Extreme rainfall data analysis for urban flood forecasting

Emanuele Artù Cassin

Master's degree in Environmental and Land engineering,
Climate Change track

Promotors:
Prof. Dr. Ir. Patrick Willems
Prof. Dr. Ir. Ricardo Reinoso-Rondinel
Prof. Dr. Ing. Stefania Tamea

Co-Promotor:
Ir. Daan Buekenhout

A.y. 2022-23

FOREWORD

This master thesis was written during my Erasmus year at KU Leuven in Belgium. I want to thank Prof. Dr. Ir Patrick Willems for giving me the opportunity to conduct research on such a fascinating topic, because I have always been interested in hydrology and climate change.

Most of the areas addressed in this master thesis were unfamiliar to me, therefore I was able to learn new abilities, particularly in data analysis and Python programming. I'd like to thank Ir. Daan Buekenhout and Prof. Dr. Ir Ricardo Reinoso-Rondinel for following me throughout the project.

I would like to thank Prof. Dr. Ing. Stefania Tamea for supervising this thesis work from Italy

The analysis and findings of this work were carried out in collaboration with another KU Leuven student, Frédéric Veithen. I'd like to express my gratitude to him because he was an amazing coworker.

CONTENTS

ABSTRACT

Extreme weather events have sadly escalated in recent years as a result of climate change. Every year, extreme precipitation causes devastating damage to communities and ecosystems, as well as loss of lives. As a result, new engineering solutions are required to forecast and mitigate the most severe effects of such disasters. The Short-Term Ensemble Prediction System (STEPS) is one of the most important nowcasting techniques to model the precipitation fields in short time ranges and can be employed and integrated in flood-control applications. One of this methodology's strengths is that it can generate both deterministic and probabilistic forecasts, allowing for an evaluation of the uncertainty associated with high precipitation peaks and hydrological balances.

This thesis has two central objectives. The first is to investigate the possibilities of py-STEPS, a Python library based on the STEPS technique for nowcasting applications. As open-source software, the functions of py-STEPS enable a wide range of applications, including motion field calculations, decomposition of precipitation structures, and the generation of both deterministic and probabilistic nowcasts. The second goal is to compare the probabilistic forecasts provided by py-STEPS and STEPS-BE, the STEPS implementation developed at Belgium's Royal Meteorological Institute (RMI). It is indeed interesting to see which STEPS-based solution conforms better to observations and whether a modular and open-source library like py-STEPS can outperform a private product designed specifically for Belgium.

Since the STEPS approach is based on radar image extrapolation, one of the main issues of the nowcasts relates to the uncertainty of observations. As a result, before forecasting the evolution of the precipitation field, it is required to validate the quality of the available radar images with additional devices such as rain gauges. In this study, the validation revealed that there is a general time shift of 10 minutes between the precipitation peak observed in situ and the one measured by the radar, and that locations 50-60 kilometers away from the closest radar are affected by smallest errors.

Three models were employed in the deterministic nowcast analysis: a simple extrapolation, S-PROG, and ANVIL. While the first approach merely translates the precipitation field detected in the most recent radar images without affecting the intensity of the precipitation, the other two can represent the growth and decay of precipitation structures within the moving cloud. An analysis of radar image subdomains that considered events classified as convective and stratiform revealed that there is no single model that can be used for all possible situations: ANVIL performed better in reproducing convective events, while S-PROG performed better when stratiform clouds were analyzed. The deterministic analysis chapter additionally focuses on an algorithm built specifically for this master thesis that is based on the parametrization of power spectra to distinguish between stratiform and convective events.

Finally, the comparison of the probabilistic nowcasts produced by STEPS-BE and py-STEPS demonstrated the latter's superiority. In almost all the cases examined, the ensemble of py-STEPS demonstrated to be better calibrated (through reliability diagrams and rank histograms) and suited as a warning system (by ROC curves). Furthermore, py-STEPS performed significantly better at *intercepting cumulative rainfall volumes throughout the events, which is crucial for flood-control systems.*

LIST OF FIGURES

LIST OF TABLES

Chapter 1: PROBLEMS RELATED TO EXTREME PRECIPITATION

As stated in one of the IPCC reports [1], it is difficult to elaborate one unique and general definition of extreme rainfall event, since in scientific literature, depending on the context and cultural background, many approaches have been used to describe and quantify such phenomenon. In general, two main methods [1] can be established to identify a precipitation extreme: (a) the use of percentiles (typically the 95th), (b) the use of absolute thresholds, defined country by country (e.g., in the United States is 50.8 mm/day). However, in order to clarify this concept in a more approximate but intuitive way, it can be stated that "heavy precipitation refers to instances during which the amount of rain or snow experienced in a location substantially exceeds what is normal" [2].

Starting from the 1950s, many regions of the world have experienced an increase of extreme precipitation events [3], with devastating consequences. In fact, heavy and intense rainfalls are associated with problems like flooding, soil erosion, crop damage, landslides triggering, water pollution, damages to ecosystems and resulting loss of lives. Therefore, finding new methods to predict and manage these kinds of events is fundamental for the future, especially considering that, on the whole, an increase of extreme rainfall events is expected around the world [4], [5]. Indeed, in a warming planet, the amount of water vapor in the atmosphere available for fostering heavy precipitations increases at a rate of 6-7 % in saturation concentration per degree rise in temperature, in compliance with the Clausius-Clapeyron relationship [4].

Nevertheless, as stated in the IPCC report of 2021 [5], the Clausius Clapeyron effect is not sufficient to give an exhaustive description of the changing patterns of extreme rainfall events on a regional scale, since both thermodynamic and dynamic processes should be accounted for. Consequently, nowadays it is still difficult to find a global and complete description of causes that trigger heavy precipitation events, since "the effect of warming-induced changes in dynamic drivers on extreme precipitation are more complicated, difficult to quantify, and are an uncertain aspect of projections" [5].

Despite this, in the report it is highlighted that the AR5 (fifth IPCC assessment report) predicts that over mid-latitudes and wet tropics the number of extreme precipitation events is *very likely* to increase due to global warming. These results are corroborated also by CMIP5 and CIMP6 simulations [5]. With particular reference to Europe, it is expected that "extreme precipitation will *likely increase* at global warming levels of 2° C and below, but *very likely* increase for higher warming levels for the region as a whole" [5]. Furthermore, other studies [6] suggest that also one single degree of warming, with respect to pre-industrial levels is enough to increase the frequency of extreme precipitation events over Europe. Considering that current global policies and energy investments are insufficient to meet the requirements of the Paris Agreement [7], the above-mentioned scenarios are likely to occur and therefore it is fundamental to elaborate proper prediction and adaptation strategies for this kind of events. Nevertheless, in the IPCC report [5] alternative scenarios associated to different global warming levels are considered and the results obtained by the CMIP6 project are shown in the figure below:

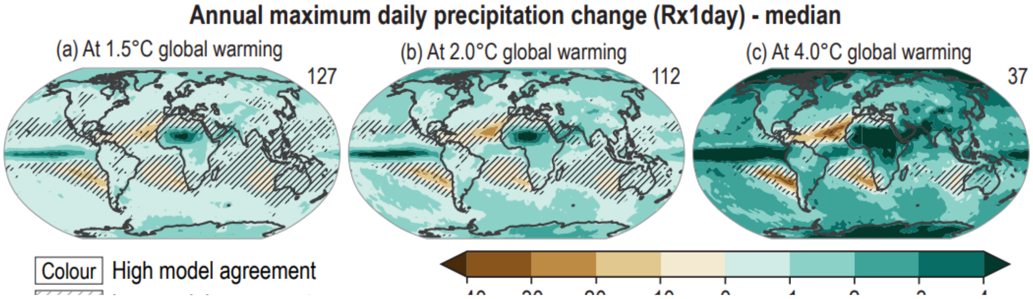


Figure 1.1. Results taken from the simulations of the Coupled Model Intercomparison Project Phase 6 (CMIP6), which were run using the Shared Socio-economic Pathway (SSP). On the upper-right part of each map it is indicated the number of simulations employed for the scenario. Diagonal lines are depicted on those regions where the modes agreement is <80%, while no signs are present in those regions where the model agreement is >80% [5].

It is easy to notice that extremes in precipitation are going to increase in all the future scenarios almost in all the regions of the world, hence there is no doubt that mitigation measures will be needed in the next decades. Eventually, it is worthy to underline that the results shown in the figure above are not valid for changes in sub-daily extremes like hourly rainfall [8], since usually these kinds of data "have only sporadic spatial coverage and are of limited length" [8] , making difficult to elaborate conclusions with a sufficient level of certainty.

Therefore, we have great uncertainties about the future development of extreme precipitation events, and this should represent another incentive to enhance methods for short-term prediction, like nowcasting techniques.

This thesis focuses on the Short-Term Ensemble Prediction System (STEPS), one of the most well-known and widely utilized nowcasting approaches for predicting rainfall extremes. Two implementations of this method will be evaluated in particular: py-STEPS and STEPS-BE. The first is an open-source Python library that permits control of all the phases and parameters included in the STEPS nowcasting technique, while the second is a private software developed at the Belgian RMI. In the following chapters, not only will some of the capabilities of py-STEPS, such as the generation of deterministic and probabilistic forecasts with a 2h lead time, be explored, but a comparison with STEPS-BE will be performed to see which is better suited for implementation in flood control systems. To accomplish these two main objectives, the thesis is structured as follows.

The second chapter contains a literature review that describes in detail all the principles underlying the STEPS method, as well as a description of py-STEPS and its main modules to highlight its similarities and differences to STEPS-BE. Moreover, the main metrics commonly used in literature and in this work to validate and compare the models are defined and discussed.

The radar images used to perform the nowcasts are examined in chapter three because STEPS, as an extrapolation-based method, is sensitive to errors in radar inputs. Hence, it is important to assess their quality to be aware of potential future radar corrections and the reliability of the model's outputs. This data validation will be performed comparing the observations of the radar network and rain gauges (considered the "ground truth") relative to a set of events extracted using an algorithm developed in Python.

Three alternative deterministic models generated using py-STEPS will be discussed and compared in chapter four. These models (simple extrapolation, S-PROG, and ANVIL) present structural differences, and the goal is to determine which is better suited to represent convective and/or stratiform precipitation. The comparison will be conducted using some of the metrics introduced in the literature review, considering both the complete radar domain and sub-areas of the radar images, to simulate a more realistic analysis inside a hydrological catchment. In addition, a section of this chapter is dedicated to an algorithm developed in python to categorize the events in convective and stratiform.

In chapter five, the same methodology used in chapter four will be used to compare probabilistic nowcasts produced by py-STEPS and STEPS-BE. This time, however, probabilistic metrics will be used to assess the skill and calibration of the ensembles of the two models. Furthermore, a section of the chapter will assess the benefits of using the ensemble mean rather than a more straightforward deterministic output.

Finally, chapter six will draw the conclusions of this work.

Chapter 2: LITERATURE REVIEW

The concept of meteorological nowcasting was firstly investigated in the ’70 [9] and at the beginning of the ’80. In a comprehensive book on the subject, Keith Browning [10] defined it as “the description of the current state of the weather in detail and the prediction of changes that can be expected on a timescale of few hours”. In 2010, the World Meteorological Organization (WMO) refined the former definition specifying that nowcasting should be characterized by a high local detail and should make forecasts over a period from the present to six hours ahead [10].

Despite the deep interest for this subject, the technological improvement in the ’70 / ’80 did not provide neither a sufficient time and spatial resolution nor advanced computer algorithms to properly describe the instantaneous state of the atmosphere and implement a real nowcasting system. Therefore, the first applications of this field were mainly limited to extrapolation of meteorological radar data for short term prediction of relatively simple atmospheric phenomena, like thunderstorm motions [9].

However, the development of communication technologies, weather data availability and data assimilation techniques opened to the implementation of increasingly advanced nowcasting systems and methods [9]. Indeed, smartphones and internet networks enable a fast and easy communication of weather information and complex graphics also to a non-expertise public, new surface and remote sensing data, communicated in real time, guarantee an accurate description of the current state of the atmosphere, while more sophisticated algorithms improve the synergic link between increasingly accurate observations and advanced models.

Since the main objective of nowcasting is to make prediction for short time periods over regional and/or local scales, the best results are obtained when instrumentation that guarantees high temporal resolution and spatial coverage is used. The best example is represented by radars, which, as previously mentioned, have been successfully employed since the very beginning of nowcasting history and still today are considered the main observational device for short time forecast. In fact, it is no coincidence that the most sophisticated nowcasting techniques are implemented in those developed countries where radar systems are more mature and robust [10].

However, the simultaneous employment of more observations (obviously referenced in the same grid system) coming from instruments like radar, satellites, lightning networks, surface stations, wind profilers and radiosondes should represent the ideal input dataset for a nowcasting system [10]. Once the optimal observational configuration for describing the real-time state of the atmosphere has been established, it comes the time to decide which nowcasting techniques suits better the scope of the meteorological prediction.

Although many nowcasting procedures have been proposed throughout the years, it is possible to subdivide them in three main categories, keeping in mind that the constant and rapid research behind this field could suddenly lead to new solutions and proposals.

The three basic groups of methods are [11] : Extrapolation Methods, Blending Methods, Artificial Intelligence-Based Methods. The accurate and extensive description of each method lies outside the scopes of this work, therefore in this section only a general summary will be given, based on the paper of Sokol et al[11]. Nevertheless, a much more exhaustive insight on a specific extrapolation/blending technique will be given in the successive paragraph.

•**Extrapolation methods** [11]

As previously mentioned, the extrapolation technique was used in the first nowcasting applications and still today represents the main baseline of advanced short-term prediction methods. The great advantage of extrapolation methods lies in their relatively simple application, especially considering that:

- The basic data can comprise only radar reflectivity. Obviously, an integrated system of different sources of observation would be preferable, but it is not essential.
- The computational cost is pretty low and therefore the calculations are pretty fast. This allows to make multiple forecasting in different short time steps.
- The use of currently measured data guarantees that the forecast is accurate for very short lead times (usually in the order of tens of minutes).

However, purely extrapolation-based methods do not take into account any physical property of the atmosphere in their modelling and are usually based on the extrapolation of radar-derived precipitation cells along Lagrangian trajectories, determined after the calculation of the Motion Field. Therefore, these kinds of methods become extremely inaccurate after leading times exceeding 10–20 minutes, because they rely on the oversimplified hypothesis that precipitation intensity does not change along the trajectories. This assumption is well applicable only in case of purely stratiform precipitations, while convective precipitations are usually characterized by significant intensity changes along the Lagrangian trajectories.

Consequently, modern extrapolation-based methods have implemented spatial stochastic models to forecast some of the unpredictable patterns of precipitation fields. A great variety of these last models has been developed, depending on the chosen approach for the determination of the Motion Field and the probabilistic techniques employed. However, it can be said that each stochastic model has the objective of filtering out/ adding small scale features to the precipitation cell, to predict in a more or less accurate manner the evolution of the forecasted event.

•**BLENDING METHODS** [11]

An alternative to the implementation of stochastic methods in extrapolation-based models is represented by the possibility of blending the results of the forecast using the outputs of numerical models based on physical atmospheric mechanisms (NWP models).

The main idea behind this technique is to combine the benefits of extrapolation methods, which are really stick to observations, and NWP models, which have the ability to make the forecast results more reliable for lead times higher than 10 minutes, thanks to their physic-based approach. Without providing complex details, it could be stated that Blending Models outputs are determined by a weighting of the two forecasts (namely, the forecast by extrapolation and the forecast by NWP model). The weighting system relies on the lead time of the forecast: the shorter the lead time, the higher the importance of the forecast by extrapolation and vice versa. Nonetheless, it is important to underline that blending methods are not always an alternative to stochastic assimilation in extrapolation-based methods: there are several examples of nowcasting techniques that implement both probabilistic and blending components in their extrapolation baseline.

- **ARTIFICIAL INTELLIGENCE-BASED METHODS** [11]

New nowcasting approaches based on machine learning (ML) are becoming increasingly widespread in the last years. These innovative methods greatly differ from the previously described techniques, as their estimations are usually independent of physical equations, and they frequently neglect the explicit statistical description of the problem. Due to this lack of physical meaning, usually ML methods are called “black box models”.

In order to structure these models, historical data of precipitations, derived from radar observations or in-situ weather stations, are used and consequently the algorithm is able to exploit the video sequence of input images to predict future rainfall fields. Therefore, these kinds of models have a self-learning ability to construct the patterns that lead to the conversion of inputs into the most logical output. Nevertheless, it can be noted that the principle exploited to obtain the outputs of ML models is similar to the one of extrapolation-based methods, since in both cases the used input is a sequence of radar images (even though, as previously mentioned, machine-learning models can use other sources of data).

In conclusion, it can be said that nowcasting methods experienced a stunning development in recent years and new technologies permitted not only a technical improvement related to the quality of observations and accuracy of model outputs, but also a more and more efficient involvement of the public. This last aspect is crucial considering that nowcasting is responsible for early detection of potentially dangerous precipitation events, which could have catastrophic repercussions on safety, health, and environment. Considering that in the future such extreme events are projected to increase almost everywhere, the implementation of advanced nowcasting systems becomes an urgent matter all around the world. Indeed, scientific literature reports many examples of more or less successful integration of nowcasting systems both in already developed [12] and developing [13] countries. Among all the possible nowcasting methods, this work focuses on STEPS (Short-term Ensemble Prediction System), which could be categorized as a hybrid form of extrapolation-based and blending models.

2.1 STEPS

STEPS is a probabilistic nowcasting method that exploits both the extrapolation and the NWP blending techniques. It has been developed by the joint work of the UK Meteo office and the Australian Bureau of meteorology to provide reliable nowcasts for lead times that extend till 6 hours ahead [14]. As previously mentioned, purely extrapolation-based models can be inaccurate for forecasting beyond 20 minutes, since they do not take into account all the phenomena of growth and decay inside the moving precipitation field, but they only make corrections to the estimation of the motion field [14]. Furthermore, another deficiency of these models is that they produce one single forecasting for each lead time, thus making impossible to estimate probabilistic metrics throughout the nowcast.

On the other hand, STEPS not only adopts an autoregressive approach to model the formation and dissociation of precipitation cells, but, thanks to a noise cascade, is also able to create an ensemble of forecasts for each lead time [14]. In addition to that, STEPS has also implemented a technique to estimate and correct the errors of the motion field, that will be explained later. Another interesting peculiarity of this method is the blending with NWP models, that allows to increase the skill of the forecast and add a “physics-based” factor that increases the accuracy of the outputs. Indeed, as explained by Seed et al. [15], STEPS can be interpreted as a merging of three different cascades: one for the extrapolation, one for the noise and another for the NWP models. Although this last point represents one of the key stones of STEPS, it is not always easy to directly implement it, for the following reasons:

1) It is not always possible to have access to the outputs of an NWP model.
2) Usually, NWP models require high computational efficiency and therefore it is difficult to obtain results in short times. This can be an issue, especially when nowcasting for short lead times (15-30 minutes) is needed.
3) NWP models have a coarse resolution when compared to the necessities of hydrological applications [16]. For example, the grid used in the NWP model by UK- Meteo Office has a resolution of 12 Km [14], but considering that usually these models cannot solve features on a scale of less than 5 times the grid length [14], the effective resolution is approximately 60 Km.
4) Even though different authors reported that blended nowcasts offer higher skill in comparison with the application of the STEPS method alone, usually the most evident benefits are obtained after 2 hours lead time [15], [17]. As a matter of fact, as shown in the figure below [18], radar-based nowcast methods can offer higher skill than NWP models for the first 2 hours of lead time:

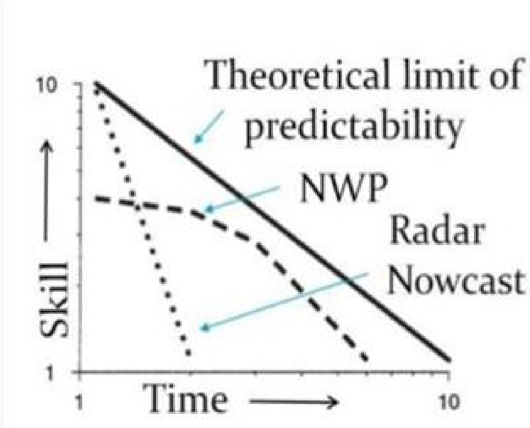


Figure 2.1. Qualitative skill comparison between extrapolation and NWP forecasting

For these motivations, not in all the implementations of STEPS the blending with NWP models is applied (like in the case of STEPS-BE, that will be explained in next sections).

Therefore, after briefly introducing the main aspects of STEPS, a detailed scheme of all the fundamental steps of the method is given below, to describe in a concise manner how all the above-described processes are carried out in practice. However, before doing this it is necessary to clarify the terminology that will be used. “STEPS” refers to the overall nowcasting method, which starts with the acquisition of radar images and ends with the production of an ensemble of forecasts; therefore, its description comprises all the 9 points reported in the scheme below. “S-PROG” is the model that carries out the operations of extrapolation of the rainfall field, thanks to an optical flow algorithm, decomposition in sub-structures, by means of a fast Fourier transform and a multiplicative cascade, and temporal evolution of the precipitation cells, through an autoregressive filter. Namely, S-PROG is the core of the STEPS method, since it models the properties of “dynamic” and “spatial scaling” of the precipitation field; in the scheme below, its description comprises points 1-3-5. Yet, the S-PROG scheme does not include the application of the stochastic noise and thus the production of an ensemble.

Since this thesis focuses on forecasting predictions for lead times of at maximum 2 hours, the blending with NWP models is not taken into account and it is not then included in the conceptual scheme reported below.

2.1.1 Estimation of the velocity field using optical flow on the last two radar rainfall images (S-PROG model)

Similarly to other extrapolation-based models, STEPS needs the estimation of the motion field of the precipitation that is therefore propagated by means of an advection scheme [14]. In the proposed model, the motion field is estimated through an optical flow algorithm, which exploits the last two observed radar images to calculate the components of the motion field from the following equation [19]:

$$\frac{dR}{dt} = \frac{\partial R}{\partial t} + u\frac{\partial R}{\partial x} + v\,\frac{\partial R}{\partial y}\ , u = \frac{dx}{dt}\ , v = \frac{dy}{dt}\ (2.1)$$

Where R represents the precipitation field and u and v are the x and y components of the motion field. It is interesting to notice that the solution of the previous equation is obtained under the hypothesis that $\frac{dR}{dt} = 0$, namely that the intensity of the precipitation field does not change in time, which is typical of traditional extrapolation-based models. In fact, in STEPS method the growth and decay of precipitation parcels is not represented through the calculation of the motion field but thanks to the decomposition of the rainfall field in different substructures, as it will be explained in the next points.

2.1.2 Data transformation

Usually, a logarithmic transformation is applied to the precipitation data, exploiting the relationship Z-R that correlates radar reflectivity (Z) and estimated surface rain rate (R) [14]:

$$Z = 10\ log_{10}(200R^{1.6})(2.1)$$

There are two main reasons why this transformation is necessary [19]. Firstly, thanks to the formula above the data are forced to follow a log-normal distribution: this allows to use established stochastic models that assume Gaussianity. Secondly, log-transformed rainfall fields simplify the decomposition of the precipitation structures when applying the fast Fourier transform, since the multiplicative cascade can be replaced by summation in the transformed space [14].

2.1.3 Decomposition of the last two observed rainfall fields into a multiplicative cascade (S-PROG model)

It has been demonstrated [14] that precipitation fields exhibit spatial characteristics that can be described by multi-fractal models. This property, known as "spatial scaling" of the precipitation field, basically assumes that similar features exist along different spatial scales, which range from very large (> 10^4 m) to small ones (< 10 m) [14].

Another relevant observation is that usually the temporal persistence of these features is broadly proportional to their size [19]. This other property, known as "dynamic scaling" of the precipitation, suggests that bigger structures of the field are easier to predict, since they take a large time to form and precipitate, while small scale features are harder to model, since their "lifetime" is relatively low [20]. These two characteristics of the precipitation field are the basis behind the development of the Spectral Prognosis (S-PROG) model [14], that represents the main core of the STEPS methodology.

Basically, the S-PROG model decomposes the precipitation fields into a multiplicative cascade by means of a fast Fourier transform (FFT), to detect and isolate the different precipitation structures. If the data transformation described in the previous point is applied, the multiplication is replaced by summation in the transformed space. In a second moment, once the process has been completed, an inverse FFT can be applied to skip from the frequency domain to a spatial one, in order to re-obtain a decomposed rainfall field. It is important to understand that the multiplicative cascade of a precipitation field does not imply the formation of distinct and independent cells of different extent, but "rather a hierarchy of precipitation structures embedded in each other over a continuum of scales" [20].

After this, the S-PROG model applies an autoregressive model to estimate the evolution of the different precipitation structure, as explained in next sub-sections.

The number of spatial scales can be calculated in the transformed space using the radial Fourier wavenumber $|k| = \sqrt{k_x^2 + k_y^2}$ through the following formula [19]

$$scale = \frac{L\Delta x}{2\ |k|} (2.2)$$

Where L is the number of rows/columns of the radar grid and Δx is the resolution. Then, for every spatial scale, the decomposition of the precipitation field can be summarized as follows [14]:

$$dBR_{i,j}(t) = \sum_{k=1}^{N} x_{k,i,j}\ (t), i = 1,..,L\ ; j = 1,..,L\ (2.3)$$

Where $dBR_{i,j}(t)$ denotes the precipitation intensity in dBR at pixel (i,j) at time t and the term $x_{k,i,j}$ indicates the precipitation intensity in dBR at pixel (i,j) for the cascade level k.

As an example, the figure below reports a decomposition of a precipitation field in seven cascade levels observed by the Finnish Meteorological Institute:

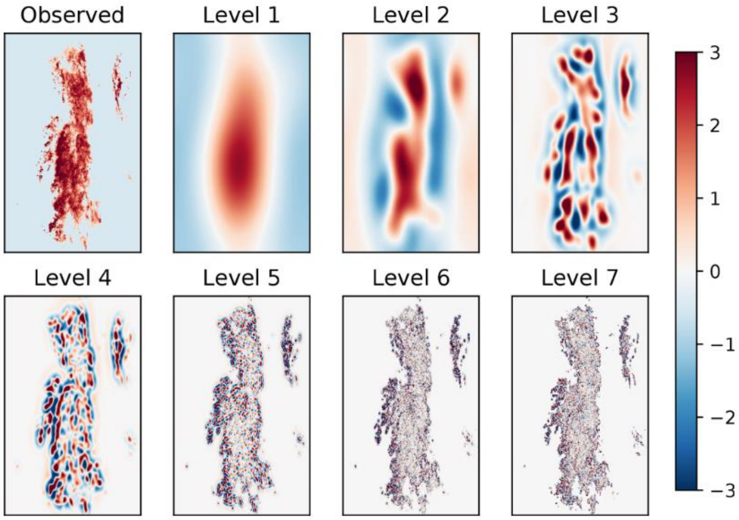


Figure 2.2. Decomposition of a precipitation field in seven cascade levels[19]

2.1.4 Generation of a cascade of spatially correlated stochastic noise

This is a fundamental step of the method, since the production of a stochastic noise is used to reproduce the phenomena of growth and decay of the precipitation field and this is what distinguishes STEPS from other nowcasting methods. The produced noise must follow a spatial correlation that is compliant with the one of the observed rainfall field detected by the radars. For this reason, usually three passages are implemented to obtain the desired perturbation [20]:

- Computing the FFT of a white noise field
- Applying a filter to the obtained components in frequency domain
- Computing an inverse FFT to switch again to the spatial domain.

The filter needed in point 2 is usually acquired by applying a 1-D or 2-D power spectrum to the most recent radar images, in order to get a spatial correlation of the noise that is similar to the precipitation field. If, during the forecasting, it is not possible to retrieve a recent radar image, the filter can be parametrized using a climatological power law [20].

2.1.5 Perturbation of the original rainfall cascade

Once the noise cascade has been obtained, it can be applied to the multiplicative cascade of the rainfall field obtained in point 3. Therefore, the perturbation is applied level by level according to the following equation, which is generalized for an autoregression of order p [19]:

$$R_j(x,y,t) = \sum_{k=1}^{p} \varphi_{j,k}\, R_j(x,y,t-k\Delta t) + \varphi_{j,0}\varepsilon_j(x,y,t) \; (2.4)$$

The symbols j and k represent the cascade level and the level of the autoregression respectively. R is the intensity of the rainfall field, $\varphi_{j,k}$ are the parameters of the autoregressive model for each structure of the precipitation estimated through the Yule-Walker equations, $\varphi_{j,0}$ is the parameter of the autoregression applied to the noise cascade, ε_j is the noise estimated for each cascade level and Δt denotes the time difference between different estimations.

By further analyzing the equation above, it can be noted that the first term corresponds to the Lagrangian (i.e. deterministic) component of the forecasting, which derives from the extrapolation process, while the second terms represents the stochastic component, that stems from the noise cascade described in the previous point. Therefore, the combination of the multiplicative cascade and the autoregressive model, together with the stochastic noise, allows the forecaster to control the temporal and spatial structure of the precipitation, considering at the same time the contribution of fast growth and decay of precipitation cells, that are usually the biggest responsible for nowcasting errors [14].

Even though the equation above is generalized for an autoregressive order p, an AR(2) model is usually employed. The Yule-Walker equations exploit the time-lagged auto-correlation functions $r(t)$ to retrieve the values of the coefficients of the autoregression, through the following system [21](in case of a second order):

$$\begin{bmatrix} r(1) \\ r(2) \end{bmatrix} = \begin{bmatrix} r(0) & r(1) \\ r(1) & r(0) \end{bmatrix} \begin{bmatrix} \varphi_{j,1} \\ \varphi_{j,2} \end{bmatrix} (2.5)$$

Where the equations $r(t)$ can be calculated from the observations and the extrapolations.

Finally, the term $\varphi_{j,0}$ is chosen in such a way that the AR(2) process is normalized to unit variance.

$$\varphi_{j,0} = \sqrt{1 - r(1)\varphi_{j,1} - r(2)\varphi_{j,2}} \quad (2.6)$$

2.1.6 Extrapolation of the cascade levels by means of an advection scheme

Another source of uncertainty in nowcasting applications stems from the propagation of the motion field for different lead times [19]. In fact, the velocity field is usually subjected to considerable noise [14].

For this reason, similarly to what has been explained in the points above, it is possible to apply a perturbation to the motion field calculated at the beginning of the forecasting (through the optical flow algorithm) to compensate the uncertainties derived from the application of the advection scheme. An indicative formula that explains this mechanism is [19]:

$$w_p(x,y) = w_0(x,y) + \alpha_{par}(t)\varepsilon_{par}(x,y)w_{par} + \alpha_{perp}(t)\varepsilon_{perp}(x,y)w_{perp} \quad (2.7)$$

Where w_{par} and w_{perp} denote the parallel and perpendicular components of the motion field w_0, ε_{par} and ε_{perp} are two random variables taken from a Laplace distribution with 0 mean and unit variance and the parameters α are used to scale the perturbations. These last parameters are usually obtained by analyzing a large sample of advection fields and are therefore tuned to get the best results for the forecasting [19]

2.1.7 Application of the autoregressive model for each lead time of the forecasting to the different cascade levels and recomposition of the cascade into the spatial domain

Exploiting what has been written in points 6 and 7, it is possible to monitor the evolution of the precipitation field in space and time for each lead time of the forecasting. Since the application of the multiplicative cascade to both the rainfall field and the precipitation returns the different rainfall features in the domain of frequency, it is necessary to apply an inverse FFT to obtain results referenced in space.

2.1.8 Probability matching of the forecast results with the observed rainfall fields

During the different passages of the STEP procedure, it is possible that some discrepancies with respect to the observed rainfall field are introduced. For this reason, it necessary to apply a post-processing of the data to match the statistical properties of the nowcasting outputs with the ones of the observed radar images (this point will be further discussed in next sections).

2.1.9 Forecast production

At the end of the process, if all the 8 point described above have been applied, it is possible to obtain a probabilistic nowcast, which comprises an ensemble of rainfall fields calculated for the desired lead time. Alternatively, it is also possible to apply only the S-PROG model and neglect the procedure of stochastic perturbation (namely, point 4 and 6), in order to retrieve a deterministic nowcast. Obviously, in the latter case all the benefits related to an ensemble forecast are lost to the advantage of a quicker and simpler method.

2.2 PY-STEPS

In this section an explanation about the implementation of STEPS in an open-source and python- based library, py-STEPS, is provided. This section offers the following insights: description of py-STEPS, metrics used in the software to evaluate the forecast's skill and main observations derived from practical applications. In the first paragraph, all the information, tables and images are taken from the official paper of py-STEPS creators [19]; therefore, if not differently specified, all the main statements refer to the same source.

2.2.1 py-STEPS DESCRIPTION

py-STEPS is an open-source and python-based library, which offers a modular framework for implementation of the STEPS method. Since it is open source, all the modules, functions and features of this platform are available to all sorts of public, ranging from meteorologists to hydrologists. This not only gives the opportunity of utilizing a well-documented and reliable software for weather prediction, but also pushes a community-driven effort for the constant improvement of the method. As a matter of fact, the source code of py-STEPS is hosted on GitHub [22], and external developers have the possibility to apply changes/integrations to the main repository via pull request. Therefore, py-STEPS not only allows free copying and redistribution of the software, but also modifications (as long as these are tracked and made available under the BSD license). Furthermore, practical examples of this library applications are available on the official website [23], together with an explanatory guide of all the modules and functions implemented in the library [24].

py-STEPS is built upon several external libraries, among which the most important ones are Numpy, SciPy and Matplotlib, that all together provide a MATLAB-like computing environment in Python. In particular, the main data structure is based on Numpy. For the sake of completeness, the table below reports all the external libraries used in py-STEPS:

Library	Website
h5py	http://www.h5py.org
netCDF4	http://unidata.github.io/netcdf4-python
PIL	https://github.com/python-pillow/Pillow
OpenCV	http://opencv.org
NumPy	http://www.numpy.org
SciPy	http://www.scipy.org
FFTW/pyFFTW	http://www.fftw.org https://github.com/pyFFTW
dask	http://dask.org
cartopy	https://github.com/SciTools/cartopy
Matplotlib	http://matplotlib.org
mpl_toolkits.basemap	http://matplotlib.org/basemap

Table 2.1.Summarizing table of external dependencies of PYSTEPS [19]

Another key feature of py-STEPS is the modularity. Multiple functions, written to perform different procedures of the STEPS method, are subdivided into different modules, which can be used in a complete interchangeable manner. This means that the user has a complete freedom of combining different methods in different orders and controlling the algorithms implemented in the library, without being forced to follow a rigid workflow.

py-STEPS implements several optical flow methods (such as Lukas-Kanade, VET, DARTS), different procedures for filtering white noise (parametric, non-parametric), as well as multiple post-processing approaches (which are based on either masking or matching the statistics of the forecast fields with the most recent radar observations). Therefore, the user can test, case by case, all the options offered by the library and make a comparison among the different obtained results, highlighting which are the main outcomes of different algorithms choice. Moreover, the forecaster can also decide to perform either a deterministic nowcast, where the application of the stochastic noise cascade to the different spatial fields is skipped, or a probabilistic forecast, which encompasses all the stages described in the previous section. Obviously, in the first case only one result will be available at the end of the forecast, while in the second one an ensemble of precipitation fields can be obtained. It is also possible to have control on the number of the ensemble members of the nowcast, basing on the requirements of the application. The figure below shows the general workflow of py-STEPS.

As it can be noticed, since py-STEPS is an implementation of the main STEPS procedure, the conceptual sequence of the different passages resembles the one described in the previous section. Inside each chart element there are two rows: the first one describes the task carried out, while the second reports the name of the py-STEPS module used. White colors indicate the operations that are performed in all nowcast methods. Green colors represent the operations that are performed when the S-PROG model is applied, namely the cascade decomposition and the application of an autoregressive filter to each cascade level. Blue colors refer to the passages that are carried out only when a probabilistic forecast is performed (i.e., they are neglected in case of a deterministic approach).

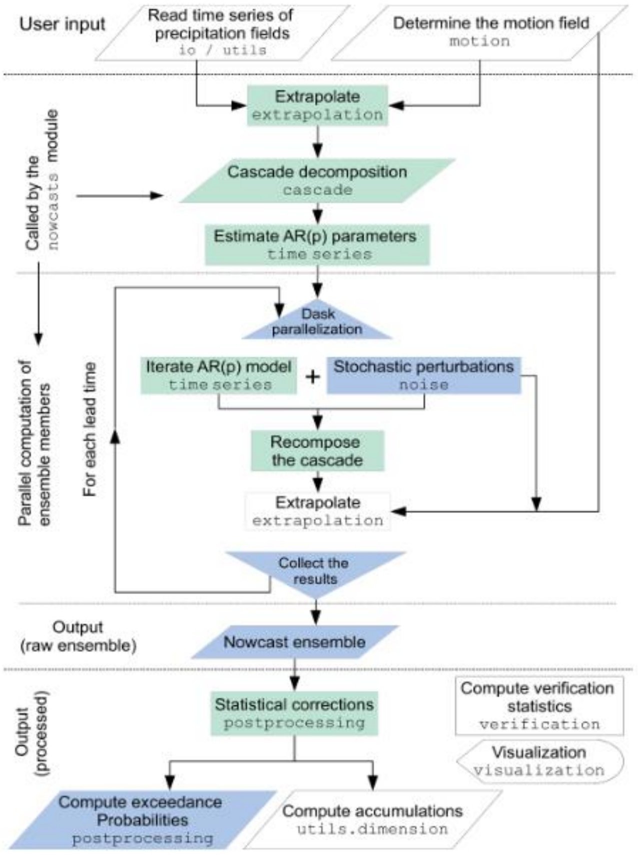


Figure 2.3. Work flow followed in py-STEPS[19]

2.2.2 Optical flow algorithms

As previously discussed, the nowcasting procedure always starts with the calculation of a motion field by means of an optical flow algorithm. In py-STEPS currently 3 methods are implemented and shortly discussed here:

- **Lucas-Kanade algorithm:** it is an easy method for estimating the mobility of interesting features in successive images of a scene. The objective is to assign a velocity vector (u, v) to each pixel in the scene by comparing two consecutive images. Considering one single pixel, the equation to solve has the form [25]:

$$I_x(x,y) * u + I_y(x,y) * v = -I_t(x,y) \ (2.8)$$

Where I_x, I_y and I_t are the derivatives of the image function $I(x,y)$ with respect to x, y and t (time), while u, v denote the x and y components of the velocity vector.

Usually, the equation above is not solved for one single pixel, but for a neighborhood of pixels which are supposed to have the same velocity, thus forming the system:

$$S * \binom{u}{v} = T \ (2.9)$$

Where S is a (n x 2) matrix containing the rows $[I_x(x_1,y_1), I_y(x_1,y_1) \ \dots \ I_x(x_n,y_n), I_y(x_n,y_n)]$ and T is a vector containing the terms $[\,I_t(x_1,y_1), \dots, \ I_y(x_n,y_n)]$. Usually, the system above is over-determined, but can be solved using the least square method:

$$S^T * S * \binom{u}{v} = S^T * T \ (2.10)$$

- **Variational echo tracking method (VET):** technique used for calculating the velocity field of radar reflectivity echoes. The entire radar map is divided into sub-regions, where a velocity vector $V_{m,n}$ is calculated by minimizing the difference in radar reflectivity between two composite maps $Z(x,y,t)$ and $Z(x,y,t+\Delta t)$ separated by a time interval Δt. Therefore, to construct the whole velocity field V, each $V_{m,n}$ should be calculated inside every sub region. However, in variational methods it is possible to compute simultaneously all the vectors by a global minimization of the following equation over the entire composite map [26]:

$$F(V) = F_Z + F_V \ (2.11)$$

Where $F(V)$ is the cost function to be minimized, completely dependent on the velocity field V and F_Z, F_V are two constraints.

The first constraint F_Z imposes the conservation of reflectivity and is defined as the sum of the square of the residuals at all locations (x, y) that arise between the map at time t and at time Δt according to the vector field V [26]:

$$F_Z = w_Z \sum_x \sum_y \{Z(x, y, t_0) - Z(x - u\Delta t, y - v\Delta t, t_0 - \Delta t)\}^2 \, dxdy \quad (2.12)$$

Where w_Z is a weight related to data quality (usually set constant) and u, v are the x, y component of the velocity vectors interpolated at every location (x, y) with respect to the $V_{m,n}$ values in every sub-area.

The second element of equation 2.11 is the so called "smoothness penalty function" and is used to limit the variability of the velocity vectors $V_{m,n}$ [26]:

$$F_V = w_V \sum_x \sum_y \{(\frac{d^2u}{dx^2})^2 + (\frac{d^2u}{dy^2})^2 + 2(\frac{d^2u}{dxdy})^2 + (\frac{d^2v}{dx^2})^2 + (\frac{d^2v}{dy^2})^2 + 2(\frac{d^2v}{dxdy})^2\} dxdy \quad (2.13)$$

As a result, to minimize the function $F(V)$ several iterations are needed. In VET, a conjugate-gradient algorithm is used to determine the optimum search direction and the step length.

- **Dynamic and Adaptive Radar Tracking of Storms (DARTS):** this method exploits reflective atmospheric data (i.e., a time series of sequential radar images) and the following flow equation to retrieve the future and current atmospheric conditions [27]:

$$\frac{\partial}{\partial t} F(x, y, t) = -U(x, y) \frac{\partial}{\partial x} F(x, y, t) - V(x, y) \frac{\partial}{\partial y} F(x, y, t) + S(x, y, t) \quad (2.14)$$

Where $F(x, y, t)$ is the scalar field of radar observations, $U(x, y)$ and $V(x, y)$ are the x-axis and y-axis motion velocity over the spatial domain and $S(x, y, t)$ represents some dynamic mechanisms occurring in the scalar field (for example, growth and decay of precipitation structures).

To find a solution of equation 2.14, a Discrete Fourier Transform (DFT) is applied to obtain a linear system that can be solved in the frequency domain. Finally, an Inverse Fourier Transform can be applied to estimate the components $U(x, y)$ and $V(x, y)$.

2.2.3 Stochastic noise generators

As previously stated, the stochastic noise used in equation 2.4 is generated by filtering a white noise field in the frequency domain. Filtering mechanisms available in py-STEPS include:

- **Parametric method:** In this method, the white noise field ε_w is filtered in frequency domain exploiting the slope of the Radially Averaged Power Spectrum (RAPS) of the most recent radar image and then transformed back to the spatial domain by means of an IFFT [19]:

$$\varepsilon(x, y) = \mathrm{F}^{-1}\left\{f(|\boldsymbol{k}|)\mathrm{F}\{\varepsilon_w\}(k_x, k_y)\right\} \, (2.15)$$

Where F indicates the FFT (and F^{-1} the IFFT) and f represents the slope of the RAPS.

To be more precise, py-STEPS uses a piecewise linear function with two spectral slopes (β_1, β_2) and one breaking point to filter the white noise field. One problem of this method is that it cannot represent anisotropic structures, as this model assumes an isotropic power-law scaling relationship.

- **Non-parametric method:** This method, as the name implies, does not rely on the parametrization of the most recent radar image, but instead directly applies their power spectrum to the white noise field in frequency domain [19]:

$$\varepsilon(x, y) = \mathrm{F}^{-1}\left\{\left|\mathrm{F}\{\boldsymbol{R}\}(k_x, k_y)\right|\mathrm{F}\{\varepsilon_w\}(k_x, k_y)\right\} \, (2.16)$$

With $\boldsymbol{R}$ indicating the power spectrum of the observed precipitation field. This method can represent anisotropic structures, but it is very sensitive to the quality of the input radar images and requires a large sample size.

- **Local non-parametric method:** This method is quite similar to the previous one, but this time the noise is generated locally through a Short Space Fourier Transform (SSFT) to account for discontinuities in the covariance structure of the last observations [19]:

$$\varepsilon(x, y) = \mathrm{F}^{-1}\left\{\left|\mathrm{F}\{\boldsymbol{R} w_h(n_1, n_1)\}(k_x, k_y)\right|\mathrm{F}\{\varepsilon_w\}(k_x, k_y)\right\} \, (2.17)$$

Where $w_h(n_1, n_1) = w_h(n_1) * w_h(n_2)$ is the outer product of two Hanning windows with sizes n_1 and n_2.

2.2.4 Postprocessing

Some remarkable discrepancies between the modeled and the observed precipitation field may arise during the intermediate stages of the nowcasting procedure. Therefore, after the production of the final precipitation field, it is customary to apply some postprocessing techniques to match the statistics of observations and modeled precipitation field. This is accomplished in py-STEPS using two complementing approaches, which are briefly detailed in this section.

- **Masking:** This procedure is implemented to avoid the formation of stochastic precipitation in no-rain areas. Currently, in py-STEPS three masking methods are available. The first one generates the mask using a precipitation threshold, and the mask is then kept constant throughout the nowcast. The second method works similarly, but the mask is progressively relaxed and changes at each lead time to better simulate the stochastic evolution of the wet area. Finally, the mask can also be obtained through S-PROG, namely an unperturbed (deterministic) nowcast.

- **Matching the statistics of the forecasted field with the most recent radar images:** two methods can be used to reach this goal. The first consists of matching the conditional mean (determined by the previously applied mask) of the forecasted field with the conditional mean of the observed one. In the second method, the Cumulative Distribution Function (CDF) of the forecasted field can be mapped to the observed field through the following expression [19]:

$$R'(x,y) = F_{obs}^{-1}\left(F\big(R(x,y)\big)\right) \quad (2.18)$$

Where F_{obs}^{-1} and F denote the CDF of the observed field and the forecasted one R.

2.3 VERIFICATION METRICS

Different metrics have been used in literature to assess the forecast's skill. Therefore, in order to provide a proper context that helps to interpret the results and the main conclusions about PYSTEPS that will be provided in the next paragraph, it is presented here a short introduction to the principal metrics. Generally speaking, it is possible to identify two main kinds of forecast's metrics: qualitative and quantitative.

In this work, "qualitative" (or categorical) refers to all those metrics that represent the ability of the forecast to correctly predict an event in comparison to the observations. For instance, given a grid's pixel and a precipitation threshold, it is possible to say that the forecast has a good qualitative skill if it predicts that the threshold will be exceeded when this is observed also in radar image. On the other hand, it can be stated that the weather model has a poor skill if it is predicted that the threshold will not be exceeded, but the contrary is observed.

Conversely, quantitative (or continuous) metrics are those that express the bias of the forecast and therefore are a representation of the accuracy of the weather model. For example, once it has been verified that in a grid's pixel both the forecast and the radar exceed the precipitation threshold, it is of interest to know by how much the forecast's precipitation differs from the one observed by the radar.

Below it is provided a more in-detail description of some of the most important metrics used in forecasting, while a summarizing table containing more information is presented at the end of this paragraph:

2.3.1 Receiver Operating Characteristic (ROC) curves

The relative operating characteristics curves are plotted considering the False Alarm Ratio (FAR) against the Hit Rate (HR). These two metrics can be defined as follows[28]:

$$FAR(\%) = \frac{FP}{FP + TN} X\ 100\ (2.19)$$

$$HR(\%) = \frac{TP}{TP + FN} X\ 100\ (2.20)$$

Where TP means "true positive", FN "false negative", FP "false positive" and TN "true negative". Considering a grid and a precipitation threshold, a true positive is obtained when both the forecast prediction and the radar observation overcome the threshold inside a given pixel, while a false negative is registered when the forecast predicts a value lower than threshold, but the radar observes an exceedance. Analogously, a true negative is obtained when neither the forecast nor the observation exceeds the threshold, while a false positive verifies when only the prediction outcomes the critical value. Therefore, considering a value threshold and a set of probability thresholds a curve like the one below can be built [28]. The bisector functions as a discriminant: an ROC curve above it (as shown in the graph) represents that the forecast has more HR than FAR and therefore can be considered reliable, while a curve below it is the result of a forecast that performs worse than a random generator, since for each probability threshold more false alarms than effective hits are given [28]. As a consequence, also the area subtended by the curves can be used to interpret the forecast skill [29]: a value equal or lower to 0.5 represents a random generator, while areas close to 1 are the results of a very accurate forecast.

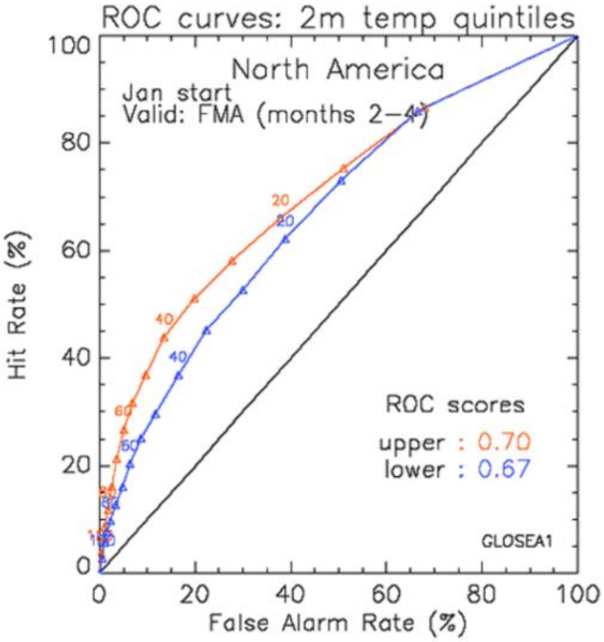


Figure 2.4. Example of an ROC curve [28]

2.3.2 Reliability diagrams

While ROC curves represent the forecast's ability to correctly predict the occurrence of an event, the reliability diagrams graphically show what is the actual chance of observing the forecasted event [30]. As shown in the picture below, this kind of diagram plots the forecast probability against the frequency of the observations. Consequently, the bisector represents the reliability diagram of a perfect weather model; for instance, if the model predicts that an event will occur with a given % of probability, this will be effectively observed exactly with the same % of probability all the times [30]. Therefore, the closer the reliability diagram approximates the bisector and the more reliable the forecast is.

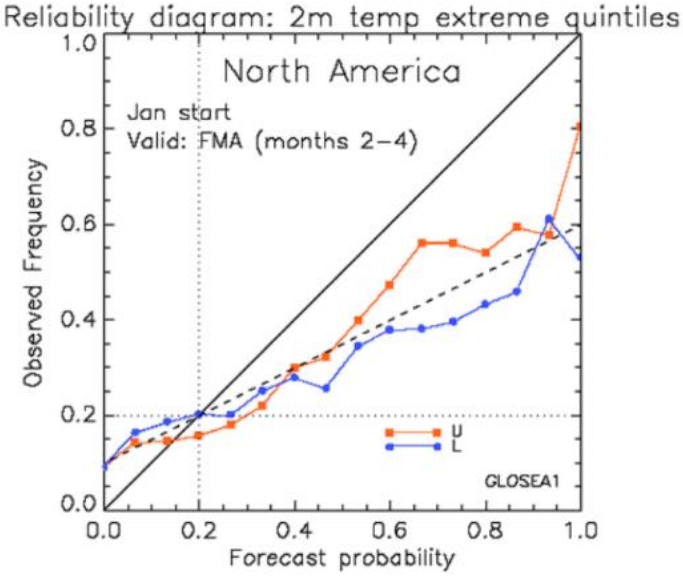


Figure 2.5. Example of a reliability diagram[30]

2.3.3 Rank Histograms

As suggested by the name, this metrics is a frequency histogram, where the class intervals defined by the forecast values are filled with observed data [31]. These histograms give a good qualitative assessment of the bias of the forecast: as a matter of fact, an accurate forecast will produce a rank histogram where all the bins have a similar frequency, while a histogram characterized by big frequency differences among the classes is symptom of an inaccurate model [31].

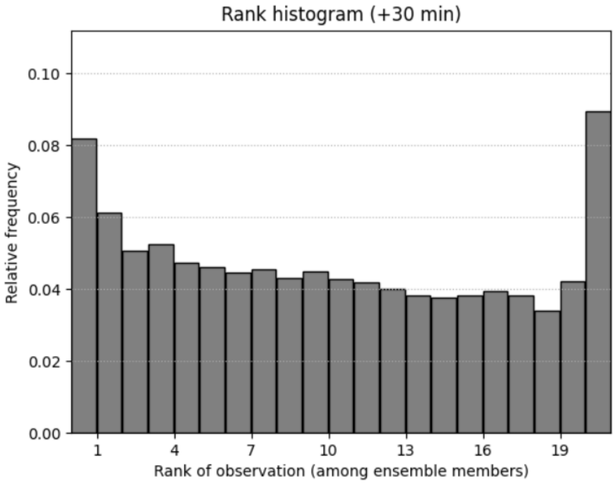


Figure 2.6. Example of a rank histogram [32]

2.3.4 Fraction skill score (FSS)

The Fraction skill score is a quantitative metric that represents the accuracy of the forecast in relation to the spatial extent [16]. It is defined as follows [33]:

$$FSS = 1 - \frac{\langle (P_f - P_o)_n^2 \rangle}{(\langle {P_f}^2 \rangle + \langle {P_o}^2 \rangle)_n} \quad (2.21)$$

Where:

- The numerator represents the MSE between the precipitation threshold frequency of the forecast and of the observations inside each grid cell (the angled brackets indicate a mean over the grid boxes for the length scale n). The numerator defined in this way is also known as fraction Brier score (FBR) [33]
- The denominator represents the FBS of the worst possible forecast for length scale n, namely the forecast that gives the highest discrepancies with respect to the observations [33]

The FSS ranges between 0 (total mismatch) and 1 (perfect forecast) and usually increases as the spatial area considered increases [16]. Therefore, this metric defines a sort of “resolution” of the forecast’s skill. For example, if considering a length scale of 8 Km the forecast shows a high skill, it means that the weather model can represent pretty well the reality at small scales. On the other hand, if it is possible to obtain a sufficient skill only for spatial extents over 50 Km, it means that the weather model gives good results only at coarse resolutions.

2.3.5 Summarizing table

After listing some of the principal metrics used in the meteorological context and in literature, below it is reported a summarizing table:

NAME	TYPE	DEFINITION	PARAMETER DEFINITION	MEANING
ROC CURVES	Qualitative	Plot of false alarm rate (FAR) ratio against hit ratio (HR), where: $FAR = \frac{FP}{FP + TN}$ [28] $HR = \frac{TP}{TP + FN}$ [28]	TP = True positive FN = False negative FP = False positive TN = True negative	Scalar measure of forecast accuracy
RELIABILITY DIAGRAM	Qualitative	Plot of the forecast probability against the observed frequencies [30]	-	Measurement of the bias and resolution of the probabilistic forecast
RANK HISTOGRAM	Qualitative	Frequency histogram where bins defined by forecast values are filled with observed values [31]	-	Qualitative assessment of the ensemble forecast reliability with respect to newly observed data
CRITICAL SUCCESS INDEX (CSI)	Qualitative	$CSI = \frac{TP}{TP + FN + FP}$ [34]	FP = False positive TP = True positive FN = False negative	Verification measure of categorical forecast performance
PROBABILITY OF DETECTION (POD)	Qualitative	$POD = \frac{TP}{TP + FP}$ [34]	TP = True Positive FP = False negative	It represents the percentage of events that are forecasted
PEARSON COEFFICIENT	Quantitative	$\rho = \frac{cov(f,o)}{\sigma_f \sigma_o}$ [17]	f = forecast rainfall in one pixel o = radar rainfall in one pixel $cov(f,o)$ = covariance of f and o over all the grid $\sigma_{f,o}$ = variance of f or o over all the grid	Description of the associations between the forecast dataset and the observation dataset

MEAN ABSOLUTE ERROR (MAE)	Quantitative	$MAE = \frac{\sum_{i=1}^{N} \lVert f - o \rVert}{N}$ [16]	N = Number of pixels f = forecast rainfall in one pixel o = radar rainfall in one pixel	Description of the discrepancies between the forecast dataset and the observation dataset (used for deterministic nowcasts)
BRIER SCORE	Quantitative	$BS = \frac{\sum_{i=1}^{N} (F_i - O_i)^2}{N}$ [16]	P_{fi} = Exceedance probability of the forecast for the i-th pixel P_{oi} = Exceedance probability of the observation for the i-th pixel N = Number of pixels	It basically represents the mean square error of the forecast. The lower the brier score, the more skillful the forecast
FRACTION SKILL SCORE (FSS)	Quantitative	$FSS = 1 - \frac{\langle (P_f - P_o)_n^2 \rangle}{(\langle {P_f}^2 \rangle + \langle {P_o}^2 \rangle)_n}$ [33]	Numerator = FBS between the forecast and the observation per length extent n Denominator = FBS of the worst possible forecast per length extent n	Accuracy of the forecast in relation to the spatial extent

Table 2.2. Summarizing table of typical meteorological verification metrics

2.4 STEPS – BE

STEPS-BE is a nowcasting system developed at the RMI (Royal Meteorological Institute) in Belgium. As the name suggests, it is built on the same principles of STEPS, but it has been optimized for urban application in Belgium [20]. For this reason, some peculiarities have been implemented to adapt this extreme rainfall prediction system to its hydrological scopes in the Belgian domain.

Firstly, since STEPS-BE has been designed for lead-times below 2 hours [20], it does not implement any NWP blending. In fact, as already remarked in previous sections, for the first two hours of the forecast extrapolation methods can offer better results than numerical weather prediction models. Therefore, STEPS-BE follows the same conceptual model explained in section 2.1 and can be categorized as an extrapolation method, where the stochastic noise cascade partially accounts for the phenomena of growth and decay of the precipitation field.

Secondly, the motion field equation (equation 2.1) is solved using a COTREC scheme. Similarly to the VET technique, the equation is computed in smaller subdomains of the radar image, with pixels collected into 10 × 10 pixel blocks. However, in order to solve the motion field equation within each block, an additional constraint is required, which in the COTREC scheme is the minimization of the velocity field divergence $\nabla^2 V$. This condition works as a smoothness constraint, since minimizing the divergence of the velocity field implies decreasing the differences among velocity vectors of adjacent blocks. However, since the minimization is performed only at the block level but not at the pixel level, some sharp discontinuities between pixels belonging to different blocks may arise. As a result, STEPS-BE includes a Gaussian Kernel smoothing that interpolates the velocity vectors in the center of the blocks, which are subsequently assigned to each pixel in the block.

Thirdly, to simplify the computations, an autoregressive model of order 1 is applied to the different cascade levels instead of the second order one used in the original STEPS implementation [20].

Fourthly, in STEPS-BE the stochastic perturbations are generated only inside the radar domain, while in the original STEPS implementation these are also produced outside the observed images when blending of NWP models, which have a coarser resolution than radars, is applied. This adaptation helps in improving the probability matching made at the end of the extrapolation process, that could be disturbed by the stochastic rain cells appearing outside the validity domain [20].

Finally, in STEPS-BE the velocity field is perturbed by simply multiplying the velocity vectors by a factor C calculated as follows [20]:

$$C = 10^{\frac{1.5N}{10}} \quad (2.22)$$

Where N is a normally distributed random variable with zero mean and unit variance. Basically, this means that the velocity field is randomly accelerated/decelerated by a single factor C, while the direction of the movement remains unperturbed.

Chapter 3: RADAR-RAIN GAUGE COMPARISON

As mentioned in the previous chapter, the STEPS model requires radar images as an input. Therefore, before computing the forecasts and making the comparison of py-STEPS with STEPS-BE, it is necessary to make some preliminary analysis on the radar images.

The considerations that will be reported at the end of this chapter allow us to have not only a deeper knowledge about the radar data, but also a better understanding of the results provided by the py-STEPS software. As a matter of fact, radar images can be affected by different sources of error, which depend on: 1) variations in the relationship between the rainfall rate and the backscattered energy (the Z-R relation previously mentioned) ; 2) changes in the precipitation before reaching the ground; 3) anomalous propagation of the beams [35] For example, it could occur that, during the precipitation phenomenon, physical mechanisms like wind speed or evaporation alter the drop-size distribution of rainfall [36], leading to a change in the Z-R relationship and thus to wrong estimation by the radar [35]. Similarly, the presence of obstacles in the surroundings of the radar or of ice-particles inside the cloud can cause scattering or attenuation of the electromagnetic signal, leading to severe underestimations [36]. Again, also errors in the hardware calibration may be present and influence the results of the detected precipitation field [35]. Finally, also heavy precipitation leads to attenuation of the radar signal [37]. Even if this short list does not provide a complete and exhaustive description of all the possible radar errors and uncertainties, it gives a general idea about all the circumstances that can affect the results of these kinds of systems.

The validation of the available data has been carried out through a comparison with rainfall intensities measured by the rain-gauges network of the Flanders region. The basic assumption behind the validation of the radar images is that the rain-gauges represent the “true” value of the precipitation, therefore any discrepancy between the values measured by the radar and the ground stations is interpreted as an error of the former. It is worth noting that assuming the rain-gauges as the “ground-truth” is a simplification, since, obviously, also these can be affected by various sources of errors in their measurements, that usually are related to factors like evaporation, advection and vertical wind motions [35].

In this chapter, in first instance, a description of the available data will be provided. Secondly, all the analysis and considerations will be explained through graphs and summarizing images. Finally, the overall conclusion will be reported.

3.1 DESCRIPTION OF RADAR DATA

The Belgian territory is covered by 4 radar stations, located in Jabbeke, Helchteren, Zaventem and Wideumont. On the overall, the composite of these stations guarantees a complete coverage of the state, as reported in the image below:

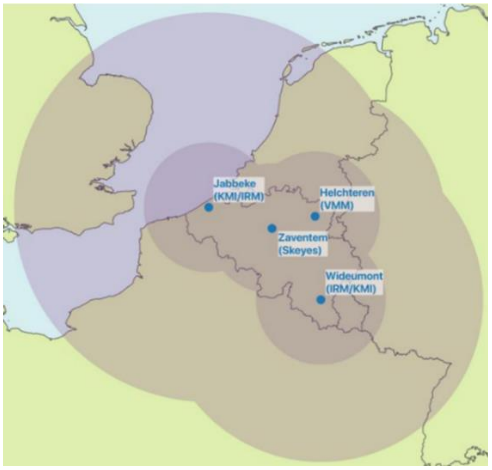


Figure 3.1. Representation of the radar coverage over the Belgian territory [38]

The radar domain is subdivided into a 512 x 512 grid, with a spatial resolution of roughly 1 Km (precisely, 1058m). The grid point locations are expressed in planar coordinates (x,y), using a stereographic projection starting from the reference system WGS84. The temporal resolution of the data is 5 minutes, while the intensity of the rainfall is expressed in [mm/5 minutes]. Finally, the data are provided in netCDF4 format. In the table reported below, all the 10 available events observed by the radar network are listed:

OBSERVATION	DATE	START / END TIME
1	18/05/2017	13:30 – 21:15
2	19/07/2017 20/07/2017	20:05 13:00
3	15/08/2017	02:05 – 11:25
4	29/08/2017 30/08/2017 31/08/2017	21:05 00:00
5	11/09/2017 12/09/2017	07:35 00:00
6	24/05/2018	12:05 – 18:20
7	27/05/2018 28/05/2018 29/05/2018	10:05 03:20
8	31/05/2018	00:05 – 22:50
9	07/06/2018	02:40 – 17:45
10	05/09/2018 06/09/2018	00:05 03:00

Table 3.1. Available radar observations

3.2 DESCRIPTION OF RAIN GAUGE DATA

The rain-station data have been acquired from the open data platform waterinfo.be and cover the years 2017 and 2018. The network comprehends 60 rain gauges, which cover the entire Flanders region, as reported in the image below:

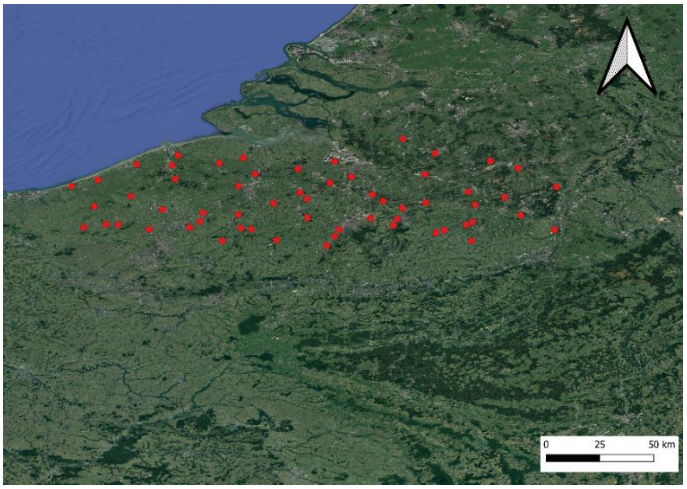


Figure 3.2. Coverage of rain gauges (red dots) along the Flanders region.

Also in this case, the data are provided with a temporal resolution of 5 minutes and are expressed in units of [mm/5 minutes]. Each station is identified with a numerical code, which is related to the geographical coordinates, expressed as longitude and latitude over the WGS84 reference system. Below, a part of the table of the rain gauges' metadata is reported.

Station name	Identity number	Station latitude	Station longitude
Herentals_P	210409042	51.16224315	4.845694177
Korbeek-Dijle_P	210396042	50.83615983	4.64327789
Sint-Joris_P	209278042	51.12657282	2.760080418
Oostkamp_P	207607042	51.13220211	3.253455485
Bonheiden_P	210397042	51.03254282	4.510399646
Liedekerke_P	210392042	50.88663188	4.094898384
Ieper_P	209277042	50.84242046	2.889729506
Maarke-Kerkem_P	210390042	50.81897472	3.670943421
Tessenderlo_P	210407042	51.05208545	5.120756298
Boekhoute_P	210381042	51.26641081	3.68645591

Table 3.2. Metadata of rain gauges

3.3 EVENT EXTRACTION

After acquiring the available observations from the radar and the rain gauges, it is necessary to define a criterion for the extraction of events. In particular, this thesis is focused on extreme precipitation and therefore the validation of radar data should be done for this kind of events. Since, as previously specified, the rain gauges are considered as the “ground truth”, it has been decided to base the event selection on their data. Moreover, it should be clarified that the extracted events must be independent one from each other since the aim of this section is to perform statistical analysis.

It has been decided to adopt a separation of events that keeps into account the functioning and the working principles of the sewage system of a city. As a matter of fact, after the occurrence of a high-intensity rainfall, the sewage system requires a certain time to completely drain. Therefore, if multiple short-duration storms occur within a short time span, there is an increased risk of flooding, since the sewer system did not completely drain in between the storm events. As a result, two or more precipitation events that occur within this time range should be considered as one single event and not distinctly. Conversely, if the separation time between the two events is long enough to guarantee the complete drainage of the system, the two events can be considered as “independent”.

Therefore, the separation and distinction of independent events is based on what is usually defined as “inter-event time” (IET). Actually, the IET depends on different factors, such as the type of the sewer system, the catchment size and the land cover and can vary dramatically around the world. Since no specific catchment was considered in this work and the independence of the events should be general, an arbitrary time of 12 hours as been chosen as IET. Consequently, a specific algorithm has been developed in Python. To clarify its description, the explanatory graph reported below will be discussed:

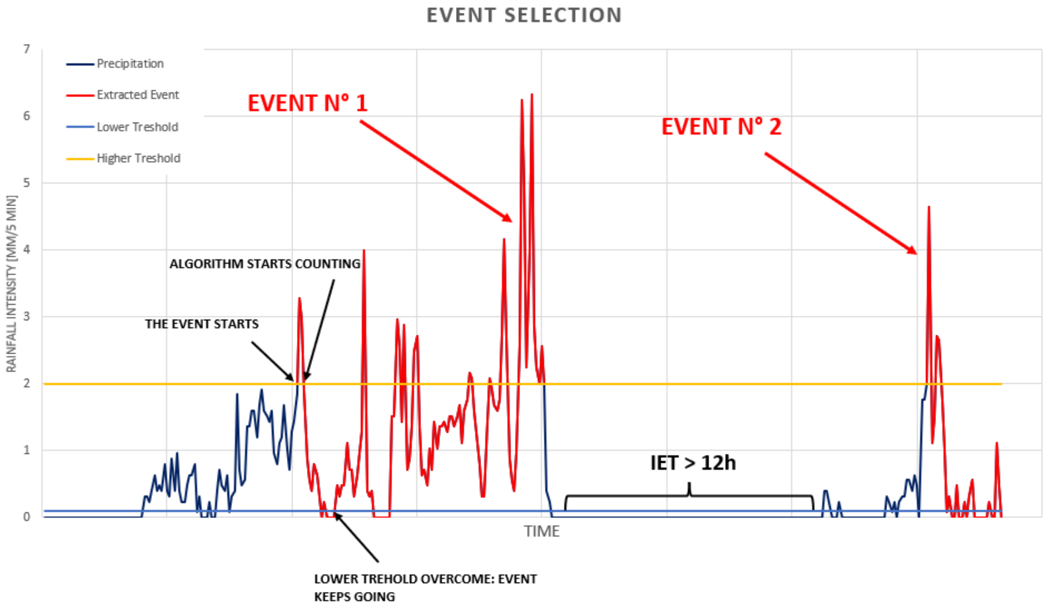


Figure 3.3. Explanatory graph for event extraction

The graph represents the value of precipitation intensity measured by one rain gauge against time. As can be noticed, one high and one low thresholds are set. The exceedance of the higher threshold defines the starting time of the event, while the lower threshold has the function of distinguishing between different events. However, basing the separation of the events only on the thresholds would not reflect the considerations made above; it is also necessary to consider the time factor. Consequently, the algorithm works as follows:

When the upper threshold (set to 2 [mm/5min] in the graph) is overcome, the event starts. Once the rainfall intensity falls below this value, the algorithm starts counting the time (as indicated by the arrow in the image). When the rainfall intensity falls below the lower threshold (set to 0.2 [mm/5min] in the graph), the system keeps registering the values until the counting of 12 hours is reached. If, during the counting, the lower threshold is not exceeded again, the algorithm will only extract the peak precipitation (namely, the values included between the two arrows "the event starts" and "algorithm starts counting"); conversely, if the lower threshold is exceeded a second time, all the precipitation values will be included in the event. Every time the upper threshold is overcome, all the mechanism previously described is reiterated (and the algorithm's counter is set to 0 and restarts).

It can be noted that the lower threshold has been set to a value slightly higher than 0: this is done to avoid that really light rain, which has not an influence on potential flooding of the sewage system, is included inside the event. The image above shows that, if the time between the exceedances of the lower threshold is higher than 12 hours, the events are extracted and classified as independent.

The two thresholds were calibrated through a sensitivity analysis in order to extract a significant number of events to perform statistical analysis. In particular, the higher and the lower thresholds were set to 3 [mm/ 5 min] and 0.1 [mm/5 min] respectively, so that 50 events were extracted after applying the algorithm to all the data of each rain gauge. In appendix A the table that summarizes the event extraction by listing the station location, start time and end time is reported.

3.4 RADAR DATA VALIDATION

After extrapolating the events through the method previously described, it is possible to make a comparison between the measurements of the rain gauges and of the radar system. In the next sections, all the analyses that have been carried out will be presented, as well as the relative conclusions.

3.4.1 Radar errors

As previously mentioned in the introduction of this chapter, the main hypothesis behind the radar data validation is that the rain gauges measure the "true value" of the precipitation; namely, every measurement that differs from what has been recorded by the stations is interpreted as an error of the radar. Furthermore, it is important to specify that, during the following comparisons, the precipitation value of the radar corresponds to the one measured inside the grid's pixel that contains the rain gauge.

Therefore, one first simple way to assess the radar's accuracy consists in plotting a cluster of points of coordinates (x,y), where the x-axis represents the value measured by the rain-gauge, while the y-axis expresses what has been recorded by the radar. The points on the graph below correspond to the precipitation intensities (unit of [mm/5min]) measured throughout all previously defined events. On the graph, all points above the bisector (red line of the graph) represent an over-estimation of the radar, while all points below imply an under-estimation.

If the radar system was really accurate, all the points should be on the bisector; yet, it can be observed that the points on the graph are dispersed evenly above and below the red line. This means that the error of the radar does not follow a precise trend, but it is rather random, with both over and under-estimations.

As can be observed from the plot, most events have relatively low precipitation intensity values for both gauge and radar observations. Consequently, absolute errors in that area are also low. For increasing precipitation, differences between gauge and radar tend to become larger. Points situated in the upper left or lower right corner are the most interesting on the graph, as they show a serious discrepancy between the observations. Probably, this dispersion is the result of the typical radar errors previously mentioned, such as the high sensitivity of the Z-R relation to the drop size or the beam attenuation occurring when high-intensity precipitations are present.

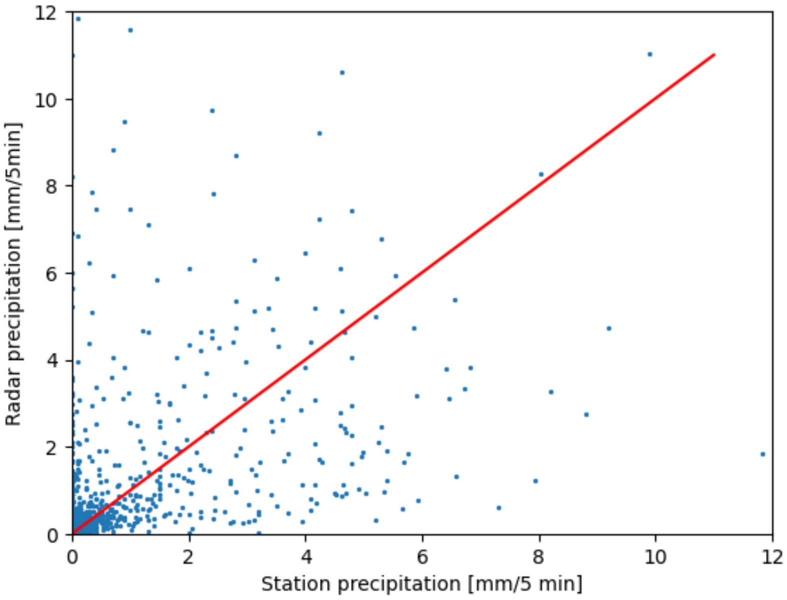


Figure 3.4. Scatterplot of the selected events

3.4.2 Time-shifting between radar and rain gauge data

In second instance, it has been decided to make a visual analysis by plotting the values measured by the rain gauges and the radars side by side for single events, as shown below:

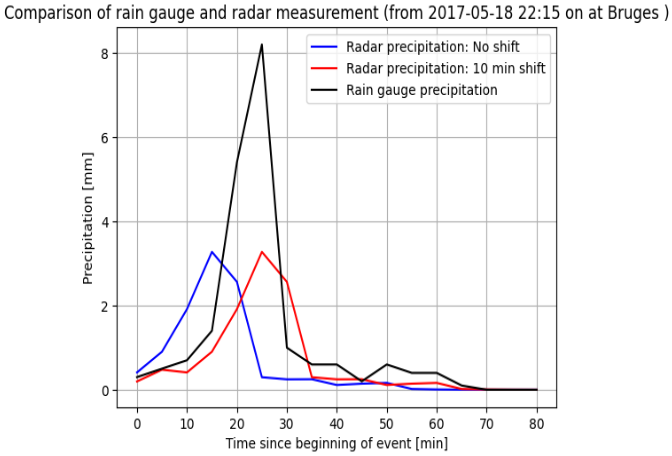


Figure 3.6. Radar and station measurements for an event occurred in May of 2017 at Bruges

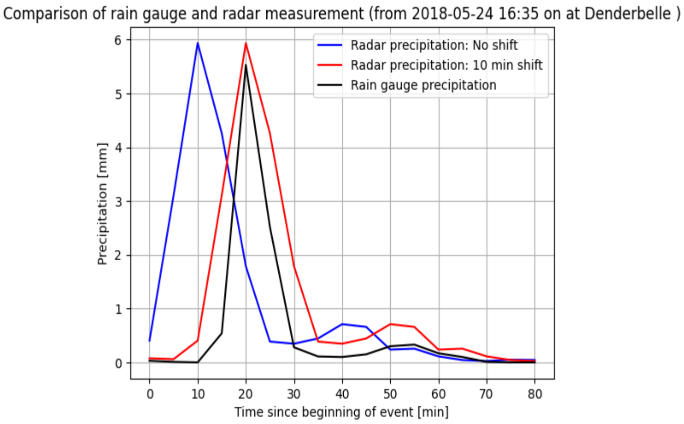


Figure 3.5. Radar and station measurements for an event occurred in May of 2018 at Denderbelle

From these graphs, it is clear that the peaks in precipitation intensity observed by the gauges and the radar are shifted in time by 10 minutes. These results, similar to what has been reported in literature [36], can be explained by considering that the precipitation field intercepted by the radar requires a certain amount of time to fall to the earth's surface and hence to be measured by a rain gauge. As a result, the peak values of the rainfall are measured by the rain gauges with a certain delay compared to the radar system.

However, the time delay between the two maximum values is not always equal to ten minutes for all the events, but a certain variability can be observed. Therefore, a frequency histogram of the time-shift between the two peaks is created, which represents how many events are characterized by a certain time delay:

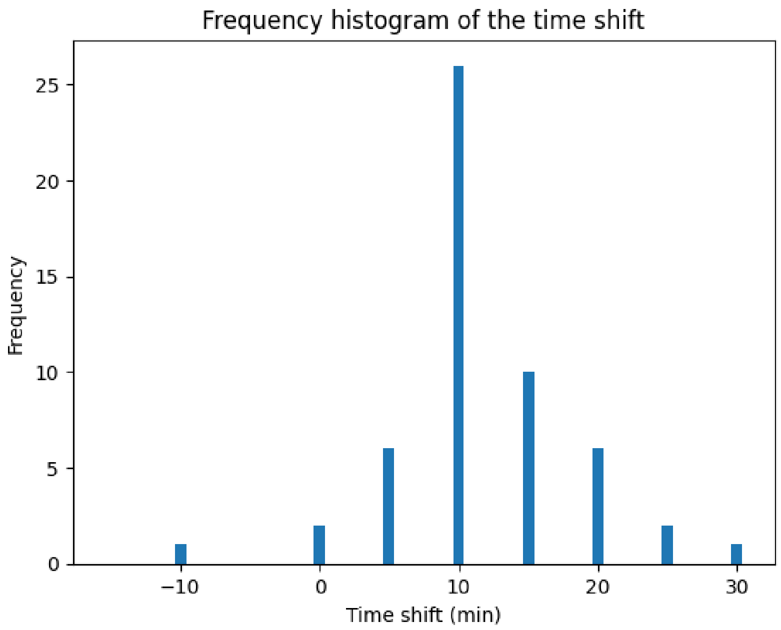


Figure 3.7. Frequency histogram of the time delay for all the considered events

As can be noticed, the events are more or less all characterized by a certain time delay, which, on average, ranges from 0 to 30 minutes. Still, the most common time delay, similarly to the example events reported above, is equal to ten minutes. This information is particularly important for two reasons.

Firstly, if the radar images, that are the required inputs for the STEPS method, represent a precipitation field that will reach the ground with an average time of 10 minutes, similarly, the results of the forecasting system will provide an advected precipitation field with similar characteristics. Namely, the lead time of each forecast based on these radar images can be extended, on average, by ten minutes.

Secondly, it is safe to say that the error between the radar and station measurements, along one single event, will be surely higher if no time shift between the two sets of data is applied compared to the case in which the radar measurements are shifted by the average time of ten minutes.

As a consequence, when some comparative metrics, such as the root mean square error (RMSE) will be calculated in the next sections, the time shift of ten minutes is applied to all the radar datasets of each event. Furthermore, it is important to specify that the time shift has been applied also to the graph previously reported in figure 3.4.

3.4.3 Values in the surrounding of the rain gauges

In the previous analysis, the comparison between radar and station data has been done considering the value measured by the radar in the grid's pixel where the station is located. However, it could be of interest to also consider the precipitation intensities measured by the radar in the surrounding of the rain gauge, to analyze if, by shifting the radar value's location, the error between the two datasets could diminish.

Therefore, two different methods to investigate this aspect have been implemented: one that extracts the maximum value among all the pixels that surround the station and another that calculates the mean of all the precipitation values close to the rain gauge. While performing this operation, it is possible to consider different distances (or radii) from the rain gauges; in this case, it has been decided to take into account the distances of 1058, 2116, 3174 and 4232 meters (in other words, the four pixels around the rain gauge have been selected). Obviously, investigating observations further than these radii would not be meaningful, since extending the gap between the rain gauges and the radar even more, would surely lead to remarked differences in the measurements of the two devices.

It has been found out that considering the maximum value of precipitation intensity in the surrounding of the rain gauge usually leads to an increasing overestimation with increasing distance, as represented in the graph below for an example event:

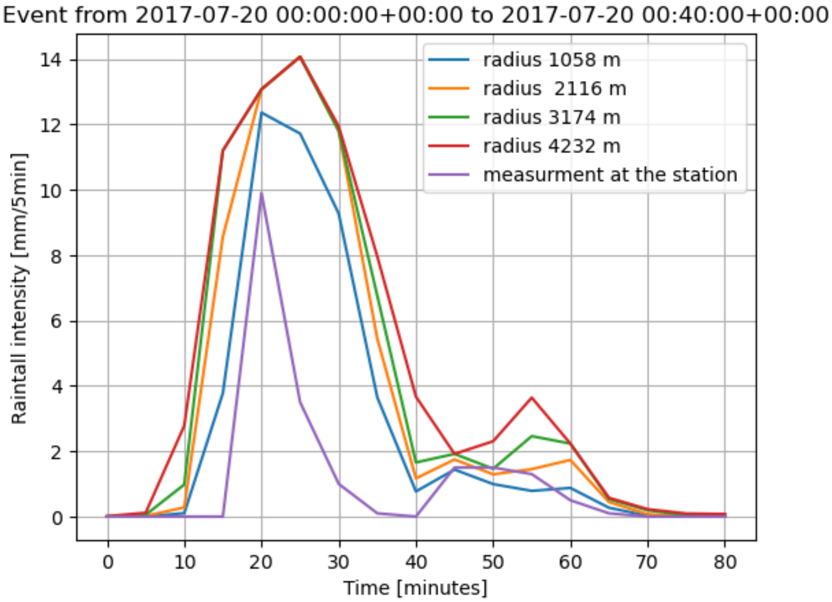


Figure 3.8. Esemple event at Boortmeerbeek

Accordingly to what has been reported in the previous section, all the curves relative to the radar measurements are shifted with 10 minutes in time. In order to understand if what has been observed for the case above is valid for all the considered events, the RMSE between the radar and the station data has been calculated for each event and for each radii. In order to summarize the findings, the boxplots of the root mean square error are reported in the image below:

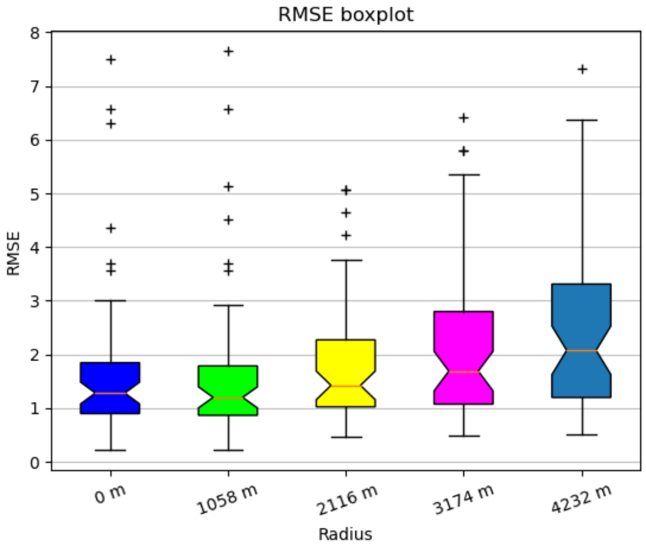


Figure 3.9. Boxplot of the RMSE with increasing distance

The RMSE clearly increases every time the distance between the rain gauge and the pixel containing the maximum precipitation value expands, as it can be appreciated from the increasing trend of the median in the boxplot. Therefore, it can be concluded that considering the maximum precipitation value in the surrounding of the rain gauge does not show any clear benefit in the validation of the radar data. Analogously, it is possible to elaborate the same boxplots considering the second method, namely the one that extracts the mean of all the cells that surround the radar within a certain radius:

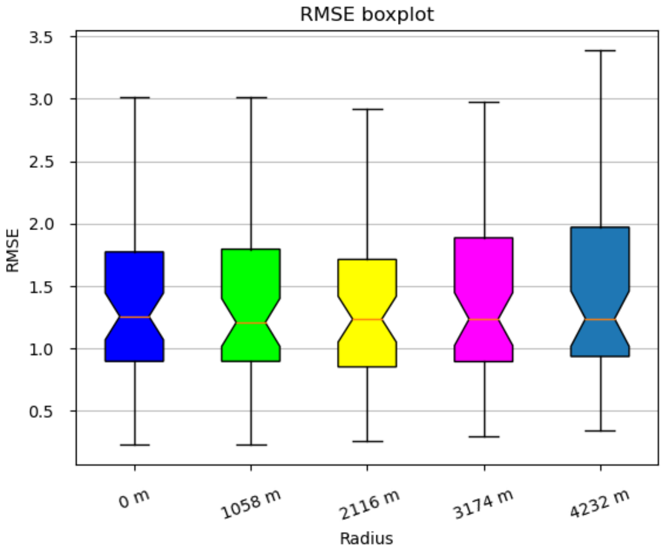


Figure 3.10. Box plots of the RMSE for the second method

For this second case, it can be observed that the RMSE tends to be constant with increasing distance. Unlike the previous case, few but really high outliers have been detected and then discarded from the boxplots.

Although the RMSE tends to be constant with increasing distance, these graphs show again that no benefit is obtained by considering the radar values in pixels that surround the rain gauges. Therefore, it seems that the best strategy for the validation is simply to extract the precipitation values in the same radar-grid cell where the station is located. Consequently, all the next analyses will follow this method.

3.4.4 Aggregation and cumulative volumes

In the previous sections, the comparison between radar and rain gauges has been performed considering the precipitation intensities, measured by the two devices, every 5 minutes. However, as it has been explained, time-shifting and different radar errors can lead to remarked over-estimations and under-estimations by the radar. Therefore, it can be of interest to also consider the total rainfall volumes cumulated during the event, expressed in [mm], in order to check if relevant differences are detected by the two sensors. As a matter of fact, evaluating the cumulated volumes allows to estimate the radar errors regardless of the time shifting that has been previously explained and not considered in the following exemplificative graph related to one event.

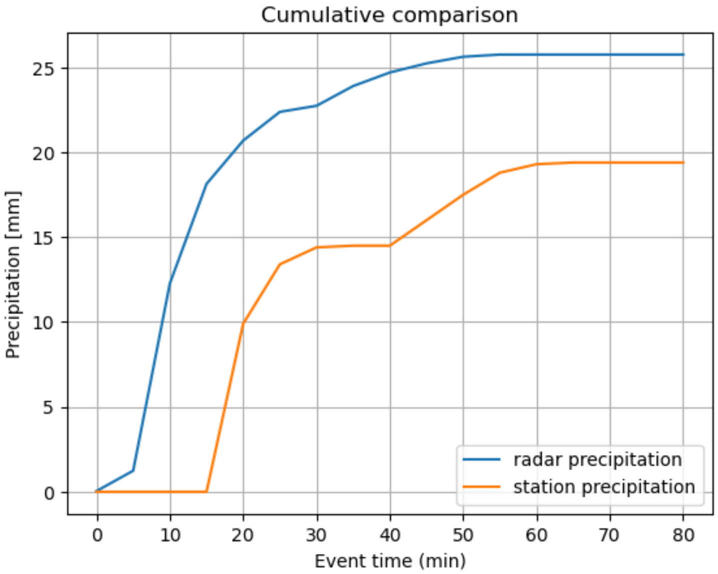


Figure 3.11. Cumulative rainfall volume for an event neglecting the time-shift

As can be observed, towards the end of the event both the radar and the station register a constant cumulative rainfall volume, that is independent of the time shift between the rainfall measured peaks. However, in this case the two total volumes do not coincide as there is an overestimation by the radar. Obviously, it is possible to perform the same operation for all the events that have been extracted and compute the difference between the cumulative volumes measured by the radar and the stations, thus obtaining a scatter plot as the one reported below.

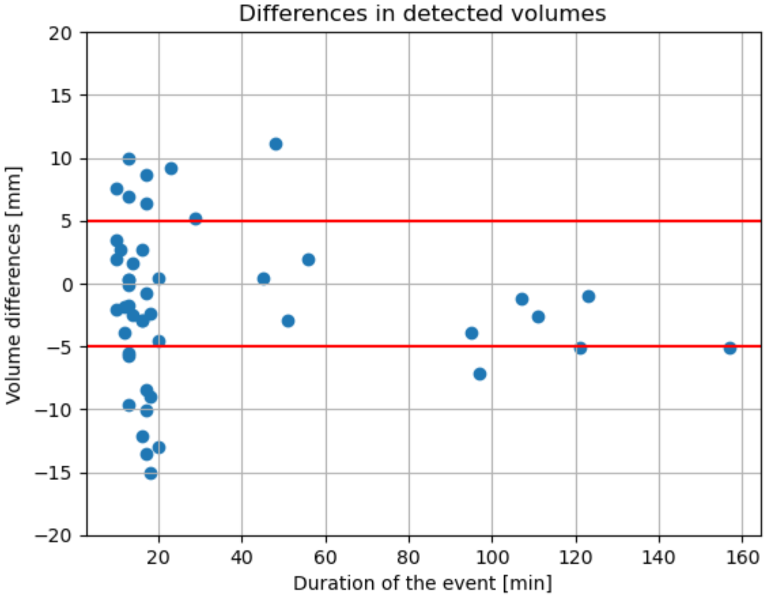


Figure 3.12. Scatter plot of the differences in cumulated volumes between the radar and the rain gauges as a function of the duration of the event

As it was expected, the graph underlines that both over-estimations (positive values) and under-estimations (negative values) are present. It emerges that, for most of the considered events, the estimation of the overall volumes is not highly biased, since the differences between the two devices range between -5 and +5 [mm] of rain (as indicated by the horizontal red lines in the graph). In particular, it can be observed that events with a longer duration are characterized by a higher accuracy of radar's measurements when compared to events that last for a short time. In fact, some events that have a duration of roughly 20 minutes can also present differences in the detected cumulative volumes (up to 15 millimeters).

Besides considering the cumulative volumes, it is also possible to make an additional comparison between the radar and rain gauges by aggregating the rainfall values for different time ranges. Therefore, a rolling window has been applied to the radar and rain gauge observations in order to aggregate the time series 5, 15, 30 and 60 minute timesteps. Obviously, all the events that have a duration shorter than the aggregation time have been excluded from the computation.

After aggregating the values, it is possible to calculate the RMSE for each event and, similarly to what has been done before, to summarize the results in the following boxplot.

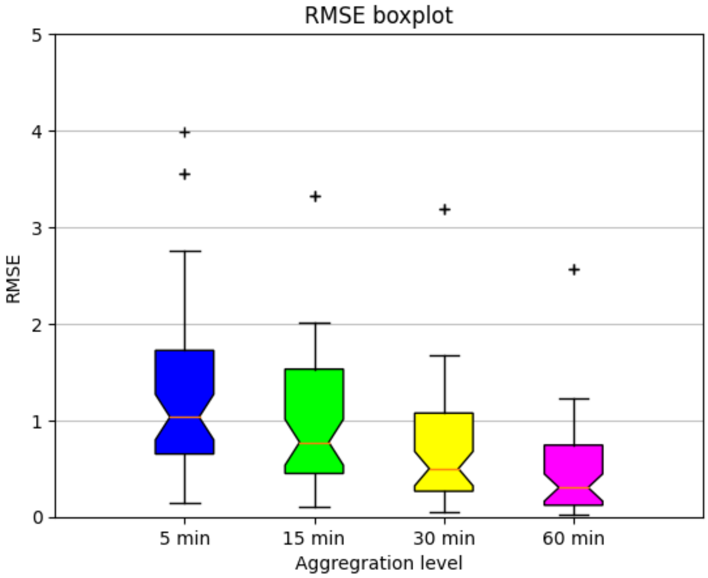


Figure 3.13. Boxplot of the RMSE for different aggregation times

It is clear that aggregating the rainfall intensities has a good influence on the radar's validation procedure. As a matter of fact, not only the RMSE tends to be pretty low for all the aggregation levels (< 2 mm/5min), but it also decreases by increasing the moving average window size. However, since the events have different time extents and can last less than 30 minutes, it must be considered that the estimation of the boxplot for durations of 30 and 60 minutes has a lower number of members (36 out of the 50 considered)

3.4.5 RMSE as a function of the duration of the event

In figure 3.12, It has been observed that events with short duration are characterized by an higher error in the estimation of the cumulative rainfall volume compared to events that last more than 60 minutes. To investigate better this aspect, it is possible to produce a scatter plot of the RMSE, in which the metric has been calculated for every event considering the rainfall intensities in units of [mm/5min], as reported in figure 3.14.

In this specific case, before calculating the RMSE, every value of the considered event has been normalized with respect to the peak observed at the rain gauge, to better highlight the decreasing of the RMSE as a function of the duration. Moreover, all the radar measurements have been shifted by 10 minutes, since, as previously stated, this operation helps in reducing the RMSE value.

As expected, the radar and the rain gauges have a better agreement for those events with long duration, while higher errors in the rainfall estimation intensities are present when the events have a short duration. However, this statement is not valid in general for all the considered events. In fact, a small "cluster" of points can be observed in the lower left part of the graph, where the duration of the event ranges between 0-20 minutes and the RMSE has values between 0.2 – 0.4: in these cases, the estimation of the radar can be considered acceptable despite the short duration.

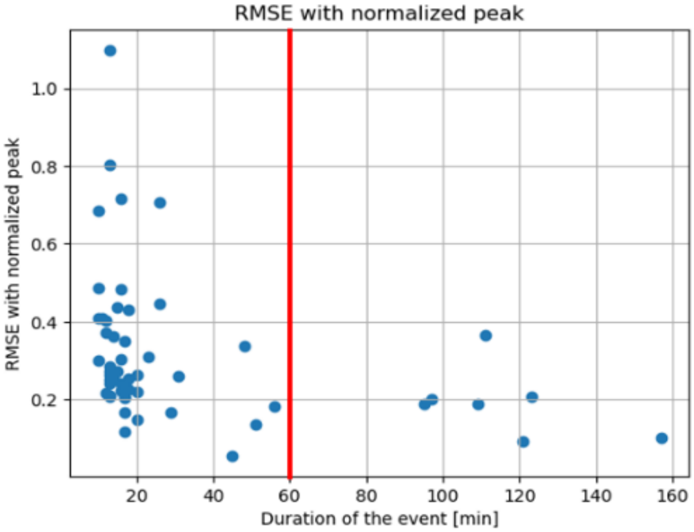


Figure 3.14. RMSE as a function of the duration with normalized peak intensity

3.4.6 RMSE in function of distance between radar and rain gauge

In order to better understand the performances of the radars as a function of space, it has been decided to evaluate the trend of the RMSE when the relative distance between a rain gauge and the nearest radar is considered. Namely, for every station at which the event occurs, both the distance from the nearest radar (among the four available) and the RMSE are calculated. The resulting graph is reported in figure 3.15.

It can be observed that the RMSE tends to decrease when the distance of the considered station with respect to the closest radar increases. Even if this result may seem counter-intuitive, different authors [36], [39] have reported that errors associated with ground clutter and wetting of the radar's radome are more remarked in the proximity of the device, especially for short duration and high intensity events. However, it would be expected that, as observed in another study [36] the RMSE increases again after a distance about 60-70 Km, since at long distances a stronger decorrelation between radar and rain gauges is usually assessed [36]. As a matter of fact, an "optimal" distance between the rain gauge and the closest radar, at which the RMSE is the lowest possible and the typical radar errors are minimized, should be found, contrary to what can be observed from the presented graph. One possible explanation could be that the considered domain is too small to assess an effective rise of the RMSE, since the maximum distance between a station and the closest radar is below 80 km. Moreover, there are only four events in our dataset with a distance higher than 60. Therefore, the available sample is too small to make concrete conclusions.

As a recommendation for future studies in this region, we would advise to retrieve a higher amount of observations, in order to build a bigger sample that allows to make more robust estimations about the RMSE for distances between rain gauge and the closest radar above 60 km.

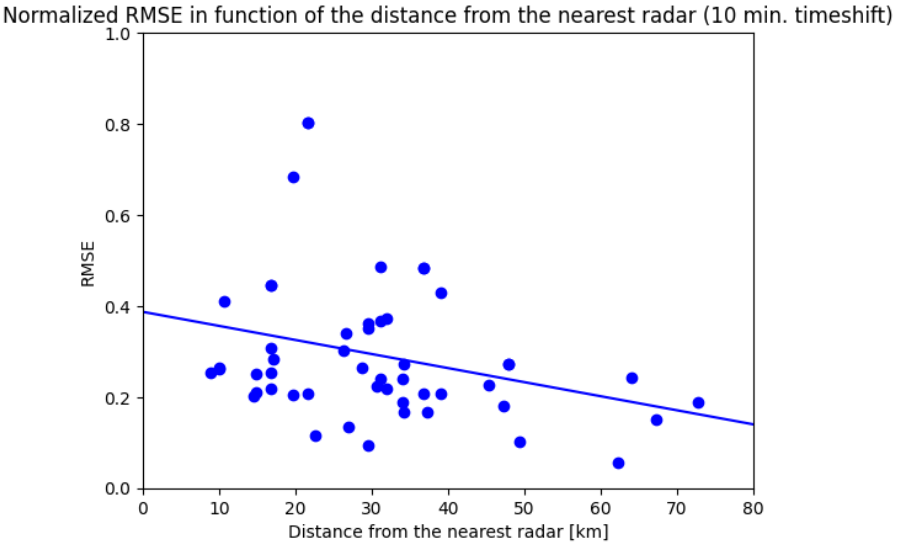


Figure 3.15. RMSE as a function of the distance between the rain gauge and the nearest radar

3.5 CONCLUSIONS

In this chapter, a validation analysis of the available radar images has been performed considering a number of 50 events that have been extracted through an algorithm programmed in Python. Analyzing a number of 50 events helped in drawing more general conclusions.

The validation of radar observations is an important step, as the measured precipitation fields can be employed not only in hydrological modelling, but also as inputs for the nowcasting methods that will be further investigated in the next chapters. Different sources of errors, explained in the introduction, can affect the performances of the radar system and thus may be propagated in the results of the hydrological/nowcast models. The aim of this master thesis is not to elaborate procedures that allow the correction of these different sources of error, but rather to present the key aspects of precipitation radar forecasting, to have a more critical interpretation of the results. However, it is highly recommended to investigate specific radar corrections in future studies.

Firstly, it has been assessed that the selected events are characterized by a time-shift between the observed peaks of the radar system and of the rain gauges. Specifically, it has been observed that the most frequent time shift is about 10 minutes on average, but also other time-delays surfaced, as underlined by the frequency histogram in figure 10. Therefore, it can be stated that, for example, nowcasts based on radar images represent a rainfall field that will take 10 more additional minutes to precipitate on the ground.

Secondly, no systematic error of the radars was observed, as both underestimations and overestimations have been assessed with the same frequency. In particular, this last statement applies not only to rainfall intensities, measured in units of [mm/5 min], but also to cumulative volumes, as depicted in figures 3.4 and 3.12. Unfortunately, since there is not a systematic error of radar images, it is difficult to predict, in case of a new rainfall event, if the radar will give an overestimation or underestimation of the cumulative volumes at the end of the event; this may have important implications for models that operate at local-catchment scales. However, as highlighted by the boxplots in figure 3.13, aggregating the rainfall intensity values may increase the overall accuracy of the radar images, since a decrease of the RMSE is observed.

Thirdly, the possibility of considering pixels in the proximity of rain gauges during the validation procedure has been taken into account. Yet, no benefits have been observed by extracting the maximum value or computing the average of rainfall intensities around the station's location. This implies that, when it comes to evaluating the precipitation intensity in a certain location, it is trivially more convenient to consider the value inside the pixel that contains the point of interest.

Finally, the accuracy of the measured precipitation as a function of the duration of the event and of the distance from the closest radar has been evaluated in terms of RMSE. As previously explained, it has been noticed that long duration events are estimated by the radar with higher accuracy with respect to short duration ones. In addition to that, image 3.15 shows that the RMSE tends to decrease as the distance from the nearest radar increases. This implies that, at least for the considered Belgian domain, rainfall images are more reliable when the considered points have a certain distance from the radar system, while observations close to the radars could be affected by ground clutter and wetting of the radar's radome, reducing the performances.

After carrying out the validation analysis and delineating the necessary conclusions, a basic understanding of the value and shortcomings of radar observations has been acquired. With this context in mind, the next steps of this study focus on the creation of precipitation forecasts, which require precipitation radar images as input.

Chapter 4: DETERMINISTIC NOWCAST

After performing the validation step, it is possible to use the radar data as inputs for the nowcasts. In particular, this chapter will focus on deterministic nowcast, while probabilistic short-term forecasts, based on STEPS and STEPS-BE methods, will be treated in the next chapter. It has been decided to evaluate the performances of three deterministic models for the current analysis: extrapolation, S-PROG and ANVIL (Autoregressive Nowcasting Vertically Integrated Liquid Method).

In chapter 2, it has already been explained how extrapolation methods advects the precipitation field in time without modeling the growth and decay of precipitation structures (i.e., following the so called "Lagrangian persistence"), while S-PROG, by decomposing the rainfall field and applying an autoregressive model, is able to simulate changes in intensity throughout the forecast. On the other hand, ANVIL is a more recent method, developed by Pulkkinen et al. [40], that brings two remarkable changes to the S-PROG scheme [40].

Firstly, the input data to ANVIL is not a rain rate, but a vertically integrated liquid (VIL) variable, namely the total mass of water that is possible to retrieve observing the reflectivity of weather radars. Secondly, an ARI (integrated autoregressive) model instead of a simpler AR is applied to each cascade level of the decomposed VIL. The ANVIL method is implemented in py-STEPS library and offers the possibility of using both VIL and rain-rates as inputs for the nowcast. As reported by the developers [41] , using the ARI model should "avoid biasedness and loss of small-scale features in the forecast field, and no statistical post-processing is needed". Therefore, ANVIL is expected to produce nowcasts that are more reliable than S-PROG, as also shown in the research conducted by Pulkkinen et al. [40]

The results of the above-mentioned methods will be evaluated by performing a sensitivity analysis on the motion field, the number of lead times and the precipitation thresholds by means of both categorical and deterministic metrics that will be discussed in the following sections. The nowcasts will be computed not only over the entire radar domain (namely, in the 512 x 512 radar grid), but also in smaller portions of area 50 x 50 km^2, that should simulate the size of a hydrological catchment. In fact, it is of interest of this thesis to verify the performance of the models in smaller areas where flooding could occur. In addition to that, a method based on the power spectrum of the precipitation field that allows to distinguish between stratiform and convective events will be presented. Thanks to this classification, it will be possible to assess which method better forecasts the convective events, which are usually associated to higher risk of flooding.

Therefore, the chapter is structured as follows. In section 4.1 the sensitivity analysis related to the motion field will be presented; in section 4.2 a first analysis of all the events retrieved in the previous chapter will be performed; in section 4.3 the method to characterize the convective nature of the events will be explained; in section 4.4 the analysis of the events will be repeated by distinguishing between convective and stratiform events; in section 4.5 the conclusions will be presented.

4.1 MOTION FIELD SENSITIVITY ANALYSIS

As explained in chapter two, computing the motion field is the first step in every nowcast model. In fact, the precipitation field needs to be advected along the Lagrangian trajectories defined by an optical flow algorithm through each lead time of the forecast. Different optical flow algorithms can be used for this purpose and currently in py-STEPS three of them are implemented: Lucas-Kanade, VET and DARTS.

Lukas-Kanade is one of the most famous algorithms in which the local features are tracked in a sequence of two radar images [19]. In the current python implementation, a final interpolation step is performed to produce a smooth field of velocity vectors. DARTS (Dynamic and Adaptive Radar Tracking of Storms) is a spectral approach based on the Fourier transform of a temporal sequence of radar fields. Contrary to Lukas-Kanade, this optical flow algorithm requires a higher number of precipitation fields to be applied (at least nine radar images) [22]. VET is a global Variational Echo Tracking approach. In VET the radar image is segmented into small regions called "echoes", where velocity vectors can be computed by minimizing the difference in the reflectivity between two radar maps separated by a certain time interval [42].

Since the motion field is potentially one of the most relevant factors that influences the nowcast's outcome, a sensitivity analysis is performed to establish which method will be applied in the following nowcasts. The following image shows the result of the three different methods applied to a precipitation field:

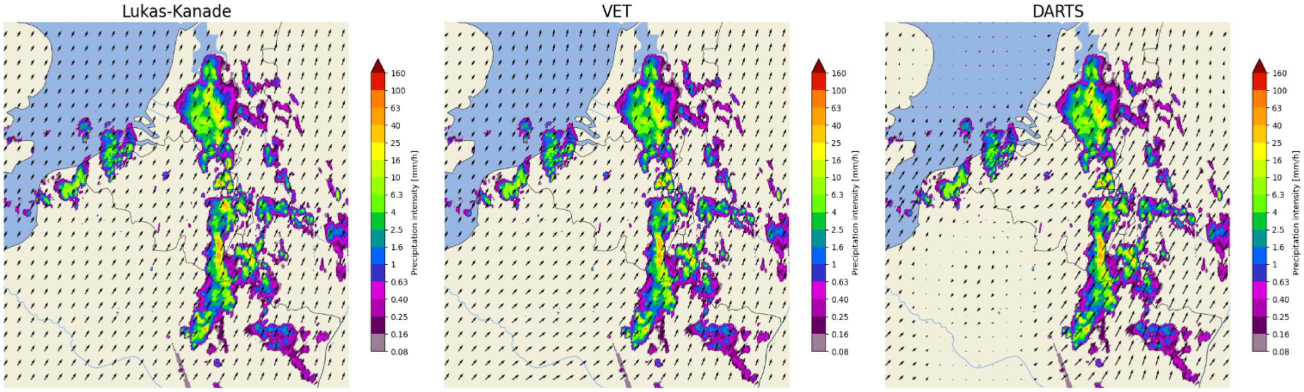


Figure 4.1. Three different motion fields estimated for the same precipitation field by Lukas-Kanade, VET and DARTS algorithms.

As it can be observed, the Lukas-Kanade and VET algorithms yield a similar result, while DARTS presents some remarked differences, especially in the zones far from the precipitation field, where the velocity field can also assume values equal to 0. This has little effect on the forecast results.

In order to perform a quantitative analysis that allows to evaluate the benefits offered by one motion field than another, two metrics have been chosen.

The first one is the Critical Success Index (CSI), already introduced in table 2.2, that expresses the ratio between the hits of the forecast divided by the sum of the hits, the false alarms and the misses. This is therefore a categorical (or qualitative) metric, and it is representative of the predicting ability of the evaluated model. The second one is the Mean Absolute Error (MAE), which expresses the mean of the absolute value of all the residuals in each pixel of the studied domain and it is therefore a continuous (or quantitative) metric that represents the forecast's accuracy.

To perform the comparative analysis, it has been decided to evaluate the results of the extrapolation method for two-hours lead time. The CSI is calculated using a threshold of 5 [mm/h], while the MAE is calculated considering all the pixels of the radar domain. It should be noted that this is not an optimal choice when it is necessary to evaluate the accuracy of a single method, as the pixels without precipitation can create a fictitious high correlation that decreases the Mean Absolute Error. However, since the scope of this analysis is to evaluate the performance of the model by varying the motion field and considering that all the forecasts have been made with the same starting conditions, this choice does not have an influence on the final comparison. In the next sections, when the accuracy of the single method (extrapolation, S-PROG, ANVIL) is more relevant, the MAE will be calculated considering different initial settings. The graphs reported below represent the average value of the metric for each lead time among all the events extracted and explained in chapter 3 (and whose summarizing table can be found in the appendix of the present thesis).

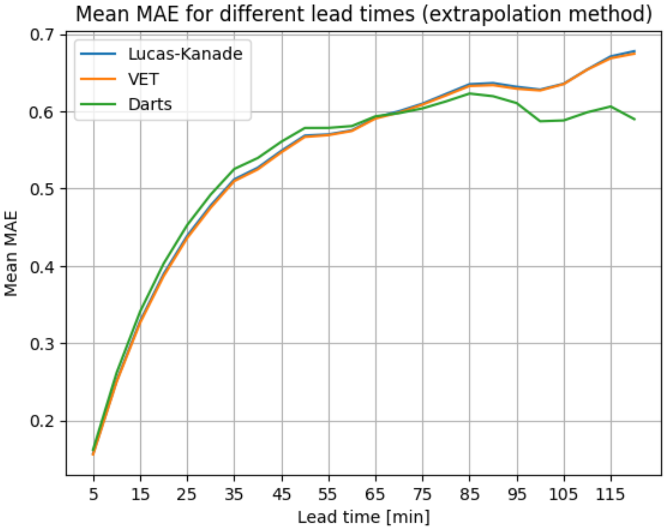


Figure 4.3. Mean MAE at different lead times for the three optical flow algorithms

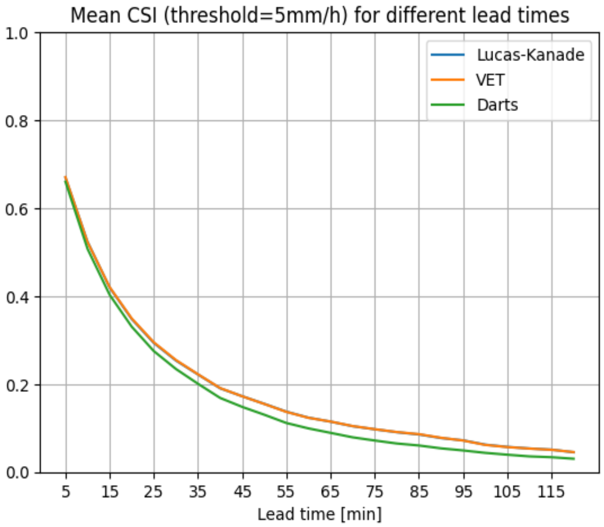


Figure 4.2. Mean CSI for 5[mm/h] threshold for the three different optical flow algorithms

As it can be observed, the choice of the optical flow method has a limited influence on the results of the nowcast, since the curves on the graphs tend to follow the same trend and to be close one to each other (the Lucas-Kanade curve is practically hidden under the VET one). However, as for figure 4.3, it can be noted that in the first 65 minutes lead time DARTS algorithm provides more biased results compared to VET and Lucas-Kanade, while after 1 hour these last two methods present a higher MAE. On the other hand, the VET and Lucas-Kanade algorithms provide a higher CSI after the first 15 minutes lead times as it can be observed in figure 4.4. These results were expected, as in figure 19 it can be appreciated that the motion fields provided by Lucas-Kanade and VET are remarkably similar.

Another relevant aspect related to these algorithms is the computational time required to produce the motion fields. In fact, in practical applications the nowcasts must be produced in the shortest possible time, especially if intense precipitations that could cause flooding has to be monitored. Therefore, it has been decided to produce a summarizing table that represents the mean computational time, among all the selected events, required to produce the motion fields associated with the three optical flow algorithms. The calculation has been performed using an Intel Core i7 – 10510U CPU with 4 cores running at 2.30 GHz, over a radar domain of grid 512 x 512 as previously specified:

	Lucas-Kanade	**VET**	**DARTS**
Mean computational time (seconds)	1.1	14.8	1.7
Relative computation time (%)	7.4	100	11.5

Table 4.1. Mean computational time required to produce the motion fields per method.

The first row of the table reports the computational time, while in the second one the computational time in percentage related to the VET algorithm is presented. As it can be observed, the Lucas-Kanade and the DARTS algorithm require a much lower computational time when compared to VET; in particular, Lucas-Kanade is the fastest one.

In conclusion, after the presented analysis, the Lucas-Kanade algorithm has been chosen as default setting for the upcoming nowcasts. In fact, this algorithm presents the lowest MAE in the first hour lead time and leads to higher CSI values with the shortest computational time when compared to the other two options.

4.2 DETERMINISTIC NOWCASTS OVER THE ENTIRE DOMAIN

In this section, a first analysis regarding the three models previously discussed (extrapolation, S-PROG, ANVIL) will be presented. The nowcasts have been performed considering the entire radar domain and by considering the events selected in chapter 3. However, it must be mentioned that not all the events that are reported in the appendix table have been used. In fact, due to the size of the dataset, that not only reports only 10 days of precipitations, but also has some lack of images during the events, it was not possible to retrieve enough radar observations to calculate the verification metrics. For this reason, it has been decided to apply a time shift of 20 minutes to all the starting dates of the events (namely, each event's start time is set 20 minutes earlier than the starting time determined by the algorithm outlined in a previous chapter). This has been done for two reasons. Firstly, this allowed to retrieve most of the observations needed for computing the verification metrics within two hours lead time. Secondly, shifting back in time the starting of the event allows to track not only the decay of the highest rainfall intensities, but also their growth. Therefore, thanks to the time shift, 46 events could be studied in this case.

For the current analysis both categorical and continuous metrics have been used. In particular, as for the continuous metrics, the Mean Absolute error (MAE) and Pearson’s correlation coefficient have been selected, while Critical Success Index (CSI), Probability of detection (POD) and False Alarm Ratio (FAR) have been chosen as categorical metrics. All these metrics have already been defined in table 2.2, therefore the reader is addressed to chapter 2 for reference.

Moreover, before presenting the results, it must be noted that the S-PROG and ANVIL nowcasts have been computed by making use of eight cascade levels, applying an AR(2) and by setting all light precipitation (i.e., below 0.5 [mm/h], that is a really small precipitation value since events that can reach up to 100 [mm/h] are considered) to 0 [mm/h] in both the observed and forecasted precipitation fields. This has been done to avoid that really light rain could fictitiously increase the level of correlation between the nowcasts and the observed precipitation fields. In addition to this, it has been chosen to use the motion field produced by Lucas-Kanade algorithm for all the forecasts, due reasons that have been explained in the previous section. All the graphs that will be presented are not referred to single events but are the results of the average computed over all the 46 considered events.

4.2.1 Continuous metrics

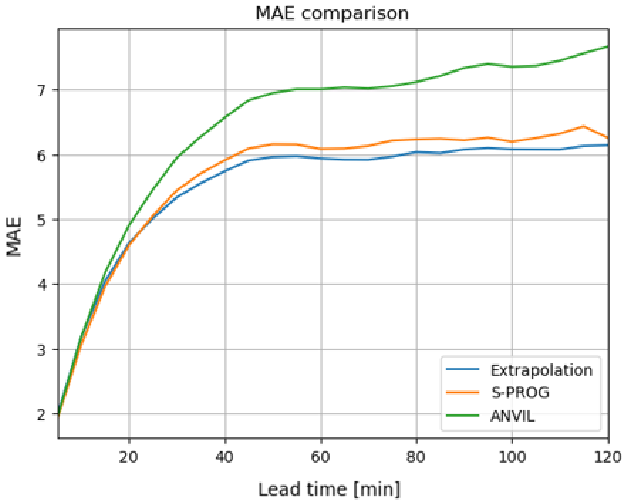


Figure 4.5. Averaged MAE for the three different nowcast models

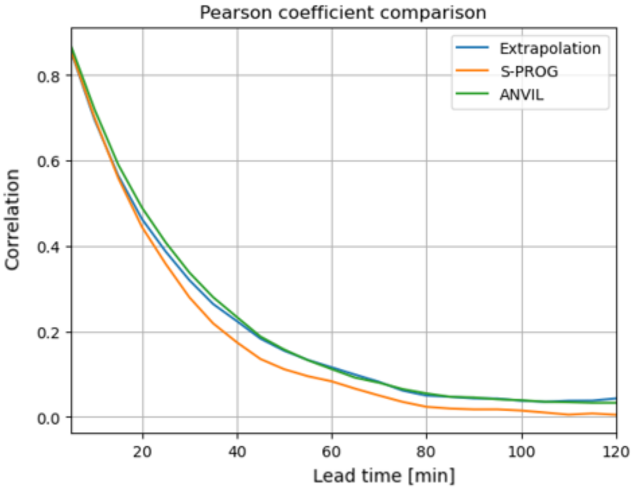


Figure 4.4. Averaged Correlation coefficient for the three different nowcast models

It must be specified that the continuous metrics were calculated by imposing a double conditioning of 1 [mm/h] threshold. This means that the MAE and the correlation coefficient were computed only in those pixels in which both the nowcast's result and the observation overcome a value of 1 [mm/h]. Again, this was done to avoid that really light rainfall values could cause an increase of the correlation coefficient or a decrease in the mean absolute error (again, 1 [mm/h] is a really small precipitation value reminding the high intensity reached by some of the events). In fact, the main aim of these nowcasting models is to intercept the higher values of precipitation to be integrated in flooding control systems.

Looking at the MAE graph, it can be noticed that at the beginning of the forecast all the models have a similar performance, while after 20 minutes lead time ANVIL presents the highest values. Surprisingly, the extrapolation shows a low Mean Absolute Error together with S-PROG, which has an average value slightly above the one of the extrapolation method. This means that, on the overall, ANVIL shows the most biased results among the three models.

On the other hand, figure 4.5 shows that ANVIL maintains the highest level of correlation after 20 minutes. However, also in this case the extrapolation method shows good performances, since the Pearson's coefficient values are slightly lower than the ones of ANVIL until 40 minutes lead time. Eventually, S-PROG shows the lowest level of correlation in comparison with extrapolation and ANVIL.

4.2.2 Categorical metrics

The three chosen categorical metrics, namely the CSI, POD and FAR, have been computed by setting three different precipitation thresholds: 1, 5 and 10 [mm/h]. The results are reported in the graphs below:

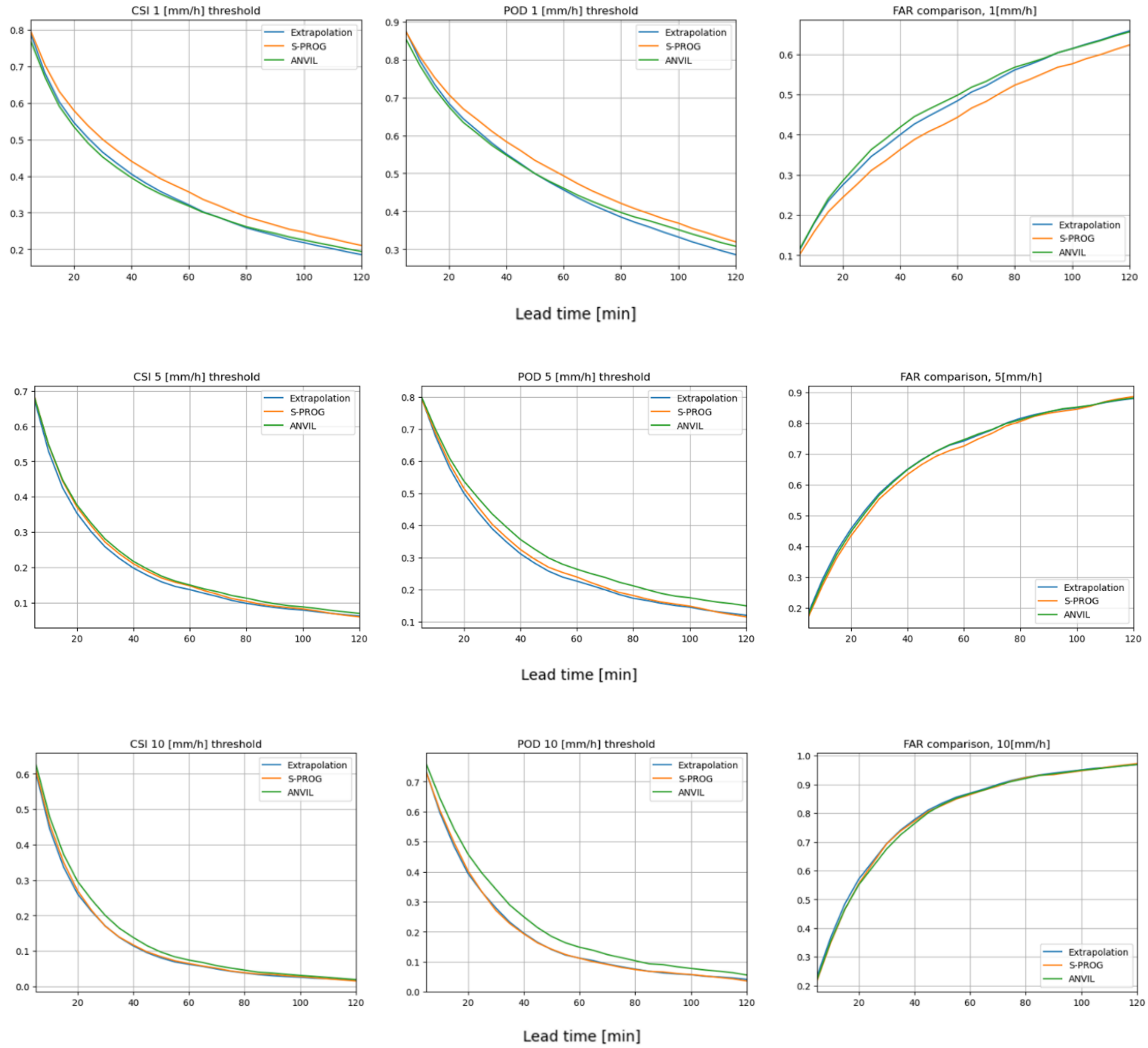


Figure 4.6. Averaged CSI, POD and FAR for the three different models at 1, 5 and 10 [mm/h] threshold.

Firstly, it can be observed that, as expected, increasing the precipitation threshold leads to a loss of skill of all the three models, with the values of CSI and POD that decrease and FAR that increases for every lead time as the threshold shifts from 1 to 10 [mm/h]. In second instance, it is pretty evident that the most skillful method depends on the applied precipitation threshold. In fact, when the threshold is set to 1 [mm/h], S-PROG provides higher values of both CSI and POD and lower estimates of FAR with respect to the other two methods, which present comparable results. Conversely, when the threshold is raised to 5 and 10 [mm/h], ANVIL seems to get the best results, with higher values of CSI and POD, even if the FAR of the three methods is similar. Therefore, even if ANVIL seems to be the more biased method in terms of MAE, it proves to be more skillful when the precipitation threshold is increased.

However, as explained in chapter two, it is interesting to evaluate the model's predictive ability not only over the entire radar domain, but also in relation to smaller scales. To evaluate the performances at smaller resolutions, the Fraction Skill scores have been computed for the three methods, as reported in the figure below:

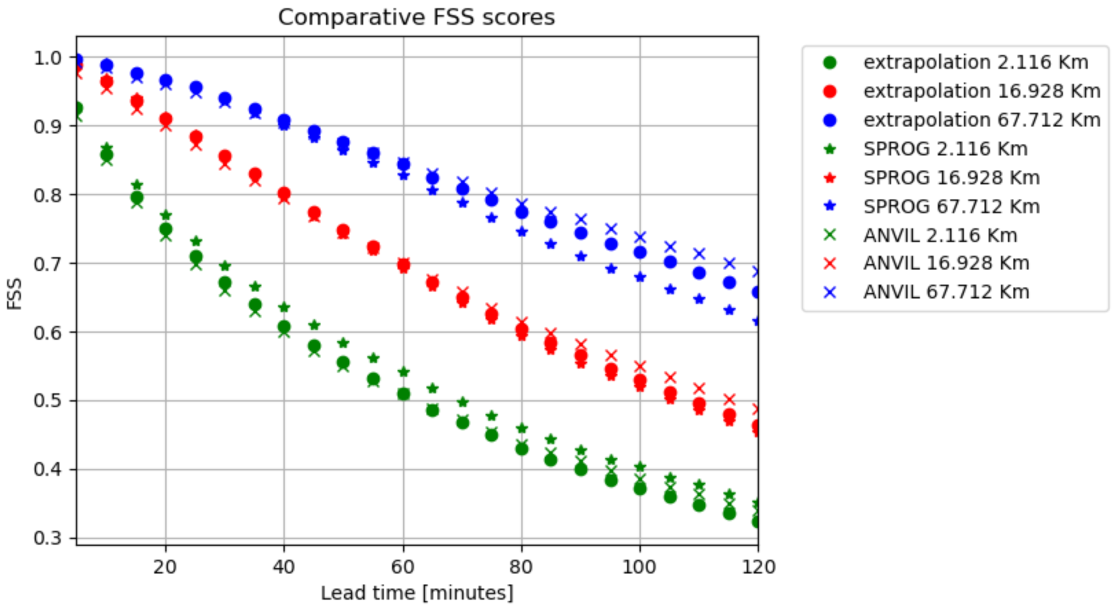


Figure 4.7. Averaged fraction skill scores at different resolutions for the three methods: ANVIL, S-PROG and extrapolation

As expected, the FSS values decreases as the analysis shifts from coarser (67.712 Km) to smaller resolutions (2.116 Km). The scales reported in the legend are dependent on the radar's grid resolution, which has a value of 1.058 Km per pixel. At smaller resolutions, S-PROG proves to have the highest skill starting from the first 20 minutes lead time, while ANVIL gets better results compared to extrapolation only toward the end of the nowcast (which is set to 2 hours). For the coarsest resolution (67.712 Km), ANVIL has the highest skill after 60 minutes lead time, while S-PROG gets the worst results after the same amount of time. However, it can be noticed that for the first hour lead time, all the methods have a similar performance. As for the intermediate resolution (16.928 Km), it can be noticed that no method exhibits a net prevalence over the other, with ANVIL that becomes more skillful only toward the end of the nowcast.

Finally, before introducing the next section, it seems appropriate to make some general considerations related to the results that have just been presented. In fact, on the one hand, the graphs suggest that autoregressive methods such as ANVIL and S-PROG are able to provide better results especially for categorical metrics at low (1 [mm/h) and high (10 [mm/h]) precipitation thresholds with respect to extrapolation. However, the MAE shows that extrapolation is able to provide forecasts with small errors. In addition to that, looking at the categorical metrics, it can be noticed that the extrapolation method does not performs so worse than ANVIL and S-PROG. These results may have different explanations.

Firstly, the number of studied events is not enough to guarantee a wide variety of different storms that allows to draw definitive conclusions about which method performs the best. Moreover, it must be considered that, due to the limited size of the dataset (that reports only 10 days of radar images), some of the studied precipitation fields belong to the same day and are separated by few minutes/hours one with respect to the other. Obviously, such fields have some level of correlation that does not help in providing a good variety of different events as mentioned above. Another explanation is that, again due the limitedness of the dataset, it is not present an high number of precipitation fields that are characterized by a remarked level of growth and decay of small and big structures, so that the true power of models such as S-PROG and ANVIL, that are able to better model such phenomena compared to extrapolation, is not completely exploited. Finally, it should also be mentioned that in this section no categorization has been applied to studied events.

However, it would be of interest to distinguish between events that have a higher or lower level of convectivity. In fact, literature [19], [29], [43] shows that the categorization of events can have a relevant impact on the performances of the nowcast models. In addition to this, it is of interest of this master thesis to focus more on the convective events which can cause major problems to water drainage systems, especially in urban environments, and lead to sudden pluvial flooding. For this reason, it has been decided to implement an algorithm in Python that helps in categorizing the events and that will be exploited to carry out further analysis on the events in the last section of the present chapter.

4.3 CATEGORIZATION OF CONVECTIVE AND STRATIFORM EVENTS

As pointed out by different authors [19], [29], [43] , the performances of nowcasting models may vary depending on the meteorological nature, the persistence and the intensity of the studied event. In fact, usually both deterministic and probabilistic nowcasting methods show better performances in reproducing events that are characterized by moderate levels of growth and decay of small precipitation structures. In this context, it seems proper to recall one sentence from the py-STEPS developers: "despite the ability of PYSTEPS to generate some new light random rain, it is not designed to represent the uncertainty related to an explosive initiation of a thunderstorm" [19]. Therefore, these results suggest that it would be appropriate to make a distinction of the events based on their meteorological nature, to test the loss/gain of skill of the three different deterministic methods analyzed in the current chapter.

In the context of hydrology, one of the most used categorizations is related to convective and stratiform precipitation phenomena. Convective events are triggered by warm and humid air masses that, in case of atmospheric instability, arise and reach the upper layers of the troposphere, where, in normal conditions, the temperatures tend to be lower than in the planetary boundary layer. When this happens, due to the temperature changes, the water vapor contained in the air condenses and forms big and vertically developed clouds such as *cumulus* or *cumulonimbus,* which are usually associated with high rainfall intensities [44]. Since these kinds of events occur when both the ground's surface temperatures and the intensity of evapotranspiration are high, their frequency is elevated especially in summer / early autumn periods. On the other hand, stratiform precipitations are associated with the collision between a warm and a cold air front, thus leading to the formation of very flat and extended clouds at low elevations, such as *altostratus* and *nimbostratus* [44]. Therefore, stratiform precipitations exhibit lower rainfall intensities which persist for a longer time compared to convective phenomena. Although this simple classification may result useful in many cases, it is not unusual to find mixed clouds, where extended stratiform rain embeds some small convective cells.

Even though the previously mentioned categorization may sound conceptually easy, it is not equally simple to build a precise methodology that allows to automatically distinguish between the convective and stratiform precipitations. In literature many different approaches can be found. Some of them are based on the measurements of rain gauges [45], [46] and mainly rely on the persistence of certain values of rainfall intensity at different timestamps; other methods exploit the Z-R relation of the weather radar measurements and base the distinction on the droplet size distribution inside the clouds [47]; in some studies satellite images have been employed [48]

However, only radar images and rain gauge measurements were available for this study and therefore one method based on this kind of data had to be found. One first idea was to define a simple parameter ("β parameter") calculated through the measurements of the rain gauges and inspired by one paper that deals with this kind of classification in Spain [45]. The simple concept behind this parameter is to define a ratio where at the denominator only the precipitation values above a certain threshold are summed for every timestamp, while at the denominator all the measured values are cumulated; the ratio value determines how convective the event is (ratio of 100% mean extremely convective, value of 0% refer to stratiform cloud).

Even though this method was conceptually easy and simple to automatize, it has the big flaw of evaluating the convectivity over only one pixel out of the entire radar domain and thus it was discarded. A second alternative was based on the 2D radar images and consisted in the detection of convective cells inside stratiform clouds. Basically, an algorithm, developed in Python, scanned all the most extreme rainfall values and, for each one, calculated a radius outside which the mean precipitation values would be significantly lower than the extreme rainfall value. Therefore, every pixel inside the radius would belong to the convective cell, while everything that lied outside this region would be classified as stratiform. However, the precipitation threshold value to define one convective cell was not pre-defined and, in addition to that, the radius that defines the precipitation cell is highly variable among different radar images. Therefore, developing the parameters of the algorithm would have required a sensitivity analysis over a larger dataset.

The last and eventually chosen alternative is based on the power spectrum of the precipitation fields. As introduced in the literature review, the precipitation field can be interpreted as a "hierarchy of precipitation structures embedded in each other over a continuum of scales" [20], which can be isolated thanks to a Fast Fourier transform. Now, convective clouds, which include different precipitation intensities in their structures, will be described by a higher power of small-scale structures than stratiform clouds, that, conversely, are usually associated with broader and smaller precipitation intensities. Therefore, the power spectra of convective and stratiform clouds should present differences that allow a neat distinction between the two. Consequently, the key idea behind this method is to retrieve two representative power spectra for convective and stratiform precipitation respectively by means of a fast Fourier transform (FFT) and a regression applied to the power spectrum of some exemplificative precipitation fields.

However, before coming to the definition of the two representative power spectra, two aspects must be clarified. Firstly, since a 2D radar image is employed, also the resulting power spectrum will be 2-dimensional. To work with a more convenient 1D power spectrum, a radially averaged power spectrum density (RAPSD) algorithm, already implemented in the py-STEPS library, has been applied to the precipitation fields. Secondly, considering the entire radar domain would not be convenient for categorization purposes. In fact, on the one hand increasing the considered space could lead to the inclusion of mixed clouds that make the distinction between stratiform and convective precipitation more difficult, and, on the other hand, applying a transform over a large domain could require high computational times. However, a discrete method such as the FFT requires a significative sample to build a representative power spectrum. Therefore, in order to make a balance between these last two requirements, it has been chosen to compute the FFT over a domain of 50 x 50 Km^2 (or of 2500 pixels), which simulates the dimension of a small hydrological catchment. In fact, as reported in the introduction, one of the objectives of this work is to evaluate the performances of different nowcasting models also in smaller areas which could be interesting under a hydrological point of view. Therefore, the power spectrum and the categorization will be carried out in the subdomains of 50 x 50 Km^2 and not over the 512 x 512 radar image.

Consequently, it is possible to proceed with the regression to produce the two representative power spectra over the subdomains. Ten different power spectra have been evaluated and are reported in the picture below on a double log scale:

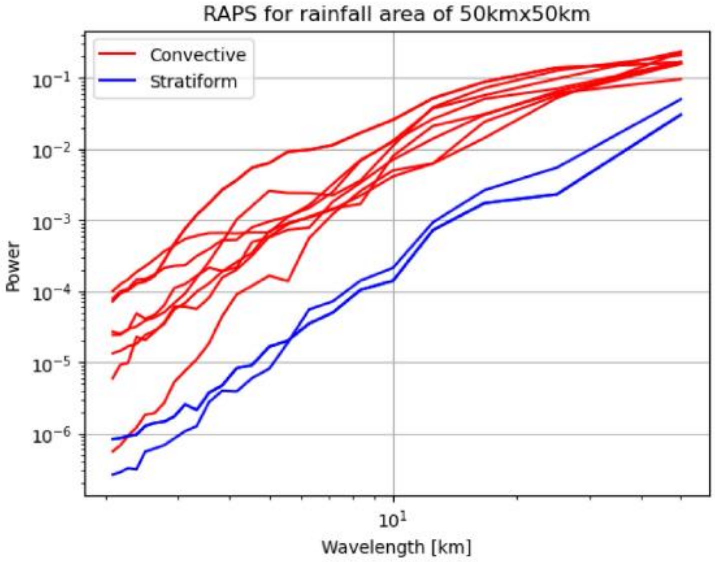


Figure 4.8. RAPSD of 10 different precipitation fields

As it can be observed, eight convective power spectra belong to convective clouds, while the remaining two are derived from stratiform clouds. It was possible to retrieve only two stratiform spectra because it was really difficult to find clouds, in the available radar images, that were purely stratiform in a spatial range of 50 x 50 Km^2. However, for the purposes of this application, the regressed power spectrum of stratiform events differs significantly from the one of convective ones.

For the sake of simplicity, the power spectra were regressed using a polynomial function, whose degree was determined according to the best possible fit. Convective spectra followed a linear regression model:

$$P = 0.037w - 0.014 \ (4.1)$$

While stratiform events were better approximated by a second-degree equation:

$$P = 2.28 * \ 10^{-5}w^2 - 0.0036w + 0.00105 \ (4.2)$$

Where P and w represent the power and the wavelength respectively. After calculating these power spectra, the algorithm works as follows.

Every time a 50 x 50 Km^2 window is applied to a portion of the image, a power spectrum is calculated and its RMSE with respect to the two equations reported above is computed; the power spectra with the lowest RMSE determines the precipitation category (namely, the algorithm returns a string that could be either “convective” or “stratiform”). Finally, in order to build a set of categorized events, the list of selected events in chapter three (whose table can be found in the appendix) has been considered. For every event in the list, the window of 50 x 50 Km^2 was centered at the station’s coordinates and the algorithm was applied for different timestamps; in other words, the algorithm has been applied every 5 minutes to the precipitation field to characterize it as convective or stratiform. If, after a certain number of timestamps and before 1 hour lead-time, the algorithm found an event that from “convective” turns into “stratiform”, that event is characterized as stratiform, conversely, it is classified as convective.

It must be noticed that all the events of the list have been classified as convective at least for the first considered timestamp; this is logical since the station list in chapter 3 was built on purpose to intercept the events with high precipitation intensities. In addition to that, it should be noticed that the provided radar observations refer only to the months of May, June, July, August and early September, that is the period in which convective events are more frequent.

Therefore, the classification of the events that has been made in this context should not be interpreted as a rigid distinction between “convective” and “stratiform” precipitation, but rather as a differentiation between events that keep a certain level of convectivity throughout their analysis and events that “degrade” into stratiform precipitation. Consequently, even if in the next sections, for simplicity, the categorized events will be referred to as “convective” or “stratiform”, these remarks should be kept in consideration.

Two distinct tables that report these kinds of events can be found in appendix B and C and will be used in the next section to provide a more in-depth analysis of the performances of the three deterministic methods. Nevertheless, one aspect related to these tables must be specified. As explained in the previous section, all the starting times of the events were shifted 20 minutes back in time to overcome the limitedness of the dataset and to better assess the growth and decay of the precipitation structures. However, in this case the time shift has not been applied, since, on average, no cloud was present in the 50 x 50 window 20 minutes before the starting time of the event and therefore it would not be meaningful to calculate verification metrics. This led to problems in retrieving, as previously explained, enough radar observations that allowed the calculation of the comparative metrics. Therefore, all the events for which not enough radar images were present, were excluded. For this reason, the number of studied convective events is 26, while the one of stratiform is 11.

4.4 ANALYSIS OF CONVECTIVE AND STRATIFORM EVENTS

In the previous section, an algorithm that characterizes the convective or stratiform nature of a precipitation field has been introduced. Thanks to its implementation, it has been possible to make a distinction between events that have a more remarked convective or stratiform nature within 1 hour lead time and whose summarizing table is reported in appendix B and C of the present work.

Now, thanks to this selection, it is possible to perform a dedicated analysis of convective and stratiform (or better, events that turn into stratiform) precipitations and to make a more precise evaluation of the performances of the three nowcast models. As reported in the introduction of the present chapter, this time the evaluation of the nowcasts will not be performed all over the radar domain, but in smaller rectangular portion of the Belgian territory. More precisely, these “small windows” of the domain have a dimension of 50 x 50 Km^2 and are built around the rain gauges that have measured the intensity of the events. This methodology has been adopted for two main reasons.

Firstly, the convectivity of every single event, as previously explained, has been determined through the algorithm inside the small windows, since it was not possible to perform this evaluation over the entire radar domain. In second instance, the objective of this section is to evaluate the performances of the nowcast models inside an area that simulates the dimension of a hydrological catchment. In fact, while literature presents a high number of forecast analyses over extended domains, there is a lack of investigations related to sub-domains of the territory. It is important to note that the size of a general hydrological catchment can vary drastically around the world, ranging from a few to millions of km^2. However, given the usual dimensions of hydrological catchments in Belgium [49], such as the Ijzer catchment (1378 km^2) and the Demer catchment (1922 km^2), 2500 km^2 appears to be a reasonable approximation for the studied domain. Therefore, the analysis of stratiform and convective events is reported below.

4.4.1 Convective events

Before presenting the results of the metrics, it is convenient to take a look at the images related to one single representative convective event in order to provide a qualitative description of the nowcast methods and their relative differences. At first, the observed precipitation field together with the motion field obtained through Lucas-Kanade algorithm is presented and then the fields forecasted by extrapolation, S-PROG and ANVIL are reported for different lead times:

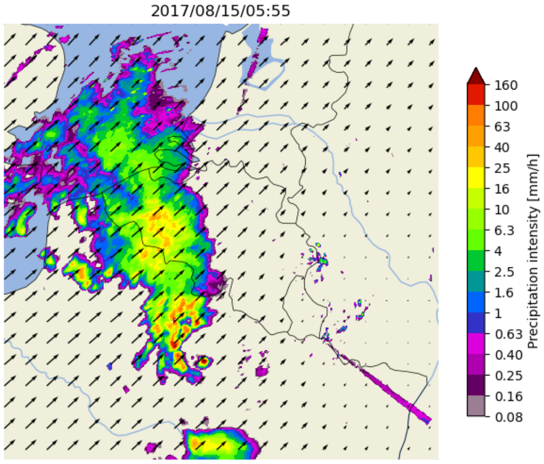
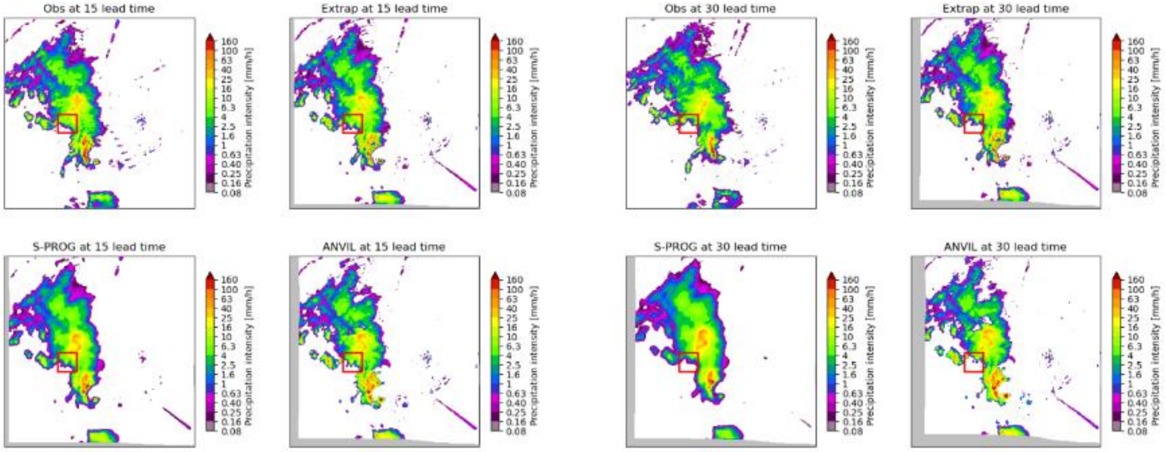
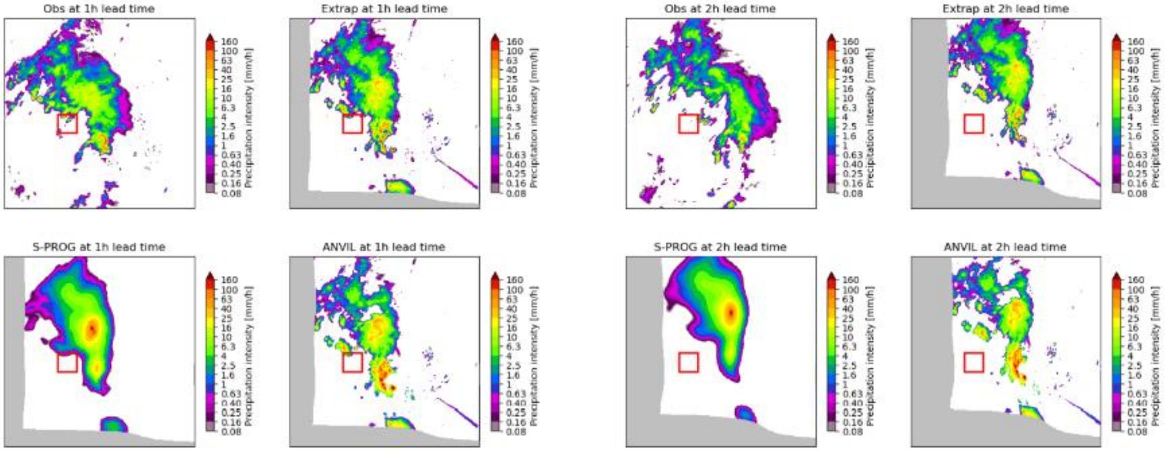


Figure 4.9. Observed convective event with Lucas-Kanade motion field.



(a) 15 min lead time (b) 30 min lead time



(c) 1h lead time (d) 2h lead time

Figure 4.10. Results of the tree different deterministic forecasts for 15, 30, 1h and 2h lead time

The red rectangle inside the different images is the simulated catchment (or "small window") inside which the forecasts' skill will be evaluated. It is possible to notice that at 1h lead time most of the precipitation field has left the area of interest due to the cloud's motion and this pattern is observed in almost all the events that have been considered. For this reason, the results of both the continuous and categorical metrics will be discussed taking into account up to 30 minutes lead time, since most of the precipitation is concentrated in these first minutes of the forecast and therefore the results are more representative. Furthermore, the image reported above suggests some interesting qualitative considerations related to the precipitation fields produced by the extrapolation, S-PROG and ANVIL.

As expected and discussed in chapter 2, the extrapolation method produces a precipitation field that does not alter the rainfall intensity and whose changes in the shape are merely dependent on the advection along the Lagrangian trajectories. However, it can be noticed that this method does not approximate badly the observed fields at least until the first hour lead time. On the other hand, the tendency of S-PROG to smooth the rainfall field is pretty evident, especially at 2 hours lead time. In fact, it can be observed that toward the end of the nowcast this method not only is prone to agglomerate all the precipitation structures in one single cloud, but also forms a concentric pattern of precipitation intensities that decrease starting from the main high-intensity core located at the center of the cloud. Finally, the precipitation field forecasted by ANVIL exhibits a shape that is more similar to the one of the extrapolation for all the duration of the event; yet, the rainfall intensities differ dramatically. It can be observed that at two hours lead time ANVIL tends to create both regions of really high intensity (that can be noticed in the lower "tail" of the precipitation field) and zones of really low or 0 intensity rainfall (this is particularly remarked in the norther part of the field). In other words, it seems that ANVIL model exaggerates both the growth and the decay of the single precipitation structures throughout the forecast when compared to the observed field and the extrapolation method.

Now, after this qualitative analysis, it is possible to carry out a quantitative study through the same continuous and categorical metrics that have been discussed in the previous sections. The MAE and Pearson correlation coefficient are reported below, using the same double conditioning of 1 [mm/h] previously seen:

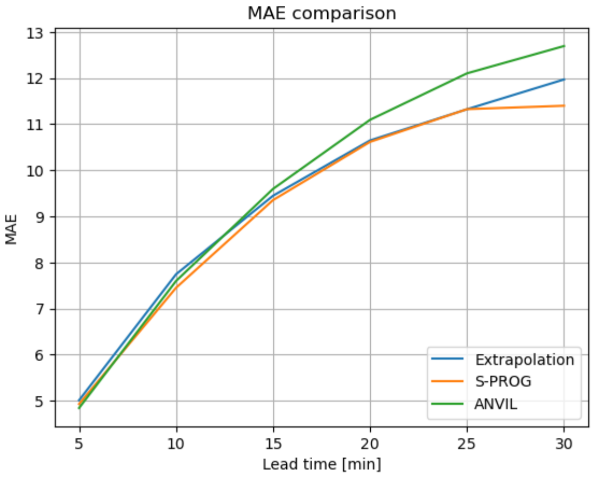


Figure 4.12. Averaged MAE over the selected convective events for the three models

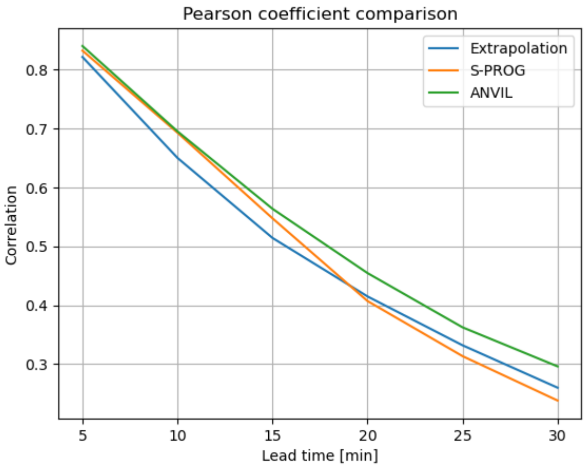


Figure 4.11. Averaged Pearson correlation coefficient over the selected convective events for the three models

These graphs provide patterns that have already been observed in the analysis computed over the entire domain, with ANVIL that is the most biased model in terms of MAE toward thirty minutes lead time, but that keeps the best correlation with respect to the observed precipitation fields. However, some differences compared to the previous cases can be reported.

In first instance, the values of Mean Absolute Error are higher than those calculated over the entire domain. In fact, while in figure 4.5 at 30 minutes lead time the MAE was about 6 [mm/h] for ANVIL and 5.5 [mm/h] for S-PROG and extrapolation, in this case values between 13 and 11 [mm/h] are reached. This was expected since convective storms are more difficult to model and predict. Secondly, it can be observed that extrapolation performs slightly worse in terms of correlation, as in the first 20 minutes lead time the Pearson coefficient is the lowest and becomes barely higher than S-PROG in the last 30 minutes.

After the continuous metrics, the categorical ones are reported in the pictures below, using precipitation thresholds of 1, 5 and 10 [mm/h]:

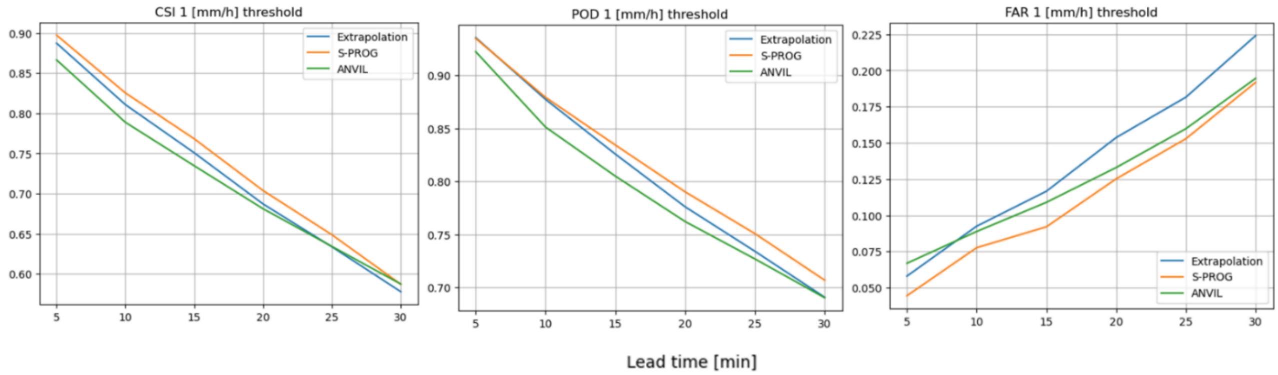


Figure 4.13. Averaged CSI, POD and FAR over the selected convective events for the three different models at 1 [mm/h] threshold

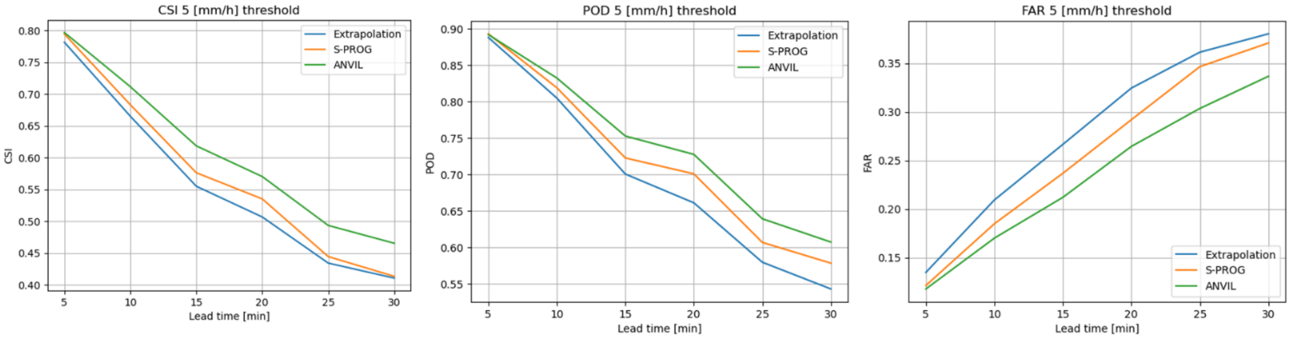


Figure 4.14. Averaged CSI, POD and FAR over the selected convective events for the three different models at 5 [mm/h] threshold

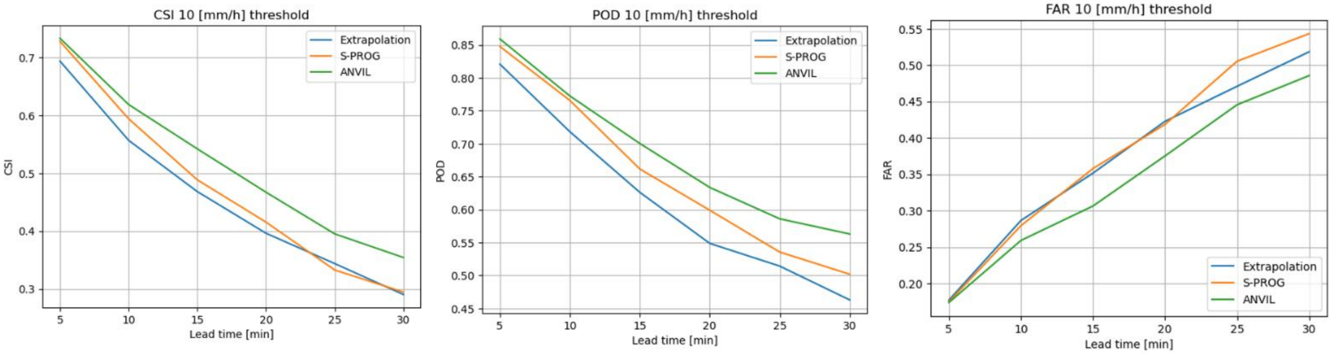


Figure 4.13. Averaged CSI, POD and FAR over the selected convective events for the three different models at 10 [mm/h] threshold

When the precipitation threshold is set to 1[mm/h], S-PROG method is clearly the most skillful in terms of CSI, POD and FAR and it is also interesting to notice that the relative values of all the three metrics are higher when computed inside the “small window” instead of all over the radar domain. However, when the precipitation threshold is set to higher values, ANVIL performs remarkably better than the other two methods and in a more definite manner when compared to the analysis over the entire radar domain. It is also worth noting that at these precipitation thresholds the extrapolation method is the one that performs the worst. Apparently, the ability of S-PROG and ANVIL to model the growth and decay of precipitation structures better suits the investigation of convective events rather than a simple extrapolation.

Nevertheless, when an analysis is performed over a hydrological catchment, one of the most relevant requirements to the models is to properly estimate the rainfall volumes. For this reason, it has been decided to report the volumes intercepted by the three models at 30 minutes, 1 hour and 2 hours lead time. This choice of lead times has been made because, if on the one hand most of the precipitation is concentrated in the first 30 minutes lead time inside the simulated catchment, on the other hand some residual rainfall at 1 and 2 hours could still be relevant when an evaluation of rainfall volumes is carried out. The obtained results are reported in the table below, where the values are calculated in percentages as: $\frac{volume\ intercepted\ by\ the\ model}{observed\ volume}\ x\ 100$

	30 lead time	**60 lead time**	**120 lead time**
Extrapolation	101%	105%	104%
S-PROG	103%	108%	110%
ANVIL	105%	112%	115%

Table 4.2. Percentages of intercepted volumes at three different lead times

First, it can be noticed that, in such sub-domains, all the three methods result in an overestimation of the rainfall volumes for all the considered lead times. As it can be seen, although the extrapolation method performs the worst for high precipitation thresholds, it is still able to intercept the volume pretty well, with an overestimation of only 4% at the second hour lead time. On the other hand, S-PROG and ANVIL present a more severe overestimation, which is probably due to the autoregressive method that tends to model a too intensive growth of rainfall intensity compared to the actual observations. Therefore, it is worth noting that, at least for these considered events, autoregressive methods such as ANVIL and S-PROG can better model the precipitation field evolution, but, due to an overestimation of the phenomena of growth and decay of the structures, could provide more biased volume estimation compared to a simpler extrapolation.

4.4.2 Stratiform events

The analysis reported in this paragraph will follow an order analogous to the one of the convective events. Also in this case, a representative stratiform event has been plotted showing the different modeling results of the three methods for two hours lead time, similarly to figure 4.10. However, since the considerations relative to extrapolation, S-PROG and ANVIL are analogous, the graph is allocated in appendix D. In addition to this, it is necessary to specify that only threshold of 1 and 5 [mm/h] have been considered, since stratiform events are characterized by less extreme rainfall intensities compared to the convective ones. Therefore, the resulting continuous metrics are reported below:

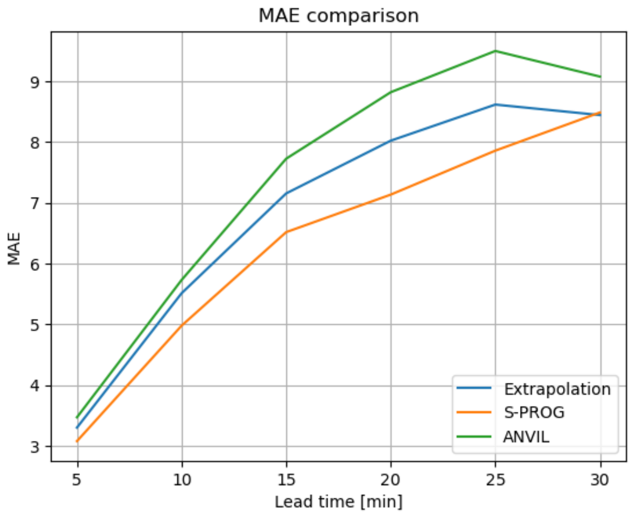


Figure 4.15. Averaged MAE over the selected stratiform events for the three models

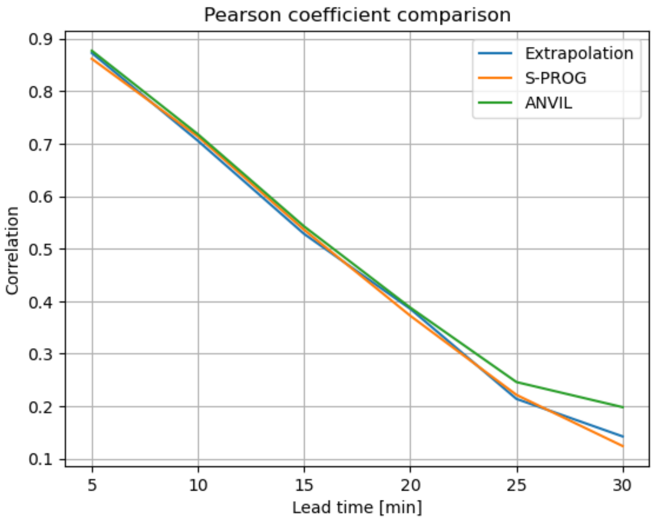


Figure 4.14. Averaged Pearson correlation coefficient over the selected stratiform events for the three models

As for the MAE graph, it can be observed that in this case the values are smaller than what has been assessed in the convective study. ANVIL is still the most biased model, but the peak value of Mean Absolute Error is slightly above 9 [mm/h] contrary to the 13 [mm/h] seen in the previous paragraph. Again, this was expected since stratiform events are easier to model than convective ones, as suggested in the previous section. In addition to this, S-PROG is remarkably the less biased model and reaches the same extrapolation value only at 30 minutes lead time, contrary to figure 4.5, where extrapolation and S-PROG method tended to have comparable results. Contrary to what has been observed in the convectivity paragraph, it is difficult this time to pick one method that offers the best correlation level, since all the three methods present practically equal values until 20 minutes lead time (after which ANVIL is proven to be slightly better). The results of categorical metrics are reported below:

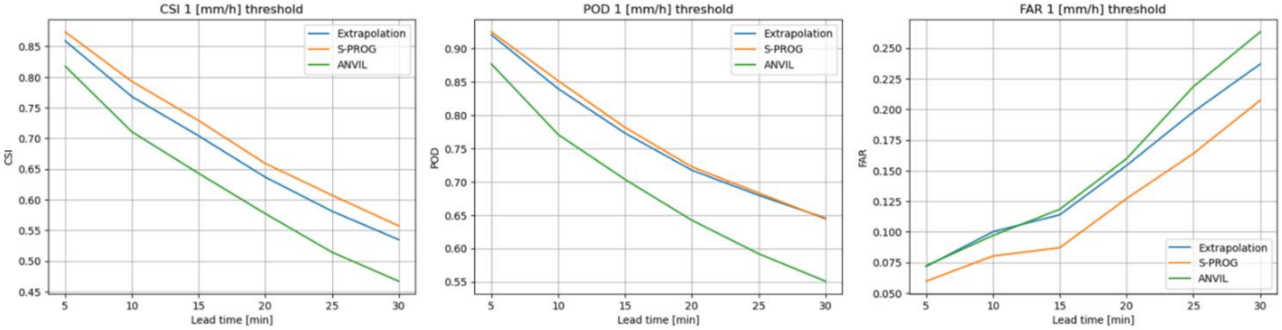


Figure 4.17. Averaged CSI, POD and FAR over the selected stratiform events for the three different models at 1 [mm/h] threshold

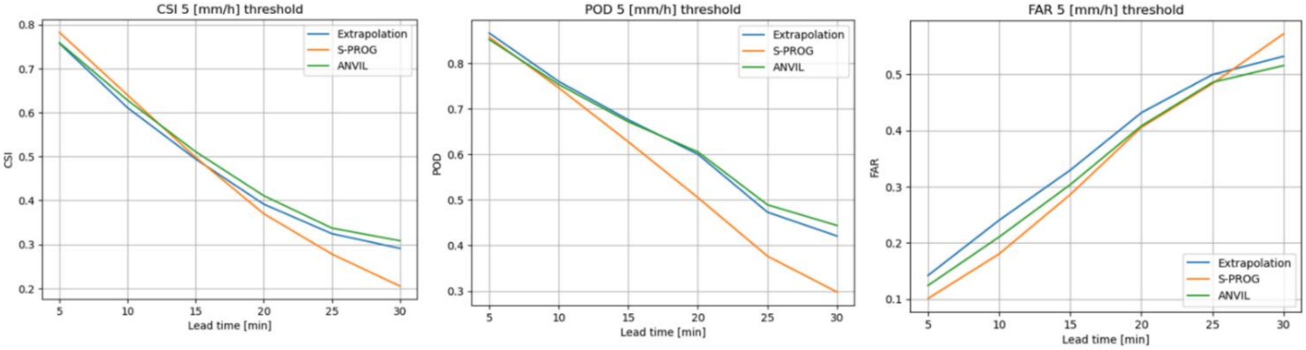


Figure 4.16. Averaged CSI, POD and FAR over the selected stratiform events for the three different models at 5 [mm/h] threshold

Similar to what has already been assessed in the previous paragraph, at 1 [mm/h] threshold the S-PROG methods proves to be the most skillful (even though it can be noticed that extrapolation offers comparable results under the POD point of view), while ANVIL offers bad performances. By increasing the precipitation threshold, extrapolation and ANVIL gain skill to the disadvantage of S-PROG; however, it is worth noting that for the study of stratiform events it is not always necessary to push the precipitation threshold above 1 [mm/h], especially if a forecaster is dealing with a persistent and low-intensity event. Therefore, on the overall, S-PROG may prove to be the most convenient methods when it comes to evaluate such phenomena.

In order to make a coherent comparison with the previous paragraph, also in this case the intercepted volumes in percentage are reported for 30 minutes, 1-hour and 2-hour lead time:

	30 lead time	**60 lead time**	**120 lead time**
Extrapolation	113%	129%	172%
S-PROG	93%	97%	108%
ANVIL	106%	111%	139%

Table 4.3. Percentages of intercepted volumes at three different lead times

For these considered events, it seems that S-PROG is the best method to intercept the rainfall volume, even though it presents a slight underestimation in the first 30 minutes and 1 hour lead time. It is also interesting to note that, contrary to what has been assessed previously, this time extrapolation considerably overestimates the rainfall volumes over the simulated catchments. Probably, the tendence of S-PROG to smooth the precipitation field has a positive effect when it is necessary to model stratiform phenomena characterized by low precipitation intensities.

4.5 CONCLUSIONS

The sensitivity analysis related to different motion fields shows that none of the algorithms clearly outperformed the others. However, Lucas-Kanade algorithm proved to be the best option in terms of computational time while still offering good results of MAE and CSI when applied to a simple extrapolation method. Therefore, this optical flow algorithm is chosen as default setting for the computation of all the nowcasts.

The analysis over the entire domain shows that ANVIL and S-PROG can perform better than extrapolation in terms of POD and CSI depending on the set precipitation thresholds. In particular, it has been observed that for the smallest precipitation threshold (1 [mm/h]) S-PROG can provide better results, while ANVIL shows better skills for the highest threshold (10 [mm/h]). This implies that S-PROG is better at intercepting smaller precipitation values but has more problems when it comes to evaluate the higher intensities. Conversely, ANVIL proved to be more skillful in forecasting higher precipitation intensities, but in some ways fails at intercepting the lighter rainfall. However, the analysis over the entire radar domain also showed that the results obtained through these two models are not so different with respect to extrapolation method, which surprisingly did not provide bad performances, especially in terms of MAE. Yet, as it has been clarified in section 4.2, several reasons could explain these outcomes, such as the limitedness and the scarce variety of the radar dataset or the lack of a categorization among the different events.

In order to make a categorization of the events, an algorithm based on power spectra and implemented in Python allowed to make a distinction between more convective and stratiform events. This opened to the possibility of making more specific analysis and to focus on smaller portions of the radar domain, which could be an interesting perspective for hydrological applications. However, as also clarified in section 4.3, the events classified as "stratiform" should not be considered as completely stratiform clouds, but more as convective precipitation that turns into stratiform clouds characterized by lighter precipitation.

The analysis in the smaller portions of the radar domain (namely, the simulated hydrological catchments) highlighted how ANVIL, despite being the most biased method in terms of MAE, has the best predictive ability for convective events, since it provided the highest categorical results for 5 and 10 [mm/h] thresholds. However, as it has been underlined in figure 4.10, it must be taken into account that this method has a certain tendency to exaggerate local zones of growth and decay, and this could lead to overestimation of the cumulated rainfall volumes (as reported in table 4.2). It is worth to notice that, despite extrapolation got worse results in terms of CSI, POD and FAR, it showed the best estimation of rainfall volumes for these kinds of events.

On the one hand, S-PROG better models the convective events that are turned into stratiform and that keep low precipitation intensities. Probably, this is the result of the tendency of this method to smooth the precipitation field, which better suits the application to more stratiform events. In addition to this, it has also been noticed that S-PROG was the best model to intercept the cumulative rainfall volume inside the sub-domains when compared to ANVIL and extrapolation, that in this case provided the worst estimation.

On the overall, it is necessary to make a balance among the three methods. Eventually extrapolation did not provide extremely biased results, especially during the evaluation over the entire radar domain. However, when it is necessary to carry out analysis over a smaller sub-domain, it is better to use methods such as S-PROG or ANVIL that are able to better simulate the growth and decay of local precipitation cells.

After all these examinations, the suggestion for forecasters should be to use ANVIL to evaluate more convective situations that are characterized by higher intensities. However, if more stratiform events are present and it is not necessary to set high precipitation thresholds, S-PROG could provide better results. In any case, when applying ANVIL, it should be kept into account that this model was found to be the most biased in terms of MAE and that it tends to exaggerate the growth and decay of precipitation cells, which could lead to overestimations in the hydrological balances. On the other hand, S-PROG has the tendency to smooth the precipitation fields, therefore it could be more suited for stratiform events or convective precipitations that degrade fast into stratiform ones, but it is not well suited for modeling events characterized by high precipitation intensities for long lead times.

Chapter 5: PROBABILISTIC NOWCAST

The nowcast methods presented in the previous chapter adopted a deterministic approach: namely, only one precipitation field was produced for each lead time of the forecast. Despite certain advantages provided by these models, such as low computational time and clear interpretation, deterministic outputs have the notable disadvantage of not accounting for uncertainty. In fact, as pointed out in literature [50], the forecast's uncertainties remarkably increase throughout the duration of the events and can affect the estimations of relevant hydrological quantities, like the cumulative rainfall volumes, which are critical for flood control systems. As a result, in recent years, the development of probabilistic models such as py-STEPS and STEPS-BE, which can give uncertainty estimation, has gained increasing interest in hydrological applications.

As noted in Chapter 2, these models can provide an ensemble of forecasts for each lead time using a perturbation scheme based on correlated noise, thus representing the uncertainty associated with the growth and decay of smaller precipitation structures.

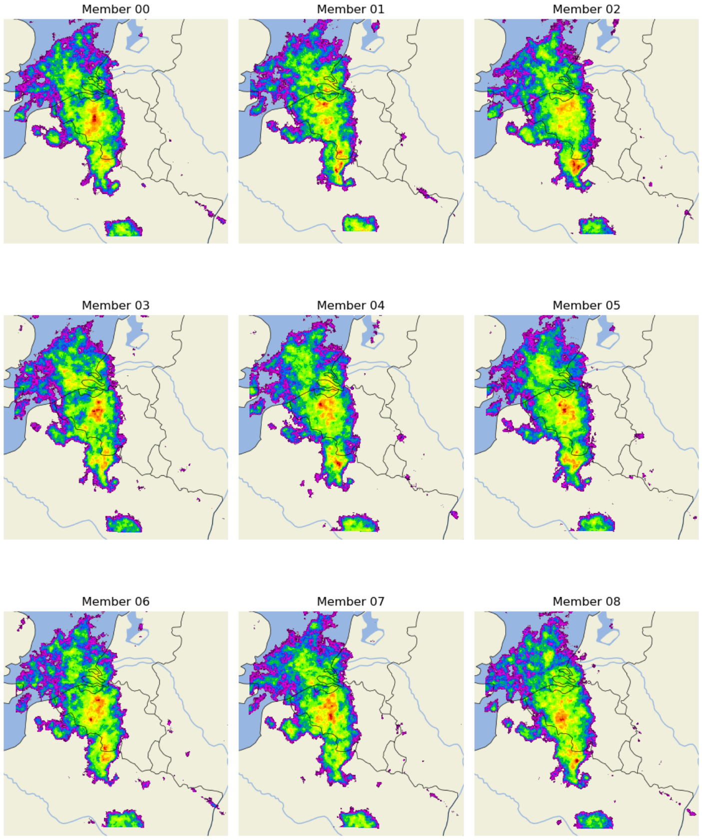


Figure 5.1. Representation of 9 ensemble members produced by py-STEPS model at 30 minutes lead time for the convective event of 8/15/2017 at 5:55 a.m

In the image above, the ensemble members provide different locations of the main centers of convectivity (indicated by the red color) and forecast different rainfall intensities across the precipitation field. It is also important to remember that probabilistic methods can add stochastic perturbations to the velocity vectors to better simulate the unpredictable evolution of the motion field. As a result, when compared to deterministic techniques, the combination of these aspects gives a tool for better addressing some of the uncertainties that arise during a nowcast.

Another great advantage of probabilistic forecasts is the ability to calculate exceedance probability maps for different precipitation thresholds across the entire radar domain, as represented in the following figure:

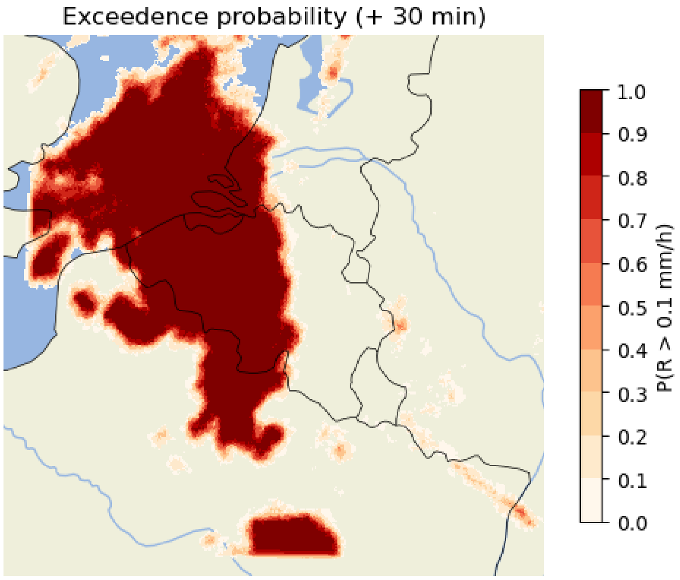


Figure 5.2. Exceedance probability for 0.1 [mm/h] threshold in the previously presented event at 30 minutes lead time

The exceedance probabilities in every pixel are calculated as the relative frequency of members that predict a precipitation intensity over the selected threshold. These exceedance probabilities can be exploited to calculate ensemble metrics such as the ROC curves by setting a defined range of probability thresholds.

The objective of this chapter is to compare the two probabilistic models py-STEPS and STEPS-BE, whose analogies and differences have already been discussed in the literature review. Indeed, it is of interest to discover whether an open-source software like py-STEPS, built by a Python community, can outperform or, at least, provide results comparable to those of a private software developed at the RMI like STEPS-BE. Given the increasing availability of radar images, the proven reliability of py-STEPS could potentially play an important role in the research of hydrologists and meteorologists, who could have access to free stochastic nowcasts without requiring any license from third parties.

To compare these two probabilistic models, three ensemble metrics have been used: ROC curves, reliability diagrams and rank histograms. The comparison of these metrics, that will be further analyzed in the next sections, provides general information about the differences in predictive ability, calibration and rank-ordering skill of the two models. Furthermore, two of the deterministic metrics used in the previous chapter, the CSI and the MAE, will be evaluated using the ensemble mean to see if the addition of stochastic noise and motion field perturbation improves the level of agreement between the observed and forecasted precipitation fields. Similarly to the previous chapter, the comparison between the two models will be carried out at two different scales: over the entire radar domain and inside the simulated hydrological catchments, where the same convective and stratiform events reported in the appendix will be considered. Finally, the ability of the models to intercept the observed cumulative rainfall volumes will be evaluated, adding the estimation of model's uncertainty.

However, before the comparative validation of the models, it has been decided to perform a sensitivity analysis on py-STEPS considering two important parameters: the number of ensemble members and the stochastic noise generator. This analysis will be carried out by assessing the results of the probabilistic metrics, to test which configuration of parameters should be used to run the model and compare its results with STEPS-BE.

It is important to note that, while the py-STEPS nowcasts were generated in Python, the STEPS-BE forecasts were delivered directly by the RMI; thus, no sensitivity analysis or control over the model's parameters was possible. For example, the approach used to generate the motion field in STEPS-BE is known as COTREC-scheme and is based on the same premise as VET in the py-STEPS library. The stochastic noise generator used to generate the ensemble of predictions is parametric, and the perturbation of the motion field consists merely of accelerating and decelerating the velocity vectors by multiplying them by a random factor. This is a significant distinction from py-STEPS, in which the motion field changes both direction and intensity. It's also worth noting that STEPS-BE makes use of an AR(1), whereas py-STEPS makes use of an AR(2).

Therefore, this chapter will be organized as follows. The sensitivity analysis will be discussed in section 5.1. The comparison of py-STEPS and STEPS-BE over the whole radar domain will be shown in section 5.2. The same analysis will be given in section inside the simulated catchments in section 5.3. Finally, concluding remarks will be made.

5.1 SENSITIVITY ANALYSIS

In this section, a sensitivity analysis on the stochastic noise generator and the number of ensemble members will be presented considering probabilistic metrics. It is important to specify that these metrics are calculated as the mean of all the events in table 1 (of the appendix) by setting a precipitation threshold of 5 [mm/h]. Furthermore, the present analysis considers the entire radar domain, and it is then not focused on the simulated hydrological catchments.

5.1.1 Stochastic noise generator

The generation of stochastic noise is one of the key points of the probabilistic STEPS nowcast. In fact, the noise cascade produced in this passage is added to the results of the autoregressive filter as perturbation to obtain the ensemble of precipitation fields for each forecast's lead time. However, applying a completely random noise (or white noise field) to the different cascade levels could be detrimental, since heavy rainfall could be formed in subdomains where actually light precipitations are observed and vice versa. Consequently, it is evident that the created stochastic noise must keep a certain level of correlation with the last radar images to produce realistic results.

Different methods are implemented in py-STEPS to produce a correlated stochastic noise and they are all based on filtering a white noise field in frequency domain after performing an FFT. In the context of py-STEPS, these methods are called "generators" and three options have been investigated in this sensitivity analysis: a parametric, non-parametric and SSFT (Short Space Fourier Transform) generators. A parametric generator filters the white noise field exploiting the parametrized slope of the RAPS of the last radar image; on the other hand, the non-parametric generator directly employs the power spectrum of the last radar image as a filter. The SSFT generator works similarly to the non-parametric one, but the filtering is applied by subdividing the domain in smaller portions rather than directly computing the PS of the entire image at once.

Theoretically, one of the advantages of non-parametric generators is that they can represent the formation/dissociation of anisotropic structures, while parametric generators can only produce isotropic structures, as shown in the image below:

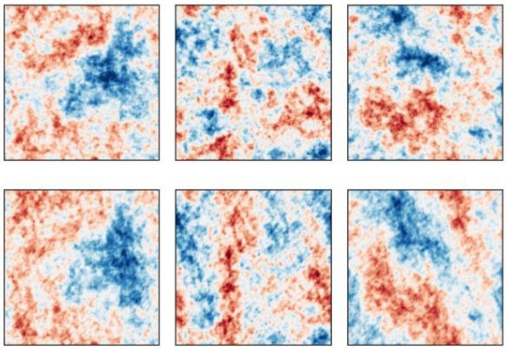


Figure 5.3. Example of noise generated with a parametric method (upper row) and a non-parametric one (lower row).

To quantitatively compare the three different methods, the average ROC curves and their AUC (Area Under the Curve) are reported in the figures below, together with the reliability diagrams:

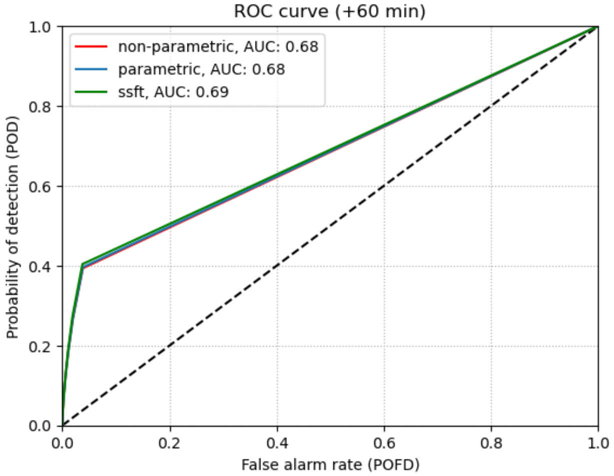


Figure 5.5. Average ROC curves considering the three different stochastic generators

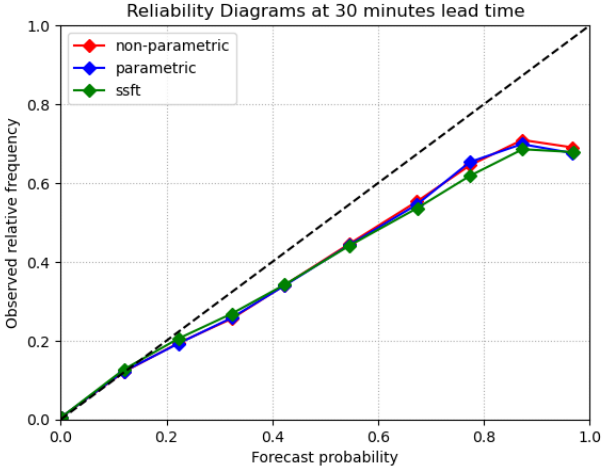


Figure 5.4. Average Reliability Diagrams considering the three different stochastic generators

It is possible to notice that, for the considered events, the choice of one method over another does not significantly improve the skill of the nowcasts. Therefore, the mean computational time (for all the events) has been considered the most relevant factor to determine the default generator for the following nowcasts. The summarizing table reported below shows that the non-parametric generator is the fastest method, while the SSFT requires high computational times (over 5 minutes); despite a small gain of AUC, this computational time is not worth the benefit.

It is important to underline that the calculations have been performed using the same processor described in table 4.1:

Stochastic generator	Computational time [s]
Parametric	206
Non-parametric	193
SSFT	312

Table 5.1. Mean computational time of the three different stochastic generators

After these results, it has been decided to adopt the non-parametric stochastic generator as default setting for all the following nowcasts.

5.1.2 Number of ensemble members

The number of ensemble members has a significant impact on the forecast's accuracy. In py-STEPS, each member of the probabilistic nowcast is generated as a new precipitation field distinguished by noise from the other outcomes. As a result, increasing the number of members should improve the ability of the model to represent a greater variety of outcomes for each lead time. However, one interesting question could be how many members are effectively needed to reach a satisfactory level of representativeness; indeed, there is no certainty that having more members leads to a greater spread of the ensemble. To test this, the following ROC curves and rank histograms are reported:

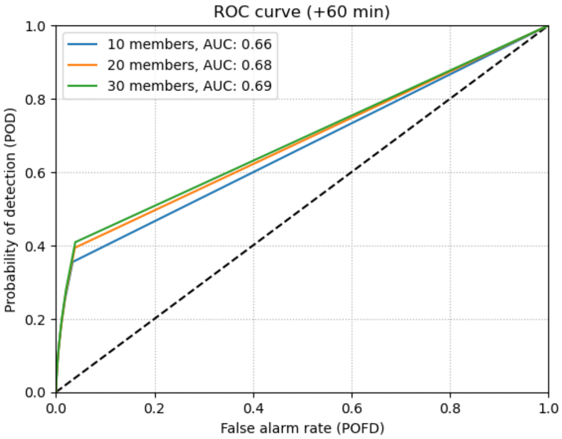


Figure 5.6. Average ROC curves considering 10, 20 and 30 ensemble members

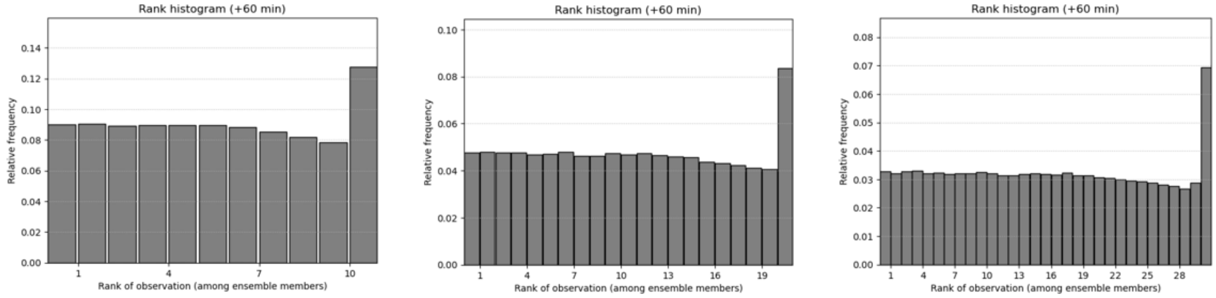


Figure 5.7. Avergae rank histograms considering 10, 20 and 30 members

The ROC curves underline that there is a small gain in the forecast’s skill by increasing the number of ensemble members, even if this improvement is more remarked when skipping from 10 to 20 members rather than from 20 to 30. Yet, the rank histograms show another relevant aspect, that answers the previous question. In fact, in all the graphs above the last bin is always the highest one, meaning that the model’s ensemble fails in representing the highest precipitation rates (namely, the probabilistic STEPS slightly underestimates the precipitation extremes).

Despite the higher number of nowcasts per lead time, the tendency in all the rank histograms above is the same, thus proving that the ensemble spread does not remarkably vary when either 10 or 30 members are produced, even though a small decrease of missing frequency is noticeable in case of 30 members. Furthermore, a higher number of ensemble members increases the computational time, a critical parameter in operational contexts. In the table below, computational times of nowcasts produced with the non-parametric noise generator are reported considering the usual Intel Core i7 – 10510U CPU with 4 cores running at 2.30 GHz:

N° of ensemble members	Computational time [s]
10	110
20	193
30	308

Table 5.2. Average computational times for producing 10, 20 and 30 ensemble members

Eventually, the choice of 20 ensemble members offers a tradeoff between a low computational time and a satisfactory gain in forecast's skill when compared to 10 members case. It is evident that the small increase in AUC of 30 members is not well balanced by the required computational time. Therefore, from now on all the nowcasts will be computed considering 20 ensemble members produced with a non-parametric noise generator. This choice also guarantees a fairer comparison with STEPS-BE, which produces 20 ensemble members as well.

5.2 COMPARISON OVER THE ENTIRE RADAR DOMAIN

In this part, the two probabilistic models will be compared across the whole radar domain. To be more specific, the primary goal is to explore the features of both the overall ensemble and the ensemble's mean.

Probabilistic metrics such as ROC curves, reliability diagrams and rank histograms provide information about the ensemble. ROC curves represent the ability of all members to correctly predict the occurrence of an event in each pixel of the domain, therefore they indicate the ability of the model as a warning system. Reliability diagrams display the observed frequency against the forecast's probability considering different probability intervals. Consequently, they measure the calibration of the model: in a well calibrated model, the observed frequencies should perfectly match the probabilities calculated through the ensemble (namely, the diagram should approximate the bisector). Finally, rank histograms express the rank-ordering skill of the ensemble. In this case, the precipitation intensities of the ensemble are ranked in ascending order and then every rank is filled with the observed relative frequency. As a result, if the ensemble is capable of entirely covering the whole precipitation range in the observed radar image, all of the bins of the rank histogram should align along a horizontal line.

On the other hand, by taking the mean of all the ensemble members, these probabilistic models can be converted into deterministic ones. Obviously, it is no longer possible to calculate an exceedance probability or estimate the uncertainty of a nowcast in this manner, thus certain advantages are lost. It should be noted, however, that the ensemble mean of these probabilistic models does not exactly correspond to the outcomes of deterministic forecasts like S-PROG or ANVIL. In fact, deterministic forecasts are generated by applying only an AR(2) to different cascade levels, whereas probabilistic models include a stochastic perturbation to both the decomposed rainfall field and the motion field. As a result, it can be worthwhile to investigate if noise has any beneficial or negative effects on the averaged rainfall field. To test this, the MAE, the CSI and the cumulated rainfall volume (compared to radar observations) will be presented and analyzed.

5.2.1 Probabilistic analysis

Firstly, the results of ROC curves for the two models are reported at 30, 60 and 120 minutes lead time considering a precipitation threshold of 5 [mm/h], that will be used to calculate also all the other probabilistic metrics.

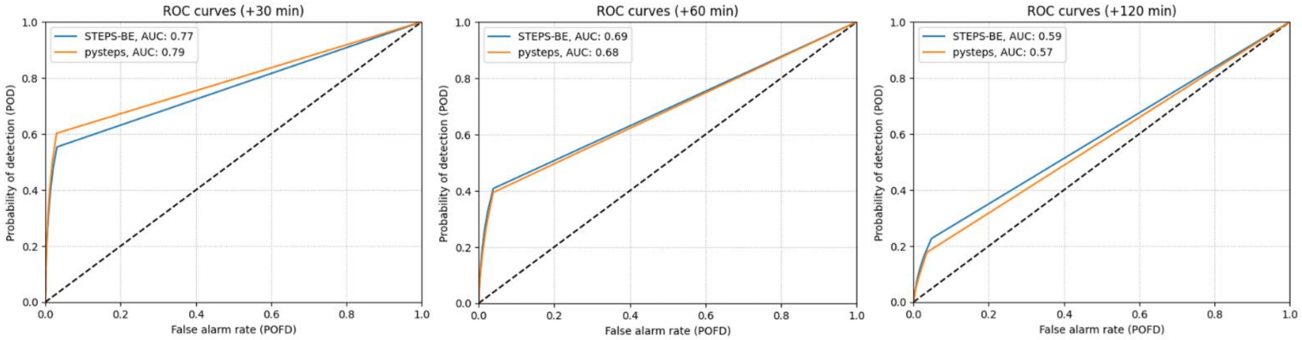


Figure 5.8. Average ROC curves of py-STEPS and STEPS-BE for 30, 60 and 120 minutes lead time

At the beginning of the nowcast py-STEPS has a higher skill than STEPS-BE, that, however, shows a higher AUC at two hours lead time. Yet, the differences between the two ROC curves are not so remarkable to establish the superiority of one model over the other. Consequently, the analysis of the reliability diagrams is reported considering the same lead times:

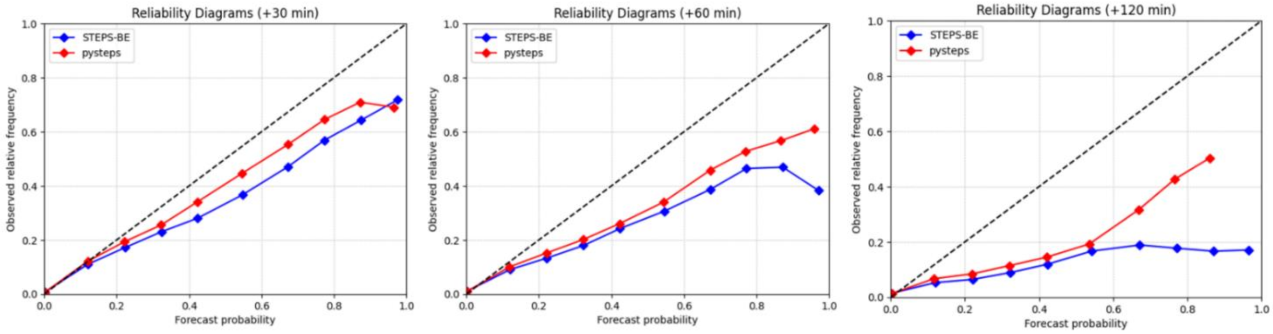


Figure 5.9. Average reliability diagrams of py-STEPS and STEPS-BE for 30, 60 and 120 minutes lead time

In this case the results of py-STEPS are clearly better than those offered by STEPS-BE, since its reliability diagram is closer to the bisector for all the lead times. However, it can be observed that both models exhibit a reliability diagram below the bisector, which is a symptom of overconfidence of the ensembles. Namely, the models forecast events with a probability that is higher than the observed frequency.

Finally, the results of the rank histograms are reported in the image below:

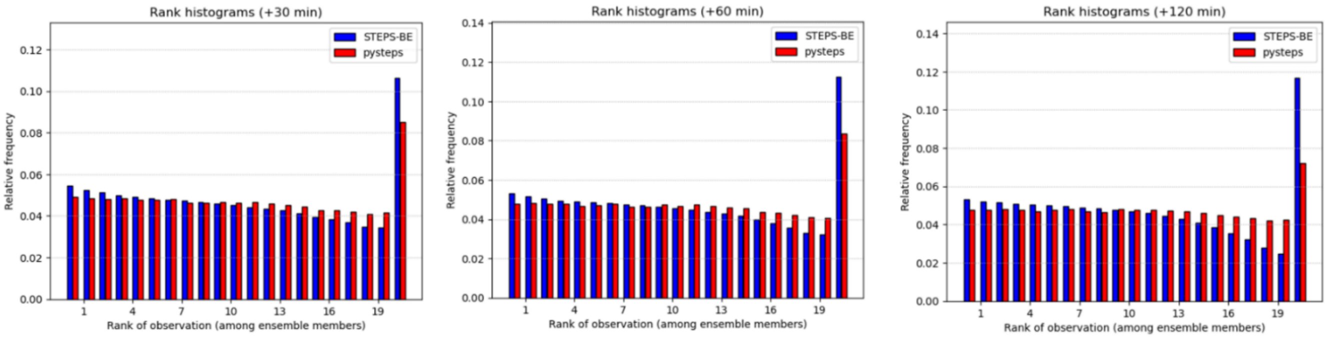


Figure 5.10. Figure 48. Average reliability diagrams of py-STEPS and STEPS-BE for 30, 60 and 120 minutes lead time

As pointed out in the previous section, a rank histogram where the last bin has a high frequency indicates that the model underestimates the highest precipitation rates. From the picture above it is possible to notice that both py-STEPS and STEPS-BE have this flaw. However, the rank histogram associated with py-STEPS has more aligned bins compared to the one of STEPS-BE, indicating that its ensemble is better spread and represents more accurately smaller rainfall intensities compared to STEPS-BE. In addition to this, it can also be noticed that the frequency of the last bin of py-STEPS decreases when the lead time increases. This result implies that the model underestimates the highest rainfall intensities less near the end of the forecast, when precipitation intensities decrease due to the decay of the most intense and convective precipitation structures. On the other hand, this tendency cannot be observed in the case of STEPS-BE, which increases the frequency of the last bin as the nowcast reaches two hours lead time. This comparison suggests that py-STEPS ensemble not only is more spread, thus better capturing the range of rainfall intensities over the domain, but also is more accurate that STEPS-BE in representing the decay of precipitation structures toward the end of the event.

5.2.2 Deterministic analysis

A qualitative assessment of some plotted precipitation fields can be presented to get a sense of the noise influence on the averaged rainfall field.. In the image below, the fields on the left refer to the observations, while the ones on the right are produced using py-STEPS for 5, 40 and 120 minutes lead time:

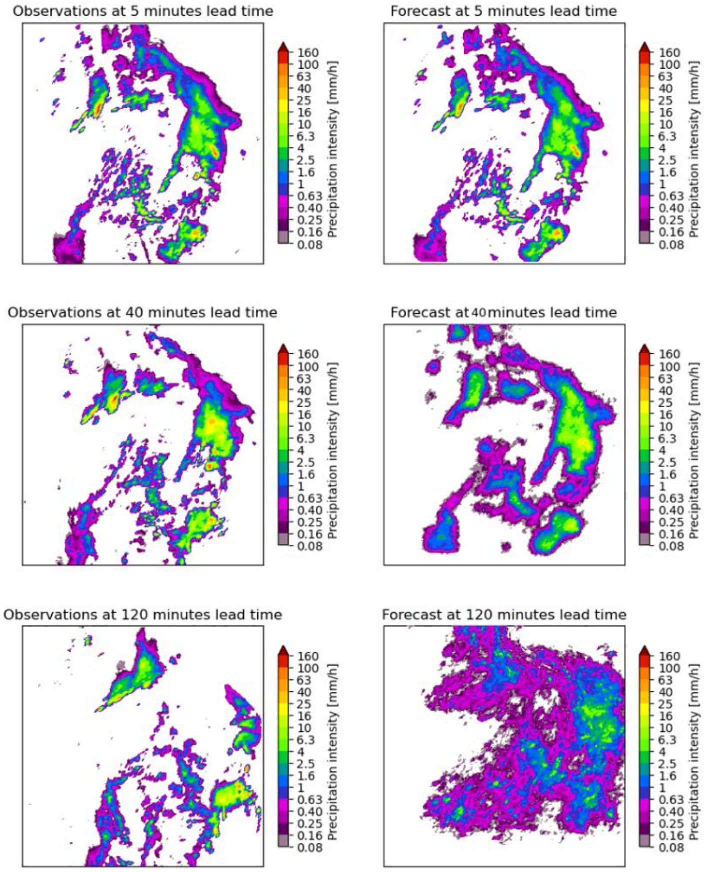


Figure 5.11. Plotted precipitation fields at 5, 40 and 120 minutes lead time. On the left the measured precipitation field by the radar, on the right the ensemble's mean produced by py-STEPS

Firstly, it can be observed that the shape of the ensemble's mean is pretty similar to the one of the observed precipitation field, also at two hours lead time. This could be due to the stochastic perturbation of the motion field, that not only accelerates/decelerates the velocity vectors, but also changes their direction. In general, it could be said that a perturbed motion field is more realistic than the static one implemented in deterministic methods. Moreover, it is interesting to notice the low rainfall intensity contours created around the main precipitation structures in the ensemble's mean, especially at two hours lead time. These shapes are due to the noise characterizing every member of the ensemble and are much more remarked than in the case of purely deterministic precipitation fields, as it can be observed in comparison to figure 4.10. Therefore, this picture underlines some qualitative differences with respect to ANVIL or S-PROG.

To compare the results of the two models, the following graphs are reported:

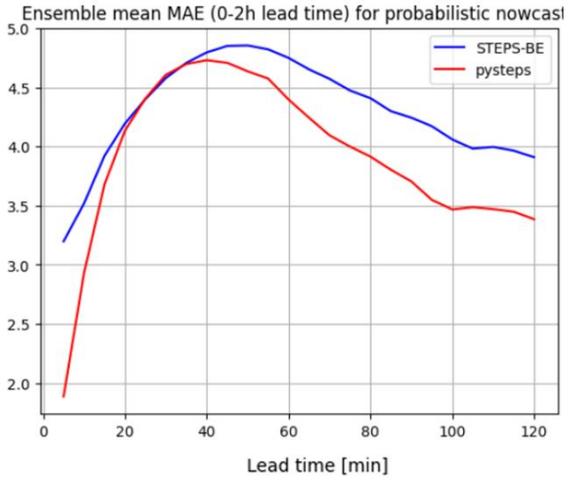


Figure 5.13. Average MAE curves of py-STEPS and STEPS-BE

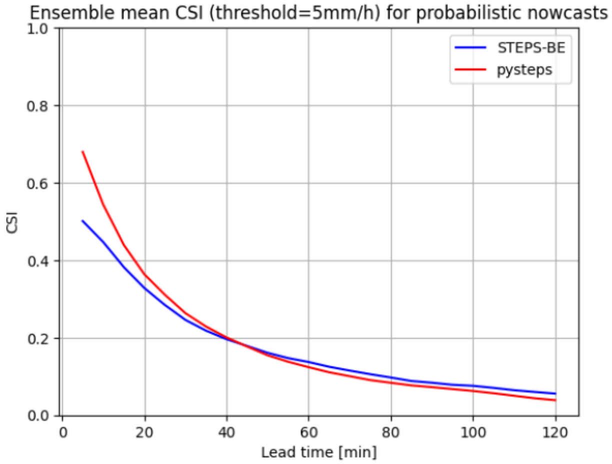


Figure 5.12. Average CSI curves of py-STEPS and STEPS-BE

Considering the MAE curves, it seems that py-STEPS performs better than STEPS-BE for almost all the lead times (except for an overlapping region between 20 and 40 minutes). STEPS-BE fields start with an higher error than py-STEPS, probably due to the different implemented optical flow algorithm and to the AR(1) that may be less effective than an AR(2). Furthermore, both models show a curious decrease of the MAE after 40 minutes lead time that is not visible when dealing with deterministic models (compare with figure 4.5). Recalling the previous considerations, this can be explained by the perturbed (and more realistic) motion field and by the noisy precipitation structures that increase the correlation level between forecasts and ensemble mean.

As for CSI values, py-STEPS outperforms STEPS-BE at least for the first 40 minutes lead times. After this time, STEPS-BE presents slightly better results than py-STEPS, but the differences are not remarkable. This tendency is observed also considering other precipitation thresholds (1 and 10 [mm/h]) and therefore are not reported here.

Finally, the cumulative rainfall volumes are reported in the graph below. When considering the entire radar domain, both models tend to underestimate the cumulative volumes. In this case, however, STEPS-BE outperforms py-STEPS, which has a more severe underestimation. Despite this, the forecasted volume of both models does not differ much from the observations, indicating that both models are well suited to hydrological balances. In fact, by the end of the nowcast, STEPS-BE intercepted 98% of the rainfall volume, while py-STEPS intercepted 90%.

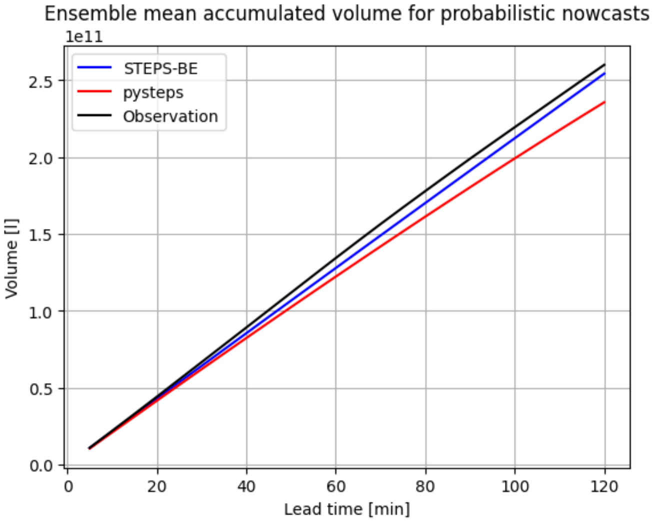


Figure 5.14. Cumulative rainfall volume comparison

Figure 5.14 shows an interesting linear trend of the cumulative rainfall volume. This can be explained considering that this analysis is extended to the entire radar domain, where local growth and decay certainly occurs, but in only two hours lead time the average precipitation intensity remains constant. Consequently, the cumulation of rainfall provides a sort of linear trend. However, this does not occur in the investigation inside the sub domains, where the smaller scale features are much more relevant. In fact, when a fixed area of interest is considered, the precipitation is usually intense in the first minutes of the nowcast, while after some time most of the precipitation field has moved and/or decayed, leaving only some residuals of rainfall. In these cases, as will be observed in the next graphs (5.22 and 5.23), the cumulative rainfall volumes follows a logarithmic trend rather than a linear one.

5.3 COMPARISON IN SUB-DOMAINS

In this section the comparative analysis will be performed inside the simulated hydrological catchments. In complete analogy with chapter 4, the same convective and (partly) stratiform events will be evaluated considering some of the probabilistic and deterministic metrics presented in the previous section. However, in this case the results will be presented considering 1 hour lead time to better underline the differences between py-STEPS and STEPS-BE. As already explained in the previous chapter, it is not useful to make evaluations for longer lead times inside the subdomains because the majority of the rainfall has already left the area of interest within 1 hour.

In the probabilistic analysis, the reliability and rank histograms will not be reported, since they provided similar results to ones seen in the previous section. Moreover, instead of building the ROC curves for different lead times, the AUC values will be plotted for the entire duration of the nowcasts. The deterministic analysis will discuss the MAE and CSI considering also the results provided by the three deterministic models presented in the previous chapter (extrapolation, S-PROG and ANVIL). Finally, instead of just computing the ensemble's mean, the cumulative rainfall volumes of the two models will be compared in the hydrological catchments reporting the uncertainties of the nowcasts as well.

5.3.1 Probabilistic analysis in sub-domains

The analyses conducted in the small sub-domains produced the following AUC graphs by setting a precipitation threshold of 5 [mm/h]:

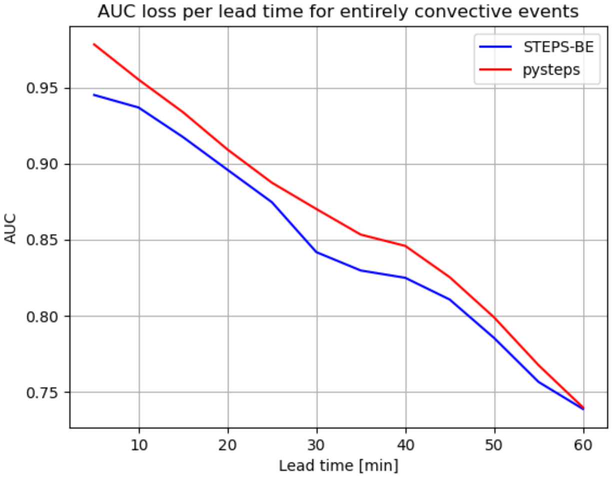


Figure 5.16. AUC loss per lead time for convective events

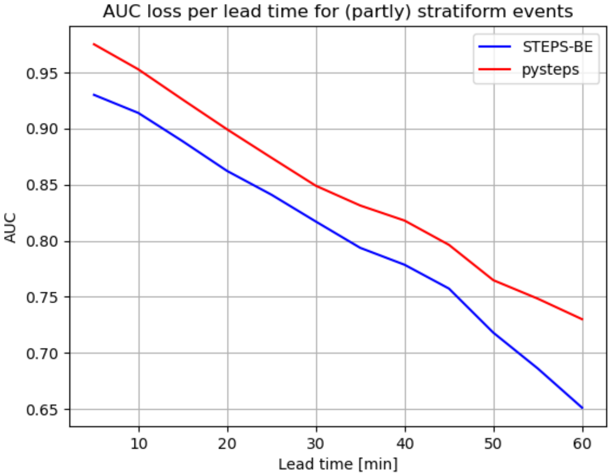


Figure 5.15. AUC loss per lead time for stratiform events

Firstly, it can be noticed that py-STEPS outperformed STEPS-BE in both convective and stratiform event analysis. This implies that, at smaller scales, the former method always functions as a better warning system. It is also interesting to notice that the values of AUC at 30 and 60 minutes lead times are always higher than those found in the analysis over the entire radar domain for both methods; for example, at 60 minutes lead time, py-STEPS had a value of 0.68, compared to a value of roughly 0.75 found in both stratiform and convective events in the sub-domains (compare with ROC curves in figure 6.8). This gain of skill at smaller scales is analogous to the increase of CSI and POD values observed for deterministic methods in chapter 4 during the study of convective and stratiform events. Therefore, this is an encouraging result for hydrological investigations, where the main interest does not concern the entire radar image but only smaller catchments.

Furthermore, it can be noticed that surprisingly the AUC values of STEPS-BE at 60 minutes lead time are higher in convective events than in stratiform ones. Yet, this can be explained recalling the considerations made during the development of the algorithm to categorize the events (section 4.3), where it has been explained that these “stratiform” events actually start as convective precipitations that “degrade” fast into stratiform rainfall. Consequently, these results suggest that STEPS-BE is less suited to represent rapid decay of smaller precipitation structures than py-STEPS, probably because the latter implements an AR(2) that is more complex than an AR(1).

5.3.2 Deterministic analysis in sub-domains

Firstly, the MAE and CSI results related to convective events are reported considering 5 and 10 [mm/h]:

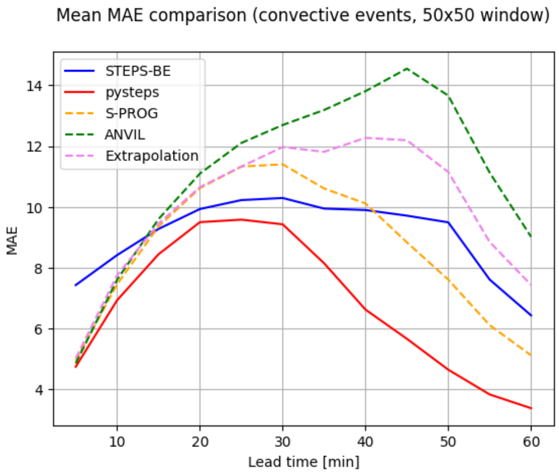


Figure 5.17. Average MAE for convective events

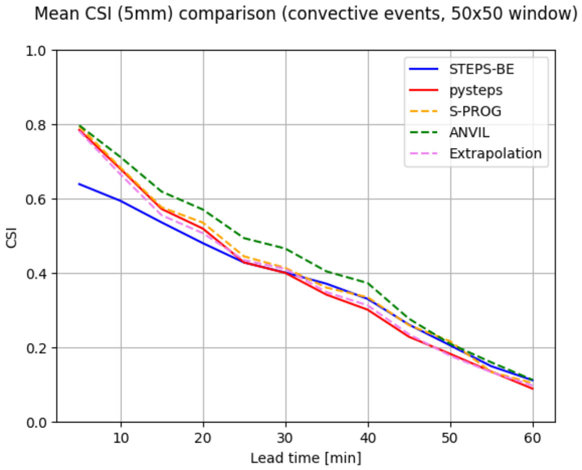


Figure 5.18. Average CSI for 5 [mm/h] thershold of convective events

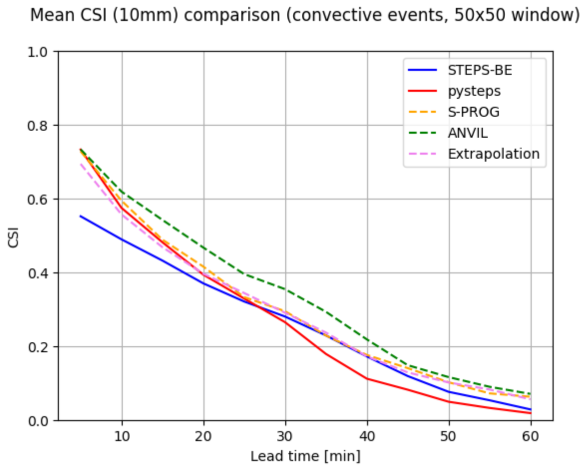


Figure 5.19. Average CSI for 10 [mm/h] thershold of convective events

In figure 6.18, at the beginning of the nowcast all the models increase their relative MAE, but after 30 minutes they all show a drop at different lead times. This tendency is due to the fact that at 1 hour lead time almost all the precipitation has left the area of interest leaving just a smaller amount of lower intensity rainfall, resulting in the observed decrease in MAE. In chapter 4 this drop did not occur because the analysis considered the first 30 minutes lead time to better show the differences among deterministic models. As previously observed, ANVIL has the highest error in terms of MAE, while the ensemble's mean of py-STEPS presents the best results. Also STEPS-BE provided a good outcome in comparison with deterministic models, even if, as already pointed out, it always starts with a relatively high MAE at the beginning of the nowcast and becomes less skillful than S-PROG after 40 minutes lead time. A possible explanation for this bias in the first minutes lead time could be that STEPS-BE exploits an AR(1) and implements its own optical flow algorithm (COTREC scheme), while the py-STEPS models have been initialized with the same AR(2) filter and optical flow algorithm (Lucas-Kanade) and thus they all start with the same MAE value. Yet, from this discussion it is possible to conclude that the ensemble's mean provides less biased results than a simpler deterministic precipitation field.

On the other hand, the CSI graphs for high precipitation thresholds (5 and 10 [mm/h]) show that ANVIL still performs better than all the other models, including the ensemble's mean of both py-STEPS and STEPS-BE. Therefore, it can be concluded that averaging the members of the probabilistic nowcasts does not provide any benefit in terms of warning systems.

The following graphs refer to the stratiform events. In this case, the CSI is calculated only for 1 [mm/h] threshold since these events, on the overall, are characterized by lower intensities:

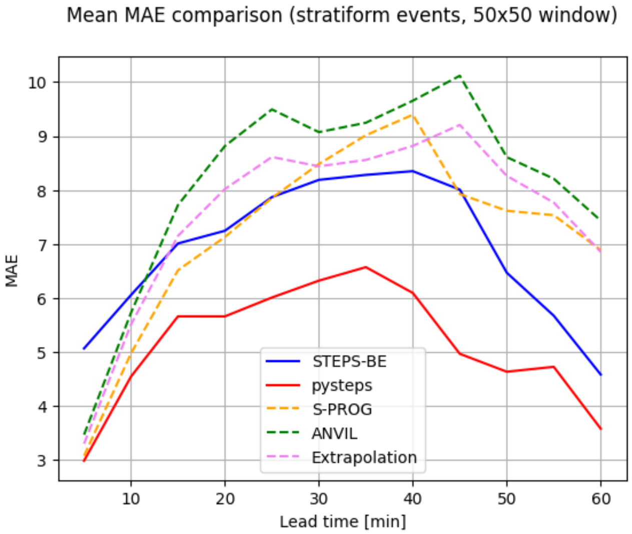


Figure 5.20. Average MAE for stratiform events

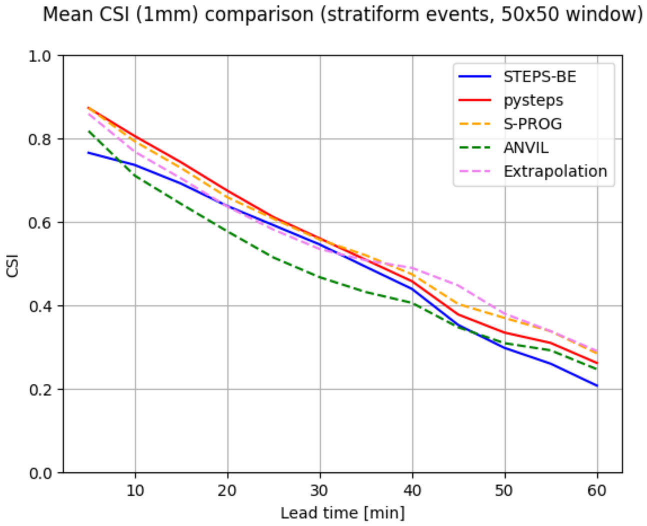


Figure 5.21. Average CSI for 1 [mm/h] thershold of stratiform events

Despite the changing nature of the events, the considerations made above remain valid. Also in this case, the ensemble's mean of py-STEPS provides the least error results, together with the one of STEPS-BE. The warning ability of py-STEPS (in terms of CSI) is comparable to the one of S-PROG in the first 30 minutes of the nowcast, but without adding any relevant improvement. Therefore, the conclusion again is that it is not useful to average the members of the probabilistic nowcast to increase the warning ability of the model.

5.3.3 Cumulative rainfall volume and uncertainties

Finally, an analysis of the cumulative rainfall volumes inside the simulated hydrological catchments is presented for convective and stratiform events. In this section, the cumulative rainfall volumes were not calculated considering the ensemble's mean, but every single member. Namely, after two hours lead time 20 different cumulative rainfall volumes, one for every member, were obtained providing the results in the picture below:

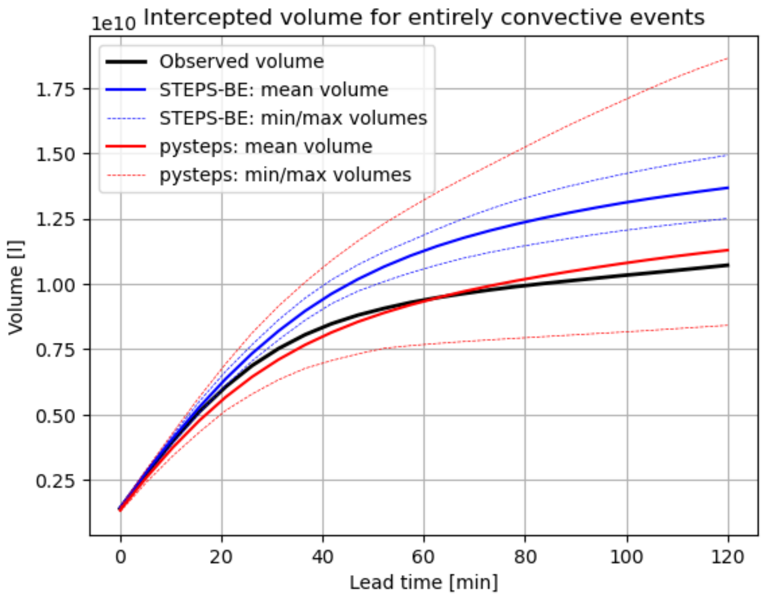


Figure 5.22. Cumulative rainfall volumes of convective events

The dark line indicates the observed cumulative rainfall volume, the solid red and blue lines refer to the average of the 20 volumes produced by the members of py-STEPS and STEPS-BE respectively and the dashed line report the maximum and the minimum cumulative volume among the ensemble members for every lead time.

The interval between the dashed lines shows the uncertainty in the nowcast, which grows with lead time. In fact, the minimum and maximum lines tend to be near within the first 20 minutes of the forecast, but as the lead time increases, the gap between the two grows wider and wider. In more rigorous terms, it can be stated that the variance of the cumulative rainfall volume is not constant in time and therefore the studied variable (cumulative volume) is heteroscedastic. It is easy to notice that py-STEPS intercepts the cumulative rainfall volume much better than STEPS-BE, that is characterized by a persistent overestimation, particularly remarked toward 2h lead time.

Another interesting aspect is related to the degree of uncertainty of the two models. In fact, although it provides much better results, py-STEPS ensemble is much more spread and provides a greater variety of volume ranges than STEPS-BE, that shows a much narrower spacing between the minimum and maximum volumes of its ensemble. Therefore, it can be remarked that py-STEPS is much more accurate than STEPS-BE in smaller domain analysis, but it is characterized by a higher degree of uncertainty.

The same methodology can be applied to the stratiform events, thus obtaining the following result:

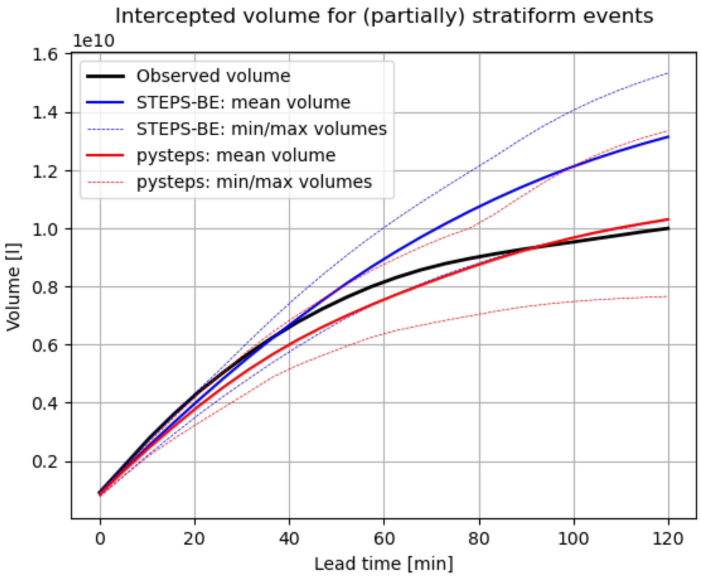


Figure 5.23. Cumulative rainfall volumes of stratiform events

Again, py-STEPS performs much better than STEPS-BE, that keeps overestimating the cumulative rainfall volume. However, in these kinds of events the degree of uncertainty of STEPS-BE is comparable to the one of py-STEPS (notice that the line of minimum values of STEPS-BE is almost hidden under the average curve of py-STEPS).

It is worth noting that the rainfall volume cumulates slower than in the previous case due to the more stratiform nature of the considered events, even if the total volume at the end of the nowcast does not differ dramatically compared to convective events.

5.4 CONCLUSIONS

This chapter focused on the comparison of the two probabilistic nowcasting models py-STEPS and STEPS-BE. Thanks to a cascade of stochastic noise applied to each level of the decomposed rainfall field, these models can produce an ensemble of forecasts for each lead time. The main differences between these two models are the application of an AR(1) (in STEPS-BE) instead of an AR(2) (in py-STEPS) and the implementation of two different optical flow algorithms (COTREC scheme vs Lucas-Kanade method). The advantage of these models with respect to deterministic options consists in the estimation of uncertainties, that is a relevant factor in modern hydrological applications.

Before running the py-STEPS model, a sensitivity analysis has been performed to set the optimal configuration to obtain the precipitation fields. In particular, the influence of two parameters was tested: the stochastic noise generator and the number of ensemble members. After evaluating the results in terms of ROC curves and reliability diagrams, it was concluded that the choice of a specific noise generator over another does not provide remarkable benefits. However, the non-parametric option was chosen since it requires the lowest computational time. On the other hand, 20 members were considered the other optimal alternative, as they offer a tradeoff between reasonable computational time of the nowcast (slightly more than 3 minutes) and ensemble's skill.

To validate and compare the two models, analyses were conducted across the entire radar domain and sub-domains (simulated hydrological catchments) considering the event categorization already discussed in chapter 4. To make a more general validation, both the ensemble and the ensemble's mean were tested through probabilistic (ROC curves, reliability diagrams, rank histograms) and deterministic (MAE, CSI) metrics.

In the analysis over the entire radar domain, py-STEPS proved more skillful than STEPS-BE, even though the latter showed some strengths. In fact, while py-STEPS got better results in terms of reliability diagrams and rank histograms, proving that its ensemble is better calibrated, STEPS-BE had higher values of AUC at 2 hours lead time (while py-STEPS still got better results in terms of ROC curves at least for the first 30 minutes of the nowcast). Considering the ensemble's mean, py-STEPS provided better results in terms of MAE and CSI (for the first 40 minutes lead time), but STEPS-BE intercepted the cumulative rainfall volumes more accurately. Therefore, when the entire radar image is studied, it is difficult to establish which method is definitely the best.

On the other hand, py-STEPS completely outperformed STEPS-BE in the analysis inside the sub-domains. The higher AUC values when forecasting both the convective and stratiform events show that py-STEPS has higher skill and adaptability of STEPS-BE at smaller scales. Moreover, also in this case the MAE values proved that the former method is less biased to evaluate rainfall values. However, the most interesting aspect is related to the hydrological balance, where py-STEPS forecasted the cumulative rainfall volumes with a much higher accuracy compared to STEPS-BE, that presented a persistent overestimation.

This chapter also underlined that the ensemble's mean can provide less biased results in terms of MAE compared to deterministic methods, thanks to the averaging of its members altered by the noise cascade. However, no relevant gains of predictive ability (measured through CSI) have been detected.

Consequently, it is necessary to make a tradeoff between deterministic and probabilistic methods to model one single precipitation field. As usually happens in the engineering field, the choice of a model must suit its application.

To draw a conclusion about the comparison between the two probabilistic models, it seems that py-STEPS represents the best choice. In fact, the small advantages offered by STEP-BE at the domain scale are less relevant than the much higher skill of py-STEPS at the hydrological scale, that is the most interesting one for practical applications. In addition to that, it must be underlined that py-STEPS offers the opportunity of changing the parameters and have a complete control of the model, while the same is not valid for the software developed at the RMI. Consequently, py-STEPS not only provides a more skillful method to forecast the precipitation fields, but is also much more adaptive and thus relevant for hydrological research then STEPS-BE.

Chapter 6: CONCLUSION

Rainfall extremes are projected to increase in the future. Indeed, the Clausius Clapeyron law states that for every 1°C increase in temperature, the atmosphere's water storage capacity increases by 7%. This means that when the earth warms, there will be a lot more water vapor available in the atmosphere to trigger convective and intense events. This is also supported by the IPCC's findings, which show that precipitation extremes will increase regardless of location in every global warming scenario. Therefore, quick and effective methods to deal with this new issue are needed.

Nowcasting is one of the instruments that can assist us in defining and comprehending the state of the atmosphere in short time spans to predict precipitation extremes. The major focus of this thesis was STEPS, which is one of the most important nowcasting methods. STEPS can provide both deterministic and probabilistic nowcasts, which means it is possible to generate one or more forecasted fields for each lead time. The most intriguing characteristic of this method is its ability to mimic both the "spatial scaling" and "dynamic scaling" properties of precipitation field. The term "spatial scaling" refers to the fact that the precipitation field can be viewed as a stack of various precipitation structures of varying sizes rather than as a continuum. The term "dynamic scaling" refers to the fact that larger precipitation structures have a longer lifetime than smaller precipitation structures, which form and degrade at a faster rate. Therefore, this is the method's fundamental core: it can reproduce spatial scaling thanks to a FFT that decomposes the field into its different structures, and then it can reproduce the growth and decay of small precipitation structures, i.e. dynamic scaling, thanks to an autoregressive filter applied to each cascade level. Furthermore, this method allows for the generation of a noise cascade, which can be added to each level of the decomposed rainfall field. If this is done, an ensemble of members for each lead time of the nowcast can be forecasted (probabilistic model); otherwise, just one field is produced for each lead time (deterministic model).

STEPS has several implementations, two of which were investigated in this thesis: py-STEPS and STEPS-BE. The first choice is a Python library that is open source and has several features. This library, for example, offers modification of settings such as the optical flow algorithm used to estimate the motion field, the order of the autoregressive filter (which is commonly set to 2, as in the standard STEPS technique published by Seed et al.), and the number of cascade levels. The other choice is the RMI version, which is a licensed private program. This suggests that none of its parameters can be controlled. This approach differs significantly from py-STEPS in several ways. For example, it implements its own optical flow method, a COTREC scheme, rather than the alternatives VET, DATRS, and Lucas-kanade of py-STPES, employs an AR(1) rather than an AR(2), and implements a different motion field perturbation. One of the main aims of this work was to establish which method is better suited to describe the evolution of precipitation extremes in the Belgian territory.

However, before performing the validation and comparison of different models, it was decided to conduct a radar data analysis, because radar images are the input to all the STEPS-based models that have been studied in this thesis. The data used were composite radar images referring to ten separate time periods, with a spatial resolution of 1058 m and a temporal resolution of 5 minutes. Data from Water Info rain gauges were used for validation. The network of these rain gauges covers the Flanders region, has a temporal resolution of 5 minutes (similar to radar images), and data for 2017 and 2018 were available. The data from the rain gauges were used to extract the events required for validation.

The event extraction method took three parameters into account: an upper threshold (to define the start and finish of the event), a lower threshold (to remove extremely light rain from the event), and the Inter Event Time, which was set to 12 hours (to extract independent events). To validate the radar results, three main strategies were evaluated: considering the pixel in which the rain gauge is located, considering the maximum in the station's neighborhood or considering the mean in the station's neighborhood. It has been found that, in general, considering the maximum in the surrounding increases the level of overestimation by the radar. To summarize, the RMSE for various distances from the rain gauge were generalized within a boxplot, where it was evident that the greater the distance from the rain gauge, the greater the error. Similarly, the RMSE boxplots of the mean in the surrounding area showed that it is not advantageous to examine pixels that are too far away from the rain gauge. Therefore, the first validation option was adopted and is recommended for future analysis in the Belgian domain.

One of the most significant sources of error was a time shift between the radar and rain gauge peaks. This shift is caused by the time it takes for precipitation to reach the ground. To determine the amount of this time shift, an analysis was performed on all the considered events, and it was discovered that the time shift is approximately 10 minutes for the majority of the occurrences. This is crucial information since it indicates that the precipitation field observed by the radar, and thus the one predicted by the model, takes an extra 10 minutes to precipitate, on average. However, despite this time shift was taken into account, the scatter plot of radar-rain gauge measurements of rain intensities showed a high level of disagreement between the two devices. Consequently, radar correction is required for future study, particularly for nowcasting applications, to acquire more reliable images.

Obviously, aggregating the results helps to increase the level of agreement, and it was discovered that after a 30-minute aggregation, the RMSE between radar and rain gauge measurements was practically half. However, the more rain intensities are aggregated, the more information is lost. Another interesting finding was that the RMSE (between the measurements of the two devices) decreases with increasing distance from the closest radar. This is because errors caused by ground clutter, for example, are more remarked in proximity to the radar. However, in theory, the RMSE should increase again after 60-70 km, due to phenomena like beam broadening or radar attenuation. This was not assessed in this study since there were insufficient rain gauges and events available with a distance between rain gauge and closest radar higher than 60 Km. Therefore, it is suggested that future research include more events at rain gauges located more than 60 kilometers from the nearest radar, to verify if an optimal distance where the values of the image are more accurate exist. This could be relevant for the Belgian domain, since it would be possible to identify locations where the radar values are more reliable, and hence where the nowcast results are more dependable as well.

When the STEPS method is applied without any stochastic perturbation (i.e., without noise cascade), one single rainfall field is produced for each lead time. This procedure is known as deterministic nowcast. In this thesis, three different deterministic models were analyzed: a simple extrapolation, S-PROG and ANVIL. The extrapolation simply advects the last radar images along the Lagrangian trajectories, without changing the local intensities of the rainfall field (Lagragian persistence). On the other hand, S-PROG and ANVIL can represent the growth and decay of precipitation structures by means of the STEPS procedure. The major difference between the two models is that S-PROG implements a simple AR(2), while ANVIL exploits an ARI (autoregressive integrated filter).

All of these models require an estimation of the velocity field, which is the first step in computing the nowcast. As a result, the three optical flow algorithms available in py-STEPS were examined to determine how they affected the forecasted precipitation fields in terms of MAE and CSI. It was discovered that, at least for this investigation, the choice of one algorithm or the other did not provide notable differences on the final results. However, it was shown that the computational time, which is a significant element in real-time forecasting, varied dramatically amongst the three different optical flow algorithms. In the end, Lucas-Kanade was chosen as the default configuration because it provided a balance between satisfying CSI-MAE results and computational time, and it is thus recommended for future investigations over the Belgian domain.

The three models were then compared using metrics such as CSI, POD, FAR, and MAE throughout the whole radar domain considering the events extracted in Chapter 3. The results revealed that ANVIL was the worst method in terms of MAE (i.e., more biased), whereas S-PROG and extrapolation performed better. In terms of categorical measures, however, ANVIL performed better on average. Nevertheless, since no classification of the events was carried out, these results were considered unsatisfactory. As a result, an algorithm based on the parametrization of power spectra within sub-domains was developed and a distinction between convective and stratiform events was identified in small radar regions. Therefore, rather than analyzing the full radar domain, a second analysis was undertaken within these smaller regions, which should better represent the size of a real hydrological catchment. Although the overall tendency was not significantly different from that seen for the entire radar domain, the results in this case were neater. When high precipitation thresholds are set in categorical metrics such as POD, CSI, and FAR for convective events, ANVIL performs better. However, because it has errors in terms of MAE, this model can result in significant cumulative rainfall overestimation. S-PROG performed best for stratiform events when smaller precipitation thresholds were used, and it was also the best approach for intercepting cumulative rainfall volumes. As a result, ANVIL is recommended for more extreme events because it exaggerates both the growth and decay of precipitation structures, whereas S-PROG works better with stratiform events or convective events that "degrade" quickly into stratiform precipitation due to its smoothing tendency.

In contrast to the deterministic approach, a probabilistic nowcast can be obtained by applying stochastic noise to each level of the decomposed rainfall field. This approach has a significant benefit over the deterministic methods because it produces various precipitation fields at each lead time, allowing the uncertainties of the nowcast to be estimated. Before comparing py-STEPS and STEPS-BE, a sensitivity analysis was undertaken to determine the configuration of py-STEPS considering the stochastic noise generator and the number of ensemble members. The non-parametric generator was chosen among the three different evaluated generators since it offered the shortest computational time (while the outcomes of the nowcasts showed a low dependence on generator choice). On the other hand, increasing the number of ensemble members is good for improving the nowcast's skill, but this gain must be accompanied by a fair computational time. Because creating 30 ensemble members for a 2 h nowcast can take up to 5 minutes, 20 members were chosen as a reasonable trade-off between skill and computational time.

In accordance with the methodology used in the previous chapter, the two STEPS implementations were evaluated using probabilistic metrics such as ROC curves, reliability diagrams, and rank histograms both in the whole radar domain and in small sub-catchments. In almost all analyses, py-STEPS outperformed STEPS-BE at both larger and smaller scales. In particular, the assessments conducted inside the simulated catchments revealed that py-STEPS has a substantially higher skill than STEPS-BE in forecasting both convective and stratiform events, as well as better intercepting accumulated rainfall volumes. The fact that an open-source software, that allows complete control of all model parameters and can use radar images from any region of the world, outperformed a private one developed specifically for Belgium suggests that researchers can have free access to a powerful tool for nowcasts. However, it is important to underline that the hydrological balance showed that py-STEPS uncertainties are rather high towards 2 hours lead time.

Finally, the impact of the stochastic perturbation was investigated further by averaging all members of the nowcast in a single precipitation field (ensemble mean). Even if the main advantage of the probabilistic technique (estimation of uncertainties) is lost, the MAE curves demonstrated that the ensemble mean has smaller errors, on average, than pure deterministic methods such as ANVIL, S-PROG, and extrapolation. This is most likely attributable not only to the noise in the averaged precipitation field, but also to the fact that probabilistic techniques incorporate a perturbation of the motion field as well, which is more realistic than the stationary field of deterministic methods. However, it was discovered that the ensemble mean's CSI is not greater than that of deterministic approaches, implying that the former has no greater predictive ability than a simpler deterministic nowcast.

To conclude, it may be interesting to underline that the Belgian RMI is currently improving its nowcasting model STEPS-BE integrating some of the features of the open-source py-STEPS library. As the authors report [51], a merging of NWP outputs, stochastic noise, and extrapolation will be developed exploiting the same flow diagram of py-STEPS to produce probabilistic nowcasts with a reliability that extends beyond 2 hours lead time. Therefore, even though STEPS-BE will still be the operational version, soon the RMI will use the benefits of the open-source and blending procedures. This is an encouraging prospect for the future of nowcasting and also shows that the advantages of py-STEPS highlighted in this thesis have a practical application in operational contexts.

APPENDIX A: EVENT LIST

Station	Start time [m/d/y hh:mm]	End time[m/d/y hh:mm]	Peak intensity [mm/h]
Poperinge_P	18/05/2017 20:15	18/05/2017 21:00	37.44
Vlamertinge OTT_P	18/05/2017 20:30	18/05/2017 21:10	58.8
Ieper_P	18/05/2017 20:35	18/05/2017 21:30	79.08
Lo-Fintele OTT_P	18/05/2017 20:45	18/05/2017 21:20	87.6
Geluwe_P	18/05/2017 21:00	18/05/2017 21:40	64.8
Zarren_P	18/05/2017 21:10	18/05/2017 21:55	142.08
Roeselare_P	18/05/2017 21:35	18/05/2017 21:55	54.12
Klemskerke_P	18/05/2017 21:40	18/05/2017 22:35	49.08
Zwevegem OTT_P	18/05/2017 21:45	18/05/2017 22:05	56.4
Brugge OTT_P	18/05/2017 22:15	18/05/2017 22:55	98.4
Liedekerke_P	19/07/2017 23:05	19/07/2017 23:40	42.24
Lot OTT_P	19/07/2017 23:35	20/07/2017 00:00	55.2
St-Pieters-Leeuw_P	19/07/2017 23:40	20/07/2017 00:00	96.48
Bonheiden_P	20/07/2017 00:00	20/07/2017 03:15	57.36
Boortmeerbeek OTT_P	20/07/2017 00:00	20/07/2017 00:40	118.8
Nossegem_P	20/07/2017 00:05	20/07/2017 00:25	78.72
Loenhout_P	20/07/2017 00:25	20/07/2017 01:20	38.64
Korbeek-Dijle_P	20/07/2017 00:25	20/07/2017 00:30	48.96
Poperinge_P	20/07/2017 00:25	20/07/2017 01:05	38.04
Heverlee_P	20/07/2017 00:30	20/07/2017 00:40	70.32
Rotselaar_P	20/07/2017 00:35	20/07/2017 00:50	43.56
Herentals_P	20/07/2017 00:40	20/07/2017 04:35	81.72
Aarschot OTT_P	20/07/2017 00:45	20/07/2017 01:00	56.4
Vosselaar_P	20/07/2017 00:50	20/07/2017 03:50	49.92

Sint-Laureins OTT_P	20/07/2017 03:05	20/07/2017 12:35	44.4
Vlamertinge OTT_P	20/07/2017 03:45	20/07/2017 12:15	42
Waregem_P	20/07/2017 04:30	20/07/2017 12:55	71.16
Zwevegem OTT_P	15/08/2017 05:55	15/08/2017 07:35	57.6
Maarke-Kerkem_P	15/08/2017 08:35	15/08/2017 08:40	36.36
Sint-Joris_P	15/08/2017 09:05	15/08/2017 09:25	68.16
Klemskerke_P	15/08/2017 09:25	15/08/2017 10:35	40.32
Brugge OTT_P	15/08/2017 09:45	15/08/2017 09:50	110.4
Geluwe_P	30/08/2017 13:40	30/08/2017 23:00	59.64
Oostkamp_P	11/09/2017 14:50	11/09/2017 22:10	64.2
Ertvelde_P	11/09/2017 20:25	12/09/2017 08:45	63.12
Vlamertinge OTT_P	11/09/2017 21:00	11/09/2017 21:05	38.4
Lommel OTT_P	11/09/2017 22:05	12/09/2017 05:15	38.4
Houthalen_P	12/09/2017 00:50	12/09/2017 01:10	41.04
Nossegem_P	24/05/2018 15:30	24/05/2018 17:20	105.72
Zingem_P	24/05/2018 16:20	24/05/2018 17:05	50.88
Denderbelle_P	24/05/2018 16:35	24/05/2018 17:15	66.36
Klemskerke_P	27/05/2018 13:30	27/05/2018 14:55	50.04
Sint-Joris_P	27/05/2018 14:35	27/05/2018 14:55	55.44
Sint-Laureins OTT_P	27/05/2018 19:20	27/05/2018 19:55	57.6
Boekhoute_P	27/05/2018 19:35	27/05/2018 20:10	63.48
St-Pieters-Leeuw_P	07/06/2018 06:05	07/06/2018 06:25	36.96
Aarschot OTT_P	07/06/2018 06:50	07/06/2018 07:05	57.6
Geluwe_P	07/06/2018 15:35	07/06/2018 16:00	59.76
Dudzele_P	05/09/2018 01:00	05/09/2018 04:30	60.36
Waregem_P	05/09/2018 01:15	05/09/2018 01:50	44.52

APPENDIX B: CONVECTIVE AND STRATIFORM EVENT LIST

Convective events:

Station name	**Starting time**	**Longitude**	**Latitude**
Bonheiden_P	7/20/2017, 00:00	366299.5343	-281839.3941
Boortmeerbeek OTT_P	7/20/2017, 00:00	371051.625	-286418.8022
Nossegem_P	7/20/2017, 00:05	365559.5114	-298380.7596
Rotselaar_P	7/20/2017, 00:35	379684.3448	-291314.4193
Herentals_P	7/20/2017, 00:40	389475.1055	-267390.5241
St-Pieters-Leeuw_P	7/19/2017, 23:40	351451.6476	-306521.7096
Heverlee_P	7/20/2017, 00:30	377222.9425	-298925.547
Liedekerke_P	7/19/2017, 23:05	337354.8354	-297933.5141
Loenhout_P	7/20/2017, 00:25	375635.7955	-303514.6869
Vosselaar_P	7/20/2017, 00:50	393788.5696	-252694.9699
Lot OTT_P	7/19/2017, 23:35	349512.6734	-311261.64
Aarschot OTT_P	7/20/2017, 00:45	389977.9522	-287293.8824
Zingem_P	5/24/2018, 16:40	306698.2663	-295564.8595
Lommel OTT_P	9/11/2017, 22:05	418181.8056	-257553.6995
Houthalen_P	9/12/2017, 00:50	424886.7581	-283177.7211
Waregem_P	9/05/2018, 01:15	289646.6383	-300182.7001
Dudzele_P	9/05/2018, 01:00	280741.7435	-253393.0434
Zwevegem OTT_P	8/15/2017, 05:55	284914.8912	-304308.5349
Geluwe_P	5/18/2017, 21:00	266820.2521	-305352.9508
Geluwe_P	8/30/2017, 13:40	266820.2521	-305352.9508
Zarren_P	5/18/2017, 21:10	259300.2322	-281762.9531
Vlamertinge OTT_P	9/11/2017, 21:00	247683.9378	-301230.2148

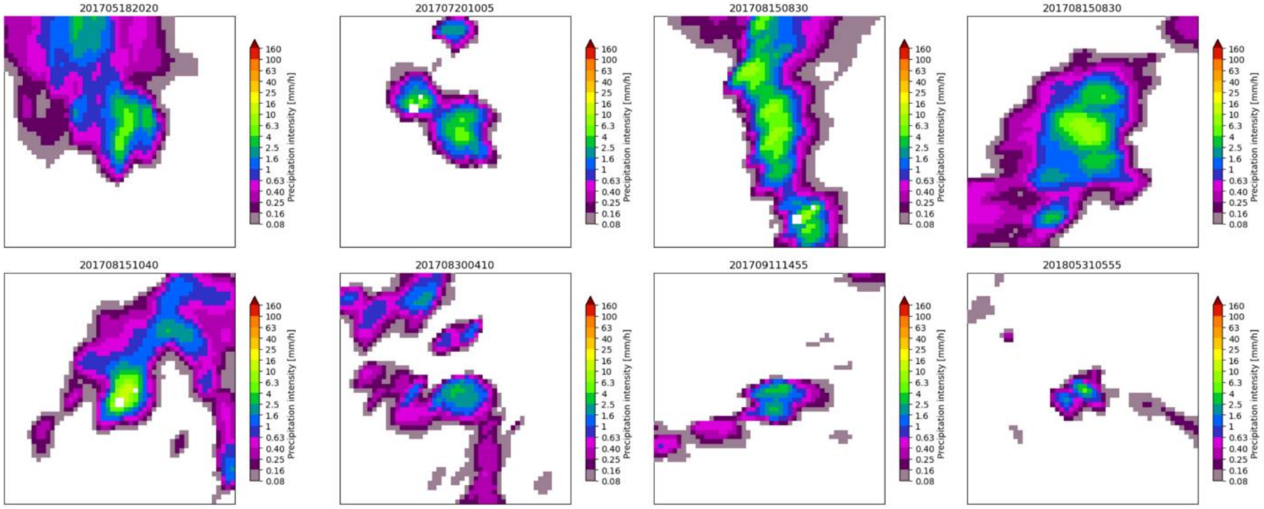
Sint-Joris_P	8/15/2017, 09:05	244998.142	-269903.1012
Lo-Fintele OTT_P	5/18/2017, 20:45	242798.981	-288522.5262
Poperinge_P	5/18/2017, 20:15	237834.1233	-303052.2863
Korbeek-Dijle_P	7/20/2017, 00:25	237834.1233	-303052.2863

Stratiform events:

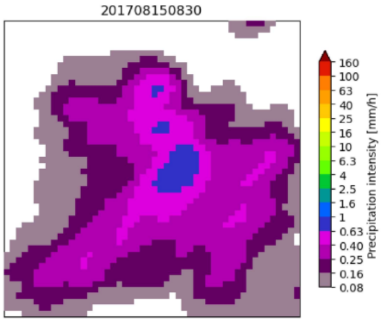
Station name	Starting time	Longitude	Latitude
Poperinge_P	7/20/2017, 00:25	379597.7127	-242559
Nossegem_P	5/24/2018, 15:30	365559.5114	-298381
Sint-Laureins OTT_P	7/20/2017, 03:05	298791.0951	-259338
Waregem_P	7/20/2017, 04:30	289646.6383	-300183
Oostkamp_P	09/11/2017 14:50	279189.4232	-270094
Geluwe_P	06/07/2018 15:35	266820.2521	-305353
Klemskerke_P	8/15/2017 9:25	262389.7548	-259429
Klemskerke_P	5/27/2018 13:30	262389.7548	-259429
Ieper_P	5/18/2017 20:35	253168.9505	-301537
Vlamertinge OTT_P	5/18/2017 20:30	247683.9378	-301230
Vlamertinge OTT_P	7/20/2017 3:45	247683.9378	-301230

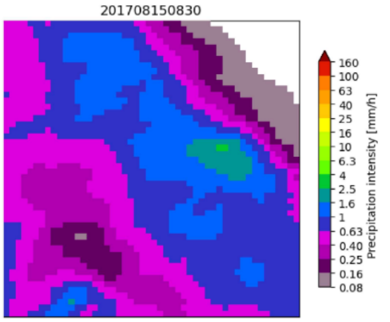
APPENDIX C: FIELDS TO PARAMETRIZE CONVECTIVE AND STRATIFORM POWER SPECTRA

Convective fields:

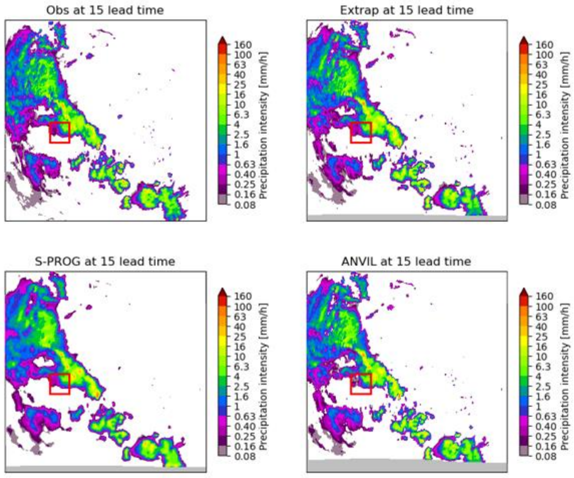


Stratiform fields:

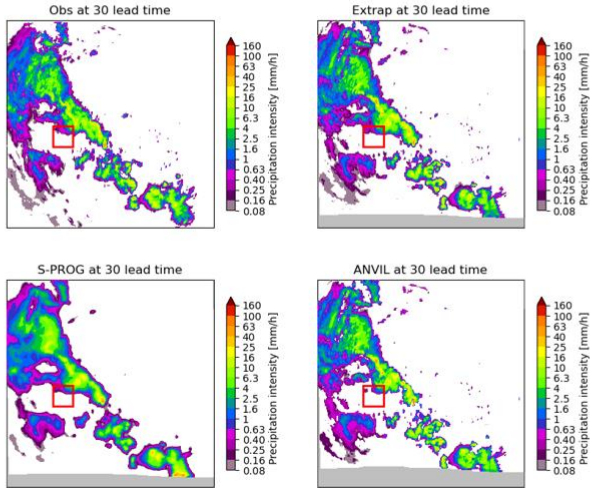




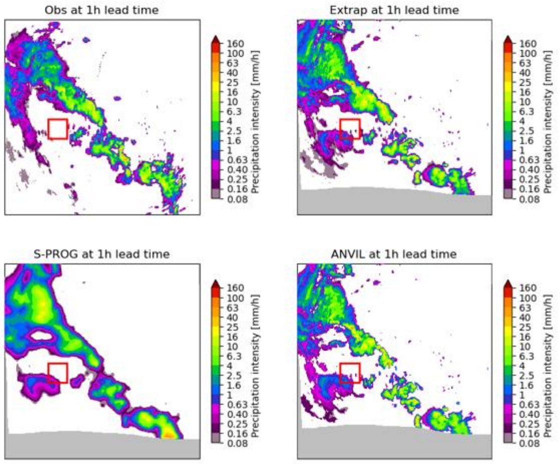
APPENDIX D: EXAMPLE OF STRATIFORM EVENTS AT DIFFERENT LEAD TIMES



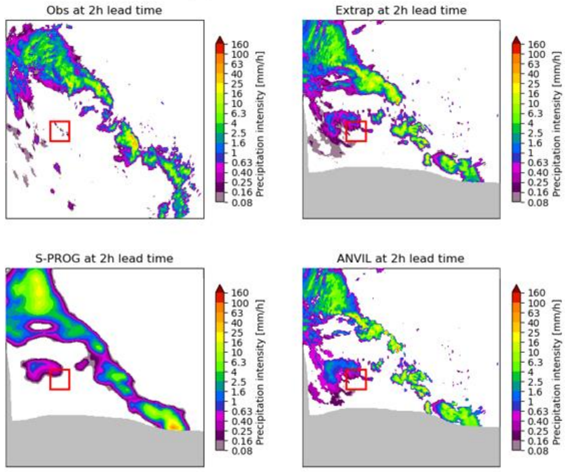
(a) 15 min lead time



(b) 30 min lead time



(c) 1h lead time



(d) 2h lead time

REFERENCES

[1] D. Easterling *et al.*, '3 - Changes in Climate Extremes and their Impacts on the Natural Physical Environment', Cambridge University Press, 2012.

[2] 'Climate Change Indicators: Heavy Precipitation | US EPA'. https://www.epa.gov/climate-indicators/climate-change-indicators-heavy-precipitation (accessed Oct. 23, 2022).

[3] 'Extreme Precipitation and Climate Change - Center for Climate and Energy SolutionsCenter for Climate and Energy Solutions'. https://www.c2es.org/content/extreme-precipitation-and-climate-change/ (accessed Oct. 23, 2022).

[4] H. Tabari, 'Climate change impact on flood and extreme precipitation increases with water availability', 123AD, doi: 10.1038/s41598-020-70816-2.

[5] S. I. Seneviratne *et al.*, 'Weather and Climate Extreme Events in a Changing Climate.', in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, and B. Z. MassonDelmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, Ed., 2021, pp. 1513–1766. doi: 10.1017/9781009157896.013.

[6] G. Myhre *et al.*, 'frequency of extreme precipitation increases extensively with event rareness under global warming', doi: 10.1038/s41598-019-52277-4.

[7] 'World Energy Outlook 2020 – Analysis - IEA'. https://www.iea.org/reports/world-energy-outlook-2020 (accessed Oct. 23, 2022).

[8] 'Explainer: What the new IPCC report says about extreme weather and climate change - Carbon Brief'. https://www.carbonbrief.org/explainer-what-the-new-ipcc-report-says-about-extreme-weather-and-climate-change/ (accessed Oct. 23, 2022).

[9] C. Mass and C. F. Mass, 'Nowcasting: The Next Revolution in Weather Prediction'.

[10] 'Nowcasting Guidelines – A Summary | World Meteorological Organization'. https://public.wmo.int/en/resources/bulletin/nowcasting-guidelines-%E2%80%93-summary (accessed Oct. 23, 2022).

[11] Z. Sokol, J. Szturc, J. Orellana-Alvear, J. Popová, A. Jurczyk, and R. Célleri, 'remote sensing The Role of Weather Radar in Rainfall Estimation and Its Application in Meteorological and Hydrological Modelling-A Review The Role of Weather Radar in Rainfall Estimation and Its Application in Meteorological and Hydrological Modelling-A', *Citation: Sokol, Z.; Szturc, J*, 2021, doi: 10.3390/rs13030351.

[12] D. Heuvelink, M. Berenguer, C. C. Brauer, and R. Uijlenhoet, 'Hydrological application of radar rainfall nowcasting in the Netherlands', *Environ Int*, vol. 136, Mar. 2020, doi: 10.1016/J.ENVINT.2019.105431.

[13] A. J. Roberts *et al.*, ‘Nowcasting for Africa: advances, potential and value’, *Weather*, vol. 77, no. 7, pp. 250–256, Jul. 2022, doi: 10.1002/wea.3936.

[14] N. E. Bowler, C. E. Pierce, and A. W. Seed, ‘STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP’, *Meteorol. Soc*, vol. 132, pp. 2127–2155, 2006, doi: 10.1256/qj.04.100.

[15] A. W. Seed, C. E. Pierce, and K. Norman, ‘Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme’, *Water Resour Res*, vol. 49, no. 10, pp. 6624–6641, Oct. 2013, doi: 10.1002/wrcr.20536.

[16] R. O. Imhoff, C. C. Brauer, A. Overeem, A. H. Weerts, and R. Uijlenhoet, ‘Spatial and Temporal Evaluation of Radar Rainfall Nowcasting Techniques on 1,533 Events’, *Water Resour Res*, vol. 56, no. 8, Aug. 2020, doi: 10.1029/2019WR026723.

[17] S. Liguori, M. A. Rico-Ramirez, A. N. A. Schellart, and A. J. Saul, ‘Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments’, in *Rainfall in the Urban Context: Forecasting, Risk and Climate Change*, Elsevier B.V., 2012, pp. 80–95. doi: 10.1016/j.atmosres.2011.05.004.

[18] L. De Cruz, L. Foresti, M. Reyniers, L. Delobbe, D. Degrauwe, and A. Duerinckx, ‘Blending STEPS-BE with high-frequency, high-resolution precipitation forecasts from a numerical weather model’, 2017.

[19] S. Pulkkinen *et al.*, ‘Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1.0)’, *Geosci Model Dev*, vol. 12, no. 10, pp. 4185–4219, Oct. 2019, doi: 10.5194/gmd-12-4185-2019.

[20] L. Foresti, M. Reyniers, A. Seed, and L. Delobbe, ‘Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium’, *Hydrol Earth Syst Sci*, vol. 20, no. 1, pp. 505–527, Jan. 2016, doi: 10.5194/hess-20-505-2016.

[21] M. Kallas, P. Honeine, C. Francis, and H. Amoud, ‘Kernel autoregressive models using Yule-Walker equations’, *Signal Processing*, vol. 93, no. 11, pp. 3053–3061, 2013, doi: 10.1016/j.sigpro.2013.03.032.

[22] ‘pySTEPS · GitHub’. https://github.com/pySTEPS (accessed Dec. 06, 2022).

[23] ‘pysteps – The nowcasting initiative’. https://pysteps.readthedocs.io/en/stable/ (accessed Dec. 06, 2022).

[24] ‘PySteps developers’, 2020.

[25] N. Sharmin and R. Brad, ‘Optimal filter estimation for Lucas-Kanade optical flow’, *Sensors (Switzerland)*, vol. 12, no. 9, pp. 12694–12709, Sep. 2012, doi: 10.3390/s120912694.

[26] A. Bellon, I. Zawadzki, A. Kilambi, H. C. Lee, Y. H. Lee, and G. Lee, ‘McGill algorithm for precipitation nowcasting by lagrangian extrapolation (MAPLE) applied to the South Korean radar network. Part I: Sensitivity studies of the Variational Echo Tracking (VET) technique’, *Asia Pac J Atmos Sci*, vol. 46, no. 3, pp. 369–381, Aug. 2010, doi: 10.1007/s13143-010-1008-x.

[27] U. Patent, ‘Sheet 1 of 22’, 2010.

[28] ‘Relative operating characteristic (ROC) - Met Office’. https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/user-guide/interpret-roc (accessed Dec. 08, 2022).

[29] S. Liguori and M. A. Rico-Ramirez, ‘Quantitative assessment of short-term rainfall forecasts from radar nowcasts and MM5 forecasts’, *Hydrol Process*, vol. 26, no. 25, pp. 3842–3857, Dec. 2012, doi: 10.1002/hyp.8415.

[30] ‘Reliability and sharpness diagrams - Met Office’. https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/user-guide/interpret-reliability (accessed Dec. 08, 2022).

[31] ‘Rank Histogram / Talagrand Diagram - Statistics How To’. https://www.statisticshowto.com/rank-histogram/ (accessed Dec. 08, 2022).

[32] ‘Example gallery’. https://pysteps.readthedocs.io/en/stable/auto_examples/index.html (accessed Aug. 12, 2023).

[33] B. Zhao and B. Zhang, ‘Assessing Hourly Precipitation Forecast Skill with the Fractions Skill Score’, *Journal of Meteorological Research*, vol. 32, no. 1, pp. 135–145, Feb. 2018, doi: 10.1007/s13351-018-7058-1.

[34] L. Han, J. Zhang, H. Chen, W. Zhang, and S. Yao, ‘Toward the Predictability of a Radar-Based Nowcasting System for Different Precipitation Systems’, *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022, doi: 10.1109/LGRS.2022.3185031.

[35] E. A. Brandes, ‘Radar Measurement ol Rainfall’, 1979.

[36] S. Stefano Almaviva SpA, F. Russo, F. Napolitano, S. Sebastianelli, F. Russo, and L. Baldini, ‘H-SAF: The EUMETSAT Satellite Application Facility on Support to Operational Hydrology and Water Management View project Comparison between radar and rain gauges data at different distances Comparison between radar and rain gauges data at different distances from radar and correlation existing between the rainfall values in the adjacent pixels Comparison between radar and rain gauges data at different distances’, *Hydrol. Earth Syst. Sci. Discuss*, vol. 7, pp. 5171–5212, 2010, doi: 10.5194/hessd-7-5171-2010.

[37] E. Goudenhoofdt, L. Delobbe, and P. Willems, ‘Regional frequency analysis of extreme rainfall in Belgium based on radar estimates’, *Hydrol Earth Syst Sci*, vol. 21, no. 10, pp. 5385–5399, Oct. 2017, doi: 10.5194/hess-21-5385-2017.

[38] K. Schoeters, ‘Early warning system for sewer floods-Case Antwerp’, 2019.

[39] C. Z. Van De Beek, H. Leijnse, P. Hazenberg, and R. Uijlenhoet, ‘Close-range radar rainfall estimation and error analysis’, *Atmos Meas Tech*, vol. 9, no. 8, pp. 3837–3850, 2016, doi: 10.5194/amt-9-3837-2016.

[40] S. Pulkkinen, V. Chandrasekar, A. Von Lerber, and A. M. Harri, ‘Nowcasting of Convective Rainfall Using Volumetric Radar Observations’, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7845–7859, Nov. 2020, doi: 10.1109/TGRS.2020.2984594.

[41] ‘pysteps/anvil.py at master · pySTEPS/pysteps · GitHub’. https://github.com/pySTEPS/pysteps/blob/master/pysteps/nowcasts/anvil.py (accessed May 03, 2023).

[42] A. Bellon, I. Zawadzki, A. Kilambi, H. C. Lee, Y. H. Lee, and G. Lee, ‘McGill algorithm for precipitation nowcasting by lagrangian extrapolation (MAPLE) applied to the South Korean radar network. Part I: Sensitivity studies of the Variational Echo Tracking (VET) technique’, *Asia Pac J Atmos Sci*, vol. 46, no. 3, pp. 369–381, Aug. 2010, doi: 10.1007/s13143-010-1008-x.

[43] S. Pulkkinen, V. Chandrasekar, A. Von Lerber, and A. M. Harri, ‘Nowcasting of Convective Rainfall Using Volumetric Radar Observations’, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7845–7859, Nov. 2020, doi: 10.1109/TGRS.2020.2984594.

[44] C. Capsoni, L. Luini, A. Paraboni, C. Riva, and A. Martellucci, ‘A new prediction model of rain attenuation that separately accounts for stratiform and convective rain’, *IEEE Trans Antennas Propag*, vol. 57, no. 1, pp. 196–204, 2009, doi: 10.1109/TAP.2008.2009698.

[45] M. C. Llasat, ‘An objective classification of rainfall events on the basis of their convective features: Application to rainfall intensity in the Northeast of Spain’, *International Journal of Climatology*, vol. 21, no. 11, pp. 1385–1400, Sep. 2001, doi: 10.1002/joc.692.

[46] Z. Rulfová and J. Kyselý, ‘Disaggregating convective and stratiform precipitation from station weather data’, *Atmos Res*, vol. 134, pp. 100–115, Dec. 2013, doi: 10.1016/j.atmosres.2013.07.015.

[47] M. Thurai, P. N. Gatlin, and V. N. Bringi, ‘Separating stratiform and convective rain types based on the drop size distribution characteristics using 2D video disdrometer data’, *Atmos Res*, vol. 169, pp. 416–423, Mar. 2016, doi: 10.1016/j.atmosres.2015.04.011.

[48] H. Feidas and A. Giannakos, ‘Classifying convective and stratiform rain using multispectral infrared Meteosat Second Generation satellite data’, *Theor Appl Climatol*, vol. 108, no. 3–4, pp. 613–630, 2012, doi: 10.1007/s00704-011-0557-y.

[49] ‘Introduction — River Basin Management Plans 2022-2027’. https://sgbp.integraalwaterbeleid.be/bekkens/dijle-en-zennebekken/kennismaking (accessed Aug. 07, 2023).

[50] Z. Liu, Q. Dai, and L. Zhuo, ‘Relationship between Rainfall Variability and the Predictability of Radar Rainfall Nowcasting Models’, doi: 10.3390/atmos10080458.

[51] R. Imhoff *et al.*, ‘EGU23-3147, updated on 07 Sep 2023 Seamless rainfall and discharge forecasting using a scale-dependent blending of ensemble rainfall nowcasts and NWP’, doi: 10.5194/egusphere-egu23-3147.