UNIVERSITÀ DEGLI STUDI DI CAGLIARI
Dipartimento di Ingegneria del Territorio

DOTTORATO DI RICERCA
IN INGEGNERIA DEL TERRITORIO
XX CICLO

PhD Thesis

ASSESSING UNCERTAINTY PROPAGATION OF PRECIPITATION INPUT IN HYDROMETEOROLOGICAL ENSEMBLE FORECASTING SYSTEMS

PhD candidate: **Giuseppe Mascaro**

Tutor: **Prof. Roberto Deidda**
Co-Tutor: **Prof. Enrique R. Vivoni**

Cagliari, December 2007

Ringrazio vivamente e calorosamente il mio tutor, prof. Roberto Deidda, che é stato per me fondamentale nel motivarmi e guidarmi in tutte le fasi di questo lavoro.

Desidero inoltre ringraziare il mio co-tutor, prof. Enrique R. Vivoni, per avermi trasmesso entusiasmo nel portare avanti il lavoro e per avermi costantemente supportato durante la mia esperienza all'estero.

Infine, desidero esprimere un vivo grazie a Maria Grazia Badas e al Prof. Enrico Piga per avermi saputo ascoltare e sostenere nei momenti di difficoltá.

A mamma e papá

Giuseppe

Abstract

Advanced hydrometeorological forecasting systems for streamflow predictions include the combined use of meteorological and hydrological models as well as of downscaling models and data assimilation systems. To account for the different sources of uncertainty involved in such complex schemes, ensemble forecasting technique has been recently adopted.

Precipitation input forecasts are a fundamental component in spatially-distributed forecasting systems. Nevertheless, the rigorous assessment of this source of uncertainty and its propagation into hydrological response have been so far barely investigated. In this work, we focus on hydrometeorological systems aimed at predicting floods in basins with short response time and we propose a systematic verification framework to evaluate (i) the uncertainty associated to ensemble precipitation forecasts and (ii) its propagation into hydrological response.

For this purpose, we preliminary design a forecasting system that starts from information at coarse scale provided by Numerical Weather Prediction (NWP) models and uses a precipitation downscaling model to generate an ensemble of spatiotemporal precipitation fields at high resolution, which are in turn utilized as meteorological forcing for a fully-distributed hydrological model.

In the first part of the work, a new verification method is proposed to test the consistency (i.e. ensemble and observation are drawn from the same distribution) of high-resolution precipitation fields forecasted by calibrated downscaling models. The method is based on a generalization of the verification rank histogram and tests the exceedance probability of a fixed precipitation threshold calculated from the observed or ensemble fields. The verification procedure is applied in numerical hindcasting experiments carried out in controlled conditions using the STRAIN (Space Time RAINfall) downscaling model and assuming no uncertainty in the coarse scale information provided by NWP models. Results permit us to conclude that: (i) ensemble members generated using model parameters estimated on the observed event are overdispersed; (ii) the adoption of a single calibration relation linking model parameters and coarse meteorological observable can lead to the generation of consistent ensemble members; (iii) when a single calibration relation is not able to explain observed events variability, storm-specific calibration relation should be adopted to return consistent forecasts.

Results of the first part of the study are then used to test how uncertainty and eventual deficiencies of ensemble precipitation forecasts affect hydrological response. Numerical hindcasting experiments are conducted again in controlled conditions by applying the proposed hydrometeorological system coupling the STRAIN precipitation downscaling model and the tRIBS (TIN-based Real Time Basin Integrator) distributed hydrological model. Uncertainty associated to basin initial state and hydrological model parameterization and structure has not been taken into account. The test basins are the Baron Fork (OK, USA) and 14 nested sub-catchments, allowing evaluation of uncertainty propagation for a wide range of catchment size (from 0.78 to 808 km^2). The STRAIN downscaling model is applied in several events to generate (i) consistent, (ii) overdispersed and (iii) underdispersed ensemble precipitation hindcasts, which are in turn utilized to force the tRIBS hydrological model. Consistency of the simulated ensemble hydrographs is evaluated, in the three cases, by means of a rigorous verification procedure ad-hoc developed and based again on the rank histogram. Results show that running the tRIBS model with either consistent or overdispersed or underdispersed precipitation leads to consistent ensemble streamflow irrespective of the basin size. This implies that basins play an important role as powerful spatio-temporal integrators of precipitation variability, at least for limited simulation time periods (1-2 days).

In conclusion, uncertainty assessment in ensemble hydrometeorological forecasting systems is a cutting-edge research topic, but few studies have so far proposed systematic verification methodologies, especially for the hydrological variables. The originality of this work stems from (i) the development of a rigorous verification framework for ensemble outputs produced in different steps of the forecasting systems and (ii) the analysis of a great number of events that allows drawing statistically significant conclusions. This research has considered only uncertainty of precipitation input; thus, future investigations should be devoted to the evaluation of the other sources of uncertainty and of their reciprocal interaction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hydrometeorological predictions are extremely important for support decisions in civil protection and water resources management. Operational agencies require forecasts to adopt the necessary measures to protect people and properties in case of flood occurrence and water managers to plan allocation of water resources. For example, in case of an intense storm, operational agencies would like to know which areas may be flooded, while water managers would like to acquire information about the possible amount of flow arriving to a reservoir and exceeding its capacity, in order to optimize the water volume to be released and save as much resource as possible. Clearly, the larger the forecast lead time, the easier and more appropriate the decisions.

As a consequence, researchers of meteorological and hydrological scientific communities, water managers and users are addressing their efforts on the development of sophisticated hydrometeorological forecasting systems for streamflow predictions. The hydrometeorological forecasting schemes firstly proposed in literature and adopted by operational centers are deterministic: a single 'best' meteorological input is produced and used to force the hydrological model which returns a single 'best' forecast without any confidence level. Therefore, in this context, support for decision process is limited.

The dramatic growth in computational power has suggested the opportunity to adopt ensemble forecasting in hydrometeorological systems to account for uncertainty and formulate probabilistic hydrological forecasts (Schaake et al. 2007). Ensemble forecasting technique has been originally developed in applied meteorology (Lorenz 1963) to deal with and predict

uncertainty of Numerical Weather Prediction (NWP) models. Ensemble forecasting comprises multiple (typically between 5 and 100) runs of NWP models which differ in the initial conditions and/or the numerical representations of the atmosphere. Subsequently, the technique has also been utilized by hydrologists in order to account for the different sources of uncertainty that are mainly due to data (input and output), state variable, parameterization and model structure.

Advanced ensemble hydrometeorological forecasting systems include the combined use of meteorological and hydrological models as well as of statistical downscaling models, land-surface models and data assimilation systems. The integrated use of all or part of such tools allows the simulation of ensemble of weather-climate forcing and land-surface states which are used respectively as inputs and initial conditions of hydrological models for the generation of streamflow ensemble.

The development of this complex forecasting systems involving several sources of uncertainty has determined the need for an accurate verification of the outputs produced in all the internal steps. At the moment, a consistent and systematic research effort has been devoted by atmospheric scientists to verification of NMP model forecasts (e.g., Murphy & Winckler 1987, Anderson 1996, Hamill & Colucci 1997, 1998, Wilson et al. 1999, Ebert & McBride 2000, McBride & Ebert 2000, Wilks 2001, 2004, Grimit et al. 2006), while few studies have been performed to develop specific verification techniques for hydrometeorological systems and, especially, for hydrological outputs (Welles et al. 2007). In most cases, the verification of ensemble hydrometeorological forecasts has been limited to the qualitative comparison between the observed and the ensemble hydrographs using simple scalar measures for few events. Such approaches are not able to provide an accurate and statistically based verification. By limiting the verification to streamflow hydrographs, the uncertainty of internal steps cannot be evaluated. Further, when a scarce number of events is used, it is not possible to infer information about system performances in other conditions. Finally, the lack of a rigorous statistical framework prevents assessing if the ensemble forecasts and observations are equally likely from the statistical point of view.

Notable exceptions are the studies of Georgakakos et al. (2004), Carpenter & Georgakakos (2004) and Franz et al. (2003), who proposed a rigorous statistical characterization of the uncertainty of ensemble streamflow, by means of methods commonly used for verification of meteorological forecasts (Wilks 2006). Nevertheless their applications were not aimed at predicting

floods through a hydrometeorological forecasting system: the former two studies were focused on the assessment of the uncertainty due to different input parameters of hydrological models, whereas the latter one analyzed the statistical properties of the U.S. National Weather Service ensemble streamflow predictions for water supply forecasting.

Precipitation input forecasts represent a source of uncertainty whose rigorous and systematic assessment has been so far scarcely investigated even if it is fundamental in spatially-distributed forecasting systems. Knowledge is further more limited regarding its propagation into hydrological response. In this work we have focused on ensemble hydrometeorological schemes aimed at predicting flood in basins with short response time and we have developed specific verification methods (i) to characterize uncertainty of precipitation ensembles used as forcing for the hydrological model, and then (ii) to evaluate how and if this uncertainty affects hydrological response.

For purpose of this study, we have preliminary designed a hydrometeorological scheme that starts form coarse information provided by NWP models and couples in cascade a statistical model for precipitation downscaling and a fully-distributed hydrological model. Given the complexity and high non-linearity of the processes involved, other sources of uncertainty, such as information at coarse scale provided by meteorological models, basin initial state and hydrological model parameterization and structure, have not been taken into account.

Setup of hydrometeorological schemes is highly dependent on the spatiotemporal scales solved by NWP and hydrological models. When hydrologic predictions are required in large-size basins ($\sim$ 10,000 km^2) with a high response time, output of Global Circulation Models (GCM), characterized by a resolution of approximately 40 km, can be directly coupled to hydrological models. The development of hydrostatic and non-hydrostatic Limited Area Model (LAM), utilizing GCM outputs as boundary conditions and initial states, has allowed hydrological modeling in watersheds with smaller sizes (e.g., Verbunt et al. 2007). Nevertheless, these spatiotemporal scales does not allow flash-flood prediction in catchments with small area ($<$ 100 km^2) and response time.

Precipitation downscaling models can then be used within the forecasting system to bridge the scale gap between the coarse scales resolved by meteorological models and the finer ones required by hydrological modeling. Starting from information at a coarse scale, downscaling models are able to produce, with minimal computational effort, an ensemble of high resolution

spatiotemporal precipitation fields that are statistically coherent with the large-scale condition. In particular, operational use of downscaling models is realized by means of calibration relations linking their few parameters with one or more meteorological observable at coarse scale, provided presumably with low uncertainty by meteorological models. For instance, the Convective Available Potential Energy (CAPE) or the precipitation volume at the large scale have been adopted in previous works by Over & Gupta (1994, 1996), Perica & Foufoula-Georgiou (1996), Deidda (2000), Deidda et al. (2004), to calibrate parameters of different downscaling models.

Physically based, distributed hydrologic models can in turn offer distinct advantages over conceptual, lumped models (i.e., models treating the watershed as a single unit) used widely for flood forecasting, once operational techniques mature. Furthermore, physically based models are distinguished from conceptual models, even if both are distributed in nature, by their capability to represent hydrologic processes at scales ranging from the hillslope to the river basin. In recent years, there has been a significant improvement in the inputs to distributed models, including digital elevation models (DEMs), land surface parameter maps, and hydrometeorological data, which are used to parameterize physically based equations at individual basin locations. In distributed models, basin runoff response can vary within the watershed according to the temporal and spatial variability in rainfall, surface properties, and antecedent wetness (Ivanov et al. 2004*a*,*b*, Vivoni et al. 2005). In particular, runoff generation via multiple physical mechanisms can be captured in a high level of detail over a complex watershed surface. This capability permits simulating basin conditions traditionally excluded from operational flood forecasts, including discharge forecasts at interior stream locations, time series of runoff generation at particular sites, and spatiotemporal fields of hydrologic response (e.g., soil moisture, runoff mechanisms, and recharge).

Uncertainty associated to precipitation forecasts provided by downscaling models and its propagation into hydrological response has been assessed by means of two verification methods, ad hoc developed for precipitation and streamflow ensembles, respectively. Both methods are based on a generalization of a graphical technique adopted in applied meteorology, the research field that has so far addressed a significant research effort to develop specific verification methods for ensemble models. The technique, known as Verification Rank Histogram (VRH) and proposed independently by Anderson (1996), Hamill & Colucci (1997), Talagrand et al. (1997), tests

consistency hypothesis (Anderson 1997), that is the degree to which the observed state is a plausible member of the forecast ensemble, of single scalar (i.e. one dimensional) outputs.

The first verification method proposed in this work tests ensemble precipitation forecasted by downscaling models. The single scalar variable for which consistency is evaluated is the exceedance probability of a fixed precipitation threshold i^*, calculated from each spatiotemporal precipitation field. This selection was motivated by several reasons: (i) each downscaled precipitation field is a multi-variate variable with high dimensionality, so that a full verification of such ensemble members is challenging; (ii) the purpose of downscaling models is the statistical characterization of precipitation at high resolution and (iii) the precipitation exceedance probability as the predictand variable satisfies the requirement for verifying the statistical properties of the precipitation field and can be calculated and tested for different values of the threshold; (iv) the method does not make any reference to the internal generation mechanism of the downscaling model and, thus, is applicable to different kinds of model.

The verification method is developed and applied using a multifractal downscaling model of precipitation, known as Space Time RAINfall downscaling model (STRAIN), proposed by Deidda et al. (1999) and refined in Deidda (2000). Nevertheless, the study results are general enough to be considered valid also for other downscaling models. Three hindcasting experiments are carried out on synthetic spatiotemporal precipitation fields that are then verified through the proposed procedure. The experiments permit investigating two key-aspects of downscaling model: (i) the effect of sampling variability on parameter estimation from the observed precipitation fields and (ii) downscaling model performances when calibration relations are used to interpret the spread of parameter estimates.

In a subsequent part of the work, results obtained by analyzing uncertainty of precipitation input have been used to evaluate how this uncertainty and possible deficiencies of downscaled precipitation fields affect hydrological response and overall performances of the hydrometeorological system. In particular, we have tried to understand which is the dominant role played by the basins, which are characterized by two opposite response mechanisms with respect to precipitation spatiotemporal variability. On one hand, basins separate the different runoff components and act as a non-linear filter emphasizing intermittency characteristics of precipitation. On the other hand, they act as complex integrators of precipitation in space and

time mitigating the spatio-temporal precipitation variability. Vegetation, soil texture, aquifer and basin geomorphometric characteristics play an important role within these opposite mechanisms.

A verification method based on the VRH has been developed to test consistency of streamflow ensemble. The method requires the definition of a fixed verification time length, dependent on the basin response time, where the accumulated streamflow at different durations can be calculated and utilized as single scalar metrics for VRH construction.

Numerical hindcast experiments have been then carried out in controlled conditions by applying a hydrometeorological system coupling the STRAIN downscaling model with the fully-distributed hydrological model known as TIN-Based Real Time Basin Integrator (tRIBS). The target basin has been the Baron Fork, located in Oklahoma (USA) and 14 nested sub-basins (areas ranging from $\sim$0.8 to 800 km^2). The possibility for testing multiple interior locations has been provided by the capability of the distributed model to provide time series of runoff in every desired sites. Experiments have been setup to investigate the following questions: (i) Which are the characteristics of ensemble streamflow simulated by the hydrological model when precipitation ensemble forcing are consistent or characterized by deficiencies? (ii) Is propagation of rainfall input uncertainty affected by catchment scale? (iii) Which is the role played by the basins? Do they emphasize precipitation variability or do they act as complex integrators?

The thesis is organized as follows. In chapter 2, the hydrometeorological forecasting system designed to test uncertainty propagation of precipitation input, is illustrated in all its steps. First, NWP models are briefly described; second, the statistical downscaling and multifractal theory together with the STRAIN model are reminded highlighting the aspects useful to understand the verification procedure for precipitation ensemble; finally, characteristics of the tRIBS distributed hydrological model are illustrated. Chapter 3 provides a review of ensemble forecasting techniques and of the main verification methods currently used in applied meteorology to test ensemble outputs, with particular regard to the VRH. The uncertainty assessment of precipitation input is discussed in detail in chapter 4, illustrating the verification method for precipitation ensemble forecasted by downscaling models and the three hindcasting experiments. Effects of uncertainty and deficiencies of ensemble precipitation input into hydrological response are analyzed in chapter 5, describing the verification procedure for ensemble streamflow and the numerical experiments based on the application of

the hydrometeorological system to the Baron Fork and nested sub-basins. Finally, conclusions are provided in chapter 6.

Chapter 2

Hydrometeorological System Coupling a Precipitation Downscaling Model and a Distributed Hydrological Model

The chapter illustrates the hydrometeorological forecasting system designed to evaluate the propagation of precipitation input uncertainty into hydrological response. Fig. 2.1 shows a scheme illustrating the system. For each event, ensemble streamflow forecasts are obtained through three steps in cascade. Precipitation maps provided by Numerical Weather Prediction (NWP) models are first used to determine precipitation accumulated in a coarse spatiotemporal domain $L \times L \times T$ containing the study watershed (1^{st} step). Subsequently, the statistical downscaling model provides an ensemble of spatiotemporal rainfall fields at a scale $\lambda \times \lambda \times \tau$ suitable for hydrological modeling (2^{nd} step). These high resolution fields are utilized to force the distributed hydrological model which in turn furnishes an ensemble of hydrographs (3^{rd} step).

The object of the thesis is focused on the uncertainty assessment of the hydrological part of the forecasting system (i.e. 2^{nd} and 3^{rd} steps), while no uncertainty has been associated to the coarse scale precipitation derived from NWP models (1^{st} step).

The chapter is organized as follows. A short description of NWP models is provided in section 2.1. The aim of section 2.2 is to briefly describe some theoretical aspects and applications of precipitation downscaling models and

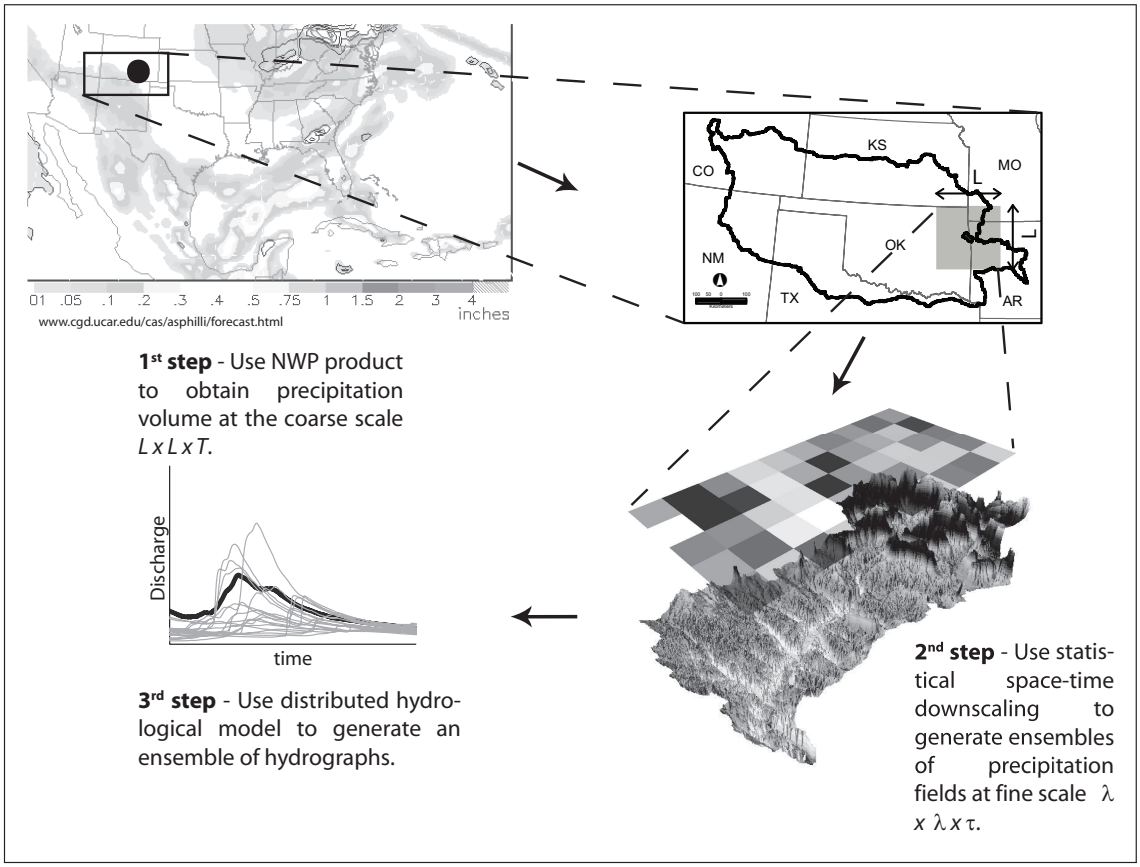

Figure 2.1: Scheme of a hydrometeorological forecasting system starting from coarse information provided by NWP models and coupling a precipitation downscaling model with a distributed hydrological model.

to clarify how the verification procedure developed in chapter 4 can be applied on the ensemble outputs provided by such models. The specific downscaling model, known as STRAIN, used to test the verification procedure is also introduced. Finally, section 2.3 contains a description of the main features of the TIN-bases Real Time Basin Simulator (tRIBS), the distributed hydrological model used in the designed forecasting system.

2.1 Numerical Weather Prediction Models

Numerical Weather Prediction (NWP) models use mathematical models of the atmosphere to predict the weather. They start from the observation of the atmosphere at a given time and use the equations of fluid dynamics and thermodynamics to estimate the state of the fluid at some time in the future. Since these equations are differential and non-linear, their solution can be only approximated by means of numerical methods, which change for the different models: global models often use spectral methods for the horizontal dimensions and finite difference methods for the vertical dimension, while regional models usually use finite-difference methods in all three dimensions.

Atmospheric processes can be described in detail or simplified using parameterizations. Given the high computational effort required in the first case, a greater detail is used especially in scientific applications while parameterizations are adopted for operational purposes.

Models are initialized using observed data from radiosondes, weather satellites, and other instruments. The irregularly-spaced observations are processed by data assimilation and objective analysis methods, which perform quality control and obtain values at locations usable by the model's mathematical algorithms (usually an evenly-spaced grid). The data are then used in the model as the starting point for a forecast. Commonly, the set of equations used is known as the primitive equations. These equations are initialized from the analysis data and rates of change are determined. The rates of change predict the state of the atmosphere a short time into the future. The equations are then applied to this new atmospheric state to find new rates of change, and these new rates of change predict the atmosphere at a yet further time into the future. This time stepping procedure is continually repeated until the solution reaches the desired forecast time. The length of the time step is related to the distance between the points on the

computational grid. Time steps for global climate models may be on the order of tens of minutes, while time steps for regional models may be a few seconds to a few minutes.

As proposed by Lorenz (1963), it is impossible to definitely predict the state of the atmosphere, owing to the nonlinear nature of fluid dynamics. Furthermore, existing observation networks have limited spatial and temporal resolution, which introduces uncertainty into the true initial state of the atmosphere. To account for this uncertainty, stochastic or ensemble forecasting is used, involving multiple forecasts created with different model systems, different physical parameterizations, or varying initial conditions.

The spatial and temporal scales provided by NWP models with a good confidence level, are strictly related to the validity range of the approximations made to solve the equations, to availability and precision of initial observations and boundary conditions as well as to the computational demand. For example, General Circulation Model (GCM) of the European Center for Medium-Range Weather Forecast (ECMWF) furnishes forecast at spatial resolution of 40 km (http://www.ecmwf.int/index_forecasts.html) with a lead time of 10 days for the deterministic forecast and an ensemble of 51 forecasts to ten days at 80 km resolution (the so called Ensemble Prediction System). These resolutions allow hydrological modeling to be roughly carried out only over large basins (size $>$ 10,000 km^2) with a significant number of grid-points of the meteorological model domain falling inside the basin. However, these large-size basins are very few all over the world. To obtain higher resolutions, national and local meteorological centers have developed the nested models known as Limited Area Models (LAM), which are in most cases based on the hydrostatic approximation and utilize the GCM outputs as boundary conditions and initial states. In principle, this kind of models can be run even up to 10 km, a resolution that can result however too coarse for hydrological applications in catchments with short concentration time. Other LAM models based on the non-hydrostatic hypothesis can solve scales of 250 m, but require observations that, in practice, are not available at this resolution.

Finally, we highlight that uncertainty associated to the products provided by NWP models increases, in general, as the spatiotemporal resolutions increase. However little research has analyzed and exactly quantified this relation that is very important to establish if and at which scales downscaling models may be used in cascade to NWP models within a hydrometeorological

system.

2.2 Precipitation Downscaling Model

2.2.1 Statistical Downscaling and Multifractal Theory

A general downscaling scheme can be summarized as follows. Suppose that a precipitation measure $\mu(D)$ or its probability distribution $P_D(\mu)$ in a domain D in $\Re^3$ (where two dimensions are in space and one in time) is observed or simulated by a NWP model. The aim of a downscaling scheme is to determine the probability distribution $P_{\delta D}(\mu)$ of the measure $\mu(\delta D)$ over a finer region δD by analyzing and then reproducing statistical properties of the measure μ observed over different scales between D and δD. In a hydrometeorological forecasting system, the domain D is a coarse spatiotemporal region (see large cube in Fig. 2.2) at which NWP models provide forecasts with low uncertainty, while the domain δD is a fine spatiotemporal region (smaller cubes of Fig. 2.2) required by a hydrological model.

Characterization of precipitation statistical properties at different scales has been carried out through multifractal theory (e.g. Lovejoy & Mandelbrot 1985, Schertzer & Lovejoy 1985, Gupta & Waymire 1993, Over & Gupta 1996, Perica & Foufoula-Georgiou 1996, Deidda 2000). This approach requires the presence of scale-invariance laws defined as:

$$< [\mu(\delta D)]^q > \cong \delta^{\zeta(q)} \cdot < [\mu(D)]^q > \tag{2.1}$$

where $< \cdot >$ denotes an average operator and q is a real number. If equation (2.1) is verified over a range of scales, the measure μ is said to be scale-invariant and if the exponent $\zeta(q)$ is a non-linear function of q, the measure is multifractal. If space-time precipitation fields display scale-invariant and multifractal properties, they can be modeled by means of a stochastic multiplicative cascade dependent on a few parameters. To investigate precipitation scale-invariance, we have to assume relations between statistically coherent scales in space and time (i.e. space-time self-similarity or self-affinity) and on the possible presence of spatial heterogeneities induced for example by orographic constraints.

Let us first discuss the case of rainfall fields displaying homogeneous properties in space. Self-similarity (or scale isotropy) represents the simplest

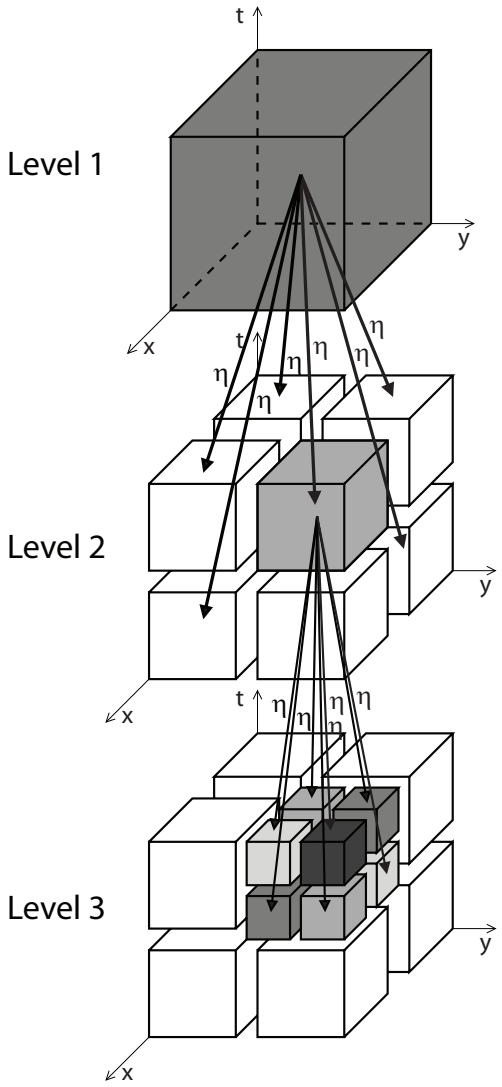

Figure 2.2: A spatiotemporal downscaling scheme, where precipitation values are represented with cubes in a three-dimensional domain (x, y refer to space and t to time). For each downscaling level, precipitation is multiplied by 8 random generators η of STRAIN model.

circumstance where the scaling law (2.1) holds true in multidimensional regions. In such situation, a linear relationship $\lambda = U\tau$ is assumed to hold between coherent space scales λ and time scale τ where the same statistical properties may be observed (Deidda 2000, Deidda et al. 2004). The scale-independent parameter U allows transferring the statistical properties observed at space scales λ to coherent time scales $\tau = \lambda/U$. We can therefore analyze rainfall variability in isotropic and homogeneous three-dimensional regions, where space-time scale-invariance can be investigated by introducing the following measure:

$$\mu_{i,j,k}(\lambda) = \int_{x_i}^{x_i+\lambda} dx \int_{y_j}^{y_j+\lambda} dy \int_{t_k}^{t_k+\lambda/U} dt \quad r(x,y,t) \tag{2.2}$$

where $r(x,y,t)$ is the rainfall rate in (x,y) location at time t, and indexes i, j and k identify the spatial and temporal position of each subregion $\lambda \times \lambda \times \tau$ in the grid partition.

As a result, in our downscaling problem, a rainfall volume $\mu(D)$ is known over an area $L \times L$ and is accumulated over a time $T = L/U$. Our aim is to predict the probability distribution $P(\mu(\lambda))$ of the measure (2.2) over smaller subregions $\lambda \times \lambda \times \tau$, where $\lambda < L$ and $\tau < T$. Scale-invariance (2.1) should be investigated in a wide range of space scales $\lambda_n = Lb_s^{-n}$ and time scales $\tau_n = \lambda_n/U = Tb_t^{-n}$, with a common branching number in space and time ($b_s = b_t$), where the integer n refers to the fragmentation level. Thus equation (2.1) can now be rewritten in terms of the partition function $S_q(\lambda)$:

$$S_q(\lambda) \;\; = < \mu_{i,j,k}(\lambda)^q > \;\sim \;\; \lambda^{\zeta(q)} \tag{2.3}$$

where $< \cdot >$ denotes an ensemble average over all the boxes $\lambda \times \lambda \times \tau$, indexed by i, j and k in the λ-partition. Multifractal exponents $\zeta(q)$ can be estimated plotting $S_q(\lambda)$ versus λ in a log-log space.

In the more general case of self-affine measures, scale-invariance laws can be investigated under anistropic space-time transformations: $\lambda \longrightarrow \lambda/b_s$, $\tau \longrightarrow \tau/b_t$, where the branching number b_s in space now differs from the temporal branching b_t. This approach, known as Generalized Scale Invariance (G.S.I.) (Lovejoy & Schertzer 1985, Schertzer & Lovejoy 1985), characterizes the degree of anisotropy by the scaling anisotropy exponent H relating branching numbers as $b_t = b_s^{(1-H)}$. Scale invariance can thus be investigated in such self-affine measures by introducing in equation (2.2)

a scale parameter $U_\lambda = U_L \cdot \left(\frac{\lambda}{L}\right)^H$, implying that the linear relationship between coherent space and time scales does not hold: $\tau = \lambda/U_\lambda \propto \lambda^{(1-H)}$. This general approach also contains the self-similar case for $H = 0$ (implying $b_s \equiv b_t$ and U constant).

The verification procedure proposed in chapter 4 to test consistency of ensemble precipitation fields generated by downscaling models, requires to calculate the Empirical Cumulative Density Function of each rainfall field at high resolution. We highlight that in case of spatial homogeneity both the simpler self-similar or the more general self-affine transformations assume that probability distribution of rainfall rates is the same in each subregion $\lambda \times \lambda \times \tau$, regardless the grid-cell position in space and/or in time. Thus, although the grid partitioning is slightly different for the self-similar and the self-affine cases, the verification procedure can be applied merging all the rainfall values observed or generated on all the $\lambda \times \lambda \times \tau$ grid-cells.

On the other hand, if rainfall fields display spatial heterogeneity, the probability distribution of rainfall rates in subregions $\lambda \times \lambda \times \tau$ may depend on the spatial location. Thus, in principle, the verification procedure should be applied separately by merging together rainfall rates observed or generated at different times in each spatial verification location. In the case that spatial heterogeneity is only due to differences in the spatial rainfall mean, observed fields may be homogeneized by means of a modulating function (Badas et al. 2006) and the verification procedure may be applied by merging rainfall rates in all the $\lambda \times \lambda \times \tau$ grid-cells as in the homogeneous case.

2.2.2 The STRAIN Multifractal Downscaling Model

Whatever the scale transformation rule holds for analyzed rainfall field, the scale-invariance analysis provides a set of multifractal exponent $\zeta(q)$ that can be used to estimate parameters of the adopted downscaling model. In this work, we apply the STRAIN (Space Time RAINfall) downscaling model (Deidda et al. 1999, Deidda 2000) to test the verification procedure proposed in chapter 4 in a homogeneous and scale-isotropic framework, but the method may be also applied in case of heterogeneity and self-affinity, according to the transformations previously described. The model is based on a log-Poisson generator $\eta = e^A \beta^y$ where β is a parameter, y is a Poisson random variable with mean c and $A = c(1 - \beta)$ is a renormalization constant. The model provides theoretical value for the multifractal exponent $\zeta(q)$ that allows

simulating the multifractal properties in real-world precipitation events. The expression $\zeta(q)$ in a d-dimensional domain ($d = 3$ in a spatiotemporal domain) is given by:

$$\zeta(q) = d \cdot q - c\frac{q(1-\beta) - (1-\beta^q)}{\log 2} \tag{2.4}$$

Equation (2.4) is used to estimate the parameters c and β for each observed event. In recent studies using radar data, the STRAIN model was calibrated and applied to reproduce observed scale-invariant properties (Deidda 2000, Deidda et al. 2004). These studies revealed that β can be assumed constant as e^{-1}, while c was found to decrease as the mean precipitation rate at the coarse scale increases. This behavior was interpreted by the following relationship:

$$c(R) = c_\infty + a \cdot e^{-\gamma \cdot R} \tag{2.5}$$

where R is the precipitation rate at the coarse scale $L \times L \times T$ and c_∞, a and γ are the parameters of the non-linear equation. Based on the mean precipitation rate R at the large scale $L \times L \times T$ obtained from a NWP model output, this relation can be used to estimate c. After parameter estimation, the STRAIN model can generate an ensemble of precipitation fields at high resolution $\lambda \times \lambda \times \tau$, which represents the equiprobable small scale scenarios corresponding to the same coarse scale condition, as depicted in Fig. 2.2.

2.3 The tRIBS Distributed Hydrological Model

The hydrological model used in this study as part of the proposed hydrometeorological system is the TIN-based Real-time Integrated Basin Simulator (tRIBS), a continuous, physically-based, fully-distributed model designed for hydrologic research and forecasting. The model explicitly considers spatial variability in precipitation fields, land-surface descriptors and is capable of resolving basin hydrologic response at very fine temporal and spatial scales. For example, it can be forced with precipitation inputs with a time step of 15' and has computational time steps of 3.75' and 30' for

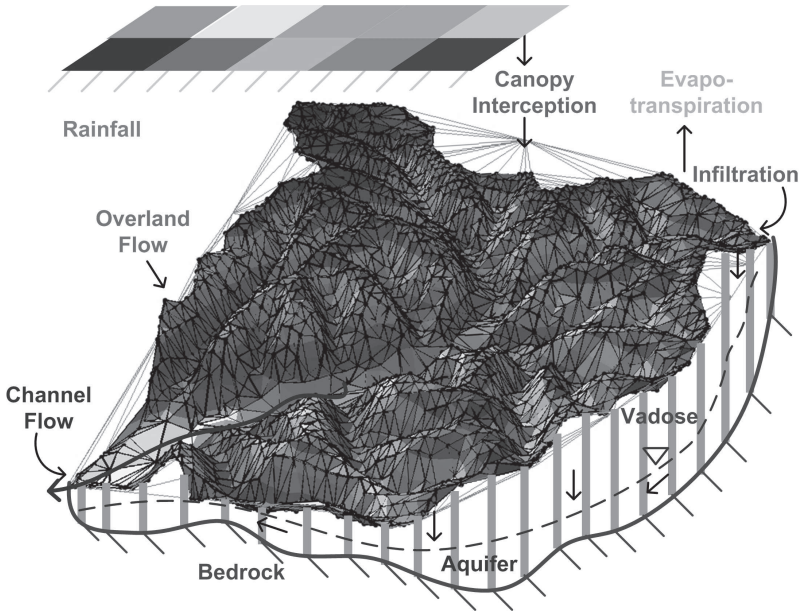

Figure 2.3: Scheme of the domain representation adopted in the TIN-based Real-Time Integrated Basin Simulator (tRIBS) hydrological model, including the parameterized hydrological processes and spatially-distributed meteorological forcing.

the subsurface unsaturated and saturated zones, respectively. tRIBS includes parameterizations of rainfall interception, evapotranspiration, infiltration with continuous soil moisture accounting, lateral moisture transfer in the unsaturated and saturated zones, and runoff routing (see Fig. 2.3 for a scheme). The model computational basis, structure, and description of processes parameterizations are given in full detail in Ivanov et al. (2004*a*).

In this section, we first describe the model domain representation and then we provide an outline of the processes parameterization, highlighting the aspects that make the model suitable to be used in a forecasting system. A summary of model parameters is shown in Table 2.1, derived from Ivanov et al. (2004*b*). The table shows, for each parameter, the units and, as an example, the range of values assumed in a specific application where the model was successfully calibrated and verified in the Baron Fork basin (OK, USA).

Parameter Symbol	Description	Units	Parameter Range	Source
	Vegetation Properties			
p	free throughfall coefficient	-	0.3-0.95	literature
S	canopy capacity	mm	0.8-1.2	literature
K	canopy drainage rate coefficient	mm h^{-1}	0.1-0.25	literature
g	canopy drainage exponent	mm^{-1}	3.2-4.3	literature
a	surface albedo	-	0.13-0.20	literature
H_v	vegetation height	m	0.1-13.0	literature
K_t	optical transmission coefficient	-	0.55-0.75	calibration
r_s	average canopy stomatal resistance	s m^{-1}	70-115	calibration
ν	vegetation fraction	-	0.1-0.65	calibration
	Soil Hydraulic and Thermal Properties			
K_{0n}	saturated hydraulic conductivity	mm h^{-1}	0.5-30.0	calibration
θ_s	saturation soil moisture content	-	0.3 0.4	literature
θ_r	residual soil moisture content	-	0.05	literature
λ_0	pore distribution index	-	0.60 2.0	literature
Ψ_b	air entry bubbling pressure	m	-0.4 to -0.1	literature
f	conductivity decay parameter	m^{-1}	0.40.9	calibration
a_r	anisotropy ratio	-	200900	calibration
n	total porosity	-	0.4 0.5	literature
k_s	volumetric heat conductivity	J m^{-1} s^{-1} K^{-1}	0.31.0	literature
C_s	soil heat capacity	J m^{-3} K^{-1}	1,200,000	literature
	Channel and Hillslope Routing Parameters			
n_e	Mannings channel roughness	-	0.3	calibration
α_B	channel width area coefficient	-	2.33	calibration
β_B	channel width area exponent	-	0.542	calibration
c_v	hillslope velocity coefficient	-	25	calibration
r	hillslope velocity exponent	-	0.4	calibration

Table 2.1: The tRIBS model parameters and their respective ranges used to simulate the hydrologic response in an application on the Baron Fork basin (OK, USA). From Ivanov et al. (2004*a*).

2.3.1 Model Domain Representation

A catchment is represented in tRIBS through a Triangulated Irregular Networks (TIN) consisting of elevation, channel, and basin boundary nodes (Vivoni et al. 2004). TINs are a piece-wise linear interpolation of a set of points, sampled from a digital elevation model (DEM), resulting in triangular facets of varying size. The triangulation represents topographically complex surfaces that include hillslopes, valleys, floodplains and ridges. A multiple resolution approach is adopted in the model to represent the complexity of topography (Vivoni et al. 2004). The stream network is composed of a set of channels ranging from headwater tributaries to large, meandering rivers. The channel cross section is established through geomorphic relations to contributing area (Ivanov et al. 2004*a*). The soil profile and shallow aquifer are bounded by a spatially distributed bedrock assumed to be an impermeable surface.

2.3.2 Precipitation Interception

The Rutter canopy water balance model (Rutter et al. 1971, 1975) is used. Canopy water dynamics is species dependent such that the parameters vary for different vegetation types.

2.3.3 Surface Energy Balance and Evapotranspiration

Short wave and long wave radiation components are simulated accounting for geographic location, time of year, aspect and slope of the element surface (Bras 1990). The combination equation (Penman 1948, Monteith 1965), gradient method (Enthekhabi 2000), and force-restore (Lin 1980, Hu & Islam 1995) method are used to estimate the latent, sensible, and ground heat fluxes at the landsurface. An optimum is sought in terms of the soil surface temperature that leads to the energy balance in the equation:

$$R_n - G \;=\; \lambda E + H \tag{2.6}$$

where R_n is the net radiation, λE, H and G are the latent, sensible and ground heat fluxes. Total evapotranspiration (ET) is determined from moist bare soil, intercepted water and plant transpiration based on soil and vegetation parameters that include vegetative fraction (v), surface albedo (a), canopy height (H_v), stomatal resistance (r_s) and an optical coefficient (K_t), in addition to atmospheric conditions (e.g., air temperature, relative humidity, pressure, wind speed) and solar radiation.

2.3.4 Coupled Unsaturated and Saturated Dynamics

Basin hydrologic response requires an appropriate depiction of the two-way interaction between surface and subsurface processes. The model accounts for moving infiltration fronts, water table fluctuations and moisture losses due to evapotranspiration and groundwater drainage. Each element consists of a sloped column of heterogeneous, anisotropic soil with an exponential decrease in saturated hydraulic conductivity (Beven 1982).

$$K_{si}(z) \;=\; K_{oi}\;\; e^{-fz} \tag{2.7}$$

where $K_{si}(z)$ is the saturated hydraulic conductivity at depth z in the normal or parallel directions ($i = n$ or p), K_{oi} is the saturated hydraulic

conductivity at the soil surface (z=0), and f is a hydraulic conductivity decay parameter. A kinematic approximation for unsaturated flow is used to compute infiltration and propagate moisture fronts in the soil column (Cabral et al. 1992, Garrote & Bras 1995, Ivanov 2002). The unsaturated moisture profile is determined from hydrostatic equilibrium using the Brooks & Corey (1964) parameterization as:

$$\theta(z) \;=\; \theta_r + (\theta_s - \theta_r)\left[\frac{\Psi_b}{z - N_{wt}}\right]^{\lambda_0} \tag{2.8}$$

where $\theta(z)$ is the soil moisture at depth z, θ_r and θ_s are the residual and saturation soil moisture contents, N_{wt} is the depth to the local water table, Ψ_b is the air entry bubbling pressure and λ_0 is the pore-size distribution index (Ivanov et al. 2004*a*).

Coupled to the vertical dynamics is lateral moisture redistribution in the vadose zone and shallow aquifer driven by gradients in surface and groundwater topography. In the unsaturated zone, horizontal flow between contiguous elements is computed over the saturated wedge and along the steepest direction. In the shallow aquifer, a quasi three-dimensional model based on the Dupuit-Forchheimer approximation redistributes groundwater from recharge zones to discharge areas. Lateral exchanges between elements are controlled by hydraulic gradient as:

$$Q_S \;=\; T\, w \tan\beta_w \tag{2.9}$$

where Q_S is the groundwater outflux, w is the flow width, $\tan\beta_w$ is the local water table slope and T is the depth averaged aquifer transmissivity:

$$T \;=\; \frac{a_r K_{on}}{f}\;\left[e^{-fN_{wt}} - e^{-fD}\right] \tag{2.10}$$

where D is the bedrock depth and a_r is the anisotropy ratio (K_{op}/K_{on}). Water table dynamics are computed from groundwater fluxes, vertical recharge and exfiltration. Overall, the water table position anchors the soil moisture profile and determines regions of saturation prior to a storm.

2.3.5 Runoff Production

The coupled nature of the unsaturated and saturated processes results in a robust set of runoff mechanisms. Four basic runoff types are simulated in the tRIBS model: infiltration-excess runoff (R_I) (Horton 1933), saturation-excess runoff (R_S) (Dunne & Black 1970), groundwater exfiltration (R_G) (Hursh & Brater 1941), and perched return flow (R_P) (Weyman 1970). Total runoff (R) is composed of the four production mechanisms:

$$R = R_I + R_S + R_P + R_G \tag{2.11}$$

where $R_I + R_S$ and $R_P + R_G$ are the surface and subsurface components. Infiltration and saturation-excess runoff are rapid surface responses as infiltration is limited by soil conditions, while perched return flow and groundwater exfiltration are slower mechanisms as subsurface flow delays the response to rainfall.

2.3.6 Hillslope and Channel Flow Routing

Runoff generated at each element is routed across an individual hillslope overland flow path and then through the channel network. The hillslope paths are defined over the edges of the triangular facets that connect a node to the closest downstream stream node (Tucker et al. 2001). A nonlinear relation is used to determine velocity over a hillslope path (Ivanov et al. 2004*a*):

$$v_h = c_v \left(\frac{Q}{A_h} \right)^r \tag{2.12}$$

where v_h is the hillslope velocity, A_h is the upslope contributing area, Q is the discharge at the downstream channel node, and r and c_v are spatially-uniform parameters of the velocity relation. Thus, overland travel time ($t_h = l_h/v_h$) is a function of discharge (Q) and hillslope path length (l_h). Overland flow from multiple hillslope nodes serves as lateral inflow into a kinematic wave, one-dimensional routing scheme solved in the channel network (Ivanov et al. 2004*a*). Channel travel time ($t_c = v_c/l_c$) depends on the channel link distance (l_c) and the discharge ($Q = v_c A_c$) through each link. For a wide, rectangular channel ($A_c = bH$), discharge for each link is:

$$Q = \frac{1}{n_e} S^{\frac{1}{2}} H^{\frac{5}{3}} b \tag{2.13}$$

where n_e is the Manning coefficient, S is the channel slope, b is the channel width, and H is the water depth. As overland travel time is faster than groundwater pathways, the partitioning of precipitation into surface and subsurface flow is critical for determining the basin response.

2.3.7 Simulation Capabilities and Model Output

The tRIBS model provides outputs ranging over a variety of spatial and temporal scales.

Point scale

At the smallest spatial scale, the Voronoi element, evolution of all the hydrological state variables can be obtained: rainfall interception, evaporation from the canopy, evolution of the infiltration fronts, dynamics of subsurface fluxes in the unsaturated and saturated zones, soil moisture conditions, runoff generation, and evapotranspiration. Analysis of these dynamics is extremely important for verifying the general physical soundness of the model performance as well as for calibrating parameters of certain hydrological processes.

Hillslope Transect Scale

A group of Voronoi cells forming a hillslope transect can be selected based on the drainage directions connecting the contiguous cells. Time-varying cross-sectional profiles of the hydrological variables can thus be obtained. If field or experimental information about temporal dynamics of the soil moisture and groundwater along the hillslope is available, the pertinent model parameters can be adjusted.

River Reach Scale

The catchment channel network can be represented with a sufficiently high accuracy by a union of segments connecting the stream nodes (Vivoni et al. 2004). For each node of the channel network, the time-series of streamflow

are provided. This offers the flexibility of tracking the spatial variability of runoff conditions in the catchment. As opposed to semidistributed modelling approaches that pre-define points of interest by partitioning the main catchment into nested sub-basins, the approach in tRIBS provides hydrologic prediction at any point of the channel network. This makes the use of tRIBS model into forecasting systems attractive, since a single model run can furnish hydrological predictions in all the sub-basins, allowing the complete definition of flood risk within the basin.

Basin Scale

The capability for reproducing internal variation of hydrologic response is among the essential features offered by distributed models. tRIBS produces spatial maps of all the major hydrological state variables (energy and water fluxes, canopy state, soil moisture conditions, runoff generation, etc.) at a specified temporal resolution. In addition to instantaneous basin states, the model generates frequency distributions and their moments for a number of hydrological variables, thus providing integral representation of site specific properties.

Chapter 3

Ensemble Forecasting Technique and Forecast Verification

The chapter is organized as follows. Section 3.1 illustrates the ensemble forecasting technique as it has been developed and applied in atmospheric science. In section 3.2, the main verification methods for ensemble forecasts are reviewed. After reminding the basic concepts of forecast verification (subsection 3.2.1), the classical methods used to verify probabilistic forecasts are described (subsection 3.2.2). Then, the consistency hypothesis and the ensemble dispersion are introduced as they are properties tested by verification methods specifically developed for ensemble model outputs (subsection 3.2.3). Finally, one of these methods, the Verification Rank Histogram, that has been widely utilized for the studies described in the next chapters, is illustrated in detail (subsection 3.2.4).

Most part of this chapter is based on the exhaustive review of the verification techniques in atmospheric science provided by Wilks (2006).

3.1 Ensemble Forecasting Technique

Ensemble forecasting is a technique originally developed in atmospheric science to deal with and predict the uncertainty associated to meteorological models. This method has also been recently used in hydrology to provide

probabilistic streamflow predictions (e.g., Ferraris et al. 2002, Franz et al. 2003, Verbunt et al. 2007) and to test uncertainty of hydrological models parameters or radar rainfall input (Georgakakos et al. 2004, Carpenter & Georgakakos 2004, Vrugt et al. 2005).

Atmospheric processes exhibit variations and fluctuations that are irregular and, consequently, weather forecast is uncertain. In order to deal quantitatively with uncertainty it is necessary to employ the tools of probability, which is the mathematical language of uncertainty.

The progress of science and the parallel advent of supercomputers has allowed the development of sophisticated models representing physics of the atmosphere and used routinely for forecasting its future evolution. In their usual forms these models are deterministic: they do not represent uncertainty. Once supplied with a particular initial atmospheric state (winds, temperature, humidities, etc.) and boundary forcings (notably solar radiation, sea surface and land conditions) each will produce a single particular result. Rerunning the model with the same inputs will not change that result.

If the description of physical processes and boundary conditions and the collection of data representing the initial atmospheric states were perfect, these model could provide forecast with no uncertainty. But this does not happen at least for two reasons. First of all, even though the models give good approximations to atmospheric behavior, they are not complete and true representations of the governing physics. An important and essentially unavoidable cause of this problem is that some relevant physical processes operate on scales too small to be represented explicitly by these models and some approximations result using only the large-scale information. For example, the problem of forecasting precipitation at high resolution in space and time is still unresolved because of the incapability of representing precipitation physical mechanisms at these small scales. Statistical downscaling models can be used to overcome this problem, starting from large-scale information provided by meteorological models.

Secondly, if all the relevant physics could somehow be included in the atmospheric models, however, we still not escape the uncertainty because of what has come to be known as *dynamical chaos*. According to this phenomenon, discovered by Lorenz (1963), the time evolution of a nonlinear, deterministic dynamical system depends very sensitively on the initial conditions of the system. If two realizations of such a system are started from two only very slightly different initial conditions, the two solutions

will eventually diverge markedly. For the case of atmospheric simulation, imagine that one of these systems is the real atmosphere and the other is a perfect mathematical model of the physic governing the atmosphere. Since the atmosphere is always incompletely observed, it will never be possible to start the mathematical model in exactly the same state as the real system. So even if the model is perfect, it will be still impossible to calculate what the atmosphere will do indefinitely far into the future. Therefore, deterministic forecasts of future atmospheric behavior will always be uncertain and probabilistic methods will always be needed to describe adequately that behavior (Wilks 2006).

Numerical Weather Prediction (NWP) models are the mainstay of weather forecasting. These models exhibit the property that solutions started from only slightly different initial conditions will yield quite different results for projections sufficiently far into the future (weeks for synoptic scale and shorter periods for mesoscale scales). In order to deal with the sensitivity to initial conditions, the so called stochastic dynamical approach can be adopted. Conventional deterministic forecast use the governing equations to describe the future evolution of a single initial state that is regarded as the true initial state. The idea behind stochastic dynamic forecast is to allow the deterministic governing equations to operate on the probability distribution describing uncertainty about the initial state of the atmosphere. The resulting model solutions provide the probability distributions describing uncertainty about the future state of the atmosphere. In addition, since governing equations do not provide a perfect description of the processes, they lead to a further contribute to uncertainty. The visualization of the initial state and forecast probability distributions is achieved through the concept of phase space, which is a geometrical representation of the hypothetically possible states of a dynamical system, where each of the coordinate axes pertains to one of the forecast variable of the system.

The practical solutions to the analytic intractability of stochastic dynamical equations is to approximate these equations using Monte-Carlo methods, as proposed by Leith (1974) and now called ensemble forecasting. Fig. 3.1 illustrates the nature of ensemble forecasting in an idealized two dimensional phase space. The ensemble forecast procedure begins by drawing a finite sample from the probability distribution describing the uncertainty of the initial state of the atmosphere (the small ellipse). Imagine that 8 members (dots) of this distribution surrounding the single best initial value (circled X) in phase space are picked randomly. Collectively, these points are

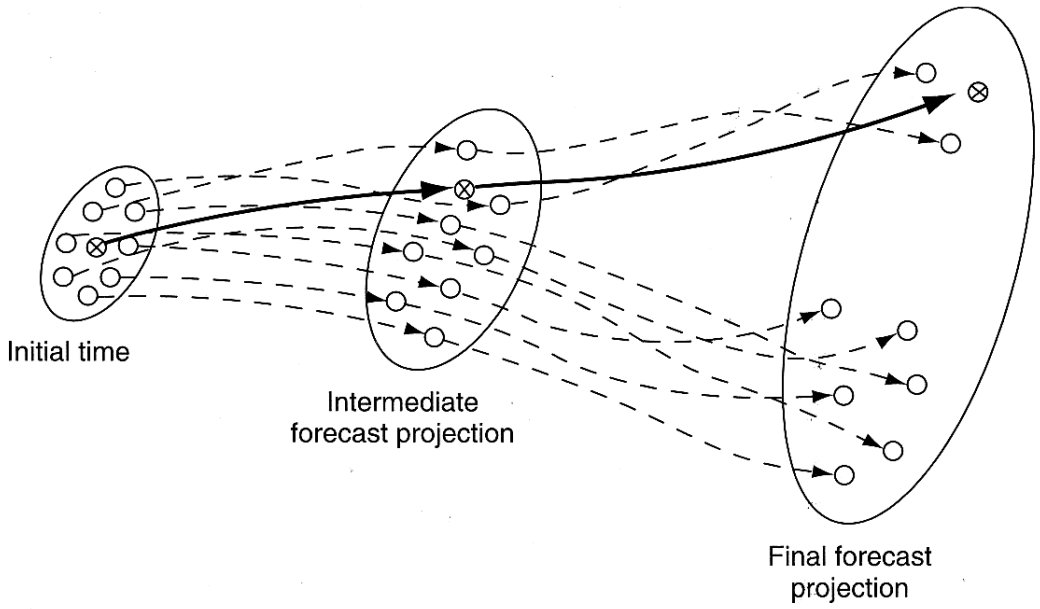

Figure 3.1: Schematic illustration of the basic concepts in ensemble forecasting plotted in terms of a two dimensional phase-space. The initial, an intermediate and the final time of the forecast are represented through ellipses. A total of eight ensemble members and one single best analysis of the initial state has been run sampling from the probability distribution of initial state (the smallest ellipse). Their evolution in time is depicted by means of dashed lines for the eight members and the heavy solid line for the single best analysis. The last ellipse provides the probability distribution of future time uncertainty. From Wilks (2006).

called the ensemble of the initial conditions and each represents a plausible initial state of the atmosphere consistent with the uncertainty in observation and analysis. The evolution of the single best forecast in the phase space, through an intermediate and then a final forecast projections, is represented by the heavy solid lines. However, the position of this point in phase space at the initial time represents only one of the many plausible initial states, which sample the probability distribution for states of the atmosphere at the initial time. The Monte-Carlo approximation to a stochastic dynamic forecast is constructed by repeatedly running the NWP model, once for each member of the initial ensemble. The trajectories through the phase space are only modestly different at first, indicating that all nine NWP integrations are producing fairly similar forecasts at the intermediate position. Accordingly, the probability distribution describing uncertainty about the state of the atmosphere at the intermediate projection would not be a great deal larger than at initial time. However, between the intermediate and the final projections the trajectories diverge markedly, with three (including the mean value of the initial distribution) producing forecasts that are similar to each other, and the remaining six members of the ensemble predicting rather different atmospheric states at that time. The underlying distribution of uncertainty that was fairly small at the initial time has been stretched substantially, as represented by the large ellipse at the time of the final projection. The dispersion of the ensemble members at that time allows the nature of that distribution to be estimated, and is indicative of the uncertainty of the forecast, assuming that the NWP model includes only negligible errors in the representations of the governing physical processes. If only the single forecast started from the best initial condition had been made, this information would have not been available.

Given the high computational time required by running the NWP for each ensemble member of the initial state, the choice of good initial state members is crucial. Without describing in detail the theory behind the choice of these members, here we say that three operational methods for the generation of medium-range initial condition ensembles have been developed in the most important meteorological offices all over the world. The U.S. National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) seek directions of rapid error growth in selective sampling procedures, known as the bred-vector perturbation method (Toth & Kalnay 1993) and the singular-vector technique (Molteni et al. 1996), respectively. The Meteorological Service

of Canada (MSC) uses the Monte Carlo like perturbed-observation approach (Houtekamer et al. 1996), in which the model physics parameterizations vary as well.

In case of precipitation fields predicted by downscaling models, the notion of *ensemble* is slightly different from the one commonly used in atmospheric science. In fact, downscaled precipitation fields are not the output of a model where initial conditions have been perturbed, but they represent a set of statistical realizations of future precipitation corresponding to the same initial condition at the coarse scale. Obviously, this initial condition should be affected by a low level of uncertainty, being, in a certain sense, a deterministic quantity; otherwise, additional uncertainty is added to the forecasts.

3.2 Verification of Ensemble Forecast

3.2.1 Basic Concepts of Forecast Verification

Forecast verification is the process of assessing the quality of the forecasts. This process has been more rigorously developed in atmospheric sciences and only recently it has become the object of systematic studies in hydrology.

All the verification techniques involve measures of the relation between forecasts and the corresponding observation(s) of the predictand. They all study the joint distribution of forecasts and observation (Murphy & Winckler 1987). In practical settings, both the forecasts and observations are discrete variables (or continuous variables that are rounded to a finite set of values). Denote the forecast by y_i, which can take on any of the I values $y_1, y_2, \ldots, y_I$, and the corresponding observation as o_j, which can take on any of the J values $o_1, o_2, \ldots, o_J$. Then the joint distribution of the forecasts and observations is denoted by

$$p(y_i, o_j) \;=\; \Pr\{y_i \cap o_j\} \qquad i = 1, \ldots, I; \;\; j = 1, \ldots, J. \tag{3.1}$$

This is a discrete bivariate probability distribution function, associating a probability with each of the $I \times J$ possible combinations of forecast and observation.

In order to allow the use of the joint distribution, which can be very complicated also in the simplest case $I = J = 2$, the definition of conditional

probability is used to factorize equation (3.1) in two possible ways that are informative about different aspects of the verification problems. The first is called the calibration-refinement factorization (Murphy & Winckler 1987):

$$p(y_i, o_j) \;=\; p(o_j \mid y_i) \cdot p(y_i) \qquad i = 1, \ldots, I; \;\; j = 1, \ldots, J. \tag{3.2}$$

One part of this factorization consists of a set of the I conditional distributions, $p(o_j \mid y_i)$, each of which consists of probabilities for all the J outcomes o_j, given one of the forecast y_i. That is, each of this conditional distribution specifies how often each possible weather event occurred on those occasions when the single forecast y_i was issued, or how well each forecast y_i is calibrated. The other part of this factorization is the unconditional (marginal) distribution $p(y_i)$, which specifies the relative frequencies of use of each of the forecast values y_i, or how often each of the possible forecast values were used. This marginal distribution is sometimes called the predictive distribution, or the refinement distribution of the forecasts.

The other factorization of the joint distribution of forecasts and observations is the likelihood-base rate factorization (Murphy & Winckler 1987)

$$p(y_i, o_j) \;=\; p(y_i \mid o_j) \cdot p(o_j) \qquad i = 1, \ldots, I; \;\; j = 1, \ldots, J. \tag{3.3}$$

Here the conditional distributions $p(y_i \mid o_j)$ express the likelihood that each of the allowable forecast values y_i would have been issued in advance of each of the observed weather event o_j. The unconditional distribution $p(o_j)$ consists simply of the relative frequencies of the J weather events o_j in the verification data set. This distributions is called the sample climatology.

As mentioned before, the use of this distribution can be very difficult also in the simplest circumstance of $I = J = 2$. Therefore, it is traditional to summarize forecast performance using one or several scalar (one dimensional) verification measures. Clearly, these scalar measures do not provide a full picture of the joint distribution and capture only some of its properties. In the following list, the main properties or attributes of forecast quality are described:

1. *Accuracy* refers to the average correspondence between individual pairs of forecasts and the event they predict. Scalar measures of accuracy

are meant to summarize, in a single number, the overall quality of a set of forecasts.

2. *Bias* measures the correspondence between the average forecast and the average observed value of the predictand.

3. *Reliability* pertains to the relationship of the forecast to the average observation, for specific values of the forecast. In other words, reliability measures summarize the I conditional distributions $p(o_j \mid y_i)$.

4. *Resolution* refers to the degree to which the forecasts sort the observed events into groups that are different from each other. It is related to reliability since it provides a measure of the properties of $p(o_j \mid y_i)$, but it differs from reliability because it pertains to the differences between the conditional averages of the observations for different values of the forecast.

5. *Discrimination* is the converse of resolution, in that it pertains to differences between the conditional averages of the forecasts for different values of the observation. Measures of discrimination characterize the conditional distributions of the forecasts given the observation $p(y_i \mid o_j)$.

6. *Sharpness* or *refinement* is an attribute of the forecast alone and characterizes the unconditional distribution of the forecasts $p(y_i)$. Sharp forecast are frequently much different from the climatological value of the predictand. They are accurate only if they also exhibit good reliability.

A last concept that it is worthy to mention is the forecast skill, which refers to the relative accuracy of a set of forecasts with respect to some set of standard control, or reference, forecasts. Common choices for the reference forecasts are the climatological values of the predictand, the persistence (values of the predictand in the previous time period), or random forecasts (with respect to the climatological relative frequencies of the forecast events o_j). Forecast skill is usually presented as a skill score, which is interpreted as a percentage improvement over the reference forecasts. In generic form, the skill score for forecasts characterized by a particular measure of accuracy A, with respect to the accuracy A_{ref} of a set of reference forecasts, is given by:

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \times 100\% \tag{3.4}$$

where A_{perf} is the value of the accuracy measure that would be achieved by perfect forecasts. If $A = A_{perf}$ the skill score attains its maximum value of 100%. If $A = A_{ref}$ then $SS_{ref} = 0\%$, indicating no improvement over the reference forecasts. If the forecast being evaluated are inferior to the reference forecasts with respect to the accuracy measure A, $SS_{ref} < 0\%$.

3.2.2 Verification of Probabilistic Forecasts

The verification techniques described in this section are referred to the simplest circumstance of probability forecasts in relation to dichotomous predictand which are limited to $J = 2$ possible outcomes. The forecast values for the predictand can instead assume a number $I > 2$ values (i.e. probabilities). In theory any real number is an allowable probability forecast, but in practice, the forecast usually are rounded to one of a reasonably small number of values.

When an ensemble model provides values for a continuous predictand (e.g. temperature or precipitation depth), it is possible to achieve this situation by introducing a threshold in the observation: if the observation is greater or smaller than the threshold, then $o_1 = 1$ or $o_2 = 0$ respectively. The probability y_i can be instead calculated as follows. The Empirical Cumulative Density Function (ECDF) of the ensemble sample is built with a plotting position formula and the exceedance probability of the considered threshold is determined and then rounded to the closest y_i ($i = 1, \ldots, I$). For example, our predictand can be: precipitation higher than 0.2 mm h^{-1} and the ensemble model provides the set of values in mm h^{-1} 0, 0, 0, 0.3, 0.35, 0.4, 0.7, 0.8, 0.9, 1. If we observe a precipitation of 0.1 mm h^{-1}, we have $y_i = 0.7$ (the exceedance probability of 0.2 mm h^{-1} within the ensemble sample) and $o_i = o_2 = 0$.

The Brier Score

The most common scalar measure for verification of probabilistic forecasts is the Brier Score (BS). The Brier Score is essentially the mean squared error of the probability forecasts, considering that observation is $o_1 = 1$ if the

event occurs and $o_2 = 0$ is it does not occur. The score averages the squared differences between pairs of forecast probabilities and the subsequent binary observations:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{3.5}$$

where the index k denotes a numbering of the n forecast-event pairs. BS is included in the interval $[0, 1]$ and is negatively oriented, with perfect forecast exhibiting $BS = 0$.

Skill score of the form (3.4) are computed for the Brier Score, leading to the Brier Skill Score:

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{3.6}$$

since $BS_{perf} = 0$.

An instructive algebraic decomposition of the Brier Score has been derived by Murphy (1973) and is related to the calibration-refinement factorization 3.2 of the joint distribution of forecast and observation.

Let be N_i the number of times each forecast y_i, ranging from $y_1 = 0$ to $y_I = 1$, is used in the collection of forecast to be verified. The total number of forecast event pairs is simply the sum of these subsample sizes:

$$n = \sum_{i=1}^{I} N_i \tag{3.7}$$

The marginal distribution of the forecasts consists of the relative frequencies:

$$p(y_i) = \frac{N_i}{n} \tag{3.8}$$

Since the observed event is dichotomous, a single conditional relative frequency defines the conditional distribution of observations given each forecasts y_i and can be expressed as the ratio between the number of o_k falling in the class i (where $o_k = 1$ if the event occurs or $o_k = 0$ if it does not) and the number of times that the forecast value was equal to y_i:

$$\bar{o}_i \;=\; p(o_1 \mid y_i) \;=\; \frac{1}{N_i} \sum_{k \in N_i} o_k \tag{3.9}$$

The sample climatology can be expressed by:

$$\bar{o} \;=\; \frac{1}{n} \sum_{k=1}^{n} o_k \;=\; \frac{1}{n} \sum_{i=1}^{I} N_i \bar{o}_i \tag{3.10}$$

After some algebra, the Brier Score in equation (3.5), can be expressed in terms of the sum of three quantities:

$$BS \;=\; \frac{1}{n} \sum_{i=1}^{I} N_i (y_i - \bar{o}_i)^2 - \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{o}_i - \bar{o})^2 - \bar{o}(1 - \bar{o}) \tag{3.11}$$

The three terms are know as reliability, resolution and uncertainty, respectively. Since BS is negative oriented, the forecaster would like the reliability to be as small as possible, while the resolution to be as large as possible.

For forecast that are perfectly reliable, when $y = y_i$ the correspondent event is observed (i.e. $o_k = 1$) and $y_i = \bar{o}_i$. Therefore, the reliability term is zero. In a real reliable forecast, the relative frequency $\bar{o}_i$ should assume small values when y_i is close to 0, large values when y_i is close to 1 and values close to 0.5 when y_i is close to 0.5.

The resolution term measure the distance between the relative frequency of the forecast and the climatology. Thus, if the forecast sort the observation into subsamples having substantially different relative frequencies than the overall sample climatology, the resolution term will be large.

The uncertainty term depends only on the observations and is not affected by the forecasts. It has minima in 0 and 1, when the climatological probability is 0 or 1, and a maximum when the climatological probability is 0.5.

The Reliability Diagram

The reliability diagram is a graphical device that shows the full distribution of forecasts and observations for probability forecasts of a binary predictand, in terms of the calibration-refinement factorization (3.2).

We remember that we are considering dichotomous predictand and the observation can be $o_1 = 1$ or $o_2 = 0$, while y_i can assume a number I of discrete values.

The first element of a reliability diagram is a plot of the calibration function, with y_i in abscissa and the corresponding $p(o_1 \mid y_i)$ in ordinate. Fig. 3.2a shows five characteristic forms of this portion of the reliability diagram, which allows immediate diagnostic of unconditional and conditional biases eventually showed by the forecast. The center panel shows a well-calibrated forecast, where $y_i \approx p(o_1 \mid y_i)$ apart from sampling variability.

The top and the bottom panels shows the typical pattern of unconditional biases. In the top panel, y_i is always greater than $p(o_1 \mid y_i)$, meaning that forecasts are overestimating probabilities of the predictand (overforecasting or wet bias). On the contrary the bottom panels shows a typical pattern of underforecasting or dry bias.

The deficiencies indicated in the left and the right panels are more difficult to be understood and indicate conditional biases. In these cases, the bias depends on the value of the forecasts themselves. In the left panel (good resolution, underconfident), there are overforecasting biases associated with smaller forecast probabilities and underforecasting biases associated with larger forecast probabilities and the opposite happens in the right panel (poor resolution, overconfident). In this last case, the values $p(o_1 \mid y_i)$ do not depend so much on the forecasts and are all near the climatological probability, revealing poor resolution. Conversely, the model verified in the left panel provides a good resolution because it is able to identify subsamples of forecast occasions for which the outcomes are quite different from each other. Nevertheless, the forecast are not well calibrated, because, for example, no events has occurred for $y = y_1$ or $y = y_I$.

An additional help to better understand the terms underconfident and overconfident is furnished by the second part of reliability diagram: the plot of the refinement distribution $p(y_i)$ (Fig. 3.2b). This plot reflects the overall confidence of the forecaster. Forecasts that do not deviate too much from their average exhibit low confidence, while forecasts spread around their average and assuming too frequently extreme values exhibit high confidence. This last circumstance produces forecasts as depicted in the right panel of Fig. 3.2a: for high values of y_i, we observe less event than expected (and $p(o_1 \mid y_i)$ is not close to 1), while, on the contrary, for lower values of y_i we observe more events than expected (and $p(o_1 \mid y_i)$ is not close to 0). However, the forecast can judge the confidence of the forecast only after the inspection

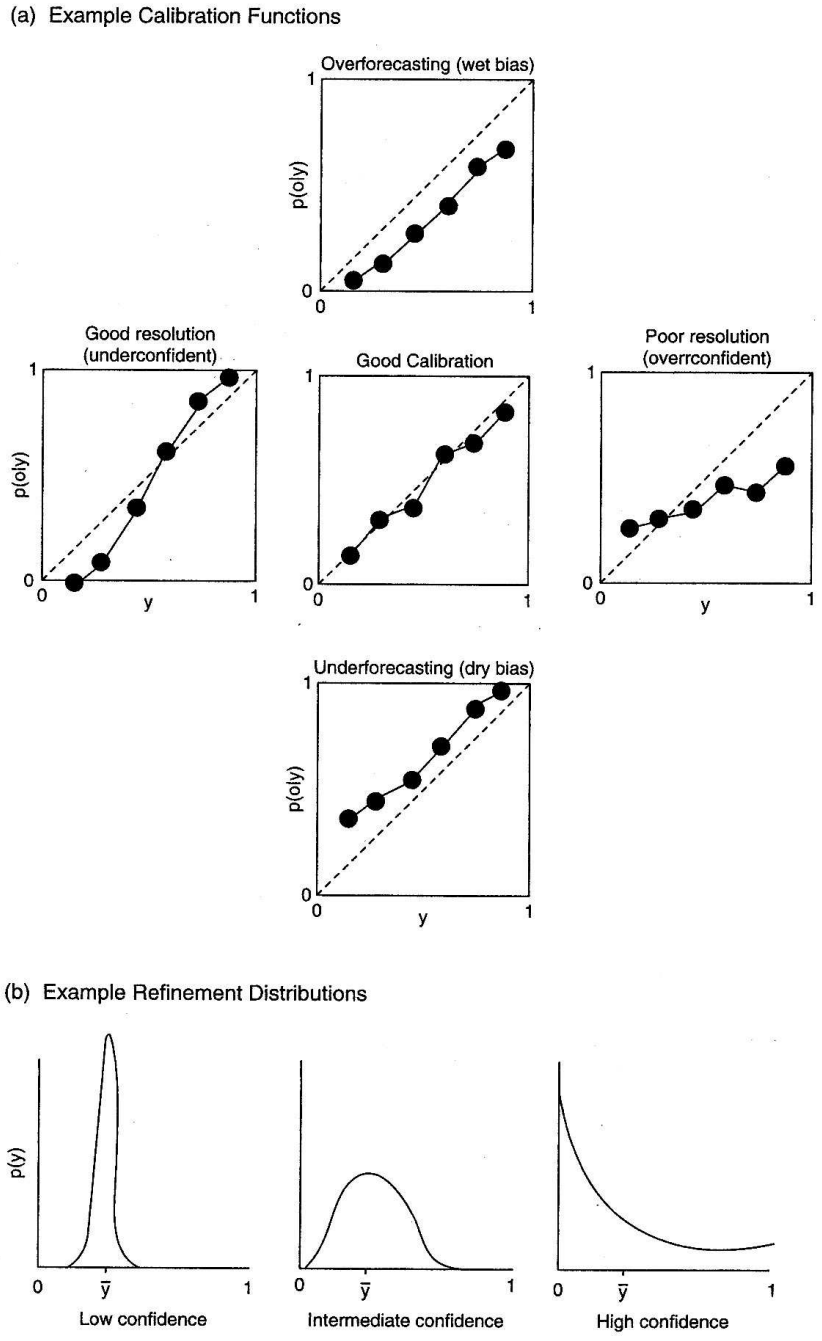

Figure 3.2: Example characteristics forms for the two elements of the reliability diagram. Panel a: calibration functions, showing $p(o_1 \mid y)$, as functions of the forecast y. Panel b: Refinement distribution, $p(y)$, reflecting aggregate forecaster confidence. From Wilks (2006).

of the calibration function for the same forecasts.

The ROC Diagram

The ROC (Relative Operating Characteristics) diagram is a discrimination-based graphical forecast verification display, although unlike the reliability diagram it does not provide a full representation of the joint distribution of forecasts and observations.

In order to illustrate the ROC diagram, it is first necessary to introduce the 2×2 contingency table. This table refers to the simplest dichotomous situation in which $I = J = 2$. In such a case we have $I = 2$ possible forecasts y_1 and y_2 if the event will occur or will not. Similarly, we have $J = 2$ outcomes: the event occurs ($o_1 = 1$) or the event does not occur ($o_2 = 0$). Referring to Fig. 3.3a, the event was successfully forecast to occur a times out of n total forecasts. These a forecast-observation pairs are usually called hits. Similarly, on b occasions, called false alarms, the event was forecast to occur but did not. There also c occasions, called misses, where the events occurred when the model predicted that they would have not occurred. Finally, on d circumstances, called correct rejection, the event was predicted not to occur and it effectively did not. The relative frequencies a/n, b/n, c/n and d/n are the estimates of the joint probabilities $p(y_1 \cap o_1)$, $p(y_1 \cap o_2)$, $p(y_2 \cap o_1)$ and $p(y_2 \cap o_2)$, respectively (Fig. 3.3b).

Two quantities used in the ROC diagram can be calculated from the contingency table. The first one if the hit rate:

$$H = \frac{a}{a+c} \tag{3.12}$$

representing the ratio of correct forecasts to the number of times this event occurred. It also represents the conditional frequency $p(y_1 \mid o_1)$.

The second quantity is the false alarm rate:

$$F = \frac{b}{b+d} \tag{3.13}$$

representing the ratio of false alarms to the total number of nonoccurrences of the event o_1. It also represents the conditional frequency $p(y_1 \mid o_2)$.

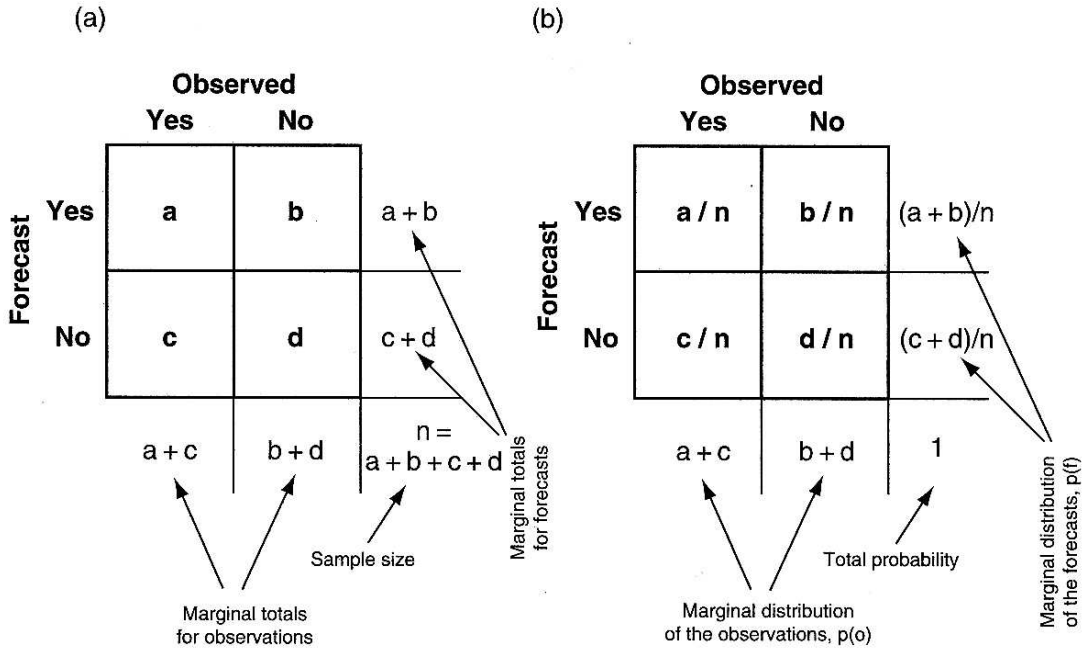

Figure 3.3: Contingency table in the simplest circumstance where $I = 2$ and $J = 2$. From Wilks (2006).

Probabilistic forecasts can be transformed into a categorical yes/no dichotomous forecasts defined by some probability thresholds and the contingency table above described can be built. In case of the ROC diagram, the thresholds are usually chosen equal to each of the values y_i. Then $(I-1)$ 2×2 contingency tables can be built: a yes forecast is imputed if the probability y_i is above the considered threshold and a no forecast is imputed if the probability y_i is below the threshold. ROC diagram is constructed by evaluating each of these $(I-1)$ contingency tables using the hit rate H (equation 3.12) and the false alarm F (equation 3.13). The resulting $(I-1)$ points (F_i, H_i) are then plotted and connected with line segments to each other and to the points $(0,0)$ corresponding to never forecasting the event, and $(1,1)$ corresponding to always forecasting the event.

Fig. 3.4 shows two examples of ROC diagram. The upper left corner represents a perfect forecast system where there are no false alarms and only hits. The closer the point is to this upper left corner the higher the skill. The lower left corner, where both hit and false alarms rate are zero, represents a system which never warns of an event. The upper right corner, represents a system where the event never occurs. A perfect forecast model exhibits always $F = 0.0$ and $H = 1.0$, so its ROC diagram is made of two line segments: the vertical left boundary and the horizontal upper boundary. Conversely, a forecast model with bad performance, for example a model providing random forecasts, is characterized by $F_i = H_i$ and the ROC diagram is given by the 1:1 line (dotted line). ROC curves for real forecasts generally fall between these two extremes (heavy solid lines). The better the forecast, the closer the ROC curve to the upper-left corner.

It can be convenient to summarize the ROC diagram through a scalar measure, given by the area A under the ROC curve. A perfect model has $A_{perf} = 1$ while random forecasts will have $A_{rand} = 0.5$. The skill score of ROC diagram can then be calculated as:

$$SS_{ROC} = \frac{A - A_{rand}}{A_{perf} - A_{rand}} = \frac{A - 1/2}{1 - 1/2} = 2A - 1 \qquad (3.14)$$

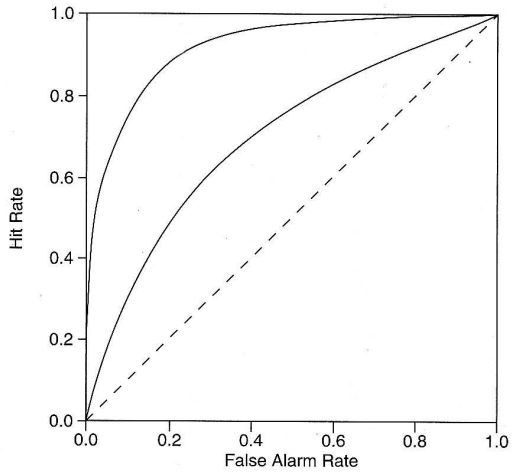

Figure 3.4: Example of two ROC diagrams. From Wilks (2006).

3.2.3 Consistency Hypothesis and Ensemble Dispersion

A consistent research effort has been focused by the meteorologic scientific community to verify the goodness of ensemble forecast. An important property for which verification methods have been developed is the consistency of the ensemble (Anderson 1997), which is the degree to which the observed state is a plausible member of the forecast ensemble.

Referring to meteorological models, if the initial state members have been chosen as a random sample from the initial-condition uncertainty Probability Distribution Function (PDF), and if the forecast model contains an accurate representation of the physical dynamics, the dispersion of the ensemble forecast represents a random sample from the PDF of forecast uncertainty. In this ideal situation, the true state of the atmosphere would be just one of the ensemble members and should be statistically indistinguishable from the forecast ensemble. Referring to precipitation downscaling models, consistency would also occur in the ideal situation where the model provides the exact probability distribution of precipitation at high resolution from which the observation is drawn.

These circumstances where the actual future atmospheric state or the future rainfall scenario behave like random draw from the same distribution

that produced the ensemble is called consistency of the ensemble.

Ensemble forecasts are probability forecasts that are expressed as a discrete approximation of a full forecast PDF. According to this approximation, ensemble relative frequency should estimate actual probability. Probability forecasts can be obtained for simple predictands, such as continuous scalar (e.g., temperature or precipitation at a single location), or discrete scalars (possibly constructed by thresholding a continuous variable, e.g., zero precipitation vs nonzero precipitation at a single location); or quite complicated multivariate predictands such as entire fields (e.g., the joint distribution of 500 mb heights at the global set of horizontal gridpoints). In any of these cases, the probability forecast from an ensemble will be good if consistency condition has been met and the observation is statistically indistinguishable from the ensemble.

A necessary condition for ensemble consistency is an appropriate degree of ensemble dispersion. If the ensemble dispersion is consistently too small, then the observation will be often an outlier in the distribution of the ensemble members, implying that ensemble relative frequency will be a poor approximation to probability. This condition of ensemble underdispersion is illustrated hypothetically in Fig. 3.5a. If the ensemble is consistently too large, as in Fig. 3.5c, then the observation may too often be in the middle of the ensemble distribution, leading again to a poor approximation of probability. If the ensemble distribution is appropriate, as illustrated by the hypothetical example in Fig. 3.5b, then the observation may have an equal chance of occurring at any quantile of the distribution that is estimated by the ensemble.

Once probability forecasts are estimated from a forecast ensemble by adopting a plotting position rule, the appropriateness of these probability assignments can be investigated through techniques of forecast verification for probabilistic forecasts. The main verification methods for probabilistic prediction been previously summarized in section 3.2.2. However, additional verification tools have been developed specifically for ensemble forecasts, many of which are aimed at testing the consistency hypothesis. The most commonly used technique is the Verification Rank Histogram.

3.2.4 The Verification Rank Histogram

The Verification Rank Histogram (VRH) is a graphical tool used for a single scalar or univariate predictand. The underlying idea of VRH is rather

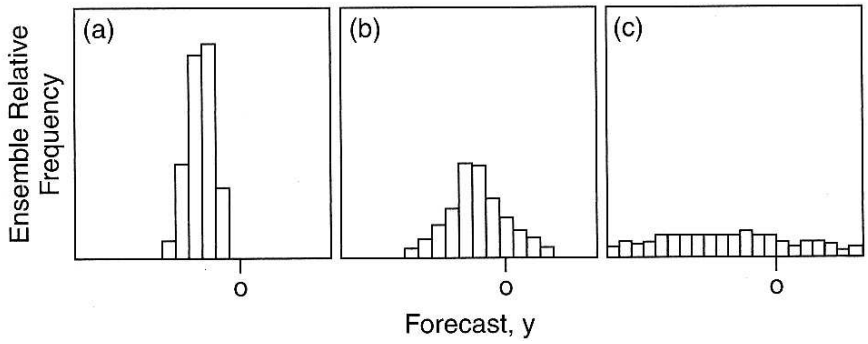

Figure 3.5: Histogram of hypothetical ensembles producing a continuous scalar, y, exhibiting relatively (a) too little dispersion, (b) an appropriate degree of dispersion, and (c) excessive dispersion, in comparison to a typical observation o. From Wilks (2006).

simple. Let S be the univariate variable to be forecasted. For each event, the ensemble model provides N_{ens} forecasts $S_1, S_2, \ldots, S_{N_{ens}}$, to predict the corresponding observation S_{obs}. If forecasts and observations are drawn from the same distribution, the rank of S_{obs} within the sorted vector $\mathbf{S}$ containing $S_1, \ldots, S_{N_{ens}}$ and S_{obs} assumes equally likely the values $1, 2, \ldots, N_{ens} + 1$. Therefore, if N_{ev} events are analyzed and ranked, the histogram built with these ranks should be uniform. Any departure from uniformity should only be due to sampling variability.

Special rules are used for assigning ranks when many ensemble members have the exact same value as the verification, as may occur, for example, with no precipitation forecast and none observed. Hamill & Colucci (1997, 1998) analyzed this problem and proposed to assign to the observation a random rank between the first and the last rank of the subsample made of ensemble members and observation with the same value.

If the VRH is not uniform, the assumptions underlying the ensemble forecasts have not been met. Fig. 3.6 shows the possible patterns of the VRH, where $N_{ens} = 8$. In each histogram, the horizontal dashed line represents the mean of a uniform distribution for that number of sample (i.e. $= (N_{ens} + 1)^{-1}$). Positive or negative unconditional biases produce overpopulation of the lowest or highest ranks and the resulting histogram are shown in the upper and in the lower panel (overforecasting and underforecasting bias,

respectively). An excess of dispersion (overdispersion) implies overpopulation of the middle ranks (left panel of Fig. 3.6): the observation is located more frequently in the middle of the ensemble members, like depicted by Fig. 3.5c, and it very rarely falls in the extremes. Conversely, a lack of variability (underdispersion) determines U-shaped histograms (right panel of Fig. 3.6). In this case, the ensemble members tends to be too much like each other and different from the observation, which occupies more frequently the extreme ranks, as in Fig. 3.5a.

Particular care should be made to interpretation of VRH, since a uniform rank histogram is a necessary but not sufficient condition for consistency. Indeed, as shown by Hamill (2001), the same histogram shape can be obtained by combining the effects of different ensemble model deficiencies, making the diagnosis of ensemble forecasts characteristics difficult. Hamill (2001) showed this effect by means of numerical experiments where he adopted normal distributions with different mean and standard deviation to simulate the truth and the ensemble model characteristics. He showed, for example, that if the true is drawn by a normal distribution with mean 0 and standard deviation 1 ($N(0,1)$) and the ensemble members are drawn with equal likelihood from either normal distributions $N(-0.5,1)$, $N(+0.5,1)$ and $N(0,1.3)$, the resulting VRH is uniform even if the ensemble is not consistent. For further examples and consequent discussion illustrating this important point, the reader is referred to Hamill (2001).

The VRH provides a measure of reliability or conditional bias of the forecast and, in this sense, it has connection with the calibration function $p(o_j \mid y_i)$. We have previously shown that the use of a threshold allows the probability y_i $(i = 1, \ldots I)$ and the correspondent observation o_j $(j = 1, 2)$ to be computed for each event from the set of ensemble forecasts $S_1, \ldots, S_{N_{ens}}$. Repeating this calculation for all the events, we can calculate the relative frequencies $p(y_i)$ and the conditional frequencies $p(o_1 \mid y_i)$ $(i = 1, \ldots, I)$ and then plot the reliability diagrams. A correspondence one-to-one can be made between the shapes of the histogram illustrated in Fig. 3.6 and the shape of the reliability diagrams shown in Fig. 3.2. When ensemble overforecasting is detected, the ensemble are too frequently centered above the verification with a majority of members above the given threshold more frequently than the observation is above that threshold. The opposite happens for the underforecasting deficiency.

In underdispersed ensembles, most or all the members will fall too frequently on one side or the other of the threshold defining a dichotomous

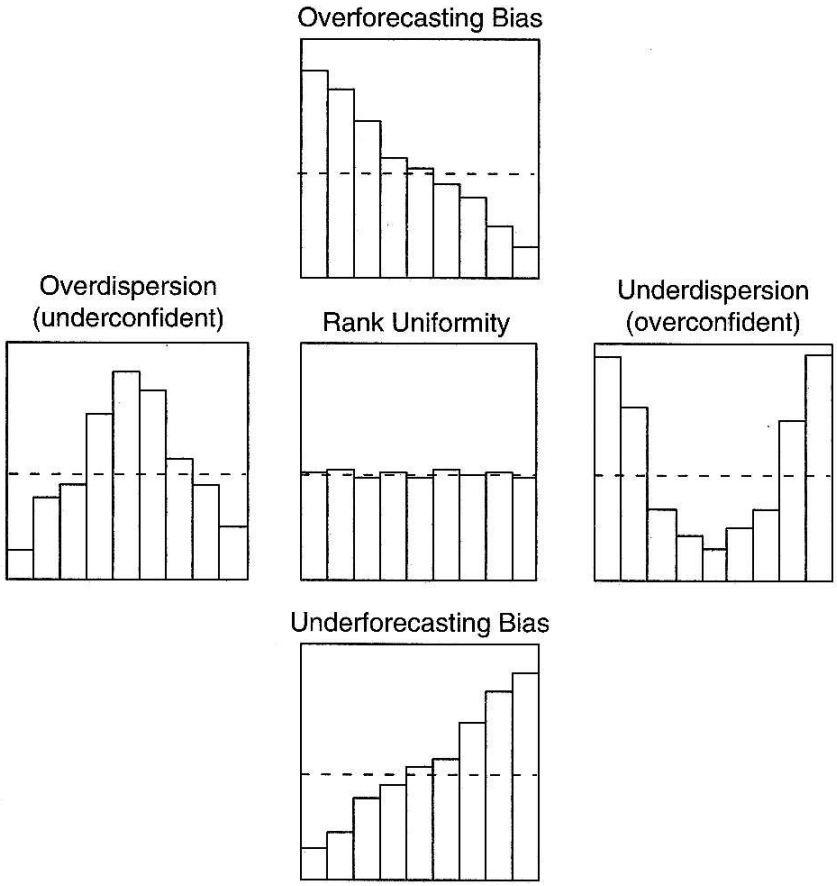

Figure 3.6: Example of Verification Rank Histogram (VRH) constructed from an ensemble with size $N_{ens} = 8$. The typical patterns of uniformity, overforecasting and underforecasting bias, overdispersion and underdispersion are shown. The horizontal dashed line in each histogram is the mean of the uniform distribution, equal to $(N_{ens} + 1)^{-1}$. From Wilks (2006).

event. The result is that probability forecast will be excessively sharp and the observation will use extreme probabilities too frequently. The probability forecast will be overconfident and the conditional event relative frequencies are less extreme than the forecast probabilities. On the other hand, overdispersed ensembles will rarely have most members on one side or the other of the event threshold and the probability forecast derived from them will rarely be extreme. These probability forecasts will be underconfident and the conditional event relative frequencies tend to be more extreme than the forecast probabilities.

The VRH does not provide a full evaluation of forecast performance, since it does not account for all the elements of the joint distribution of forecast and observation. In particular, it is not able to give information about the refinement or sharpness of ensemble forecasts, but it indicates only if the forecast refinement is appropriate relative to the degree to which the ensemble can resolve the predictand.

Applications of the VRH to meteorological model outputs are provided by Hamill & Colucci (1997, 1998) who tested the reliability of single scalar outputs predicted by Eta-RSM short range model.

In this work, the VRH has been used as the basis to develop specific verification methods to test consistency of two ensemble outputs of a hydrometeorological system: (i) precipitation fields generated by downscaling models and (ii) hydrographs outputted by hydrological models.

Chapter 4

Precipitation Downscaling Model Verification

The chapter is devoted to the assessment of uncertainty associated to ensemble precipitation forecasts produced by downscaling models. A verification method aimed at testing consistency of downscaled precipitation ensemble is first described in detail and then applied in three numerical hindcast experiments where the STRAIN multifractal downscaling model is used. Experiments allow testing the working of the verification procedure and drawing general conclusions about downscaling model performances when different calibration modes are adopted.

The chapter is organized as follows. In section 4.1, we propose the verification procedure: the determination of the precipitation exceedance probability of downscaled rainfall fields is first described, followed by the description of how the Verification Rank Histogram (VRH) is built for the exceedance probability; finally, we discuss the presence of randomly assigned ranks that artificially affect histogram shape and propose a graphical method for their interpretation. In section 4.2, we describe the three hindcast experiments and illustrate results, while in section 4.3 we discuss conclusions.

4.1 Methods

4.1.1 Construction of Rank Histograms to Test Precipitation Exceedance Probability

Let us consider a spatiotemporal precipitation field at fine resolution $\lambda \times \lambda \times \tau$ included in a coarse domain $L \times L \times T$, as depicted in (Fig. 4.1a). Assuming isotropy in space and time, the exceedance probability $S(i^*) = \Pr\{I > i^*\}$ of a fixed threshold i^* can be calculated from the entire set of the M high resolution precipitation values in each $\lambda \times \lambda \times \tau$ grid-cell. $S(i^*)$ can be derived by the Empirical Survival Function (*ESF*) (Evans et al. 2000) of the rainfall rates i_j in each grid-cell j, $(j = 1, \ldots, M)$, without reference to their position in the cube. The *ESF* can be estimated using the ranks of the order statistics $i_{(j)}$ as the complement of the Empirical Cumulative Frequency $F(i_{(j)})$:

$$S(i_{(j)}) = 1 - F(i_{(j)}) = 1 - \frac{j - 0.5}{M} \qquad j = 1, \ldots, M \tag{4.1}$$

where $F(i_{(j)})$ is estimated through Hazen plotting position formula. As shown in Fig. 4.1b, the exceedance probability $S(i^*)$ corresponding to a generic precipitation threshold i^* is computed by linear interpolation of the two closest values of $S(i_{(j)})$, when $i_{(1)} \leq i^* \leq i_{(M)}$, while it is set to 1 or 0 when i^* is smaller than $i_{(1)}$ or greater than $i_{(M)}$. In the case of heterogeneous rainfall fields which cannot be homogeneized by a modulating function, the *ESF* should be built only with rainfall rates in the verification location.

The construction of a VRH for the exceedance probabilities of a fixed threshold i^* is straightforward. For each event to be forecasted, $(N_{ens} + 1)$ *ESFs* can be built, (i.e. N_{ens} from the ensemble members and 1 from the observed or verification event). The correspondent $(N_{ens} + 1)$ exceedance probabilities of the selected threshold i^*, $S_1(i^*), S_2(i^*), \ldots, S_{N_{ens}}(i^*)$ and $S_{obs}(i^*)$, can be calculated and sorted in increasing order in a vector indicated with $\mathbf{S}$. Finally the position of $S_{obs}(i^*)$ in the vector $\mathbf{S}$ is tabulated. This procedure is repeated for all the N_{ev} events obtaining N_{ev} ranks. In previous applications, the VRH is constructed by plotting the integer ranks from 1 to $(N_{ens} + 1)$. Here, we modify this approach by introducing a normalized rank r defined as the Empirical Cumulative Frequency of $S_{obs}(i^*)$ in the sample $\mathbf{S}$, estimated using Hazen plotting position formula:

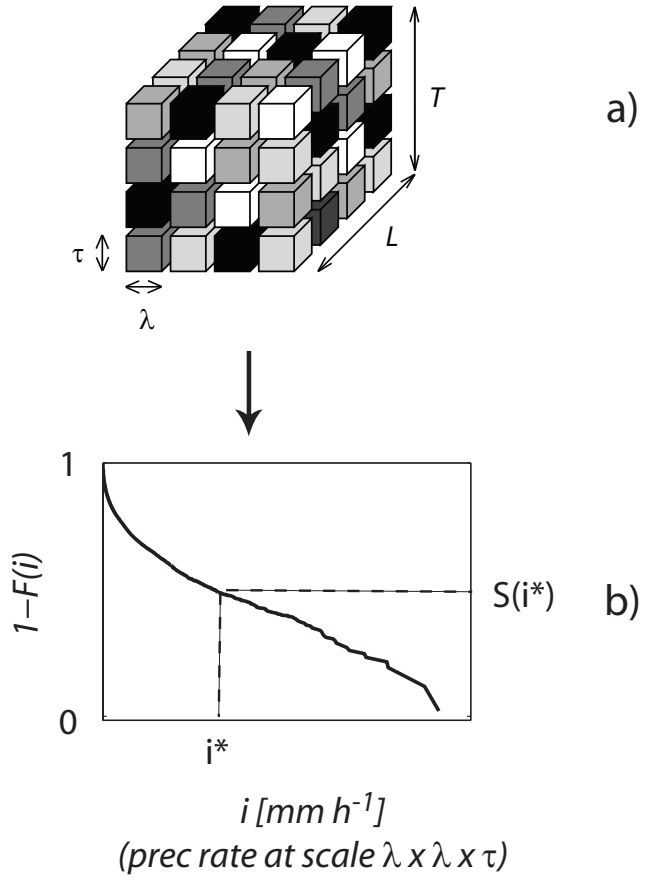

Figure 4.1: Panel a: spatiotemporal grid of a high resolution precipitation field (observed or simulated). Panel b: determination of the exceedance probability $S(i^*)$ from the Empirical Survival Function built with the entire set of precipitation rates i_j, $(j = 1, \ldots, M)$ at the fine scale $\lambda \times \lambda \times \tau$ (case of spatial homogeneity).

$$r = \frac{p - 0.5}{N_{ens} + 1} \tag{4.2}$$

where p is the position of $S_{obs}(i^*)$ within the sample $\mathbf{S}$. The use of this plotting position formula assures that the probability values assigned to the normalized rank are uniformly distributed in the interval $(0, 1)$ regardless N_{ens}. This modification does not lead to conceptual difference with the standard procedure and it allows keeping track of portion of the VRH that has been randomly assigned, according to the graphical method proposed in the following subsection.

From a fixed precipitation threshold i^*, it is possible to build a VRH testing the exceedance probability of that threshold. The procedure can be repeated to test several precipitation thresholds of interest in the hydrometeorological forecasting system.

4.1.2 Rank Assignment and Histogram Interpretation

For high values of the threshold i^*, it is possible that the exceedance probabilities for the observed event and for one or more ensemble members are equal to zero. In such a case, the position p of $S_{obs}(i^*)$ in the vector $\mathbf{S}$ is randomly assigned. When this occurs for several events, the VRH will be populated by many random values and its shape will be artificially affected, disguising the presence of model forecast deficiencies, if present. Therefore, in this subsection, we first discuss whenever the position p of $S_{obs}(i^*)$ can be unequivocally or randomly assigned and then we propose a graphical method providing guidance for histogram interpretation.

The unequivocal or random assignment for the position p is illustrated in Fig. 4.2, where each panel compares the *ESF* of the observed precipitation field (in black) with the ensemble *ESFs* of synthetic fields (in gray) predicted by downscaling models with different forecast skills. An important issue that affects the determination of p is the value of i^*, which can be smaller or greater than the maximum observed precipitation value. For this reason the six panels are divided into two groups (left and right panels) where different precipitation thresholds i_a^* and i_b^* are used:

1. In panels a, b and c, the precipitation threshold i_a^* is smaller than the maximum observed precipitation value, so that S_{obs} is always

greater than 0 and its position p within the vector $\mathbf{S}$ is unequivocally determined in all the 3 cases. In particular, in panel a, S_{obs} is included between S_{min} and S_{max}, which are the minimum and maximum exceedance probabilities of the ensemble members (i.e. $1 < p < (N_{ens} + 1)$). In panels b and c, S_{obs} occupies the first and the last position in the sorted vector $\mathbf{S}$ (i.e. $p = 1$ and $p = (N_{ens} + 1)$).

2. In panels d, e and f, the precipitation threshold i_b^* is greater than the maximum observed precipitation value, so that S_{obs} is always equal to 0. In this case, if one or more exceedance probabilities of the ensemble members are also equal to 0, the position p in $\mathbf{S}$ is randomly assigned following a similar approach as in Hamill & Colucci (1998). In order to better illustrate this, let us indicate with N_0 the number of ensemble *ESFs* for which the exceedance probability is 0. In panel d, $N_0 = N_{ens}$ and $\mathbf{S}$ is a sequence of zero values. The position p of S_{obs} in vector $\mathbf{S}$ takes randomly one of the integer values $1, 2, ..., (N_{ens} + 1)$. In panel e, $1 \leq N_0 < N_{ens}$, so that $\mathbf{S}$ contains a sequence of (N_0+1) elements equal to zero and $(N_{ens} - N_0)$ values greater than 0. The position p of S_{obs} is again randomly assigned, but within the limited number of integers $1, 2, \ldots, N_0 + 1$. Finally, in panel f, $N_0 = 0$ and S_{obs} is unequivocally placed in the first position of $\mathbf{S}$, being the only element equal to zero.

Note that the position may also be randomly assigned in the cases depicted in panels a, b and c when $S_{obs} > 0$ and other ensemble exceedance probabilities have the exact value as S_{obs} (Hamill & Colucci 1998), but this rarely occurs and thus does not affect the shape of the resulting histograms. For this reason, we do not consider random assignments occurring when $S_{obs} > 0$.

We remark that for small values of i^*, the ranks are usually unequivocally determined as in Fig. 4.2a, b and c. As the threshold i^* increases, the chance of encounter $S_{obs} = 0$ and some $S_j = 0$ in the ensemble members increases (Fig. 4.2d and e). Thus, depending on N_0 values, a smaller or larger portion in the left side of the VRH will be randomly populated, making the detection of model forecast deficiencies more difficult. Therefore, it is convenient to store the occurrence of unequivocal and random assignments for the whole set of N_{ev} events used to verify the model.

For this purpose, we can associate to the VRH the Empirical Cumulative Density Function (*ECDF*) of a variable $\tilde{r}_k$ calculated for each forecasted precipitation event $k = 1, \ldots, N_{ev}$ and defined as:

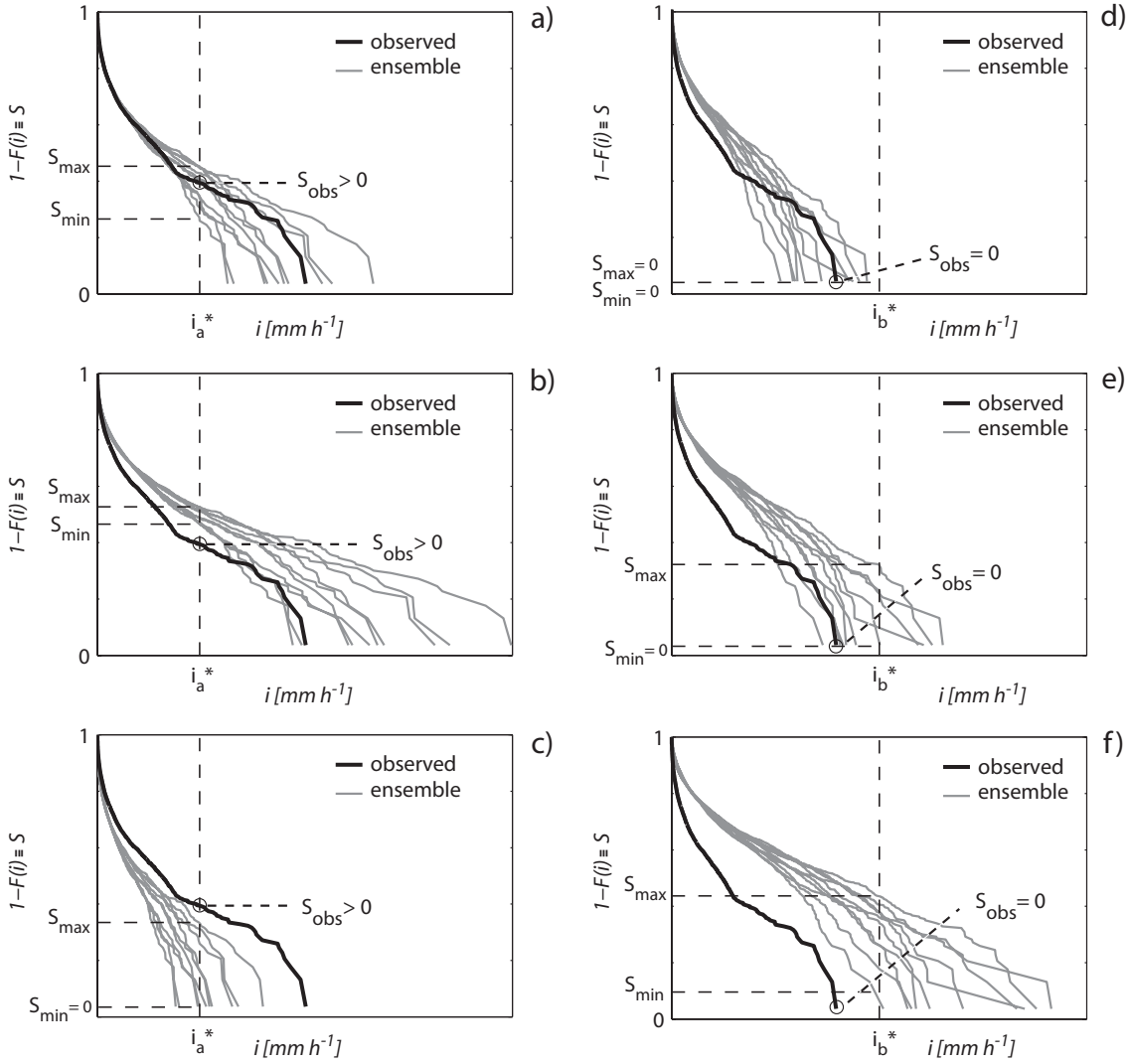

Figure 4.2: Determination of the position p of the observed exceedance probability S_{obs} within the vector $\mathbf{S}$ containing S_{obs} and the ensemble exceedance probabilities S_j ($j = 1, \ldots, N_{ens}$) sorted in increasing order. Each panel shows the Empirical Survival Function (*ESF*) of the same observed event (in black) forecasted by different downscaling models (whose *ESFs* are plotted in gray). Panels are divided into two groups: (i) in panels a, b and c, $S_{obs} > 0$ and p is always unequivocally assigned; (ii) in panels d, e and f, $S_{obs} = 0$. If a number $N_0 \geq 1$ of S_j are also equal to 0, p is randomly determined among the integers $1, \ldots, (N_0 + 1)$ (panels d and e, where $N_0 = N_{ens}$ and $1 \leq N_0 < N_{ens}$ respectively). If $N_0 = 0$ (panel f), $p = 1$.

$$\tilde{r}_k = \begin{cases} 0 & S_{obs}(i^*) > 0 \\ & \quad or \\ & \left(S_{obs}(i^*) = 0 \quad and \quad N_0 = 0\right) \\ \dfrac{(N_0+1) - 0.5}{N_{ens}+1} & S_{obs}(i^*) = 0 \quad and \quad 1 \leq N_0 \leq N_{ens} \end{cases} \tag{4.3}$$

The variable $\tilde{r}_k$ is set to 0 when the rank is determined in unequivocal way: this happens either if $S_{obs}(i^*) > 0$ irrespective of the values of the ensemble members (Fig. 4.2a, b and c) or if $S_{obs}(i^*) = 0$ and all the exceedance probabilities of the ensemble members are non zero, thus $N_0 = 0$ (Fig. 4.2f).

On the contrary, $\tilde{r}_k$ assumes a positive value when $S_{obs}(i^*) = 0$ and there is at least one ensemble member with a zero exceedance probability of precipitation ($1 \leq N_0 \leq N_{ens}$), so that a random assignment occurs (Fig. 4.2d and e). In this case, $\tilde{r}_k$ is defined as the cumulative frequency of the $(N_0+1)^{th}$- zero value in vector $\mathbf{S}$, since, accordingly to equation (4.2), the normalized rank r will be randomly determined within the interval $(0, \tilde{r}_k)$. As $\tilde{r}_k$ increases, the interval $(0, \tilde{r}_k)$ becomes wider until it reaches the whole range $(0, 1)$.

The *ECDF* of the sample $\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_{N_{ev}}$ provides a graphical summary of how the normalized rank has been assigned on the entire set of N_{ev} precipitation events. Fig. 4.3 shows how the *ECDFs* of $\tilde{r}_k$ may change as the precipitation threshold i^* increases. For lower values of i^*, no random assignment occurs, thus all the $\tilde{r}_k$ are equal to 0 and the *ECDF* represents an impulse concentrated on 0 (case A). As i^* increases, part of the ranks may be randomly assigned and the *ECDF* contains some of the $\tilde{r}_k$ values equal to 0 and the others greater than 0 (cases B and C). In particular, in *ECDF B*, the normalized ranks are randomly assigned in intervals $(0, \tilde{r}_k)$ with $\tilde{r}_k < 1$, while in *ECDF C* there are also some $\tilde{r}_k = 1$, meaning that the corresponding ranks are randomly assigned in the whole interval $(0, 1)$. If i^* further increases, all the N_{ev} ranks may be randomly determined and the corresponding $\tilde{r}_k$ result always greater than 0 (cases D and E). *ECDF* D refers to the situation where part of the ranks are randomly assigned in intervals $(0, \tilde{r}_k)$ with $\tilde{r}_k < 1$ and part in the whole interval $(0, 1)$, while *ECDF* E represents the extreme case where all the ranks are randomly determined in the interval $(0, 1)$. Although the VRH is populated only by uniformly distributed random values, this last

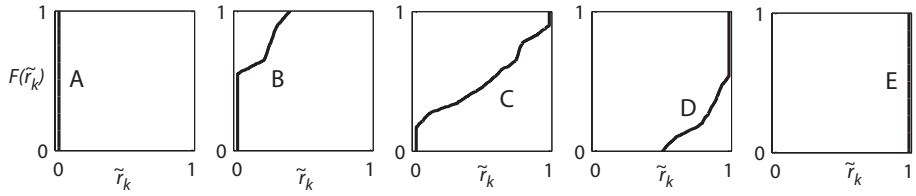

Figure 4.3: Possible behaviors of the Empirical Cumulative Density Function of the variable $\tilde{r}_k$ ($k = 1, \ldots, N_{ev}$) calculated applying the verification procedure for increasing precipitation thresholds i^* (from left to right).

case indicates that precipitation values at the fine scale greater than or equal to the threshold i^* do not represent critical situation (i.e. a zero exceedance probability) for the available observations and ensemble members.

The use of *ECDFs* of $\tilde{r}_k$ permits us to evaluate how much the shape of the histogram depends on real model forecast characteristics and how much it is artificially affected by random assignment. VRHs and *ECDFs* of $\tilde{r}_k$ obtained for the different thresholds should be interpreted together to better detect forecast deficiencies, if any. This aspect is illustrated in the following section, where the verification procedure is applied and tested on synthetic precipitation fields generated by the STRAIN multifractal model.

4.2 Numerical Hindcast Experiments

The verification procedure is applied on three series of experiments based on space-time rainfall events generated by the STRAIN model under controlled conditions, with the aim to better understand and correctly interpret the verification results and to evaluate its capability in detecting downscaling model deficiencies when different calibration approaches are adopted. Hindcast experiments are carried out in the following way. First, we generate a set of N_{ev} precipitation events applying the STRAIN model with selected values of the c and β parameters. Although synthetically generated, the N_{ev} events are assumed as 'observed events' and then used to estimate STRAIN parameters to be adopted in the hindcasting phase.

In Experiment 1 (Constant Parameters), the 'observed events' are generated with the constant parameters c_0 and β_0 with the aim of analyzing the influence of intrinsic model sampling variability in parameters

estimation. In Experiment 2 (Single Calibration Relation), 'observed events' are generated from a calibration relation between STRAIN parameters and precipitation rate at the coarse scale, like those presented in previous works and previously mentioned in the paper. Finally, Experiment 3 (Multiple Calibration Relations) is built by mixing 'observed events' coming from two different calibration relations, with the aim to mimic the case of events originated from different meteorological conditions.

4.2.1 Experiment 1: Constant Parameters

Experiment 1 allows us to show how the ensemble downscaled fields can be affected by overdispersion if intrinsic model sampling variability is not taken into account in parameters estimation.

A set of $N_{ev} = 400$ high resolution precipitation events are generated through a Monte Carlo approach by downscaling a coarse precipitation rate of 1 mm h^{-1} over 5 downscaling levels with fixed values of STRAIN model parameters $c_0 = 0.7$ and $\beta_0 = e^{-1}$. Thus, each event is drawn from the same distribution and the downscaling levels vary, for example, from a coarse scale with $L = 128$ km and $T = 5$ h and 20 min to a fine scale with $\lambda = 4$ km and $\tau = 10$ min (Deidda et al. 2004). These 400 events are assumed to be the set of 'observed events'. Subsequently, we do not assume any knowledge of the method used to generate these 'observed events' and we use them first to calibrate STRAIN parameters for the hindcasting phase and then as verification for the ensemble hindcasting members.

The STRAIN parameters are estimated for each 'observed event' in the following way. First, we compute partition functions $S_q(\lambda)$ with equation (2.3) for different λ scales and q moments. Secondly, sample multifractal exponents $\zeta(q)$ are estimated by the slope of the linear regression between $\log S_q(\lambda)$ and $\log \lambda$, for each moment q. Finally, the c parameter is estimated by fitting equation (2.4) to sample multifractal exponents $\zeta(q)$, while the β parameter is kept constant at e^{-1} (Deidda 2000, Deidda et al. 2004).

Let c_k^{est} be the estimate of the c parameter on the k-th 'observed event'. Because of sampling variability, the 400 different c_k^{est} estimates result in a Gaussian-like distribution around a mean value c_{mean} close to c_0 (i.e. a quasi-unbiased estimator).

In order to show the importance of accounting for intrinsic model sampling variability, two approaches called 'event-based' and 'mean-based' calibration modes for determining the downscaling parameters are compared.

The 'event-based' calibration mode is derived from the notion that c should be estimated from the same 'observed event', since, at first sight, this appears the most obvious approach to simulate each event. As a result, for each event k, the parameter c_k^{est} is used to generate the ensemble members. Note that this calibration mode can only be used for hindcasts. If a forecast is required, c_k^{est} is unknown and the parameter should be determined from past events. In contrast, the 'mean-based' calibration mode is based on the average behavior of the entire set of events using the same parameter c_{mean} and thus it is suitable for forecasts.

In both cases, $N_{ens} = 100$ ensemble members are simulated to hindcast each 'observed event' (total of 40,000 synthetic fields for each approach) and VRHs are constructed for thresholds 10, 15, 20, 25, 30, and 35 mm h^{-1} (selected to span the potential range of *ECDFs* of $\tilde{r}_k$ behavior). Results are shown in Fig. 4.4 and 4.5 for the 'event-based' and 'mean-based' calibration modes, respectively. Each panel contains the VRH for a precipitation threshold, plotted using 10 bins to group the 400 ranks, and the respective *ECDF* of $\tilde{r}_k$. To distinguish between true deviations from uniformity and sampling variations, the 5%, 25%, 50%, 75% and 95% quantiles of a uniform distribution are plotted using horizontal lines.

The following results can be summarized for the 'event-based' calibration mode (Fig. 4.4). For the lowest threshold ($i^* = 10$ mm h^{-1}), the *ECDF* of $\tilde{r}_k$ is concentrated on zero implying ranks have been unequivocally determined, while the histogram is more populated in the middle ranks (i.e. overdispersed forecasts). As the precipitation threshold increases ($i^* = 15$, 20 mm h^{-1}), the number of non-zero $\tilde{r}_k$ values increases leading to random assignments in the interval $(0, \tilde{r}_k)$, where $0 < \tilde{r}_k \leq 1$. The small ranks in the histograms are artificially more populated, but overdispersion is still visible. When i^* further increases, ($i^* = 25, 30$ mm h^{-1}), both the number and magnitude of non-zero $\tilde{r}_k$ values increase so that several ranks are randomly assigned and the interval length $(0, \tilde{r}_k)$ becomes wider. This implies that even the high ranks are randomly populated. As extreme case, when the i^* is higher than observed and synthetic precipitation ($i^* = 35$ mm h^{-1}), the standardized ranks are all randomly assigned in the interval $(0, 1)$ leading to a uniform histogram.

In Fig. 4.5 results are shown for the verification of ensemble members generated with the 'mean-based' calibration mode. In this case, the VRHs are uniform despite the expected sampling variability, whatever the value of the precipitation threshold i^*. This means that the consistency condition is

respected.

The effects of overdispersion and uniformity in the histograms resulting respectively from the 'event-based' and 'mean-based' calibration modes can be explained as follows. The consistency condition requires that, for every predicted event, observations and forecast ensemble behave like random draws from the same distribution. This requirements is satisfied in the 'mean-based' mode because ensemble members are generated using the parameter c_{mean} which is close to c_0 (used to generate the 'observed events'). In this situation, the *ESFs* of the 'observed events' are placed equally likely as the ensemble hindcasts, leading to uniform VRHs for the exceedance probabilities.

When ensemble members are generated using the parameter c_k^{est} ('event-based' calibration mode), the consistency condition is not respected because ensemble and observations belong to different distributions and an effect of overdispersion is produced. This effect is explained by a 'centering' of the variability around the event k. Thus, the *ESF* of the observed event k is placed in the center of the ensemble hindcast *ESF*s and the probabilities of exceedance occupy intermediate positions. For this reason, even if we were able to know the value of c_k^{est} in a forecasting framework, results show that the best choice is to generate the ensemble members using c_{mean}.

In conclusion, we highlight the importance of interpreting the VRHs looking also to the *ECDF* of $\tilde{r}_k$. In fact, in the 'event-based' hindcasts, when the lower thresholds are analyzed, the histogram shape is not artificially affected (most of $\tilde{r}_k = 0$) and overdispersion can be detected. As the threshold increases, the histograms shape becomes more uniform and overdispersion cannot be easily detected. In such cases, the *ECDF* of $\tilde{r}_k$ informs us that the uniform shape has been artificially caused by random assignments of the rank.

4.2.2 Experiment 2: Single Calibration Relation

Experiment 2 is aimed at showing how calibration relations between model parameters and a meteorological observable at coarse scale can take into account model sampling variability leading to consistent ensemble members.

In this case, we generate a set of 'observed events' with different intermittency properties, adopting the calibration relation between STRAIN parameter c and precipitation rate at the coarse scale R provided by equation (2.5). For purpose of this study, we select a calibration relation found to be

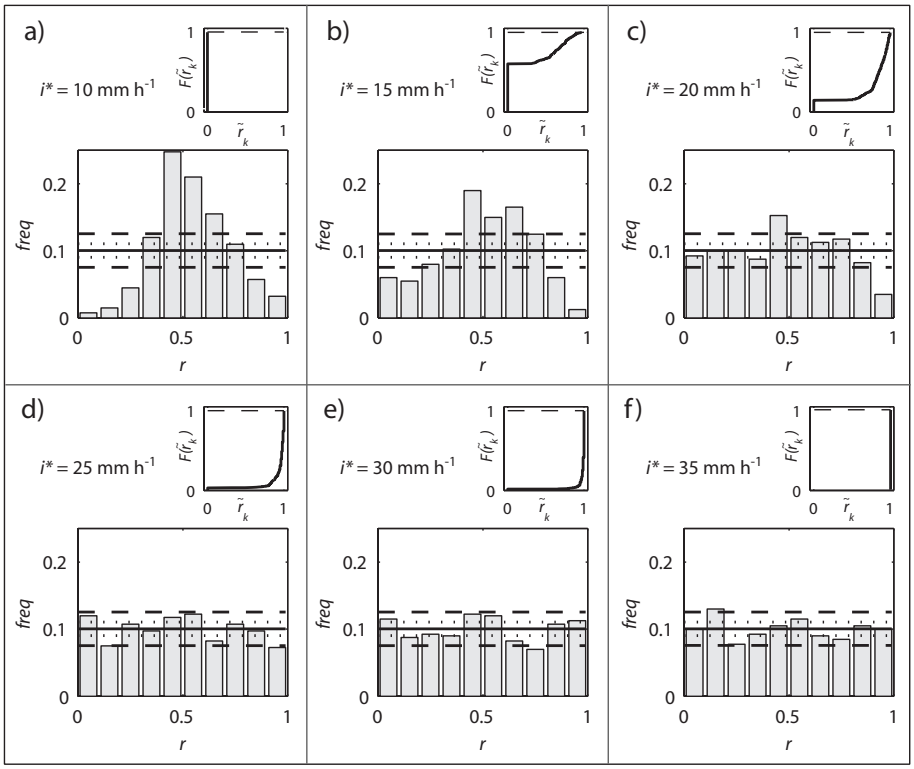

Figure 4.4: Experiment 1, 'event-based' calibration mode. Verification Rank Histograms of exceedance probabilities are built for $i^* = 10$, 15, 20, 25, 30 and 35 mm h^{-1} and plotted using $N_{bins} = 10$ bins to group the 400 ranks. The horizontal lines represent the 5%, 25%, 50%, 75% and 95% confidence intervals of a uniform distribution. In each panel the *ECDF* of $\tilde{r}_k$ is associated to each rank histogram. The histograms corresponding to the lower thresholds (panels a, b and c), where the *ECDF* of $\tilde{r}_k$ reveals that most part of the ranks has been unequivocally determined, shows an effect of overdispersion. When i^* increases (panels d and e), the number and magnitude of non-zero $\tilde{r}_k$ values increase and the histograms become artificially more uniform. As extreme case, when the precipitation threshold is higher than observed and ensemble precipitation values (panel f), all the ranks are always randomly assigned in the interval $(0, 1)$ and the histogram is drawn from a uniform distribution.

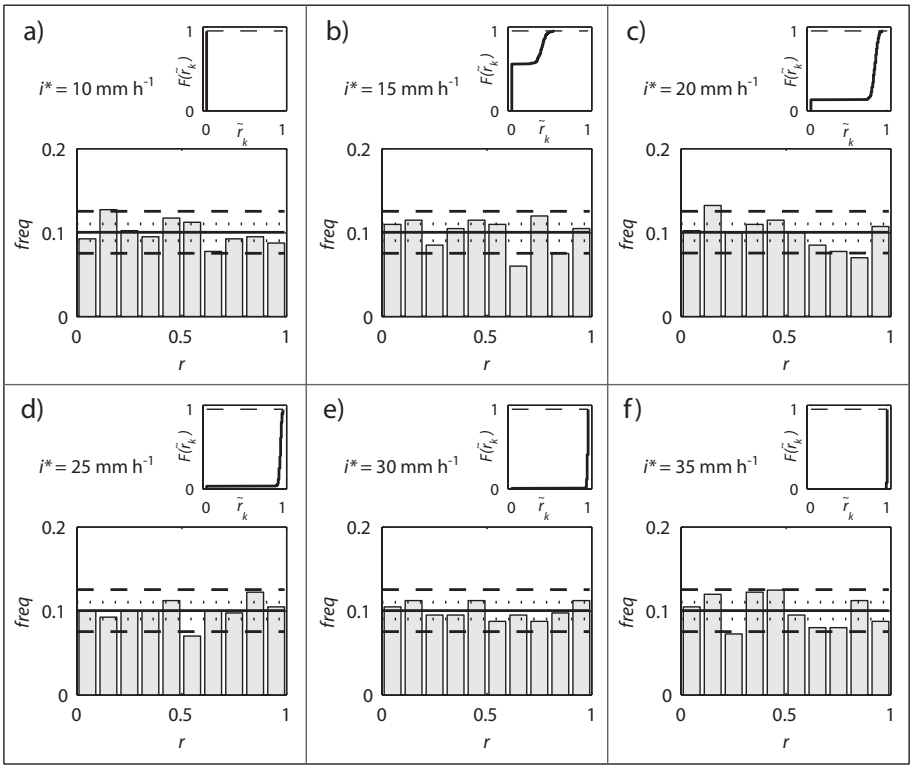

Figure 4.5: Experiment 1, 'mean-based' calibration mode. All the histograms result uniform, whatever the value of the precipitation threshold i^*. As a result, ensemble consistency is achieved.

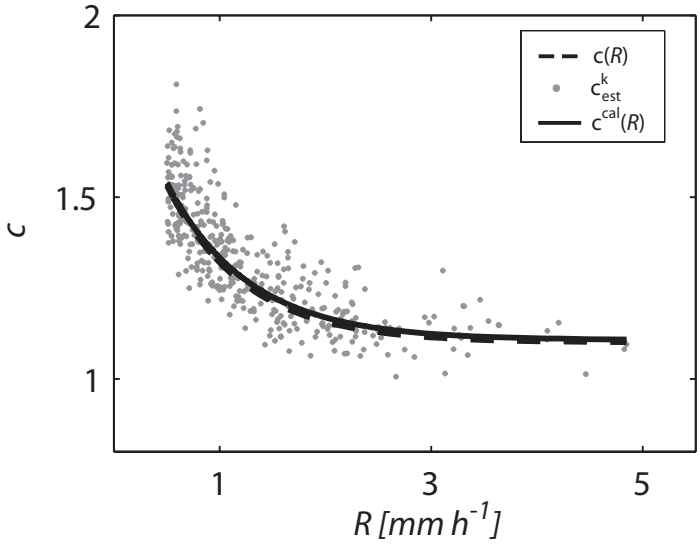

Figure 4.6: Relation between STRAIN parameter c and coarse rain rate R. The dashed black line represents the calibration relation $c = c(R)$, dots represent the parameters c_k^{est} estimated on each event and the solid black line is the calibration relation $c = c^{cal}(R)$.

valid by Deidda et al. (2004), where $c_\infty = 1.1$, $a = 0.85$ and $\gamma = 1.35$. Parameter β was found to be fairly constant at e^{-1}. This calibration relation $c = c(R)$ is plotted in Fig. 4.6 through a dashed black line. In this experiment, the set of 'observed events' is generated using the calibration relation, in the following way. First, 400 precipitation rates at the coarse scale R_k ($k = 1, \ldots, 400$) are randomly drawn, in an interval included between 0.25 and 5 mm h^{-1}, from an exponential distribution fitted to the R values analyzed by Deidda et al. (2004) to mimic the occurrence of observed large-scale events. Then, the set of parameters $c_k = c(R_k)$ is calculated using equation (2.5) and used to generate 400 high resolution precipitation events through STRAIN model with 5 downscaling levels.

As Experiment 1, given the 'observed events', we attempt to calibrate STRAIN parameters that will be used in the hindcasting phase, assuming no knowledge about their origin. For this purpose, parameter c is estimated on each observed event k. In Fig. 4.6, the 400 c_k^{est} estimates are plotted versus the R_k of the correspondent k-th event through dots. Because of intrinsic

model sampling variability, each c_k^{est} estimated on the event k is different from the parameter $c_k = c(R_k)$ used for the generation of the event k itself. In particular, c_k^{est} estimates result in a Gaussian-like distribution around c_k. The set of 400 c_k^{est} is then used to fit equation (2.5), obtaining the parameters $c_\infty = 1.1$, $a = 0.82$ and $\gamma = 1.29$. This new calibration relation, $c = c^{cal}(R)$, plotted in Fig. 4.6 with a black solid line, is very close to the calibration relationship $c = c(R)$ used for the generation of the 'observed events'.

In Experiment 2, hindcasts of the 400 observations are carried out according to the 'event-based' and 'mean-based' calibration modes. In addition, we test the 'functional-based' calibration mode, where the ensemble members hindcasting each event k are generated using the parameter value $c_k^{cal} = c^{cal}(R_k)$. In all modes, 100 ensemble members are generated to hindcast each observation and VRHs of the exceedance probabilities are then constructed for precipitation thresholds $i^* = 50$, 75, 100, 150, 250 and 500 mm h^{-1} (selected to obtain all the possible *ECDF* of $\tilde{r}_k$ behaviors). Results are shown in Fig. 4.7, 4.8 and 4.9 for the 'event-based', 'mean-based' and 'functional-based' calibration modes, respectively.

Results for the 'event-based' calibration mode (Fig. 4.7) display similar behavior as detected in Experiment 1 (overdispersion), due to the same 'centering' effect. In the 'mean-based' calibration mode (Fig. 4.8), the histograms for the lower thresholds (i^*= 50, 75 and 100 mm h^{-1}) are more populated in the lowest and in the highest ranks (U-shaped), revealing an effect of forecast underdispersion. The histograms shape for the higher thresholds ($i^* = 150$, 250 and 500 mm h^{-1}) is affected by randomly assigned ranks. Finally, the 'functional-based' calibration mode results (Fig. 4.9) display uniform histogram shape for every precipitation thresholds and the consistency condition is respected.

The effect of underdispersion for the 'mean-based' calibration mode implies that ensemble members are more frequently close to each other and distant from the observation. Indeed, the sampling variability of STRAIN model with a single parameter c_{mean} is not able to fully explain the variability of the 400 'observed events'. As a result, the *ESF* of several 'observed events' will result far away from the corresponding set of ensemble hindcasting *ESF*s.

In contrast, when the parameter c_k^{cal} is used to generate the ensemble members ('functional-based' calibration mode), the consistency condition is achieved because observations and ensemble belong to the same distribution. In fact, the new calibration relation $c = c^{cal}(R)$ is very close to $c = c(R)$ used to generate the 'observed events', implying $c_k^{cal} \approx c_k$, for each k-th event.

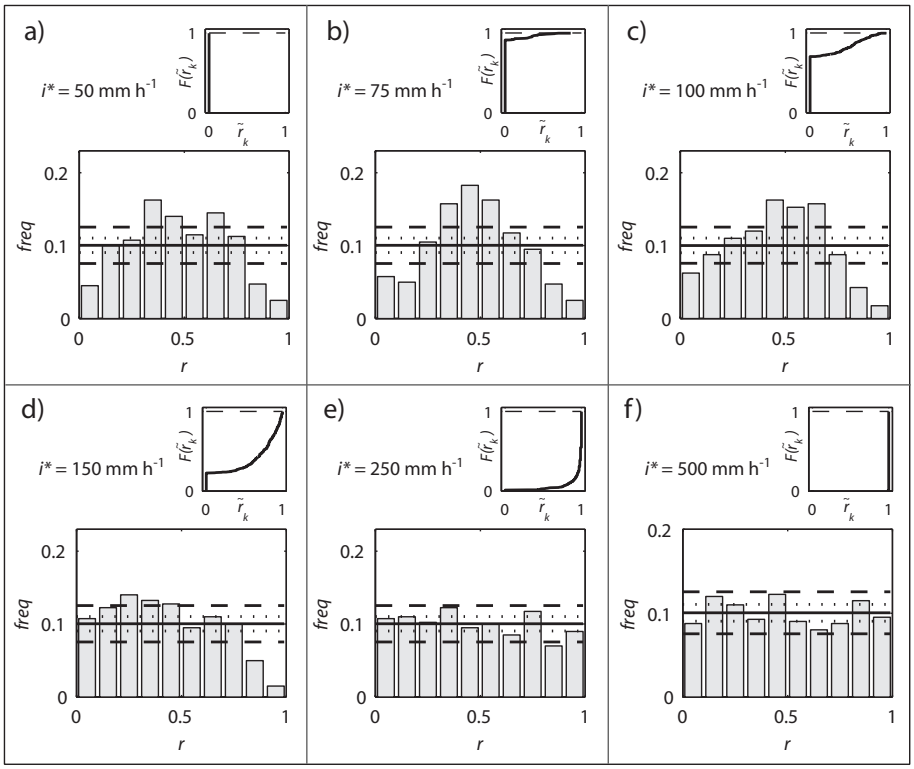

Figure 4.7: Experiment 2, 'event-based' calibration mode. In the histograms correspondent to the lower precipitation thresholds (panels a, b and c), an effect of overdispersion is detected, while, for the higher thresholds (panels d, e and f), the *ECDFs* of $\tilde{r}_k$ reveal that most of the ranks have been randomly determined and thus the histograms are artificially uniform.

In summary, this synthetic experiment was set to mimic the observed link between downscaling model parameters and coarse scale rainfall rates. In this case, the 'functional-based' calibration mode was the only one able to capture downscaling model sampling variability and to generate consistent ensemble members.

4.2.3 Experiment 3: Multiple Calibration Relations

Precipitation events can have different physical origins (e.g. convective or stratiform) and, in principle, one could expect that different calibration relations should be determined. Therefore, Experiment 3 was designed to

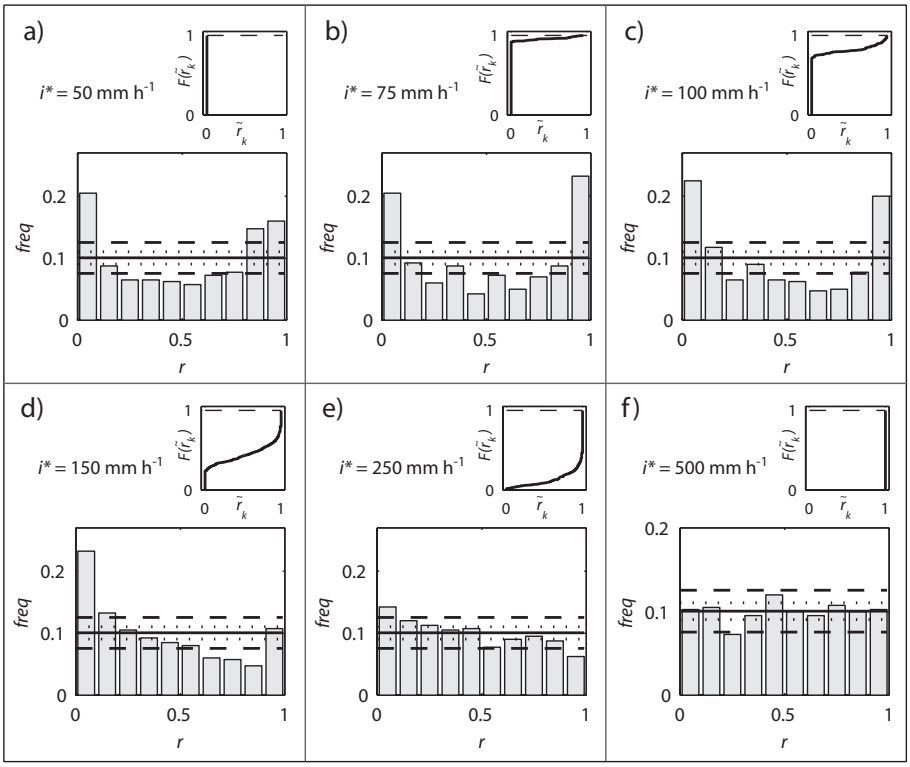

Figure 4.8: Experiment 2, 'mean-based' calibration mode. The histograms corresponding to the lower thresholds (panels a, b and c), where the *ECDF* of $\tilde{r}_k$ reveals that most of the ranks has been unequivocally determined, show an effect of underdispersion. As i^* increases the shape of the histograms are artificially more uniform (panels d, e and f).

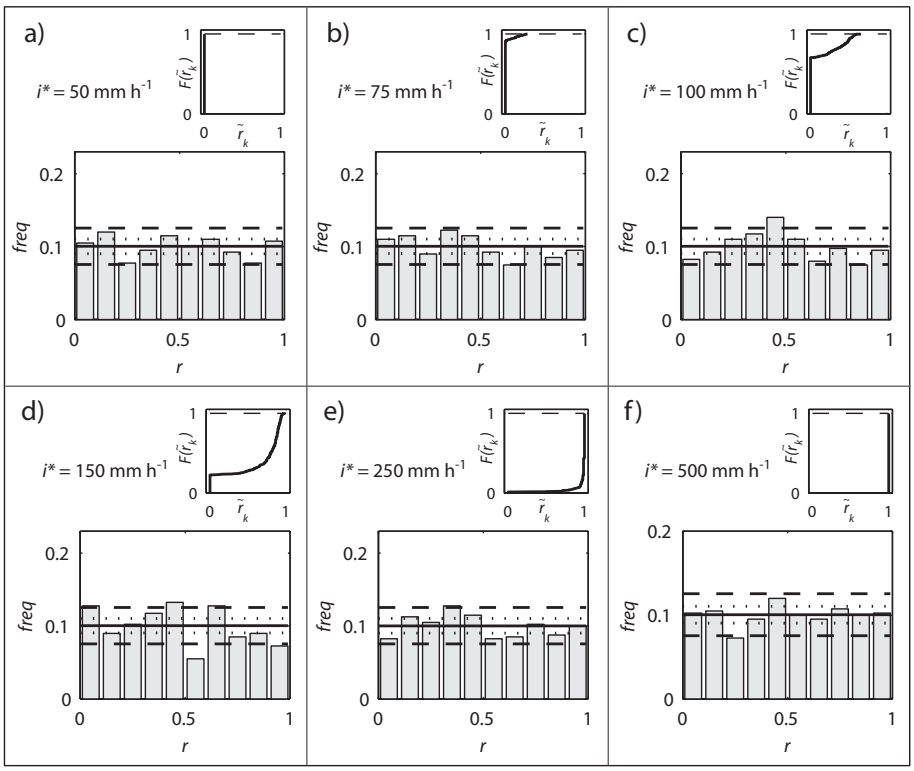

Figure 4.9: Experiment 2, 'functional-based' calibration mode. The histograms are uniform, for all precipitation thresholds i^*. As a result, ensemble consistency is achieved.

analyze the working of the VRH in this scenario. The set of 'observed events' is generated starting from two different calibration relations $c = c^1(R)$ and $c = c^2(R)$, both described by equation (2.5) with parameters $c_\infty = 1.13$, $a = 0.88$, $\gamma = 1.49$, and $c_\infty = 0.90$, $a = 0.55$, $\gamma = 1.00$ (Fig. 4.10). We selected these calibration relations to mimic different precipitation mechanisms. However, this selection is used only to explain the potential impact of multiple relationships and the values selected for c_∞, a, and γ are not representative of real climatologies as studies investigating this phenomenon have not been yet conducted. A total of 400 'observed events' is generated using STRAIN model with 5 downscaling levels starting from 400 precipitation rates at the coarse scale R_k drawn by the same exponential distribution simulating the large-scale events occurrence adopted in Experiment 2. The set of 400 parameters c is here determined by introducing 200 R_k values in calibration relationship $c = c^1(R)$ and the other 200 R_k in calibration relationship $c = c^2(R)$.

Assuming a no a-priori knowledge of the method used to generate the set of 'observed events', we then use these events to calibrate STRAIN parameters for the subsequent hindcasting phase, adopting the 'event-based', 'mean-based' and 'functional-based' calibration modes.

In the 'event-based' calibration mode, we adopt the parameters c_k^{est} estimated in each event k, to generate the set of ensemble hindcasts. The values c_k^{est} are plotted versus R_k in Fig. 4.10 using circles and asterisks for events generated by relations $c = c^1(R)$ and $c = c^2(R)$. In the 'mean-based' calibration mode, each 'observed event' is hindcasted using the mean value c_{mean} of the 400 c_k^{est} estimates. In the 'functional-based' calibration mode, we interpret the behavior of the estimates c_k^{est} with respect to R by estimating only a single calibration relationship $c = c^{cal}(R)$, which ignores differences in precipitation type. This relation is then used to determine the set of 400 parameters $c_k^{cal} = c^{cal}(R_k)$ for the generation of the ensemble members (solid line in Fig. 4.10).

For each calibration mode, 100 ensemble members are produced to hindcast each observation and VRHs of the exceedance probabilities are constructed for the same precipitation thresholds tested in Experiment 2. Results of the verification procedure for the 'event-based' and 'mean-based' calibration modes (not shown) are very similar to those obtained in the second experiment.

Results of the 'functional-based' are shown in Fig. 4.11. Histograms for low i^*, whose shape is not affected by random assignment of the ranks, are

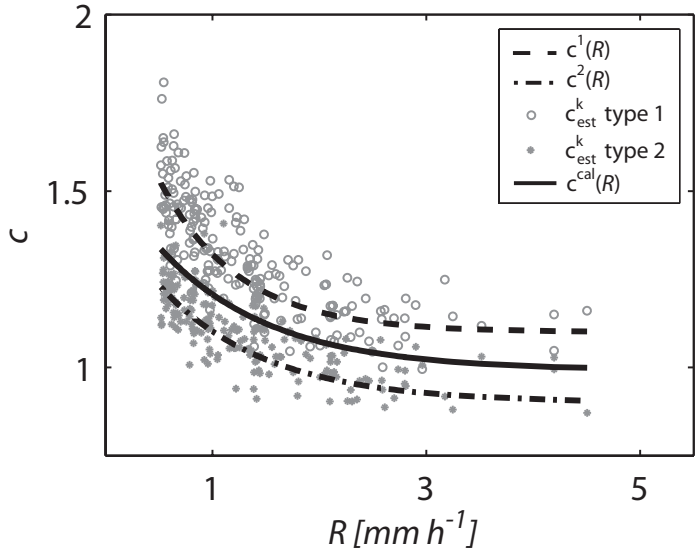

Figure 4.10: Variation in calibration relations with precipitation type. Dashed and dashed-dotted lines represent $c = c^1(R)$ and $c = c^2(R)$ used to generate 400 'observed events' (200 events for each relation). Parameters c_k^{est} are plotted with circles and asterisks if the correspondent 'observed events' come from $c = c^1(R)$ (type 1 events) and $c = c^2(R)$ (type 2 events). The solid black line represents the calibration relation $c = c^{cal}(R)$, which ignores differences in precipitation type.

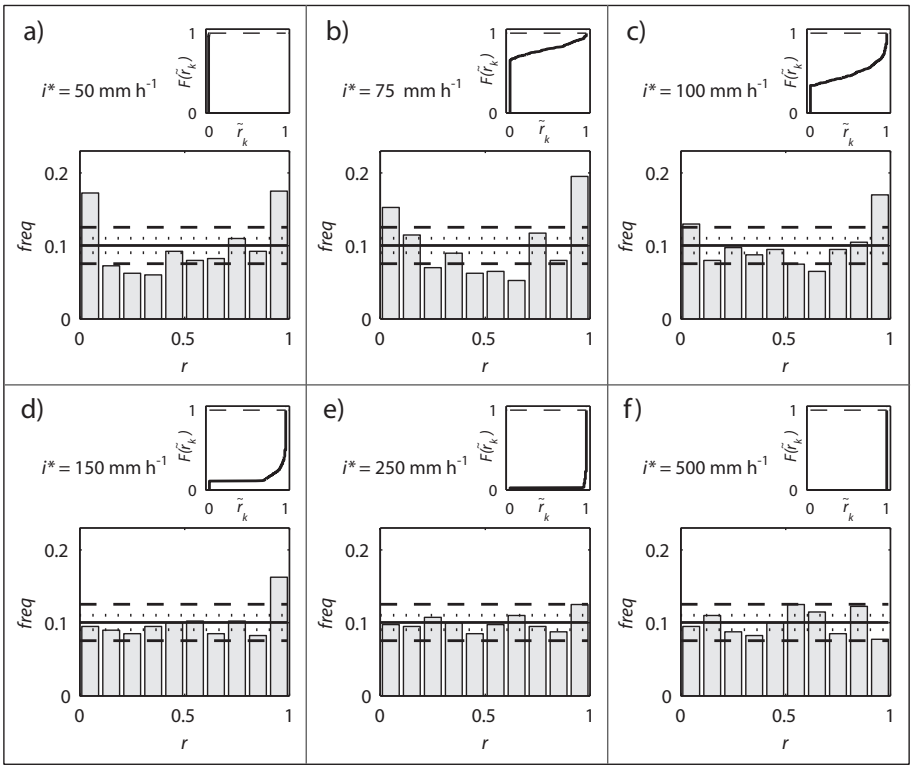

Figure 4.11: Experiment 3, 'functional-based' calibration mode. Histograms corresponding to the lower thresholds (panels a, b and c), where the *ECDF* of $\tilde{r}_k$ reveals that most part of the ranks has been unequivocally determined, show an effect of underdispersion. As i^* increases the shape of the histograms are artificially more uniform (panels d, e and f).

more populated in the lowest and the highest ranks (U-shaped), revealing an effect of underdispersion. In particular, the 'observed events' from calibration relations $c = c^1(R)$ and $c = c^2(R)$ cause a population mainly of high or low ranks, since the observed *ESF* is placed far away from the ensemble *ESFs*. Therefore, a downscaling model adopting the 'functional-based' calibration mode based on a single relation, is not able to fully interpret the proper variability of events drawn by different calibration relations.

In the case that different precipitation mechanisms lead to different calibration relations, ensemble consistency should be reached by classifying the available 'observed events' according to their physical origin (e.g.

stratiform or convective or on the basis of the synoptic pattern), and then by estimating different calibration relations able to simulate the variability of each family of events. Each 'observed event' should be then forecasted using the storm-dependent calibration relation.

4.3 Discussion and Conclusions

A verification method for ensemble precipitation fields generated by downscaling models has been presented. The method is based on the VRH and tests consistency of the exceedance probability of a fixed precipitation threshold i^*. Once i^* has been fixed, the exceedance probabilities of ensemble members and observation are calculated for each event, the rank of the observation is tabulated and the VRH is built. For high values of i^*, several ranks may be randomly assigned affecting the shape of the VRH. Therefore, a graphical method accounting for random assignments of the rank has been also developed.

The verification procedure has been applied on three series of numerical experiments, using the STRAIN downscaling model, with the aims of (i) testing how the verification procedure works and (ii) evaluating downscaling model deficiencies when different calibration modes are adopted to estimate model parameters. The analysis of the results of the three experiments permit us to draw the following conclusions:

1. If we consider a hindcast framework and we generate the ensemble members adopting the parameter c_k^{est} estimated on the event k to be hindcasted (at a fist sight, the best possible solution), the model returns overdispersed forecasts. This is due to the fact that model sampling variability is not accounted for in parameter calibration producing a centering of ensemble members around the observation.

2. The intrinsic variability of downscaling model when a single parameter value c_{mean} (averaged of the estimates c_k^{est}) is used, may not be able to capture the variability of observed events and underdispersed forecasts are produced.

3. The use of a calibration relation linking model parameter with meteorological observable at coarse scale may allow model sampling variability to be taken into account leading to consistent members.

4. When precipitation events depend on different physical origins (i.e. convective or stratiform) or are generated by different synoptic conditions, a single calibration relation may not be able to explain the variability of the entire set of events and thus may return underdispersed forecasts. In order to reach consistency, in such a situation, it would be necessary first to classify the events according to their physical origin and then to estimate storm-dependent calibration relations.

Chapter 5

Uncertainty Propagation Into Hydrological Response

In chapter 4 we showed how downscaling models can be used to generate consistent ensemble precipitation forecasts through calibration relations between model parameters and coarse meteorological observable. We also showed that if other calibration modes are adopted to select model parameters, overdispersed or underdispersed forecasts are instead produced. In this chapter the STRAIN downscaling model is coupled with the tRIBS distributed hydrological model to analyze how uncertainty and possible deficiencies of downscaled precipitation fields affects hydrological response and performances of hydrometeorological forecasting systems.

Given the complexity and high non-linearity of the processes involved, the study has been focused only on the uncertainty caused by precipitation predictions, while all the other sources of uncertainty of the forecast system, such as coarse precipitation uncertainty, basin initial state and hydrological model parameterization and structure, have not been taken into account.

Hindcast experiments have been carried out in controlled conditions, with the same philosophy adopted in chapter 4, applying the hydrometeorological system described in chapter 2 over the Baron Fork basin (Oklahoma, USA), a sub-basin of the Arkansas Red River basin (Fig. 5.1). Sets of spatiotemporal precipitation fields covering several summer periods have been first generated through the STRAIN downscaling model with selected values of parameters c and β (i.e. known statistical and intermittency properties). These fields have been assumed as 'observed' precipitation input and used to force the tRIBS model. The resulting hydrographs have been considered in

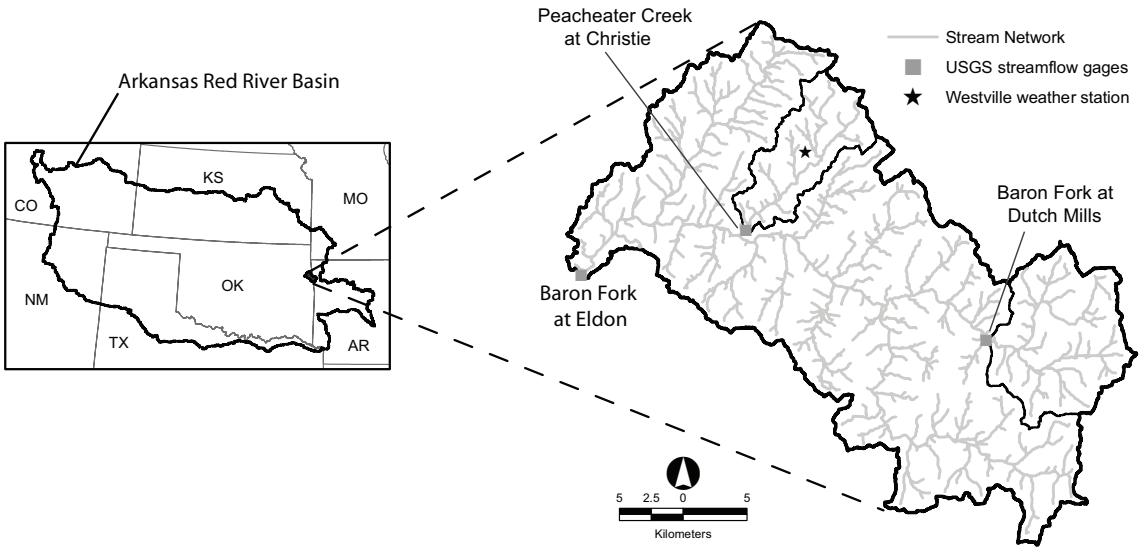

Figure 5.1: Location of the study basin Baron Fork. Left panel: Baron Fork basin position in relation to the Arkansas Red River basin. Right panel: Baron Fork basin boundaries with the two nested sub-basins Peacheater Creek at Christie and Baron Fork at Dutch Mills boundaries, including stream network, U. S. Geological Survey streamflow gages and Westville weather station.

turn as ground-truth and utilized for the hydrological verification phase. Subsequently, once the 'observed' precipitation and streamflow database have been built, performances of the forecasting system have been verified in several hydrometeorological events by forcing the tRIBS model with ensemble hindcasts of consistent, overdispersed and underdispersed precipitation fields. To test if and how the hydrological response is affected, an event-based verification procedure for ensemble hydrographs based on the Verification Rank Histogram (VRH) is proposed and applied.

The chapter is organized as follows. Section 5.1 illustrates the study basin and how the 'observed' precipitation and streamflow database have been generated. In section 5.2, the verification method for ensemble streamflow is illustrated. Hindcast experiments are described in section 5.3, while discussion of results is presented in section 5.4.

5.1 Study Area and 'Observed' Database Generation

5.1.1 Study Basin and Spatial Attributes Representation

The target basin of the study is Baron Fork at Eldon (Oklahoma, USA). Basic topographic and hydrologic characteristics for this catchment and two nested sub-basins Peacheater Creek at Christie and Baron Fork at Dutch Mills, monitored by United States Geological Survey (USGS), are summarized in Table 5.1, based on Slack et al. (2001), USGS streamflow, and USGS DEM data.

Basin/USGS gauge #	A [km^2]	H/C_{vH} [m]/[-]	L [km]	S_L [m km^{-1}]	S_A [m km^{-1}]	P [mm]	Q [mm cm^{-1}]
Baron Fork at Eldon (USGS 0719700)	808.39	346/0.462	65.2	4.35	15.3	1130	371/9.43
Sub-basin Peacheater Creek at Christie (USGS 07196973)	65.06	328/0.352	18.1	6.74	11.7		368/0.75
Sub-basin Baron Fork at Dutch Mills (USGS 07196900)	106.91	408/0.559	17.8	8.26	20.8		388/1.29

Table 5.1: Basic topographic and hydrologic characteristics of the test basin Baron Fork and of two nested sub-basins monitored by USGS. Symbols: A, basin drainage area; H, basin mean elevation ([m] above NGVD29); C_{vH}, coefficient of variation of elevation as a ratio of standard deviation to the difference between the mean and minimum elevation of the basin; L, maximum distance of channel flow; S_L, average slope of the longest channel; S_A, average slope of channel drainage network; P, mean annual precipitation; Q, mean annual flow. From Ivanov et al. (2004*b*).

The terrain of Baron Fork catchment is characterized by gently rolling relief at the basin headwaters and quite rugged terrain in its lower areas. The watershed has significant vegetation cover: about 52% of the area is occupied by deciduous and evergreen forests, 46% is occupied by croplands and orchards. The surface soil texture is primarily silt loam (94%) and fine sandy loam (6%).

Land-surface characteristics (topography, landuse/ vegetation and soils) describing the interior watershed structure, has been represented in tRIBS

model computational domain through an irregular spatial discretization based on TINs (see also section 2.3.1). Topography for the test basin was derived from USGS 30 m DEM (Fig. 5.2a) using the hydrographic TIN procedure described in Vivoni et al. (2004). The approach provides high resolution in areas with significant elevation gradient. River floodplains, resolved at a high detail, were also integrated into the TIN terrain models. Through the TIN implementation, the quantity of computational elements was significantly reduced. Compared to the 30 m resolution DEM, the amount of computational elements was 7.22% (64,836 nodes) of the original number of grid cells and the equivalent grid cell sizes, i.e. the pixel size in the grids with the same number of computational elements, is correspondingly 112 m (Fig 5.2b). A comparative analysis of the TIN accuracy relative to the highest DEM available was conducted by Vivoni et al. (2004).

Information about landuse, vegetation cover, and soils is required for proper parameterization of energy and water fluxes at the land-surface. Spatial heterogeneity of these properties in tRIBS is accounted for by assigning the relevant landuse/soil texture type to the nodes of the computational domain. The USGS Land Use and Land Cover (LULC) data were used in the current study to represent landuse and vegetation cover (Fig. 5.3). Soil Conservation Service (SCS) State Soil Geographic Database (STATSGO) soils provides soil characteristic for all U.S. territory. Nevertheless, since the STATSGO data showed essentially homogeneous soil texture types in the study basin, spatial non-uniformity of soil hydraulic properties that can affect the infiltration regime was obtained following an alternative approach. Vegetation and landuse classes were combined to define grassy, forested, and urbanized sites that were used as a surrogate representation of soils spatial variability (Ivanov et al. 2004*b*).

5.1.2 Generation of 'Observed' Precipitation Database

As first step of the study, a database of 'observed' spatiotemporal precipitation series with know statistical properties has been synthetically generated and used as meteorological forcing for the hydrological model. The STRAIN downscaling model (see section 2.2) has been utilized for this purpose, starting from precipitation values at the coarse scale obtained by radar estimates provided by NWS Next-Generation Weather Radar (NEXRAD) system of the Arkansas Red River Forecasting Center (ABRFC) during 9 years (1997-2005). Generation of 'observed' data is based on some

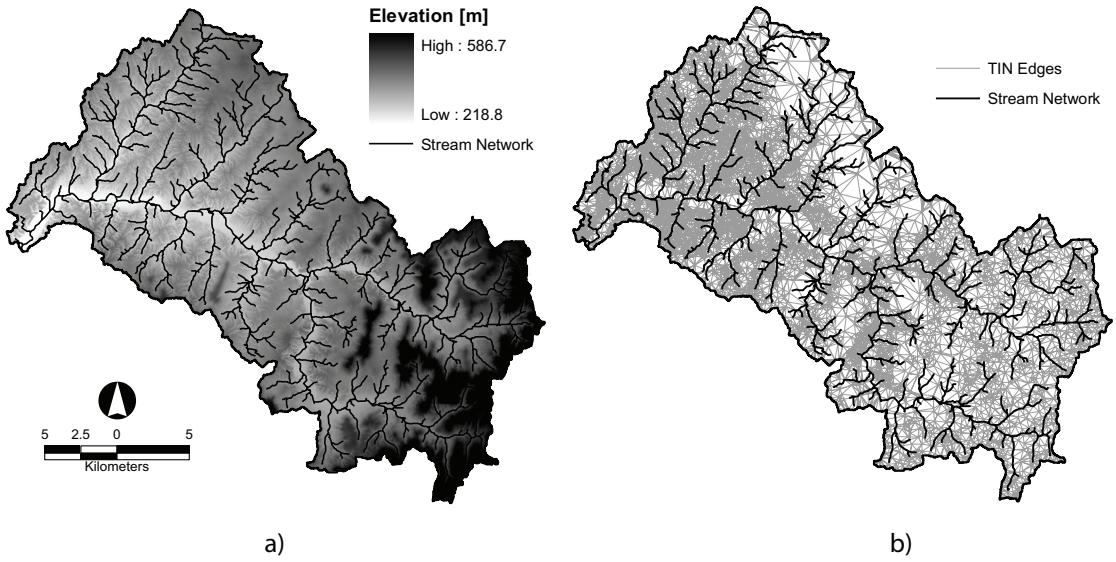

Figure 5.2: Panel a: U.S. Geological Survey 30-m digital elevation model (DEM) for Baron Fork basin. Panel b: Terrain representation using a TIN derived from the 30-m DEM shown in panel a, where the higher triangle density corresponds to more rugged topography (Vivoni et al. 2004).

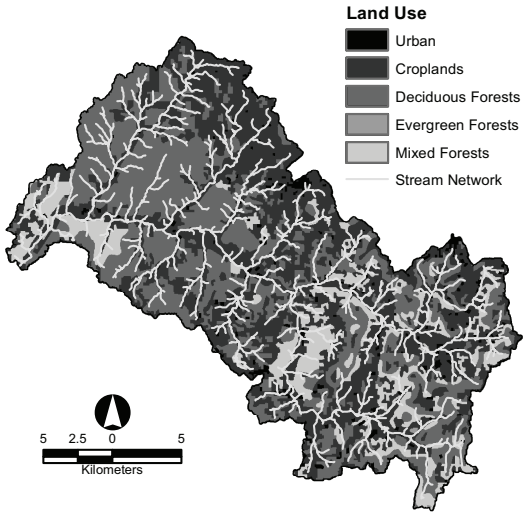

Figure 5.3: Spatial distribution of land cover within the basin (deciduous, evergreen and mixed forest, croplands, and urban).

preliminary hypotheses. First of all, the existence of scale invariance laws has been assumed between the coarse scales $L = 256$ km, $T = 16$ h and the fine scales $\lambda = 4$ km, $\tau = 15'$, implying that temporal and spatial scales have been homogeneized by a velocity parameter $U = 16$ km h^{-1}. Then, it has been chosen to focus the analysis only on summer period (June, July and August) in order to simulate events that have been likely caused by similar meteorological signatures. Thus, it has been assumed that a single calibration relation $c = c(R)$ exists between coarse precipitation rate R and STRAIN parameter c, provided by equation 2.5 with parameters $c_\infty = 0.675$, $a = 0.907$ and $\gamma = 0.764$. These values were estimated in a previous application on radar data by Deidda (2000) for events in the same hypothesized range of scales. Parameter β was kept constant to e^{-1}.

We highlight that the assumptions mentioned above have not been verified through real data collected in this location (also because no radar data at 15' temporal resolution were available), but that they have been made only with the aim of creating a precipitation database with known intermittency and statistical properties, allowing uncertainty evaluation within the forecast system to be better controlled. Nevertheless, we remember that Deidda (2000) and Deidda et al. (2004) found the presence of scale invariance laws

in the same and in a very similar range of scales on real radar data of different locations.

Fig. 5.4 illustrates the database generation steps:

1. An area of $L = 256$ km $\times$ $L = 256$ km centered on Baron Fork basin (in Universal Transverse Mercator, UTM, coordinate system) has been first identified (panel a). Subsequently each summer of years $l = 1997, \ldots, 2005$ has been divided into consecutive T =16-hour long events with a total of 138 events per summer starting from June, 1^{st} at 00:00 UTC and ending at September 1^{st} at 00:00 UTC. The observed precipitation rates $R_{i,l}$ $(i = 1, \ldots, 138)$ at the coarse scale $L \times L \times T$ have been then extracted for each event from NEXRAD database of ABRFC. Processes of weather radar data consisted of coordinate transformations from the Hydrologic Rainfall Analysis Project (HRAP) to the UTM coordinate system and selection of data corresponding to the geographic extent of the spatial coarse scale $L \times L$.

2. For each 16-hour long event i in year l, STRAIN parameter $c_{i,l} = c(R_{i,l})$ has been determined by means of the calibration relation 2.5 and utilized to generate one synthetic field at resolution 4 km $\times$ 4 km $\times$ 15' (panel b). No downscaling has been performed if $R_{i,l} = 0$ and zero precipitation at the finest resolution has been assumed throughout the 16 hours.

3. As a result, for each year l, the 'observed' database has been built by concatenating the downscaled fields coming from $R_{1,l}, R_{2,l}, \ldots R_{138,l}$. For example, panel c shows the spatial fields at resolution of 4 km of the first two and the last 15' time intervals for the 16-hour long event starting in June 21^{st} 2000 at 00:00 UTC (i.e. the rainfall fields downscaled from $R_{31,2000}$).

To further illustrate the generation procedure, Fig. 5.5 shows, in the top panel, the time series of the $R_{i,2000}$ coarse precipitation rates (time step $\Delta t = 16$ hours), and, in the bottom panel, the time series of the Mean Areal Precipitation (MAP) over Baron Fork basin for the correspondent downscaled fields ($\Delta t = 15'$). A zoom on one of the 16-hour long event is also shown in the right part of the figure for both the coarse and the downscaled rainfall.

As last point of precipitation database generation through STRAIN model, the 'observed' spatiotemporal fields have been converted as ASCII grid (ESRI, 1992) to be inputted to tRIBS model.

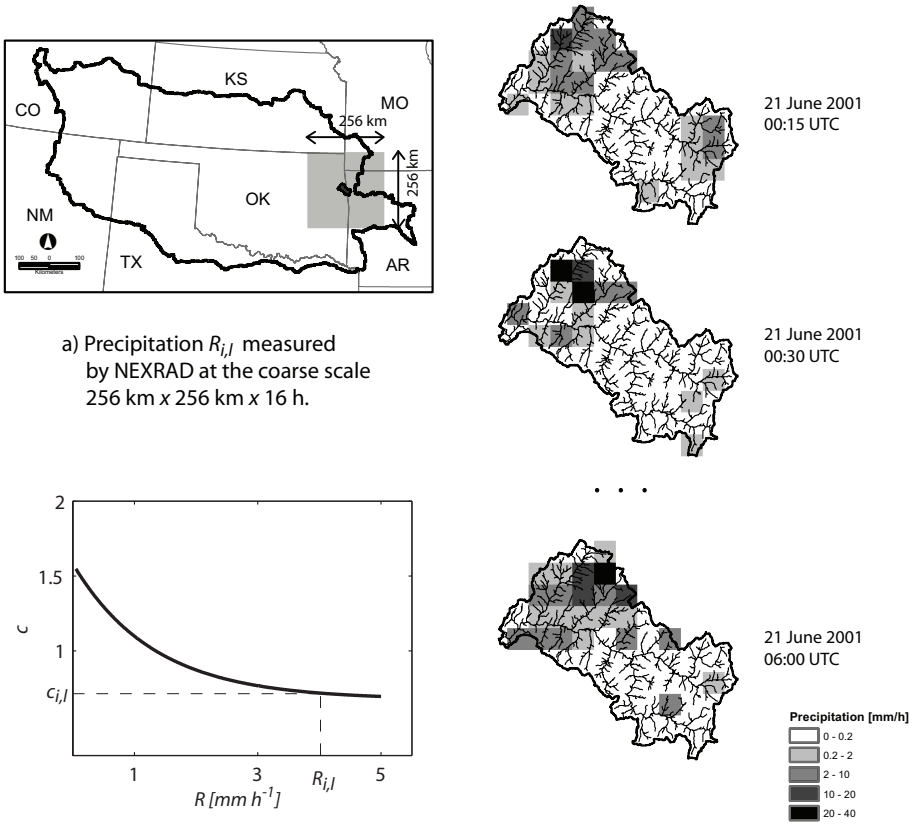

Figure 5.4: Generation of 'observed' precipitation database. Panel a: NEXRAD radar estimates are used to obtain precipitation values $R_{i,l}$ averaged in the shaded area of 256 km $\times$ 256 km over Baron Fork basin and along each $i = 1, \ldots, 138$ consecutive 16-hour long interval covering summers of years $l = 1997, \ldots, 2005$. Panel b: parameter $c_{i,l} = c(R_{i,l})$ is obtained using calibration relation 2.5 with parameters $c_\infty = 0.675$, $a = 0.907$ and $\gamma = 0.764$. Panel c: example of precipitation spatial fields at high resolution (4 km, 15') obtained by downscaling $R_{31,2000}$.

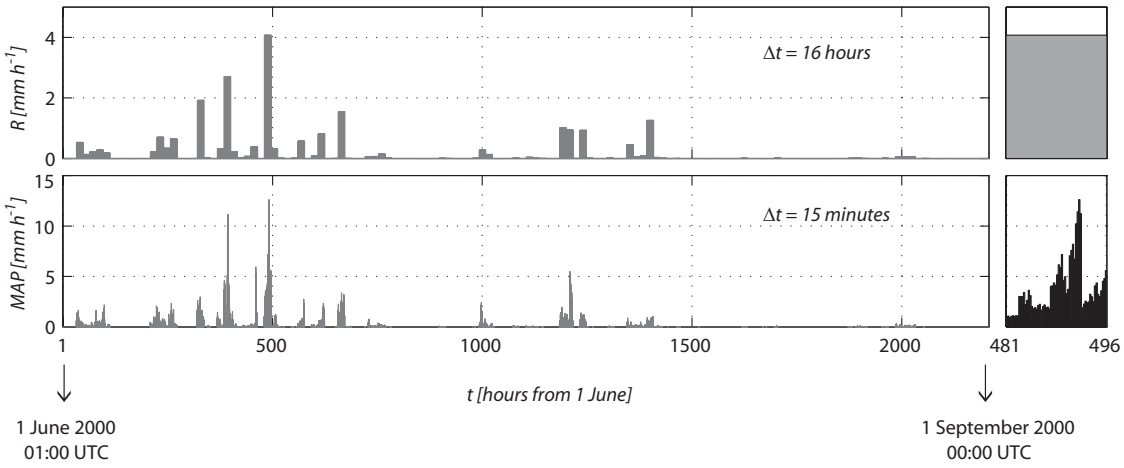

Figure 5.5: Example of 'observed' precipitation database for summer 2000. Top panel: time series of $R_{i,2000}$ ($i = 1, \ldots, 138$). Bottom panel: time series of the Mean Areal Precipitation (MAP) over Baron Fork basin calculated from each precipitation fields at high resolution downscaled from the correspondent coarse value $R_{i,2000}$.

5.1.3 Generation of 'Observed' Streamflow Database

The database of 'observed' precipitation grids at resolution 4 km, 15' of summers 1997-2005 have been used to force the tRIBS distributed hydrological model, which returned streamflow values assumed as ground truth and then used for verification in the subsequent hindcast phase.

Since the tRIBS model has the capability of reproducing multiple flood forecasts across a range of nested basin scales, hydrographs have been simulated at the outlet and at 14 nested sub-basins ranging in area from 0.78 to 808 km^2. Characteristics of nested sub-basins and outlet are summarized in Table 5.2 while their location is shown in Fig. 5.6. These interior locations have been selected basing on a study by Vivoni et al. (2006), who evaluated the influence of catchment scale and forecast lead time on the predictability of flood events through the combined use of radar nowcasting and the tRIBS distributed hydrological model. One of the goals of this study is therefore to further investigate and deepen results found by Vivoni et al. (2006).

Again, to preserve a better control on the simulations, simplifying assumptions have been made on the initial condition and the physical and

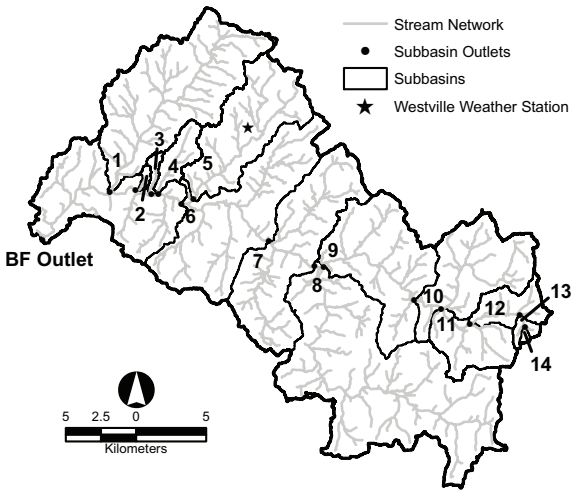

Figure 5.6: Boundaries and outlets of Baron Fork basin and the 14 nested sub-basins listed in Table 5.2.

hydrological parameterization of the catchment. In particular, hydrological model parameters have been considered always fixed for all the summers and the same initial ground water table position has been assumed at the beginning of each summer. Model parameters (listed in Table 2.1) have been selected according to baseline simulations for the Baron Fork basin over a 7-yr period (1993-2000) reported by Ivanov et al. (2004*b*). In addition, in order to obtain reliable parameter values for the specific period of the year here analyzed, a manual calibration experiment on summer 2000 has been carried out using multiple-gauge observations at the outlet Eldon and at two nested locations Peacheater Creek and Dutch Mills. Following the calibration strategy suggested by Ivanov et al. (2004*a*), a nested calibration experiment has been carried out forcing the model with NEXRAD precipitation data (4 km, 1h) and tuning model parameters of each sub-basin (e.g. soil, aquifer, and channel properties) prior to calibrating the overall watershed response. Only minor modifications have been made to the baseline calibrations. Results of the calibration experiment, shown in Fig. 5.7, reveal good model performances for the three stations. This provides confidence in the distributed hydrological model as a numerical laboratory and its capability for streamflow prediction in gauged basins. Performance at the gauged sites is also sufficiently accurate to test the model capabilities in forecast mode at ungauged sites in the watershed.

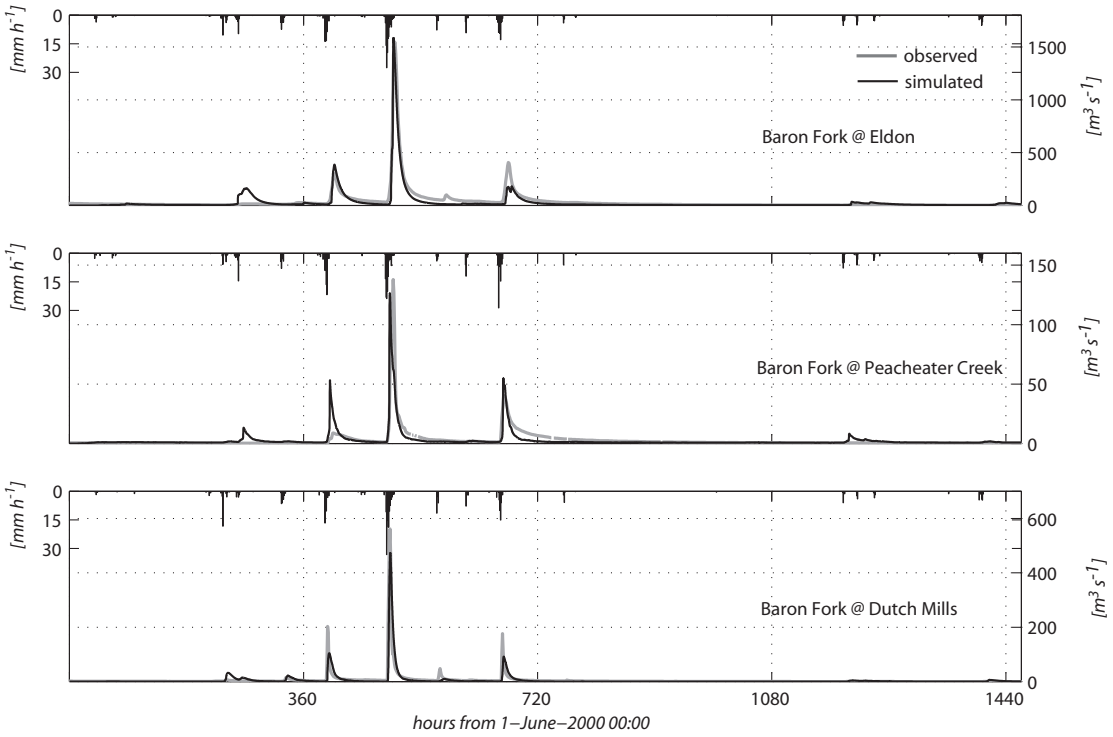

Figure 5.7: An excerpt from the run of summer 2000 illustrating simulation skills of the tRIBS model at the outlet Eldon and nested locations Peacheater Creek and Dutch Mills after parameter calibration.

Basin	A [km^2]	L [km]	S [m km^{-1}]	D_d [km^{-1}]	T_c [h]
1	108.23	25.73	6.06	0.9895	5.78
2	1.41	2.59	34.01	0.8264	0.51
3	2.67	4.52	21.44	0.7701	0.93
4	12.14	8.06	14.94	0.8059	1.67
5	65.06	19.90	9.26	0.8293	4.03
6	610.60	50.33	6.81	0.8355	9.26
7	450.26	40.01	8.11	0.8352	7.25
8	365.25	35.03	9.09	0.8209	6.27
9	182.91	29.78	9.49	0.8230	5.44
10	106.91	18.64	13.41	0.8370	3.32
11	49.07	12.72	19.10	0.8692	2.16
12	21.18	9.03	24.92	0.8700	1.50
13	4.29	3.53	51.27	0.7720	0.55
14	0.78	1.33	112.77	0.3033	0.19
15 (outlet)	808.39	67.26	5.47	0.8630	12.59

Table 5.2: Baron Fork sub-basins characteristics: area (A), maximum distance to the sub-basin outlet (L), relief ratio (S), drainage density (D_d); time of concentration (T_c) from Kirpich (1940): $T_c = 0.000325\, L^{0.77}\, S^{0.385}$, where units are L [m] and S [m m^{-1}].

With the aim of preserving the real climatology of area and season, hourly meteorological data: air and dew point temperature, cloudiness, wind speed and atmospheric pressure collected by Westville station (Fig. 5.6) belonging to the Oklahoma MESONET network, have been utilized in tRIBS to compute the surface energy fluxes and evaporation potential.

Example of the hydrographs obtained forcing tRIBS model with 'observed' precipitation are shown in Fig. 5.8 for summer 2000 at the outlet and Peacheater Creek and Dutch Mills sub-basins. Comparison among the calibrated hydrographs of Fig. 5.7 (the black ones) and the 'observed' hydrographs of Fig. 5.8 reveals good agreement between their shapes (apart for two missed peaks in Peacheater Creek). However, the 'observed' streamflow values of Fig. 5.8 systematically underestimate the correspondent calibrated hydrographs of Fig. 5.7. This is likely due to a deficiency of the adopted calibration relation in reproducing intermittency of

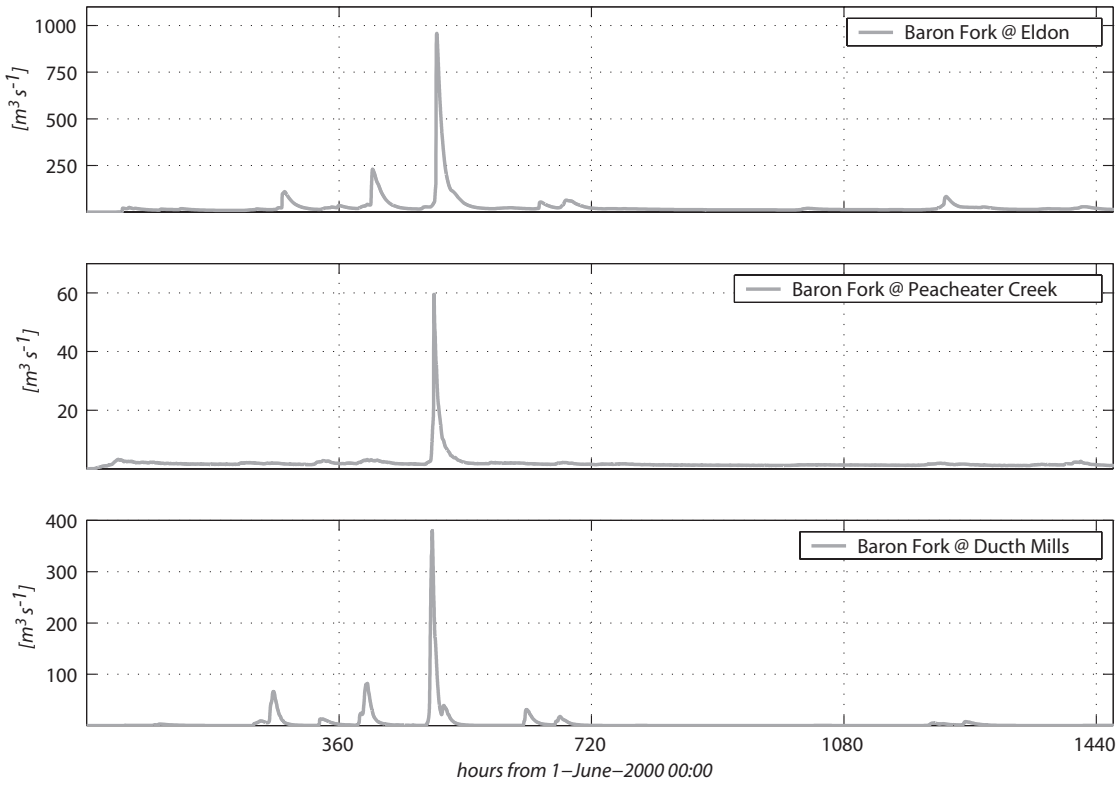

Figure 5.8: An excerpt from the run obtained forcing tRIBS with 'observed' precipitation data (4 km, 15') of summer 2000, for the outlet Eldon and nested locations Peacheater Creek and Dutch Mills.

high-resolution precipitation fields.

Nevertheless, we highlight once again that generation of 'observed' hydrographs was not made to simulate the USGS streamflow measures, but with the only purpose of having more control while testing uncertainty propagation in the hydrometeorological system.

5.2 A Method to Verify Consistency of Ensemble Streamflow

Before describing the numerical experiments carried out to evaluate uncertainty propagation in a hydrometeorological system, a methodology based on the VRH (illustrated in section 3.2) is proposed in this section to test consistency of ensemble streamflow with respect to the observed hydrograph. Although developed to test hydrometeorological forecasts, the verification procedure has more general validity and may be used whenever ensemble streamflows are produced. For example, it can be used to test consistency of ensemble hydrographs simulated with the purpose of evaluating uncertainty of hydrological model parameters or basin initial state.

The procedure is illustrated in Fig. 5.9. Suppose to know the time series of duration T_{hydro} for the observed and N_{ens} ensemble streamflow. The method requires first to fix a time interval T_{ver} to identify the events where ensemble and observed streamflow values will be postprocessed and the rank of the observation will be calculated. Panel a shows the observed time series of duration T_{hydro} where N_{ev} time intervals of duration T_{ver} have been selected.

Panel b shows observed and ensemble hydrographs within a generic event k belonging to these N_{ev} events of duration T_{ver}. From each time series, it is possible to extract a specific metric Q^m such as the maximum accumulated streamflow at different time durations. Specifically, N_{ens} values Q^m_j $(j = 1, \ldots, N_{ens})$ of the metric are extracted from the ensemble members and one, Q^m_{obs}, from the observation.

Subsequently, the vector $Q^m_1, \ldots, Q^m_{Nens}, Q^m_{obs}$ is sorted in increasing order obtaining the vector $Q^m_{(1)}, \ldots, Q^m_{(Nens+1)}$ and the Empirical Cumulative Distributive Function (ECDF) of this vector is built by means of Hazen plotting position formula (panel c). If p is the position of Q^m_{obs} in the sorted vector, the rank r_k of Q^m_{obs} is given by $\frac{p-0.5}{N_{ens}+1}$.

The described procedure can be repeated for each event and the VRH can be populated by N_{ev} ranks. Consistency implies a uniform histogram; otherwise, deficiencies may be detected.

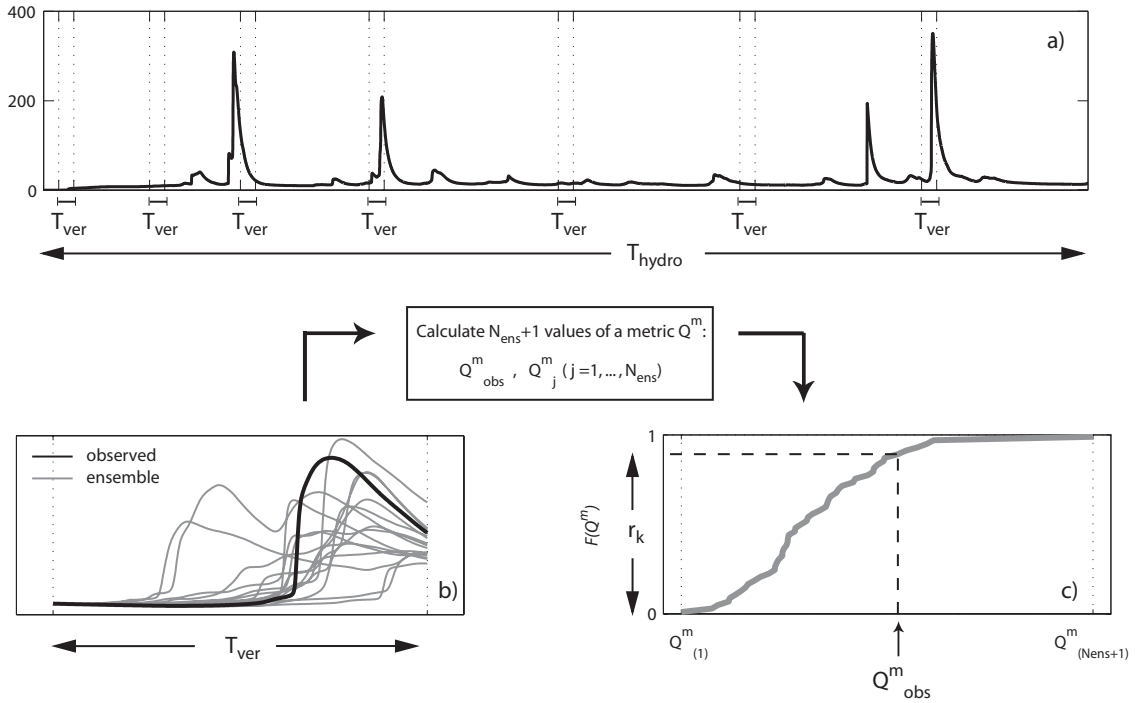

Figure 5.9: Ensemble streamflow verification method. Panel a shows a time series of observed streamflow with duration T_{hydro} and N_{ev} time intervals of length T_{ver} selected throughout the series. Panel b shows observed and N_{ens} ensemble hydrographs for a generic event k of length T_{ver}, from which N_{ens} metrics Q^m_j and the observed Q^m_{obs} are calculated. Panel c contains the Empirical Cumulative Density Function of the vector $Q^m_{(1)}, \ldots, Q^m_{(Nens+1)}$ returning the rank r_k of Q^m_{obs}.

5.3 Evaluation of Uncertainty Propagation in a Hydrometeorological System

The propagation of uncertainty and/or deficiencies of ensemble precipitation fields into hydrological response has been tested by means of three hindcast experiments. Each experiment consists of applying the hydrometeorological system to produce ensemble streamflow hindcasts for $N_{ev} = 100$ precipitation events selected among the 'observed' precipitation database. For each event, starting from information at the coarse scale, ensemble precipitation hindcasts are first generated by the STRAIN model according to a specific calibration mode, producing either consistent or overdispersed or underdispersed precipitation fields. These fields are then used in cascade to force the tRIBS model and the resulting ensemble streamflows are postprocessed together with the correspondent observation (furnished by the 'observed' hydrographs database) using the verification procedure described in the previous section.

5.3.1 Choice of T_{ver} and Event-Based Experiments Setup

Hindcast experiments have been setup according to an event-based approach aimed at applying the ensemble streamflow verification method. As already mentioned, the use of the verification procedure requires to preliminary fix the duration T_{ver} of the event where the rank is calculated and the time length of each ensemble precipitation member used to force the hydrological model. Provided that our purpose is to evaluate uncertainty propagation of precipitation forecasts covering T hours, the value of T_{ver} should be large enough to contain the hydrological effect caused by rainfall events occurring within T. In order to account for the effect of rainfall predicted in the last hours of T, T_{ver} should include the entire duration T of the downscaled event plus the basin response time, which is the time-lag between precipitation storm and streamflow occurrence for that basin. This last time interval is related to the catchment scale and characteristics and can be empirically estimated by the basin concentration time T_c (reported in Table 5.2 for the outlet and 14 sub-basins here analyzed). As a consequence, T_{ver} should be at least equal to $(T + T_c)$.

It is worthy to notice that, in theory, the basin response time is not a fixed parameter but it also depends on the initial state of the basin. In general, for a wet basin, we expect smaller response time and viceversa in case of a drier basin. In our study we do not account for variation of T_c with the basin conditions and we have assumed a constant value approximated with T_c. However, we acknowledge that specific analysis accounting for this aspect should be carried in future works.

The value assumed by T_{ver} determines the minimum duration required to the simulated hydrograph. As a result, the time length of precipitation ensemble used to force the hydrological model needs to be equal to or greater than T_{ver}. Since we test a wide range of basin scales with different T_c (ranging from 0.2 to 12.6 hours), we should in principle identify different values for T_{ver} in each sub-basin and, consequently, different durations of the input precipitation ensemble. However, the value of T_{ver} in the case of the basin outlet is surely greater than all the possible $(T + T_c)$ of the nested sub-basins and can be adopted as the only fixed value used to apply the verification procedure for all the catchment scales.

Selection of T_{ver} and of precipitation input duration used in our experiments have been made considering two possible approaches. As a first option, we could set $T_{ver} = (T + T_c)$ in the basin outlet and force the tRIBS model with downscaled precipitation fields for the first T hours and then add other T_c hours of zero rainfall (zero padding).

The second approach, that we preferred and adopted, is derived from the following operative consideration aimed at utilizing all the available information at the coarse scale provided by NWP models. Each prediction furnished by these models has a forecast lead time, let say T_{meteo}, which can be larger than T. Starting from NWP model output, we can calculate the mean precipitation value in the coarse spatial domain $L \times L$ km^2 for a number M of subsequent T-hour long periods, obtaining the values $R_1, R_2, \ldots, R_M$. These values can be in turn separately downscaled and the resulting high resolution precipitation fields can be concatenated covering a duration of $M \times T$ hours. In our exercise, $T = 16$ and $T_c = 12.59$ hours for the Baron Fork outlet, implying $T_{ver} \geq 28$ hours. For purpose of this study, every ensemble precipitation member used as input for tRIBS has been built by downscaling two consecutive coarse precipitation values and T_{ver} has been set equal to the entire duration of the precipitation forecast (i.e. $T_{ver} = 2T = 32$ hours).

Fig. 5.10 illustrates the event-based approach adopted in the experiments.

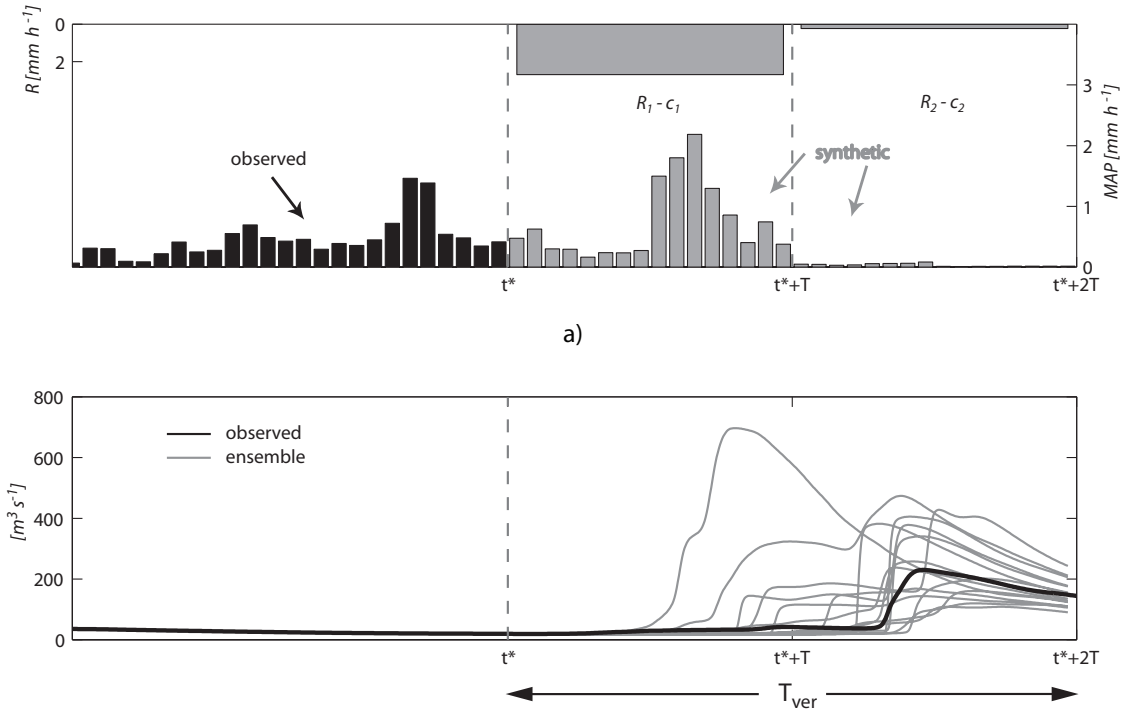

Figure 5.10: Setup of the event-based hydrological simulations. Panel a: the tRIBS is forced with observed precipitation up to t^*, the time when coarse scale information R_1 and R_2 are provided, and with two consecutive synthetic precipitation fields downscaled from R_1 and R_2 using STRAIN parameters c_1 and c_2. Panel b: ensemble and observed hydrographs in the interval T_{ver} used to calculate the rank of the observation according to the verification procedure for ensemble streamflow.

The tRIBS model is forced with 'observed' precipitation up to t^*, the time when coarse rainfall forecast (in our case, hindcast) is provided (panel a). Subsequently, two consecutive precipitation values R_1 and R_2 at the coarse scale are extracted, and the corresponding STRAIN model parameters c_1 and c_2 are determined according to a specific calibration mode. Parameters c_1 and c_2 are used to downscale precipitation in the time intervals $[t^*, t^* + T]$ and $[t^* + T, t^* + 2T]$, respectively; finally, the two synthetic spatiotemporal fields are concatenated. Ensemble rainfall fields of duration $2T$, each made of two T hour-long downscaled fields, are generated and used to force the tRIBS model, which in turn produces the ensemble hydrographs shown in panel b. Observed and ensemble streamflow values within $T_{ver} = 2T$ are then used to apply the proposed verification procedure to test ensemble streamflow consistency. We remember once again that in this exercise, we have assumed no uncertainty in the coarse precipitation values R_1 and R_2 and therefore we do not use NWP model products and use instead the NEXRAD radar estimates upscaled to the coarse scale resolution.

The adoption of this event-based approach has some implications. Assuming that T_c represents exactly the basin response time, streamflow values contained in T_{ver} can be generated by precipitation events observed in $[t^* - T_c, t^*]$ and by precipitation events forecasted in the first or/and in the second downscaling time intervals $[t^*, t^* + T]$ and $[t^* + T, t^* + 2T]$ respectively. In particular, the larger the basin response time, the smaller the influence of the second downscaled event. If R_1 and R_2 were extracted from NWP forecasts, they would be characterized by different uncertainty levels because forecast skill of meteorological models varies with time (Golding 1998). As a consequence, uncertainty of hydrological response would include the combined effect of two kinds of uncertainty associated to coarse scale information. However, in our study we do not account for uncertainty in the coarse scale information and examine streamflow values whose uncertainty depends only on the characteristics of precipitation fields simulated by the downscaling model, which are in turn related to the calibration mode used to select the parameters. Thus, if the same calibration mode is utilized to generate both synthetic fields corresponding to R_1 and R_2, the resulting ensemble hydrographs depend only on the same type of uncertainty.

5.3.2 Hindcast Experiments

Three hindcast experiments have been carried out in 'controlled conditions' using 'observed' precipitation and streamflow database as verification. Every experiment has been conducted on $N_{ev} = 100$ events according to the approach depicted in Fig. 5.10. In each event, the two coarse rainfall R_1 and R_2 have been set equal to two consecutive precipitation values $R_{i,l}$ and $R_{i+1,l}$, respectively, selected among the 138×9 precipitation values used to build the 'observed' database. Events were selected in order to mimic the occurrence of large-scale event. In this way, we did not focus only on those events generating storms and then flood in the study basin, but we evaluated performances of the forecasting system in all the possible situations regarding precipitation and flood occurrence/non occurrence in the basin starting from the only information at the coarse scale.

STRAIN model parameters c_1 and c_2 have been selected in each experiment according to the same calibration mode and a different mode has been adopted in the three experiments to produce ensemble precipitation hindcasts with different characteristics. In particular, the 'functional-based', 'event-based' and 'mean-based' calibration modes, illustrated in section 4.2.2, have been utilized to generate consistent, overdispersed and underdispersed precipitation ensemble in experiments called CONS, OVER and UNDER, respectively.

A procedure analogous to the one described in section 4.2.2 has been followed to determine STRAIN parameters in the three calibration modes. A total of $N_{ev} = 100$ high resolution precipitation 'observed' events, corresponding to the coarse values $R_{i,l}$ selected in the experiments, have been used to estimate the calibration relation linking STRAIN parameter c and coarse scale rainfall R, assuming a no a-priori knowledge about how these events have been generated. Fig. 5.11 shows the calibration relation $c = c(R)$, given by equation 2.5 with $c_\infty = 0.675$, $a = 0.907$ and $\gamma = 0.764$ and utilized to generate the 'observed' fields (dashed line). Grey asterisks represent the 100 parameters $c_{i,l}^{est}$ estimated on the observed events coming from the correspondent selected $R_{i,l}$, while the black line is the new calibration relation $c = c^{cal}(R)$ fitted on the $c_{i,l}^{est}$. It is apparent that the two calibration relations are very close to each other.

In all the experiments, $N_{ens} = 50$ synthetic precipitation fields have been generated to hindcast each event. In experiment CONS, STRAIN parameters of the two consecutive downscaled fields have been set to $c_1 = c^{cal}(R_{i,l})$

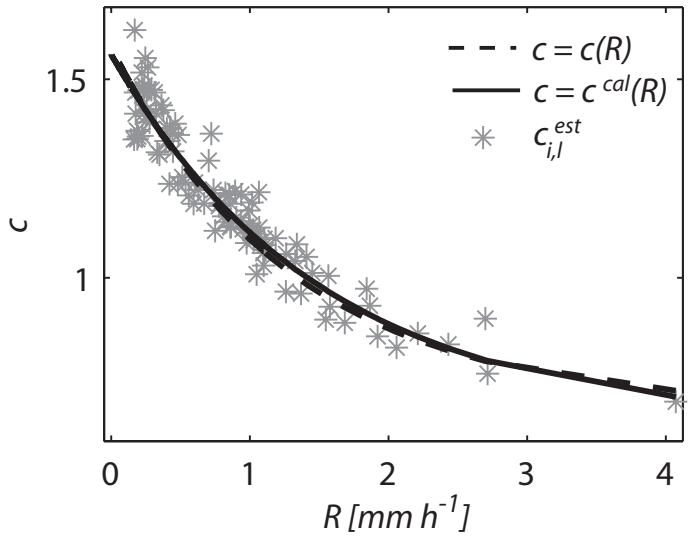

Figure 5.11: Calibration relation between STRAIN parameter c and coarse rain rate R. The dashed black line represents the calibration relation $c = c(R)$, asterisks represent the parameters $c_{i,l}^{est}$ estimated on each 'observed' event and the solid black line is the calibration relation $c = c^{cal}(R)$ fitted on $c_{i,l}^{est}$.

and $c_2 = c^{cal}(R_{i+1,l})$; in experiment OVER, $c_1 = c_{i,l}^{est}$ and $c_2 = c_{i+1,l}^{est}$, while in experiment UNDER, the mean value c_{mean} of the $c_{i,l}^{est}$ has been always adopted. Consistency, overdispersion and underdispersion of the synthetic fields generated according to the 'functional-based', 'event-based' and 'mean-based' calibration modes have been verified through the graphical method for precipitation ensemble described in chapter 4. The resulting VRHs are shown in panels a, b and c of Fig. 5.12 for the three cases.

Subsequently, for a given experiment, the tRIBS model has been forced in the $N_{ev} = 100$ events by the ensemble precipitation hindcast and $N_{ens} = 50$ hydrographs have been outputted at the 15 locations in each event. This has resulted in $(N_{ens} = 50) \times (N_{ev} = 100) \times (N_{exp} = 3) = 15,000$ hydrological simulations requiring a significant computational effort for which a Linux cluster with 64 processors has been used. A capability of tRIBS model that has enormously reduced the time required by the simulations, is the so called RESTART option, which allows the user to save the simulation state at a given time, t^* in our case, and then to run N_{ens} T_{ver} hour-long simulations including only synthetic precipitation, instead of running each simulation starting from the beginning of the summer.

The verification procedure for ensemble streamflow has been applied to test consistency of the hydrological response in all the nested sub-basins, selecting the accumulated streamflow at durations 1, 16 and 32 hours (Q_{1h}, Q_{16h},Q_{32h}) as metrics used to determine the rank of the observation. The resulting VRHs for experiments CONS, OVER and UNDER are reported in Fig. 5.13, 5.14 and 5.15, respectively. For sake of clarity, each figure shows results only for basins 13, 9, 7, 6 and 15 (see Table 5.2), whose sizes span the entire range of basin scale and each column refers to a specific metric. In every histogram, the 100 ranks are grouped in 10 bins and the 5%, 10%, 25%, 50%, 75%, 95% quantiles of a uniform distribution are plotted using horizontal lines. Results reveal that, in all the VRHs, hypothesis of uniformity cannot be discarded, implying that consistency is achieved, in all the experiments, for all the basins and the metrics.

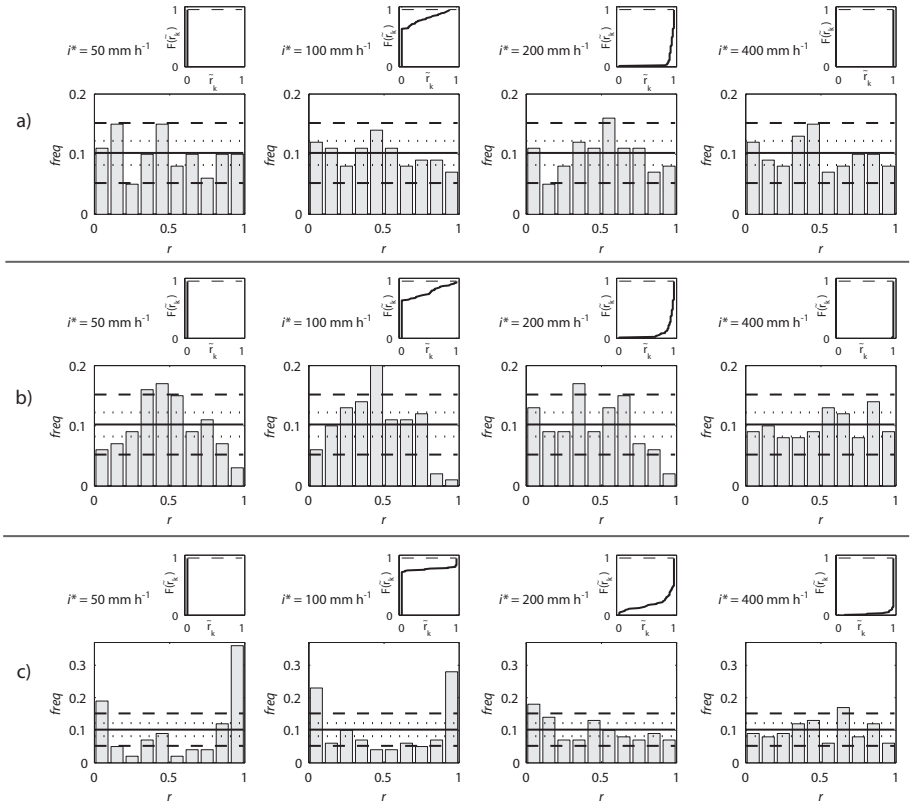

Figure 5.12: Verification Rank Histograms constructed, according to the verification method described in chapter 4, from rainfall ensembles used to force the tRIBS model in the three hindcast experiments. Panel a: consistent ensembles generated with the 'functional-based' calibration mode. Panel b: overdispersed ensembles generated with the 'event-based' calibration mode. Panel c: underdispersed ensembles generated with the 'mean-based' calibration mode.

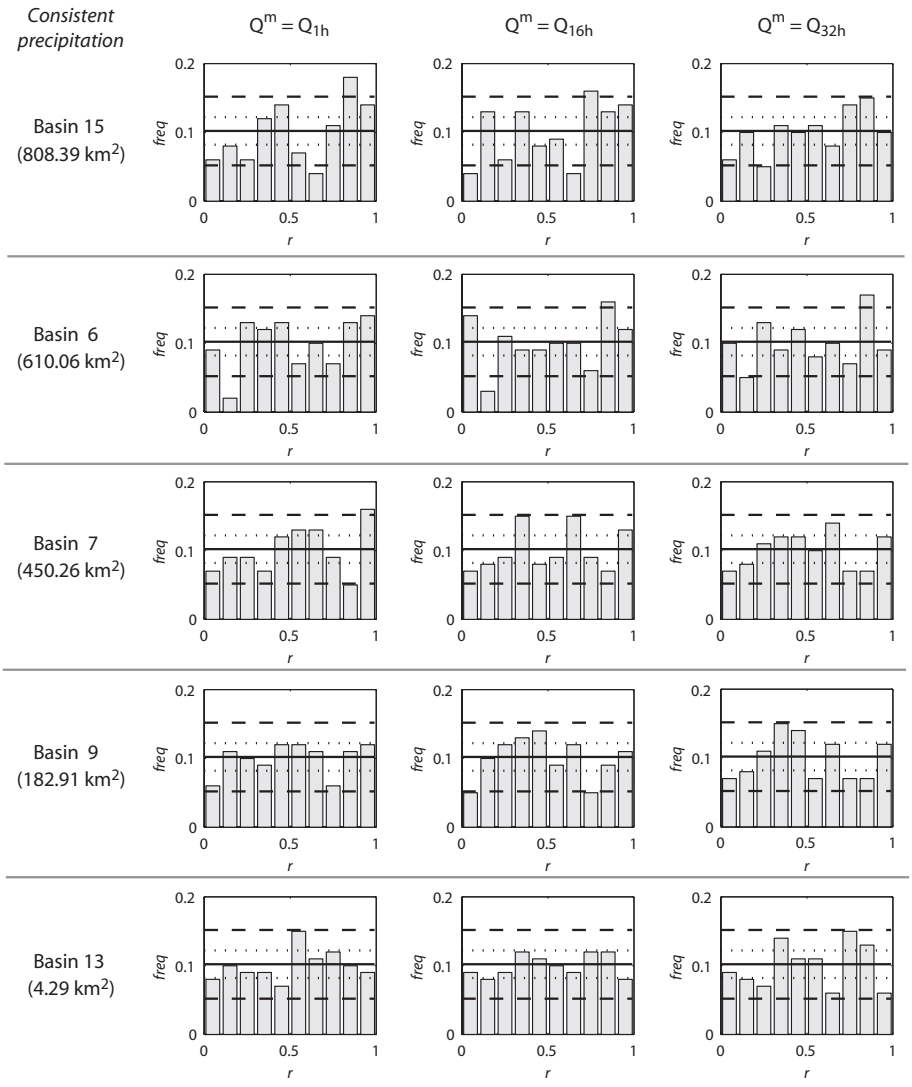

Figure 5.13: Experiment CONS: Verification Rank Histograms built from the ensemble streamflows obtained forcing the tRIBS model with consistent precipitation ensemble. Results are shown for basins 15, 6, 7, 9 and 13 covering the entire range of basin scales and for the metrics Q_{1h}, Q_{16h} and Q_{32h}. Consistency is achieved in all the cases.

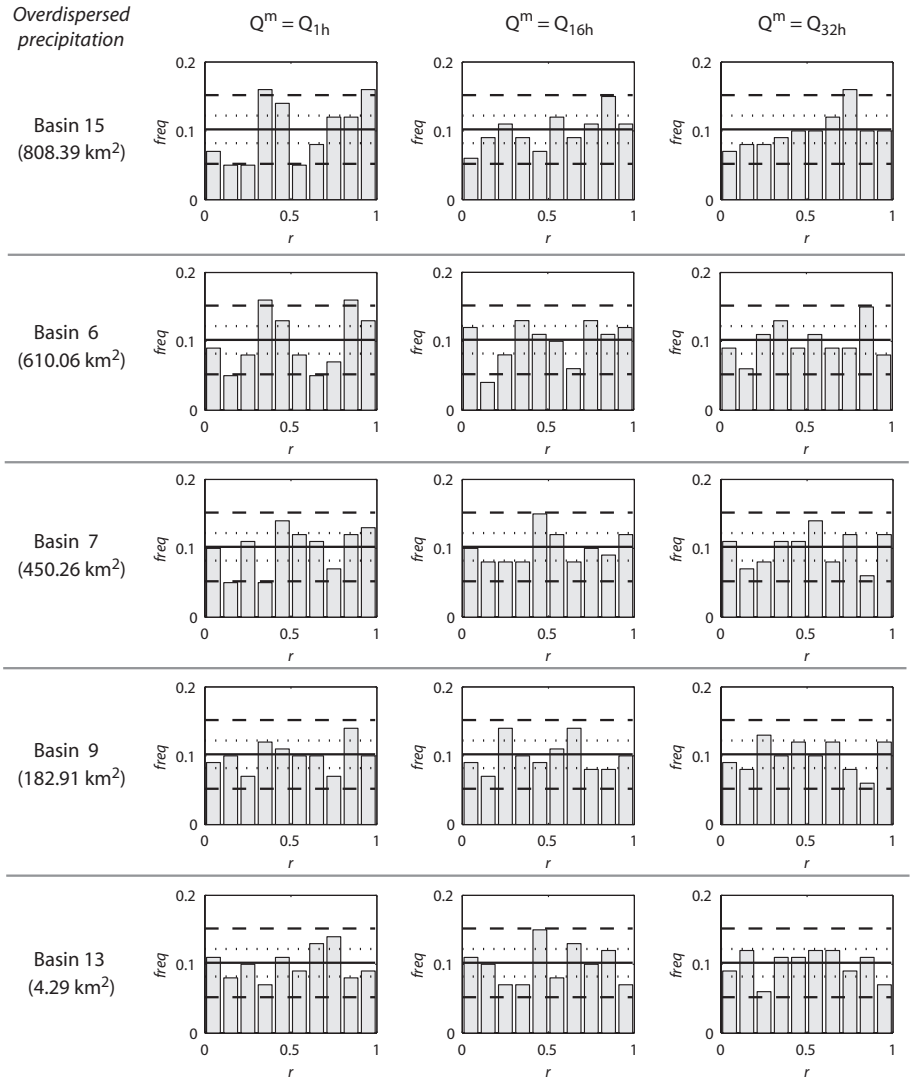

Figure 5.14: Experiment OVER: Verification Rank Histograms built for the ensemble streamflows obtained forcing the tRIBS model with overdispersed precipitation ensemble. Consistency is achieved in all the cases.

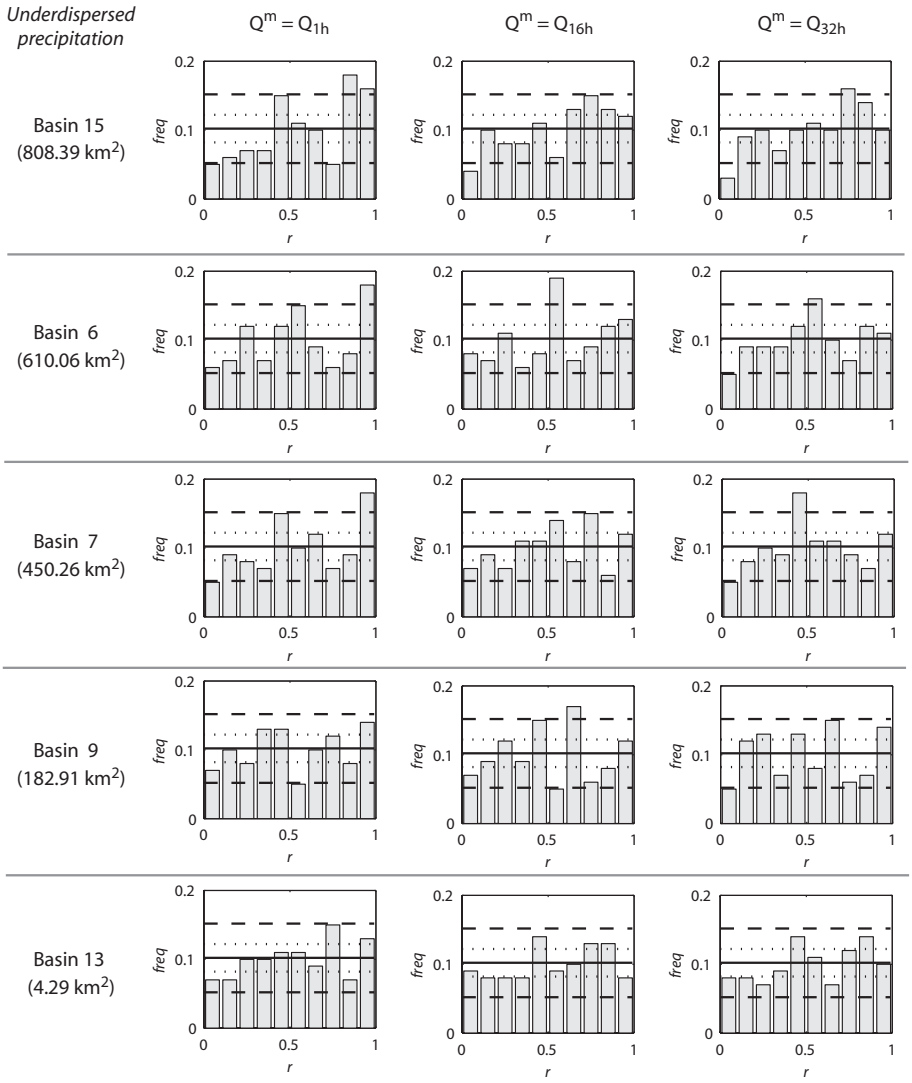

Figure 5.15: Experiment UNDER: Verification Rank Histograms built for the ensemble streamflows obtained forcing the tRIBS model with underdispersed precipitation ensemble. Consistency is achieved in all the cases.

5.4 Results and Discussion

Analysis of the VRHs obtained applying the hydrological verification procedure in the three hindcast experiments, permits us to conclude that running the tRIBS model with either consistent or overdispersed or underdispersed precipitation always leads to consistent ensemble streamflow.

This result can be further investigated by comparing, in each event, the values of the ranks calculated in the three experiments. Fig. 5.16 contains 5 panels, relative to the basins 15, 6, 7, 9 and 13, showing the ranks of the metric Q_{1h} obtained from ensemble hydrographs produced by experiments CONS, OVER and UNDER, plotted versus the logarithm of coarse precipitation value of the first downscaled field $R_1 = R_{i,l}$, which has in general the largest influence in streamflow generation. For sake of clarity, only 10 events spanning the range of values of $R_{i,l}$ are shown, but analysis carried out on the other events and metrics provides similar results. Inspection of Fig. 5.16 suggests that: (i) there is no relationship between the ranks and the coarse precipitation of the first downscaled field for any of the experiments; (ii) for a given event, the ranks are always very close one each other, indicating that the ECDFs of the metric are very similar in all the three experiments. In particular, if we calculate in each event the absolute difference between the two farthest ranks, we obtain an average value of approximately 0.09 in all the basins, meaning that the three ranks very often fall in the same bin of the respective VRH. An example of the ECDFs of the metric Q_{1h} is reported in Fig. 5.17 for the event with $R_1 = 1.86$ mm h^{-1} and for the basins 13, 7 and 15. The three lines are always very close to each other implying similar values for the ranks of Q_{obs}.

Additional analyses have been focused on ensemble dispersion, which has an influence on ensemble consistency and refers to the distribution of ensemble members irrespective of the correspondent observation. In particular, we have studied how dispersion of ensemble precipitation varies with the catchment scale and the calibration modes and how it affects dispersion of the correspondent ensemble streamflow. For this purpose, we have defined and extracted one metric from each member of ensemble precipitation (a spatiotemporal field with time length T_{ver}), and another metric, dependent on the first one, from each member of ensemble streamflow (a time series of duration T_{ver}). For a given sub-basin and experiment, we could then build the two empirical probability distributions for the two

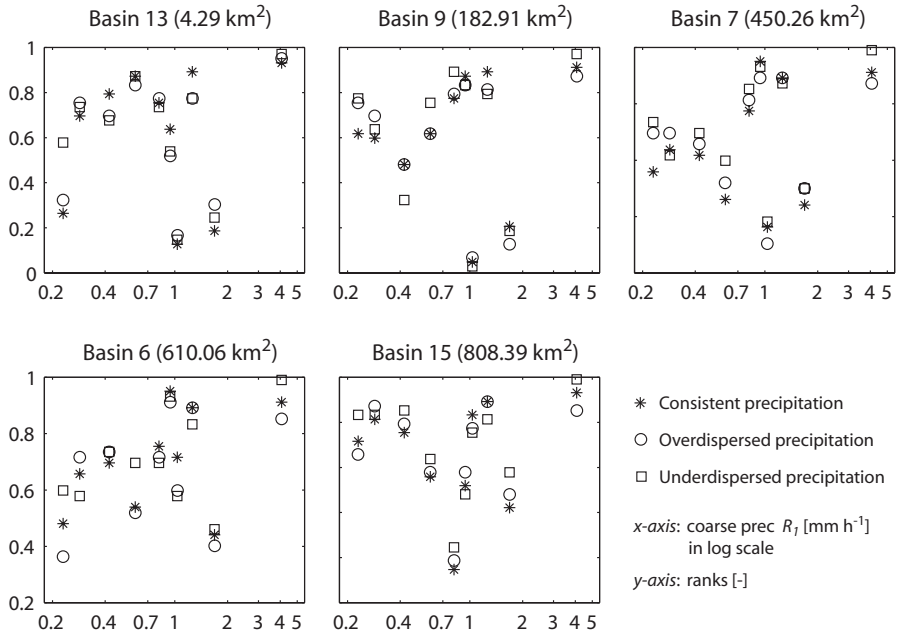

Figure 5.16: Ranks of Q_{1h} obtained from ensemble streamflows produced by consistent (asterisk), overdispersed (square) and underdispersed (circle) precipitation ensembles in 10 events at basins 13, 9, 7, 6 and 15.

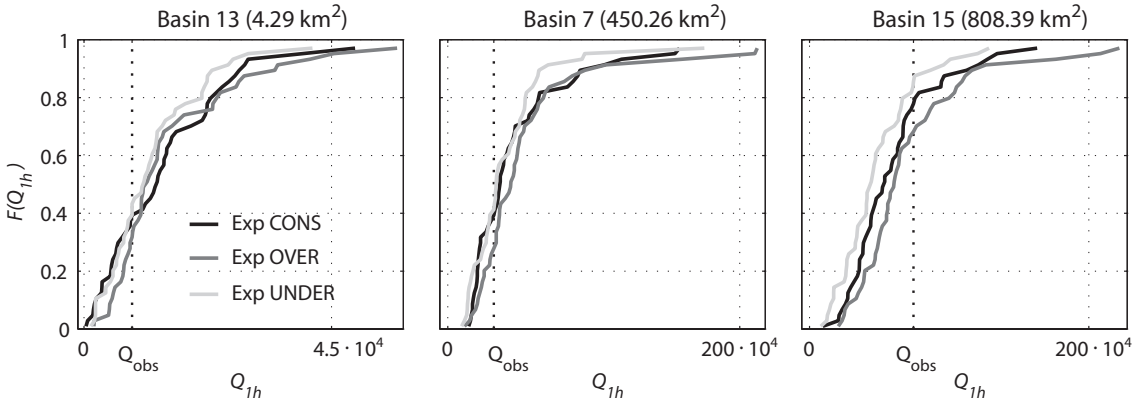

Figure 5.17: Empirical Cumulative Density Functions of the metric Q_{1h} ([m ^{3} h^{-1}]) obtained from the ensemble streamflows produced by experiments CONS, OVER and UNDER. Panels from the left to the right are referred to the sub-basins 13, 7 and 15 with increasing area. The ranks of Q_{obs} for the three experiments always assume very similar values.

metrics and analyze their dispersion.

The first metric has been selected as follows. For a sub-basin of size A, we have considered the rainfall values of each T_{ver}-hours long ensemble member falling over the Baron Fork basin and determined the maximum precipitation accumulated over 1 hour in a square of size S [km], where S is multiple of the spatial fine scale resolution and comparable with the square root of A (i.e. $S \times S \approx A$). In particular, we have analyzed basins 13, 9, 7, 6 and 15 (see Tab. 5.2) with respective areas $A = 4.29$, 182.91, 450.26, 610.06 and 808.39 km^2, implying the corresponding spatial scales $S = 4$, 12, 20, 24 and 28 km. The second metric has been instead extracted from the ensemble streamflow obtained in each sub-basin and has been set equal to the hourly maximum streamflow Q_{1h} within the time length T_{ver}.

In summary, for each experiment and sub-basin and for a given event, we have extracted $N_{ens} = 50$ hourly precipitation maxima at scale S and the corresponding $N_{ens} = 50$ hourly streamflow maxima Q_{1h}. From these two samples, we have built the empirical probability distributions and we have measured their dispersion by means of the Coefficient of Variation (CV), which is the ratio between the standard deviation and the average of the sample. As a result, from each experiment we have obtained $N_{ev} = 100$ CVs characterizing dispersion of ensemble precipitation and the correspondent $N_{ev} = 100$ CVs characterizing dispersion of the ensemble streamflow.

Fig. 5.18 and 5.19 illustrate results for precipitation and streamflow respectively. Panel a of both figures shows the average $< CV >$ of the CVs calculated from the N_{ev} events plotted versus the basin size (which is represented by $S \times S$ or A in the two figures) for the three experiments CONS, OVER and UNDER. Panel b focuses instead on the N_{ev} values of the CV calculated for sub-basin 7 ($A = 450.26$ km^2 and $S \times S = 400$ km^2) and shows the distribution of the relative frequency of their occurrence in the three experiments. Sub-basin 7 has been chosen as an example but similar behaviors have been detected in the other sub-catchments.
Results of Fig. 5.18 and 5.19 reveal that:

- For all the experiments, the $< CV >$ of precipitation maxima is almost constant for all the sub-basins, meaning that dispersion of ensemble precipitation does not change with catchment size (panel a of Fig. 5.18).

- In contrast, $< CV >$ of Q_{1h} decreases with basin scale, implying a much higher ensemble dispersion for the smaller basin (panel a of

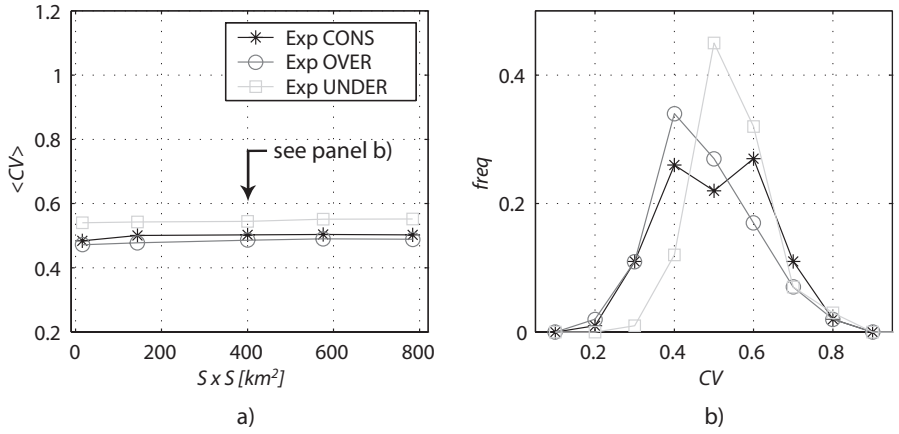

Figure 5.18: Dispersion of ensemble precipitation hindcasting each event, measured by the CV of the $N_{ens} = 50$ hourly precipitation maxima at spatial scale S [km]. Panel a shows, for each experiment, the average $< CV >$ of the $N_{ev} = 100$ CVs at spatial scale S versus the corresponding area $S \times S$. Panel b reports the distributions of the relative frequency of occurrence for the N_{ev} CVs obtained in the case $S \times S = 400$ km^2.

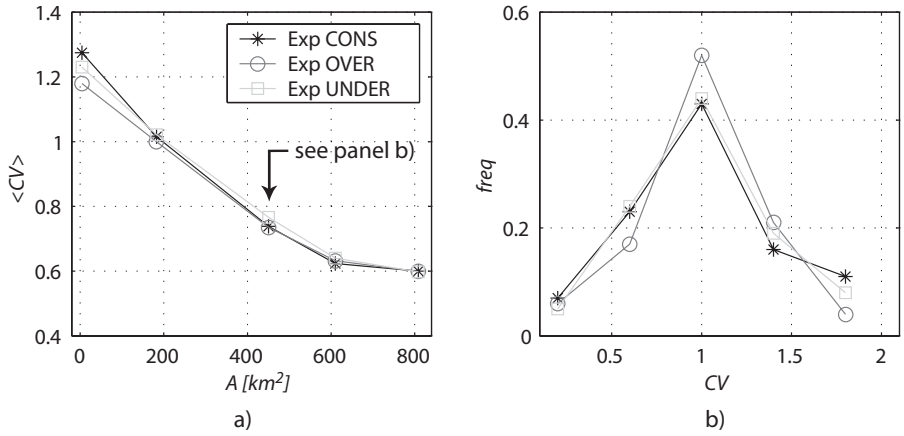

Figure 5.19: Dispersion of ensemble streamflow, measured by the CV of the $N_{ens} = 50$ Q_{1h}. Panel a shows, for each experiment, the average $< CV >$ of the $N_{ev} = 100$ CVs for sub-basins 13, 9, 7, 6 and 15 of Tab. 5.2 versus the respective area A [km^2]. Panel b reports the distributions of the relative frequency of occurrence for the N_{ev} CVs obtained for basin 7.

Fig. 5.19). This implies that sensitivity of the basins to differences in precipitation input increases as their size decreases and resulting streamflow ensemble are characterized by higher uncertainty.

- In the case of ensemble precipitation, the $< CV >$ for experiments CONS and OVER is almost the same in each catchment, while it always assumes the highest value for experiment UNDER (panel a of Fig. 5.19). Further, panel b of Fig. 5.18 reveals that the distribution of the N_{ev} CVs obtained for experiment UNDER are more concentrated around the mean, while experiments CONS and OVER return more spread values of CV with larger standard deviations. This is explained considering that parameter c of STRAIN model directly controls the CV of the synthetic fields: the CVs of underdispersed precipitation fields assume a more constant value throughout the N_{ev} events because the same parameter c_{mean} has always been adopted, while CVs of consistent and overdispersed ensembles are characterized by more variability since they have been produced using different values of c in each event.

- Conversely, the $< CV >$ of streamflow ensemble obtained from the three experiments, are almost equal in each sub-basin (panel a of Fig. 5.19) and panel b of Fig. 5.19 shows that even the distributions of the relative frequencies of the N_{ev} CVs in a given sub-basin are very close one each other. This implies that the basins act as filters for precipitation fields with different characteristics, returning similar behavior for dispersion of ensemble hydrographs in all the cases.

The study described in this chapter aimed at assessing propagation of uncertainty of precipitation input into hydrological response, has two important implications:

1. The first implication is hydrological and is related to the role played by the basins, which are characterized by two opposite response mechanisms with respect to precipitation spatiotemporal variability. On one hand, basins separate the different runoff components and act as a non-linear filter emphasizing intermittency and multifractal characteristics of precipitation. On the other hand, they act as complex integrators of precipitation in space and time mitigating the spatio-temporal precipitation variability. Vegetation, soil texture,

aquifer and basin geomorphometric characteristics play an important role within these opposite mechanisms. Results achieved in this study for the Baron Fork basin have revealed that, for simulations covering limited periods of 32 hours, the second mechanism is dominant and the basin behaves like a powerful spatio-temporal integrator. This behavior has been detected not only for the larger scales (i.e. $\sim 10^3$ km^2) but also for the smallest ones covered by a single radar pixel (~ 16 km^2).

2. The second implication is instead related to downscaling model calibration. We demonstrated how calibration relations linking downscaling parameters and coarse meteorological observable can be able to account for intrinsic variability of model parameters and lead to the generation of consistent ensemble fields. The study carried out in this chapter, hypothesizing the existence of a single calibration relation, has revealed that ensemble streamflow produced by consistent precipitation are consistent too. Further, it has shown that consistency of ensemble hydrographs can be achieved also when the hydrological model is forced by precipitation fields affected by underdispersion or overdispersion deficiencies. This suggests a certain flexibility in the use of downscaling model in forecasting system, because even a calibration returning precipitation ensembles that are not perfectly consistent, situation that can likely occur when real-world data are used, can lead to the simulation of consistent ensemble streamflows. However, this aspect requires a specific and careful evaluation, case by case, regarding both the downscaling model and the characteristics of the basin where hydrological predictions are made.

Chapter 6

Conclusions

Ensemble forecasting technique has been originally developed in meteorology and more recently adopted in hydrology to account for the different sources of uncertainty, mainly due to data (input and output), state variable, parameterization and model structure.

In the last years, ensemble technique has also been used to provide probabilistic predictions in hydrometeorological forecasting systems, which are extremely important for civil protection and water resources management. Advanced hydrometeorological systems for ensemble streamflow forecasts are based on schemes that include the combined use of meteorological and hydrological models as well as downscaling models and data assimilation systems. The complexity of such schemes requires the creation and testing of rigorous verification methods to evaluate the sources of uncertainty involved.

Assessment of precipitation forecast uncertainty and its propagation into hydrological response is fundamental in spatially-distributed forecasting systems, but is has been so far barely analyzed. In this work we have proposed systematic verification methods to evaluate propagation of precipitation input uncertainty within a hydrometeorological forecasting system for flood prediction in catchments with short response time.

For purpose of this study, we have designed a forecasting system that starts from output at coarse scale provided by NWP models and couples in cascade a precipitation downscaling model with a fully distributed hydrological model. Given the complexity and high non-linearity of the

processes involved, the study has been focused only on the hydrological part, including precipitation downscaling model and distributed hydrological model. No uncertainty has been instead associated to output at the coarse spatiotemporal scale provided by NWP models.

Precipitation downscaling models start from coarse scale information, furnished presumably with low uncertainty by NWP models, and provide ensembles of spatiotemporal precipitation fields at high resolution. A particular class of statistical downscaling models is based on the multifractal theory and are able to reproduce observed intermittency and small scale variability. Their operation is usually assured by calibration relations linking their few parameters with a coarse meteorological observable or predictant, such as the mean rainfall rate.

Physically based, distributed hydrological models can in turn offer advantages over conceptual, lumped models, widely used for flood forecasting, since they are able to capture hydrological processes in a wide range of scales. In addition, they have the capability of simulating discharge forecasts at interior locations, time series of runoff generation at particular sites and spatiotemporal fields of hydrological response.

A first part of the study has been devoted to uncertainty characterization of ensemble precipitation fields forecasted by precipitation downscaling models. For this purpose, we have proposed a verification procedure based on the generalization of the verification rank histogram (VRH), a graphical device used in applied meteorology to test the consistency hypothesis (i.e. ensemble and observation are drawn from the same distribution) of univariate variables. Since downscaling models reproduce the probability of precipitation at high resolution, they cannot be verified in a deterministic way but should be tested by evaluating their ability in reproducing the statistical behavior of precipitation. Therefore, the univariate variable adopted here is the probability of exceedance of a fixed precipitation threshold i^* calculated from each spatiotemporal field.

The generalization of verification rank histograms has been performed as follows. First, a precipitation threshold i^* is fixed. Then, for each event, the exceedance probabilities of i^* are calculated for the N_{ens} ensemble and the observed fields and the position p of the observed exceedance probability is found. Finally, the rank histogram is built with the normalized ranks r, defined as the cumulative frequency of p computed for each verification event. The procedure is repeated for different values of i^*, spanning a range interesting for hydrometeorological applications. As i^* increases, the

observed and ensemble exceedance probabilities can be equal to zero in a certain number of verification events, so that the ranks are randomly assigned. As a consequence, the histogram shape becomes artificially uniform, making the detection of model forecast deficiencies more difficult. To avoid possible erroneous evaluation of model performances, we have introduced a graphical method based on the interpretation of the ECDF of a variable $\tilde{r}_k$ accounting for random assignment.

The verification procedure has been applied and tested using the STRAIN downscaling model. The model depends on two parameters c and β and is able to simulate homogeneous precipitation fields in a self-similar framework. Three numerical experiments have been carried out in controlled conditions according to the following approach. STRAIN model has been first used to generate 'observed events' with selected values of parameters c and β. These 'observed events' have been used to calibrate STRAIN parameter assuming a no a-priori knowledge of the method used to generate them. Finally, each 'observed-event' has been hindcasted using STRAIN according to three calibration modes for parameters determination: 'event-based', 'mean-based' and 'functional-based'. Results of the three experiments permit us to derive the following conclusions relative to the first part of the work:

1. If we consider a hindcast framework and we generate the ensemble members adopting the parameter c_k^{est} estimated on the same event k to be hindcasted (at a first sight, the best possible solution to simulate the observed event), the model returns overdispersed forecasts. This is due to the fact that model sampling variability is not accounted for in parameters calibration and a centering of ensemble members around the observation is produced.

2. The intrinsic variability of downscaling model when the average of the estimates c_k^{est} is used, may not be able to capture the variability of observed events and underdispersed forecasts are produced.

3. The use of a calibration relation linking model parameter with a meterological observable at coarse scale may allow model sampling variability to be taken into account leading to consistent members.

4. When observed events display a large variability that a single calibration relation is not able to explain, underdispersed forecasts are produced. For example, this variability can be due to different physical

origins or different synoptic conditions generating the events. In order to reach consistency, it would be necessary first to classify the events according to their physical origin and then to estimate storm-dependent calibration relations.

We remark that systematic analyses of the effects of precipitation type on scale-invariance statistical properties have not been yet conducted. We believe that this can be an interesting topic for future research and the proposed verification method can be an useful tool to assess the need for single or multiple calibration relations.

We also highlight that the verification method, tested with a homogeneous and isotropic model, can be applied whatever the downscaling method used since the method does not refer to the generation mechanism of the model. A slight modification needs to be adopted only in case of a downscaling model reproducing spatial heterogeneity, because, in this case, the analysis should be carried out in each location rather than in the entire spatial domain.

In a second part of the study, we have used results of the previous part to analyze how uncertainty and deficiencies of ensemble downscaled precipitation forecasts affect hydrological response. A verification method based again on the VRH has been first developed to test consistency of streamflow ensembles. The method requires the fixation of a verification time length, dependent on the basin response time, where a certain metric is extracted from ensemble and observed hydrographs and then used to build the VRH.

Three numerical hindcast experiments have been then carried out applying the hydrometeorological system coupling the STRAIN precipitation downscaling model and the tRIBS distributed hydrological model in the Baron Fork basin and 14 nested sub-basins (areas ranging from 0.78 to 808 km^2). Hindcast experiment have been conducted in controlled conditions, to assure an easier control and assessment of uncertainty propagation. First, a database of 'observed' precipitation fields covering several summer periods have been generated through the STRAIN model with selected values of parameters c and β (i.e. known statistical properties). For this aim, existence of scale invariance laws has been assumed in a range of scales found in past applications on real data, and a single calibration relation linking coarse rainfall and STRAIN parameters has been adopted. To preserve climatology of the studied area and period of the year, the precipitation values at the coarse scale used in the generation have been extracted from

the radar estimates of the NEXRAD network. The 'observed' precipitation have been utilized to force the tRIBS hydrological model. Since uncertainty of hydrological model parameters has not been taken into account, tRIBS has been calibrated in one summer and the resulting parameter values have been kept fixed in all the model runs for all the summers. The simulated hydrographs have been considered as ground truth and used as verification for the three hindcast experiments.

Subsequently, we have assumed a no a-priori knowledge about the origin of precipitation and streamflow 'observations' and produced N_{ens} ensemble streamflow hindcasts for N_{ev} precipitation events selected among the 'observed' database. An event-based approach has been setup to permit application of the proposed verification procedure in all the sub-basins. For each event, starting from information at the coarse scale, ensemble precipitation hindcasts have been first generated by the STRAIN model according to a specific calibration mode: the 'functional-based', 'event-based' and 'mean-based' calibration modes have been adopted to produce consistent, overdispersed and underdispersed precipitation hindcasts, respectively, in the three experiments. These fields have been then used in cascade to force the tRIBS model and the resulting streamflow ensembles at the 15 locations have been utilized together with the correspondent observation (furnished by the 'observed' hydrographs database) to build the VRH according to the proposed verification procedure.

Inspection of the VRHs for the three experiments shows that running the tRIBS model with either consistent or overdispersed or underdispersed precipitation always leads to consistent ensemble streamflow, irrespective of the basin scale.

Additional analyses have been focused on ensemble dispersion, which refers to the distribution of ensemble members without considering the corresponding observation. In particular, we have studied how dispersion of ensemble precipitation varies with the catchment scale and the calibration modes and how it affects dispersion of the correspondent ensemble streamflow. For this purpose, we have defined two metrics, related one each other, and calculated from each member of precipitation and streamflow ensemble, respectively: (i) the maximum hourly precipitation value in an area comparable to the size of the analyzed basin; (ii) the correspondent maximum hourly accumulated streamflow. Thus, for each event, we have built the empirical probability distribution functions for the two metrics extracted from the N_{ens} members of ensemble precipitation and streamflow and we have

measured their dispersion using the Coefficient of Variation (CV). Analyses carried out for the three experiments and for different catchment sizes, shows that:

- Dispersion of ensemble precipitation does not change with catchment scale, while, in contrast, dispersion of ensemble streamflow is higher for the smallest basin and decreases as basin area increases. This means that sensitivity of basins to differences in precipitation input is higher as their size decreases and resulting streamflow ensemble are characterized by a bigger level of uncertainty.

- For a given basin, underdispersed ensemble precipitation are characterized by a very similar level of dispersion in all the N_{ev} events (i.e. the N_{ev} CV are very close to their mean value), while the degree of dispersion has a larger variability (i.e. the N_{ev} CV have higher standard deviation) in the case of consistent and overdispersed ensemble. In contrast, these effects are not present anymore in the ensemble streamflow, whose dispersion results very similar in all the three cases.

In light of the analyses and preliminary considerations made in the second part of the study aimed at assessing propagation of uncertainty into hydrological response, two main conclusions can be drawn:

1. For the Baron Fork and its nested sub-basins and for simulations covering limited time period (32 hours), the basins always behave as complex integrators of precipitation in space and time, mitigating precipitation intermittency. Thus, ensemble precipitation input with different characteristics lead to the same characteristic for ensemble streamflow (i.e. consistency).

2. A certain flexibility in the use of downscaling models within forecasting systems is suggested. In fact, even if downscaling models calibration does not lead to perfectly consistent members, consistent ensemble streamflow may be however produced when the downscaled rainfall forecasts are used as forcing for the hydrological model. Nevertheless, this effect requires to be verified and confirmed with further investigations on real data and on basins with different characteristics.

Finally, we remark that, at our knowledge, this is one of the first studies that try to verify performances of a hydrometeorological forecasting system using a rigorous and systematic framework over a great number of events, causing or not a flood, instead of analyzing just few test cases for which no statistically significant conclusion can be drawn.

Bibliography

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

Anderson, J. L. (1997), 'The impact of dynamical constraints on the selection of initial conditions on ensemble predictions: low-order perfect model results', *Mon. Wea. Rev.* **125**, 2969–2983.

Badas, M. G., Deidda, R. & Piga., E. (2006), 'Modulation of homogeneous space-time rainfall cascades to account for orographic influences', *Nat. Hazards Earth Syst. Sci.* **6**, 427–437.

Beven, K. J. (1982), 'On subsurface streamflow: an analysis of response times', *Hydrol. Sci. J.* **27**, 505–521.

Bras, R. L. (1990), *Hydrology: an introduction to hydrologic science*, Addison-Wesley/Longman, Reading, MA/London.

Brooks, R. H. & Corey, A. T. (1964), Hydraulic properties of porous media, *in* 'Hydrol. Pap. 3', Colo. State Univ., Fort Collins.

Cabral, M. C., Garrote, L., Bras, R. L. & Entekhabi, D. (1992), 'A kinematic model of infiltration and runoff generation in layered and sloped soils', *Adv. Water Res.* **15**, 311–324.

Carpenter, T. M. & Georgakakos, K. P. (2004), 'Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model', *J. Hydrol.* **298**, 202–221.

Deidda, R. (2000), 'Rainfall downscaling in a space-time multifractal framework', *Water Resour. Res.* **36**, 1779–1794.

Deidda,
R., Badas, M. G. & Piga, E. (2004), 'Space-time scaling in high-intensity Tropical Ocean Global Atmosphere Coupled Ocean-Atmosphere Response Experiment (TOGA-COARE) storms', *Water Resour. Res.* **40**, W02056.

Deidda, R., Benzi, R. & Siccardi, F. (1999), 'Multifractal modeling of anomalous scaling laws in rainfall', *Water Resour. Res.* **35**, 1853–1868.

Dunne, T. & Black, R. D. (1970), 'An experimental investigation of runoff production in permeable soils', *Water Resour. Res.* **6(2)**, 478–490.

Ebert, E. E. & McBride, J. L. (2000), 'Verification of precipitation in weather systems: Determination of systematic errors', *J. Hydrol.* **239**, 179–202.

Enthekhabi, D. (2000), *Land Surface Processes: Basic Tools and Concepts*, Department of Civil and Environmental Engineering, MIT, Cambridge, MA.

Evans, M., Hastings, N. & Peacock, B. (2000), *Statistical distributions*, 3rd edn, New York: Wiley.

Ferraris, L., Rudari, R. & Siccardi, F. (2002), 'The uncertainty in the prediction of flash floods in the nothern mediterranean environment', *J. Hydrometeo.* **3**, 714–727.

Franz, K. J., Hatmann, H. C., Sorooshian, S. & Bales, R. (2003), 'Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin', *J. Hydrometeo.* **4**, 1105–1118.

Garrote, L. & Bras, R. L. (1995), 'A distributed model for real-time flood forecasting using digital elevation models', *J. Hydrol.* **167**, 226–306.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J. & Butts, M. B. (2004), 'Towards the characterization of streamflow simulation uncertainty through multimodel ensembles', *J. Hydrol.* **298**, 222–241.

Golding, B. W. (1998), 'Nimrod: a system for generating automated very short range forecasts', *Meteorol. Appl.* **5**, 1–16.

Grimit, E. P., Gneiting, T., Berrocal, V. J. & Johnson, N. A. (2006), 'The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification', *Quart. J. Roy. Meteor. Soc.* **132**, 2925–2942.

Gupta, V. K. & Waymire, E. C. (1993), 'A statistical analysis of mesoscale rainfall as a random cascade', *J. Appl. Meteor.* **32**, 251–267.

Hamill, T. M. (2001), 'Interpretation of rank histograms for verifying ensemble forecasts', *Mon. Wea. Rev.* **129**, 550–560.

Hamill, T. M. & Colucci, S. J. (1997), 'Verification of Eta-RSM short-range ensemble forecasts', *Mon. Wea. Rev.* **125**, 1312–1327.

Hamill, T. M. & Colucci, S. J. (1998), 'Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts', *Mon. Wea. Rev.* **126**, 711–724.

Horton, R. E. (1933), The role of infiltration in the hydrological cycle, *in* 'Trans. AGU', Vol. 14, pp. 446–460.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H. & Mitchell, H. L. (1996), 'A system simulation approach to ensemble prediction', *Mon. Wea. Rev.* **124**, 1225–1242.

Hu, Z. & Islam, S. (1995), 'Prediction of ground surface temperature and soil moisture content by the force-restore method', *Water Resour. Res.* **31**, 2531–2540.

Hursh, C. R. & Brater, E. F. (1941), Separating storm-hydrographs form small drainage-areas into surface and subsurface flow, *in* 'Trans. AGU', Vol. 22, pp. 863–870.

Ivanov, V. Y. (2002), A continuous Real-time Intercative Basin Simulator (RIBS), Master's thesis, R. M. Parson Lab. for Water Resour. and Hydrodyn., Mass. Inst. of Technol., Cambridge.

Ivanov, V. Y., Vivoni, E. R., Bras, R. L. & Entekhabi, D. (2004*a*), 'Catchment hydrologic response with a fully distributed triangulated irregular network model', *Water Resour. Res.* **40**, W11102.

Ivanov, V. Y., Vivoni, E. R., Bras, R. L. & Entekhabi, D. (2004*b*), 'Preserving high-resolution surface and rainfall data in operational-scale basin hydrology: a fully-distributed physically-based approach', *J. Hydrol.* **298**, 80–111.

Kirpich, Z. P. (1940), 'Time of concentration of small agricultural watersheds', *Civ. Eng.* **10**(362).

Leith, C. E. (1974), 'Theoretical skill of Monte-Carlo forecasts', *Mon. Wea. Rev.* **102**, 409–418.

Lin, J. D. (1980), 'On the force-restore method for prediction of ground surface temperature', *J. Geophys. Res.* **85**, 3251–3254.

Lorenz, E. N. (1963), 'Deterministic nonperiodic flow', *J. Atmos. Sci.* **20**, 130–141.

Lovejoy, S. & Mandelbrot, B. B. (1985), 'Fractal properties of rain, and a fractal model', *Tellus Series A* **37**, 209–232.

Lovejoy, S. & Schertzer, D. (1985), 'Generalized scale invariance and fractal models of rain', *Water Resour. Res.* **21**, 1233–1250.

McBride, J. L. & Ebert, E. E. (2000), 'Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia', *Wea. Forecasting* **15**, 102–121.

Molteni, F., Buizza, R., Palmer, T. N. & Petroliagis, T. (1996), 'The ECMWF Ensemble Prediction System: methodology and validation', *Quart. J. Roy. Meteor. Soc.* **122**, 73–119.

Monteith, J. L. (1965), 'Evaporation and environment', *Symp. Soc. Exp. Biol.* **122**, 226–234.

Murphy, A. H. (1973), 'A new vector partition of the probability score', *J. Appl. Meteor.* **12**, 595–600.

Murphy, A. H. & Winckler, R. L. (1987), 'A general framework for forecast verification', *Mon. Wea. Rev.* **115**, 1330–1338.

Over, T. M. & Gupta, V. K. (1994), 'Statistical analysis of mesoscale rainfall: dependence of a random cascade generator on large-scale forcing', *J. Appl. Meteor.* **33**.

Over, T. M. & Gupta, V. K. (1996), 'A space-time theory of mesoscale rainfall using random cascades', *J. Geophys. Res.* **101**, 26319–26332.

Penman, H. L. (1948), 'Natural evaporation from open water, bare soil and grass', *Royal Society of London Proceedings Series A* **193**, 120–145.

Perica, S. & Foufoula-Georgiou, E. (1996), 'Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions', *J. Geophys. Res.* **101**, 26347–26362.

Rutter, A. J., Kershaw, K. A., Robins, P. C. & Morton, A. J. (1971), 'A predictive model of rainfall interception in forests. 1. Derivation of the model from observation in a plantation of Corsican pine', *Agric. Meteorol.* **9**, 367–384.

Rutter, A. J., Morton, A. J. & Robins, P. C. (1975), 'A predictive model of rainfall interception in forests. 2. Generalization of the model and comparison with observation in some coniferous and hardwood stands', *J. Appl. Ecol.* **12**, 367–380.

Schaake, J. C., Hamill, T. M., Buizza, R. & Clark, M. (2007), 'HEPEX, the Hydrological Ensemble Prediction Experiment', *Bull. Amer. Meteor. Soc.* . in press.

Schertzer, D. & Lovejoy, S. (1985), Generalised scale invariance in turbulent phenomena, *in* 'Phys. Chem. Hydrodyn.', Vol. 6, pp. 623–635.

Slack, J. R., Lumb, A. M. & Landwehr, J. M. (2001), 'USGS water-resources investigations report, 93-4076'.

Talagrand, O., Vautard, R. & Strauss, B. (1997), Evaluation of probabilistic systems, *in* 'Proceedings, ECMWF Workshop on Predictability', Vol. 125, ECMWF. Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.

Toth, Z. & Kalnay, E. (1993), 'Ensemble forecasting at NMC: the generation of perturbations', *Bull. Amer. Meteor. Soc.* **74**, 2317–2330.

Tucker, G. E., Lancaster, S. T., Gasparini, N. M., Bras, R. L. & Rybarczyk, S. (2001), 'An object-oriented framework for distributed hydrologic and geomorphologic modeling using triangulated irregular networks', *Comp. Geosci.* **27(8)**, 959–973.

Verbunt, M., Walser, A., Gurtz, J., Montani, A. & C., S. (2007), 'Probabilistic flood forecasting with a Limited-Area Ensemble Prediction System: selected case studies', *J. Hydrometeo.* **8**, 807–909.

Vivoni, E. R., Entekhabi, D., Bras, R. L., Ivanov, V. Y., Van Horne, M. P., Grassotti, C. & Hoffman, R. N. (2006), 'Extending the predictability of hydrometeorological flood events using radar rainfall nowcasting', *J. Hydrometeo.* **7**, 660–677.

Vivoni, E. R., Ivanov, V. Y., Bras, R. L. & Entekhabi, D. (2004), 'Generation of triangulated irregual networks based on hydrological similarity', *J. Hydrol. Eng.* **9**, 288–303.

Vivoni, E. R., Ivanov, V. Y., Bras, R. L. & Entekhabi, D. (2005), 'On the effects of triangulated terrain resolution on distributed hydrologic model response', *Hydrol. Process.* **19**, 2101–2122.

Vrugt, J. A., Gupta, H. V., O Nuallain, B. & Bouten, W. (2005), 'Real-time data assimilation for operational ensemble streamflow forecasting', *J. Hydrometeo.* **7**, 548–565.

Welles, E., Sorooshian, S., Carter, G. & Olsen, B. (2007), 'Hydrologic verification. A call for action and collaboration', *Bull. Amer. Meteor. Soc.* pp. 503–511.

Weyman, D. R. (1970), 'Throughflow on hillslopes and its relation to the stream hydrograph', *Hydrol. Sci. Bull.* **15**, 25–33.

Wilks, D. S. (2001), 'A skill score based on economic value for probability forecasts', *Meteorol. Appl.* **8**, 209–219.

Wilks, D. S. (2004), 'The Minimum Spanning Tree (MST) histogram as a verification tool for multidimensional ensemble forecasts', *Mon. Wea. Rev.* **132**, 1329–1340.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd edn, Academic Press. 627 pp.

Wilson, J. L., Burrows, W. R. & Lanzinger, A. (1999), 'A strategy for werification of weather element forecasts from an Ensemble Prediction System', *Mon. Wea. Rev.* **127**, 956–970.

ACKNOWLEDGMENTS

I want to thank the referees Prof. P. Furcolo and Prof. V. Venugopal for their suggestions that helped me to improve the quality of the thesis.