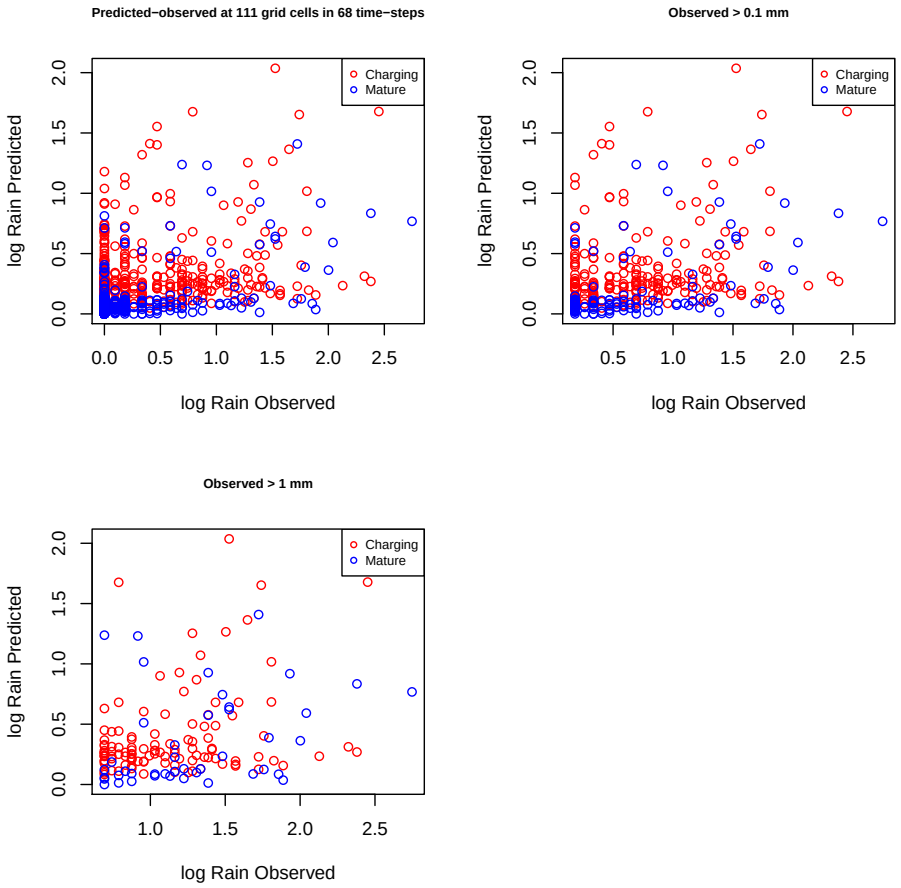


Figure 6.13. Predicted versus observed values at validating cells for 9-May-2006-15min-k4 model subdivided into Charging and Mature phases for three levels of precipitation quantity.

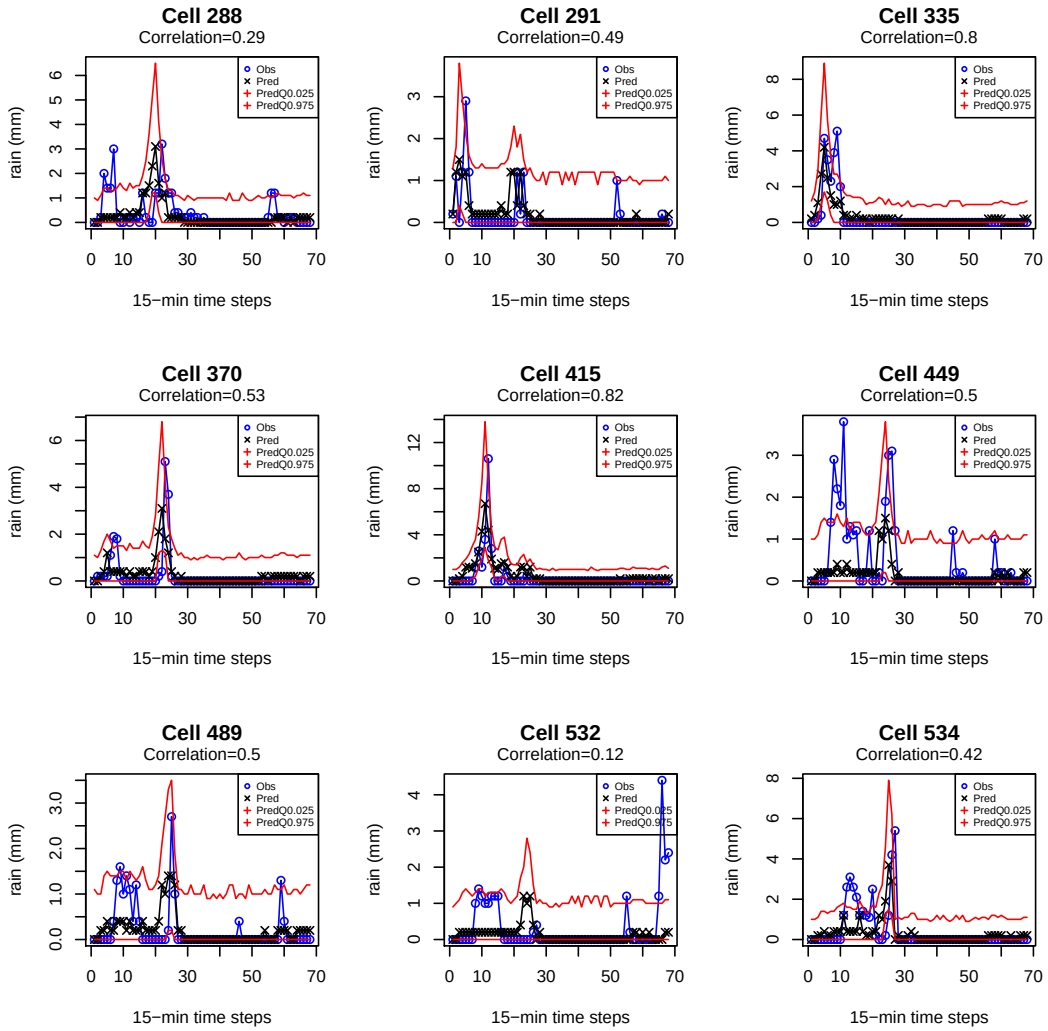


Figure 6.14. Time series of rainfall predicted values at validating cells for 9-May-2006-15min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

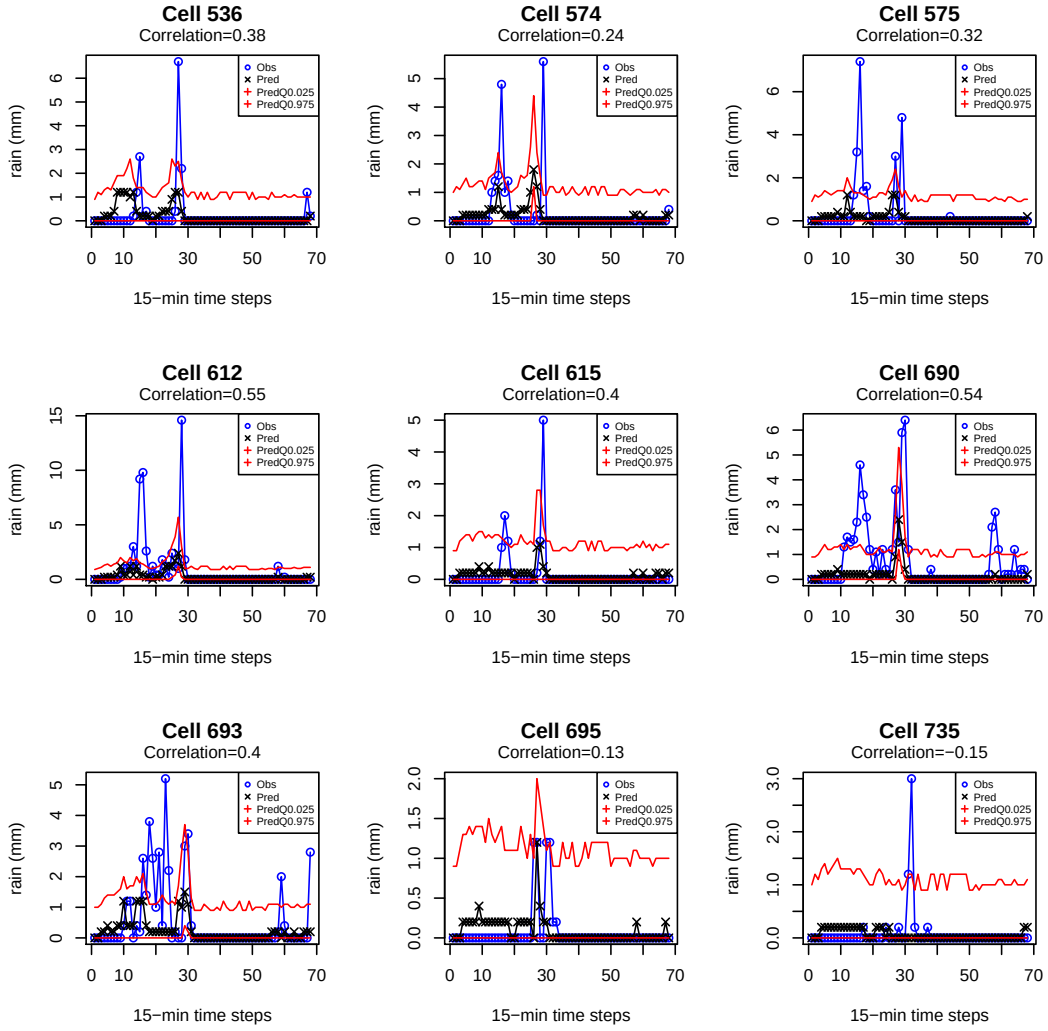


Figure 6.15. Time series of rainfall predicted values at validating cells for 9-May-2006-15min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

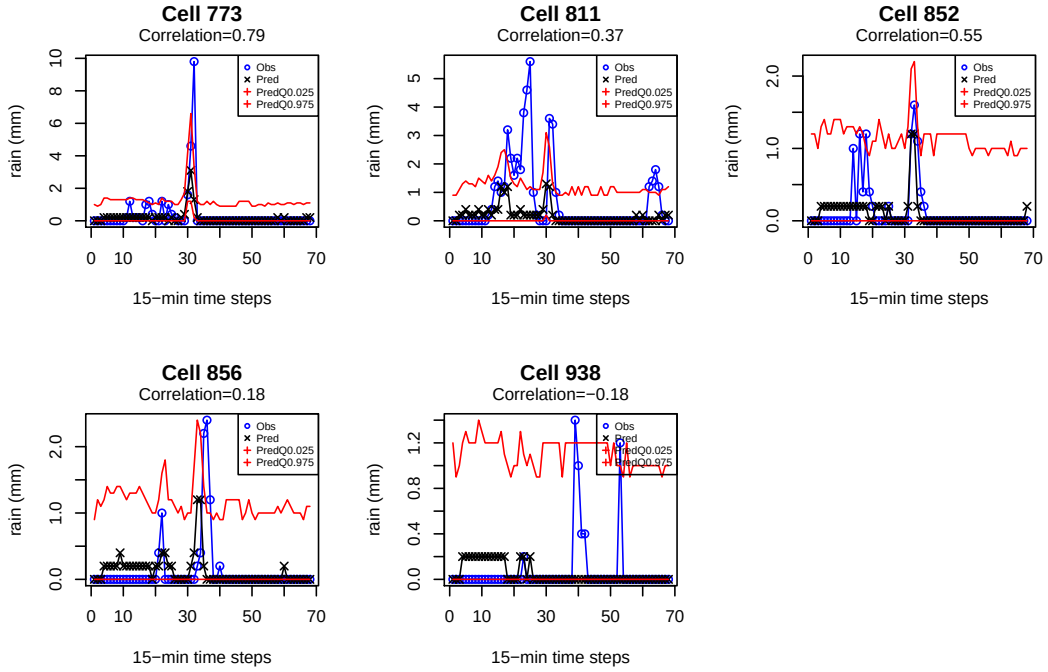


Figure 6.16. Time series of rainfall predicted values at validating cells for 9-May-2006-15min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

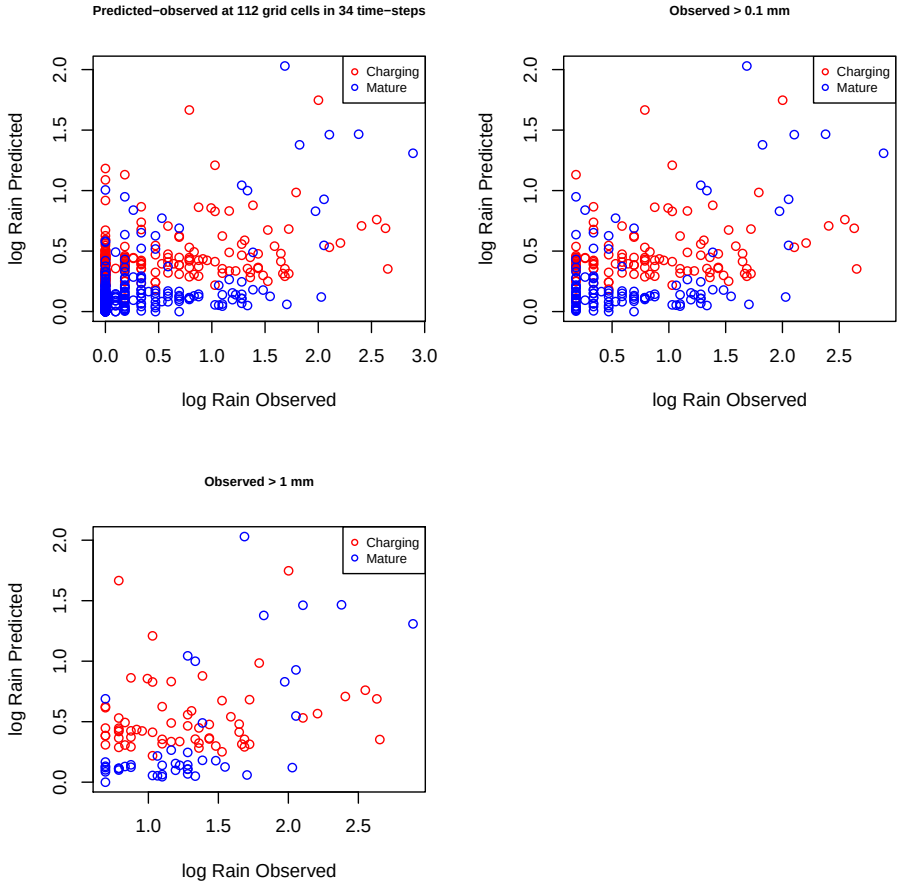


Figure 6.17. Predicted versus observed values at validating cells for 9-May-2006-30min-k4 model subdivided into Charging and Mature phases for three levels of precipitation quantity.

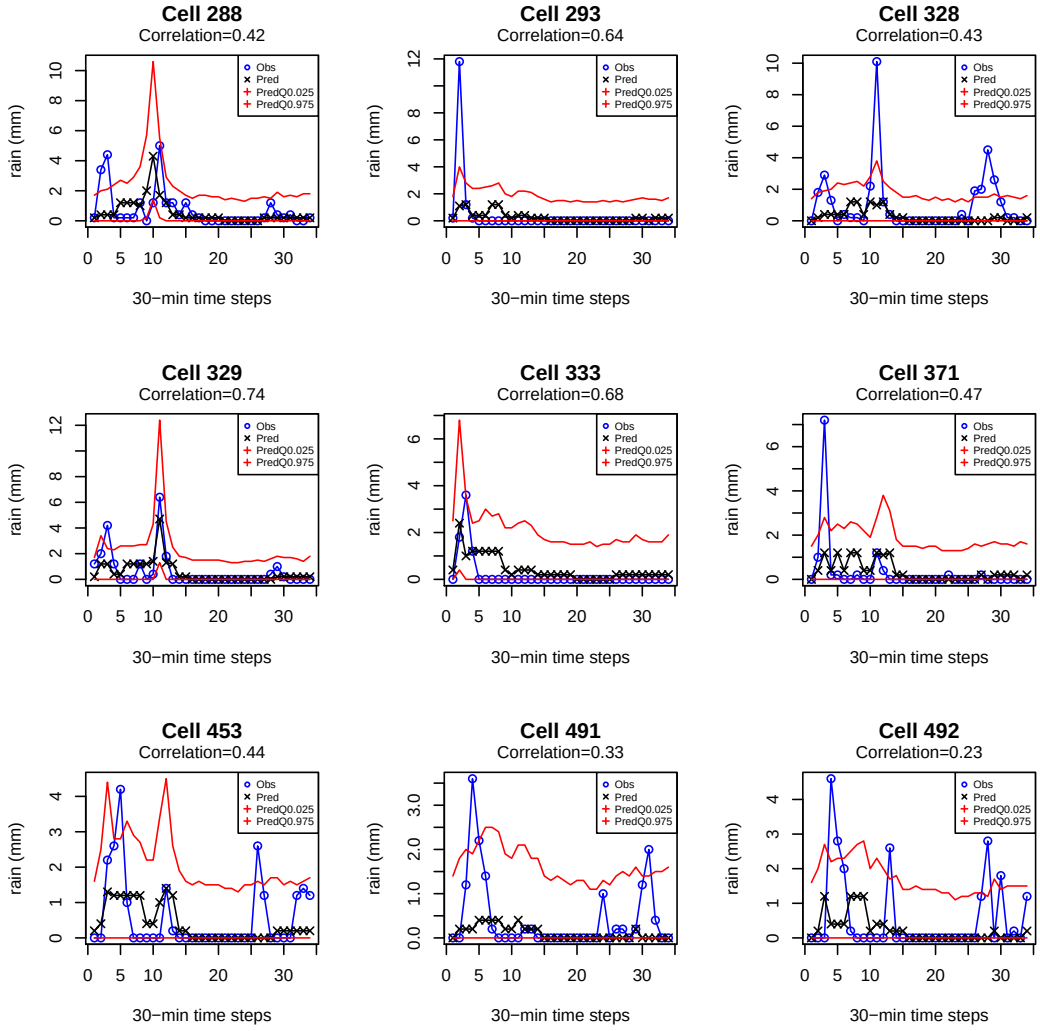


Figure 6.18. Time series of rainfall predicted values at validating cells for 9-May-2006-30min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

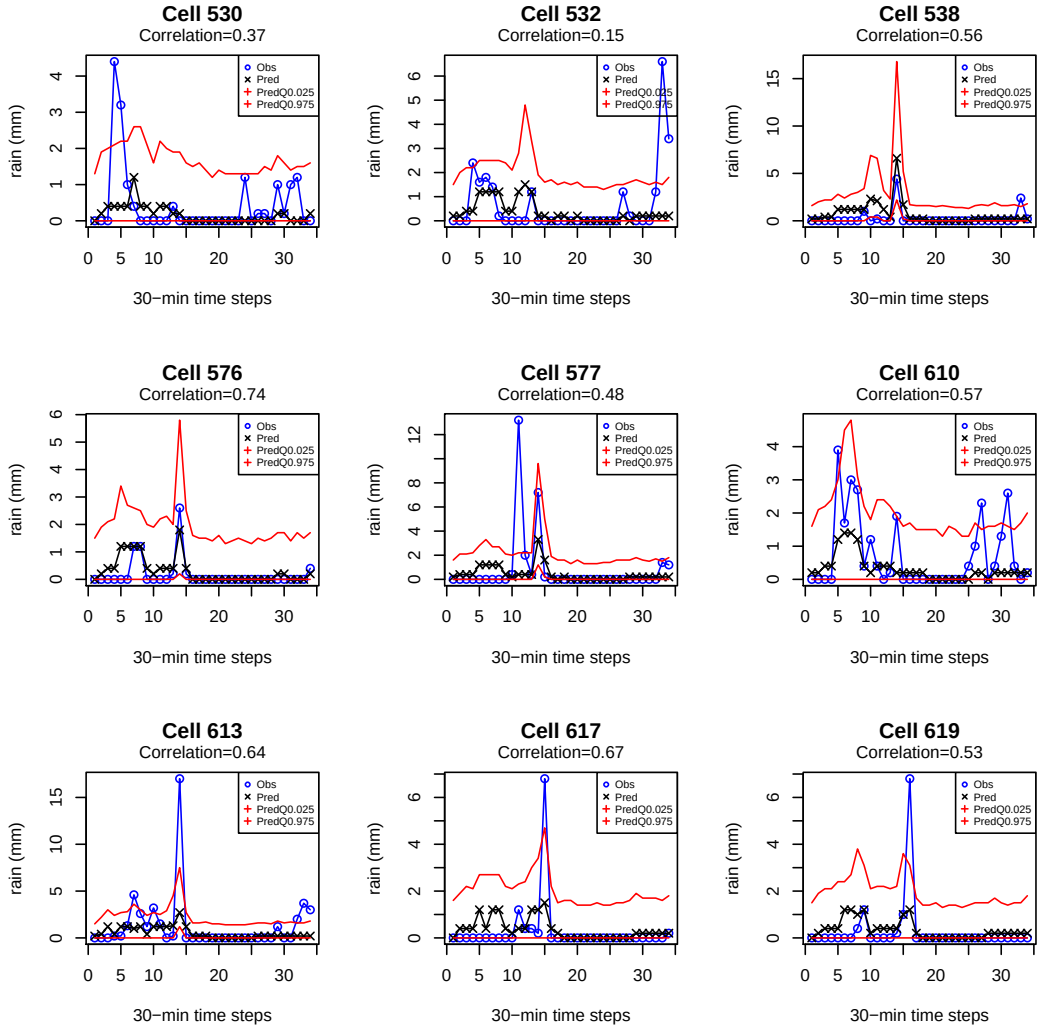


Figure 6.19. Time series of rainfall predicted values at validating cells for 9-May-2006-30min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

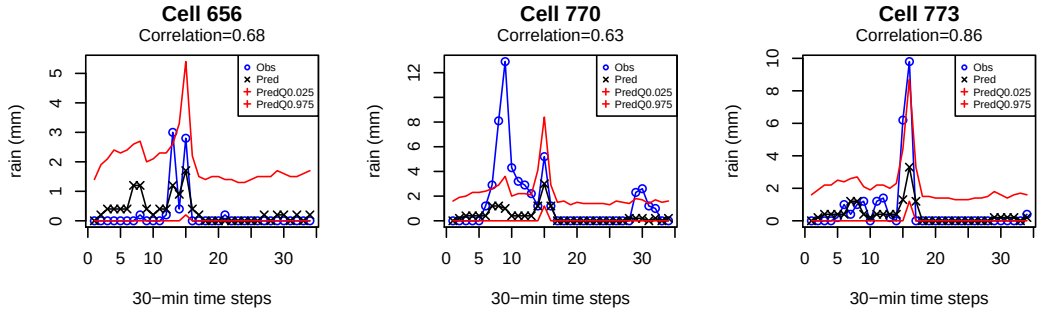


Figure 6.20. Time series of rainfall predicted values at validating cells for 9-May-2006-30min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

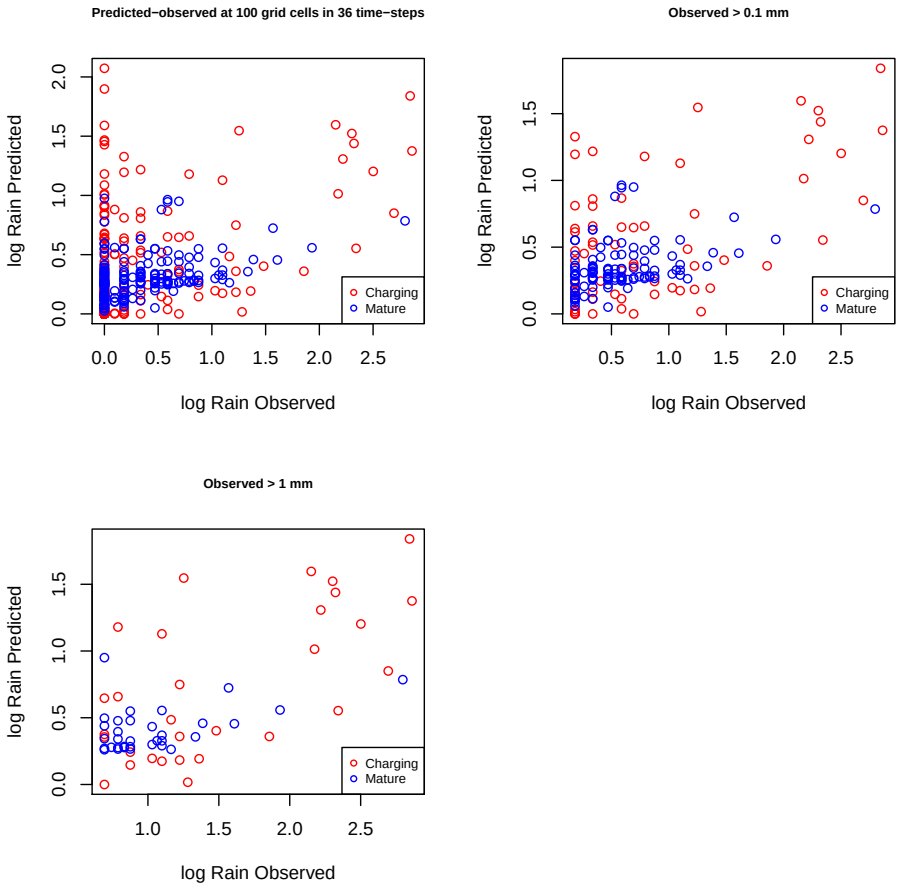


Figure 6.21. Predicted versus observed values at validating cells for 5-Aug-2004-15min-k4 model subdivided into Charging and Mature phases for three levels of precipitation quantity.

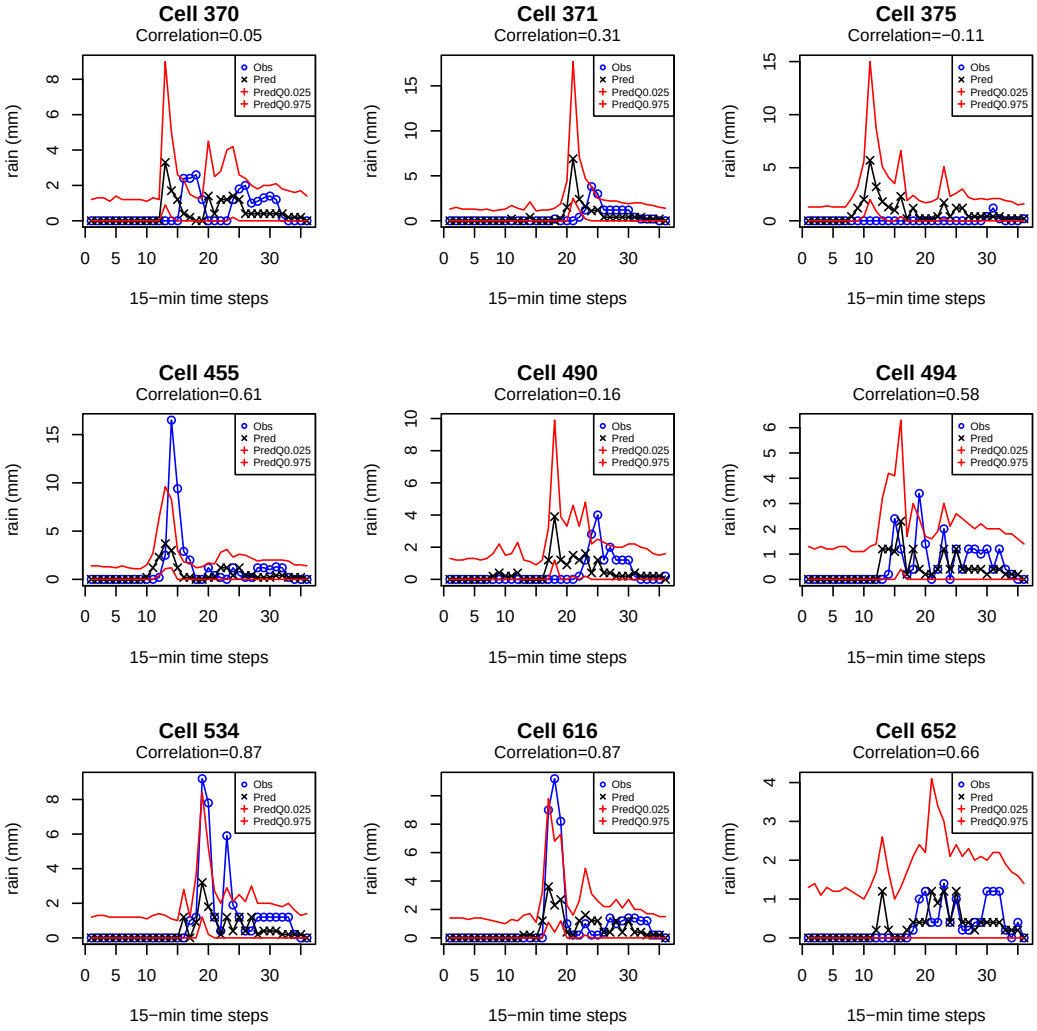


Figure 6.22. Time series of rainfall predicted values at validating cells for 5-Aug-2004-15min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

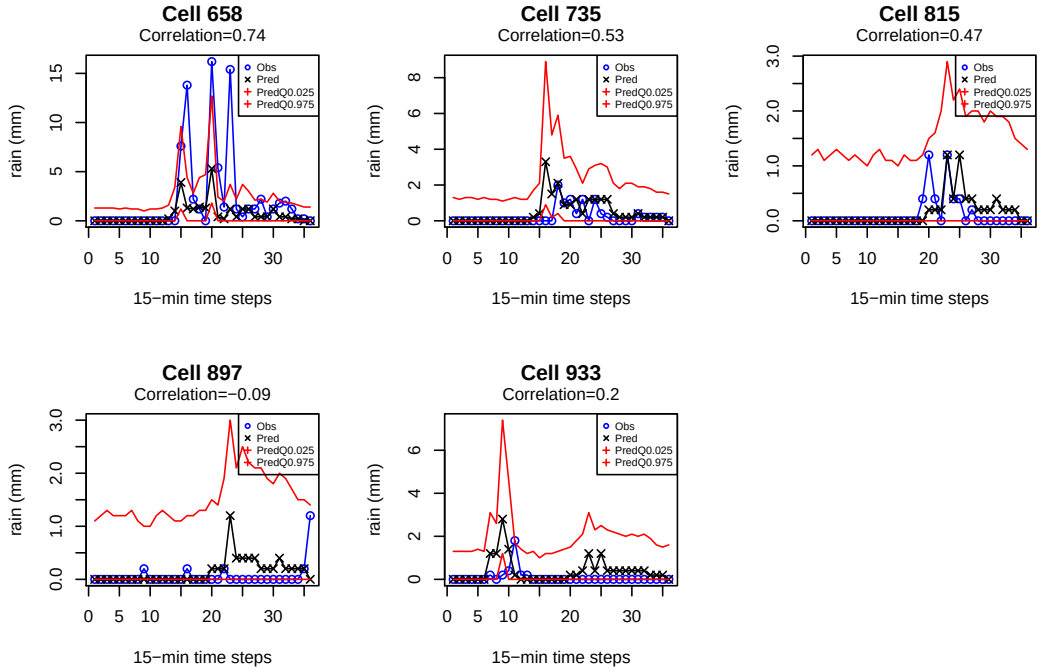


Figure 6.23. Time series of rainfall predicted values at validating cells for 5-Aug-2004-15min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

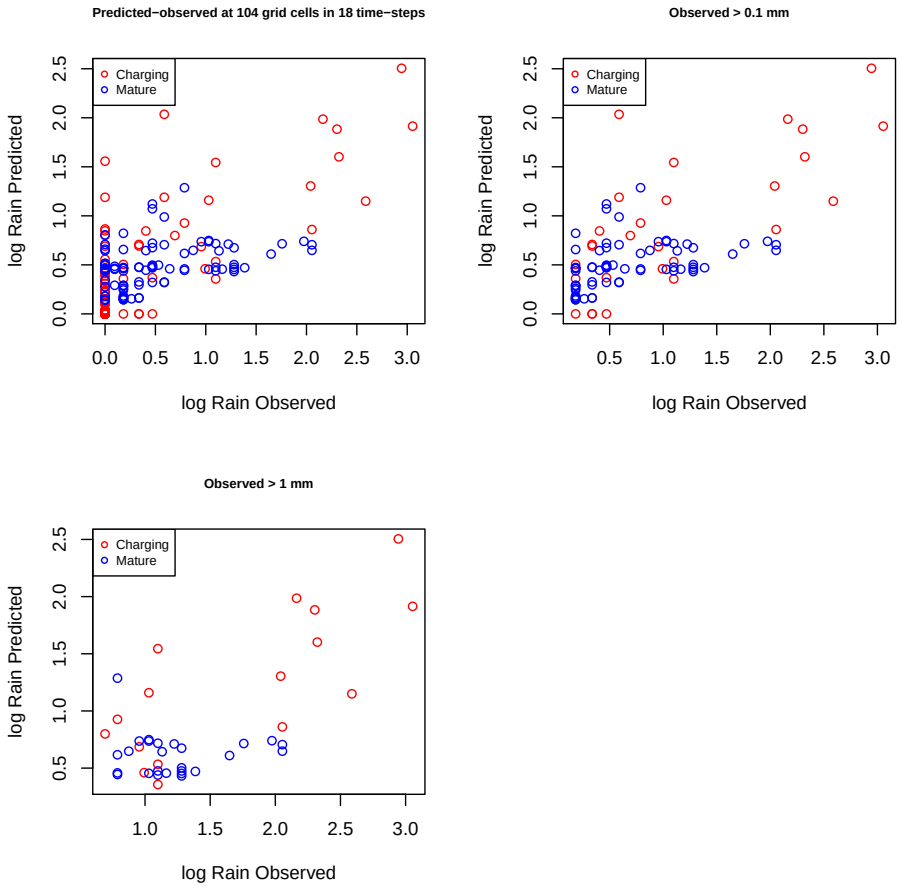


Figure 6.24. Predicted versus observed values at validating cells for 5-Aug-2004-30min-k4 model subdivided into Charging and Mature phases for three levels of precipitation quantity.

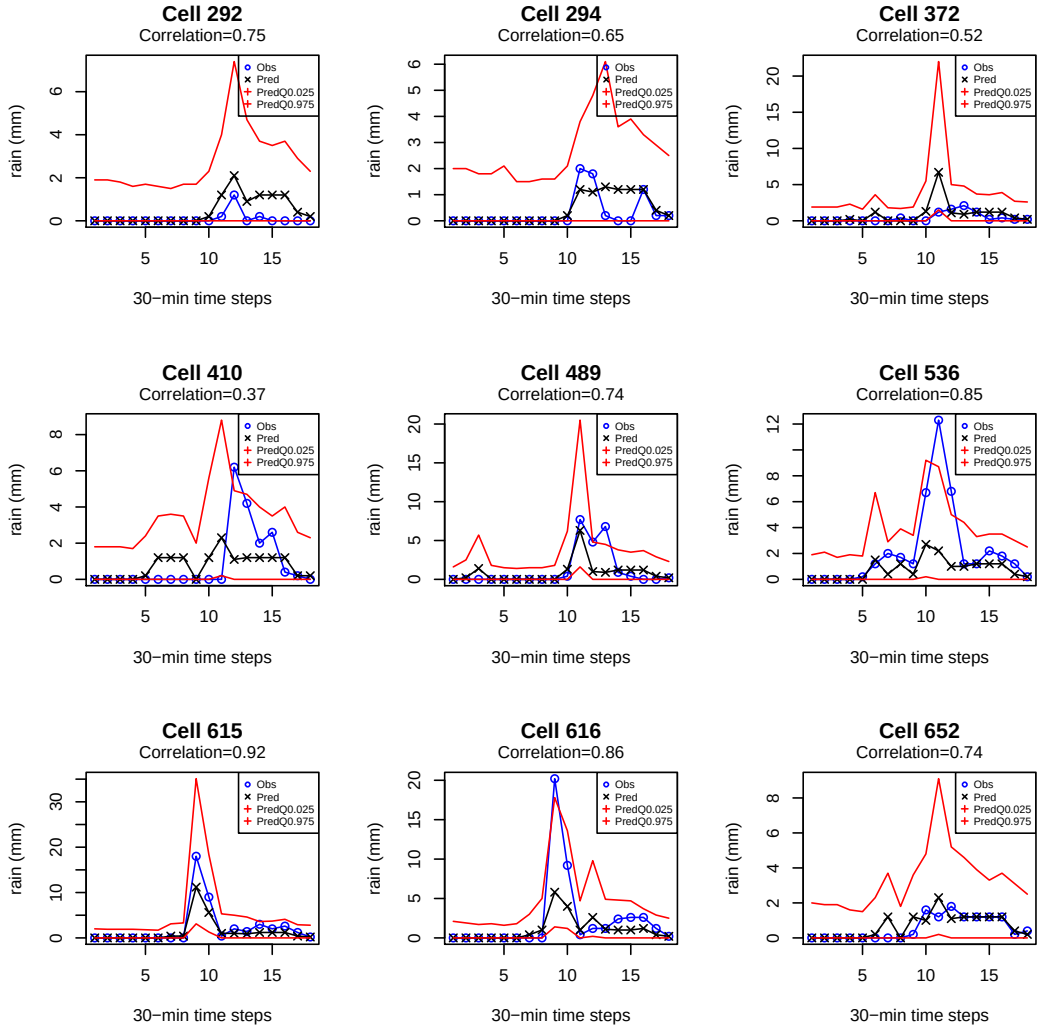


Figure 6.25. Time series of rainfall predicted values at validating cells for 5-Aug-2004-30min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

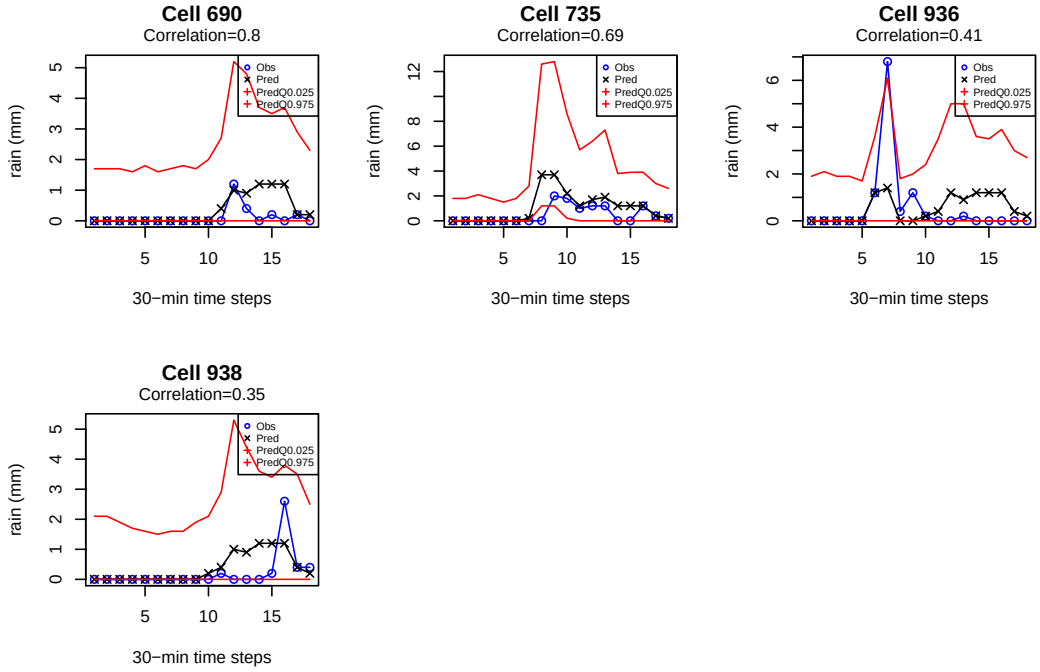


Figure 6.26. Time series of rainfall predicted values at validating cells for 5-Aug-2004-30min-k4 model: predictions (black crosses), observed values (blue rings), 0.025 and 0.975 quantiles (red lines).

Extended tables

Model	Cross correlation	Empirical Coverage (%)
A1	0.46	90.7
A1-k2	0.52	90.9
A1-k2-lag1	0.45	91.1
A2	0.49	91.2
A2-k2	0.49	91
A2-lag1	0.49	91.2
B1	0.48	92.8
B2	0.70	92.4
B2-lag1	0.67	91.1

Table 6.15. Cross correlation and Empirical coverage.

	RMSE (mm)		
Model	global	observed $\geq 0.2\,mm$	observed $\geq 1\,mm$
A1	0.410	0.954	1.679
A1-k2	0.388	0.865	1.380
A1-k2-lag1	0.414	0.962	1.542
A2	0.541	1.019	1.742
A2-k2	0.541	1.016	1.603
A2-lag1	0.543	1.016	1.602
B1	0.541	0.725	1.493
B2	0.524	0.710	1.085
B2-lag1	0.549	0.757	1.134

Table 6.16. RMSE of predicted against observed for 3 classes of rain: rain and no rain (global), positive precipitation (rain$\geq 0.2mm$) and rain$\geq 1mm$.

Bibliography

[Aonashi et al. 2009] Aonashi, K., Awaka, J., Hirose, M., Kozu, T., Kubota, T., Liu, G., Shige, S., Kida, S., Seto, S., Takahashi, N., Takayabu, Y.N.: GSMaP passive, microwave precipitation retrieval algorithm: Algorithm description and validation. J. Meteor. Soc. Japan. **87A**, 119-136 (2009).

[Baddley 2010] Baddeley, A.: Multivariate and Marked Point Processes. In Gelfand A.E., Diggle P.J., Fuentes M., Guttorp P. (eds) Handbook of Spatial Statistics, pp.371-402. Chapman&Hall/CRC, Boca Raton (2010)

[Banerjee et al. 2004] Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, Boca Raton, FL (2004).

[Barnes et al. 2009] Barnes, L.R., Schultz, D.M., Gruntfest E.C., Hayden, M.H., Benight, C.C.: CORRIGENDUM: False Alarm Rate or False Alarm Ratio? Weather and Forecasting, 24,1452-1453. DOI: 10.1175/2009WAF2222300.1, (2009).

[Berliner 1996] Berliner, L.M.: Hierarchical Bayesian time series models. In: Hanson, K., Silver, R. (Eds.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, pp. 15-22, (1996).

[Berrocal et al. 2008] Berrocal, V.J., Raftery, A.E., Gneiting, T.: Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. The Ann of Appl Stat 4:1170-1193. DOI: 10.1214/08-AOAS203, (2008).

[Besag 1974] Besag, J.E.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B, 36, 192âĂŞ236, (1974).

[Bivand and Rundel 2012] Bivand, R., Rundel, C.: rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.2-5 (2012).
http://CRAN.R-project.org/package=rgeos

[Bivand et al. 2013] Bivand, R., Pebesma, E., Gomez-Rubio, V.: Applied spatial data analysis with R, Second edition. Springer, NY (2013).

[Brook 1964] Brook, D.: On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. Biometrika, 51, 481-483, (1964).

[CESI-Sirf] CESI-Sirf: CG-Lightnings database (2010).
http://www.fulmini.it

[Consorzio Lamma] Consorzio Lamma: Regione Toscana and CNR Ibimet Institute of Biometeorology, Weather stations database (2010). http://www.lamma.rete.toscana.it

[Cressie 1993] Cressie, Noel A. C.: Statistics for Spatial Data. Wiley-Interscience,revised edition edition, (1993).

[Diggle et al. 1998] Diggle, P. J., Tawn, J. A., Moyeed, R. A.: Model-based geostatistics. Journal of the Royal Statistical Society. Series C, 47(3), 299âĂŞ350, (1998).

[Di Giuseppe et al. 2013] Di Giuseppe, E., Jona Lasinio, G., Esposito, S., Pasqui, M.: Functional clustering for Italian climate zones identification. Theoretical and Applied Climatology, 114, 1-2, pp. 39-54. DOI: http://dx.doi.org/10.1007/s00704-012-0801-0, (2013).

[Ester et al. 1996] Ester, M., Kriegel, H., Jorg, S., Xiaowei Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 226–231 (1996)

[Fierro et al. 2012] Fierro, A. O., Mansell, E. R., Ziegler, C. L., MacGorman, D. R.: Application of a Lightning Data Assimilation Technique in the WRF-ARW Model at Cloud-Resolving Scales for the Tornado Outbreak of 24 May 2011. Monthly Weather Review, 140(8), 2609-2627 (2012).

[Fuentes et al. 2008] Fuentes M., Reich B., Lee G.: SpatialâĂŞtemporal mesoscale modeling of rainfall intensity using gage and radar data. The Ann of Appl Stat 4:1148âĂŞ1169. DOI : 10.1214/08âĂŞAOAS166, (2008).

[Gelman and Rubin 1992] Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences, Statistical Science, 7, 457-511, (1992).

[Geman and Geman 1984] Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6, 721-741, (1984).

[Goodman and MacGorman 1986] Goodman S.J. and D.R. MacGorman: Cloud-to-ground lightning activity in mesoscale convective complexes. Mon. Wea. Rev., 114, 2320-2328 (1986).

[Grecu et al. 2000] Grecu, M., Anagnostou, E. N., Adler, R. F.: Assessment of the Use of Lightning Information in Satellite Infrared Rainfall Estimation. J. Hydrometeor. **1**, 211-221 (2000).

[Henning, 2010] Hennig, C.: fpc: Flexible procedures for clustering. R package version 2.0-3 (2010) http://CRAN.R-project.org/package=fpc

[Hoff 2009] Hoff, P. D.: A First Course in Bayesian Statistical Methods. (G. Casella S. Fienberg I. Olkin, Ed.) Springer Dordrecht Heidelberg London New York (2009). doi:10.1007/978-0-387-92407-6

[Kubota et al. 2007] Kubota, T., Shige, S., Hashizume, H., Aonashi, K., Takahashi, N., Seto, S., Hirose, M., Takayabu, Y. N., Nakagawa, K., Iwanami, K., Ushio, T., Kachi, M., Okamoto, K.: Global Precipitation Map using Satelliteborne Microwave Radiometers by the GSMaP Project: Production and Validation. IEEE Trans. Geosci. Remote Sens. **45(7)**, 2259-2275 (2007).

[Jona Lasinio et al. 2007] Jona Lasinio, G., Sahu, S.K., Mardia, K.V.: Modeling rainfall data using a Bayesian Kriged-Kalman model, Bayesian Statistics and its application, Anamaya Publisher, New Delhi, India; (2007).

[Lee 2004] Lee Peter, M.: Bayesian Statistics: an introduction. Edward Arnold ed., III edition, London (2004).

[Lee and Zawadzki 2005] Lee, G. W. and Zawadzki, I.: Variability of drop size distributions: Time-scale dependence of the variability and its effects on rain estimation. Journal of applied meteorology, 44(2), 241-255 (2005).

[Levizzani et al. 2010] Levizzani, V., Pinelli, F., Pasqui, M., Melani, S., Laing, A. G., Carbone, R. E.: A 10-year climatology of warm-season cloud patterns over Europe and the Mediterranean from Meteosat IR observations. Atmos. Res., 97(4), 555-576 (2010).

[Marin and Robert 2007] Marin J.M., and Robert C. P.: Bayesian Core: A Practical Approach to Computational Bayesian Statistics. (G. Casella, S. Fienberg, I. Olkin, Eds.) (2007).

[Morales and Anagnostou 2003] Morales, C. A., Anagnostou, E. N.: Extending the capabilities of high-frequency rainfall estimation from geostationary-based satellite infrared via a network of long-range lightning observations. Journal of Hydrometeorology, 4(2), 141-159 (2003).

[Morel and Senesi, 2002] Morel, C. and Senesi, S.: A climatology of mesoscale convective systems over Europe using satellite infrared imagery. 2nd: Characteristics of European mesoscale convective systems. Quarterly Journal of the Royal Meteorological Society, **128**(584), 1973-1995 (2002) doi:10.1256/003590002320603494

[Neyman and Scott 1958] Neyman, J., Scott, E. L.: A statistical approach to problems of cosmology. R. Stat. Soc.B. **20**, 1-43 (1958).

[Okamoto et al. 2005] Okamoto, K., Iguchi, T., Takahashi, N., Iwanami, K. and Ushio, T.: The Global Satellite Mapping of Precipitation (GSMaP) project, 25th IGARSS Proceedings, pp. 3414-3416 (2005).

[Palmeira et al.] Palmeira, F. L. B., Morales, C. A., FranÃğa, G. B., Landau, L.: Rainfall estimation using satellite data for ParaÃŋba do Sul Basin (Brazil). In The XXth ISPRS International Society for Photogrammetry and Remote Sensing Congress (2004).

[Parker and Johnson 2000] Parker, M. D. and Johnson, R. H.: Organizational modes of midlatitude mesoscale convective systems. Monthly weather review, 128(10), 3413-3436 (2000).

[Parker et al. 2001] Parker, M. D., Rutledge, S. A., Johnson, R. H.: Cloud-to-ground lightning in linear mesoscale convective systems. Monthly weather review, 129(5), 1232-1242 (2001).

[Plummer 2003] Plummer, M.: JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20-22, Vienna, Austria. ISSN 1609-395X (2003).

[Rajala 2012] Rajala, T.: spatgraphs: Graphs for spatial point patterns. R package version 2.60 (2012).
http://CRAN.R-project.org/package=spatgraphs

[Rodriguez-Iturbe et al. 1987] Rodriguez-Iturbe, I., Cox, D. R., Isham, V.: Some models for rainfall based on stochastic point processes. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 410(1839), 269-288 (1987).

[Sahu et al. 2005] Sahu, S.K., Jona Lasinio, G., Orasi, A., Mardia, K.V.: A comparison of spatio-temporal Bayesian model for reconstruction of rainfall fields in cloud seeding experiment, Journal of Mathematics and Statistics; 1(4); pp. 272-280, (2005).

[Schmidt and Migon 2009] Schmidt, A.M. and Migon, H.S.: Modelling zero-inflated spatio-temporal processes, Statistical Modelling; 9(1), (2009).

[Soula and Chauzy 2001] Soula, S., Chauzy, S.: Some aspects of the correlation between lightning and rain activities in thunderstorms. Atmospheric Research **56**, Issues 1-4, 355-373, doi:10.1016/S0169-8095(00)00086-7 (2001).

[Tapia et al. 1998] Tapia, A., Smith, J. A., Dixon, M.: Estimation of convective rainfall from lightning observations. J. Appl. Meteor. **37**, 1497–1509 (1998)

[Ushio et al. 2009] Ushio, T., Kubota, T., Shige, S., Okamoto, K., Aonashi, K., Inoue, T., Takahashi, N., Iguchi, T., Kachi, M., Oki, R., Morimoto, T., Kawasaki, Z.: A Kalman filter approach to the Global Satellite Mapping of Precipitation (GSMaP) from combined passive microwave and infrared radiometric data. J. Meteor. Soc. Japan. **87A**, 137-151 (2009)

[Van Delden 1992] Van Delden, A.: The dynamics of meso-scale atmospheric circulations. Amsterdam, North-Holland, 1992.

[Wheater et al. 1999] Wheater, H. S., Isham, V. S., Cox, D. R., Chandler, R. E., Kakou, A., Northrop, P. J., Rodriguez-Iturbe, I.: Spatial-temporal rainfall fields: modelling and statistical aspects. Hydrology and Earth System Sciences, 4(4), 581-601 (1999).

[Yu-Sung and Masanao 2012] Yu-Sung, S., Masanao, Y.: R2jags: A Package for Running jags from R. R package version 0.03-08 (2012) http://CRAN.R-project.org/package=R2jags

Appendix A

Publications

Part I

The work described in Part I has been presented at *11th International Meeting on Statistical Climatology* and *Gfkl-Cladag Joint Meeting 2010* and published in:

`Journals:` **Di Giuseppe E., Jona Lasinio G., Esposito S., Pasqui M., 2013**: Functional clustering for Italian climate zones identification, *Theoretical and Applied Climatology*, 114(1-2),*pp. 39-54*, http://dx.doi.org/10.1007/s00704-012-0801-0, Springer Vienna.

`Proceedings:` **Di Giuseppe E., Jona Lasinio G., Pasqui M., Esposito S., 2010**: Functional clustering of Temperature and Precipitation data for Italian climate zones determination, *Gfkl-Cladag Joint Meeting 2010 Book of Abstract*, 8-10 September 2010, Firenze, Italy, *pp. 151-152.*

`Proceedings:` **Di Giuseppe E., Jona Lasinio G., Esposito S., Pasqui M., 2010**: A functional data approach for climate zones identification, *11th International Meeting on Statistical Climatology Program& Abstracts*, 12-16 July 2010, Edinburgh, Scotland, *pp. 140-141.* URL: http://imsc.seos.uvic.ca/proceedings.shtml.

Part II

The work described in Part II has been presented at *Bayesian Young Statisticians Meeting 2013* and *IX Conference on Geostatistics for Environmental Applications GeoENV2012* and published in:

`Book section:` **Di Giuseppe E., Jona Lasinio G., Pasqui M., Esposito S., 2014**: Predicting rainfall fields from lightnings records: a hierarchical Bayesian approach, in *The Contribution of Young Researchers to Bayesian Statistics*, Volume=63, Springer Proceedings in Mathematics & Statistics, Editor=Lanzarone, E. and Ieva, F., ISBN 978-3-319-02083-9, DOI=10.1007/978 − 3 − 319 − 02084 − $6_1 9$, *pp. 95-99.*

`Proceedings:` **Di Giuseppe E., Jona Lasinio G., Pasqui M., Esposito S., 2012**: Point Processes for modeling lightning data, *IX Conference on Geostatistics for Environmental Applications GeoENV2012*, Valencia, Spain, September 19-21, Editor: J. Jaime Gòmez-Hernàndez ISBN:978-84-8363-924-5, *pp. 87-88*.