Alma Mater Studiorum
Università di Bologna

DOTTORATO DI RICERCA IN
Metodologia Statistica per la Ricerca Scientifica

XXVII Ciclo

Settore Concorsuale di afferenza: 13/D1

Settore Scientifico disciplinare: SECS-S/01

RAINFALL SPATIAL PREDICTIONS: A TWO-PART MODEL AND ITS ASSESSMENT

Presentata da: Elena Scardovi

Coordinatore Dottorato
Prof.ssa Angela Montanari

Relatori
Prof.ssa Daniela Cocchi
Prof.ssa Francesca Bruno
Prof. Tilmann Gneiting

Esame finale anno 2015

Contents

List of Figures

List of Tables

Summary

Rainfall is a phenomenon difficult to predict, due to its strong irregularity in space and time. Knowledge of fallen precipitation is fundamental for water resource planning and risk monitoring. Applications of rainfall prediction span from agriculture to the insurance field; public agencies may benefit from precise quantification of rainfall when decisions about safety have to be taken. Short temporal resolution allows knowledge of the features and dynamics of single events, and creates challenges due to the abundance of zero values. Moreover, right-skeweness of the distribution of positive amounts is a peculiar feature of the phenomenon; the Gaussian assumption is thus not adequate when modelling rainfall accumulation. Chapter 1 provides an introduction to the problem of rainfall spatial prediction and an overview of the existing literature. Rain gauge interpolation methods are presented in Section 1.1. Additional information is often available, for example radar or satellite maps; its exploitation is often crucial. Approaches for radar calibration aiming at correcting indirect measurements on the basis of direct observations are reviewed in Section 1.2. When different data sources are exploited, the spatial supports may not coincide, or they can even have different nature (e.g. point vs grid); the so called "change of support problem" is introduced is Section 1.3.

Chapter 2 presents the motivating problem of the thesis, i.e. the spatial reconstruction of rainfall fields in the Emilia-Romagna Region, in Italy. Section 2.1 illustrates details on data: about 300 rain gauges are available in the area under study, where radar information is also retrieved in the form of grids with 1 km × 1 km resolution. The ARPA-SIMC Emilia-Romagna service collects and preprocesses the data, which is thus available in the form of hourly

accumulated amounts. Some preliminary exploratory analysis on the data are provided. Special attention is devoted to the detection of rainfall occurrence, explaining the inadequateness of a definition of a deterministic threshold; stochastic modelling of rainfall probability emerges as a better choice, with the two-part semicontinuous approach becoming a leading theme of our work. Section 2.2 presents our original contribution as regards modelling, consisting in a three-stage Bayesian hierarchical model aiming at calibrating radar by exploiting rain gauges as reference measure. The relationship between the two instruments is modelled in locations were they are both available. Both when addressing rain probability and rain amounts, a linear relationship in the log scale is assumed, with spatial correlated Gaussian effects capturing the residual information; a probit link is used for addressing rainfall probability. The two steps are joined via a two-part semicontinuous model, which directly specifies the probability of occurrence, and ensures flexibility in dealing with rainfall accumulation, allowing the exploitation of an arbitrary continuous distribution defined on the positive real semiaxis, like the Gamma or the Lognormal. Section 2.2.2 deepens the investigation of the change of support problem, sketching how rain gauge and radar data are matched in our proposal. Three model specifications are presented: Model "base" simply associates rain gauge locations to the radar pixel where they fall; Model "mean" adds the mean over the 8 neighbouring pixels as a further covariate; Model "SW" proposes a stochastic weighting of all radar pixels, driven by a latent Gaussian process defined over the whole grid, aiming at efficiently exploiting the entire radar map and correcting for the potential misalignment between the two instruments. Estimation is performed via Markov chain Monte Carlo procedures; in particular, Gibbs sampling with Metropolis-Hastings steps is implemented in C, linked to R software, in order to guarantee computational efficiency (also through BLAS and LAPACK algebraic libraries) while preserving simplicity in arranging data and displaying results. Details about estimation and prediction are illustrated in Sections 2.3 and 2.4.

The Bayesian approach provides whole posterior predictive distributions. They constitute probabilistic forecasts, which are the most complete and desirable kind of predictions, since they carry full information about the uncertainty originated by the stochastic nature of the model and parameter estimation. For specific applications, single numbers representing the forecaster's best guess are required: in this case, point predictions can be obtained as syntheses of the probabilistic forecasts via the application of suitable functionals like the mean or a quantile; moreover, they can be accompanied by predictive intervals for quantifying uncertainty at a desired level. Chapter 3 defines probabilistic forecasts as opposed to point forecasts, introduces confidence intervals, and extensively reviews literature about the evaluation of predictive performance. In particular, the concept of probabilistic calibration is explained; the Probability Integral Transform (PIT) histogram is the main graphical tool for its assessment, and proper scoring rules are the correct numerical instruments allowing fair comparisons between competing forecasts. Consistent scoring functions are introduced for the evaluation of point forecasts; finally, the concepts of sharpness and coverage are briefly summarized for the purpose of assessing confidence intervals. The mixed discrete-continuous nature of precipitation creates challenges in the evaluation of predictions, since some standard tools are incorrect when applied to two-part semicontinuous models. For example, even in case of ideal forecasts the PIT histogram is not uniform and the coverage of the predictive intervals exceeds the nominal level. Chapter 4 discusses this issue by investigating the application of the tools presented in Chapter 3 to two-part semicontinuous models for probabilistic quantitative precipitation forecasts, and proposes modifications of the standard techniques when necessary. In particular, a non-randomized PIT histogram for dealing with two-part semicontinuous models is suggested; this also provides a straightforward correct computation of interval coverage. Section 4.2 is devoted to the communication of predictions, with a specific focus on precipitation forecasts; some best practices are reviewed regarding the communication to a non-specialized public, based on joint studies of psychologists and statisticians. Chapter 5 shows the results of the different model specifications proposed in Chapter 2 when applied to Emilia-Romagna hourly rainfall data of September-October 2010. A simple model in which spatial effects are assumed independent is introduced as a benchmark. Section 5.1 analyses the estimates for some basic features of the model, like the coefficient for radar measurements in regressions, the weights for radar pixels in Model “SW”, and the spatial processes. Sections 5.2, 5.3 and 5.4 report predictive results on the basis of the evaluation tools suggested in Chapters 3 and 4; probabilistic, point and interval forecasts are considered. Reliability plots show the success in predicting rain probability, while sharpness histograms investigate boldness in detecting rainfall occurrence. PIT histograms address the probabilistic calibration of the model. Performances are also assessed and compared via numerical tools, with synthetic tables providing results obtained on 8 rainfall events; both event-specific scores and global results summarising the behaviour on the whole analysed period are presented. In particular, the Brier Score is employed for the assessment of rain probability, and the Continuous Rank Probability Score (CRPS) for probabilistic forecasts. The Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) provide evaluation of point forecasts for the predictive mean and median, respectively. The Brier Score Plot and the Quantile Decomposition Plot help in detecting differences between model specifications according to the threshold or quantile level considered. Uncertainty information is provided via 90% and 50% confidence intervals.

Conclusive remarks are summarized in Chapter 6, together with some possible future developments.

Chapter 1

Rainfall spatial prediction

Rainfall measurements are essential for public authorities, being the basis for hydrological models and risk monitoring: knowledge of precipitation fields can be useful for water resource planning and management, and might enable public agencies to alert citizens when extreme events occur. Rainfall data are exploited both as an input for hydrological models and for validation of weather forecasts. Also, the detection of exceedances over a certain threshold is critical in the insurance field, when assigning responsibilities for damages occurred during a storm.

According to specific purposes, different precipitation accumulation times can be of interest: yearly or monthly rainfall amounts are used for climate research (see for example Adler *et al.* 2003), daily and hourly measurements are the starting point for flood monitoring (Cooley *et al.* 2007) or agricultural planning (Stern and Coe 1984), while hourly and instantaneous data allow to study single rain events.

Direct measurements are provided by rain gauges, which are instruments collecting rainfall in sparsely distributed locations at ground. Rain amounts are read either manually or by automatic weather stations. Several types of rain gauges exist, with tipping bucket ones being the most common: they consist in a funnel that collects and channels precipitation into a small seesaw-like container. After a pre-set amount of precipitation has fallen, the lever tips, dumping the collected water and sending an electrical signal. “Instantaneous” rain gauges allow the detection of rainfall rate; otherwise, rainfall quantities are accumulated over a certain time interval. Limitations in rain gauges functionality are well known in case of violent convective events, due to the possibility of rain drops to bounce off the gauge, or during floods, since excessive amounts may not be collected; moreover, strong wind can affect the angle of incidence toward the ground and cause dispersion of part of rainfall, and non-heated rain gauges are not able to work properly in case of snowy precipitation. Nevertheless, apart from rare circumstances, rain gauges provide reliable measurements, which can be treated as ground truth. Automatic and manual monitoring systems check the operation of the network and invalidate malfunctioning or blocked rain gauges.

1.1 Rain gauge interpolation

The sparse nature of the rain gauge network only allows to collect information in locations where a rain gauge resides. The need for a more complete knowledge of rainfall over a region has urged meteorologists and statisticians to look for interpolating methods able to fill the missing spatial information. Rainfall spatial prediction has many challenges; strong spatial and temporal heterogeneity in fact characterizes rainfall events. Different possible physical processes can generate the weather front, giving birth to stratiform, convective or mixed events. Abrupt changes are common, leading to strong irregularities, most of all when dealing with short accumulation times like one hour. With such temporal granularity, the probability of finding dry locations is non-negligible, thus making the presence of a relevant amount of zero values a main feature of the phenomenon. Finally, a symmetric modelling of positive rainfall amounts is not suitable, their distribution being highly right-skewed. Deterministic or stochastic approaches can be adopted for interpolating rain gauge measurements, with kriging being a widespread and powerful choice (Matheron 1963, Cressie 1990). Rainfall distribution is known to be asymmetric and skewed, which is in contrast with the assumption of normality; Erdin *et al.* (2012) propose a Box-Cox transformation, which does not completely solve the problem. Moreover, basic kriging approaches do not directly model the probability of rain, thus often resulting inadequate when dealing with relevant amounts of zero measurements; variogram estimation is heavily affected by this feature, and spurious rainfall predictions are produced. Several methods have been proposed for handling the large number of zero measurements. Goovaerts (1997) constructs threshold-exceedance probability maps using indicator kriging. Yoo and Ha (2007) propose bivariate mixed distributions, Kim and Ahn (2009) exploit neural network classifiers for rainfall occurrence, Li *et al.* (2010) adopt nonlinear Markov Chain random fields incorporating interclass dependences through transiograms. In the framework of ensemble forecast calibration, Sloughter *et al.* (2007) model the rainfall distribution at a specific site as a two-part semicontinuous model with Gamma distribution for positive amounts; successively, Berrocal *et al.* (2008) improve the model by including two spatial Gaussian processes driving precipitation occurrence and accumulation, respectively.

1.2 Radar calibration

Rainfall can also be detected and measured by other instruments like satellites and radar, which provide continuous spatial information in the form of gridded maps, often at a high resolution; such technologies produce indirect estimates of rainfall, deriving from the measurement and transformation of variables that are correlated with the rainfall rate. For example, radar localizes rainfall by exploiting the reflection of its beams when they encounter precipitation particles, and quantifies precipitation intensity by measuring reflectivity. Knowledge of the orography of the terrain is fundamental for determining the minimum vertical angle for radar beams; moreover, corrections for beam blocking caused by mountains are often performed, together with other deterministic procedures aiming at reducing biases caused by the morphology of the region or by the meteorological conditions, like clutter, anomalous propagation and the presence of bright bands. After this preprocessing of radar data, the reflectivity Z is converted into the rainfall rate R. The well known Z-R conversion formula (Marshall and Palmer 1948) establishes an exponential deterministic relationship between the two quantities:

$$Z = a\,R^b, \quad a = 200,\ b = 1.6. \tag{1.1}$$

Parameters a and b are usually kept fixed, but studies reveal the values in (1.1) change slightly according to the meteorological condition. A large number of experimental corrections to this relationship have been proposed, for example by Marshall and Palmer (1948), Gunn and Marshall (1958) and Battan (1973), among many others; in recent years, the National Oceanic and Atmospheric Administration (NOAA) in the United States changed these parameter values to $a = 300$ and $b = 1.4$ after the installation of new radars. Thanks to the Z-R conversion, from each radar scan indirect instantaneous rainfall rates can be obtained. Accumulation in time finally provides rainfall amounts (mm); for this purpose, the Emilia-Romagna environmental agency ARPA-SIMC uses pattern recognition tools for comparing successive radar maps in order to detect cloud movement and assess rainfall transportation, obtaining hourly accumulated rainfall. Hourly radar data can be made available on fine-pixel grids, thus overcoming the problem of sparseness of the rain gauge network. However, such data consist of indirect measurements and are affected by stochastic biases which are not removed by deterministic preprocessing; for this reason, they are not reliable for the assessment of rain amounts, and require calibration. Similarly, satellites provide rainfall maps in the form of images; they are less direct then radar measurements but have the advantage of complete coverage over oceans, mountainous regions, and sparsely populated areas where other sources of rainfall data are not available.

Merging rainfall data provided by different instruments is a topic that received considerable attention in the hydrological and statistical literature. In particular, we focus on the problem of radar calibration on the basis of rain gauge observations: the aim is to correct radar maps for providing reliable rainfall predictions in locations where rain gauges are not available. Mean or local adjustment factors can help in removing the radar bias (see for example Koistinen and Puhakka 1981; Amorati *et al.* 2012). Several model-based methodologies were proposed, adapting kriging and co-kriging techniques to radar-rainfall calibration. Among others, Seo and Smith (1991a, b) integrate kriging in a Bayesian framework incorporating a priori information on past rainfall observations, for improving classical lognormal co-kriging and nonparametric methods. A Bayesian approach is also adopted in Pilz and

Spöck (2008), using trans-Gaussian kriging. Kriging with External Drift (KED, Wackernagel 2003) allows to incorporate radar measurements in the large scale component of a spatial model. Goudenhoofdt and Delobbe (2009) show that this procedure outperforms common deterministic procedures for gauge-radar combination, like mean or local bias correction, range dependent and Brandes spatial adjustment, and Sinclair and Pegram's conditional merging (2005); a comparison with two bias correction techniques (Amorati *et al.* 2013) confirms KED's superiority when working on Emilia-Romagna hourly rainfall data. Parametric estimation of the variogram, with a preference for the exponential form in many rainfall applications (see for example Leung and Law 2002, Pathac and Vieux 2007), is a widespread approach. A non-parametric technique based on Bochner's theorem and Fast Fourier Transform was proposed by Yao and Journel (1998) and further developed for automatically defining a valid correlogram from radar maps without the need for the specification of an analytical expression (Velasco-Forero *et al.* 2009, Schiemann *et al.* 2011). Applications of KED to rainfall field reconstruction in Italy are illustrated in Orasi *et al.* (2009), where the method is applied to a cloud seeding experiment in the Puglia region, and in Scardovi *et al.* (2012 b), where a rainfall event in the Emilia-Romagna region is analysed. In both cases, KED provides smaller prediction errors with respect to kriging performed only on rain gauge data. Despite the ability of external information in driving the prediction, probabilistic performances of this method are affected by the inadequacy of the normality assumption and by the presence of many zero values, as anticipated in Section 1.1. Transformation approaches have been proposed for mitigating the former problem, while two-steps procedures can address the mixed discrete-continuous nature of rainfall, predicting the presence or absence of rainfall in a first stage, and rain accmulation in a second one, conditioning on rain occurrence. In both cases, care is required when assessing uncertainty in a multi-step procedure. This is particularly awkward when indicator kriging, providing spatial predictions of rain probability, is joined with a further kriging step for the prediction of positive rainfall amounts. Plain single stage KED, without directly addressing the abundance of zero measurements, outperforms multi stage approaches which need an arbitrary threshold for the definition of rainfall presence. Fully Bayesian approaches should be adopted for keeping track of all modelling and estimating steps and correctly combining the resulting uncertainties.

Methods alternative to kriging for exploiting radar information when dealing with rainfall modelling and prediction have been proposed in the literature. Brown *et al.* (2001) introduce a pioneering calibration approach building a multivariate state-space time series model for modelling reflectivity radar against gauge measurements. In a similar framework, Costa and Alpuim (2011) provide a thorough study based on state-space models and the Kalman filter; Sahu *et al.* (2005) adopt Kalman filtering in a Bayesian framework. Fuentes *et al.* (2008) follow a different approach, developing a spatio-temporal model where a latent process, corresponding to the true rain amount, drives the probability of precipitation occurrence and the rainfall accumulation; both radar and rain gauge data are modelled according to the common latent process.

The problem of radar calibration falls within the wider framework of multi-source combination. In particular, the availability of numerous data sources arises the necessity for a comprehensive approach able to account for differences in the spatial supports on which measurements are defined. In the statistical literature, this issue goes under the name of “change of support problem” (COSP); the next section provides an introduction to the problem and an overview of the available techniques.

1.3 Merging data sources

Direct observations usually consist of point data, sparsely distributed in space according to the discrete structure of the monitoring network. In many fields, there is a growing interest in supplementing such measurements with additional information, deriving for example from numerical models, radars and satellites, to increase the availability of data across space and time. Such information is organized on grids; nevertheless, the richness of continuous maps, which can span large spatial domains with no missing values, is compensated by their need for calibration and for an assessment of uncertainty. In fact, deterministic models mathematically approximate the underlying physical processes, while instruments working on the basis of scans or images provide indirect measurements of the phenomenon; discretisation of space and time adds further approximations. Such data have been derived with a deterministic approach; therefore, they do not convey any information about their inherent uncertainty. Calibration and the assessment of uncertainty are possible by combining different data sources; nevertheless, errors can be caused by misalignment and by inconsistences in the nature of the supports. The increasing availability of data sources induced the search for data assimilation tools. Kalnay (2003) reviews methods for combining observational data on the current state of the atmosphere with a short-range forecast. Algorithmic and ad hoc methods are usually employed in atmospheric data assimilation. The so called "change of support" problem was originated from the need for reconciling differences in spatial resolution under a statistical approach (see, e.g., Cressie 1993; Gotway and Young 2002; Banerjee, Carlin, and Gelfand 2004). Downscaling or upscaling approaches can be followed: the former combines the sources of information for obtaining improved predictions at point level, while the latter models station data in order to provide predictions at point-level and on grid.

Block kriging (Cressie 1993; Chilès and Delfiner 1999; Banerjee, Carlin and Gelfand 2004) predicts the average value of a process over a block exploiting point observations. Carroll *et al.* (1995) link block and ordinary kriging to develop a geostatistical method combining ground-based observations with areal block measurements; Gotway and Young (2007) extend block kriging and develop a flexible geostatistical method able to handle several change of support problems at the same time, allowing both upscaling and downscaling.

Fully model-based solutions to the change of support problem have been proposed in the literature. Wikle and Berliner (2005) develop a Bayesian hierarchical model that allows combination of data observed at different spatial scales. The main assumption underlying their model is the existence of a true unobserved process related with observations via a measurement error model. Such underlying process is equipped with a spatial correlation structure and specified at a spatial scale that is different from the one characterizing observations. In a similar fashion, Fuentes and Raftery (2005) present a Bayesian model combining point-referenced air pollution observations with block average predictions obtained by an air-quality model. The model consists in an application of the Bayesian melding method developed by Poole and Raftery (2000), and continues the work of Cowles *et al.* (2002) and Cowles and Zimmerman (2003), who use systematic sampling and numerical integration techniques to combine point and areal data. As in Wikle and Berliner (2005), Fuentes and Raftery (2005) assume the existence of an underlying unobserved spatial process driving both observational data and numerical model output. However, instead of modelling the true process at areal unit scale, Fuentes and Raftery (2005) specify it at point level. A measurement error model links the unobserved process to observations; the relation with computer model data is linear, accounting for potential bias in the model output. Since the computer model output is specified in terms of block averages, the linear model is expressed in terms of stochastic integrals. The Bayesian melding model of Fuentes and Raftery (2005) has gained considerable attention and has already been used in several applications (see for example Smith and Cowles 2007; Foley and Fuentes 2008). However, it is computationally intensive, due to the large number of stochastic integrals needed to account for the abundance of grid cells. As in Fuentes and Raftery (2005), McMillan *et al.* (2009) propose a spatio-temporal fusion model postulating the existence of a true spatial process related to both observational data and numerical model output. However, they specify the true process at block rather than at point level; in this way, upscaling is addressed instead of downscaling, and the computational burden of Bayesian melding is avoided, thus allowing spatio-temporal applications.

Relevant contributions focusing on spatial aspects have been provided in the literature. Guillas *et al.* (2008) and Liu, Le, and Zidek (2008) use a two-stage regression with spatial interpolation of the coefficients of the linear regression. Berrocal *et al.* (2010a, 2010b) propose univariate and bivariate hierarchical downscaler models. They take the numerical model output as data and relate observations and numerical model output via a linear regression with spatially-varying coefficients (Gelfand *et al.* 2003). These are, in turn, modelled as correlated spatial Gaussian processes exploiting coregional-

ization (Schmidt and Gelfand 2003; Gelfand *et al.* 2004). These models offer the advantage of local calibration of the numerical model output without incurring problems due to the dimensionality of the computer model output, being only fitted at the numerical model grid cells where the monitoring stations reside; moreover, they allow straightforward prediction at point level, thus offering a fully model-based solution to the problem of downscaling. Statistical downscaling has been successfully applied to air quality simulations (Berrocal *et al.*, 2010a, b; Zhou *et al.*, 2011), climate model output (Berrocal *et al.*, 2012; Zhou *et al.*, 2012), and remotely-sensed satellite images (Liu *et al.*, 2009; Kloog *et al.*, 2011). Reich *et al.* (2014) develop a multiscale statistical downscaler for dealing with different spatial resolutions, utilizing the spectral representation of spatial processes. Fassó and Finazzi (2013) develop a space-time multivariate data fusion model addressing ground level point observations and remote sensing pixel data over Europe, handling missing information without the need for data imputation. Covariates and latent variables acting as space-time varying coefficients for the covariates allow the adjustment of the model to local conditions; D-STEM (Distributed Space Time Expectation Maximization) software provides a parallel and distributed implementation of the EM algorithm for model estimation.
Neighbour-based extensions of the downscalers were firstly provided by Berrocal *et al.* (2011), with the inclusion of information belonging to grid cells near the one where the location lies, thus directly addressing the potential problem of misalignment between stations and putatively associated grid cells. An adaptive smoothing of the computer model output is provided, allowing for stronger association with the observed station data and resulting in improved spatial interpolation. For this purpose, a Gaussian Markov Random Field (GMRF) is employed to smooth the computer model. An alternative consists in introducing spatially varying weights driven by a latent Gaussian process to accomplish smoothing.

This thesis proposes a Bayesian Hierarchical three-stage model for predicting the probability and amount of rain exploiting radar information, with the aim of calibrating radar and reconstructing the whole rainfall field. Several variants are tested on hourly rainfall data of the Emilia-Romagna region provided by the agency for the environment ARPA-SIMC. A detailed description of the data and methods is performed in Chapter 2. In the first formulation of the model, each rain gauge site is associated with the radar grid cell where it is located. For addressing COSP, most approaches introduced in the present Section can not be straightforwardly applied since they rely on the Gaussian assumption; we develop a model able to address this problem with non-Gaussian data. Several model specifications are developed and tested, with increasing accuracy in handling the potential misalignment between rain gauges and radar.

Chapter 2

Motivating problem: rainfall field reconstruction in the Emilia-Romagna Region

The motivating example of the work is the reconstruction of hourly rainfall fields in the Emilia-Romagna Region in Italy. Data have been provided by the environmental agency ARPA Emilia-Romagna, SIMC division (Servizio Idro-Meteo-Clima).

2.1 The data

ARPA environmental agency has hourly rain gauge data available on a very dense monitoring network. Data quality control is performed by ARPA, eliminating malfunctioning or blocked rain gauges from the network until maintenance. Moreover, ARPA produces radar data exploiting two polarimetric doppler C-band radars. We focus on the radar circle with 125 km radius centred in San Pietro Capofiume, near Bologna; this area is equipped with 317 rain gauges. A complex pre-processing is performed by ARPA aiming at removing errors characterizing radar based measurements. More precisely, hourly accumulated amounts are obtained through the following procedure: reflectivity radar measurements, recorded every 15 min on a 1 × 1 km grid cell resolution (including about 49,000 pixels) are corrected for systematic and occasional biases due to the morphology of the region and atmospheric conditions (see for example Fornasiero *et al.* 2004). Then, reflectivity data are transformed into rain rates by means of the Z-R conversion formula (1.1) proposed by Marshall and Palmer. Finally, in order to accumulate data over an hour, pattern recognition tools are exploited for comparing successive maps and finding pointwise vectors describing the movement of rain masses; this information is used to integrate precipitation in time (advection, Hannesen and Gysi 2002). The procedure sketched above makes radar and gauge data comparable and allows to model a calibration equation under a consistent physical framework. The accurate preprocessing of ARPA on radar measurements provides high quality data which, however, are still far from being reliable for assessing rainfall accumulation: they are affected by temporally and spatially varying bias. To this regard, a thorough exploratory analysis has been performed at the beginning of the work; early results are reported in Scardovi *et al.* (2012). The main achievements of the analysis are briefly summarized. The discordance between the two instruments in detecting rainfall presence was a central issue. While the rain gauges threshold for discriminating between rainfall and moisture or dew is commonly fixed at the rain gauge precision, i.e. 0.2 mm (meaning that only measurements strictly higher then 0.2 mm are considered as rain), it is not straightforward to determine radar precision, nor to establish a threshold which helps in assessing where rainfall occurs when only radar data are examined. The use of the same thresholds for the two instruments shows disagreement in 6% of the cases, most of which attributable to radar measuring rain in dry locations. A set of alternative thresholds was thus tried for radar, spanning from 0.2 to 1.2 mm, and comparisons were made on the basis of the agreement between the two instruments, via the skill scores that are common in meteorology. We denoted with Y (Yes) the detection of rain and with N (No) its opposite, i.e. measurement less or equal to the threshold; then we built a 2×2 contingency table reporting the detection of rainy or non rainy measurement from the two instruments, and indicated the four possible entries of the contingency table on n observations as couples of Y/N with rain gauge result as first entry and radar as second one. The following scores were computed:

- H = hit rate = (YY+NN)/n = relative frequency of coherent detections; it lies between 0 and 1, with better results when it is near 1

- TS = threat score = YY/(YY+NY+YN) = probability of coherence in detecting rain calculated excluding the NN case, which characterizes the majority of the cases; it lies between 0 and 1, with better results when it is near 1

- POD = probability of detection = YY/(YY+YN) = probability of concordance when rain gauge detects rain; it lies between 0 and 1, with better results when it is near 1
- FAR = false alarm rate = NY/(YY+NY) = probability of observing a dry hour when radar detects rain; it lies between 0 and 1, with better results when it is near 0
- BIAS = (YY+NY)/(YY+YN) = number of cases in which radar detects rain divided by the number of cases in which the rain gauge detects rain; it is not limited and denotes a better skill when near 1.

Results revealed that the disagreement between the two instruments is minimized when the threshold for radar is set near 0.9 mm. This result encouraged further investigations of the disagreement. The inspection of the geographical distribution of the indices detected the presence of an area in the Po river delta where FAR was particularly high. In that region, several secondary trips were found: they consist in reflectivity spots which are incorrectly localized. More precisely, the frequency with which radar beams are sent is equal to the time the beam requires for reaching the boundary of the selected circle (125 km of radius) and returning to the center. When rain masses are found out of the boundary, the beam is reflected but returns only after another beam has been sent; so, its detection is misunderstood, being considered as the result of the lastly sent beam; this causes the localization within the circle of storms which are actually out of it. Secondary trips can be detected by comparing the short radius map (125 km radius) with the medium radius one (250 km radius): if rain masses are seen in the small circle and not in the medium one in the same position, but can be detected in the anulus between 125 and 250 km from the center, then they are secondary trips and must be removed. On the basis of this reasoning, we implemented an automatic algorithm of correction. It checks each rainy pixel P_S in the short radius grid, looking for the pixel P_M in the medium radius grid corresponding to the same location: if neither P_M nor the 8 neighbours detect rain, then P_S is suspicious. Since short radius data are the most used, ARPA-SIMC meteorological service performs a number of further corrections with respect to the larger one; thus, in principle, P_S might be the result of a beam blocking correction and must not be corrected. In order to check whether P_S has to be removed, the position of the rain mass in the anulus between 125 and 250 km that should have generated the secondary trip is computed; if one of the 25 pixels around that location detects rain, P_S is considered as containing a secondary trip and removed. The amplitude of the neighbourhoods which are analysed for the comparisons between small and medium circle (9 pixels in the first phase, 25 in the second one) has been chosen after several trials; a simple pixel-to-pixel comparison would not be adequate, also due to a slight temporal misalignment between short and medius radius radar scans. The simple procedure just summarized noticeably corrects short radius maps (see Figure 2.1), and convinced ARPA to implement an operative algorithm. The analysis of the 2×2 tables regarding the agreement/disagreement of the corrected radar and rain gauges about rainfall detection confirmed the reduction of the problem of anomalous FAR values in the Po river delta, which were caused by frequent storms in Istria. As a consequence the optimal threshold value for rainfall definition according to radar can be decreased to 0.8 mm. After corrections, however, the preliminary analysis revealed a complex discordance between the two instruments, consisting in overestimation of rainfall presence detected by radar, which suggested to abandon the usual deterministic approach for establishing rain occurrence. Relying on rain gauge detections and modeling rainfall probability appeared a more promising solution than establishing empirical exogenous thresholds.

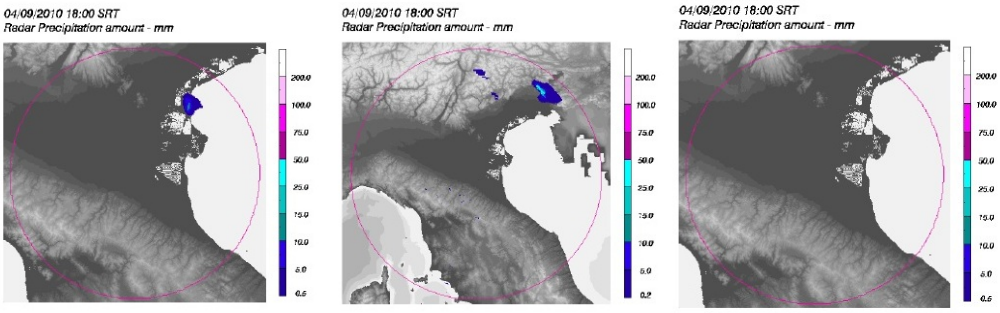


Figure 2.1: Correction of secondary trips at 6 p.m., September 4^{th}, 2010: short radius circle before the correction in the left hand panel; medium radius radar circle before the correction in the central panel, with the rain mass causing the secondary trip out of the short circle; short radius circle after the correction in the left hand panel.

The period under study is September-October 2010. In particular, the main 8 rainfall events of such period were chosen, denoted in Table 2.1 as E1-E8. The table reports the duration in hours of each rain event, the percentage of zero values collected by rain gauges, and some quantiles of the distribution of rain accumulation measured by rain gauges when rain occurred. In the last column, the value of the linear correlation between rain gauges and radar measurements is reported. Rain events are characterized by different durations (ranging from 5 to 16 hours) and variable rainfall intensity, both in terms of average and maximum amounts. Our rain gauges have 0.2 mm precision. Several papers take explicit account of the discreteness of rain gauge measurements in model building. As an example, Sahu *et al.* (2005) consider this feature in a case study on a dry region characterized by low rainfall amounts and short accumulation times (10 min). In contrast, in each of the 8 rain events analysed here, a considerable amount of rainfall is observed and the accumulation is made with hourly resolution, making the effect of discretisation less relevant. Observed zero values are not due to censoring but correspond to a reliable no-rain detection; their amount ranges from 24% to 67% of the observations, suggesting that modelling rain accumulation needs to be coupled with appropriate modelling of rain occurrence. The correlation between radar and rain gauge data, reported in Table 2.1 as

Event	H.rs	Zero %	Q1	Med	Q3	Max	Corr
E1	6	24	0.4	1.0	2.4	19.4	0.81
E2	9	49	0.6	2.0	4.4	31.2	0.80
E3	10	30	0.4	0.8	2.2	34.4	0.55
E4	6	49	0.6	1.8	3.8	27.0	0.59
E5	7	53	0.2	0.8	3.4	37.2	0.76
E6	5	61	0.2	0.6	1.8	12.4	0.79
E7	11	67	0.2	0.8	2.0	15.0	0.46
E8	16	43	0.4	0.8	1.6	10.8	0.43

Table 2.1: Descriptive statistics of 8 rainfall events in September-October 2010: Event ID (Event); number of hours characterizing the Event (H.rs); percentage of zero measurements (Zero %), first (Q1), second (Med) and third (Q3) quartiles of positive amounts, in mm; maximum (Max), in mm; correlation between rain gauge and radar measurements (Corr).

a proxy of the quality of radar data, shows variable but generally high values. The fact that the lowest values occur in the longest Events (E7 and E8) is not completely surprising, since these events are characterized by many near zero values that radar does not accurately detect. confirming the

When building models for rainfall field reconstruction starting from radar and rain gauge data, a basic concern is to understand if, for each event, the relationship between the two measurements can be kept constant or varies along time. Figure 2.2 reports hourly scatterplots of rain gauge against radar data for Events E1 and E4 as an explorative tool in this direction. Event E1 shows higher dispersion in the radar-rain gauge relationship with respect to Event E4. Moreover, the slopes of OLS regression lines show variability along time, both between and within rainfall events. This suggests that the proposal of time-specific models for the radar-rain gauge relationship is appropriate; previous work (Bruno *et al.* 2014) confirmes the superiority of hour specific parameterisation over a common specification on an event, after comparing model fit and predictive performances obtained in the two cases. As an example of data spatial representation, Figure 2.3 shows radar and

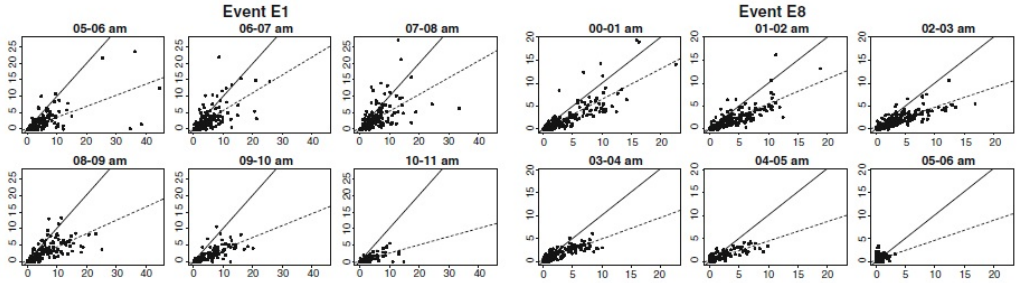


Figure 2.2: Scatter plots of rain gauge (y axis) versus radar data (x axis) in rain Events E1 (left panel) and E4 (right panel); the solid line is the bisector, the dashed line is the OLS regression line.

rain gauge measurements for the first hour of Event E4. The left panel shows the bubble plot of rain gauge data: at each rain gauge site, the size and color of the points are related to the amount of rainfall accumulation. The right panel displays the observed radar map, with rain gauge locations marked with black dots. Since predictive performances of the calibration procedures need to be assessed, 50 randomly selected rain gauge sites (red squared marks in the right-hand panel of Figure 2.3) will be excluded from model estimation and used for validation. The eastern part of the radar circle covers the Adriatic Sea, where no rain gauges are available: in this area, prediction relies solely on radar data. The spatial overview provided by these maps confirms the high concordance between the two instruments in identifying and localizing rain masses.

2.2 A two-part three-stage Bayesian model

A strategy for radar-rainfall calibration in a Bayesian hierarchical framework is proposed, relying on a two-part semicontinuous model in order to properly handle zero and positive values. The main aim is not modelling the rainfall process, neither time forecasting, rather, model construction is mainly focused on the spatial features of the relationship between radar information and rainfall probability and amounts in sites where both radar and

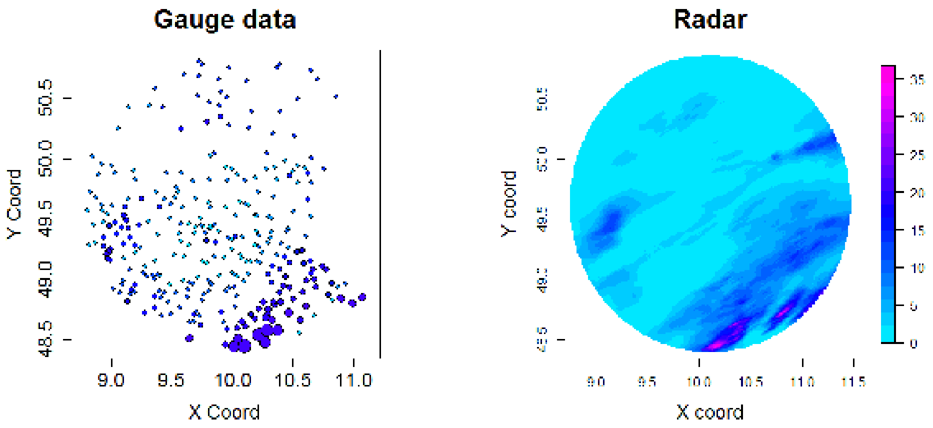


Figure 2.3: Rain gauge data (left panel) and radar data (right panel) for the first hour of Event E4. In the left panel, point size and color are related to values of rain accumulation. In the right panel, rain gauge sites are identified by black dots, sites left aside for validation are marked with a red square.

rain gauge measurements are available, in order to reconstruct rainfall fields starting from an observed radar map. Several alternatives are investigated. We now introduce the main concepts; model structure is described in detail in Section 2.2.1, and details about how to address the change of support problem are presented in Section 2.2.2.

The classical problem of calibration (see for example Brown 1994) involves measurements of the same quantity, along time and space, simultaneously obtained by a reference and an equivalent measuring instrument. In our calibration proposal, we face the problem of rainfall field reconstruction merging radar and rain gauge data, in the spirit of Brown *et al.* (2001): rain gauges represent our reference measures, whereas radar-based rainfall estimates are treated as measures produced by the equivalent, uncalibrated instrument, and are calibrated on the basis of rain gauge data, in order to provide accurate spatial predictions of hourly rainfall.

We denote with S_G the set of rain gauges locations, $\#S_G = 317$, and with S_R the set of all the available pixels in the radar grid, $N_R = \#S_R = 48047$; the radar circle covers the whole area in which rain gauges lie. As already mentioned in Section 1.2, joint modelling of radar and rain gauge data, directly including all radar data in the model, is proposed in Fuentes *et al.* (2008), where the elicitation of the spatial structure for the joint process requires the specification of a $N_R \times N_R$ -dimensional covariance matrix, considering all available information in the model. We propose a different, less complex approach, that focuses only on sites $s \in S_G$ in order to learn about the relationship between rain gauges and radar measurements. Successively, spatial prediction of rainfall is performed starting from radar data for any $s \in S_R \setminus S_G$. The high density of the monitoring network, the smoothness of the radar surface (see Fig. 2.3 as an example) and the major interest in a calibration procedure that allows to efficiently transform radar measurements into rain amounts, motivate our choice to build the model by including only data measured at sites $s \in S_G$. Following this approach, the model turns out to be characterized by a manageable computational complexity, preserving a satisfactory efficiency of the predictions. Thus, we focus on conditional modelling of rain gauge data (reference measure) Y on radar data R (uncalibrated measure), i.e. on the distribution

$$p(Y_s|\{R_P,\ P \in S_R\}),\ s \in S_G \tag{2.1}$$

where subscripts indicate the location. Expression (2.1) allows rainfall in a certain location to depend on the whole radar map. In fact, several approaches are proposed in this thesis: a punctual one, in which Y_s only depends on the radar value in the grid cell containing location s, for each $s \in S_G$; a simple improvement, taking also the nearest neighbouring cells into account; and a more sophisticated one, where Y_s depends on all R_P (with only P in a neighbourhood of s being relevant, properly weighted). Details about these model specifications are provided in Section 2.2.2. For the moment, radar information is generically denoted with R; it will be further specified according to the chosen downscaler.

Notice that expression (2.1) does not include temporal subscripts. The present work does not address modelling of the temporal evolution of the relationship between rain gauges and radar measurements; we work on each hour separately. Several reasons have driven this choice. Early checks, reported in Bruno *et al.* (2014), showed that the joint modelling of successive hours worsens model performance: the flexibility gained by separately modelling each hour turned out to be more decisive than the common features shared within the same rainfall event. Moreover, model estimation is easier and faster if performed on single hours, and permits a real time processing of data hour by hour. In the following we thus ignore the temporal dimension and always refer to a single hour.

2.2.1 Model structure

The empirical distribution of rainfall measurements shows high percentages of zero values (see Table 2.1); on the other hand, the continuous nature of the phenomenon and the high sensibility of the instruments suggest a continuous distribution is suitable for modelling positive amounts. A review of possible approaches for dealing with this kind of data, called semicontinuous, can be found in Frees (2009) and Neelon et al. (2014). A flexible choice is constituted by the two-part model, which was proposed in Sloughter *et al.* (2007) for modelling rainfall; it constitutes the first level of the hierarchy in the following proposed models. This likelihood allows, at the higher levels of the hierarchy, to model both rain probability and, conditionally on rain occurrence, rain accumulation. As a difference from tobit model, in which a single latent process drives both the zeros and the positive values, a two-part approach ensures the widest possible flexibility, addressing zero values and positive amounts with separate ad-hoc modeling. Conditionally on rainfall probability and radar data, rain gauge data are therefore independently distributed as:

$$p(Y_s|R,\pi_s) = P_{0\,s} I_{Y_s=0} + (1 - P_{0\,s}) p(X_s) I_{Y_s>0}, \quad s \in S_G \tag{2.2}$$

where I is the indicator function, $P_{0\,s}$ is the probability of zero at location s and $p(X_s) = p(Y_s|Y_s > 0)$ is the distribution of the hourly rain accumulation when rain occurs.

The second level of the hierarchy regards rain occurrence. We propose a spatial probit regression where rain probability is modelled as a function of log-transformed radar data plus a spatial adjustment specified as a Gaussian spatial process with exponential covariance function:

$$\text{probit}(1 - P_{0\,s}) = \gamma_1 + \gamma_2 \log(R_s) + \epsilon_s \qquad \text{where} \tag{2.3}$$

$$\boldsymbol{\epsilon}|\sigma^2_\epsilon, \phi_\epsilon \sim MVN(\mathbf{0}, \sigma^2_\epsilon \boldsymbol{\Sigma}_\epsilon) \quad \text{and} \quad \boldsymbol{\Sigma}_\epsilon(s, s') = \exp(-\phi_\epsilon \, d_{s,s'}), \quad s, s' \in S_G \tag{2.4}$$

where $\boldsymbol{\Sigma}_\epsilon(s, s')$ denotes the (s, s')-entry of the spatial covariance matrix of the random effects $\boldsymbol{\epsilon}$, and $d_{s,s'}$ indicates the Euclidean distance between sites s and s'. Parameters σ^2_ϵ and ϕ_ϵ denote the sill and the decay parameter of the spatial covariance function, respectively. Dropping the spatial random effect from the model would imply that radar measurements efficiently explain rain occurrence along space, which is not the case, as shown in Bruno *et al.* (2014). R_s denotes radar information[1] associated with location s; in a basic formulation, it can consist of the radar measurement provided by the cell containing location s, but other choices are possible. A detailed discussion of gauge-radar matching is provided in Section 2.2.2, in which several specifications of the model are proposed.
Probit link has been preferred against logit for simplicity in implementation; no relevant differences distinguish the results obtained with the two different specifications.

With regard to the conditional distribution of rain accumulation given rain occurrence, the main features to be respected are the positive support and the right-skewness. Gamma distribution is thus a suitable choice, often used for modelling precipitation (see for example Sloughter *et al.* 2007; Berrocal *et al.* 2008), thanks to it asymmetry and flexibility. We parameterized it with the second parameter representing the rate:

$$X_s|\mu_s, \tau \sim Gamma(\tau, \tau/\mu_s), \quad s \in S_G \tag{2.5}$$

in order to have

$$\text{E}[X_s|\mu_s, \tau] = \mu_s \quad \text{Var}[X_s|\mu_s, \tau] = \mu^2_s/\tau. \tag{2.6}$$

An alternative can be the Lognormal distribution (see for example Fuentes *et al.* 2008); we do not treat the Lognormal specification here, since in Bruno *et*

[1] Since modeling is performed in the log scale, max(0.01,R_s) is taken

al. 2014 we showed it returns sightly worse results than Gamma distribution in terms of predictive performance.
According to (2.6), the model is heteroscedastic but characterized by a constant coefficient of variation. In the original scale, the variance is expressed as an increasing function of the mean. Parameter τ is devoted to capture measurement errors and to accommodate for the spatial misalignment between rain gauge point data and radar data measured on a raster grid; it can be interpreted as a tuning parameter adjusting the mean-variance relationship.

The third level of the hierarchy is explicitly devoted to the calibration of radar measurements. The calibration equation for rain amounts is specified in the log scale as follows:

$$\log \mu_s = \alpha_s + \beta_1 + \beta_2 \log(R_s). \tag{2.7}$$

The inclusion of a spatial random effect is needed in order to capture the influence of unobserved confounding factors on the reliability of radar measurements as proxies of rain gauges. This is the role of terms α_s, each modelled as a Gaussian spatial process, i.e.:

$$\boldsymbol{\alpha}|\sigma^2_\alpha, \phi_\alpha \sim MVN(\mathbf{0}, \sigma^2_\alpha \boldsymbol{\Sigma}_\alpha) \quad \text{and} \quad \boldsymbol{\Sigma}_\alpha(s, s') = \exp(-\phi_\alpha \, d_{s,s'}), \quad s, s' \in S_G. \tag{2.8}$$

The covariance function is assumed to be exponential with sill σ^2_α and decay parameter ϕ_α. The inclusion of the spatial effect translates the simple idea of continuity of radar bias, suggested by the smoothness of radar surface (see Fig. 2.3). If radar data are positively/negatively biased at a rain gauge site, then they will be positively/negatively biased in the neighbourhood.

Equations 2.7 and 2.3 might in principle be enriched with the addition of further geographical or meteorological covariates. For what concerns our case study, the preprocessing performed by ARPA on radar information already addressed the removal of the influence of orography and meteorological conditions on the measurements. Moreover, explorative analysis revealed that altitude is not relevant in explaining discrepancies between radar and rain gauges, and no evidence is found in literature supporting correlation between altitude and rainfall when short cumulation periods (like an hour) are considered.

The model hierarchy is completed by hyperpriors specification: we assume Normal independent priors N(0, 1000) for parameters Γ_1, Γ_2, β_1, β_2. Independent small parameter Gamma distributions, i.e. Gamma (0.001, 0.001), are assumed for parameters τ, σ^2_ϵ and σ^2_α. The decay parameters ϕ_ϵ and ϕ_α are fixed, in order to guarantee an approximate range of 100 and 150 km for the spatial effects driving rainfall probability and rain accumulation respectively; further details about this choice are provided in Section 2.3.

2.2.2 Alternative model specifications

Both rain occurrence and rain accumulation are modelled by exploiting radar measurements as a covariate. In Bruno *et al.* (2014) the matching between rain gauges and the corresponding radar information is performed by relating Y_s to the radar measurement $R_{P(s)}$, where $P(s)$ is the grid cell containing location s; this corresponds to take $R_s = R_{P(s)}$ in Equations (2.3) and (2.7). From now on this formalization will be denoted as Model "base"; it represents a simple downscaler. The spatial effects correct for potential inconsistencies due to the different spatial supports on which the data are available.

In order to directly face misalignment and to take advantage of the abundance of radar spatial information, a covariate $R_{\overline{P}_8(s)}$ can be added, representing the mean of the 8 cells surrounding $P(s)$; this corresponds to substitute Equation (2.7) with

$$\log \mu_s = \alpha_s + \beta_1 + \beta_2 \log(R_{P(s)}) + \beta_3 \log(R_{\overline{P}_8(s)}). \tag{2.9}$$

In this way we preserve the privileged role of pixel $R_{P(s)}$ containing location s, and enrich the calibration equation with radar neighbouring information. This simple procedure can of course be applied to Equation (2.3) as well by setting $\text{probit}(1 - P_{0\,s}) = \gamma_1 + \gamma_2 \log(R_{P(s)}) + \gamma_3 \log(R_{\overline{P}_8(s)}) + \epsilon_s$; nevertheless, results show the basic model is already very skilled in predicting rain occurrence, as confirmed by diagnostic tools such as the Brier Score (see Bruno *et* *al.* 2014 and Section 5.2.1). Therefore, we we focus on modelling the positive amounts of rain, looking for enrichments which can improve predictive performance.

Following Berrocal *et al.* (2011), the mean of radar cells can be made more flexible and effective by introducing stochastic weights and extending the mean over the whole radar map. As in the previous case, we only adopt this method when modelling rainfall amounts. More precisely, R_s in Equation (2.7) takes the form of an s-specific weighted mean of radar values over the grid S_R:

$$R_{\overline{P}s} = \sum_{P \in S_R} w_{P,s} R_P, \quad s \in S_G. \tag{2.10}$$

This model will be denoted as Model "SW" (Stochastic Weighting) in the following. The weights w are stochastic, relying on a unique latent N_R-dimensional Gaussian process Q, defined over the grid; its exponential covariance function has decay parameter ψ_Q. The influence of each component Q_P of such a process on a specific location $s \in S_G$ is smoothed according to the distance between s and the centroid c_P of pixel P using an exponential kernel $\mathcal{K}$:

$$w_{P,s} = \frac{\mathcal{K}(s, c_P) \exp(Q_P)}{\sum_{P' \in S(R)} \mathcal{K}(s, c_P) \exp(Q_{P'})}, \quad \mathcal{K}(s, c_P) = \exp(-\psi_Q(d_{s,c_P})). \tag{2.11}$$

The decay parameters ψ_Q and ψ_K for the covariance function of Q and for $\mathcal{K}$ respectively are exogenously set in order to make negligible the influence of the pixels exceeding a distance of 5 km.

While the kernel $\mathcal{K}$ attributes more weight to radar pixels near the location of interest s, determining a symmetric decrease with increasing distance, the latent GP Q is defined on the radar grid and is able to capture local effects, but its value is common to each location s; thus, when multiplying $\mathcal{K}$ and $\exp(Q)$, asymmetric patterns can be produced, allowing weights to be directional, possibly generating different shapes of the weighting schemes when moving from site to site. When the pixel containing location s is the most representative, as expected in most of the cases, the flexible weighting driven by this method is able to attribute it a leading role. For this reason, as a difference from Model "mean", there is no need to keep $R_{P(s)}$ as an additional

covariate; thus, we simply take $R_s = R_{\overline{P}s}$.
From (2.11), the Q process is not identified, since a shift in its center leaves the weights unchanged; as suggested in Berrocal *et al.* (2011), we thus impose a sum to zero constraint, implementing it on the fly during model fitting. Moreover, in order to alleviate computation when working with the $49,000$-dimensional process Q, as suggested in Banerjee *et al.* (2008) a predictive process Q^* is actually estimated instead of Q, defined on a rougher grid (1 pixel every 16 in both directions). The Gaussian process specification is therefore imposed only on Q^*, i.e. on the knots of the rougher grid; this trick allows a reduced rank approach, requiring the inversion of smaller matrices. Q^* consists in the projection of the original spatial process onto the smaller-dimensional space; at the other locations, Q is replaced by the conditional expectation given the knots, obtained by exploiting the analytic properties of the multivariate normal distribution. Little sensitivity on knots selection is proved, particularly if the knots are chosen on a regular grid with small spacing relatively to the range of the parent process, as in our case.

The three aforementioned model specifications are the result of a selection made on a wider set of tested models. In particular, some modifications of Model “mean” were tested, trying for example to exploit wind information for driving the deterministic weighting scheme of radar pixels. Nevertheless, wind is a very complex phenomenon with a high variability also in the vertical direction, due to the differences in the flows at different altitudes, and simple approaches did not provide conclusive improvements with respect to the presented models.

2.3 Model estimation and computational issues

Parameter estimation is performed through Markov chain Monte Carlo algorithms, implementing a Gibbs sampler with Metropolis-Hastings steps. OpenBUGS (Thomas *et al.* 2006) codes were replaced by own MCMC samplers for a better control and understanding of the estimation steps. R software is employed for data manipulation and visualization and for handling the results, while the computational core is written in C for allowing fast computation. The link between C and R is performed via the loading of pre-compiled dynamic libraries, both in Windows and in Linux Operating Systems, generating .dll or .so files, respectively. The exploitation of BLAS and LAPACK algebraic libraries determines a further dramatic improvement in computational speed, which is now compatible with real time application of the method. In fact, Model "base" only takes approximately ten minutes for 200,000 MCMC iterations. The computational times we refer to correspond to mean performances tested on several machines (among which Intel(R) Core(TM) i5 2.27 GHz, AMD Phenom(tm) II X4 945 Processor 3.0 GHz, and on ARPA computer) with R generic BLAS and LAPACK versions, for guaranteeing portability of the code; they are further reduced, also noticeably, if optimized algebraic libraries are chosen according to the machine under use. Most of the computational time is spent by the estimation of the probability of rain, which follows the implementation for the probit model suggested by Holmes and Held (2006). In particular, let $\tilde{Y}_s$ be a random variable denoting rain occurrence at location s, i.e. $\tilde{Y}_s \sim \text{Bernoulli}(1 - P_{0\,s})$. Then, Equations (2.3) and (2.4) have the following equivalent representation exploiting an augmented approach with auxiliary variables:

$$\tilde{Y}_s = \begin{cases} 1 & \text{if} \quad Z_s > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

$$\begin{aligned} &Z_s = \gamma_1 + \gamma_2 \log(R_s) + \epsilon_s + \nu_s, \quad \nu_s \text{ i.i.d. } \sim N(0,1) \\ &\boldsymbol{\epsilon}|\sigma_\epsilon^2, \phi_\epsilon \sim \; MVN(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{\Sigma}_\epsilon) \quad \text{and} \quad \boldsymbol{\Sigma}_\epsilon(s,s') = \exp(-\phi_\epsilon \, d_{s,s'}), \quad s, s' \in S_G. \end{aligned} \tag{2.13}$$

In this formulation, $\tilde{Y}_s$ is deterministic conditionally to the sign of the auxiliary variable Z_s. The unit variance of the white noise effects ν_s guarantees the identifiability of the model; more details for the spatial case with correlated spatial effects are provided in Gelfand *et al.* (2000). Marginalization over the spatial effects allows faster convergence at the cost of many non-diagonal matrices inversions; algorithms suggested by Rue and Held (2005) help in quickly extracting correlated Gaussian random variables. The estimation of rainfall accumulation is fast in the basic model, only taking a few minutes; this is made possible by the exploitation of efficient algebraic functions and by the need for estimation only involving the rainy sites. The additional effort required by Model “mean” is negligible, thanks to an efficient matrix implementation. Model “SW” is instead more computationally intensive, requiring about an hour of estimation for each run over an hour of data with 200000 iterations. As explained in Section 2.2.2, a predictive process Q^* instead of the whole Q is estimated for reducing dimensionality; moreover, updating of Q and its sill parameter is performed via a block Metropolis-Hastings step in 1 MCMC iteration every 10.

A burnin of 50,000 iterations is removed, corresponding to a reasonable set of pre-convergence iterations; convergence has been checked by graphical examination of the trace plots of the chains sample values versus iterations, and of the autocorrelation plot for each chain. Thinning of the chains is performed, keeping only one over 100 MCMC iterations in order to reduce autocorrelation; tests on the persistent autocorrelation show a desirable behaviour. In this way, a final set of 1,500 iterations is obtained for each parameter, representing an appropriately sized sample from the posterior distribution. Simplicity in the predictive procedure and code optimization allow a reconstruction of the whole rainfall field with a 1 km × 1 km resolution in a couple of minutes. It is relevant to highlight the effort towards an efficient implementation, which allowed to reach a 15-minute version on 200,000 estimating iterations against an initial (non-optimized) full-R implementation requiring about one hour for the estimation and one hour for the prediction with one tenth of the iterations.

Estimation of rain probability is performed via Gibbs sampler exploiting full conditionals for each parameter. When dealing with rain accumulation instead, closed form full conditional is only available for the variance of the spatial effect σ_α. As anticipated, Metropolis-Hastings steps compensates for this lack; tuning algorithms help in retrieving a correct mixing and accelerating the convergence, by modifying the dispersion of the proposal distributions via the successive modification of appropriate tuning parameters, in order to reach acceptance rates around 40% for singly updated parameters and 25% for vector parameters, as common working rules suggest.

A separate mention is needed for the choice of the decay parameters ϕ_ϵ and ϕ_α. An empirical Bayes approach was firstly followed (Sahu *et al.* 2010): a two-dimensional grid with varying values of the decay parameters for rain probability and accumulation is built, and an optimal value is chosen on the basis of a predictive criterion. In particular, we considered a set S_V of validation sites, and chose the couple ϕ_ϵ and ϕ_α which minimized the mean square prediction error. This approach relies on two main justifications: first of all, spatial interpolation is sensitive only to the product between the decay parameter and the variance parameter (Stein 1999) that are weakly identifiable when contemporaneously included in the model as unknowns; secondly, setting the decay parameters allows to speed-up computation. The search for optimal range values was driven by the assumption that events may be characterized by different ranges, according to the nature and the extension of the perturbation. The analysis on the chosen set of rainfall events was conducted on a two dimensional grid containing the values 0.01, 0.02, 0.03 and 0.04 corresponding to spatial ranges of 300, 150, 100 and 75 km. These range values permit to appreciate the influence of random effects in the spatial domain, since the maximum distance between rain gauge sites is 250 km. Spatial ranges of 100 and 150 km for rainfall probability and rain accumulation, respectively, turned out to be appropriate for all events in terms of MSPE (see Bruno *et al.* 2014). This result was confirmed by other approaches for the estimation of ϕ_ϵ and ϕ_α, such as the adoption of discrete or continuous priors. These techniques reflect that convective events are usually more localized, showing smaller values for the ranges; nevertheless, the employment of such estimated values does not improve results. The explanation is twofold: on one hand, it is well known that flexibility in modelling does not always guarantee better results, even when convergence is reached, since simpler formulation can improve stability; on the other hand, wide spatial dependence may characterize all events despite the effective extension of the nucleus of the storm. All these trials suggested that the model can be run with the fixed chosen values for ϕ_ϵ and ϕ_α; this also reduces the computational burden.

2.4 Sampling from the predictive distribution

Major attention is devoted to predictions. They are firstly employed for model checking in the validation sites, allowing comparison between different specifications. Then, reconstruction of the rainfall fields is achieved by computing predictions at all the pixels of the radar grid. Sampling from the predictive distribution of rain accumulation at an unmonitored site s_0 needs to take account of both the probability of rain occurrence and the rain accumulation in case of rain occurrence. Let K be the size of the postconvergence MCMC sample. Samples of rain probabilities for the k-th MCMC iteration $(k = 1, \ldots, K)$ are obtained as follows:

1. Predict $\epsilon_{s_0}^{(k)}$, on the basis of (2.4), by sampling from

$$\epsilon_{s_0}^{(k)}|\boldsymbol{\epsilon}^{(k)}, \sigma_\epsilon^{2\ (k)} \sim \mathcal{N}(\sigma_\epsilon^{-2\,(k)}\omega_\epsilon^{(k)\prime}\Sigma_\epsilon^{-1}\boldsymbol{\epsilon}_s^{(k)},\ \sigma_\epsilon^{2(k)} - \sigma_\epsilon^{-2\,(k)}\omega_\epsilon^{(k)\prime}\Sigma_\epsilon^{-1}\omega_\epsilon^{(k)}) \tag{2.14}$$

   where $\omega_\epsilon^{(k)}$ is an n-dimensional vector with elements $\sigma_\epsilon^{2\,(k)}\exp(-\Phi_\epsilon d_{ss_0})$, $s \in S_G$

2. According to (2.3), generate a realization of a Bernoulli variable with parameter

$$1 - P_0 = \Phi(\Gamma_1^{(k)} + \Gamma_2^{(k)}\log(R_{P(s_0)}) + \epsilon_{s_0}^{(k)})$$

   with Φ denoting the CDF of the standard normal distribution.
   If the outcome is zero, then $\hat{Y}_{s_0}^{(k)} = 0$. Otherwise, sampling from the predictive distribution of the rain accumulation consists in the two steps below:

3. Predict $\alpha_{s_0}^{(k)}$ following a procedure analogous to (2.15):

$$\alpha_{s_0}^{(k)}|\boldsymbol{\alpha}^{(k)}, \sigma_\alpha^{2\ (k)} \sim \mathcal{N}(\sigma_\alpha^{-2\,(k)}\omega_\alpha^{(k)\prime}\Sigma_\alpha^{-1}\boldsymbol{\alpha}_s^{(k)},\ \sigma_\alpha^{2(k)} - \sigma_\alpha^{-2\,(k)}\omega_\alpha^{(k)\prime}\Sigma_\alpha^{-1}\omega_\alpha^{(k)}) \tag{2.15}$$

   where $\omega_\alpha^{(k)}$ is an n-dimensional vector with elements $\sigma_\alpha^{2\,(k)}\exp(-\Phi_\alpha d_{ss_0})$, $s \in S_G$

4. Generate a realization from a Gamma$(\tau^{(k)}, \tau^{(k)}/\mu_{s_0}^{(k)})$, where
$$\log(\mu_{s_0}^{(k)}) = \alpha_{s_0} + \beta_1 + \beta_2 \log(R_{s_0}).$$

In step 4, R_{s_0} assumes the value $R_{P(s_0)}$, $R_{P(s)} R_{\overline{P}_8(s_0)}^{\beta_3^{(k)}}$ or $R_{\overline{P}s_0}^{(k)}$ for Model "base", "mean" and "SW" respectively. In the last case, the apex k is needed for specifying the stochastic weights corresponding to the estimation of the Q process available at the k-th iteration:

$$R_{\overline{P}s_0} = \sum_{P \in S_R} w_{P,s_0}^{(k)} R_P, \quad w_{P,s_0}^{(k)} = \frac{\mathcal{K}(s_0 - c_P) \exp(Q_P^{(k)})}{\sum_{P' \in S(R)} \mathcal{K}(s_0 - c_{P'}) \exp(Q_{P'}^{(k)})}.$$

Let us note that, differently from the assessment of the spatial effects $\boldsymbol{\epsilon}$ and $\boldsymbol{\alpha}$, no spatial interpolation is required for the process Q, which is not s_0-specific: it is defined on the grid S_R and each value Q_P corresponding to a cell P is appropriately weighted for the prediction in location s_0 according to its distance from s_0 via the deterministic kernel $\mathcal{K}$.

The presented predictive procedure provides a chain composed of both zeros and positive values. The probability of rain, predicted at each iteration and thus available as a chain, can also be approximated as the percentage of zero values in the chain drawn from the posterior predictive distribution.

Chapter 3

Forecast evaluation

Once a forecasting model is available, the evaluation of its predictive performance is a crucial issue. In fact, the analysis of the results reveals whether the goals of the study have been reached; further efforts might be necessary for reducing the weaknesses that affect predictions. Moreover, comparison between competing models allows to rank the available forecasting procedures on the basis of a desired performance criterion, detecting the best solution according to the specific purpose at hand. Defining the focus of the study is fundamental for correctly understanding, evaluating and comparing all results. Are we interested in producing point predictions, or do we seek for predictive distributions reproducing the true data generating process? In the first case, do we want to penalize single huge errors? Are we aware of the uncertainty associated with the forecasts? This Chapter provides a review of the main concepts about the form in which forecasts can be provided and presented, the features which ideal predictions should possess and how to evaluate predictive performance. First of all, the distinction between point and probabilistic predictions must be made clear. The former consist in a single number representing the forecaster's best estimation of the quantity of interest; the latter provide a whole predictive distribution for each location and time instant. Predictive distributions communicate complete information about the forecast, and can be reduced to simple point forecasts by the application of suitable functionals, like the mean or a quantile.

The assessment of the predictive ability of forecasters, and the comparison and ranking of competing forecasting methods, are critical issues. Relevant methodology is present in the literature for meteorology (Jolliffe and Stephenson, 2003) and econometrics (Diebold and Mariano, 1995; Christoffersen, 1998; Diebold *et al.*, 1998; Corradi and Swanson, 2006). This Chapter provides guidelines for the assessment of predictive performances and the ranking of competing predictive models; specific tools for precipitation predictions are investigated in Chapter 4.

3.1 Probabilistic forecasts

Due to the inner uncertainty characterizing predictions of non-deterministic phenomena, forecasts should be probabilistic in nature, taking the form of probability distributions (Dawid, 1984). Predictive distributions completely specify the process of interest. In simple cases, they may be available in analytic form. In the classical framework, this is often obtained via the plug-in of parameter estimates ; in contrast, Bayesian solutions aim at a correct assessment of the variability of the results, taking the uncertainty on parameters into account. In case of complex models, advances in Markov chain Monte Carlo methodology have led to explosive growth in the use of predictive distributions, mostly in the form of Monte Carlo samples from the posterior predictive distribution; every summary of the distribution can be obtained from this sample with the desired level of accuracy. Finally, even in the field of numerical models, the awareness of uncertainty in input data and of imprecisions in the mathematical and physical formulations urged forecasters to run numerical models many times, with different boundary conditions corresponding to perturbations of the best estimate of the existing state, and several specifications in the model; this procedure provides an ensemble, consisting in a sample of point forecasts with the aim of approximating the predictive distribution.

In the statistical literature, the diagnostic approach for dealing with probabilistic forecasts faces a challenge, in that the predictions take the form of probability distributions whereas the observations are real valued.

Following Gneiting *et al.* (2007), I denote with G_i the true unknown data-generating process, in the form of a predictive Cumulative Distribution Function, for individual i, $i = 1, \ldots, n$ (the subscript may refer to time, space and/or subjects, and no sequentiality is assumed), with y_i the corresponding observation , and with F_i the predictive CDF; an ideal forecaster would choose $F_i = G_i$ for all i. In accordance with Dawid's (1984) prequential principle, probabilistic forecasts need to be assessed on the basis of the forecast-observation pairs (F_i, y_i) only, regardless of their origins.

Probabilistic forecasts must be statistically consistent with the observations. This concept goes under the name of "calibration", which assumes a different and more specific meaning with respect to Chapters 1 and 2, where it denoted the consistency of radar deterministic information with rain gauge observations. Subject to calibration, sharp forecasts are desirable (see Gneiting *et al.* 2007).

Gneiting *et al.* (2007) propose a formalization of the concept of calibration providing several definitions corresponding to different asymptotic properties; we focus on the most widely studied form of calibration, which goes under the name of "probabilistic calibration":

Definition 3.1. The sequence $F_1, \ldots, F_n$ is **probabilistically calibrated** relative to the sequence $G_1, \ldots, G_n$ if

$$\frac{1}{n}\sum_{i=1}^{n} G_i \circ F_i^{-1}(p) \to p \quad \forall p \in (0,1) \tag{3.1}$$

where "$\circ$" denotes the composition operator.

The following sections review the existing literature about the assessment of probabilistic forecasts, presenting the main tools for checking probabilistic calibration and evaluating predictive distributions. Both graphical and numerical tools are provided; they are presented in Section 3.1.1 and 3.1.2 respectively.

3.1.1 PIT histogram

For assessing probabilistic calibration, Dawid (1984) and Diebold *et al.* (1998) proposed the use of the Probability Integral Transform (PIT). It consists in the value that the predictive CDF attains at the observation: $\mathrm{PIT}(y_i) = F_i(y_i)$. If the forecasts are ideal and F_i is continuous, then the PIT has a uniform distribution. The connection to probabilistic calibration is established by substituting the empirical distribution function $\mathbb{1}(y_i \leq y)$ for the data-generating distribution $G_i(y)$, in the probabilistic calibration condition (3.1), and noting that $y_i \leq F_i^{-1}(p)$ if and only if $\mathrm{PIT}(y_i) \leq p$. Hence, checks for PIT uniformity constitute a tool for forecast evaluation. An exploratory approach is usually adopted, by plotting the empirical CDF of the PIT values and comparing it with the CDF of the uniform distribution. As an alternative, histograms of the PIT values can be displayed, with 10 or 20 histogram bins being generally adequate (Diebold *et al.* 1998, Gneiting *et al.* 2005); this kind on display is recommended when the sample size is large and departures form uniformity are small. Underdispersion, overdispersion or bias in the predictive distribution can be detected by visual inspection, since they give raise to U-shaped, inverse U-shaped or triangle-shaped PIT histograms, respectively. Formal tests of uniformity can be employed, but non negligible efforts are required for their definition in case of complex dependency structures, and they are subject to potential fallacies (Hamill 2001). When dealing with ensemble forecasts, PIT histogram is substituted by the verification rank histogram, which exploits the empirical CDF of the ensemble values as predictive distribution. This tool, also called Talagrand diagram, was proposed independently by Anderson (1996), Hamill and Colucci (1997) and Talagrand *et al.* (1997), and represents the principal device for assessing calibration. It consists in the histogram of the rank of the observations when pooled within the ordered ensemble values; interpretation follows the same rationale as for PIT.

In case of a discrete distribution, PIT is no longer uniform even under the hypothesis of an ideal forecast, due to the jumps in the CDF which prevent PIT from assuming certain values. To remedy this, several authors have suggested a randomized PIT (Smith, 1985; Frühwirth-Schnatter, 1996; Liesenfeld *et al.* 2006; Brockwell, 2007). Gneiting and Ranjan (2013) propose a generalised formulation of the randomized PIT, which holds for every CDF F:

$$\mathrm{PIT}(F, y) = \lim_{w \uparrow y} F(w) + V(F(y) - \lim_{w \uparrow y} F(w)), \quad V \sim \mathrm{Unif}(0, 1). \tag{3.2}$$

Nevertheless, its random nature causes slight changes in the PIT behaviour according to the specific realization of V in (3.2). As an alternative, Czado *et al.* (2009) proposed a non-randomized version of the PIT histogram, still preserving uniformity when dealing with count data. Let us denote the cumulative distribution function with $(P_k)_{k=0}^{\infty}$, and define $P_{-1} = 0$; then the non-randomized PIT histogram is obtained by calculating the conditional CDF F^{PIT} of the randomized PIT given the observed count $y \in \mathbb{N}$:

$$F^{PIT}(u|y) = \begin{cases} 0 & u \leq P_{y-1} \\ (u - P_{y-1})/(P_y - P_{y-1}) & P_{y-1} \leq u \leq P_y \\ 1 & u \geq P_y \end{cases} \tag{3.3}$$

The letter i, $i = 1 \ldots, n$ has been dropped for simplicity of notation, but it should be attached both to F^{PIT} and any instance of y in formula (3.3).
The mean PIT is thus obtained aggregating the PIT CDFs $\{F^{PIT}(u|y_1), \ldots, F^{PIT}(u|y_n)\}$ over the conditioning observations $\{y_1, \ldots, y_n\}$:

$$\bar{F}(u) = \frac{1}{n} \sum_{i=1}^{n} F^{PIT}(u|y_i), \quad 0 \leq u \leq 1. \tag{3.4}$$

Calibration can be assessed directly comparing $\bar{F}$ to the CDF of the standard uniform law, that is, the identity function, or by creating a non-randomized PIT histogram as follows: once chosen the number of bins, named J, the frequencies are computed as

$$f_j = \bar{F}\Big(\frac{j}{J}\Big) - \bar{F}\Big(\frac{j-1}{J}\Big), \quad j = 1, \ldots, J \tag{3.5}$$

and used as heights for equally spaced bins. If the predictive distributions coincide with the true ones, the histogram is uniform.
The non-randomized PIT provides a fixed and reproducible diagnostic tool, not affected by random fluctuations.

3.1.2 Proper scoring rules

Scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performances, in that they address calibration and sharpness simultaneously (Gneiting *et al.* 2007). A scoring rule is a function $S(F^{pred}, y^{obs})$ of the predictive distribution F^{pred} and of the observation y^{obs}, and consists in a penalty the forecaster aims to minimize. A validation approach is adopted, keeping n observations $y_1^{obs}, \dots, y_n^{obs}$ out of the estimating set, using the model for predicting them, and comparing the predictions F_i^{pred} with the effective outcomes y_i^{obs}; a single score for a set of predictions is obtained as the mean over the cases, i.e.

$$\overline{S}(\mathbf{F}^{pred}, \mathbf{y}^{obs}) = \frac{1}{n}\sum_{i=1}^{n} S(F_i^{pred}, y_i^{obs}) \tag{3.6}$$

where $\mathbf{F}^{pred}$ denotes the set of predictive distributions $\{F_1^{pred}, \dots, F_n^{pred}\}$ and $\mathbf{y}^{obs} = \{y_1^{obs}, \dots, y_n^{obs}\}$.
A fundamental property for a scoring rule is propriety: in fact, it guarantees fairness in the evaluation of results.

Definition 3.2. A scoring rule is **proper** if

$$E_F S(F, Y) \leq E_F S(G, Y) \ \forall F,\, G.$$

Propriety implies that if the true distribution is F, the scoring rule is minimized on average when the proposed forecast is exactly F. Strict propriety if achieved when the minimum is unique. Proper scoring rules thus encourage honest and sharp forecasts (see Winkler 1977), representing decision theoretically justifiable tools for probabilistic forecasts evaluation.
We now provide an overview of the most relevant and well-known scoring rules.
A well known proper scoring rule for checking binary events is the **Quadratic** or **Brier Score**

$$\mathrm{BS}(F^{pred}, y^{obs} \in \Omega) = (P_F^{pred}(Y \in \Omega) - I_{\{y^{obs} \in \Omega\}})^2 \tag{3.7}$$

which compares the predicted probability of the realization of an event with the effective outcome, being Ω the set of values corresponding to the realization of the event; notice that it is bounded between 0 and 1.

When dealing with continuous variables, the **logarithmic score** is the negative of the logarithm of the predictive density f evaluated in the observation (Good, 1952; Bernardo, 1979). It is proper and has many desirable properties (Roulston and Smith, 2002), for example it is the only local proper scoring rule, meaning it only depends on the value of the predictive density in the observation. Some of its drawbacks are lack of robustness (Selten, 1998; Gneiting and Raftery, 2007) and the need of explicit predictive density f^{pred}. In case the predictive distribution is available in the form of a big sample (as with Markov chain Monte Carlo procedures), we denote with $F^{pred,k}$ the predictive CDF corresponding to the k^{th} iteration, for $k = 1, \dots, K$, and with $f^{pred,k}$ the respective density. Thus, the logarithmic score can be approximated as

$$\mathrm{LS}(F^{pred}, y^{obs}) = -\log\Big(\frac{1}{K}\sum_{k=1}^{K} f^{pred,k}(y^{obs})\Big) \tag{3.8}$$

(see e.g. Czado and Gschlößl 2007).
Another interesting scoring rule for predictive distributions on $\mathbb{R}^m$ depending only on the estimated mean vector $\boldsymbol{\mu}$ and the dispersion or covariance matrix Σ is the **Dawid and Sebastiani score** (Dawid and Sebastiani 1999, Gneiting and Raftery 2007):

$$\mathrm{DS}(\mathbf{F}^{pred}, \mathbf{y}^{obs}) = -\log(\det\Sigma) - (\mathbf{y}^{obs} - \boldsymbol{\mu}^{\boldsymbol{pred}})^T\Sigma^{-1}(\mathbf{y}^{obs} - \boldsymbol{\mu}^{\boldsymbol{pred}}). \tag{3.9}$$

It is proper relative to for probability measures with finite first two moments, but strictly proper only if such moments fully define the distribution (as in the Gaussian case).

The **Continuous Ranked Probability Score** (CRPS) is the integral of the Brier Scores associated with the binary events describing the exceeding of all possible thresholds in the set of outcome values:

$$\begin{aligned}\mathrm{CRPS}(F^{pred}, y^{obs}) &= \int_{(-\infty,+\infty)} \mathrm{BS}(F^{pred}, \{y^{obs} \le x\})\, dx \\ &= \int_{(-\infty,+\infty)} (F^{pred}(x) - I_{\{y^{obs} \le x\}})^2\, dx\end{aligned} \tag{3.10}$$

where F^{pred} is the predictive CDF and y^{obs} is the observation. When the predictive distribution is represented by a sample, possibly based on MCMC output or ensemble forecasts, the following alternative representation of the CRPS is useful for its calculation (Gneiting and Raftery 2007):

$$\mathrm{CRPS}(F^{pred}, y^{obs}) = E_{F^{pred}}|Y - y^{obs}| - \frac{1}{2}E_{F^{pred}}|Y - Y'| \tag{3.11}$$

where Y and Y' are independent copies of a random variable with CDF F^{pred} and finite first moment. If we consider the discrete approximation of F^{pred} via the sample $\{F^{pred,1}, \ldots, F^{pred,k}\}$, the second expectation in (3.11) implies a double summation, which can be time demanding if the sample is big; nevertheless, the following equivalence holds:

$$\sum_{i,\,j=1}^{K} |x_i - x_j| = 2\sum_{i=1}^{K-1} i(K-i)(x'_{i+1} - x'_i)$$

where $x'_1 \leq \ldots \leq x'_K$ are obtained after reordering $\{x_1, \ldots, x_K\}$ (see Hersbach 2000 and Scheuerer 2013).

Representation (3.11) also shows that the continuous ranked probability score generalizes the absolute error, to which it reduces if F is a point forecast; moreover, it highlights the fact that the CRPS is reported in the same unit as the observations.

The continuous ranked probability score is proper. As anticipated in (3.6), ranking of competing forecasting procedures is performed via its average:

$$\begin{aligned} \overline{\mathrm{CRPS}}(\mathbf{F}^{pred}, \mathbf{y}^{obs}) &= \int_{(-\infty,+\infty)} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathrm{BS}(F_i^{pred}, \{y_i^{obs} \leq x\})\, dx \\ &= \int_{(-\infty,+\infty)} \overline{\mathrm{BS}}(\mathbf{F}^{pred}, \{\mathbf{y}^{obs} \leq x)\}\, dx \end{aligned} \tag{3.12}$$

where $\overline{\mathrm{BS}}(\mathbf{F}^{pred}, \{\mathbf{y}^{obs} \leq x\})$ denotes the mean Brier Score, averaged over the forecasting cases, corresponding to the threshold x. Formula (3.12) provides a decomposition of the mean CRPS according to the thresholds: besides calculating the single overall CRPS value, a plot of the mean Brier Score versus the thresholds may be a useful diagnostic tool, providing a deeper insight into the behavior of competing predictions (Schumacher *et al.* 2003 call it prediction error curve, Gneiting *et al.* 2007 name it as Brier Score plot). It can reveal variations in the ranking of the predictions on different intervals. If our interest is not uniform over the domain of the possible outcomes, it is possible to apply a nonnegative weighting function w to the CRPS threshold decomposition formula (4.4), obtaining the (still proper) **weighted CRPS** (see Gneiting and Ranjan 2011):

$$\mathrm{wCRPS}(F^{pred}, y^{obs}) = \int\limits_{(-\infty,+\infty)} w(x)(F^{pred}(x) - I_{\{y^{obs} \le x\}})^2 \, dx. \tag{3.13}$$

The CRPS can also be decomposed and plotted with respect to quantiles (see Gneiting and Ranjan 2011):

$$\begin{aligned} \mathrm{CRPS}(F^{pred}, y^{obs}) &= 2 \int_0^1 (I_{\{y^{obs} \le (F^{pred})^{-1}(\alpha)\}} - \alpha)((F^{pred})^{-1}(\alpha) - y^{obs}) \, d\alpha \\ &= \int_0^1 \mathrm{QS}_\alpha((F^{pred}), y^{obs}) \, d\alpha \end{aligned} \tag{3.14}$$

where $\mathrm{QS}_\alpha((F^{pred}), y^{obs}) = 2\,(I_{\{y^{obs} \le (F^{pred})^{-1}(\alpha)\}} - \alpha)((F^{pred})^{-1}(\alpha) - y^{obs})$ is the quantile score for the quantile forecast $(F^{pred})^{-1}(\alpha)$ at the level $\alpha \in (0, 1)$. When averaging over the forecasting cases, it gives

$$\begin{aligned} \overline{\mathrm{CRPS}}(\mathbf{F}^{pred}, \mathbf{y}^{obs}) &= 2 \int_0^1 \frac{1}{n} \sum_{i=1}^n (I_{\{y_i^{obs} \le (F_i^{pred})^{-1}(\alpha)\}} - \alpha)((F_i^{pred})^{-1}(\alpha) - y_i^{obs}) \, d\alpha \\ &= \int_0^1 \overline{\mathrm{QS}}_\alpha((\mathbf{F}^{pred})^{-1}(\alpha), \mathbf{y}^{obs}) \, d\alpha \end{aligned} \tag{3.15}$$

with $\overline{\mathrm{QS}}_\alpha((\mathbf{F}^{pred})^{-1}(\alpha), \mathbf{y}^{obs})$ denoting the mean of $\mathrm{QS}_\alpha((F^{pred}), y^{obs})$ over the forecasting cases.
Notice that the quantile score is not bounded, neither is CRPS.

3.2 Point forecasts and predictive intervals

Probabilistic forecasts provide complete information about the prediction and its uncertainty and are therefore the best choice. Nevertheless, many practical situations require single valued point forecasts, which can be obtained via the application of a suitable functional like the mean or a quantile. Error measures $S(y^{pred}, y^{obs})$ are exploited for quantifying the distance between point forecasts and observations; following Gneiting (2011), we call

S scoring function. When predicting n cases, the average is taken; thus, the performance criterion takes the form

$$\bar{S}(\boldsymbol{y}^{pred}, \boldsymbol{y}^{obs}) = \frac{1}{n}\sum_{i=1}^{n} S(y_i^{pred}, y_i^{obs}).$$

S is usually taken as negatively oriented: higher $\bar{S}$ values correspond to worse predictions according to the chosen error measure. As anticipated, the predictions $\{y_i^{pred}, i = 1, \ldots, n\}$ are usually obtained as the result of the application of a functional T on the predictive distribution F_i, like the mean or the median. Establishing which summary of the predictive distribution is to be taken is thus a fundamental issue, strictly related to the way in which the forecasts are assessed and compared. Consistency links the choice of the functional to the one of the scoring function.

Definition 3.3. A scoring function S is consistent for a functional T if

$$E_F S(t, Y) \leq E_F S(l, Y)$$

whatever is the true distribution F of Y, the chosen value $t \in T(F)$ and the potential prediction l belonging to the domain of the outcomes.

Thus, concistency means that if the cumulative distribution of the observations is F, no other point prediction can be better on average, according to the chosen scoring function, then the one obtained by applying the functional T to F; in other words, T is the functional that minimizes the expected score. As a consequence, the values of a chosen scoring function S corresponding to competing predictions must be calculated on the point forecasts $\{T(F_1), \ldots, T(F_n)\}$ where T is a functional for which S is consistent; and vice versa, if the functional T is chosen, the point forecasts $\{T(F_1), \ldots, T(F_n)\}$ must be compared according to a consistent scoring function S.
The most common choices for S are the absolute error and the squared error

$$S(y^{pred}, y^{obs}) = |y^{pred} - y^{obs}| \quad \text{and} \quad S(y^{pred}, y^{obs}) = (y^{pred} - y^{obs})^2$$

which lead to the Mean Absolute Error (MAE) and Mean Square Error (MSE) as average scoring functions:

$$\text{MAE}(\boldsymbol{y}^{pred}, \boldsymbol{y}^{obs}) = \frac{1}{n}\sum_{i=1}^{n} |y_i^{pred} - y_i^{obs}| \tag{3.16}$$

$$\mathrm{MSE}(\boldsymbol{y}^{pred}, \boldsymbol{y}^{obs}) = \frac{1}{n}\sum_{i=1}^{n}(y_i^{pred} - y_i^{obs})^2. \tag{3.17}$$

Consistent functionals for these choices are known to be the median and the mean of the predictive distribution, respectively. A main advantage of MAE and MSE is their easy interpretation; if the square root of the MSE is taken (RMSE), they are both expressed in the scale of the observations (without affecting consistency). MSE gives more weight to bigger errors, as desirable in some cases. On the other hand, the median has a straightforward explanation as a quantile (1/2 probability of obtaining a higher or lower amount according to the predictive distribution); moreover, it is robust with respect to fluctuations in the right tail of the distribution. Gneiting (2011) proves that the only consistent scoring functions for an α-quantile have the form of generalized piecewise linear (GPL) scoring functions:

$$\mathrm{GPL}(q^{pred}, y^{obs}) = (I_{\{q^{pred} \geq y^{obs}\}} - \alpha)(g(q^{pred}) - g(y^{obs})) \tag{3.18}$$

where g is a nondecreasing function on the domain. In particular, taking the identity function as g, we obtain the piecewise linear quantile score (see e.g. Koenker and Machado 1999, Gneiting and Raftery 2007):

$$\mathrm{PLQS}(q^{pred}, y^{obs}) = (I_{\{q^{pred} \geq y^{obs}\}} - \alpha)(q^{pred} - y^{obs}). \tag{3.19}$$

Since point forecasts are rather poor, only conveying prediction in the form of a single number, they are often accompanied by predictive intervals, which provide information about the uncertainty.

Credibility intervals are not uniquely defined by the percentage of the predictive distribution they contain; an information about their location is also necessary for identifying them. Lower or upper intervals can be chosen, whose left or right extreme coincides with the left or right boundary of the predictive domain respectively. **Central intervals** are the most common choice; they leave out the same probability in the two tails. When dealing with symmetric unimodal distributions, their center coincides with both the mean and the median of the distribution; in that particular case, they also constitute higher posterior density (HPD) intervals, consisting in the narrower (sets of) intervals among the ones containing the same amount of probability. In case of asymmetric or multimodal distributions, the centered intervals may not be HPD.

Credibility intervals are also used to assess sharpness, via numerical and graphical summaries of their width. In real world applications, conditional heteroscedasticity often determines considerable variability in the width of the prediction intervals; box plots of interval widths represent an instructive graphical device (Bremnes 2004).

Finally, the correctness of predictive intervals is checked via the computation of the coverage, which should be close to the nominal level.

Chapter 4

Evaluation and communication of quantitative probabilistic precipitation forecasts

Probabilistic Quantitative Precipitation Forecasts (PQPF) consist in predictions providing numerical and probabilistic characterization of precipitation. The form in which such forecasts are provided changes according to the purposes and the modelling framework. For example, Seo *et al.* (2000) focus on a set of quantiles, and National Oceanic and Atmospheric Administration (NOAA) defines PQPF as a "form of QPF that includes an assigned probability of occurrence for each numerical value in the forecast product". We adopt a comprehensive approach based on the specification of a predictive Cumulative Distribution Function (CDF) which respects the nature of the phenomenon allowing flexibility in modelling both the exceedance of the zero threshold and the continuous distribution for positive amounts (as in Sloughter *et al.* 2007).

This Chapter investigates how to evaluate (Section 4.1) and communicate (Section 4.2) PQPFs, since some modifications to the standard tools are necessary for handling the positive probability of no rain. The procedures we propose can be adopted regardless of the method used for obtaining the predictions and of the temporal or spatial structure of the predicted values. In Chapter 5 we will show their application to the predictions obtained with the model developed in Chapter 2, but they are not exclusive for that specific model; the only fundamental assumption, which we consider suitable in most cases when dealing with precipitations, is the two-part semicontinuous structure of the predictive distribution with a spike of probability on zero (see (4.1) in the following).

We point out that the focus of this Chapter is on univariate predictions, which are compared with the observed outcomes. On the other hand, the purpose of investigating how to assess field forecasts, with main attention to the spatial aspect, was addressed by the Spatial Verification Method Intercomparison Project (ICP; Gilleland *et al.* 2009); it provided an insight on the available methods, which are classified as filtering methods (neighbourhood or field separation) and displacement methods (feature-based or field deformation). Such tools essentially look at the prediction in its totality and try to detect and quantify displacement and stretching errors avoiding the double-penalisation implied by usual tools. However, our interest here is on the assessment of precisely geolocated predictions. The ground truth is only known in locations where a rain gauge is available; thus, the comparison of predictions with the observed values can only be performed in a sparse set.

Rigorous methods have been proposed (see Chapter 3), but they need for revision when dealing with rainfall. Precipitation distribution is characterized by a potentially positive probability of zero and by skewness on the positive real semiaxis. Such features must be taken into account both in modelling and in presenting and assessing predictive performance. In particular, we focus on the case of rainy precipitations, even if the proposed methodologies are general and can be exploited also for other kinds of precipitation like snow.

Models for rainfall vary according to purposes, cumulation time, available data and covariates, and the chosen framework for the estimation (classical or Bayesian). As explained in Section 2.2.1, a flexible choice for addressing the positive probability of no rain and the asymmetry of the distribution of the positive amounts is the two-part model. It corresponds to specifying a likelihood with a CDF in the following form:

$$F(y) = P(Y \leq y) = P_0 + (1 - P_0)\, F_c(y) \tag{4.1}$$

where P_0 is the probability of no rain and F_c is the CDF of a continuous random variable with support on the positive real semiaxis.

For what concerns the probability of rain, numerous modelling choices are available. Probit or logit regressions are suggested when significative covariates are available; tobit or other approaches involving left censoring of a real valued distribution belong to our framework as a special case, corresponding to imposing a constraint on modelling dry locations.

Gamma distribution has often been suggested as suitable for modelling positive rainfall amounts (see for example Sloughter *et al.* 2007), sometimes showing a slight superiority over the competing Lognormal (see Bruno *et al.* 2014); left-censored GEV constitutes an appealing alternative allowing for flexibility in modelling high quantities due to its heavy tail (see Scheuerer 2014). While some approaches, like EMOS for ensemble postprocessing, impose that the predictive distribution is of the form (4.1) specifying a single parametric distribution for the positive amounts, this is not a general rule: for example, Bayesian Model Averaging output is a mixture of distributions, while in complex Bayesian hierarchical models the analytic expression of the posterior predictive distribution may not be available at all. Nevertheless, in all of these cases a predictive distribution of the form (4.1) emerges after assuming a two-part semicontinuous model.

4.1 Forecast evaluation

Due to the mixed discrete-continuous nature of the distribution, care is required when assessing performances of rainfall prediction. In particular, specific tools are needed for addressing rainfall occurrence; moreover, for what concerns the whole distribution, some of the tools presented in Chapter 3 cannot be exploited when dealing with precipitation, due to the positive probability of zero. Modifications are proposed in this Section in order to guarantee the correctness of the evaluation procedures when a two-part semicontinuous model with a spike of probability on zero is employed. If useful in other contexts, generalization of the proposed procedures for a two-part semicontinuous specification with the spike on a non-zero value is straightforward.

4.1.1 Probabilistic forecasts

Graphical tools for evaluating rainfall probability

A first goal of precipitation predictions is to assess the presence or absence of rain. When probabilistic forecasts are available, graphical tools can be exploited for investigating the ability in discriminating between the two cases. **Reliability diagrams** (or calibration curves, Sanders, 1963; Pochernich 2009) consist in plots of the Hit Rate computed on successive bins of predictive rainfall probability: the conditional observed frequency of rainfall is plotted against its forecasted probability. For perfect reliability the forecasted probability and the frequency of occurrence should be equal, resulting in points lying on the diagonal.
Sharpness diagrams consist in the histogram of predicted rainfall probability: high frequencies of probabilities near zero or one denote a bold forecast characterized by a high level of certainty, which is a preferable attitude under calibration. Both tools are provided by the function "verify" in R package "verification" (Pochernich 2009).

Graphical tools for assessing probabilistic calibration: the PIT histogram

PIT histogram for rainfall predictions needs to share features from both the discrete and the continuous formulations, due to the mixed nature of precipitation distribution, which is made up of a probability mass at zero and a continuous density for positive amounts. An easy explanation of why the standard PIT for continuous distributions cannot be exploited in the two-part semicontinuous case relies on the fact that, while the CDF in the former case assumes values in [0,1], in the latter it only takes values in $[P_0, 1]$, where

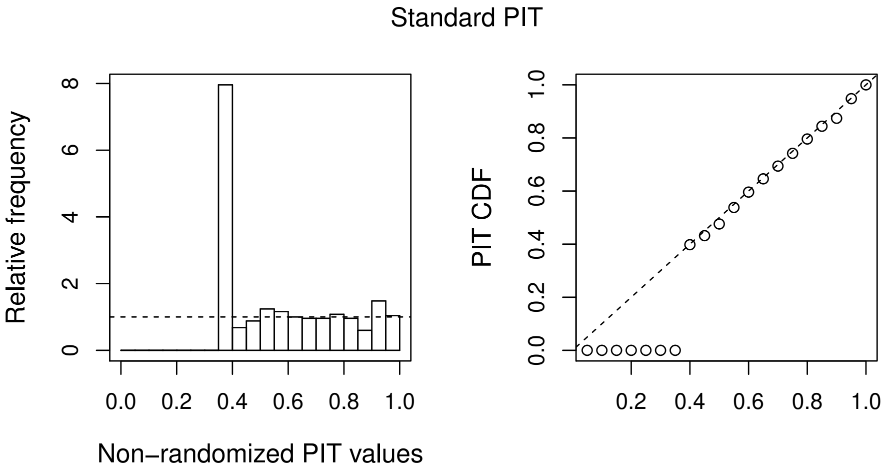


Figure 4.1: Standard PIT for continuous distributions applied to a two-part model with Gamma specification on positive values, with P_0=0.4, shape=2 and scale=0.2.

P_0 is the probability of zero; therefore, values of the CDF smaller then P_0 appear as a cumulated peak on 0 in the PIT histogram, leaving the interval $[0, P_0)$ empty. An example of this phenomenon is shown in Figure 4.1. It shows the PIT histogram and the CDF of the PIT in a simulated example: we took the forecast coinciding with the true data generating process, but standard PIT is not uniform.

A correction in correspondence to zero observations can be applied either with a randomized or a non-randomized approach. The former solution simply randomizes the PIT in zero, substituting the predicted probability of zero P_0 with $v\, P_0$ where $v \sim U[0, 1]$; in practice, this corresponds to draw the value of PIT in zero from a uniform distribution on $[0, P_0]$. In this way, the peak on P_0 of the PIT CDF evident in the left panel of Figure 4.1 is redistributed on [0,P_0], as shown in Figure 4.2.

Following a non-randomized approach instead, the whole CDF of the PIT is to be computed for each observation: when $y = 0$ the proposal of Czado *et al.* (2009) is maintained (see Section 3.1.1), while in correspondence to positive observations the degeneration of (3.3) to a step function is considered:

$$F^{PIT}(u|y, y=0) = \begin{cases} 0 & u \leq 0 \\ u/P_0 & 0 \leq u \leq P_0 \\ 1 & u \geq P_0 \end{cases} \qquad F^{PIT}(u|y, y \neq 0) = \begin{cases} 0 & u < F(y) \\ 1 & u \geq F(y) \end{cases} \tag{4.2}$$

We recall from (4.1) that F is the predictive CDF of the whole likelihood of the two-part semicontinuous model. The mean PIT is thus obtained by aggregating over the observations $\{y_1, \ldots, y_n\}$ and the respective PIT CDFs $\{F^{PIT}(u|y_1), \ldots, F^{PIT}(u|y_n)\}$:

$$\bar{F}(u) = \frac{1}{n}\sum_{i=1}^{n} F^{PIT}(u|y_i), \quad 0 \leq u \leq 1. \tag{4.3}$$

Then, a number J of bins for the PIT histogram is chosen, usually 20, and the frequencies f_j, $j = 1, \ldots, n$ are calculated according to (3.5) as suggested in Czado *et al.* (2009). As already explained, while the randomized PIT slightly changes according to the draws of the U variable, the non-randomized PIT provides a fixed and reproducible diagnostic tool, not affected by random fluctuations. Figure 4.3 reports the non-randomized PIT for the previous

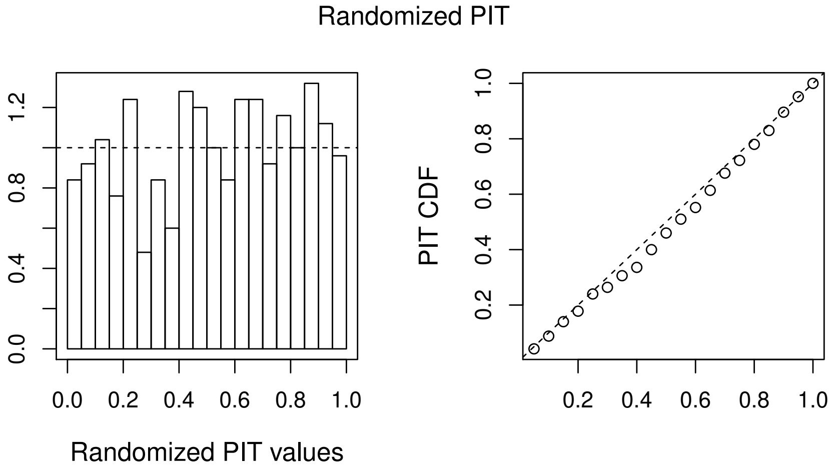


Figure 4.2: PIT randomized in zero applied to a two-part model with Gamma specification on positive values with P_0=0.4, shape=2 and scale=0.2.

simulated example.

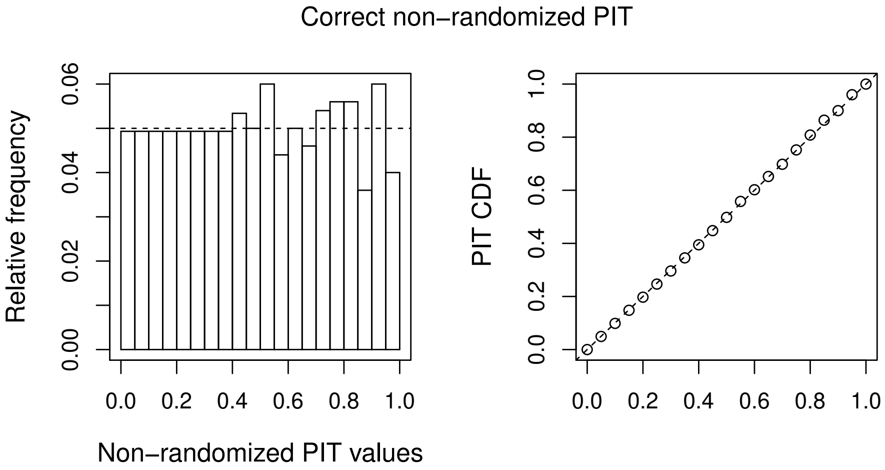


Figure 4.3: Non-randomized PIT applied to a two-part model with Gamma specification on positive values with P_0=0.4, shape=2 and scale=0.2.

Numerical tools for assessing probabilistic calibration: proper scoring rules

In this paragraph we review the proper scoring rules examined in Section 3.1.2 for the special case of precipitation. When studying this phenomenon, a proper Lebesgue density cannot be defined due to the positive probability on zero; nevertheless, the Logarithmic Score can be computed by taking $f^{pred}(0) = P_0$ and $f^{pred}(y^{obs}|y^{obs} \neq 0) = F'(y^{obs})$, with F being differentiable on the positive semiaxis.

In contrast, Dawid and Sebastiani score is not suitable for assessing rainfall predictions, since the two-part semicontinuous nature of predictive distributions allows to assign probability one to the value zero, thus making the variance-covariance matrix not invertible.

Brier Score is widely used for verifying the ability in detecting the presence of rain ($\Omega = (0, +\infty)$). Its integral over all possible thresholds, i.e. the continuous ranked probability score, is the most recommendable scoring function for assessing overall predictive performances. Due to the domain of precipitation distribution, the support of the integral is reduced to the positive semiaxis:

$$\mathrm{CRPS}(F, y) = \int\limits_{[0,+\infty)} \mathrm{BS}(F, \{y \leq x\})\, dx = \int\limits_{[0,+\infty)} (F(x) - I_{\{y \leq x\}})^2\, dx. \quad (4.4)$$

Brier Score plot and Quantile Decomposition plot provide useful deeper insights into CRPS results.

4.1.2 Point forecasts and predictive intervals

Precipitation distributions are asymmetric, skewed and subject to the presence of outliers. The mean is thus not a good choice, since it lacks in robustness; therefore, even though the MSE can be useful for giving more weight to higher errors, in case of precipitation forecasting the median can be a better synthesis than the mean. In any case, the use of MSE and MAE computed on the predictive mean and median, respectively, is correct, since they are consistent with these functionals also in the two-part semicontinuous case. The use of quantiles is widespread in the field of precipitation forecasting. Consistent scoring functions are proposed in Section 4.1.2 and can also be used without the need for modifications. In many meteorological applications skill scores are used for normalizing results with respect to a reference method. Nevertheless, skill scores are in general not consistent; it is thus recommendable to rely on the raw scores, which also maintain full information about the results.

The assessment of predictive intervals for precipitation is challenging. In particular, the computation of coverage requires care, due to the mass of probability on zero. In fact, when dealing with continuous distributions, the inclusion of the extremes of the interval has no interest since the probability of an observation being equal to one of them is always zero. In case of precipitation forecasting instead, the probability P_0 of observing zero is positive, and the lower bound of credibility intervals is often zero. This generates troubles when calculating the coverage: it is necessary to define whether the left extreme is included or not, and if the wrong decision is taken, the coverage will not equal the nominal level α even when the predictive distribution is the true one. For example, focussing on centered intervals, the inclusion of the left extreme in the credibility interval implies the inclusion of the left tail, which corresponds to a probability of $(1-\alpha)/2$ that should not be ascribed to the interval.

One possible solution is the randomization of the inclusion of the left extreme when it coincides with zero. This corresponds to compute the percentage P_{0in} of zero observations that must be ascribed to the interval; then, a Bernoulli variable with probability P_{0in} of "1" exit is extracted, and only the observations corresponding to a 1 exit are ascribed to the credibility interval. More precisely, when computing the coverage of a centered interval of level α, three different situations can occur:

a) if $P_0 \leq (1-\alpha)/2$, then the confidence interval entirely falls into the continuous part of the density; thus no problems arise and the coverage is computed as usual as the percentage of observations falling into the interval;

b) if $(1-\alpha)/2 < P_0 < 1-(1-\alpha)/2 = (1+\alpha)/2$, then $P_{0in} = (P_0 - (1-\alpha)/2))/P_0$, in order to remove the left tail;

c) if $P_0 \geq (1+\alpha)/2$, then the interval degenerates to zero, and the discrete peak of probability on zero also includes both the left tail and part of the right tail of the distribution that should be left apart; for this reason, $P_{0in} = (P_0 - (1-\alpha)/2 + ((1-\alpha)/2 - (1-P_0)))/P_0 = \alpha/P_0$.

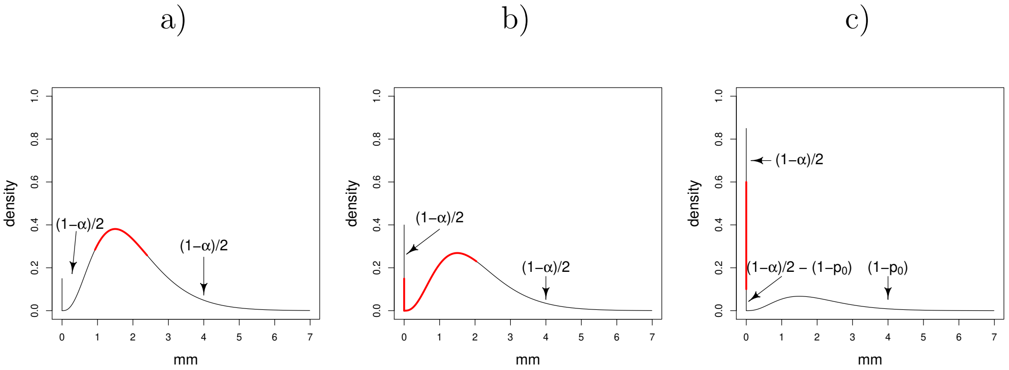


Figure 4.4: Issues arising when computing the coverage of a centered interval in case of a two-part model with Gamma specification on positive values: attention must be put to the left tail in cases b) and c) (blue), and to the right tail in case c) (red and green).

Nevertheless, randomization procedures provide non-fixed results; we rather suggest a non-randomized approach based on non-randomized PIT. Coverages in fact can be computed through calculation of the area underlying the desired portion of [0,1]. For example, in case of centered intervals of level α, the coverage can be recovered as

$$\text{Coverage}_\alpha = \overline{F}\left(\frac{1+\alpha}{2}\right) - \overline{F}\left(\frac{1-\alpha}{2}\right). \tag{4.5}$$

Finally, in case of precipitation forecasting, the assessment of the performances of a model in terms of lower prediction intervals coincides with checking the ability in predicting quantiles; as showed in Section 4.1.2, this can be done using piecewise linear scoring functions.

4.2 Communicating forecasts

A crucial issue for every forecaster is the communication of results. Predictions can be provided in the form of point or probabilistic forecasts, according to specific needs; in the former case, a quantification of uncertainty is fundamental, while in the latter it is necessary to establish guidelines explaining how to efficiently illustrate and summarize the distributions. In 2005 in USA the National Weather Service (NWS) commissioned the National Research Council (NRC) to provide recommendations for effective estimates and communication of uncertainty in weather and climate forecasts; this resulted in the publication of "Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts" (2006). Interdisciplinary studies, involving experts in meteorology, physics, statistics and psychology, have been developed in the following years for providing reliable predictions in easy-to-understand formats. The University of Washington Probability Forecast project (PROBCAST, see Mass *et al.* 2009) was a test bed for exploring the best approaches for communicating high-resolution uncertainty information to a large and varied user community. Its web site (online at www.probcast.com) is the front end of a sophisticated modelling and postprocessing data system, and provides an easy-to-interpret interface for the probabilistic forecasts.

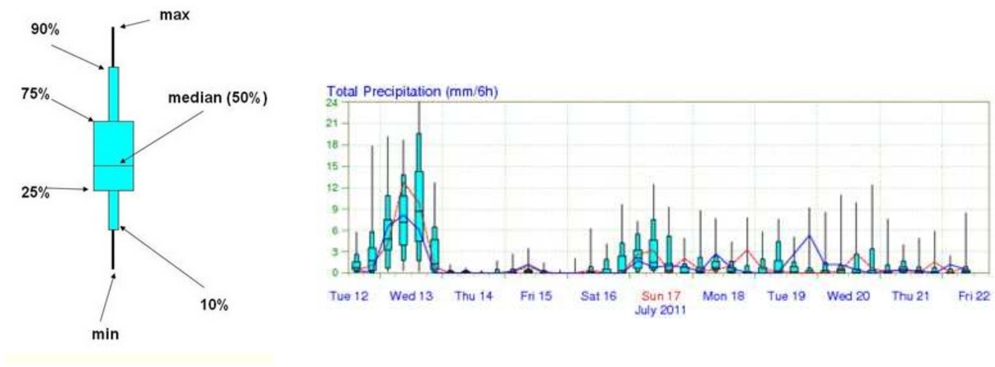
In this Section, we briefly revise main achievements in the field of rainfall forecasts communication; probabilistic, point and interval forecasts are investigated.

4.2.1 Predictive distributions

The quest for good probabilistic forecasts has become a driving force in meteorology over the past two decades (Gneiting and Raftery, 2005). As sketched in Chapter 1, models for precipitation forecasts can be developed and estimated in the frequentist and in the Bayesian framework. In both cases, the mixed discrete-continuous nature of the distribution often encourages to split modelling (and thus the estimation and the prediction) in two steps: firstly, the presence or absence of rain is assessed; then, in the lo-

cations and instants in which it is raining, the positive amount of rain is analysed. We remark that an appropriate Bayesian approach is preferable since it ensures to include all the steps in a comprehensive framework and to keep track of all the uncertainties, through the use of full distributions (also for the parameters).
When ensemble forecasts are produced, they can be seen as (small) samples from the predictive distribution. Some postprocessing techniques work individually on the ensemble members, thus leaving the dimensionality unchanged (see for example Diomede *et al.* 2013); on the other hand, most statistical procedures, like BMA and EMOS, perform calibration of the whole ensemble turning it into a predictive distribution, whose analytical form is known. The present work focusses on this second case.
Thus, in each location and each time instant, the output consists in a predictive distribution, which is available as a big sample or analytically. Before trying to summarize it, statisticians have the task of communicating the customers the richness deriving from the availability of a whole distribution: every aspect of the phenomenon can be investigated and evaluated when the complete information about the outcome, and its uncertainty, is provided. Predictions are made on uncertain events, and thus can not be deterministic by definition; awareness of the necessity of probabilistic forecasts is growing (see Gneiting and Katzfuss 2014 for an overview) but still needs to be strengthened. When possible, communicating the whole posterior distribution is the best solution. Softwares often provide only some "relevant" summaries; however, what is relevant depends on the customer's interest and, possibly, on the circumstances, since new situations or needs may require new summaries, which can be easily obtained if all information is available. As a general suggestion, it would be useful to always provide a function for calculating quantiles and extracting a sample from the predictive distribution; such tools would also allow a straightforward exploitation and analysis of transformed variables.

When results must be communicated to a wide and diversified public, possibly not possessing a specific education, clear graphical displays can be useful. EPS-grams are the tool ECMWF uses for summarising ensemble predictions: they consist in a sequence of Box and Whisker plots corresponding to successive time instants, showing the median, the 25^{th} and 75^{th} percentiles in the thicker part, the 10^{th} and 90^{th} percentiles in the narrower boxes, and the maximum and minimum (see Figure 4.5). Examples can be found in the official website (`http://old.ecmwf.int/products/forecasts/guide/Ten_day_EPSgrams.html`). Schefzik *et al.* (2013) develop a similar tool in the right panel of their Figure 4, dropping the first and third quartiles but inserting marks denoting the predictions from all ensemble members and the realized outcome. The same additional information is included when plotting the predictive density at a single time instant, and helps the forecasters in understanding how the predictions and their correction via calibration are performing. EPS-grams are a common tool in meteorology. In some cases, successive Box plots are connected by lines highlighting the temporal development of the forecasts, or predictions provided by different sources are reported (see for example the Meteorological Service of Canada website `http://weather.gc.ca/ensemble/naefs/EPSgrams_e.html`). Similar visualization tools have been developed for functional forecasts; in that case, a definition of the rank of the curves is needed in order to provide confidence



(a)

(b)

Figure 4.5: Structure of a Box and Whisker plot (a) and an EPS-gram (b) referring to Strasbourg, 12-22 July 2011; figures are taken from ECMWF website.

bands (see for example Sun and Genton 2011).
As outlined in the introduction, our main interest relies in univariate predictions; nevertheless, some suggestions for displaying predictions corresponding to different locations at the same time can be useful. A map with geolocated Box and Whisker plots can be drawn, but it will result too crowded in case of many analysed locations, and even illegible in case of grid predictions. In this case, we suggest to split the EPSgrams information in three or four different maps (for each leading time), reporting a relevant quantile (.05, .5, .95) or the mean of the predictive distribution. PROBCAST website for example produces a variety of forecast maps. Moreover, determining the best way to communicate the probability of precipitation was a main task of PROBCAST project. Web-based surveys and psychological studies (see for example Morss *et al.* 2008) revealed that users benefit of direct access to uncertainty information which prevents them from relying on their own subjective estimates of uncertainty. Extensive testing of several visual presentation formats revealed that interpretation errors are reduced when the chance of observing zero is made explicit, i.e. reporting the probability of facing a dry hour together with the one of rain occurrence; in fact part of the public is not familiar with the concept of probability, but a simple and complete communication allows to deliver information about the uncertainty of rain occurrence.

4.2.2 Point forecasts and predictive intervals

Despite the superiority of probabilistic forecasts, many practical situations require single valued point forecasts. At the same time, uncertainty estimates associated to forecasts are fundamental for a correct assessment and understanding of predictions. Moreover, explicit uncertainty information, both in the form of probability of exceeding a certain threshold and of credibility intervals, benefits users in everyday decision-making allowing to better distinguish between situations in which a target event is likely or unlikely. Psychological tests revealed that people have the background knowledge necessary to understand explicit uncertainty forecasts, being aware of the uncertainty inherent in deterministic forecasts as well as of the factors that tend to increase uncertainty, including lead time and deviations from climatology (see for example Morss *et al.* 2008, Joslyn and Savelli 2010). Uncertainty forecasts, in the form of calibrated predictive intervals for instance, can provide a better understanding of where to expect potential uncertainty and allow users to tailor the forecast to their own tolerance for risk. Moreover, predictive intervals increase trust in the forecast with respect to traditional point forecasts (Joslyn and Savelli 2013). Since users without a specific education might encounter difficulties in interpreting credibility intervals, displays of forecasts should be simple, and simple text format seems to be more effective than visualization tools in this case (Joslyn *et al.* 2013). The concept of complementary probability is not always straightforward, as already explained in the previous paragraph; moreover, people mistrust predictive interval forecasts as they mistrust deterministic forecasts, tending for instance to expand the forecasted range, considering it as an underestimation of the real one. Interestingly, presenting the probability of obtaining a result beyond a certain threshold, or out of a certain interval, turns out to be more effective than defining the probability of falling within a range.

When credibility intervals are communicated, the centered ones are the most common choice. In case of two-part semicontinuous models, HPD intervals can not be easily considered, due to the mixed nature of the distribution. More precisely, HPD intervals would be well defined if the chosen density on positive amounts is decreasing; otherwise, a single point (zero) with positive probability may be part of the "interval", thus requiring an appropriate measure definition. For example, in the case of Gamma specification for positive rainfall amounts, a straightforward use of HPD intervals is possible if the shape parameter is greater then one.

Chapter 5

Results for rainfall spatial prediction in the Emilia-Romagna Region

The three model specifications ("base", "mean" and "SW") presented in Chapter 2 are applied to the 8 selected rainfall events in Emilia-Romagna (see Section 2.1 for details). This Chapter examines the results obtained by exploiting the tools introduced in Chapters 3 and 4. For a more comprehensive understanding of the proposed framework and of the relevance of its spatial connotation, an additional benchmark model is added, denoted by "No Sp" (No Space); it corresponds to a simple model in which the spatial random effects are supposed to be uncorrelated:

$$\boldsymbol{\epsilon}|\sigma^2_\epsilon, \phi_\epsilon \sim MVN(\mathbf{0}, \sigma^2_\epsilon Id), \quad \boldsymbol{\alpha}|\sigma^2_\alpha, \phi_\alpha \sim MVN(\mathbf{0}, \sigma^2_\alpha Id) \tag{5.1}$$

with Id denoting the identity matrix. Equation (5.1) replaces equations (2.4) and (2.8).

The analysis of the results is organized as follows. Section 5.1 provides an overview of the obtained estimates for the most interesting parameters; more precisely, the role of radar in the regressions in the various model specifications is investigated, and examples of maps of the spatial random effects are provided. Sections 5.2, 5.3 and 5.4 are devoted to the assessment of the predictive performances on 50 randomly selected validation sites for each hour;

probabilistic, point and interval forecasts are evaluated and compared, and some reconstructed rainfall fields are shown.

5.1 Analysis of relevant model components

In this section, an insight into the estimates for some coefficients is provided. Our motivating problem is rainfall prediction. Therefore, investigation of the single model components is intended as a tool for achieving a better understanding, which may help for future development. After verifying that convergence has been reached by all the posterior chains (see Section 2.3 for details), precise knowledge of the value of all the parameters is not of interest, also due to the difficulty in their interpretation deriving from the complexity of the hierarchical structure. We focus on understanding the influence of radar in the various steps of rainfall reconstruction. The aim is to gain knowledge about the way in which the several model specifications are able to exploit radar information, in order to highlight interesting behaviour or trends. We recall from Section 2.2.1 and 2.2.2 that the coefficients γ_1 and γ_2 quantify the influence of the logarithm of radar in the probit regression for rainfall probability, while coefficients β_1 and β_2, and β_3 in Model “mean” determine the effect of radar on rain accumulation in the log scale; remember they are all hour-specific. Our Bayesian approach provides whole posterior distributions for these coefficients, carrying information about their uncertainty. We choose boxplots to summarize their behaviour for each of the 69 analysed hours. We specify that each of the 8 chosen events corresponds to a block of successive hours, but events are separated one from the other; thus, a temporal evolution of the coefficient might be followed within an event but not across them.

Figures 5.1 and 5.2 show the effect of radar on rain probability in Models “No Sp” and “base”; results for Models “mean” and “SW” are not reported here since they coincide with Model “base” for what concerns rain probability. In fact, all the four model specifications share Equation (2.3), but while “base”, “mean” and “SW” provide for spatial correlation in the random effects according to Equation (2.4), the benchmark Model “No Sp” assumes $\boldsymbol{\epsilon}$

is uncorrelated, following Equation (5.1). Both figures show the coefficients are significant, with boxplots non crossing the axis $y = 0$ (apart form the last one), reported as a dashed horizontal line. Differences are evident between the events, but also within them, confirming the choice of modelling each hour separately was appropriate: a common event-specific coefficient would hide precious hourly variability. Imposing a temporal evolution for the coefficient within an event might be an interesting attempt in the future, even if the boxplots reveal a non-negligible irregularity, with some hours performing very differently for the previous and successive ones. Model "base" provides sharper posterior distributions if compared to Model "No sp". Spatial correlation thus turns out to be a relevant feature; capturing it via Equation (2.4) helps in ascribing more precise coefficients to the covariate containing radar information. The effect of the inclusion of correlated spatial effects on the value of the coefficients is not uniform in time. In some hours, like the second one of Event E2, a larger γ_2 in Model "base" denotes an enhancement of the role of radar; nevertheless, in the majority of the cases, the values of radar coefficient is smaller in the more structured models where the spatial random effects $\boldsymbol{\epsilon}$ are able to capture a lot of information, causing a reduction of radar relevance as a side effect. Later in this Section, examples of maps of the random effects are provided; their complex spatial patterns in Models "base", "mean" and "SW" confirms their cardinal role in modelling rainfall probability, while in Model "No Sp" they only consist in random errors which would not be able to drive the modelling of rainfall probability. We specify that the decrease in the width for γ_2 intervals is not compensated by an increase in the overall uncertainty, as shown in the following (Section 5.4).
We finally notice the coefficients are all positive, as expected, reflecting a positive correlation between rain probability and radar measurements.

In a similar fashion, Figures 5.3, 5.4, 5.6 and 5.5 show boxplots of the coefficient for the logarithm of radar in the regression for rainfall accumulation. Comments for Models "No Sp", "base" and "SW" are similar to the previous ones, with coefficients generally being different from zero, and showing differences across and within events. Also in this case, the coefficients are positive. A decreasing pattern, sometimes preceeded by a short peak, might be driven by the evolution of the rainfall event. Future development of the

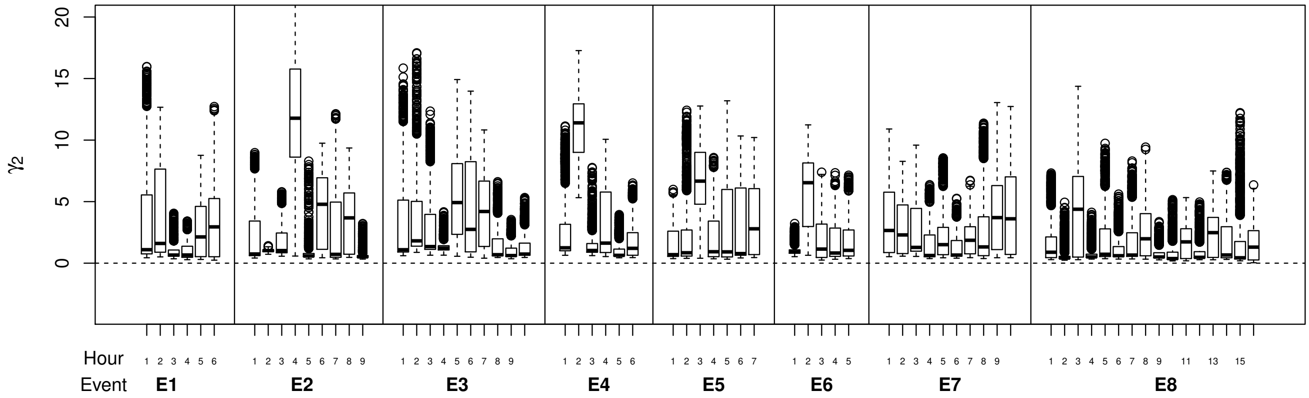


Figure 5.1: Boxplots of hourly coefficient γ_2 obtained with Model “No sp”.

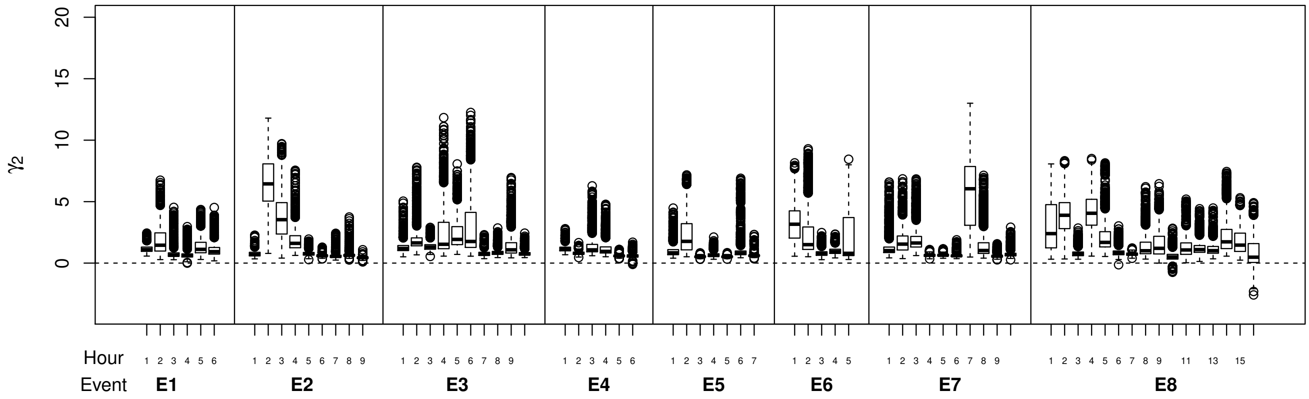


Figure 5.2: Boxplots of hourly coefficient γ_2 obtained with Model “base”.

work may try to address temporal modelling of the parameters; such attempt would be not trivial, due to the different behaviours in the various events, with E2, E5 and E6 being quite irregular.

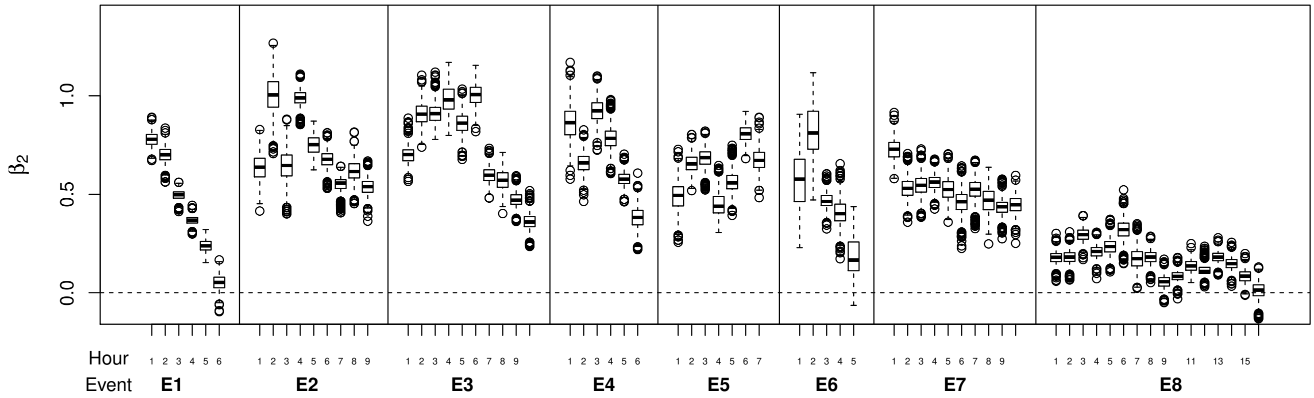


Figure 5.3: Boxplots of hourly coefficient β_2 obtained with Model "No sp".

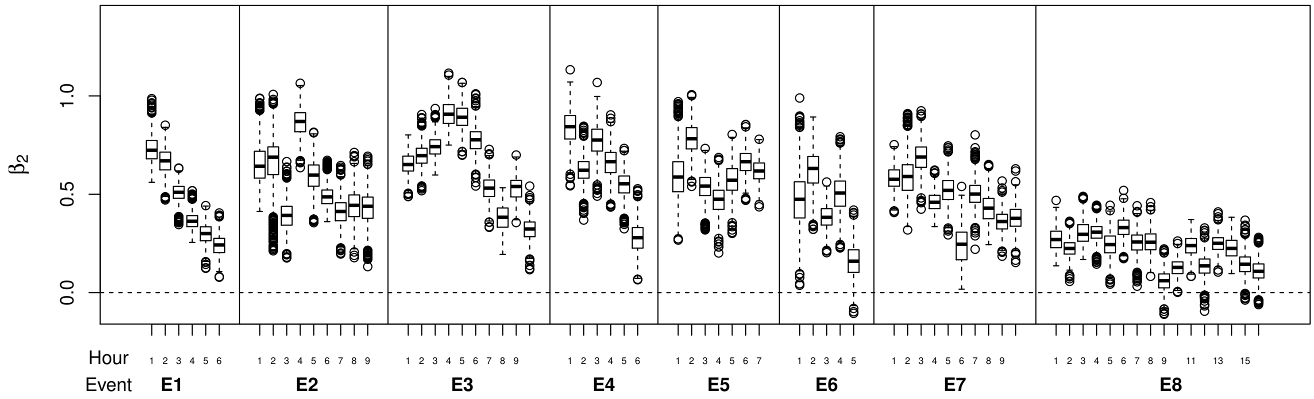


Figure 5.4: Boxplots of hourly coefficient β_2 obtained with Model "base".

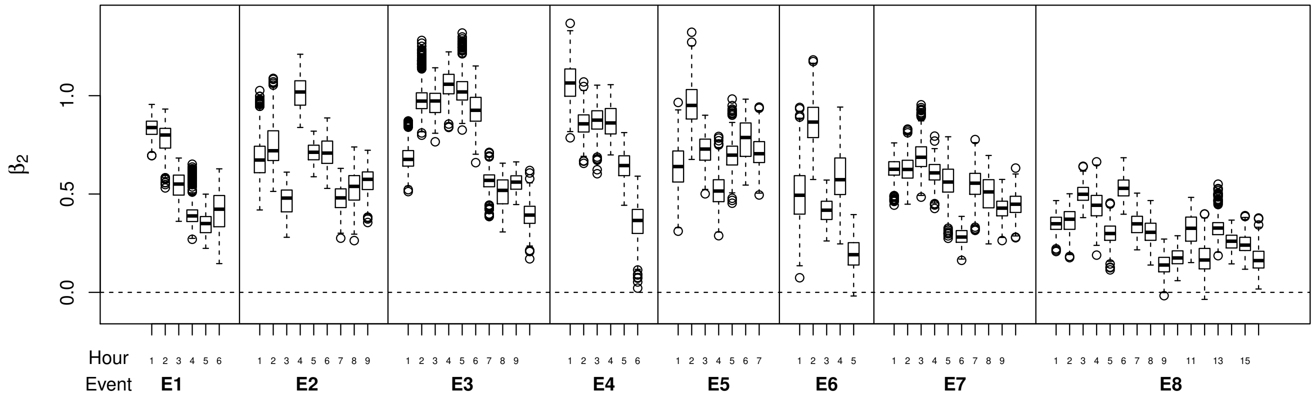


Figure 5.5: Boxplots of hourly coefficient β_2 obtained with Model "SW".

A separate discussion is needed for Model "mean", for which the two coefficients β_2 and β_3 are shown, the former referring to radar value in the pixel containing the location of interest, the latter referring to the mean over the 8 pixels surrounding that location. The boxplots in Figure 5.6 and 5.7 are partly beyond, partly below, and sometimes crossing the x-axis, with a main common feature: when a big coefficient β_2 is estimated for the contingent pixel, the corresponding coefficient β_3 for the neighborhood is small, often negative, and vice-versa; moreover, when one of the two coefficients is not significant, the other one is positive. This behaviour reflects the split of relevant radar information into the two parts, corresponding to the contingent radar pixel, and the mean over its 8 neighbors, respectively. A highly variable behaviour both across and within the events is observed: no general rule can be easily inferred about the predominance of one of the two elements over the other.

Finally, since Model "SW" is the least intuitive but most powerful of the four presented specifications, we provide some examples of the estimated weights for the radar pixels. We recall from Section 2.2.2 that this model takes as a covariate a weighted mean of the raw radar map values; since such weights are stochastic, and driven by a spatial process defined on the whole grid, the weighting scheme is allowed to be asymmetric and also to change across the map. This means that, for example, more weight can be assigned to the pixels standing on the West of a certain location of interest, but when looking at another location (in the same hour) its Northern pixels can be the most relevant ones and gain the biggest weights. Moreover, not only the orientation but also the shape of the weighting scheme can vary across the map. Figure 5.8 shows the posterior mean of the latent Gaussian process Q for the 18^{th} September 2010 at 5 p.m., defined over the radar grid, driving the weights for that hour. Figure 5.9 highlights nine of the fifty validation sites on the radar map. The weights for the radar pixels in the neighbourhoods of such locations is shown in Figure 5.10: it reports the posterior means for the weights of the 49 nearest pixels, the central one containing the validation site (the other pixels in the 49000-dimensional map have a non relevant weight). Notice that the weight of the central pixel significantly changes across the map; in particular, in panel "6" the pixels on the East of the validation site

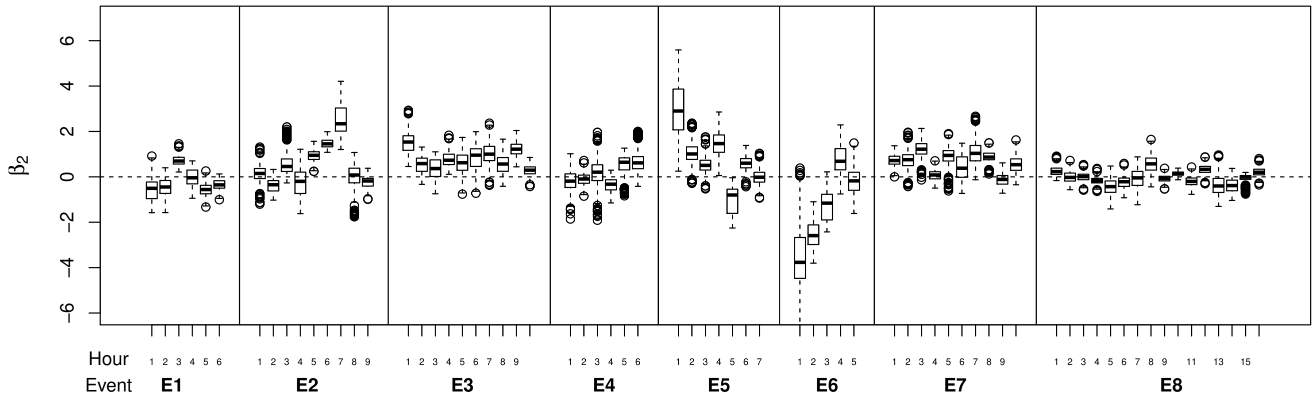


Figure 5.6: Boxplots of hourly coefficient β_2 obtained with Model "mean".

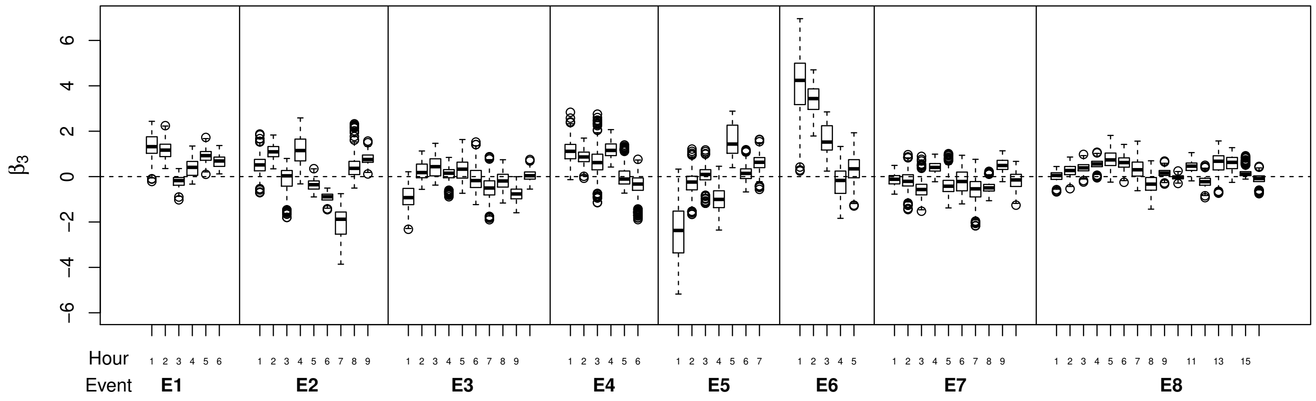


Figure 5.7: Boxplots of hourly coefficient β_3 obtained with Model "mean".

has the same relevance as the central one.

The flexibility ensured by model "SW" can help in correcting errors due to the topographic differences in the region of interest, and potentially to the displacement of rainfall measurements caused by wind. Its positive impact on the predictive performances of the Model is discussed in Sections 5.2 and 5.3.

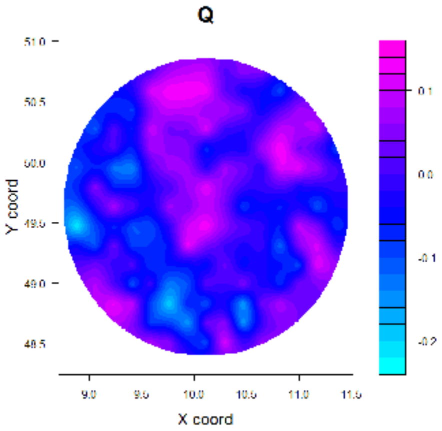


Figure 5.8: Latent Gaussian process Q on 18/09/2010 at 5 p.m.

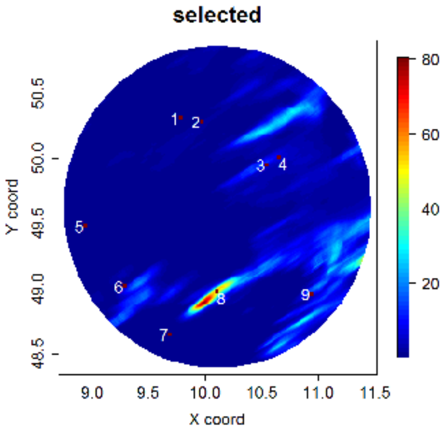


Figure 5.9: Radar map on 18/09/2010 at 5 p.m.; 9 locations are numbered, for which the weighting scheme is reported in Figure 5.10.

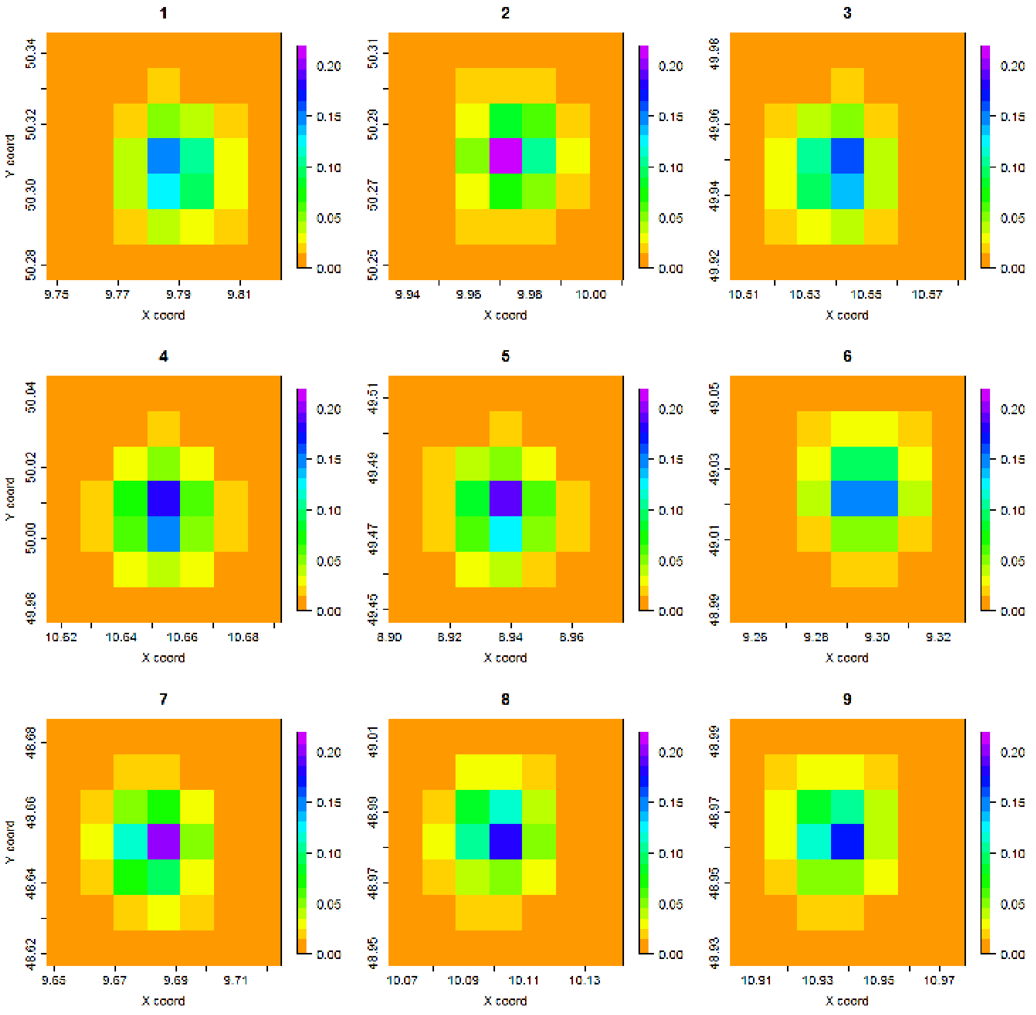


Figure 5.10: Radar weighting on 18/09/2010 at 5 p.m. for the 9 validation sites highlighted in Figure 5.9.

To conclude the Section, we provide some insights into the random effects characterizing the four model specifications. First of all, we recall $\boldsymbol{\epsilon}$ and $\boldsymbol{\alpha}$ are defined on the spatial domain, assuming a random value ϵ_s and α_s on each location s of interest. Therefore, when reconstructing the whole rainfall field, their posterior mean can be visualized via maps; as an example, in Figures 5.11, 5.12, 5.13 and 5.14 we report their estimated mean values in two hours belonging to Event E1 and E3. Rainfall data for these two hours are shown in Figures 5.23 and 5.27 respectively. As already noticed, Models "mean" and "SW" coincide with "base" for what concerns rain probability, thus we only report $\boldsymbol{\epsilon}$ once. According to Equation 5.1, in Model "No Sp" the components of $\boldsymbol{\epsilon}$ are spatially uncorrelated, thus generating random noise, as shown in Figures 5.11 and 5.12. Spatial correlation is instead evident in the panels corresponding to Model "base". The same rationale holds for α, with Model "No Sp" retrieving uncorrelated random noise. The almost uniform maps in the left panels of Figures 5.13 and 5.14 are the result of a unique legend for four models; Figure 5.15 shows the maps of $\boldsymbol{\alpha}$ for Model "No Sp" in the two selected hours with more appropriate colors. Spatial patterns characterize the other three models; differences in the regression determining rainfall amounts, with an increasing exploitation of radar information, have consequences on the spatial effect. In particular, while Models "base" and "mean" are very similar, the patterns for Model "SW" are slightly more irregular and noisy; this is coherent with the role of $\boldsymbol{\alpha}$, which represents the spatial information not explained by the covariates. More precisely, in Model "base" $\boldsymbol{\alpha}$'s task is burdensome, since it adjusts for main local discrepancies between the observed value and the estimated linear transformation of the contingent radar value (in the log scale); when radar influence is more effectively included, as in Model "SW", $\boldsymbol{\alpha}$ should only capture residual patterns and adjust for local fluctuations. Notice that $\boldsymbol{\epsilon}$ and $\boldsymbol{\alpha}$ assume both positive and negative values.

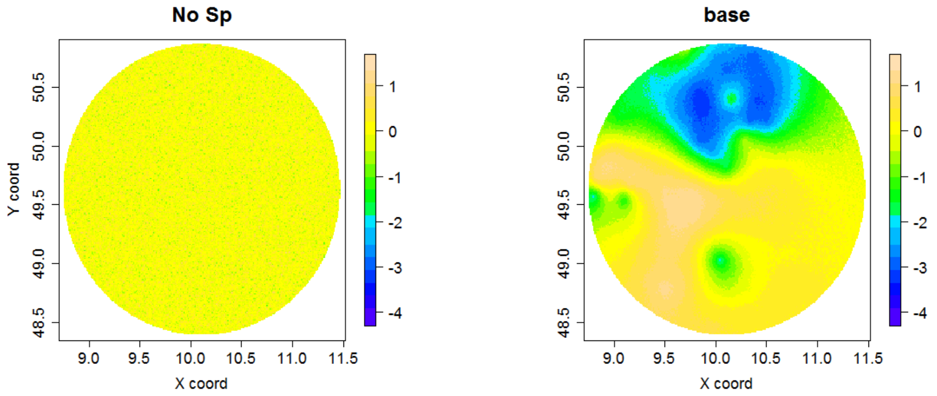


Figure 5.11: Spatial effect for rainfall probability obtained with Models “No Sp” and “base” on 19/09/2010 at 05 p.m.: mean of the posterior predictive distribution of $\boldsymbol{\epsilon}$.

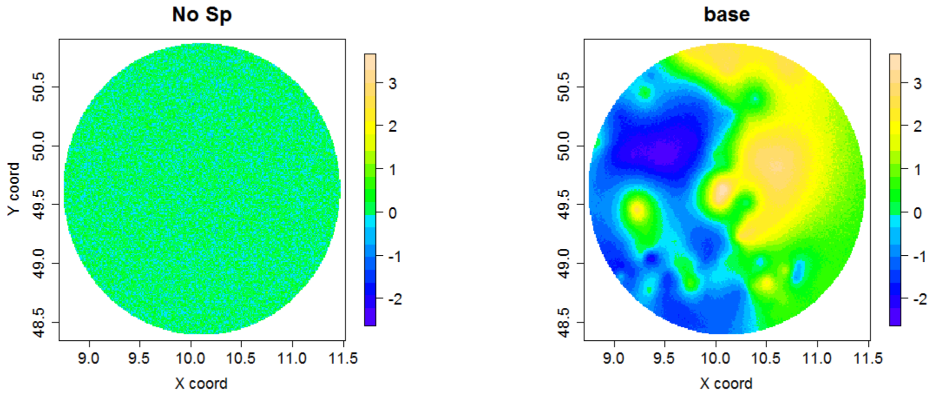


Figure 5.12: Spatial effect for rainfall probability obtained with Models “No Sp” and “base” on 18/09/2010 at 01 a.m.: mean of the posterior predictive distribution of $\boldsymbol{\epsilon}$.

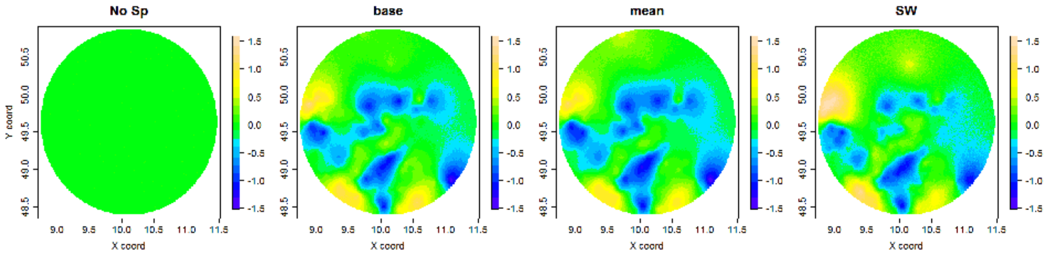


Figure 5.13: Spatial effect for rainfall amounts obtained with the four models on 19/09/2010 at 01 a.m.: mean of the posterior predictive distribution of $\boldsymbol{\alpha}$.

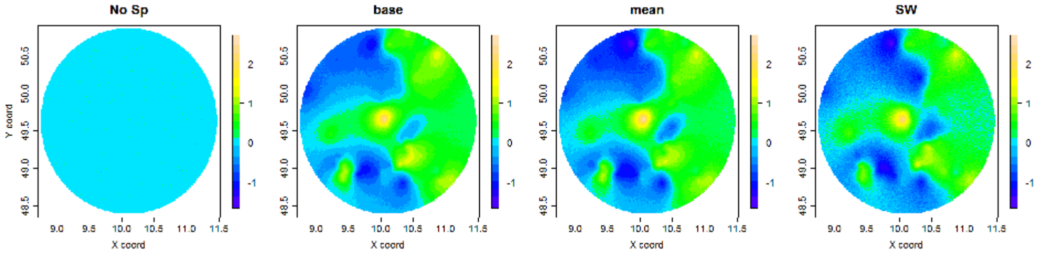


Figure 5.14: Spatial effect for rainfall amounts obtained with the four models on 18/09/2010 at 5 p.m.: mean of the posterior predictive distribution of $\boldsymbol{\alpha}$.

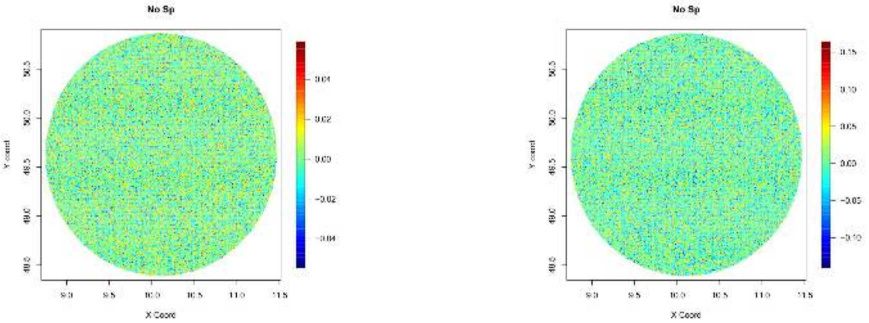


Figure 5.15: Spatial effect for rainfall amounts obtained with Models "No Sp" on 19/09/2010 at 01 a.m. and on 18/09/2010 at 5 p.m.: mean of the posterior predictive distribution of $\boldsymbol{\alpha}$.

5.2 Probabilistic prediction

In this Section and in the following ones, an investigation of the predictive performances is provided by focussing on each of the 8 events individually, preserving the natural grouping of the hours and highlighting the differences in performance in the various cases; moreover, results on the whole pool of hours are shown, in order to give an overall idea of how the different model specifications are performing, and simplify their comparison. In particular, this Section is devoted to the analysis of probabilistic predictions, which are available in the form of a big sample from the posterior predictive distribution, as explained in Section 2.4. Section 5.2.1 focuses on the probability of rain and the prediction of rainfall occurrence, while Section 5.2.2 addresses the whole predictive distributions, investigating probabilistic calibration.

5.2.1 Assessment of rainfall probability

The three model specifications presented in Chapter 2 all share Equations (2.3) and (2.4), thus they all predict the same probability of rain; the benchmark Model "No Sp" instead assumes the random effect $\boldsymbol{\epsilon}$ follows Equation (5.1). Therefore, in this Section we only show the performances of Models "base" and "No Sp" in terms of prediction of rainfall probability (the other two models are equal to "base").
First of all, Figure 5.16 reports the result of function "reliability.plot" of the R package "verification". The reliability plots show consistence of the predicted versus the conditioned rainfall occurrence, with the red dots distributed near the bisector. A deeper insight is given by Figure 5.17, which shows results for each event; Model "base" shows an overall improvement with respect to Model "No Sp". Sharpness histograms are reported in the right-bottom of the figure and reveal a bold (and thus desirable) behaviour, with probabilities mainly concentrated on 0 and 1; this feature is more evident when correlated spatial effects are included in the model.

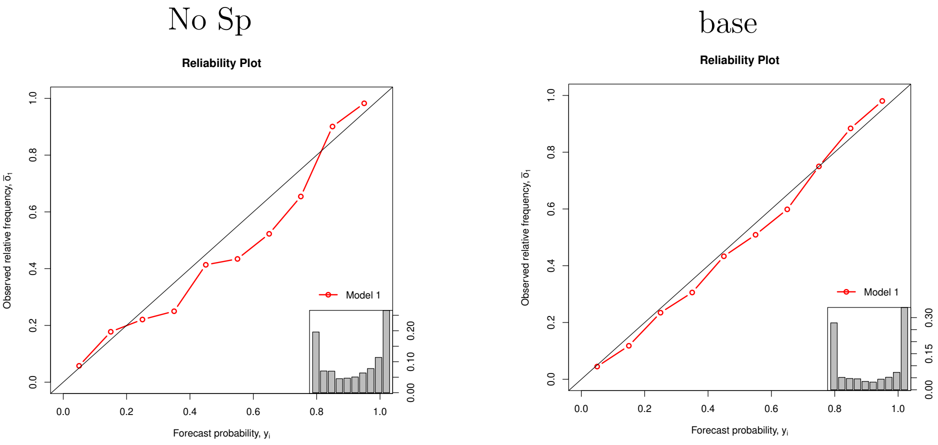


Figure 5.16: Reliability plots obtained with the different model specifications on all the analysed hours pooled together.

Table 5.1 reports the Brier Scores relative to the zero threshold, showing the inclusion of spatially correlated random effects in the model improves the ability in predicting rainfall occurrence, as remarked in Bruno *et al.* (2014).

BS	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	0.099	0.086	0.098	0.083	0.115	0.116	0.108	0.172	0.117
Base	0.079	0.069	0.094	0.070	0.108	0.090	0.091	0.086	0.086

Table 5.1: Brier Score computed on the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours together (tot).

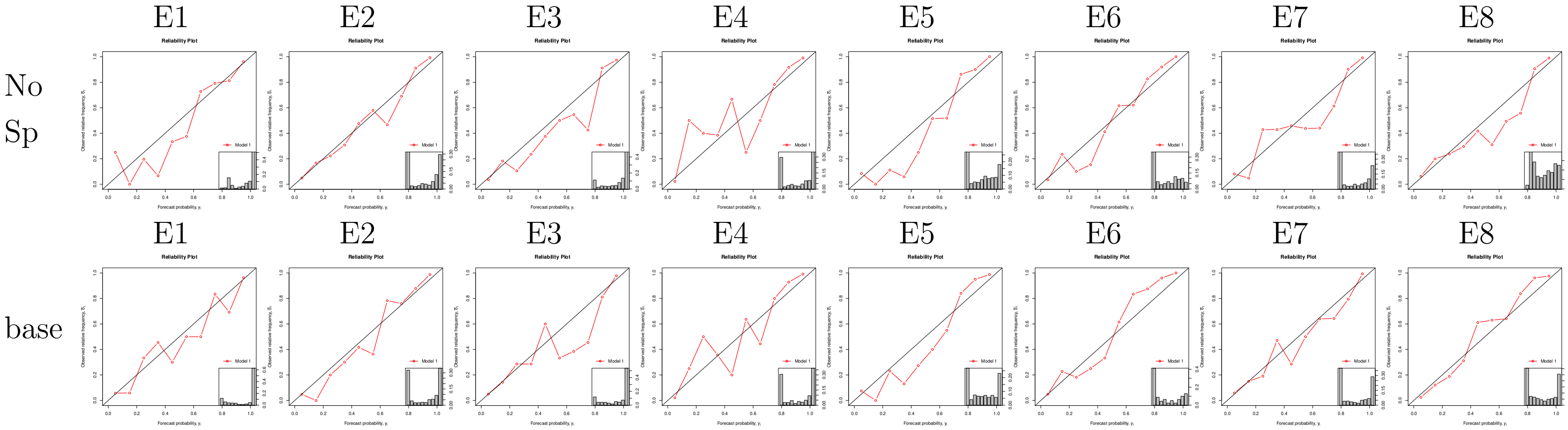


Figure 5.17: Reliability plots obtained with the different model specifications on the 8 Events.

5.2.2 Probabilistic calibration

Probabilistic calibration is achieved, as confirmed by the nearly uniform shape of non-randomized PIT histograms displayed in Figure 5.18. No relevant differences distinguish the four model specifications, neither main discrepancies emerge from the separate analysis of the 8 events (Figure 5.19).

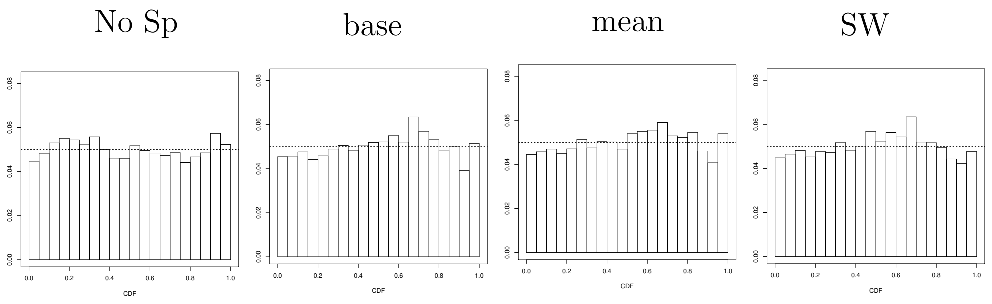


Figure 5.18: Non-randomized PIT histograms obtained with the different model specifications on all the analysed hours pooled together.

Numerical quantification of both probabilistic calibration and sharpness is provided by the Continuous Rank Probability Score (CRPS), which is shown in Table 5.2. It allows comparison between models: spatial correlation of the random effects turns out to be a fundamental feature, determining a noticeable reduction of CRPS when included in the model. The relevance of neighbouring information, in particular when exploited through stochastic weighting, emerges more evidently from Figure 5.20, where the quantile and threshold decompositions of the CRPS are shown; the blue line associated to Model "SW" is lower than the others, denoting smaller values of the Brier Scores corresponding to different thresholds (left-hand panel) and of the Quantile Scores corresponding to several quantile levels (right-hand panel). Main differences around the central quantiles are evident; the superiority of the median obtained with "SW" over the other Models will also be confirmed by the analysis of Mean Absolute Error in Section 5.3. Finally, Figures 5.21 and 5.22 report the Quantile and Brier Score plots analysing the events separately.

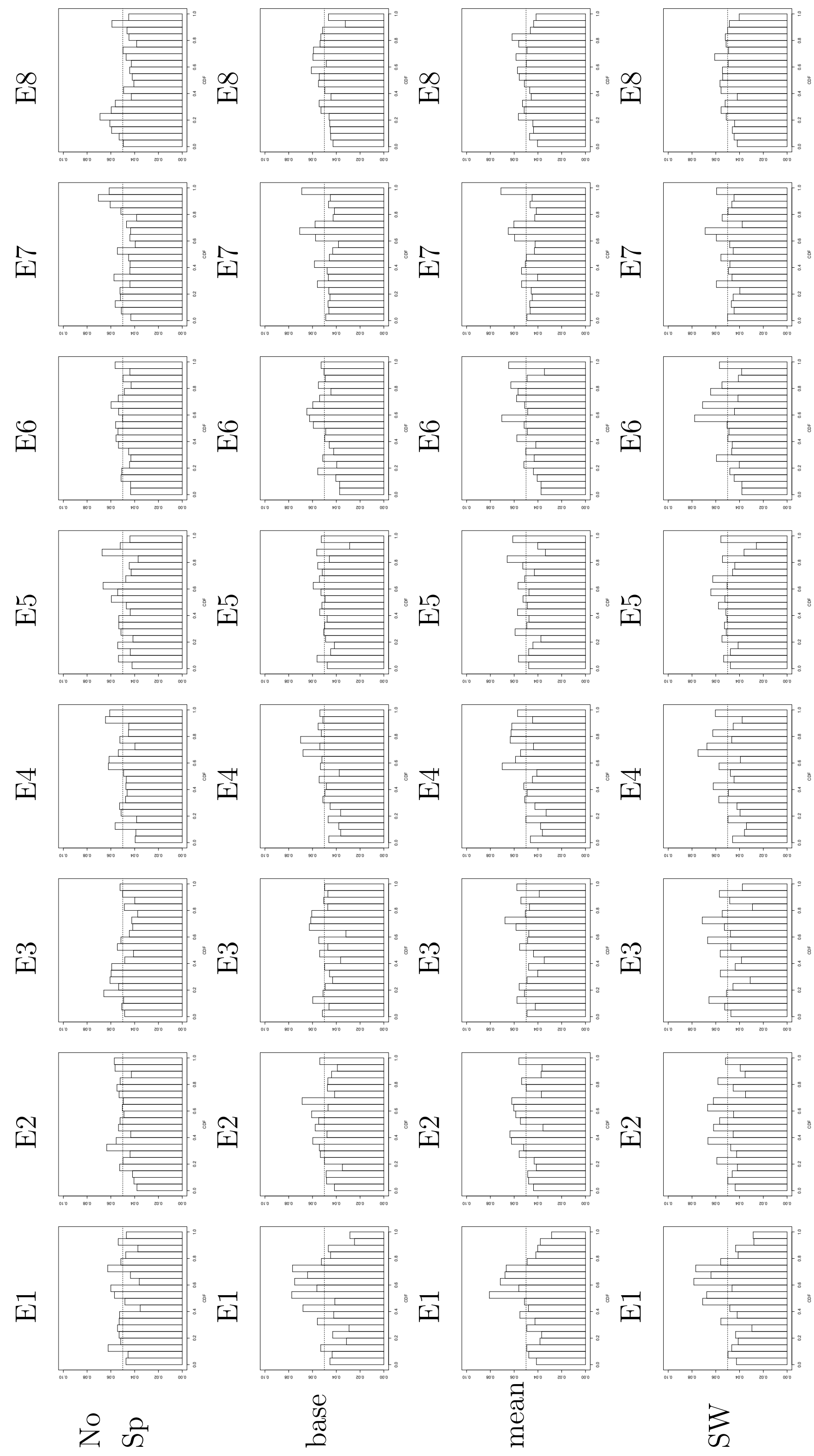
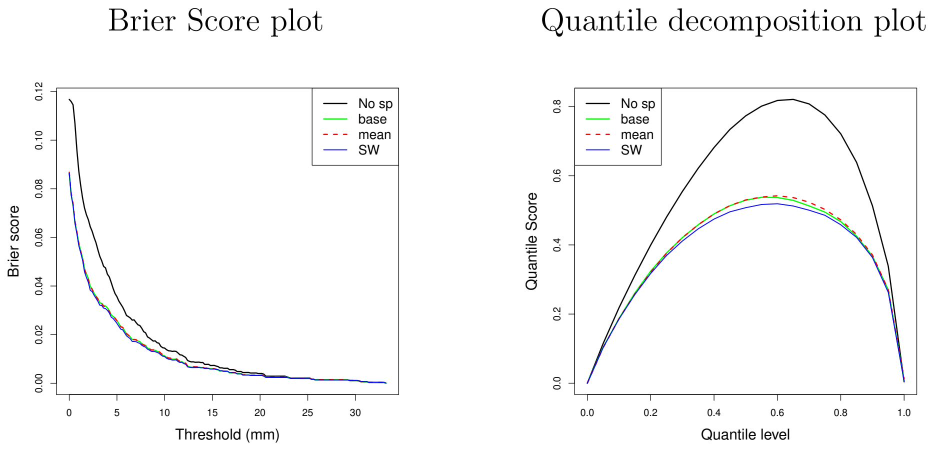


Figure 5.19: Non-randomized PIT histograms obtained with the different model specifications on the 8 Events.

CRPS	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	0.440	1.011	0.594	0.697	0.826	0.262	0.491	0.297	0.560
base	0.249	0.665	0.419	0.563	0.684	0.184	0.360	0.174	0.393
mean	0.248	0.656	0.421	0.553	0.705	0.185	0.363	0.176	0.395
SW	0.251	0.661	0.409	0.519	0.658	0.182	0.348	0.175	0.383

Table 5.2: Continuous Rank Probability Score computed on the 50 selected validation sites separately on each of the 8 events (E1-E8) and on all the analysed events together (tot).


Brier Score plot

Quantile decomposition plot

Figure 5.20: Comparison of the Brier Score plots and quantile decomposition plots obtained with the different model specifications on all the hours and events pooled together.

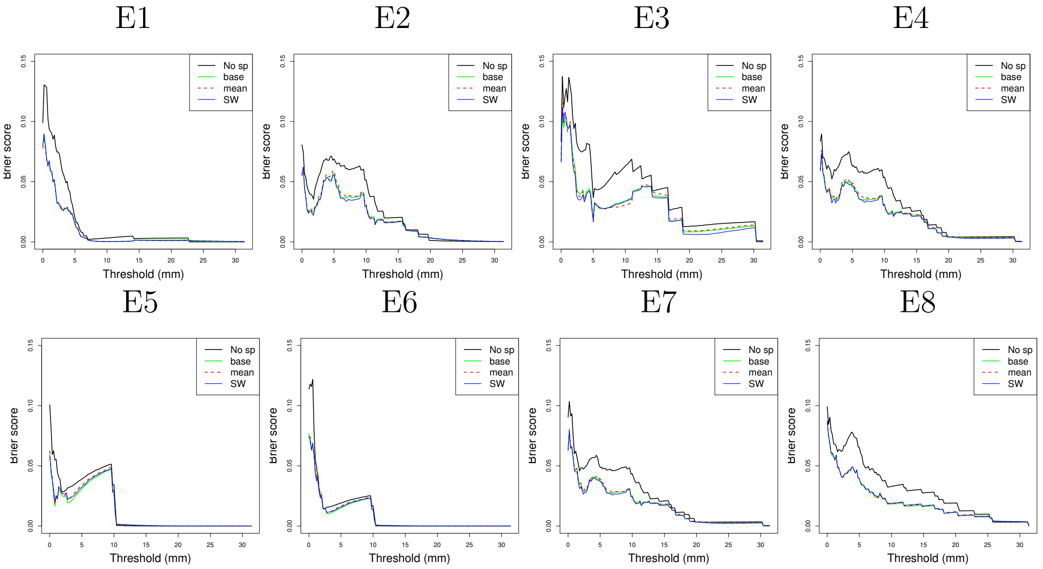


Figure 5.21: Comparison of the Brier Score plots obtained with the different model specifications on the 8 Events.

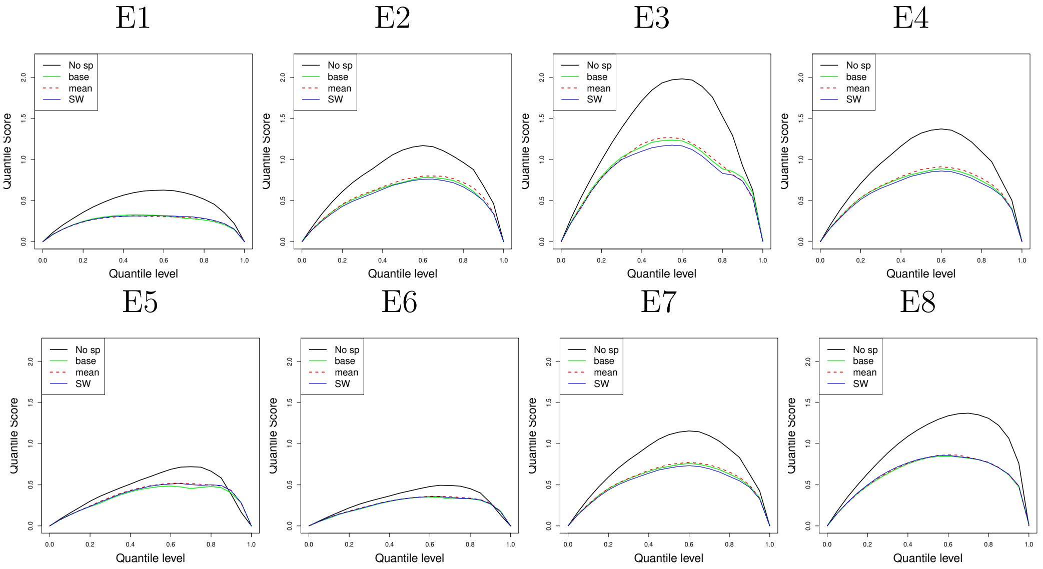


Figure 5.22: Comparison of the quantile decomposition plots obtained with the different model specifications on the 8 Events.

5.3 Point predictions

Both the mean and the median of the posterior predictive distribution are taken as point predictions, and appropriately evaluated via Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), respectively; predictions are compared with the amounts observed by rain gauges on the 50 considered validation sites for each hour. Event-specific and global results are reported, showing overall satisfying performances; we remark that they are expressed in mm. Differences among events reveal difficulties in predicting higher rainfall amounts, most of all in cases in which radar information is far from the gauge measurement. Nevertheless, in all cases radar calibration (with the meaning of correction of indirect measurements, as in Section 1) has been overall successful. The first rows of Tables 5.3 and 5.4 report the scores computed on radar information; more precisely, for each validation site s, we take radar value $R_{P(s)}$ in the pixel containing s as point prediction. Each of the proposed models outperforms raw radar in terms of MAE, and the three models with spatial correlated effects also perform noticeably better in terms of RMSE. The inclusion of neighbouring information turns out to be relevant, with Model "SW" returning the lowest scores, in particular in Events 4 and 5, in which more complex rainfall patterns are observed, due to the convective nature of the phenomenon.

RMSE	E1	E2	E3	E4	E5	E6	E7	E8	tot
radar	1.603	2.928	2.249	3.660	3.457	1.216	2.034	1.222	2.346
No Sp	1.247	3.260	1.717	2.308	2.828	0.878	1.404	0.783	1.925
base	0.589	2.306	1.434	2.052	2.648	0.728	1.077	0.602	1.546
mean	0.595	2.283	1.435	2.032	2.696	0.712	1.076	0.603	1.547
SW	0.626	2.281	1.400	1.939	2.566	0.729	1.039	0.586	1.506

Table 5.3: Root Mean Square Error computed on the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours pooled together (tot).

MAE	E1	E2	E3	E4	E5	E6	E7	E8	tot
radar	0.935	1.400	1.376	1.854	1.540	0.678	0.917	0.630	1.109
No Sp	0.618	1.366	0.817	0.923	1.185	0.366	0.689	0.415	0.773
Base	0.326	0.890	0.571	0.773	0.937	0.232	0.475	0.234	0.529
Mean	0.313	0.880	0.575	0.753	0.966	0.244	0.479	0.240	0.531
SW	0.314	0.883	0.555	0.692	0.860	0.228	0.458	0.234	0.508

Table 5.4: Mean Absolute Error computed on the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours together (tot).

Examples of reconstructions of the rainfall field are provided in Figures 5.24 and 5.28 for the four model specifications on the two hours analysed in detail in Section 5.1; the original rain gauge and raw radar data are shown in Figures 5.23 and 5.27. The maps consist in the means of the posterior predictive distributions in the centroids of radar pixels. A visual comparison shows that Model “No Sp” operates a simplification of the rainfall patterns identified by radar, while the other model specifications are able to smooth radar while preserving the richness of its spatial information. More in detail, Models “base” and “mean” originate very similar results, the latter being slightly smoother, thanks to the additional information about the neighbouring pixels; the map obtained with Model “SW” better succeeds in capturing spatial patterns and high intensities. It is interesting to highlight that the proposed models also allow to predict rainfall in areas where no rain gauges are available, as in the Eastern part of the radar circle which covers part of the Adriatic Sea. Figures 5.25 and 5.29 show the predictive bias in the validation sites. Predictive errors are small and around zero in the former example, while in the latter a point of relevant underestimation is detected, in which radar overestimates rainfalland the model correctly intervenes in reducing its value, but an excessive correction causes underestimation in the nucleus of the storm.

Maps of some quantiles from the predictive distribution for the same two hours are provided in Figures 5.26 and 5.30.

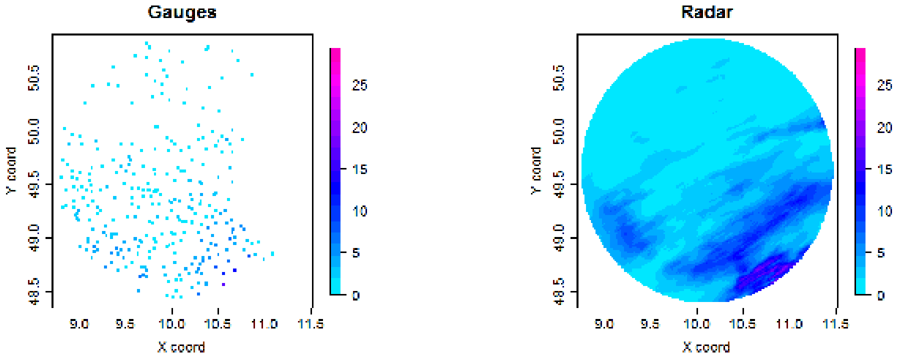


Figure 5.23: Gauge measurements (left panel) and radar grid (right panel) on 19/09/2010 at 01 a.m.

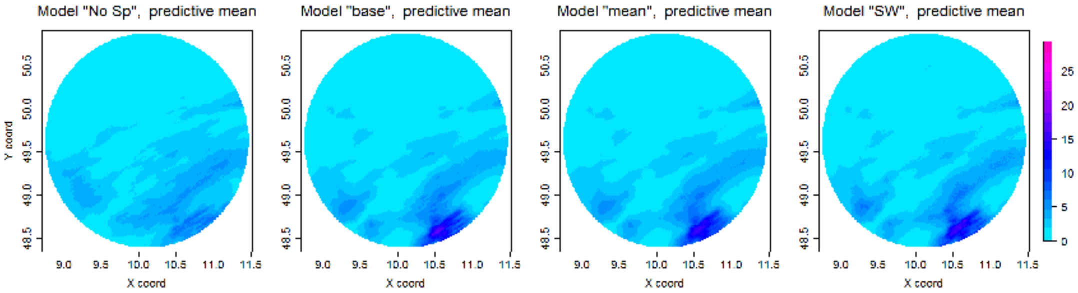

Model "No Sp", predictive mean
Model "base", predictive mean
Model "mean", predictive mean
Model "SW", predictive mean
Y coord
X coord
9.0
9.5
10.0
10.5
11.0
11.5
48.5
49.0
49.5
50.0
50.5
0
5
10
15
20
25

Figure 5.24: Reconstructed field with the four model specifications on 19/09/2010 at 01 a.m.

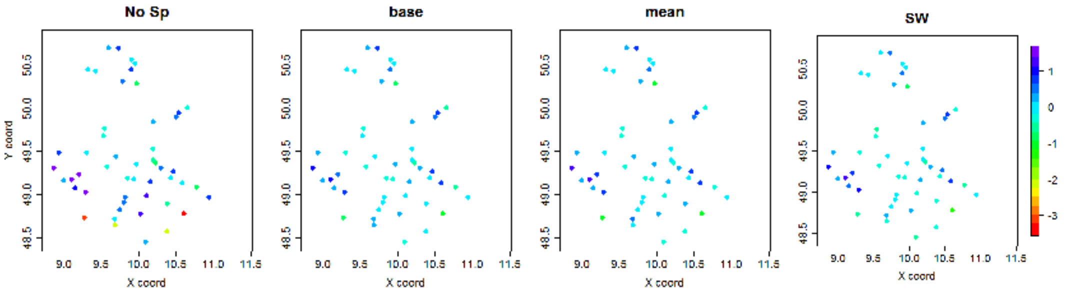


Figure 5.25: Predictive bias in the validation sites with the four model specifications on 19/09/2010 at 01 a.m.

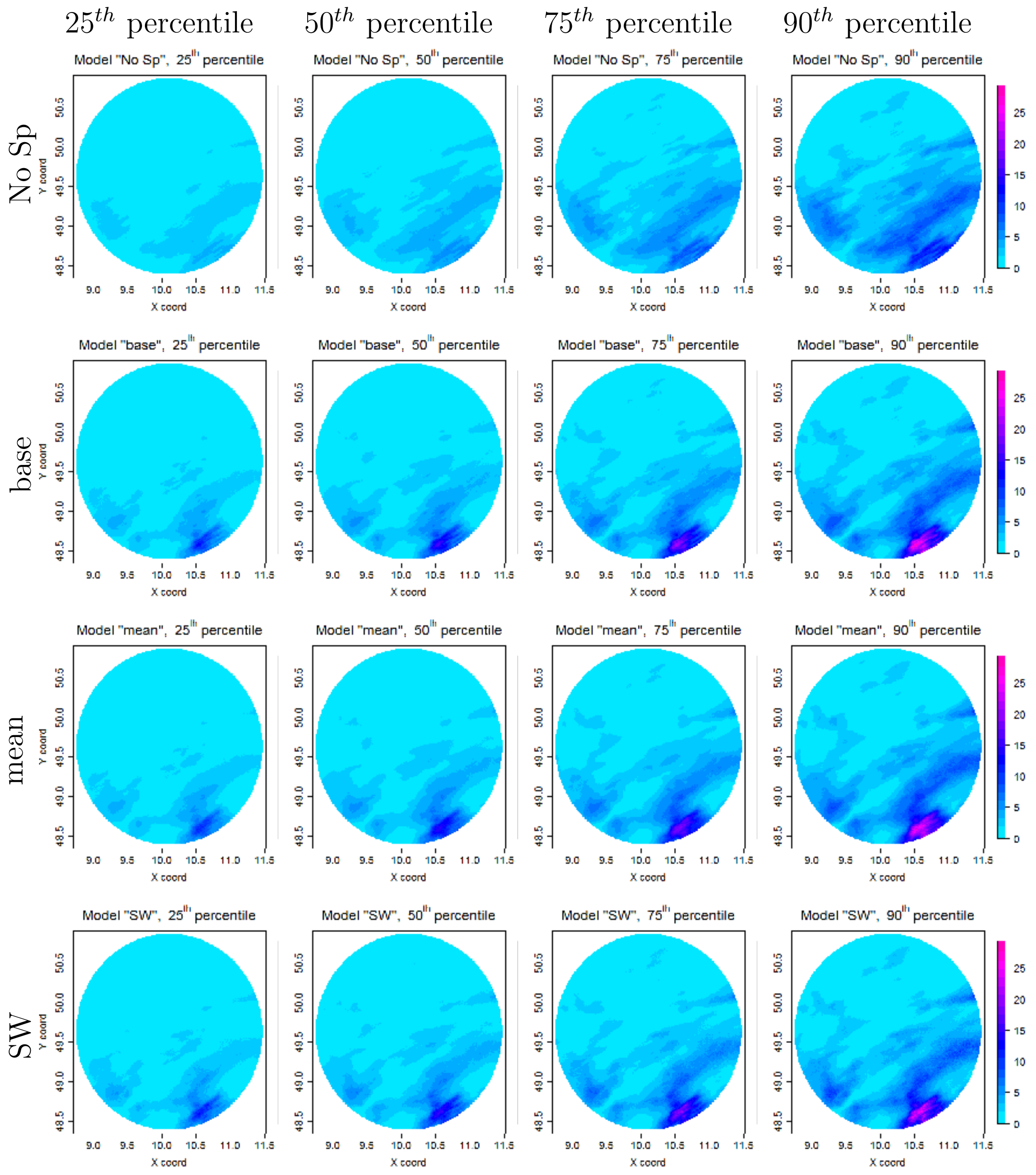

25th percentile
50th percentile
75th percentile
90th percentile
No Sp
base
mean
SW
Model "No Sp", 25th percentile
Model "No Sp", 50th percentile
Model "No Sp", 75th percentile
Model "No Sp", 90th percentile
Model "base", 25th percentile
Model "base", 50th percentile
Model "base", 75th percentile
Model "base", 90th percentile
Model "mean", 25th percentile
Model "mean", 50th percentile
Model "mean", 75th percentile
Model "mean", 90th percentile
Model "SW", 25th percentile
Model "SW", 50th percentile
Model "SW", 75th percentile
Model "SW", 90th percentile
Y coord
48.5
49.0
49.5
50.0
50.5
X coord
9.0
9.5
10.0
10.5
11.0
11.5
0
5
10
15
20
25

Figure 5.26: Map of the $25^{th}, 50^{th}, 75^{th}$ and 90^{th} predictive percentile with the four model specifications on 19/09/2010 at 01 a.m.

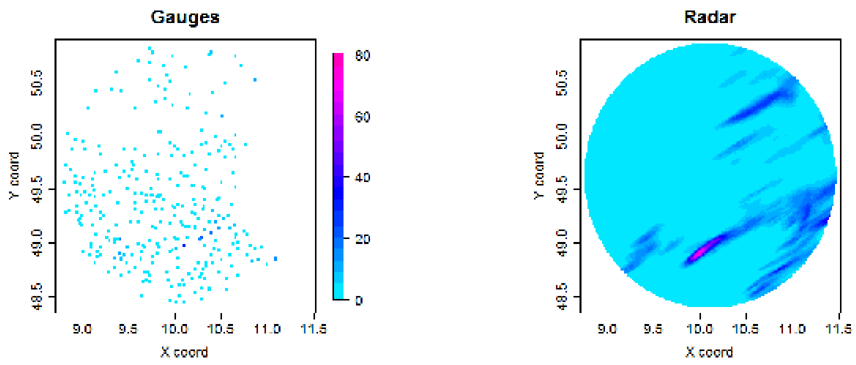


Figure 5.27: Gauge measurements (left panel) and radar grid (right panel) on 18/09/2010 at 05 p.m.

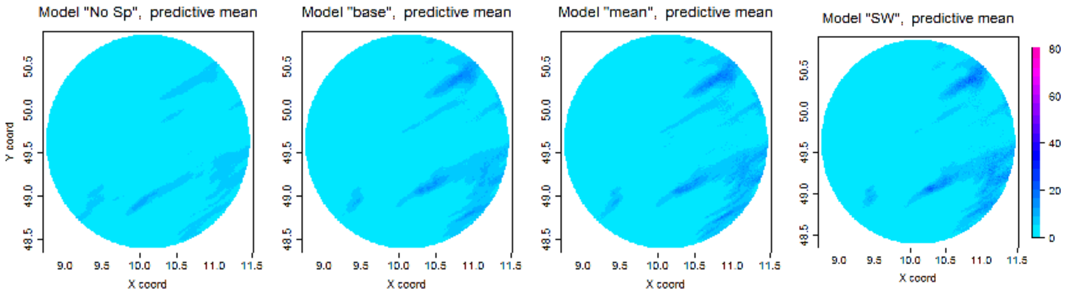

Model "No Sp", predictive mean
Model "base", predictive mean
Model "mean", predictive mean
Model "SW", predictive mean
Y coord
50.5
50.0
49.5
49.0
48.5
X coord
9.0
9.5
10.0
10.5
11.0
11.5
80
60
40
20
0

Figure 5.28: Reconstructed field with the four model specifications on 18/09/2010 at 05 p.m.

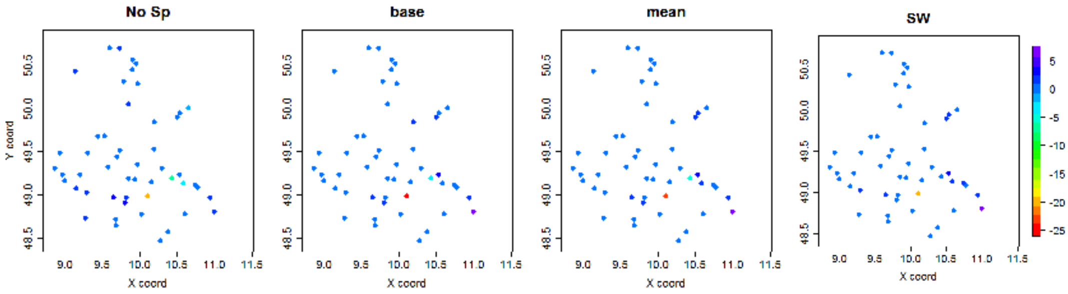


Figure 5.29: Predictive bias in the validation sites with the four model specifications on 18/09/2010 at 05 p.m.

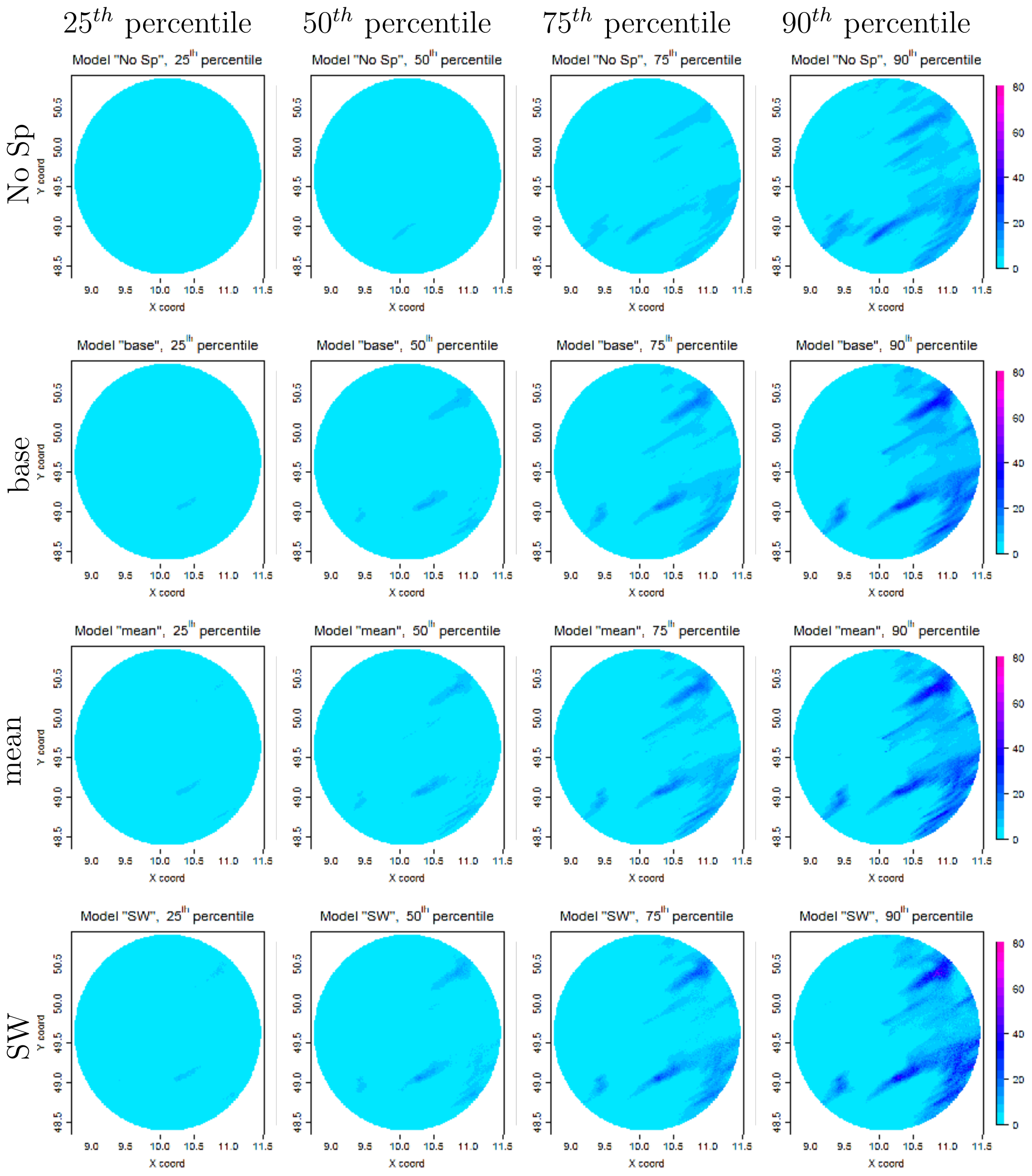

25th percentile
50th percentile
75th percentile
90th percentile
No Sp
base
mean
SW
Model "No Sp", 25th percentile
Model "No Sp", 50th percentile
Model "No Sp", 75th percentile
Model "No Sp", 90th percentile
Model "base", 25th percentile
Model "base", 50th percentile
Model "base", 75th percentile
Model "base", 90th percentile
Model "mean", 25th percentile
Model "mean", 50th percentile
Model "mean", 75th percentile
Model "mean", 90th percentile
Model "SW", 25th percentile
Model "SW", 50th percentile
Model "SW", 75th percentile
Model "SW", 90th percentile
Y coord
X coord
48.5
49.0
49.5
50.0
50.5
9.0
9.5
10.0
10.5
11.0
11.5
0
20
40
60
80

Figure 5.30: Map of the $25^{th}, 50^{th}, 75^{th}$ and 90^{th} predictive percentile with the four model specifications on 18/09/2010 at 05 p.m.

5.4 Predictive intervals

Tables 5.5 and 5.6 report the average width of the 90% and 50% central predictive intervals, showing that modelling correlation in the spatial effects allows a reduction in the amplitude of the intervals, thus ensuring sharper predictions. Differences between models “base”, “mean” and “SW” are not very relevant, with “SW” retrieving slightly bigger intervals. This behaviour is not unexpected since the uncertainty in the results obtained with the Bayesian procedure derives from the consideration of all the sources of uncertainty, and the last model also includes the stochastic process Q. Moreover, such small increase in the interval is also associated with an increase in the coverage, computed here from the non-randomized PIT as explained in Section 4 and shown in Tables 5.7 and 5.8; coverages are overall close to the nominal level as desired. A further insight into the amplitude of the predictive intervals is provided by the boxplots of their width, reported in Figures 5.31 and 5.32; large intervals associated to huge rainfall amounts can be object of investigation for future development.

avgwd 90%	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	2.573	5.056	3.325	3.942	4.042	1.582	2.830	1.876	2.805
Base	2.036	4.031	2.619	3.226	3.857	1.279	2.246	1.207	2.118
Mean	2.032	4.029	2.633	3.133	3.758	1.211	2.235	1.181	2.089
SW	2.130	3.994	2.694	3.248	4.318	1.303	2.267	1.217	2.180

Table 5.5: Average width of the 90% centered credibility intervals for the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed events together (tot).

avgwd 50%	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	0.930	1.631	1.083	1.423	1.203	0.536	0.953	0.634	1.020
Base	0.761	1.451	0.947	1.143	1.268	0.442	0.793	0.433	0.868
Mean	0.764	1.469	0.952	1.131	1.251	0.427	0.789	0.426	0.866
SW	0.802	1.452	0.976	1.16	1.439	0.455	0.804	0.436	0.898

Table 5.6: Average width of the 50% centered credibility intervals for the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours together (tot).

cov 90%	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	0.906	0.905	0.899	0.899	0.914	0.900	0.895	0.906	0.903
Base	0.926	0.905	0.899	0.900	0.900	0.910	0.882	0.911	0.903
Mean	0.930	0.900	0.891	0.897	0.891	0.898	0.878	0.918	0.901
SW	0.929	0.904	0.915	0.894	0.897	0.905	0.891	0.918	0.908

Table 5.7: Non-randomized coverage of the 90% centered credibility intervals for the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours together (tot).

cov 50%	E1	E2	E3	E4	E5	E6	E7	E8	tot
No Sp	0.503	0.514	0.504	0.507	0.520	0.525	0.463	0.474	0.469
Base	0.586	0.545	0.487	0.513	0.522	0.538	0.522	0.539	0.530
Mean	0.589	0.542	0.493	0.507	0.513	0.521	0.518	0.519	0.523
SW	0.580	0.529	0.511	0.542	0.539	0.536	0.519	0.530	0.532

Table 5.8: Non-randomized coverage of the 50% centered credibility intervals for the 50 selected validation sites separately on each of the 8 Events (E1-E8) and on all the analysed hours together (tot).

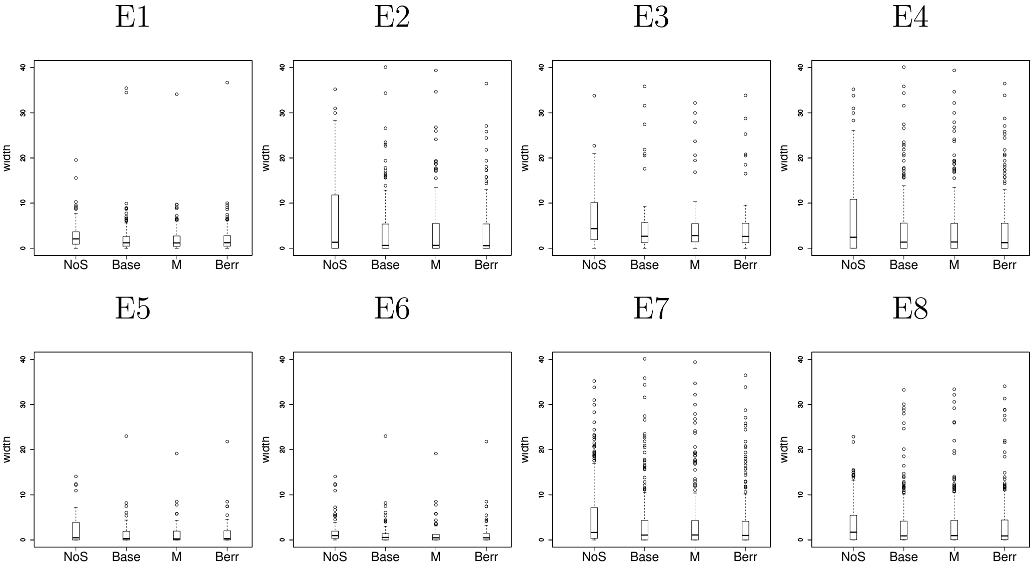


Figure 5.31: Boxplots of the widths of the 90% credibility intervals obtained with the different model specifications on the 8 Events.

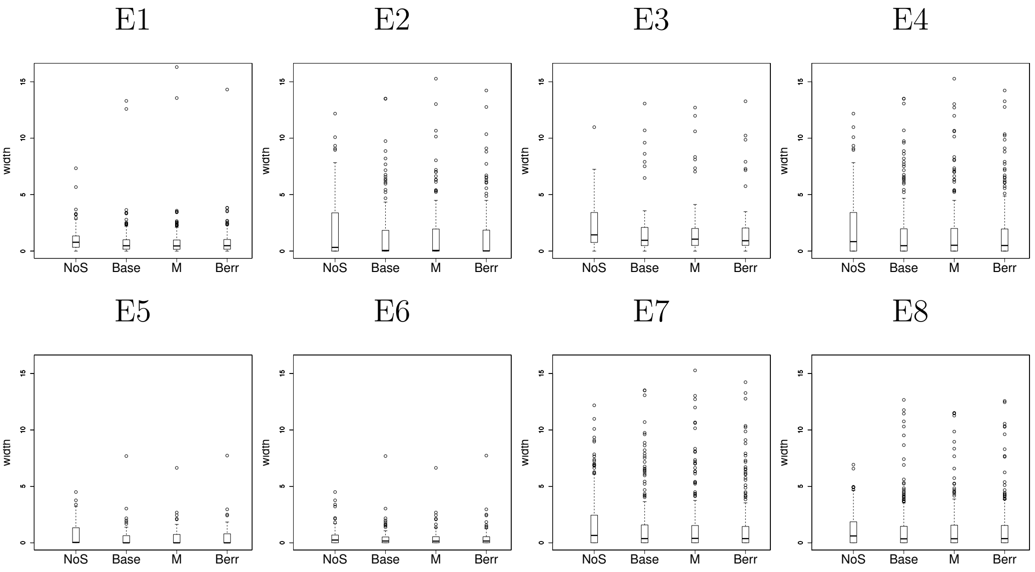


Figure 5.32: Boxplots of the widths of the 50% credibility intervals obtained with the different model specifications on the 8 Events.

Chapter 6

Conclusions

Spatial prediction of hourly rainfall via radar calibration is addressed in the thesis. The existing literature is reviewed in Chapter 1: geostatistical models exploiting radar information, like Kriging with External drift, outperform methods only relying on rain gauge interpolation. The task of merging data provided by different sources is challenging, especially when spatial supports do not coincide; the change of support problem (COSP) has become a very debated issue in recent years. The increasing availability of data of various nature motivated statisticians to develop tools for addressing misalignment and efficiently exploiting all available information. The downscaler approach proposed by Berrocal *et al.* (2011) provides an appealing solution, thanks to its flexibility in assigning spatially varying stochastic weights to a gridded covariate; its definition at a point level allows to restrain the complexity, thus preserving computational feasibility.
Most of the existing literature, including Kriging and the aforementioned downscaler with stochastic weights, relies on the hypothesis of normality; nevertheless, such assumption is not adequate when dealing with hourly precipitation. Gamma distribution is a rather common and suitable choice for modelling rainfall amounts, thanks to its flexible shape driven by two parameters, which allows to reproduce the right skewness typical of rainfall distributions. Nevertheless, a relevant amount of zero values characterizes hourly rainfall and cannot be explained by a continuous distribution; a mixed discrete-continuous specification should be exploited instead. Two-part mod-

els constitute an appealing solution: their likelihood consists of a (positive) probability on zero and a continuous distribution modelling positive amounts. This tool has not been deeply explored when complex spatial processes are involved. Part of our original contribution consists in the construction of a Bayesian hierarchical model for hourly rainfall spatial prediction, with a two-part semicontinuous specification. It allows direct modelling of rain probability without the need for a deterministic threshold; radar is exploited as a covariate in this stage, and correlated random effects capture spatial patterns. At the same time, positive amounts of rain are modelled via a Gamma distribution, whose mean is driven by radar via a spatial regression. In its basic formulation, introduced in Bruno *et al.* (2014) and here denoted as Model "base", rain gauges are punctually associated with the radar cell containing their location. Nevertheless, the model allows for enhancements addressing the change of support problem in this non-Gaussian setting. Neighbouring information is added via the mean of the 8 surrounding pixels (Model "mean"), or through a stochastic weighting of radar cells driven by a latent Gaussian process defined on the whole grid (Model "SW", inspired by Berrocal *et al.* 2011). Details about the model structure are provided in Chapter 2. The model is defined on rain gauge locations only, thus containing the complexity. Estimation is performed via Markov chain Monte Carlo algorithms; its implementation in C language, together with the use of BLAS and LAPACK algebraic libraries, allows for computational speed, with estimation of Model "base" on 300 rain gauges completed in about ten minutes, and reconstruction of the whole rainfall field ($\sim$49,000 pixels) requiring only one minute. Model "SW" is more computationally demanding, but an efficient implementation, with its additional efforts restrained only to some of the iterations, make estimation and prediction feasible in operational time, too. R software is essentially exploited for managing and visualizing data and results.

The aim of the work is rainfall prediction. Literature from Dawid (1984) encourages to produce and communicate full probabilistic predictions, rather than point forecasts, since they deliver complete distributional information. Our model succeeds in this, by providing large MCMC samples from the posterior predictive distribution. We remark that our fully Bayesian approach guarantees a reliable assessment of uncertainty taking all the estimating steps into account. Chapter 3 recalls the most important desired features for probabilistic forecasts, like calibration and sharpness, and explains how to evaluate them starting from a literature review. The assessment of probabilistic forecasts is not trivial, due to the necessity of comparing predictive distributions with observed point data; moreover, competing forecasts must be ranked with fair procedures favouring the data-generating process over the alternatives. Graphical and numerical tools are proposed for assessing, quantifying and comparing probabilistic forecasts, PIT histogram and Continuous Rank Probability Score being the most common. Moreover, since practical applications often require point forecasts, the concept of consistent scoring functions is recalled; they represent the correct tool for a fair model comparison. Uncertainty information can be associated to point forecasts via predictive intervals. Most of the tools available from the literature have been developed for continuous distributions, with some exceptions for count data; two-part semicontinuous models create further challenges, due to their mixed discrete-continuous nature. Chapter 4 explains why some of the standard tools for evaluating predictions give incorrect results when applied to the two-part semicontinuous case, and proposes alternatives when necessary. In particular, a non-randomized PIT histogram for two-part semicontinuous models is proposed, ensuring uniformity when the ideal forecast is considered. Moreover, both randomized and non-randomized procedures for computing the coverage of predictive intervals are illustrated, correcting standard methods that would retrieve erroneous results due to the positive probability of zero. The presented tools are suitable for the evaluation of every kind of rainfall prediction relying on a two-part semicontinuous model. For example, they can be applied to time forecasts obtained from ensemble postprocessing via Bayesian Model Averaging or Ensemble Model Output Statistics (EMOS). It is worth noticing that no restriction is put on the continuous distribution for positive values: the proposed procedures need not to be adapted for this. Moreover, they can be easily adapted for assessing semicontinuous predictions obtained with different approaches. In this work, we focused on the two-part assumption since it allows to model rain probability independently from rain accumulation as a difference from other approaches like tobit. Further research might investigate whether the modelling of correlation among the regression coefficients or the spatial effects in the two stages would constitute a useful enrichment of the proposed model (see for example Su *et al.* 2009, Neelon *et al.* 2011). Different modelling choices can also be considered; for example, left censoring consists in the attribution to zero of all the probability corresponding to negative outcomes of a suitable distribution defined on the real axis. Such approach of course limits the flexibility in modelling rainfall probability, but may be suitable in scenarios where the focus is on high rainfall amounts; the left censored generalized extreme value distribution (see Scheuerer 2014), with its heavy right tail, might be an adequate choice for that purpose.

Chapter 5 presents the results for the three models, plus a simple regression model without spatial correlation which constitutes a benchmark (Model “No Sp”). Hourly precipitation regarding 8 distinct rainfall events in the Emilia-Romagna Region is analysed, with separate model runs on each hour. Literature often suggests to pool the available data together for obtaining a big training dataset; in our case, the richness of the spatial information eliminates this necessity. Moreover, the estimates of the regression parameters, showing irregularity in the behaviour in the radar-gauge relationship, support the decision of separately addressing successive hours. All model specifications thus focus on the spatial aspects, which are predominant; further advancements may address the temporal development, linking successive hours by modelling the evolution of the regression parameters. When the mean of the neighbouring pixels is included as a further covariate (Model “mean”), the influence of radar is split between the contingent pixel and its neighbourhood, with a change of predominance of the former or the latter.

Examples of maps of the spatial random effects reveal their role in identifying spatial patterns. Model “SW”, with its efficient exploitation of the whole radar map, delegates part of this task to the covariate, relying slightly less on the structured errors. Model “No Sp”, lacking of instruments for capturing and reproducing spatial trends, is outperformed by the three proposed model specifications, on the basis of the aforementioned tools for forecast assessment. All predictive performances are evaluated on a randomly chosen set of 50 validation sites for each hour; both event-specific and global scores and plots are shown, in order to provide a detailed insight into results and simplify interpretation. In terms of probabilistic forecasts, Model "SW" provides the best results, indicated by the smallest Continuous Rank Probability Score values. Quantile decomposition plot reveals that the main differences between competing specifications are collected around the central quantile levels. The same ranking is obtained when analysing point predictions, with "SW" providing the best predictive performances. Root Mean Square Error and Mean Absolute Error, computed after choosing the posterior predictive mean or median, respectively, are shown; comparison with raw radar confirms the three spatial specifications succeed in calibrating the indirect measurements, providing smaller values of the scoring functions.

Analysis of the credibility intervals confirms the probabilistic consistency between predictions and observations, with coverages close to the nominal value; modelling spatial correlation has a positive effect also in terms of sharpness, reducing interval widths.

A relevant peculiarity of our models is the estimate of rainfall probability, obtained as a by-product but interesting per se. Specific tools for evaluating the prediction of rainfall probability, such as the Brier Score and the Reliability plot, confirm good performance also in this specific task; sharpness histograms highlight low uncertainty when predicting rainfall occurrence, with a predominance of probabilities near zero or one. Recent joint studies from statisticians and psychologists (Mass *et al* 2009, Joslyn and Savelli 2010) suggest to communicate uncertainty in meteorological forecasts, also when rainfall occurrence is the object of the study and when a non specialized public is addressed. Probabilistic results can improve the effectiveness of communication and increase users' confidence. Since we provide full information about the predictive distribution, the probability of exceeding a certain threshold can be computed, which can for example be useful for quantifying the risk of floods. Probabilistic calibration ensures that the probabilities obtained with our three spatial models are reliable.

Our models produce spatial predictions in every location where radar is available; this allows to overcome the absence of rain gauges in the Eastern area covered by the Adriatic Sea.

As a conclusion, the proposed model allows reconstruction of the rainfall field correcting radar information. Simple exploitation of neighbouring radar information via a deterministic mean (Model “mean”) does not improve predictions, but the estimates of the coefficients reveal that the pixel containing the location of interest is not always the most relevant. A more sophisticated methodology, consisting in stochastically weighting the neighbouring pixel (Model “SW”), improves point and probabilistic forecasts. While the basic model fast implementation is compatible with an operational exploitation, Model “SW” is more computationally demanding. For this reason, since differences between the performances of the two models are moderate, we suggested ARPA the use of Model “base” for operational every day purposes, and to run Model “SW” when investigating more complex events, where the improvement of this model can be more valuable.

Future model development may address the detection of high rainfall amounts localized in small areas where a few rain gauges (or none) are available; this scenario is not rare in convective events, which are often characterized by major discrepancies between radar and rain gauges. Possible extensions of the work may address the distinction between stratiform and convective rainfall events, based on an automatic method for classification based on the analysis of the patterns of radar maps; distinct model specifications should then be developed for the two scenarios, and possibly also for the mixed stratiform-convective case. An alternative approach which may help in detecting heavy rainfall can rely on quantile regression, which allows the influence of the covariate to change across the quantile levels of the response variable; an approximation of the whole predictive distribution might be recovered in that case by considering a large number of quantiles.

Acknowledgements

This research work was funded by an ARPA-SIMC / Department of Statistical Sciences agreement (nr. 20/2012). Special thanks to Roberta Amorati, Valentina Pavan, Pier Paolo Alberoni and ARPA-SIMC who provided us data, knowledge about the phenomenon and precious cooperation and comments.

I need to thank my supervisor Daniela Cocchi, who guided my Ph.D. project with expertise and dedication, and my co-supervisors Francesca Bruno and Tilmann Gneiting, who were fundamental for the development of the work. My sincere thanks also go to our co-author Fedele Greco.

This work is part of the FIRB 2012 project no. RBFR12URQJ, with title "Statistical modelling of environmental phenomena: pollution, meteorology, health and their interactions", funded by the Italian Ministry of Education, Universities and Research. I would like to thank all the colleagues who feed this project with experience and enthusiasm and who built a network which allows exchange of ideas and fruitful cooperation.

Acknowledgements to the Ph.D. committee (Marco Alfó, Livio Finos, Luigi Ippoliti and Roberta Amorati) for their observations and suggestions.

Finally, a big thank you to all my German colleagues and friends, and everyone who helped me in Heidelberg under the scientific and human point of view. And, last but not least (at all), an enormous thank you to my family, who always supports me in every possible way, to Riccardo and all friends who were close to me in these three years.

Bibliography

Adler, R. F, Huffman, G. J, Chang, A., Ferraro, R., Xie, P. P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P. and Nelkin, E. (2003). The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *Journal of Hydrometeorology*, **4**, 1147–1167.

Amorati, R., Bruno, F., Cocchi, D. and Scardovi, E. (2013). Radar and rain gauge merging methods for operational hourly precipitation estimates. *11th International Precipitation Conference, Ede-Wageningen, The Netherlands, 30 June 2013–3 July 2013.*

Amorati, R., Alberoni, P. P. and Fornasiero, A. (2012). Operational bias correction of hourly radar precipitation estimate using rain gauges. *ERAD 2012 - the 7th European Conference on Radar Meteorology and Hydrology.*

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, **9**, 1518-1530.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall/CRC, Boca Raton, FL.

Battan, L. J. (1973). *Radar Observation of the atmosphere.* University of Chicago Press, Chicago.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, **7**, 686–690.

Berrocal, V., Craigmile, P. and Guttorp, P. (2012). Regional climate model assessment using statistical upscaling and downscaling techniques. *Environmetrics*, **23**, 482–492.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010a). A bivariate spatio-temporal downscaler under space and time misalignment. *Annals of Applied Statistics*, **4**, 1942–1975.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010b). A spatiotemporal downscaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176–197.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2011). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics*, **68**, 837–848.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–1193.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **136**, 192–236.

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP Model Output. *Monthly Weather Review*, **132**, 338–347.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statistics and Probability Letters*, **77**, 1473–1478.

Brown, P. J. (1994). Measurement, regression and calibration. *Oxford University Press.*

Brown, P. E, Diggle, P. J, Lord, M. E. and Young, P. C (2001). Space-time calibration of radar rainfall data. *Applied Statistics*, **50**, 221–241.

Bruno, F., Cocchi, D., Greco, F. and Scardovi E. (2014a). Spatial reconstruction of rainfall fields from rain gauge and radar data. *Stochastic Environmental Research and Risk Assessment*, **28**, 1235–1245.

Bruno, F., Greco, F. and Scardovi, E. (2014b) Assessment of Bayesian models for rainfall field reconstruction. *Proceedings of the* 47^{th} *Scientific Meeting of the Italian Statistical Society, Cagliari, Italy, June 11–13.*

Carroll, S. S., Day, G., Cressie, N. and Carroll, T. R. (1995). Spatial modeling of snow water equivalent using airborne and ground-based snow data. *Environmetrics*, **6**, 127–139.

Chilès, J. P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* Wiley, New York.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, **39**, 841–862.

Chumchean, S., Seed, A. and Sharma, A. (2006). Correcting of real-time radar rainfall bias using a Kalman filtering approach. *Journal of Hydrology*, **317**, 123–137.

Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts Board on Atmospheric Sciences and Climate Division on Earth and Life Studies (2006). *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts.* The National Academies Press.

Cooley, D., Nychka, D. and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, **497**, 824–840.

Corradi, V. and Swanson, N. R. (2006). *Predictive density evaluation.* In Elliott, G., Granger C. W. J. and Timmermann, A. (eds), Handbook of Economic Forecasting. **1**, Elsevier, Amsterdam, 197–284.

Costa, M. and Alpuim, T. (2011). Adjustment of state space models in view of area rainfall estimation. *Environmetrics*, **22**, 530–540.

Cowles, M. K. and Zimmermann D. L. (2003). A Bayesian space-time analysis of acid deposition data combined from two monitoring networks. *Journal of Geophysical Research: Atmospheres* **108**, (D24), 1984–2012.

Cowles, M. K., Zimmermann D. L., Christ, A. and McGinnis, D. L. (2002). Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems. *Journal of Agricultural, Biological and Environmental Statistics*, **7**, 536–557.

Cressie, N. (1990) The origins of kriging. *Mathematical Geology*, **22**, 239–252.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, 2nd edition.* Wiley.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society Series A*, **147**, 278–292.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, **27**, 65–81.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39**, 863–883.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13**, 253–263.

Diomede D., Marsigli C. and Paccagnella T. (2013). Calibration of limited-area ensemble precipitation forecasts using reforecasts. *Monthly Weather Review*, **142**, 2176–2197.

Erdin, R., Frei, C. and Künsch, H. (2012). Data transformation and uncertainty in geostatistical combination of radar and rain gauges. *Journal of Hydrometeorology*, **13**, 1332–1346.

Fassó, A. and Finazzi, F. (2013). A varying coefficients space-time model for ground and satellite air quality data over Europe. *Statistica & Applicazioni*, Special Issue, 45–56.

Foley, K. M. and Fuentes, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane wind prediction. *Journal of Agricultural, Biological and Environmental Statistics*, **13**, 37–59.

Fornasiero, A., Amorati, R., Alberoni, P. P., Ferraris, L. and Taramasso, A. C. (2004). Impact of combined beam blocking and anomalous propagation correction algorithms on radar data quality. *Proceedings of ERAD 2004, the 3th ERAD conference held in Gotland, Sweden, Copernicus GmbH 2004*, 216–222.

Fraley, C., Raftery, A., Gneiting, T., Sloughter, M. and Berrocal, V. (2011). Probabilistic weather forecasting in R. *R Journal*, **3(1)**, 55–63.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, Cambridge.

Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics*, **3**, 291–309.

Fuentes, M. and Raftery, A. E. (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**, 36–45.

Fuentes, M., Reich, B. and Lee, G. (2008). Spatial–temporal mesoscale modeling of rainfall intensity using gage and radar data. *The Annals of Applied Statistics*, **2**, No. 4, 1148–1169.

Gelfand, A. E., Kim, H. J., Sirmans, C. F. and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387–396.

Gelfand, A. E., Ravishanker, N., Ecker, M. D. (2000). *Modeling and inference for point-referenced binary spatial data.* In: Dey, D. K., Ghosh, S. K., Mallick, B. K.(eds.), Generalized Linear Models: A Bayesian Perspective. Marcel Dekker, Inc., New York, 373–386.

Gelfand, A. E., Schmidt, A. M., Banerjee, S. and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, **13**, 263–312.

Gilleland, E. (2014), Verification: Forecast verification utilities. R package, version 1.40.

Gilleland, E., Ahijevych, D. and Brown, B. J. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, **24**, 1416–1430.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual*

Review of Statistics and its Application, **1**, 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting, T. and Ranjan, R. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society, Series B*, **72**, 71–91.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted proper scoring rules. *Journal of Business and Economic Statistics*, **29**, 411–422.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69**, 243–268.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B*, **14**, 107–114.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford University Press.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**, 632–648.

Gotway, C. A. and Young, L. J. (2007). A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, **16**, 115–135.

Goudenhoofdt, E. and Delobbe, L. (2009). Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrology and Earth System Sciences*, **13**, 195–203.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in insurance. *Scandinavian Actuarial Journal*, **3**, 202–225.

Guillas, S., Bao, J., Choi, Y. and Wang, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmospheric Environment*, **42**, 1338–1348.

Gunn, K. L. S. and Marshall, J. S.(1958). The Distribution With Size of Aggregate Snowflakes. *Journal of Meteorology*, **15**, 452–461.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312-1327.

Hannesen R. and Gysi, H. (2002). An enhanced precipitation accumulation algorithm for radar data. *Proceedings of the* 2^{nd} *European Radar Conference, Delft, The Netherlands, 18-22 November 2002.*

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–570.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**, (1), 145–168.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather and Forecasting*, **22**, 637–650.

Jolliffe, I. T. and Stephenson, D. B. (2003). *Forecast Verification: A*

Practicioner's Guide in Atmospheric Science. Wiley, Chichester, U.K.

Joslyn, S. and Savelli, S. (2010). Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorological Applications*, **17**, 180–195.

Joslyn, S., Nadav-Greenberg, G. and Nichols, R. M. (2009). Probability of precipitation. Assessment and enhancement of end-user understanding. *Bulletin of the American Meteorological Society*, **90**, 185–193.

Joslyn, S., Nemec, L. and Savelli, S. (2013). The benefits and challenges of predictive interval forecasts and verification graphics for end users. *Weather, Climate, and Society*, **5**, 133–147.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press.

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

Kloog, I., Koutrakis, P., Coull, B., Lee, H. and Schwartz, J. (2011). Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, **45**, 6267–6275.

Koenker, R. and Machado, J. A. F. (1999). Goodness-of-fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296–1310.

Koistinen, J. and Puhakka, T. (1981). An improved spatial gauge-radar adjustment technique. *Proceedings of the 20th Conference On Radar Meteorology, Boston, American Meteorological Society*, 179–186.

Kyung, M. and Ghosh, S. K. (2010). Bayesian inference for directional conditionally autoregressive models. *Bayesian Analysis*, **4**, 675–706.

Leung, J. K. Y., Law, T. C. (2002). Kriging Analysis on Hong Kong Rainfall Data. *The Hong Kong Institution of Engineers, Transactions 9.*
Li, W., Zhang, C. and Dey, D. K. (2010). Estimating threshold-exceeding probability maps of environmental variables with Markov chain random fields. *Stochastic Environmental Research and Risk Assessment*, **24**, 1113–1126.

Liesenfeld, R., Nolte, I., and Pohlmeier, W. (2006). Modeling financial transaction price movements: A dynamic integer count data model. *Empirical Economics*, **30**, 795–825.

Liu, Z., Le, N. and Zidek, J. V. (2008). Combining measurements and physical model outputs for the spatial prediction of hourly ozone space-time fields. *The University of British Columbia, Department of Statistics, Technical Report no. 239.*

Liu, Y., Paciorek, C. and Koutrakis, P. (2009). Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, **117**, 886–892.

Lu, H. and Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, **36**, 265–285.

Marshall, J. and Palmer, W. (1948). The distribution of raindrops with size. *Journal of Meteorology*, **5**, 165–166.

Matheron, G. (1963). Principles of statistics. *Economic Geology*, **58**, 1246–1266.

Mass, C., Joslyn, S., Pyle, J., Tewson, P., Gneiting, T., Raftery, A., Baars, J., Sloughter, J. m., Jones, D. and Fraley, C. (2009). PROBCAST: A web-based portal to mesoscale probabilistic forecasts. *Bulletin of the American Meteorological Society*, **90**, 1009–1014.

McMillan, N., Holland, D. M., Morara, M. and Feng, J. (2009). Combining numerical model output and particulate data using bayesian space-time modeling. *Environmetrics*, **21**, 48–65.

Neelon, B. and O'Malley, A. J. (2014). *Two-part models for zero-inflated and semicontinuous data.* To appear in Boris Sobolev and Constantine Gatsonis (eds.) Handbook of Health Services Research, Springer.

Neelon, B., O'Malley, A. J. and Normand, S. L. T. (2011). A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity. *Biometics*, **67**, 280–289.

Orasi, A., Jona Lasinio, G. and Ferrari, C. (2009). Comparison of calibration methods for the reconstruction of space-time rainfall fields during a rain enhancement experiment in Southern Italy. *Environmetrics*, **20**, 812–834.

Pathac, C. S. and Vieux, B. E.(2007). Geo-spatial analysis of NEXRAD rainfall data for central and South Florida. *K. C. Kabbes (eds). World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat. ASCE Publications. Proceedings of the 2007 World Environmental and Water Resources Congress, Tampa, Florida, May 15–19, 2007.*

Pilz, J. and Spöck, G. (2008). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, **22**, 621–632.

Pocernich, M. (2009). Verification: Forecast verification utilities. R package, version 1.29.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation Models: the bayesian melding approach. *Journal of the American Statistical Association*, **95**, 1244–1255.

Reich, B. J., Chang, H. H. and Foley, K. M. (2014). A spectral method for spatial downscaling. *Biometrics*, DOI: 10.1111/biom.12196.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, **130**, 1653–1660.

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, **55**, 16–30.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* Chapman & Hall/CRC, Boca Raton, FL.

Sahu, S. K, Jona Lasinio, G., Orasi, A. and Mardia, K. V. (2005). A comparison of spatio-temporal Bayesian models for reconstruction of rainfall fields in a cloud seeding experiment. *Journal of Mathematics and Statistics*, **1**, 273–281.

Sahu, S. K, Gelfand, A. E. and Holland, D. M. (2010). Fusing point and areal level space-time data with application to wet deposition. *Journal of the Royal statistical Society Series C*, **59**, 77–103.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191–201.

Savelli, S. and Joslyn, S. (2013b). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, **27**, 527–541.

Scardovi, E., Alberoni, P. P., Amorati, R., Cocchi, D. and Pavan, V. (2012a). Uso integrato dei dati di pioggia radar-pluviometro: analisi esplorativa dei dati orari. *Quaderno Tecnico ARPA*.

Scardovi, E., Bruno, F., Amorati, R. and Cocchi, D. (2012b). Rainfall spatial modeling from different data sources. *Gonçalves, A. M., Sousa, I., Machado, L., Pereira, P., Menezes, R., Faria, S. (eds).*

Proceedings of the VI International Workshop on Spatio-Temporal Modelling (METMA6). Guimarães, Portugal, 12–14 September 2012, CMAT-Centro de Matemática da Universidade do Minho. ISBN: 978-989-97939-0-3.

Schefzik, R., Thorarinsdottir T. L. and Gneiting T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling *Statistical Science*, **28**, 616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1086–1096.

Schiemann, R., Erdin, R., Willi, M., Frei, C., Berenguer, M. and Sempere-Torres, D. (2011). Geostatistical radar-raingauge combination with nonparametric correlograms: methodological considerations and application in Switzerland. *Hydrology and Earth System Sciences*, **15**, 1515–1536.

Schmidt, A. M. and Gelfand, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant Data. *Journal of the Geophysical Research: Atmospheres*, **108**, 8783.

Schumacher, M., Graf, E. and Gerds, T. (2003). How to assess prognostic models for survival data: A case study in oncology. *Methods of Information in Medicine*, **42**, 564–571.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, **1**, 43–62.

Seo D. J., Perica S., Welles E., Schaake J. C. (2000). Simulation of probabilistic quantitative precipitation forecast *Journal of Hydrology*, **239**, 203–229.

Seo, D. J. and Smith, J. A. (1991 a). Rainfall estimation using raingages and radar - A Bayesian approach: 1. Derivation of estimators. *Stochastic*

Hydrology and Hydraulics, **5**, 17–29.

Seo, D. J. and Smith, J. A. (1991 b). Rainfall estimation using raingages and radar - A Bayesian approach: 2. An application. *Stochastic Hydrology and Hydraulics*, **5**, 31–44.

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

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, **4**, 283–291.

Smith, B. J. and Cowles, M. K. (2007). Correlating point-referenced radon and areal uranium data arising from a common spatial process. *Applied Statistics*, **56**, 313–326.

Spiegelhalter, D. J., Best, N., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*, **64**, 583–639.

Stein, M. L. (1999). *Interpolation of Spatial Data. Some Theory for Kriging.* Springer.

Stern, R. D. and Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society Series A*, **147**, 1–34.

Su, L., Tom, B. D. M., Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, **10**, 374–389.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316–334.

Talagrand, O., Vautard, R. and Strauss, B. (1997). Evaluation of probabilistic prediction systems. *Proceedings of the ECMWF Workshop on Predictability*, 1–26.

Thomas, A., O'Hara, B., Ligges, U. and Sturz, S. (2006). Making BUGS open. *R News*, **6**, 12–17.

Velasco-Forero, C. A., Sempere-Torres, D., Cassiraga, E. F. and Gómez - Hernández, J. J. (2009). A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Advances in Water Resources*, **32**, 986–1002.

Yoo, C. and Ha, E. (2007). Effect of zero measurements on the spatial correlation structure of rainfall. *Stochastic Environmental Research and Risk Assessment*, **21**, 287–297.

Wackernagel, H. (2003). *Multivariate Geostatistics. An Introduction with Applications.* Springer.

Wikle, C. K and Berliner, L. M. (2005). Combining information across spatial scales. *Technometrics*, **47**, 80–91.

Winkler, R. L. (1977) *Rewarding expertise in probability assessment.* In H. Jungermann and G. de Zeeuw, eds., Decision Making and Change in Human Affairs. Dordrecht, Holland: D. Reidel, 127–140.

Yao, T. and Journel, A. G. (1998). Automatic modeling of (cross) covariance tables using fast fourier transform. *Mathematical Geology*, **30**, 589–615.

Zhou, J., Chang, H. and Fuentes, M. (2012). Estimating the health impacts of climate change with calibrated model output. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 377–394.

Zhou, J., Fuentes, M. and Davis, J. (2011). Calibration of numerical model output using nonparametric spatial density functions. *Journal of*

Agricultural, Biological, and Environmental Statistics, **16**, 531–553.