Technical University of Denmark

MAITE MONICA LOVRING (S102545)

COMPARISON OF RAINFALL FORECASTS AND EVALUATION OF FLOOD PREDICTIONS

MASTER'S THESIS - JULY 2016

DTU Environment
Department of Environmental Engineering

KRÜGER VEOLIA

English title:
Comparison of Rainfall Forecasts and Evaluation of Flood Predictions

Dansk titel:
Sammenligning af regn forecasts og evaluering af forudsigelser af oversvømmelser

Author:
Maite Monica Lovring (s102545)

Supervisors:
Peter Steen Mikkelsen, Professor, DTU Environment
Roland Löwe, Postdoc, DTU Environment
Vianney Augustin Thomas Courdent, PhD Student, DTU Environment
Steen O. Petersen, Project Manager, Krüger A/S

Project period: 8th February 2016 − 28th July 2016
Size: 30 ETCS
Education: MSc in Environmental Engineering
University: Technical University of Denmark
Institute: Department of Environmental Engineering
Classification: Not confidential

DTU Environment
Technical University of Denmark
Bygningstorvet
Bygning 115
2800 Kongens Lyngby
Denmark
Email: info@env.dtu.dk
Phone: 45 25 16 00

Abstract

Climate change and growing urbanization increase the pressure on the drainage systems and thereby the risk of pluvial flooding in Copenhagen. The city has experienced extreme rainfall events in the last years, highlighting the need for adaptation. OMOVAST is a development project conducted by Krüger A/S, in cooperation with DMI, HOFOR, BIOFOS and Nordvand A/S, to improve warnings of heavy rain and potential flooding in the Greater Copenhagen area. The project resulted in the tool SURFF, which generates flood predictions based on precipitation forecasts. This study builds on the OMOVAST project and has investigated the potential of three precipitation forecast products as input for flood prediction in Greater Copenhagen: a Radar Nowcast (Radar), a deterministic Numerical Weather Prediction with radar data assimilation (NWP) and an Ensemble Numerical Weather Prediction (Ensemble), by comparing them against rain gauge measurements. The study is separated in two parts: A general precipitation forecast analysis, aiming at comparing long term performance of the precipitation forecasts at different temporal and spatial scales, to identify high performance situations relevant for flood forecasting; and a more specific event forecast analysis, aiming at investigating the potential of the products in flood prediction, based on precipitation and flood forecasting, during four events: 1) 2nd of July 2011, 2) 30th-31st of August 2014, 3) 4th of September 2015 and 4) 15-16th of June 2016.

In Part I, investigating the forecast products performance over lead time proved that Radar has the potential for providing more precise forecasts than NWP and Ensemble on short horizons. However, as the quality of the product deteriorates quickly, a processing time of 30 min reduces the possible gains. A clear underestimation of the rain was documented for Radar, and thus correction might improve its performance further. NWP and Ensemble performed similarly, however better performance was sometimes seen for Ensemble for longer lead time suggesting a longer spin-up time. Applying a conservative approach to the ensemble members further increased its performance, however at the cost of false alarms. Finally, expanding the warning area temporally and spatially improved the performance of all forecast products. In Part II, visual investigations revealed that NWP underestimates rainfall amounts and intensities while Ensemble overestimates, although influenced by the selected approach. Simple correction of Radar for underestimation did not suffice, as high, narrow peak intensities were observed, in spite of an overall underestimation. Difficulties of the extrapolation method in predicting development of events also became clearer, which made Radar less desirable for flood prediction. A simple investigation of flood prediction using MIKE FLOOD revealed similar trends as seen for the precipitation forecasts, and indicated that Ensemble might be the superior product despite its lower temporal resolution. Based on these results a combination of the three products is proposed as the best solution for warning systems, as the three forecast products have different optimal lead times and all showed potential as input for flood forecasting. The study indicates a possible advantage for flood prediction in including Ensemble as input, as it would enable the extension of the forecast horizon but involve a compromise on temporal and spatial resolution. Further analysis is therefore needed to highlight possible drawbacks of temporal and spatial upscaling and to quantitatively compare the performance of the three products in flood prediction.

Preface & Acknowledgments

This thesis constitutes to the completion of the Master of Science degree in Environmental Engineering at the Technical University of Denmark, department of Environmental Engineering. The study was carried out from the 8th of February to the 28 of July 2016, and corresponds to a workload of 30 ECTS. The thesis was conducted under supervision of Professor Peter Steen Mikkelsen, Postdoc Roland Löwe and PhD Student Vianney Augustin Thomas Courdent, at DTU Environment as well as External Supervisor Steen O. Petersen at Krüger A/S. The study was conducted in cooperation with the company Krüger A/S. Precipitation data was provided by the Danish Meteorological Institute, DMI, and a MIKE URBAN model of Greater Copenhagen was provided by HOFOR.

First of all, I would like to thank my supervisors, Peter Steen Mikkelsen, Steen O. Petersen, Roland Löwe and Vianney Augustin Thomas Courdent for excellent meetings and support. Special thanks to Roland Löwe for many helpful talks and guidance throughout the study, and to Vianney Augustin Thomas Courdent for providing Ensemble Numerical Weather Predictions and sharing his knowledge and experience regarding weather forecasting.

Thanks to the team at Krüger A/S for providing a workspace with a nice atmosphere, allowing me to be involved in the OMOVAST project and showing interest in my study. Thanks to DMI for providing the data I needed, and a special thanks to Henrik Vedel for his great interest, helpful attitude and supervision throughout the project.

Finally, big thanks to my friends and family for proofreading and support throughout the study. Special thanks to Cécile Kittel for always knowing the optimal sentence.

Kongens Lyngby, July 2016

Maite Monica Lovring (s102545)

Abbreviations

CSI	Critical Success Index
DEM	Digital Elevation Model
DMI	Danish Meteorological Institute
DHI	Danish Hydraulic Institute
EM	Ensemble Member
EPS	Ensemble Prediction System
FAR	False Alarm Rate
FARO	False Alarm Ratio
FBI	Frequency Bias Index
FC	Forecast
HR	Hit Rate
IDW	Inverse Distance Weighting
LTS	Lead Time Steps
NSE	Nash-Sutcliffe Efficiency coefficient
NWP	Numerical Weather Prediction
PC	Proportion Correct
PSS	Peirce Skill Score
RG	Rain Gauge
RMSE	Root Mean Square Error
WBE	Water Balance Error

Definitions

Forecast (FC): Every time a forecast product generates a new prediction of future precipitation this prediction is called a forecast. Thus it may also be referred to as the prediction. Each forecast is considered as a separate dataset. The word may also be used as synonym for forecast products.

Forecast frequency: The time between each generated forecast.

Forecast product: The type of forecast data generated by a certain model is referred to as a forecast product.

Horizon: The temporal length of the forecast into the future. Can also be referred to as the lead time of the forecast.

Lead time: The temporal length of the forecast into the future. Can also be referred to as the horizon of the forecast.

Lead time step (LTS): The interval in time between each value in a forecast. Can also be referred to as the temporal resolution of the forecast.

Prediction: The same as a forecast. The two words are used interchangeably in this study.

Spatial resolution: As the forecasts are gridded the spatial resolution is the resolution of the grid for each forecast product.

Temporal resolution: The interval in time between each value in a dataset. Also referred to as the lead time steps for the forecast products or just as the time steps for measured data. In some cases similar to aggregation level, as the datasets are aggregated into desired temporal resolutions.

Table of Contents

CHAPTER 1

Introduction

1.1 Background

To forecast is attempting to avoid complete ignorance of the future. Prediction of weather patterns has long been an important research field for meteorologists, and many different techniques have been developed in this area. Forecasting of precipitation is especially important in an urban context, as many parts of the human infrastructure are affected by rain. Rainfall turns into runoff, which has to find its way out of the city either by artificial or natural paths. However, increasing urbanization removes the possibilities of natural infiltration, as the amount of impervious area increases, and other drainage paths are needed.

Combined sewer systems are used in Copenhagen and its suburbs to transport the mixed rain- and wastewater to wastewater treatment plants, where it is treated and discharged to surface water bodies. All components of a combined sewer system are thus affected by precipitation. System saturation may have many consequences including combined sewer overflows to surface water bodies, decreased efficiency of waste water treatment plants, changes in nutrient loads and risk of flooding on the surface. Flooding is often caused by large rain events, exceeding the capacity of the sewer systems. Investments in improving the capacity of drainage systems and other innovative solutions to remove water from the system are conducted in several cases to minimize the risks of flooding. There will however always be events which exceed the capacity of the system. Being able to predict surface runoff and flow in the drainage system allows prediction of risks of flooding, which is relevant in minimizing the costs related to flood damages.

The Intergovernmental Panel on Climate Change estimates that climate change will affect precipitation patterns resulting in more frequent and extreme precipitation events in Northern Europe (Barros et al., 2014). This, coupled with increasing urbanization and thus increasing pressure on the drainage systems, will result in increasing risk of pluvial flooding. This risk has been stressed by a number of extreme rain events in the previous years causing flooding in the Copenhagen area. In August 2010 and July 2011 Copenhagen was hit by two cloudbursts with measured rain amounts up to 96,8 and 132 mm in 24 hours (Thomsen, 2011, 2012). In August 2014 another large event was observed with 135 mm in 24 hours (Hansen and Pedersen, 2014), and in September 2015 Copenhagen once again experienced large amounts of rain with 44 mm in five hours (Siewertsen, 2015). The maximum observed mean intensities over 10 minutes (min) for the four events were respectively 26 $\mu m/s$, 52 $\mu m/s$, 33 $\mu m/s$ and 29 $\mu m/s$. This means that all four events can be classified as events with a return period of close to or above 20 years (Thomsen, 2011, 2012,

2015, 2016). It should be noted that while these 10 min values give a good impression on the rarity of the events, much longer time horizons are relevant when considering runoff, especially in a larger urban catchment like Copenhagen. In general, DMI noted that rain events in Denmark are becoming more and more intense and extreme (Cappelen and Scharling, 2010). Increase in extreme events, means an increase in the needed capacity of the drainage systems, however as this is often not possible through simple investments, it also increases the importance of predicting risk of flooding in an attempt to avoid damages.

1.2 From Rainfall to Urban Runoff

Urban hydrologist are very interested in precipitation as it is the most important input variable in describing runoff and flow processes. The path from rainfall to runoff is however complex and thus looking at rainfall characteristics alone does not provide answers on urban hydrology as the spatial/temporal distributions of causes and effects are not the same (Schilling, 1991).

Urban catchments have faster response times than natural catchments as there is less natural retention. More detailed rainfall input data is needed to describe urban catchments and understand their hydrological storage processes, as large variability can occur in runoff production in one catchment. Schilling and Fuchs (1986) proved that higher spatial resolution of rain data improves runoff model results, however depending on the interest of the hydrologist, different resolution levels are necessary. For analysing and operating existing systems a higher resolution is needed compared to the design or evaluation of a new system. According to Schilling (1991) the minimum spatial and temporal resolutions needed for this type of tasks are 1 km and 1-5 min, respectively. Using too long time steps results in systematic underestimation of peak runoff, especially if the catchment has a fast response rate (Schilling, 1991). This can be a problem for example for real time control of a system as both the runoff volume and peak intensity are important factors. Ochoa-Rodriguez et al. (2015) investigated different combinations of spatial and temporal resolutions of radar observations for nine different events on seven small urban catchments. They concluded that the impact of the input resolution decreases rapidly with increasing catchment size. However, they also expressed the need for higher resolution of precipitation products for the high resolution runoff models used for urban catchments, and pointed out that this is currently the area where improvement is most needed. It is thus clear, that a good cooperation between hydrologists and meteorologists is necessary to improve rainfall-runoff modelling.

Error in input data is one type of errors in flow simulations along with the simplification of the physical setup and model parameters describing unknown conditions that cannot be measured (Schilling and Fuchs, 1986). The majority of the input error is caused by the difficulty of measuring precipitation throughout the entire catchment. Schilling and Fuchs (1986) proved that introducing more precipitation measurement points reduced runoff value and peak flow errors more than modifying model components. Input resolution is the most limiting factor on model accuracy and they therefore suggested that using a faster, more simplified model but increasing input resolution might provide better results for real time applications, where computation time often is a limiting factor. It should be noted that temporal and spatial resolutions are linked, i.e. for a given temporal resolution there is a spatial resolution which minimizes the error (Fabry et al., 1994, as cited in (Schellart et al., 2012)). This was also observed by Ochoa-Rodriguez et al. (2015), who found a strong interaction between the spatial and temporal resolution of rainfall input.

From these observations it is clear that high resolution rainfall input is important in urban flood simulations. Precipitation forecasts products are necessary inputs to create flood predictions, and the previously mentioned studies justify the need of investigating the potential of forecast products

with different temporal and spatial resolutions for urban flood prediction. To model flow in drainage systems the software MIKE URBAN by DHI is traditionally used in Denmark. MIKE URBAN is a software that can be used for modelling hydraulics in urban drainage systems (DHI, 2014c). By extending this model to include a 2D hydrodynamic surface model, the flooding on the surface can be simulated. This is done using the MIKE FLOOD tool, which integrates the 1D model of MIKE URBAN with a 2D surface model of MIKE 21 (DHI, 2014b). Input data to these models are precipitation time series. Thus by using forecast data as input, the expected flow and flooding can be computed and a warning system can be based on these calculations. This is the basis for the OMOVAST project conducted by Krüger A/S, which will be described in section 1.3.

1.3 The OMOVAST Project

OMOVAST (Operativ Model til Varsling og Styring - Operational Model for Warning and Control) is a development project, which began in 2014, and is conducted in cooperation between DMI, HOFOR, BIOFOS and Krüger A/S. The aim of the project is to improve warnings of heavy rain and possible flooding for the Greater Copenhagen area and thereby contribute to secure the city against climate change. The project includes improving the forecast of rain events, the development of a tool to estimate flood extents, providing flood forecasts with up to 6 hours lead time every hour, and a flood warning system. Besides this, part of the OMOVAST objective is also to provide a tool for simulating the effect of climate change adaptation measures. The project demonstrates the use of the developed tool on two selected catchments - the drainage areas of the Lynetten and Spildevandscenter Avedøre treatment plants (Brødbæk et al., 2015). The second phase of the OMOVAST project began in the end of 2014 by the inclusion of Nordvand A/S in the project, and thus a third area was included in the model around Gladsaxe and Søborg. The second phase focuses on improving and expanding the OMOVAST product (Krüger, 2015).

The tool developed in the OMOVAST project is called SURFF. It was developed in Python, and executes 1D-2D hydraulic simulations in MIKE FLOOD. SURFF includes MIKE URBAN models of the drainage systems of the three catchments, and by extending this with 2D MIKE FLOOD models of the relevant areas, a flood forecast animation can be created showing the extent of the resulting flooding of a certain precipitation event. Precipitation forecasts provided by DMI are used as input data. The forecasts are based on a Numerical Weather Prediction model (NWP) with a 3x3 km grid and a lead time of 8 hours. A new forecast is generated every hour, however as it takes two hours to generate the forecast, only a lead time of 6 hours is available for OMOVAST (section 2.1.3). SURFF also uses sea level and rain gauge measurements to calculate initial conditions for the model. Each time a precipitation forecast is received from DMI, the model generates a flood forecast. If flooding is forecasted a warning message is generated and sent to selected users. The model is operational and runs automatically all year round. Registered users can access the weather and flood predictions online at `surff.dk`.

1.4 Aim and Objectives

The overall aim of this study is to investigate the potential of different precipitation forecast inputs for flood forecasting in an urban context. More specifically, the study aims at comparing long term performance of different precipitation forecast products over different temporal and spatial scales, to shed light on the strengths and weaknesses of the forecast products in different situations and combinations. Putting these investigations in the light of flood prediction, the performance of the products, both in terms of precipitation and flood forecasting, is analysed for selected flooding events to investigate the potential and usefulness of the forecast products in flood prediction.

The project is therefore separated in two parts: Part one is a thorough, general analysis of the forecast products aiming at identifying differences in performance, while part two is a sorter, scenario based analysis aiming at documenting the identified trends for cases relevant in the context of this study. The individual objectives are described in the following subsections.

1.4.1 Part I: Long Term Precipitation Forecast Analysis

The main objective of this part of the study is to systematically investigate and compare three different types of rainfall forecasts: a Radar Nowcast (Radar), a deterministic Numerical Weather Prediction with radar data assimilation (NWP) and an Ensemble Numerical Weather Prediction system (Ensemble). The forecast products are compared against rain gauge (RG) observations, at different forecast horizons and at different temporal and spatial scales, using data from a longer historical time period. A major objective of this part is to identify in which situations and for which forecast horizons the individual forecast products are preferable over the others. Focus is on situations relevant for use in flood warnings and control, and thus especially comparing change in performance over lead time is relevant to get indications on which time scales the different products are more or less reliable. Finally, it is an objective to investigate effects of applying different spatial and temporal perspectives relevant for warning systems.

1.4.2 Part II: Event Precipitation & Flood Forecast Analysis

A natural extension from the broader investigations conducted in Part I is an event analysis focusing on extreme precipitation events and the resulting flooding. Thus for the second part, a number of high intensity rain events, where forecast quality could be evaluated, were selected: 1) 2nd of July 2011, 2) 30th-31st of August 2014, 3) 4th of September 2015 and 4) 15-16th of June 2016. The main objectives of the second part of the study is therefore to 1) visually investigate the performance of the precipitation forecast products in time and space during these events, and 2) evaluate the quality of flood forecasts of these events by comparing MIKE FLOOD simulations based on the three forecast products against simulations based on rain gauge input.

The study area of this project is the area relevant for OMOVAST: Greater Copenhagen. The spatial boundaries of the used precipitation products was chosen to ensure that the entire Greater Copenhagen was covered with some buffer distance, thus covering the three SURFF catchments. For the flood modelling, focus is put on one of the three catchments included in the OMOVAST project: the Lynetten catchment. This catchment covers Central Copenhagen and areas that were flooded in the selected events. The extent of the different datasets can be seen on Figure 1.1 along with the Lynetten catchment.

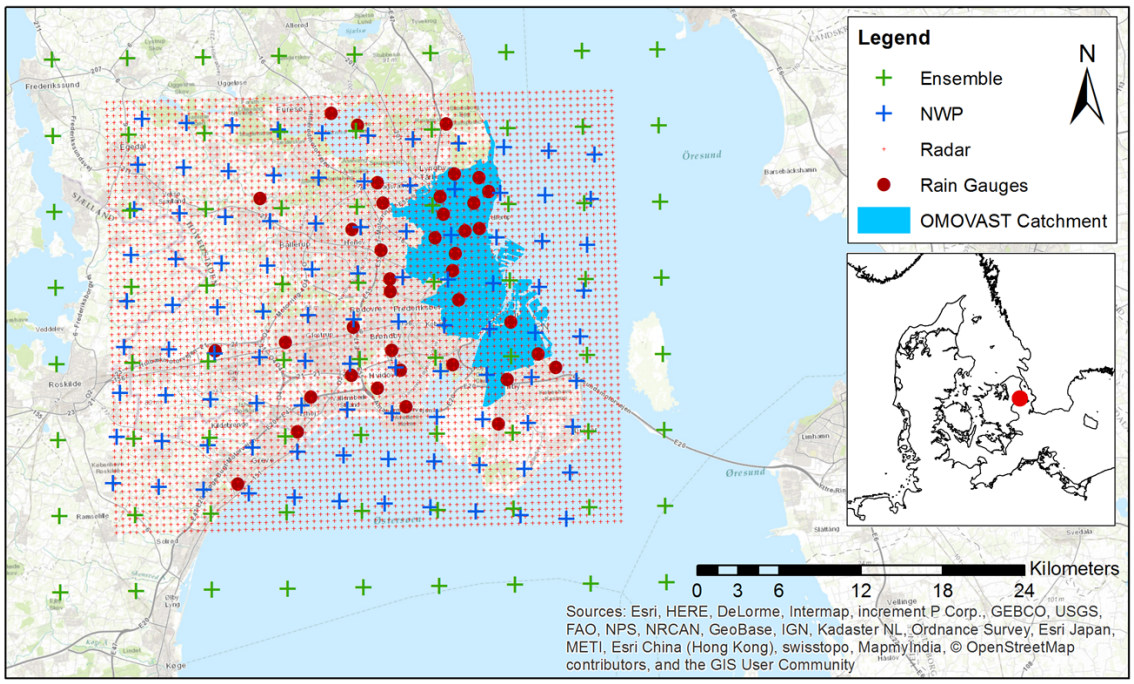


Figure 1.1: *Map of Greater Copenhagen with the centerpoints of the NWP, Radar and Ensemble Grid, as well as the RG and the catchment used in SURFF.*

For the first part of the study all data handling and calculations are done using R version 3.2.3 with R Studio version 0.99.489. In the second part a MIKE URBAN and MIKE FLOOD model of the selected catchment is used, along with data handling in R.

1.5 Outline of Thesis

The analyses conducted in this thesis are separated in two parts, however both parts investigate the same forecast products. The thesis is therefore structured as follows: Chapter 1 is the introduction. Chapter 2 describes the precipitation data used in this study and the different data conversions applied. Chapter 3 describes the methods used for Part I while Chapter 4 presents the results, discussions and conclusions of this part. Chapter 5 describes the methods used in Part II while Chapter 6 presents the results, discussions and conclusions of this part. Finally, the results and discussions from the two parts of the project are combined and overall conclusions are drawn in Chapter 7.

Working with several forecast products involves many spatial and temporal scales as well as many similar expressions. To avoid confusion and ensure consistency in the use of words, common terms have been defined on page v, and all abbreviations used in this report can be found on page iv.

CHAPTER 2

Precipitation Data

2.1 Data Types

Four types of precipitation data are used in this study: 1) Measured Rain Gauge data, 3) Radar Nowcast data, 2) Numerical Weather Prediction data with data assimilation from radar observations, and 4) Ensemble Numerical Weather Prediction data. The four data types are described in the following sections. An overview of relevant information about the different precipitation datasets can be seen in Table 2.1. The four types are hereafter referred to as RG, Radar, NWP and Ensemble in this study.

Table 2.1: *Overview of the four different precipitation data used in this study, with information on original data characteristics.*

	RG	**Radar**	**NWP**	**Ensemble**
Data period	07/05 2014 05:12 - 09/02 2016 09:30	04/08 2015 19:00 - 25/01 2016 09:20	01/07 2014 00:00 - 25/01 2016 08:00	01/09 2014 00:00 - 31/01 2016 18:00
Spatial resolution	Point obs.	500 m	3 km	5 km
Forecast frequency	-	10 min	1 hour	6 hours
Forecast lead time	-	180 min	8 hours (2 + 6 hours)	53 hours
Temporal resolution	1 min	2 min (first hour) + 10 min (next hours)	10 min	1 hour
Unit	mm/min	mm/h	mm acc.	mm/h

2.1.1 Rain Gauge data

DMI has a network of rain gauges (RGs) across Denmark in collaboration with the Water Pollution Committee of the Society of Danish Engineers (Spildevandskomitéen, SVK). The network consists of 154 RGs (as of 2014) of the tipping bucket type with 20 cm diameter and a tip resolution of 0.2 mm (Thomsen, 2015). Of these, 38 are located in the area covered by the OMOVAST project (Figure 1.1). Five are used as hotstart stations, i.e. to calculate initial conditions for the SURFF model, for the Copenhagen city area: 5725, 5740, 5745, 5765 and 5705 (hereafter referred to as the five Copenhagen stations), one as hotstart for the Avedøre area: 5804, and three for the Gladsaxe

and Søborg area: 5645, 5655 and 5660. It should be noted that station 5628, Mølleåværket, was moved from outside the area of this study to Mølleåværket in 2014, which is located within the area relevant for the OMOVAST project. The station was thus given the coordinates as described in Thomsen (2015) and included in the analyses. A list of the stations with numbers, names, start date, height and location as well as a zoomed map with all RG numbers can be found in Appendix A.

The measured data from the RGs is provided by DMI in a format including information on the quality of the data, as all data has been through an automatic quality control (Thomsen, 2015). From the control, doubtful data is marked with letters indicating whether the doubt is caused by either maximum deviation from the nearest meter, technical error, interruption, snow or extreme precipitation intensities ($> 2mm/min$). The snow indication automatically applies when the temperature gets below 3°C. The extreme precipitation intensities are manually assessed by a climatologist afterwards (Thomsen, 2016). For this study, the data was converted to a continuous time series by inserting 0 values for dry periods and removing data marked with low quality or as faulty, except data marked as extreme rainfall or snow. Indeed, these might be flawed but are important in the context of this study, while keeping in mind that snow will result in delayed runoff. Due to the measurement method it is not possible to know the exact start of a precipitation event, however they are assumed to start at the first tip of the bucket. Thus the first value of an event is always $3.333\mu m/s$ (0.2 mm in one min). These values are kept in the data even though they can be misleading, as removing them would mean removing water from the system, possibly affecting water balance comparisons.

The RG data is provided in a temporal resolution of 1 min, and was aggregated to the following levels: 2 min, 10 min, 30 min and 1 hour (h) resolution. Each time step was marked with the timestamp of the end of the step, e.g. the 1 min values from 12.00-12.29 were aggregated to one value with the stamp 12.30. By using the timestamp of the end of the aggregation, one value represents the average intensity observed over the period up to that time. When aggregating to larger time steps, empty values (Not Available, NA) are excluded unless all values within the aggregated step are NA. Preliminary investigation of the data showed that the majority of the NA values (78 %) originated from station 5715, Bispebjerg Hospital. As it was evident that this station did not measure any rain for the period with all forecast data available, it was decided to exclude the station, and thus only 37 stations are included in the study.

The RG measurements are used as reference data for analysing the quality of the historical forecast data. They are thus assumed to represent reality, however uncertainties are related to the measurement method and quality check of the data. The quality control of the data is essential in removing faulty values and obtaining a reference dataset as close to reality as possible. However, one must still take into account the random and systematic errors related to the measurement method, which might move the data in a certain direction compared to the truth, e.g. it is commonly observed that gauges underestimate rain due to wind or wetting losses (Schilling, 1991). The ability of a gauge to detect the precipitation also varies with the type and intensity of the precipitation. Systematic errors can be accounted for, however uncertainties related to the correction methods are also present (Thomsen, 2016). In the end, this means that achieving an exact match between the forecasts and the RGs is not only impossible but also not desirable.

2.1.2 Radar Nowcast data

The Radar Nowcast data used in this study is also provided by DMI and created using a radar extrapolation model (Jensen et al., 2015). It is referred to as a nowcast, due to the short forecast length of 3 hours. The radar observations used to create the nowcast are obtained from a radar station at Stevns (approx 50 km from Copenhagen). It is a C-Band Doppler radar, and the data is

quality checked and treated for external noise prior to the extrapolation. The extrapolation process is described in Jensen et al. (2015) and Jensen (2015). It should be noted, that the radar observations are converted using standard Marshall-Palmer coefficients, which provides the best results for a certain type of droplets and thus for a certain type of rain event (Jensen et al., 2015).

The Radar data has a spatial resolution of 500x500 m and covers approx. the same area as the NWP, with a total coverage area of approx 1100 km^2 (Figure 1.1). It has a lead time of 3 hours, with a temporal resolution of 2 min for the first hour and 10 min for the following two hours of prediction. A new forecast is provided every 10 min. The Radar data thus has a higher spatial and temporal resolution than the NWP model, however the lead time is shorter, and prediction quality is also expected to deteriorate quickly as it is based only on an extrapolation of the current situation and without taking meteorological processes into account (Jensen, 2015).

As the original resolution changes over the lead time, it complicates the comparison process. To overcome this, a 10 min aggregation was done of the first hour, resulting in a temporal resolution of 10 min for the entire forecast horizon. The Radar data was also aggregated to 30 min and 1 hour temporal resolutions. Spatially, the Radar data was aggregated onto the NWP grid to reduce computation time for the long term analysis conducted in Part I. This was done using a spatial interpolation with inverse distance weighting of Radar data within a cutoff distance of 2 km from each NWP grid point. For further discussion of this procedure, see section 2.3.

Investigation of NA values showed that in 33 forecasts one time step contained NA values, however these were removed in the aggregation process. Forecasts with NA values were inserted to patch the historical time series, thus ensuring a continuous time series with one forecasts every 10 min. About 28 % of the forecasts were missing from the entire data period, and it was noted that a quite high percentage of the total dataset was now NA values.

Finally it was noted that the Radar Nowcast is delivered to the user around 30 minutes after the start time of the forecast, due to processing time. For this reason, it was decided to consider the first 30 minutes unavailable for forecasting. However, it should be noted that, in contrary to the NWP data, the Radar nowcast model is not adjusted during the forecasts horizon, and thus the delay only reflects processing time in this case.

2.1.3 OMOVAST NWP data

The Numerical Weather Prediction (NWP) used for the OMOVAST project is also provided by DMI. NWP models at DMI are based on the HIRLAM system (High-Resolution Limited Area Model, Unden et al. (2002)), which is an international cooperation of European meteorological institutes, see `hirlam.org` for more information. The OMOVAST NWP model is an operational hydrostatic model that runs every hour and produces an 8 hour deterministic forecast with a temporal resolution of 10 min. At DMI the model is referred to as HIRLAM RA3, however in this project it will be referred to as NWP. The model covers Greater Copenhagen with a grid of 3x3 km spatial resolution and covering an area of approx. 900 km^2 in total (Figure 1.1).

The generation of the NWP forecasts consists of two parts: a data assimilation part, where meteorological data from the last 1.5 hours is collected and treated in the model, and a forecasting part, which includes nudging of cloud observations and Radar observations into the first half hour of the prediction period. The technique of nudging Radar data into NWP models has been described in (Korsholm et al., 2015). Thus the first two hours are used to generate the prediction and are not available for forecasting, resulting in an available forecast length of 6 hours (Olsen et al., 2015). However in this study, most analyses are conducted on the entire lead time, as this allows to see the

effects of data assimilation and nudging in the first two hours. Besides this it should be noted that NWP models need spin up time, and the forecast quality is therefore not unlikely to improve with increasing lead time (Thorndahl et al., 2012).

The data is provided as accumulated values with 49 time steps marked with timestamps of 10 min intervals (from 0 to 480 min). Thus the data was transformed into average values with a temporal resolution of 10 min. The forecasts were also aggregated into 30 min and 1 hour resolutions using the same procedure as for the RG data. The provided format of the data left room for doubt about how to interpret the first data point (marked with timestamp 0 min). To ensure proper data handling, it was investigated which way of reading the data resulted in the best performance, by using similar procedures as for the binary threshold test (section 4.2.1) on the 10 min data. The investigation can be seen Appendix B.1. It gave sufficient evidence, that the first data point should be considered as an initialization step, similar to the format of the Ensemble data (section 2.1.4), meaning that the first data point can be disregarded. As with the Radar data, the NWP was aggregated to 30 min and 1 hour temporal resolutions.

Investigation of the data showed that the forecast generated at 2014-08-30 22:00 included only 1/3 of the stations, and the rest was thus replaced with NA values. Besides this, no NA values were observed in the other forecasts. The historical time series of hourly forecasts was not continuous, as about 3 % of the forecasts were missing. Thus to obtain continuous historical time series of hourly forecasts, the time series was patched with forecasts consisting of NA values for all missing forecasts, as done for the Radar data.

Finally, it should be noted that DMI changed the setup of the NWP model at some point in the period of the provided data. It was expected to be around September 2015 however, to investigate whether this change resulted in noteworthy changes in model performance, the average monthly quality was computed using a similar procedure as the one described in section 4.2.1. The results of the test can be seen in Appendix B.2. They showed no visible difference or trend in quality on a monthly average, when comparing data from 2014 to data from 2015. Thus it was concluded that the change in model setup did not affect the model performance significantly. It was also noted that the expected time of change in the model is outside the period where data is available for all forecasts products, and thus the shift will not affect most analyses performed in this study.

2.1.4 Ensemble NWP data

The last forecasts product is an Ensemble Prediction System (EPS) with 25 scenarios or Ensemble Members (EMs). The system is based on the DMI-HIRLAM-S05 NWP model which runs every 6th hour at 00:00, 06:00, 12:00 and 18:00 UTC producing a forecast with a lead time of up to 53 hours and a resolution of 1 hour. The spatial resolution of the grid is approx. 5x5 km. The system covers Scandinavia and Northern Europe, however only the part of the grid covering Greater Copenhagen including a buffer distance was included in this study, as shown in Figure 1.1, resulting in a total covered area of about 1700 km^2.

The 25 scenarios are based on five different initial conditions generated from measured atmospheric conditions, and two different model schemes for modelling convection and condensation processes both run with or without stochastic physics. The last five EMs include parameters for studying the interaction between land cover and the atmosphere with randomly chosen values. For further description of the EPS and the difference between its members see Feddersen (2009). As the 25 EMs have different model setups a systematic difference between the members could be expected, especially in relation to their prediction of a certain type of events like high intensity events. Thus in some cases the EMs were investigated individually to check for differences in performance.

Having 25 EMs allows alternative ways of using the product, and thus different investigations are useful for this forecast product compared to the ones used for the other two products, as will be described in section 3.3.

The Ensemble data has an initialization step of 144 seconds, which is excluded from the analysis. The model provides both data on total precipitation and convective precipitation, however only the former is of interest for this study. The processing time of the EPS is unknown, however a delay due to computation time can be expected, which means that the data is not available to the user at the start time of the forecast. A new forecast is produced only every 6th hour indicating longer computation time, thus it was assumed that the first three hours of a forecast is not available to the user. It should be noted that there is no radar data nudging involved in the Ensemble as was the case for NWP (Feddersen, 2009).

All faulty forecasts were removed prior to delivery, thus no NA values were present in the data, however to create a historical time series with a forecast every 6th hour, the same procedure as for Radar and NWP was followed and the time series was patched with forecasts with NA values. Approx. 6 % of the forecasts were missing in the provided period and replaced by NA values.

2.2 Obtaining Comparable Formats Temporally

Identical data formats both temporally and spatially are necessary to apply a common analysis procedure and enable comparison of the data. Fist of all, the datasets must be compared using the same historical period. The radar data is the limiting factor on the period to include in the long term forecast performance analysis. A common period was thus defined from 05/08 2015 at 00:00 to 25/01 2016 at 00:00, as this period has data available for all datasets. This is hereafter referred to as the common period. For the NWP a longer period was however also defined: 01/07 2014 at 08:00 to 25/01 2016 at 08:00, exploiting the availability of the larger amount of data to investigate whether trends seen from the shorter period are representative for the forecast product over a longer period.

The temporal resolution of the Ensemble data is the lowest and thus the limiting factor for obtaining a common temporal scale. Most investigations were thus made on a 1h temporal resolution for all three forecast products. However for both Radar and NWP data, analyses were also made using 10 min and 30 min temporal resolutions. All comparisons were done against RG data with the same aggregation level, as shown in Figure 2.1 and 2.2.

Working with historical forecasts from a certain time period involves two temporal dimensions: 1) The forecast horizon of each generated forecasts with a certain temporal resolution, and 2) The time series of generated forecasts. In order t consider all generated forecasts from the period as a historical time series, one must decide on how to navigate between the two temporal scales. In this study two different time series generation methods were applied depending on the purpose of the analysis: A lead time step approach and a continuous time series approach.

2.2.1 Lead Time Step Time Series

In the case of the lead time step time series one data point in the time series corresponds to the chosen lead time step (LTS) of each forecast. The approach has been conceptualized in Figure 2.1 showing the case for generating a time series using the first available LTS for the three forecast products and the three temporal resolutions. This setup was used for analyses investigating quality performance over lead time, as it highlights the performance of a certain LTS for a forecast product.

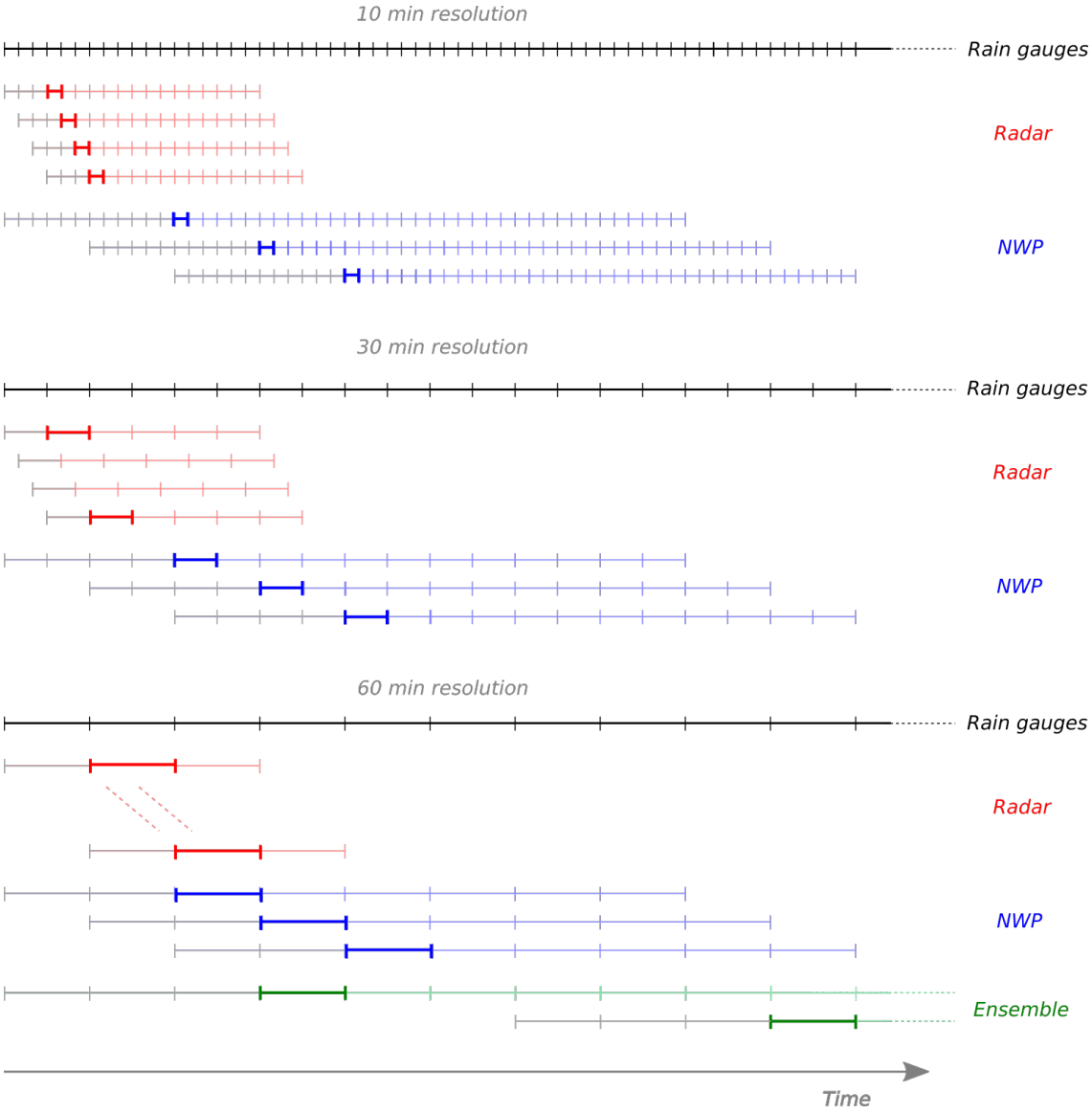


Figure 2.1: *Conceptual drawing of the Lead Time Step Time Series approach for the three forecast products and the three temporal resolutions using the first LTS. Each coloured horizontal line symbolizes a forecast. Grey parts are unavailable for forecasting due to processing time. The slanted dotted lines for 1 hour resolution Radar data indicates the 6 skipped forecasts.*

The drawback of this type of time series is that it is not necessarily continuous due to differences in forecast generation interval and temporal resolution. The forecast generation interval varies for the different products. For Ensemble, a new forecast is generated every 6th hour, and thus only every 6th hour can be compared to the RGs for every LTS thereby only using 1/6 of the RG dataset for these comparisons. For NWP a similar problem occurs for the 10 and 30 min LTSs, as a new forecasts is only generated every hour, and thus only 1/6 or 1/2 of the RG datasets were used for these comparisons.

2.2.2 Continuous Time Series

To achieve continuous time series for the forecasts the necessary lead time was included, meaning that for the Radar only one LTS of 10 min was included as a new forecast is created every 10 min, while for NWP and Ensemble more LTSs were included depending on the temporal resolution

investigated. The setup has been conceptualized in Figure 2.2 for the three temporal resolutions and the three forecast products. By including the first 6 LTSs for the Ensemble time series, a continuous time series with hourly values can be created. Similarly for NWP, by considering the first six or the first two LTSs of each forecast a continuous time series with respectively 10 min and 30 min values can be created.

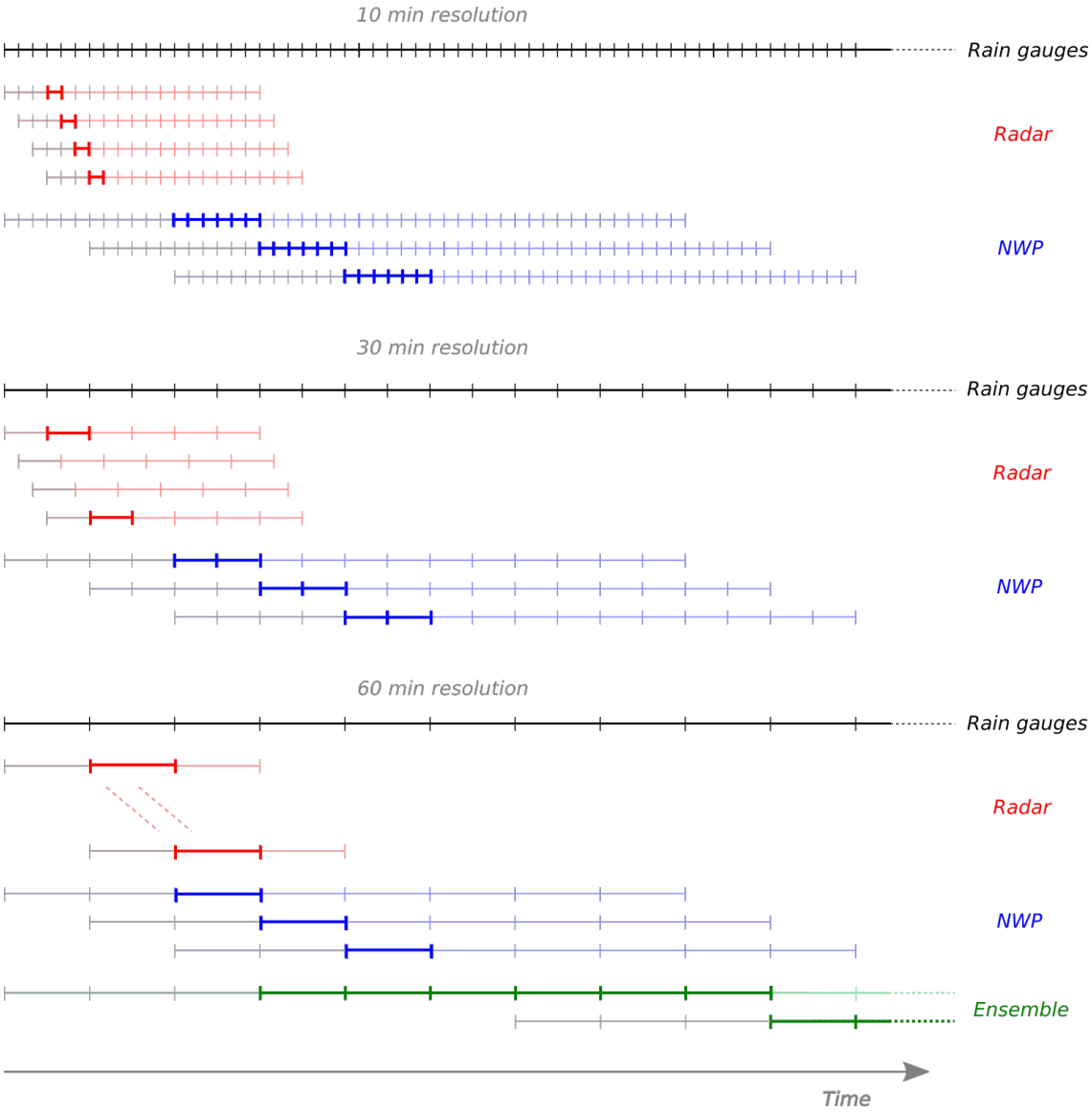


Figure 2.2: *Conceptual drawing of the Continuous Time Series approach for the three forecast products and the three temporal resolutions using the first LTSs available. Each coloured horizontal line symbolizes a forecast. Grey parts are unavailable for forecasting due to processing time. The slanted dotted lines for 1 hour resolution Radar data indicates the 6 skipped forecasts.*

This procedure was used to increase the size of the datasets in more general analyses comparing forecasts products. It highlights the performance of the lead time covering the period between forecasts, when applied on the first available LTSs. This means that for NWP the 3rd forecasts hour is used, for Radar the 4th or the 2nd LTS is used, and for Ensemble the first 4-9 lead time hours are used, as shown in Figure 2.2. These time series cannot be assumed to be representative for the entire forecasts. However they do give a good indication on the highest performance of the forecast products, when disregarding any spin-up time and assuming decreasing performance over

lead time. This should be kept in mind when interpreting results based on these time series. The procedure could however also be used on later time steps but this was not applied in this study.

The continuous time series procedure was also used to create time series of radar observations for the four selected events in the second part of the study. To do this, the first LTSs of the provided radar forecasts were used. This is actually a 10 min extrapolation of the current situation, which is however assumed to be close to reality and reflect the observed situation at the next time step.

For Radar another problem occurs when the lead time is aggregated to 30 min or 1h steps. As a new forecast is generated every 10 min, this leads to a double comparison in time. Thus it was decided to include only every 3rd or every 6th of the forecasts respectively in the analyses, as shown in Figure 2.1 and 2.2. This way, a sample of the forecast data dependant on the chosen start time is included. These samples should be representative for the entire forecast dataset, however this is not necessarily the case, especially due to the short period of data available. To ensure no statistical significance of the choice of start time, a test comparing the results obtained using different samples of the dataset could be done. However, this was considered outside the scope of this project.

2.3 Obtaining Comparable Formats Spatially

As the reference data and the forecasts are obtained by different methods and at different scales a legitimate variance between the RG data and the forecast products is expected without invalidating the forecasts. One major difference in the datasets is that the RGs are point measurements while the forecasts estimate rain over an area. To enable direct comparison between the gridded forecasts and the point observations, one must choose a matching method and convert to a common spatial format.

It is difficult to assess which is the most correct approach to achieving a common spatial resolution, as interpolating the data will always include some simplifications. Spatial interpolation with a defined cutoff distance and inverse distance weighting (IDW) was chosen as matching method in this project. An alternative interpolation is the Nearest Neighbour method, however this was considered too simple as it only allows contribution from one grid cell to each point or vice versa. Another commonly used matching method is to compute a weighted average of the four surrounding grid points for each point (Jolliffe and Stephenson, 2012). This method is easily modified to a spatial interpolation of all grid cells or points within a chosen distance. This was considered the most fitting approach for this study, as it is easily adapted to include fewer or more points or cells, and thus to the different interpolations needed for the different parts of the study. The approach can be described with equation 2.1 and 2.2.

$$U(x) = \frac{1}{N} \sum_{i=1}^{N} U_i w_i(x) \tag{2.1}$$

where U is the interpolated value at a given point x, U_i is the sample to interpolate, w_i is the weight defined by

$$w_i(x) = \begin{cases} \dfrac{1}{D(x,x_i)^2} & \text{if} D(x,x_i) \leq \text{cutoff} \\ 0 & \text{if} D(x,x_i) > \text{cutoff} \end{cases} \tag{2.2}$$

where $D(x, x_i)$ is the distance between the interpolated point and the interpolating point, and cutoff is the chosen cutoff distance. The specific interpolations used in the two parts of this study are described in the following subsections.

2.3.1 Spatial Interpolations for Part I

In Part I of this study it was decided to spatially interpolate the three forecast products from their grid scales to the RGs points, as the RGs are the reference data and their spatial distribution was not an investigated factor in this part. A cutoff distance of 5 km was chosen as the distance for relevant grid points for each RG, yielding around 300 contributing Radar grid points, around 7 contributing NWP grid points and 2-4 contributing Ensemble grid points to each RG. Again the Ensemble is the limiting factor, as this cutoff was the lowest possible that would still allow contribution from more than one nearest neighbour for all forecasts products. The spatial interpolation was done for each forecast product for each RG. A map illustrating the procedure for one RG can be seen in Figure 2.3.

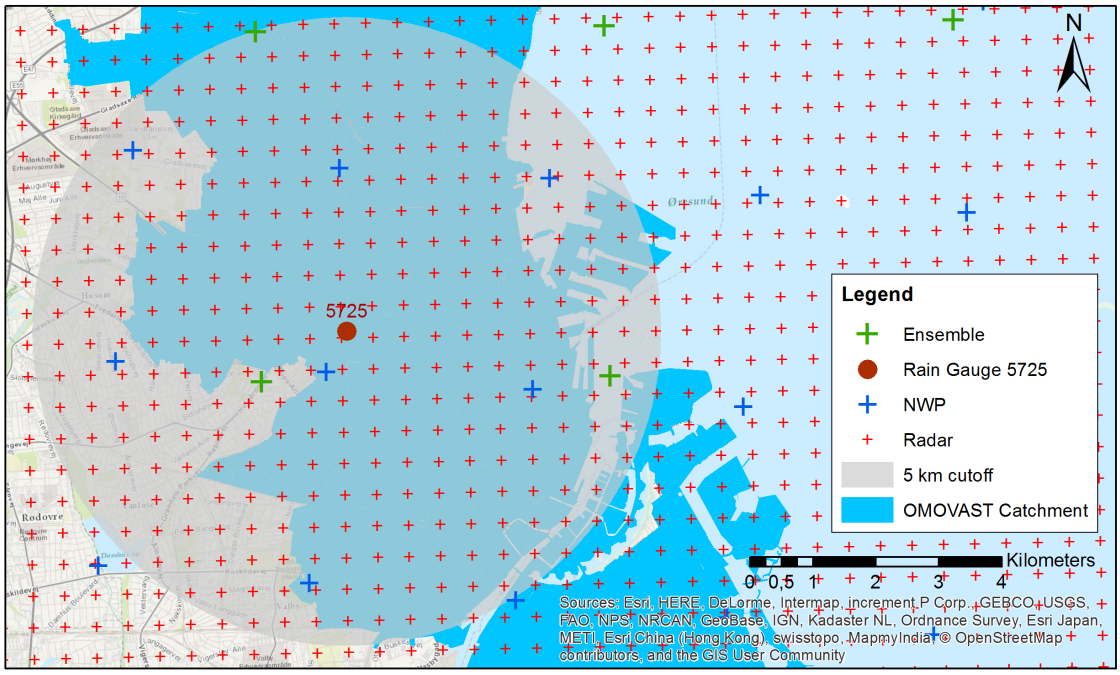


Figure 2.3: *Illustration of 5 km cutoff distance for the RG 5725 with the Radar, NWP and Ensemble grid centerpoints. All centerpoints within the grey circle would be included in the spatial interpolations for this RG and weighted according to their distance to the RG.*

It was also considered, whether to use the same distance for the three different forecast grids or to ensure a common number of contributing cells, as this could potentially provide very different results. Using the same cutoff distances results in more smoothing of data for higher spatial resolutions, as more grid points are included and averaged. However using the same number of contributing cells means that the area considered in the interpolation varies which may also alter the results as more local changes are modelled on a higher resolution grid. A simple investigation of the difference between the two methods was done for the Radar data, as the largest differences related to the choice of method were expected for this dataset. The results of the investigation can be seen in Appendix B.3 and revealed that over the entire Radar period a residual of 0.4 mm is seen between the two investigated interpolations. From accumulation plots and Quantile-Quantile

plots, it was concluded that the choice of distance for the spatial interpolation does not have an important impact on the results. Thus it was decided to keep the distances fixed at 5 km for all spatial interpolations to the RGs.

For the more indirect comparisons including distance as a parameter, as the ones described in section 3.4, no spatial interpolation was done for NWP and Ensemble. However as mentioned in section 2.1.2 the Radar data was spatially interpolated to the NWP grid to reduce computation time while providing a common spatial scale for the two products. Using a cutoff of 2 km resulted in an average of 50 Radar grid points contributing to each NWP grid point.

2.3.2 Spatial Interpolations for Part II

For the second part of the study, the spatial distribution of the events were an important parameter, as a spatial scale is necessary to relate the precipitation data to flooding. Thus the opposite approach was used, interpolating the RG measurements to the three forecast grids to obtain a spatially distributed dataset, however the same method with a cutoff distance and IDW was also used in this case.

The distribution of the gauges is not very equal, and within the relevant area, some parts are poorly covered by RGs. Using a cutoff distance of approx 15 km the entire forecast grids of Greater Copenhagen could have been covered with RG data. However this would provide a false impression on the availability of data in poorly covered areas. Taking into account the very local differences in precipitation amounts that are often observed during cloud bursts, interpolating with a too large cutoff distance might result in assigning incorrect values to grid points too far away from a RG. Based on these considerations a cutoff distance of 5 km was chosen to avoid misleading data but still ensure a complete coverage of the central part of the investigated area.

2.4 Data for Event Analysis in Part II

For the second part of this study, a selection of high intensity events was needed. Only few high intensity events occurred in the common data period from August 2015 to January 2016. It was therefore decided to look beyond this period and select historical extreme events and recent high intensity events that resulted in flooding in the Copenhagen area. Based on data availability, the following four events were selected: 1) 2nd of July 2011, 2) 30th-31st of August 2014, 3) 4th of September 2015 and 4) 15-16th of June 2016.

Data was collected for the four events, including data for about 9 hours up to and after the events. The data used for the event analysis were generally similar to the data used in Part I, but with the following differences:

Radar: The model did not exist in 2011, however a re-run of the event was done by DMI using the standard setup. This is considered an important historical event and often used as example event, however a re-run of the event in 2014 was not conducted, and thus no Radar data was obtained for Event 2.

NWP: The model did not exist in 2011 and the data is therefore a re-run of the event with a slightly different setup than for the OMOVAST project (Vedel, 2016): 1) The boundaries are from a different boundary model used in 2011. 2) As it is a re-run the model was run with the most recent setup, where the radar data assimilation has been updated to include time interpolated radar data (see Nielsen et al. (2014) for description of concept). This is not included in the version providing data for the OMOVAST project. The NWP data for this event might thus be worsened by one

change and improved by another and changes may to some extent counteract each other. As it is not possible to obtain the exact same setup for the 2011 event these differences were considered sufficiently small to be ignored.

Ensemble: The EPS did exists in 2011 but with a prediction horizon of 47 hours compared to the current 53 hours. The quality of the data is however assumed to be similar for all four events. It was decided to extend the studied grid of the Ensemble data to cover all of Zealand for all four events based on the following reasoning: Firstly, it was considered interesting to expand the area included in the event analysis to improve impression on movement and development of the events and ensure proper coverage of the predicted extremes. Secondly, as the Ensemble is the product with the lowest resolution the increase in extent was considered most relevant for this product. The original and the extended grid coverage can be seen in Figure 2.4.

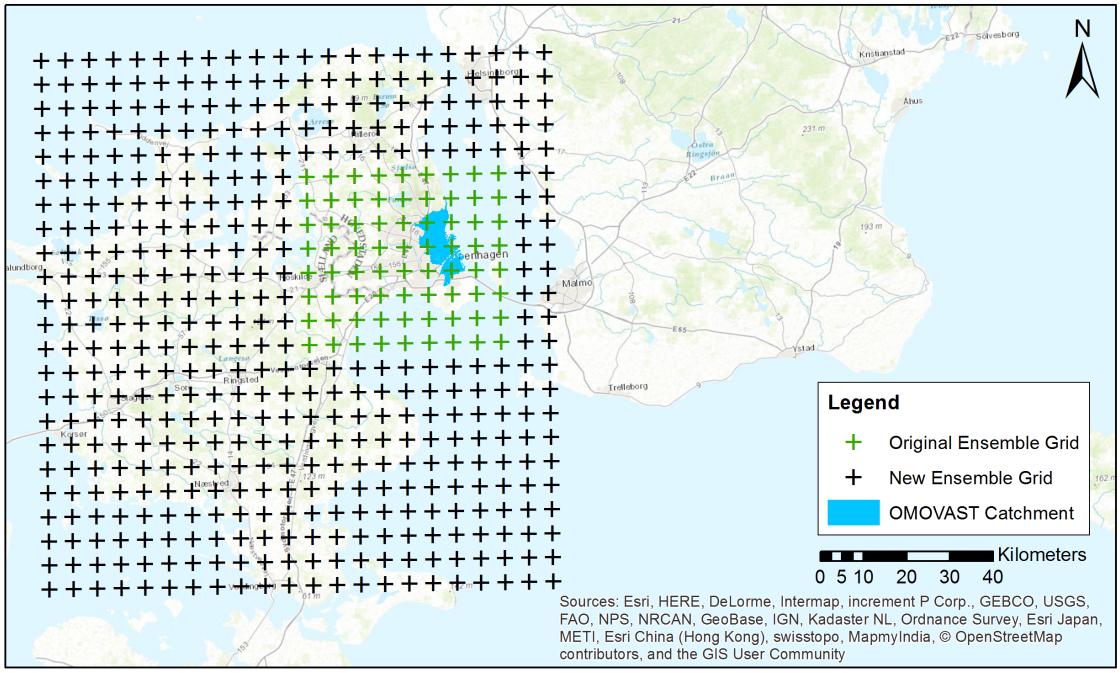


Figure 2.4: *Map of Zealand, Denmark with the original and the extended Ensemble grid centerpoints and the catchment used in SURFF.*

For Ensemble, the maximum value at each time step of the 25 EMs was used in several cases instead of the individual EMs. This can be seen as similar to the most conservative approach using a so called "weight of evidence" of 1 EM, and counting a hit when only one EM predicts rain. This approach to an EPS is described in detail in section 3.3.3. Being conservative increases the hit rate but also the false alarm rate, which has to be taken into account before applying such a solution in warning systems. This is however the most relevant approach when attempting to simulate flooding, as this results in the most extreme prediction and thus the most flooding. However, it should be noted that the procedure generates a fictive time series with more rain than any of the individual EMs predict, which must be taken into account when analysing the results. Alternatives to this approach would be to select the EM predicting the highest intensity or the highest amount of rain. However with these approaches one excludes more correct predictions of shape or timing by the other forecasts. Thus it was considered most relevant to include as much information as possible by considering the maximum at each time step, even though it might increase the risk of overestimating.

2.5 Radar Correction Factors for Part II

The Radar data was in Part I of this study found to underestimate the rain (section 4.4). As the rainfall amounts are important for flood simulation results, it was decided to identify correction factors on event basis. Thus an appropriate correction factor was found for the Radar data for the individual events, and applied on the forecasts prior to flood simulation. A simple procedure was used to identify suitable factors:

1. The areal average was computed for the first LTS of each Radar forecast and for the RG data spatially interpolated to the Radar grid.
2. The sums of the areal averaged Radar and RG observations were computed for each event.
3. The fraction of the two sums was found for each event.

The Radar data is thus adjusted to the RG measurements. The predicted amount of rain for the three events and the estimated correction factors can be seen in Table 2.2. Figures of the areal average rainfall profiles for the three events with RG and Radar data can be seen in Appendix F.2.

Table 2.2: *Sum of the areal averaged observed rainfall according to the spatially interpolated RG and the Radar observations, as well as the resulting correction factors, for the tree events with Radar data available.*

	Event 1	Event 3	Event 4
Sum RG $[mm]$	96.33	13.54	39.32
Sum Radar $[mm]$	66.39	7.05	21.87
Correction Factor	1.45	1.92	1.80

CHAPTER 3

Methods for Part I - Long Term Analysis

3.1 Approach of Part I

Forecast Verification is an research area with many different tools for assessing the performance of forecast products (Jolliffe and Stephenson, 2012). In this part of the study a general assessment is done based on approximately six months of data, with focus on aspects relevant for flood predictions. The following approach was therefore applied in the long term forecast analysis:

1. A visual investigation was first conducted, where the forecast products were analysed based on visualization the data to identify patterns. Three common visualization approaches were used: 1) Scatter plots, 2) Accumulation plots and 3) Quantile-Quantile plots.
2. A performance investigation was then conducted using different threshold analyses to investigate performance at different levels of rain and with different lead times. The threshold results were then combined with other performance indicators to get a thorough impression of the performance of the forecasts over lead time.
3. A resolution investigation was then conducted, investigating the effects of applying different spatial and temporal perspectives on the data, utilizing the resolutions of the forecast products in ways relevant for warning systems.

The long term precipitation forecast analysis is separated in sections according to these three distinct investigations.

3.2 Visual Investigation Methods

The visual investigation is the primary analysis of the datasets before calculating various indicators from it. They are referred to as visual investigations as they are based on graphical representations of the data instead of statistical measures. Three different methods are applied:

1. **Scatter plots**, relating a forecast dataset to the RG measurements by showing all data as points located based on their values in the two datasets. A scatter plot is a concurrent comparison of the data, relating a forecast dataset to the RG measurements by plotting all data as points located by their values in the two datasets. Thus variation of the scatter of points from the 1:1 line indicates differences in the two datasets.

2. **Accumulation plots**, showing the accumulated values over the common period for all datasets. This gives an indication on whether one forecast product might have a general tendency to over- or underestimate rain.

3. **Quantile-Quantile plots** (QQplots), relating a selected number of quantiles of one forecast product to the quantiles of the reference data. Quantiles sort data points of a sample into intervals of equal probability, and a quantile thus indicates how often in percent of time the precipitation exceeds a certain value. This means that QQplots compare the distribution of datasets and not the individual data points, unlike the accumulation and scatter plots, which are direct comparisons of the data. This allows datasets of different sizes and even periods to be compared, and thus the results are not directly affected by the amount of NA values, as these are left out. Variations from the 1:1 line indicate differences between the two distributions and whether the forecasted distributions over- or underestimate the measured distribution.

3.3 Performance Investigation Methods

After a visual investigation of the data different verification procedures were used to investigate the quality of the forecast products. These methods are described in the following sections. From the different statistical analyses, an impression of a forecast products performance is obtained and the products can be compared. When analysing the forecasts, two different approaches are relevant: 1) evaluating the average skill of the forecast products across forecast horizons, and 2) distinguishing between forecast skill for different horizons, thus evaluating the development of performance over selected horizons.

3.3.1 Binary Threshold Analysis

The binary threshold test is commonly used to investigate the performance of simple binary events (e.g. rain / no rain). It is a concurrent comparison method, comparing pairs of observed and modelled data points. Keeping the hydrological perspective of this study in mind, interest lies on larger rain events that may fill the drainage system. Thus, the method is applied for two different thresholds: 1 mm/h and 3 mm/h. These thresholds taking into account the short time span of the available data and thus low presence of large rain events. A 2x2 contingency table, as the one showed in Table 3.1, with the number of hits (H), misses (M), false alarms (FA) and correct rejections (CR) was set up for each threshold, each RG and each LTS.

The sum $H + FA + M + CR = n$ is the sample size, thus by dividing the counts with n relative frequencies can be obtained. The probability of an event occurring is defined as $s = (H + M)/n$ which is also called the base rate (Jolliffe and Stephenson, 2012).

Table 3.1: *Categorical Contingency Table for deterministic forecasts with the number of occurrences in each category represented as Hits (H), False Alarms (FA), Misses (M) and Correct Rejections (CR).*

	Event Observed		
Event Forecasted	Yes	No	Total
Yes	Hits (H)	False Alarms (FA)	H+FA
No	Misses (M)	Correct Rejections (CR)	M+CR
Total	H+M	FA+CR	n

Based on this table some descriptive statistics can be calculated. Relevant performance measures and their definition can be seen in Table 3.2. The two main measures commonly used are the Hit Rate and False Alarm Rate. Besides these, a bias measure and a skill score are relevant for this study. Thus, the included measures are:

- Hit Rate (HR), also know as the probability of detection (POD) (Jolliffe and Stephenson, 2012).
- False Alarm Rate (FAR), also known as the probability of false detection (POFD) (Jolliffe and Stephenson, 2012).
- Frequency Bias Index (FBI), also known as just the bias, meaning the relation between predicted event and observed events. This is an important measure from a hydrological perspective as good estimates of rainfall amount is important in obtaining accurate runoff predictions.
- Peirce Skill Score (PSS) was chosen as the skill score measure for this project. It is equivalent to the subtracting the FAR from the HR. This skill score was one of the first defined in the forecast verification field and is also applicable to multi-category analyses (Jolliffe and Stephenson, 2012), which will be described in Section 3.3.4.

Table 3.2: *Summary table of binary verification measures based on Jolliffe and Stephenson (2012).*

Name of measure	Definition	Range
Hit Rate, HR	$HR = \frac{H}{H+M}$	[0,1]
False Alarm Rate, FAR	$FAR = \frac{FA}{FA+CR}$	[0,1]
Frequency Bias Index, FBI	$FBI = \frac{H+FA}{H+M}$	$[0, \infty[$
Peirce Skill Score, PSS	$PSS = \frac{H \times CR - FA \times M}{(FA+CR)(H+M)} = HR - FAR$	[-1,1]

Other relevant scores were also tested and are described in Appendix D.1. As they provide similar results as FAR and PSS they were only used as support and the results can be seen in Appendix D.7.

3.3.2 Visualization of Binary Results

To investigate the performance of the forecasts, a **Relative Operating Characteristic** (ROC) diagram (FAR vs HR) can be plotted. This diagram illustrates how well a forecast differentiates between the occurrence and not occurrence of an event. The perfect forecast is thus located in the top left corner at (0,1). If located at (0,0) the forecast never forecasts an event, while if located at (1,1) the forecast constantly forecasts events wrongly. Thus points above the 1:1 line performs better than a random forecasts with no skill, and a higher and shaper bend on the curve implies a better forecast product.

Another common visualisation method is a **Skill-Bias diagram** (e.g. FBI vs PSS) (Jolliffe and Stephenson, 2012). Both diagrams was used in this study to visualize the results. It should be noted that if forecast systems with different base rates are plotted together in a Skill-Bias or ROC diagram one should take care not to compare their location on the diagram as the bias and skill are not uniquely defined (Jolliffe and Stephenson, 2012). Thus for forecasts with different base rates, e.g for different thresholds, no inter-comparison should be conducted.

3.3.3 Weight of Evidence Approach for Ensemble Data

Using binary statistics for an EPS enables a new dimension of investigation. An important factor of forecasting is to obtain a good balance between hits and false alarms, which depends on the purpose of the forecast product. A larger number of EMs predicting an event is a stronger evidence that the event will occur, and a more conservative, and less reliable, prediction is made when using a lower number of predicting EMs as evidence for an event. Thus for Ensembles, one can extend the binary analysis by introducing a so called "weight of evidence" as the decision threshold, meaning the fraction of EMs which must predict an event for it to be counted as a hit and trigger a decision. In this case a hit is counted when the chosen number of members predict above the threshold, and a 2x2 contingency table can be created for each weight of evidence (corresponding to the number of members in the EPS).

Similar to the ordinary binary test, a ROC diagram can be made based on the weight of evidence approach ranging from a high weight of evidence (all EMs should predict the event) as the point furthest to the left and a low weight of evidence (only 1 EM is required to predict the event) as the point furthest to the right. By letting the decision threshold vary trough each possible weight of evidence, the ROC curve can be extended compared to what is obtained when using the EMs individually. This method enables the possibility of choosing a weight of evidence that provides an appropriate balance between hit rate and false alarm rate for the purpose of the forecast.

3.3.4 Multi-Category Analysis

The previously described binary analysis (section 3.3.1) can be extended to a multiple category analysis using several thresholds. Five thresholds were chosen for this analysis: 0.1 mm/h, 0.5 mm/h, 1 mm/h, 3 mm/h and 5 mm/h. As we are mainly interested in the performance of the forecasts during rain events, the thresholds were used to define five categories in mm/h: [0.1,0.5[, [0.5,1.0[, [0.1,3.0[, [3.0,5.0[and [5.0, ∞[, thereby excluding all the data point with less than 0.1 mm/h. Thus a 5x5 multi-category contingency table was created for each RG for each LTS. Measures relevant for the multi-category analysis can be seen in Table 3.3. K is the number of of categories, p_i is the distribution of the observations in the categories, $\hat{p}_i$ is the distribution of the forecasts in the categories, p_{ii} is thus the probability of a hit for each threshold.

Table 3.3: *Summary of multi-category measures, based on Jolliffe and Stephenson (2012).*

Name of measure	Definition	Range
Proportion Correct, PC	$PC = \sum_{i=1}^{K} p_{ii}$	[0,1]
Frequency Bias Index, FBI	$FBI = \frac{\hat{p_i}}{p_i}, i = 1, ..., K$	$[0, \infty[$
Hit Rate, HR	$HR = \frac{p_i}{p_{ii}}, i = 1, ..., K$	[0,1]
Peirce Skill Score, PSS	$PSS = \frac{\sum_{i=1}^{K} p_{ii} - \sum_{i=1}^{K} p_i \hat{p_i}}{1 - \sum_{i=1}^{K} p_i p_i}$	[-1,1]

As seen in the table, one value can be obtained for the PC and PSS for each RG and each LTS, however for the FBI and HR one value is obtained for each threshold. In this study, the multi-category analysis is used to assess whether the performance seen for the two thresholds tested in the binary threshold tests are also evident for other ranges. Thus it is mainly used as support for the results obtained from the binary test.

3.3.5 Residual Quality Measures

The binary statistics are direct value comparisons, comparing pairs of observed and forecasted data points based on a threshold. However categorical measures do not account for the quantitative discrepancies between the observed and forecasted rainfall. Considering the discrepancies is useful to describe the goodness-of-fit of the data, and relevant when coupling the forecast to flow and flood modelling. Thus other measures are relevant to compare model performance and investigate the behaviour of the entire datasets based on residual errors. The following quantitative accuracy measures were used: Mean Error or Water Balance Error (WBE), Root Mean Square Error (RMSE), the Nash-Sutcliffe Efficiency Coefficient (NSE), and the correlation coefficient (R). An overview of the different statistical indicators can be seen in Table 3.4. These residual indicators can be combined with the binary measures from section 3.3.1 to describe the quality of the forecasts for specific cases.

The NSE was originally developed for runoff modelling (Nash and Sutcliffe, 1970) and care should be taken when using it for other purposes. Because it gives indication on how the forecasts perform compared to the mean of the observations (a NSE higher than 0 show that the forecast performs better than the mean of the observations) it was still included here as a secondary measure supporting the conclusions from the other indicators. Concluding anything from the magnitude of the value obtained is not possible as it is not runoff.

Table 3.4: *Residual Measures to investigate forecast model behaviour. x_i is the observed rain and $\hat{x}_i$ is the forecasted rain at time step i.*

Name of measure	Definition	Range
Water Balance Error, WBE	$WBE = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i - x_i$	$]-\infty;\infty[$
Root Mean Squared Error, RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}$	$]-\infty;\infty[$
Nash-Sutcliffe Efficiency coefficient, NSE	$NSE = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$	$]-\infty;1]$
Correlation Coefficient, R	$\frac{\sum_{i=1}^{n}(\hat{x}_i - \overline{\hat{x}})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(\hat{x}_i - \overline{\hat{x}})^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$	$[-1;1]$

3.4 Resolution Investigation Methods

As they are products with a spatial and temporal resolution, forecasts may predict a certain event correctly but dislocate the event in time or space. When evaluating the quality of a forecast product, one may thus choose to investigate its performance on different spatial and temporal scales to include events forecasted with spatial or temporal displacements, as they are still useful as warnings. It is however difficult to evaluate a gridded forecasts using only traditional verification scores such as the ones described above (section 3.3). A slight offset in position or time of an event will lead to double penalization in these investigations: once for not occurring in the right place or at the right time and once for occurring in the wrong place or time. This is also known as the dual penalty problem. This situation occurs more frequently for forecast products with a higher spatial/temporal resolution, and may thus result in lower scores for higher resolution products than for their lower resolution counterparts.

One way to account for differences in location is the neighbourhood method. This verification method is a filtering method allowing forecasts located in the neighbourhood of the observation to be counted as correct (Jolliffe and Stephenson, 2012). By doing this, one may obtain higher hit rates at the cost of higher false alarm rates. Similarly, a temporal analysis investigating the increase in prediction quality when including more lead time can give an impression of how much of a forecast could be included in a warning system to obtain the best balance between hits and false alarms. It is a trade-off between an increase in hits and false alarms, which should be balanced depending on the purpose of the forecast and warning system.

It is thus relevant to investigate the performance of the forecast products on different temporal and spatial scales and combinations of these. The methods used for the spatial and temporal investigations are described in the following subsections.

3.4.1 Spatial Analysis

For the spatial analysis, a procedure based on Atger (2001) was applied. The method is a neighbourhood method which allows the comparison of gridded forecasts with point observations. The five

thresholds used in the multi-category analysis (section 3.3.4) were kept for this analysis and four different distances from the RGs were selected: 5 km, 10 km, 15 km and 20 km. The minimum distance was chosen to be 5 km, as this is the resolution of the Ensemble grid and thus the minimum distance applicable to all three forecast products. Taking into account the total covered area of approx 900 km^2 for the smallest forecast grid (NWP), the maximum relevant distance was set to 20 km, and distances at a 5 km interval from the selected minimum to maximum was evaluated. The situation has been exemplified for a selected RG in Figure 3.1, showing the area included in each of the four cases. 10 min, 30 min and 1 hour LTSs were used for Radar and NWP, while the original resolution of 1 hour was kept for the Ensemble. The Radar data was interpolated to the NWP grid as described in section 2.3.

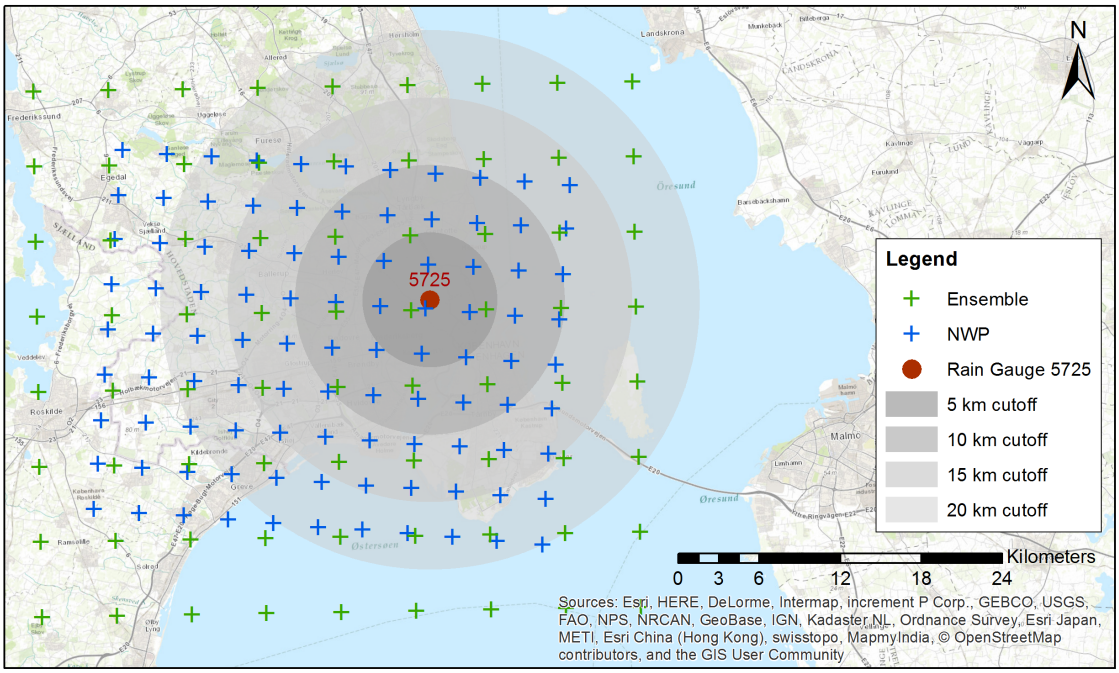


Figure 3.1: *Illustration of the four cutoff distances 5, 10, 15 and 20 km for the RG 5725 with the NWP and Ensemble grid centerpoints. All centerpoints within the grey circles are included in the analysis for the respective distance for this RG and contribute to the hit count.*

A spatial binary contingency table was then created for each RG for each LTS and varying the threshold for the observations and the forecasts as well as the distance. A hit is counted as when at least one forecast grid point within the selected distance is above the selected threshold when the RG is above the threshold. The contingency tables thus reflects for a given observed or not observed event (above/below the respective threshold) the number of times a similar event was predicted or not a least once within the different distances from the observed event. For Ensemble this was conducted for each EM.

It should be noted, that for many RGs increasing the distance on this scale results in the included area reaching beyond the grids, which is also evident for the RG on the figure. This of course does not result in the same changes as would be obtained if the grid covered the entire included area. Ensemble has a slightly larger grid and might thus be less affected by this.

3.4.2 Temporal Analysis

For each forecast the temporal precision is investigated by including more and more of the forecast lead time in the analysis, as the prediction of the rain event might have a temporal offset. Thus a similar analysis as the one described above is done using a spatial cutoff distance of 5 km, and with the same five thresholds as before. Instead of varying the distance, the data is analysed over the lead time, including more and more. The following temporal cutoffs were chosen independent of the aggregation level: the first LTS and after each hour of lead time. As the spatial analysis it was conducted using the 10 min, 30 min and 1 hour data for Radar and NWP, and using 1 hour data for Ensemble. Again the Radar data was spatially interpolated to the NWP grid. A conceptual drawing of the variation in included lead time can be seen in Figure 3.2 using Radar as example. A similar procedure was used for NWP and Ensemble but with more steps due to their longer forecast horizons.

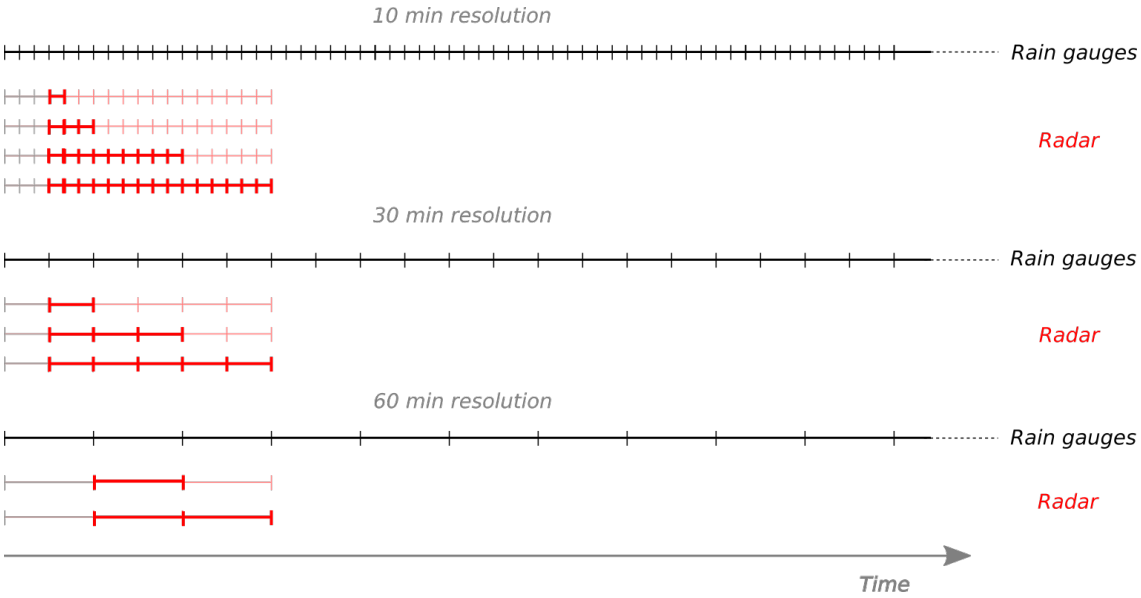


Figure 3.2: *Conceptual drawing of the temporal analysis exemplified for Radar with the used temporal cutoffs. Each red horizontal line symbolizes an analysis of the same forecast but with varying lead time inclusion.*

One major drawback of this method lies in the way it compares to the RG data. The way it is implemented here, the forecast starting at the time of the event is used to investigate the quality of the forecast. Thus this analysis introduces a way of accounting for delay in forecasts. However if a forecast product has a tendency to predict events in advance, it is not rewarded in this analysis. This will be investigated further for specific events in the second part of the project, and was thus not considered here. The current analysis still provides an overview of the effects of increasing the considered lead time.

CHAPTER 4

Results of Part I - Long Term Analysis

4.1 Visual Investigation

To achieve an initial understanding of the data, three different kinds of visual investigations were made: Scatter plots, Accumulation plots and QQplots. Details on data used in the presented results, such as temporal aggregation levels, are described for the individual analyses in the following sections.

A simple investigation of the data reveals that according to the RGs it is raining approx. 7 % of the time (for an aggregation level of 10 min). According to Radar, NWP and Ensemble it is raining respectively approx 70 % of the time, approx. 45 % of the time and approx. 50 % of the time (aggregation level 10 min for Radar and NWP and 1 hour for Ensemble). The big difference between RGs and forecast products is related to the tipping bucket method, meaning that very small drizzles will not make the bucket tip and thus these events do not contribute to the total time with rain. Also, the extra high value for Radar can be explained by remaining noise after the data processing prior to use. If a limit of $0.001 \mu m/s$ is used instead of 0 the following percentages are achieved: Radar: 15 %, NWP: 38 %, Ensemble: 40 % and RGs 7 % for the same aggregation levels as before. Thus some noise in the Radar data is removed and a much lower value is achieved, however the Ensemble and NWP still seem to overestimate the time with rain. RGs get the same value, which supports the expected results of a tipping bucket gauge.

4.1.1 Scatter Plots

Scatter plots were made for all three forecast products, to investigate the correlation between forecast and RG data and its change over the forecast horizon. Hourly LTSs were plotted for the three precipitation products with data for all RGs, and can be seen in Figure 4.1, Figure 4.2 and 4.3. For Ensemble, the mean of the EMs is shown. Similar plots were made using the all EMs and the maximum of the EMs and can be seen in Appendix C. The correlation coefficient R^2 was included in the plots to help evaluate the overall performance, as the scatter is difficult to interpret for low intensities.

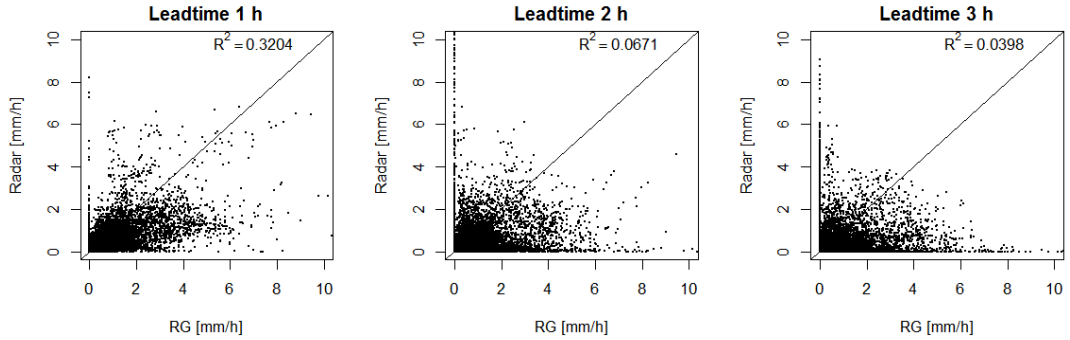


Figure 4.1: *Scatter plots of Radar vs RG for each hour of lead time for the 1 hour temporal resolution with all RG. The correlation coefficient R^2 is shown in the top right corners of the plots.*

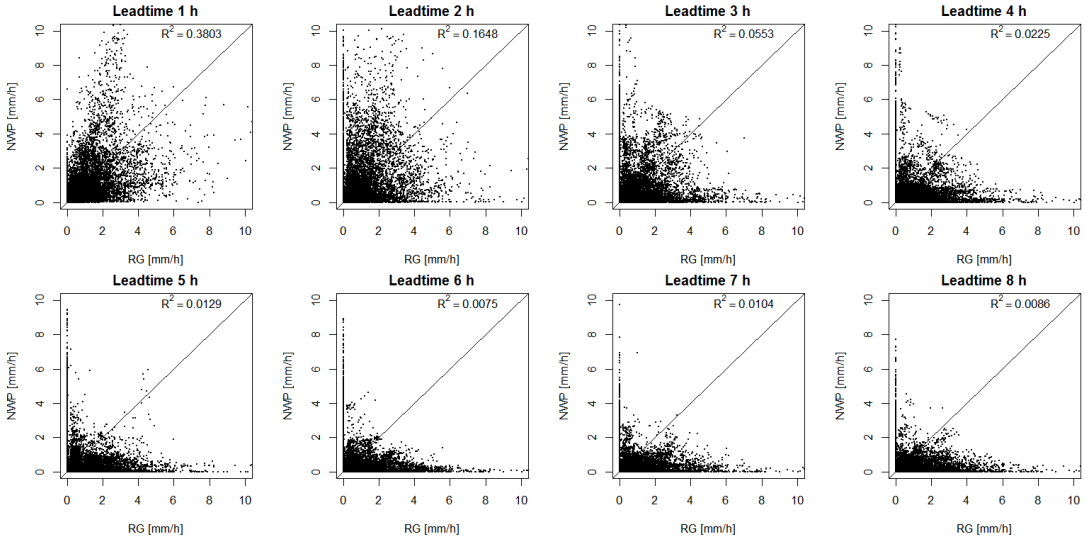


Figure 4.2: *Scatter plots of NWP vs RG for each hour of lead time for the 1 hour temporal resolution with all RG. The correlation coefficient R^2 is shown in the top right corners of the plots.*

From all three figures it is evident that the short period of data affects the shape of the scatter. Neither of the plots show a high correlation between the two datasets, however for the Radar and NWP data, the correlation decreases over increasing lead time. This tendency is less clear for Ensemble, possibly caused by a longer spin-up time. For all three products both complete misses and incorrect prediction of events can be seen from the dots along the two axes. Systematic divergence from the 1:1 line suggests unmodelled behaviour, however no clear systematic trends are evident from the plots. Especially for higher LTSs the majority of the scatter seems to be below the 1:1 line for all three products suggesting an increasing tendency to underestimate further out in the future. This is however not evident from the max of the EMs (Apppendix C). For NWP the two first lead time hours are clearly different from the rest, reflecting the data assimilation and nudging of the model. For Ensemble, the mean, max and all EMs plots generally show similar trends (Appendix C). Due to the many points in the plots it is not possible to visually identify trends for the low intensity rain events. They are however still reflected in the correlation coefficient, which was considered sufficient as these events are not within the area of interest of this study.

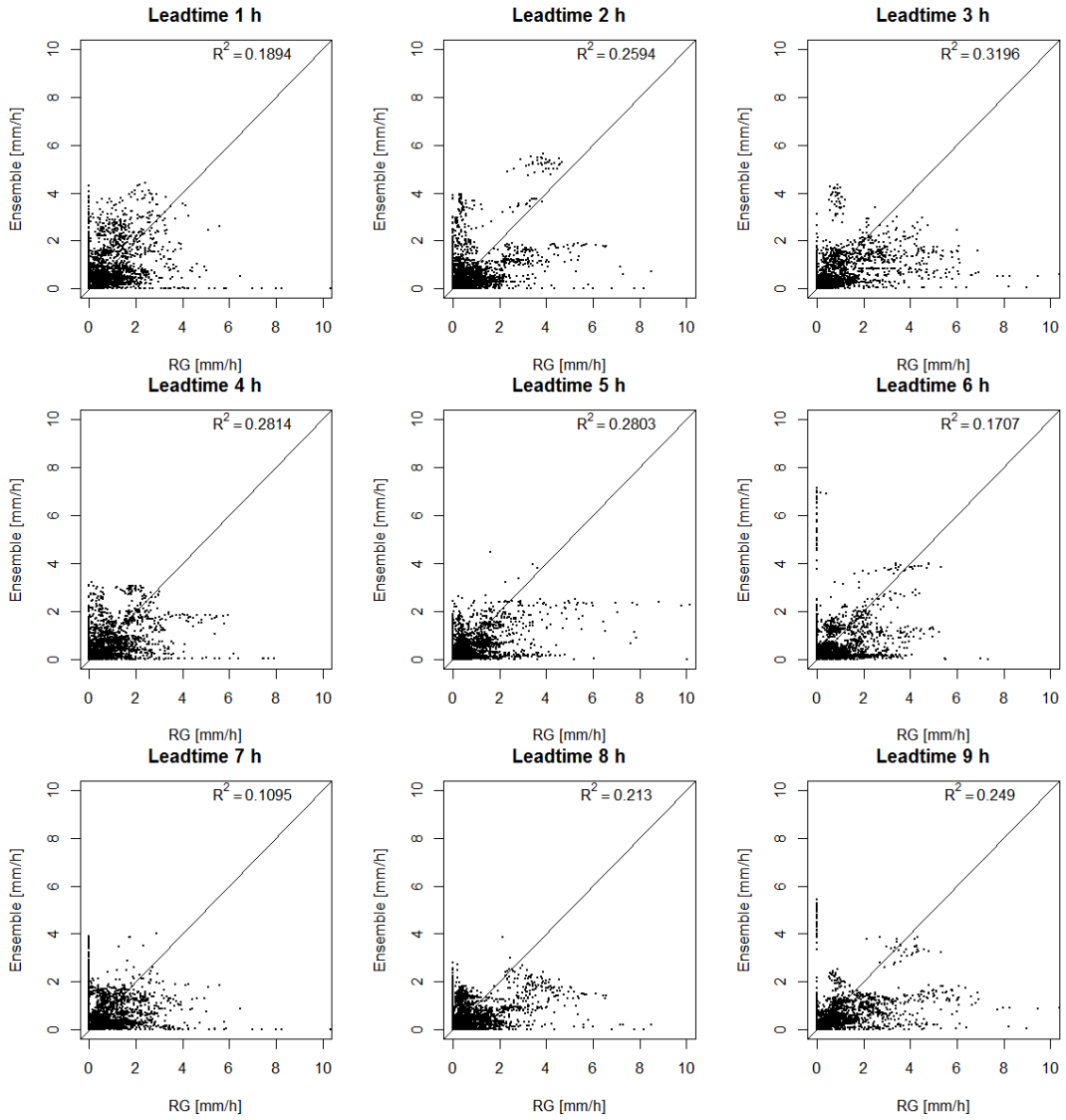


Figure 4.3: *Scatter plots of Ensemble vs RG for each hour of lead time with all RG and the mean of EMs.*

For NWP, scatter plots using data from the entire NWP period were also made and can be seen in Figure C.1 in Appendix C. Compared to the scatter plots obtained using only the common data period, a more spread-out scatter is seen for all lead times reflecting the increased amount of data. However, the overall trends seen here are also observable for the longer period of data, and thus no undetected behaviour was revealed by including more data.

For Ensemble an investigation of the individual EMs performance was done, by comparing the first 6 hours of lead time of each forecast (and thus creating a continuous time series) against hourly aggregated values for the RGs for each EM. The results can be seen in Figure C.4 in Appendix C. From this it was seen that even though there are differences in the performance of the EMs, all members show similar trends, as the same shape of the scatter is seen in most of the plots. Thus no member reflect a tendency to have an overall higher or lower correlation than the other members.

4.1.2 Accumulation Plots

The accumulated rain was calculated for the three forecasts and the RG data for the common period using 10 min temporal resolution for RG, Radar and NWP and the 1 hour resolution for Ensemble. The continuous time series setup was used in this case, and thus the accumulation plots show the behaviour of the part of the forecasts covering the time between forecasts.

The results can be seen in Figure 4.4 for the five Copenhagen stations and the mean of the RGs. For Ensemble all 25 EMs are plotted. From the figure it is seen that Radar underestimates the rain amount, NWP overestimates, and most of the EMs seem to estimate the rain amount quite well. It should be noted that due to missing values, which especially for the Radar data represent a high percent of the total data (28 %), the accumulation can be misleading as some rain is missing. This contributes to the underestimation seen for Radar, however it is assessed that an underestimation would still be evident with all data available, as underestimation is seen early in the period while most of the NA values are located later in the period. This situation could be avoided by only including periods with data available for all three forecast products, however this would mean excluding useful data for both NWP and Ensemble from the analysis.

The overestimation of NWP is not evident if the first lead time hour is used instead of the third, as seen in Figure C.5 in Appendix C. This shows that the overestimation is caused by the data assimilation and nudging performed in the start-up period of the model. This is however counter-intuitive, as data assimilation should move the model towards the data, but might be explained by the overestimation of drizzle seen in the QQplots (section 4.1.3).

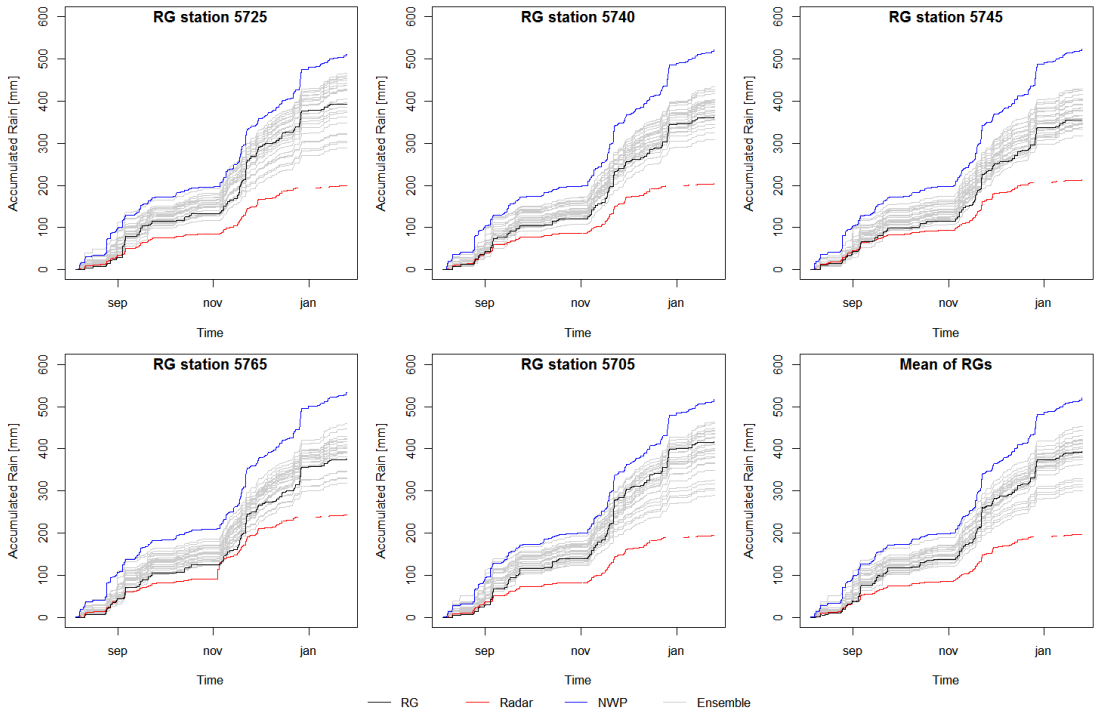


Figure 4.4: *Accumulated rain for the Radar period for the three forecasts and the RG data for the five Copenhagen stations and the mean of the RGs.*

4.1.3 Quantile-Quantile plots

1000 quantiles (steps of 0.1 %) were calculated for all data types using the spatially interpolated forecast data. The high number of quantiles was chosen to ensure good representation of the higher, relevant values which only correspond to a small percentage of the datasets. QQplots were created for the five Copenhagen stations and for the mean of the RGs, and can be seen in Figure 4.5. Keeping in mind that the aggregation level affects the magnitude of the values and that the period is shorter than a full year and thus reflects the included seasons, the quantiles were computed for the common period, using the 10 min temporal resolution for RG, when comparing to Radar and NWP (10 min aggregation levels), but the 1h temporal resolution for RG when comparing to Ensemble. The continuous time series approach was used, thus including the lead time between forecasts in the analysis.

From the figure it is seen that the NWP quantiles follow the RG quantiles best, however overestimating small and large rains, but capturing the middle rains quite well. The EMs are mainly located below the 1:1 line, showing a tendency to underestimate. It is seen that both NWP and Ensemble have a tendency to overestimate smaller rain events, and as these are much more frequent than the larger rains which they estimate better, over time the many small overestimations can explain why most EMs and NWP seem to overestimate in the accumulation plots in section 4.1.2, but not in the QQplots. The Radar data clearly underestimates the precipitation and is even lower than the lowest of the EMs. This supports the conclusion from the accumulation plots (section 4.1.2), that the underestimation of the Radar is not only caused by the large quantity of NA values, but rather by a general tendency to underestimate rain. For all three forecasts it is seen that they predict rain when none is measured, this is again to be expected when using tipping bucket gauges.

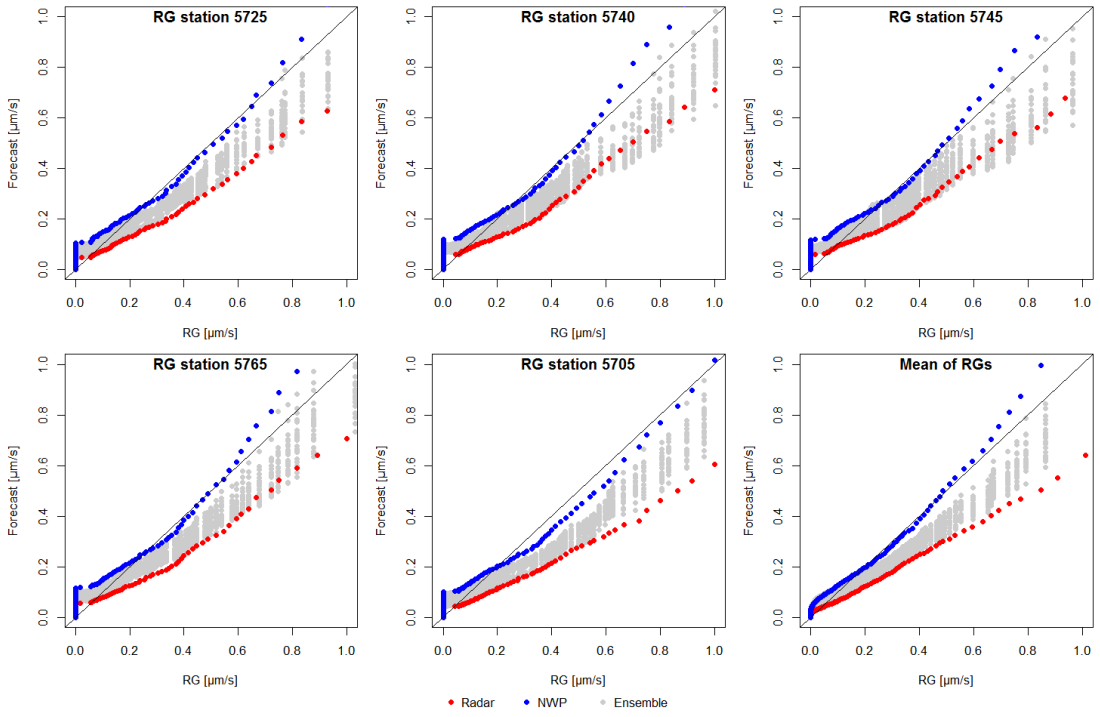


Figure 4.5: *QQplots for the five Copenhagen stations and for the mean of all RGs, with Radar, NWP and Ensemble quantiles.*

4.2 Performance Investigation

From the visual investigations a first impression on the performance and trends of the forecast products related to the RG was obtained. Most importantly it was observed that the Radar data underestimates, while the NWP has a small tendency to overestimate and the Ensembles on average seem to capture the rain amounts quite well. To thoroughly investigate the performance of the forecast products in different situations and particularly over lead time, further analysis is needed.

The performance investigation analyses this further using different statistical measures calculated for the forecasts for different temporal resolutions and different LTSs. All tests were performed on the forecasts spatially interpolated to the RGs for the common data period.

4.2.1 Binary Threshold Tests

Binary threshold analyses were made using thresholds of 1 mm/h and 3 mm/h. As the values get more smoothed for higher aggregation levels, a lower number of hits will be seen for the 1h temporal resolution. For this resolution and a threshold of 1 mm/h, the RG data is above the threshold at least 108 times out of 4153 observations (approx. 2.6 %, for the RG with the minimum value). However with a threshold of 3 mm/h the minimum number of times a RG is above the threshold is only 17. For Radar, NWP and Ensemble the minimum number of events above the thresholds were respectively 124, 276 and 76 for 1 mm/h and 23, 11 and 4 for 3 mm/h. Considering 3 mm/h was thus clearly pushing the limit of what the threshold could be and still provide usable results.

For Radar and NWP the analysis was done for the 10 min, 30 min and 1 h aggregated data. To illustrate the performance of the forecasts over lead time Skill-Bias plots were made showing the five Copenhagen RGs and the mean of the statistics for all RGs, and can be seen in Appendix D.2. The results for the mean of the RGs for the three temporal resolutions and for the two thresholds can be seen in Figure 4.6 and 4.7, and show a tendency of lower skill and bias for the lower temporal resolutions. For NWP 10 min the initialization step was included in the 10 min plots. This data has a low skill and bias compared to the rest of the LTSs, supporting the conclusion that it reflects a shorter period and is an initialization step (section 2.1.3).

For the Radar data, a tendency of fast decreasing skill over lead time is seen in Figure 4.6 for the 1 mm/h threshold, however the bias remains stable to some extent just above 0.5, which indicates that the Radar in general underestimates the rain as was also seen from the visual investigations (Section 4.1). It is clear from this analysis and from the previous section 4.1 that the Radar data should be corrected for underestimating. It is common that radar data underestimates the rain and they are thus often multiplied with a correction factor (Thorndahl et al., 2014). It should be noted that conducting a threshold analysis on the data before correction does provide poorer results. However, the current threshold analysis helps investigate the needed magnitude of the correction. As a bias of about 0.55 is seen, this would indicate a correction factor of 1.9 could be applied on the dataset. However this is only based on one threshold and might thus not be representable for the entire dataset.

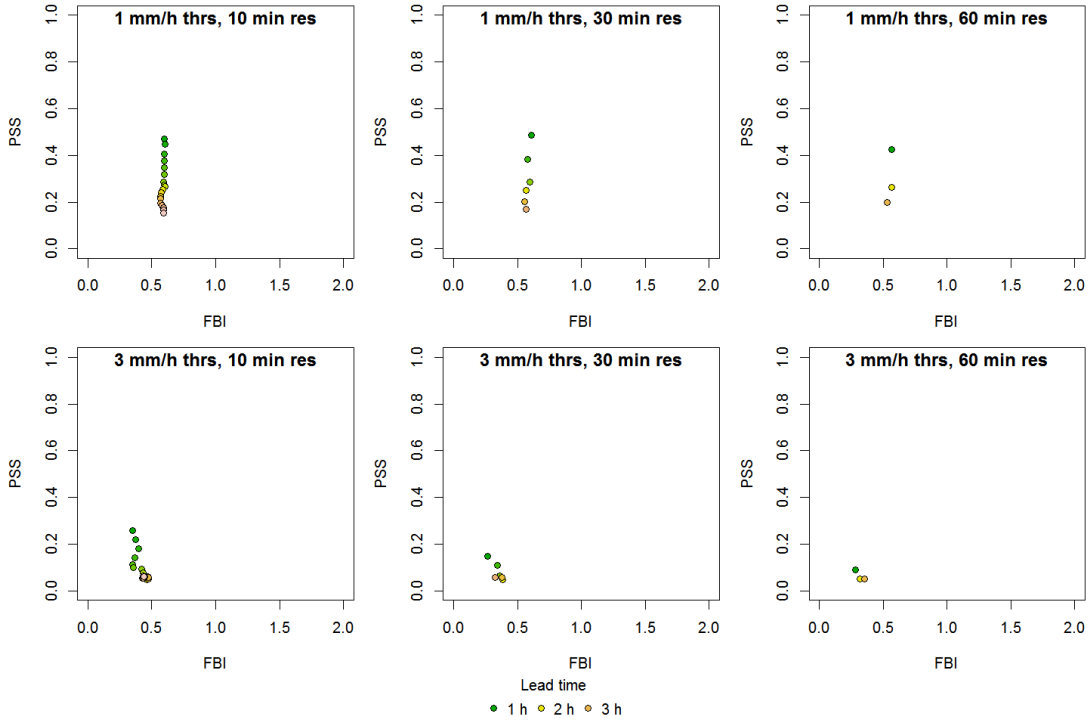


Figure 4.6: *Skill-Bias plots for Radar for the three temporal resolutions 10 min, 30 min and 1 hour and the two thresholds 1 mm/h (top) and 3 mm/h (bottom).*

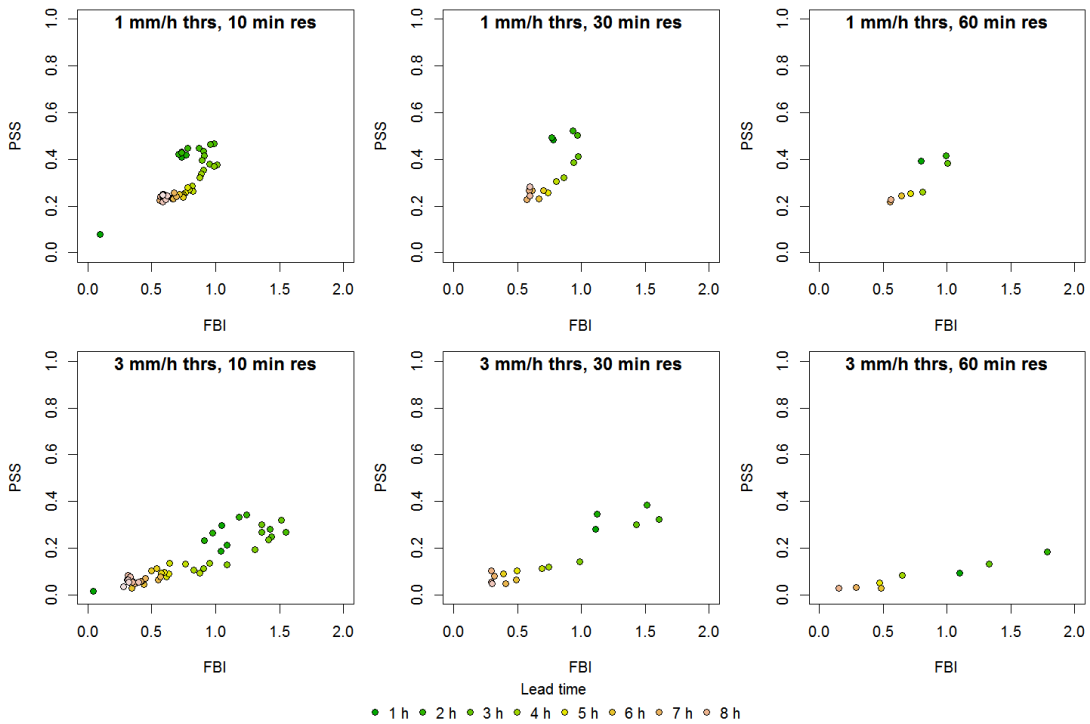


Figure 4.7: *Skill-Bias plots for NWP for the three temporal resolutions 10 min, 30 min and 1 hour and the two thresholds 1 mm/h (top) and 3 mm/h (bottom).*

From Figure 4.7 it is seen that the skill and bias of the NWP data for the threshold of 1 mm/h increase in the first two hours of lead time. This is caused by the data assimilation and nudging processes described in section 2.1.3, however apart from this, a tendency of decreasing skill and bias with lead time is seen. The bias is close to one in the data assimilation period, but it is seen that the assimilation is not enough to push the model towards generating correct water amounts throughout the forecast. From the results of the accumulation plots (4.4) one might expect to see a bias larger than one for some LTSs. However as seen in the QQplots (4.5), the overestimation by NWP is mainly caused by the very small rains, which are all below the thresholds used in this analysis and therefore not included in the calculations.

For the 3 mm/h threshold for NWP a similar trend is seen as for the 1 mm/h threshold but with generally lower skill scores, down to 0 for the furthest predictions. Besides this, the points are more spread as the bias ranges from close to 0 to almost 2, thus indicating that the forecasts actually overpredicts for the first couple of lead time hours at this threshold. For Radar the points are not on a vertical line for the 3 mm/h threshold, however they still exhibit a trend of decreasing skill with lead time and a fairly constant bias. Both skill and bias are again lower for the higher threshold. When comparing the results of the two thresholds in general, it is seen that the forecasts perform better for the low threshold, which was expected not only because more data is included but also because of what was seen in the visual investigations (4.1). This proves that the choice of threshold affects the results greatly, and thus this is investigated further in section 4.2.3.

For the NWP the procedure was also conducted for the full NWP period using the 10 min data, and the equivalent plots can be seen in Appendix D.2, Figure D.1 and D.2. No clear difference was seen from including more data, supporting the argument that the trends observed in the shorter period are representative for the NWP data. However it is also seen that increasing the included period leads to more consistent performance.

For Ensemble, the mean of the statistics for all EM for the first 9 hours lead time is plotted in Figure D.15 and D.16 in Appendix D.2 for both thresholds. As no clear pattern of decreasing quality with lead time was seen in these plots, it was tested if this could be improved by using the mean of the EM to calculate the statistics, rather than taking the mean of the individual statistics for the individual EM (Appendix D.3), however no clear pattern of the measures obtained for the different LTSs was seen, and thus the ordinary averaging method was kept. The results for the mean of the RGs and EMs can be seen in Figure 4.8. For the 1 mm/h threshold the best performing hour is the third one, supporting the argument that the models need spin-up time. For the 3 mm/h threshold higher bias and higher skill is seen for some of the points, however the pattern in concerned LTS seems random, and thus it might be due to the shortness of the period.

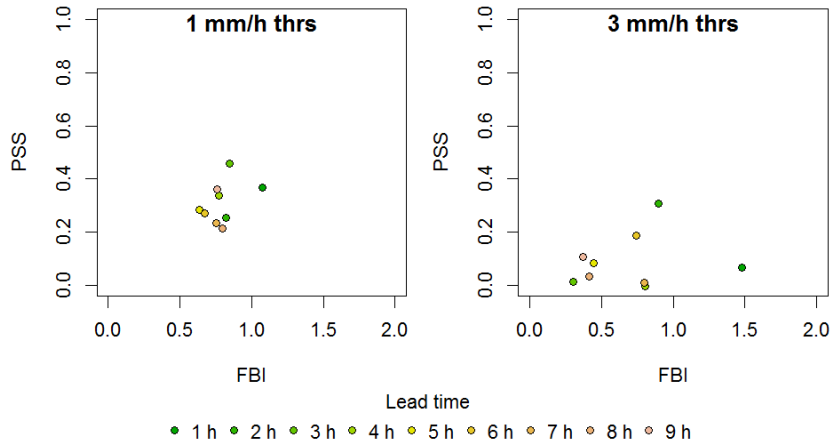


Figure 4.8: *Skill-Bias plots for Ensemble 1 hour resolution for the two thresholds 1 mm/h (left) and 3 mm/h (right), for the mean of the RGs and the EMs.*

Finally, when comparing the three forecast products it was noted that some EMs have higher skill than Radar and NWP for some LTSs for both thresholds, however most of the LTS are located in the same area for the three products. In general all three products perform worse for the high threshold and both NWP and Ensemble show a much more varying bias for the high threshold indicating more randomness in the results. These comparisons are seen more clearly in Figure D.18 in appendix D.4.

4.2.2 Weight of Evidence Approach for Ensemble

Besides the ordinary binary tests, the weight of evidence approach was applied for the first 9 LTSs of the Ensemble and with the 1 mm/h threshold. The results can be seen in Figure 4.9. The mean of the results obtained for the first 9 hours lead time can be seen in Appendix D.5, showing that on average over the first 9 lead time hours, a decrease in hit rate and false alarm rate for increasing EM inclusion is achieved. The method was not applied for the 3 mm/h threshold as the previous analysis showed poor results for this, indicating too little data might be above the threshold to provide proper results.

From Figure 4.9 it is seen that depending on the weight of evidence a higher hit rate can be obtained for the all LTSs, at the cost of a higher false alarm rate. Based on these results it is evident that there is room for improvement of the forecasts if one uses the weight of evidence approach with a low weight of evidence. This would provide better forecast quality over lead time, and thus improve the grounds for using this product in a warning system. It it also noticeable that the performance varies over lead time, and that the 9th lead time hour performs better than many of the other LTSs.

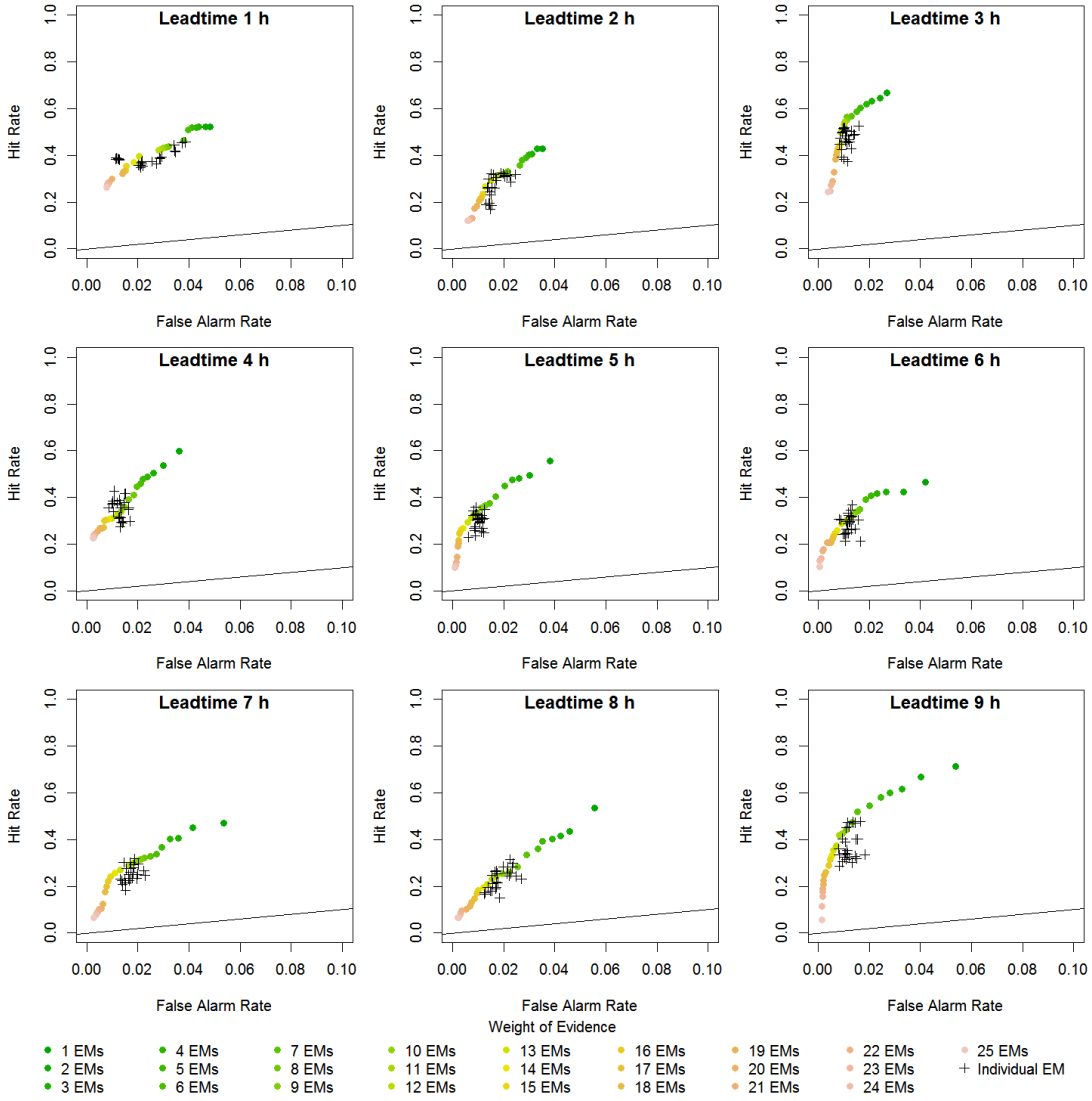


Figure 4.9: *ROC curves for Ensemble using the weight of evidence approach and the results obtained for the individual EMs, for the first 9 LTSs, and using the average of the statistics obtained for the RG.*

4.2.3 Multi-Threshold Tests

Multi-threshold contingency tables were made for the three different forecasts products, using the following limits in mm/h: [0.1,0.5[, [0.5,1.0[, [0.1,3.0[, [3.0,5.0[and [5.0, ∞[. As seen from the thresholds, everything less than 0.1 mm/h was left out of the analysis, as the focus of this study is on effective rain events, and thereby avoiding interference from events with little or no rain.

Contingency tables were computed for all LTSs (using the 10 min data for Radar and NWP) and for all RGs. To illustrate the results, the average of the contingency tables for the relevant lead time and all RGs was computed and the resulting contingency tables with biases and hit rates can be seen in Table 4.1 and 4.2. For Ensemble, Table 4.3 represents the mean table for the first 9 hours of lead time, for all RG and all EMs. Similar tables were also generated for the 30 min and 1 hour aggregation level for Radar and NWP and the averages of the tables can be seen in Appendix D.6, however the same trends were seen for the 10 min and the higher aggregation levels. To get an impression on the size of the samples, the average sample size (N) was calculated for all

aggregation levels and can be seen in the top left corner of the tables. From this, it can be seen that especially for Ensemble, the sample size is quite small, worsened by its tendency to underestimate.

Table 4.1: *Average 5x5 Contingency table for Radar 10 min aggregation level over all LTSs and all RGs. The table shows the percentage of data in each category together with sums, bias and Hit Rate. F indicates forecast and O observation. The intervals are numbered from 1 to 5 from lowest to highest. The percent of hits in each category is indicated with bold.*

N=677	O1	O2	O3	O4	O5	sum
F1	**15.61**	12.53	15.10	2.03	1.27	46.54
F2	4.86	**5.89**	9.91	1.91	0.93	23.50
F3	3.25	4.13	**12.03**	3.58	1.74	24.73
F4	0.46	0.41	1.48	**0.70**	0.50	3.55
F5	0.17	0.18	0.71	0.35	**0.27**	1.69
sum	24.35	23.14	39.24	8.56	4.72	100.00
bias	1.92	1.01	0.63	0.43	0.41	
HR	0.64	0.25	0.30	0.08	0.05	

Table 4.2: *Average 5x5 Contingency table for NWP 10 min aggregation level over all LTSs and all RGs. The table shows the percentage of data in each category together with sums, bias and Hit Rate. F indicates forecast and O observation. The intervals are numbered from 1 to 5 from lowest to highest. The percent of hits in each category is indicated with bold.*

N=226	O1	O2	O3	O4	O5	sum
F1	**16.80**	11.80	14.96	1.83	1.00	46.39
F2	6.07	**6.08**	10.95	1.46	0.65	25.22
F3	4.12	4.73	**10.20**	2.07	1.04	22.17
F4	0.48	0.53	2.09	**0.56**	0.33	4.00
F5	0.22	0.21	1.11	0.42	**0.27**	2.23
sum	27.69	23.35	39.31	6.35	3.29	100.00
bias	1.74	1.09	0.57	0.67	0.71	
HR	0.61	0.26	0.26	0.09	0.08	

Table 4.3: *Average 5x5 Contingency table for Ensemble 60 min aggregation level over all LTSs and all RGs. The table shows the percentage of data in each category together with sums, bias and Hit Rate. F indicates forecast and O observation. The intervals are numbered from 1 to 5 from lowest to highest. The percent of hits in each category is indicated with bold.*

N=42	O1	O2	O3	O4	O5	sum
F1	**20.53**	14.36	13.43	0.83	0.22	49.36
F2	7.22	**6.54**	9.34	1.11	0.22	24.43
F3	3.86	5.62	**10.26**	2.67	0.67	23.08
F4	0.51	0.58	1.08	**0.51**	0.05	2.73
F5	0.05	0.03	0.09	0.22	**0.01**	0.40
sum	32.17	27.13	34.20	5.34	1.17	100.00
bias	1.63	0.96	0.70	0.65	0.15	
HR	0.65	0.24	0.30	0.09	0.02	

Looking at the average multi-category contingency table for all three forecast products (Table 4.1 - 4.3), it is seen that in general all three forecasts have problems with predicting the correct intensities. The number of hits is related to the number of observed events in each category. Thus, a decreasing number of hits is seen for increasing threshold, except for the third threshold, which has a higher average number of hits than the second threshold for all three forecasts products, as there are more observations falling within the third interval.

All forecast products have a tendency to underpredict as more counts are seen above the hit diagonal than below, indicating more misses than false alarms. This is also evident from the bias, which is

around 1 or below for all thresholds except the lowest. This indicates that the products overpredict the smallest rain events but otherwise underpredict. Similar trends to what was seen in the QQplots in section 4.1.3 can be identified here. The same tendency to decrease with increasing threshold can be seen for the HR, where the best rate of around 60 % is obtained for the lowest threshold for all three forecast products. Finally, to investigate the performance over lead time, the results of the multi-threshold tests can also be seen in Appendix D.6 for the 1 hour temporal resolution for all LTS, including PC and PSS. From this no clear tendencies in performance over lead time was seen.

Based on this analysis it was concluded that the performance of the products is in general better for rain in the range from 0.1 - 1 mm/h than for larger rain events, however as we are interested in the rain amounts that might have hydraulic affects, focusing lower than 1 mm/h is not appropriate. On the other hand the analysis showed a fast decrease in performance for thresholds above 3 mm/h, which could be related to the low amount of data in this range.

4.2.4 Residual Statistics and Quality over Lead Time

An important aspect of this study is the performance of the forecasts products over lead time. Both the binary statistics and the residuals statistics were investigated over lead time for the three different temporal resolutions. Figure 4.10 and 4.11 show the performance of Radar and NWP over lead time using a temporal resolution of 10 min, and assuming that the first half hour of the Radar data and the first two hours of the NWP data are unavailable for forecasting.

The following trends were noted from the plots (Figure 4.10 and 4.11):

1. Radar has a lower RMSE and a higher NSE than NWP for approx. the first two hours of available lead time, even though no correction factor has been applied. This indicates that even though the Radar underestimates the rain amounts, the error is less significant than the error of the NWP.
2. The correlation coefficient for Radar starts much higher than for NWP but at the time the forecast is available it is almost at the level of NWP.
3. According to the WBE Radar underestimates the rain amount while NWP overestimates the amount, which supports the results from section 4.1.2 and 4.1.3. FBI and WBE agree on the variations over lead time.
4. A lower FAR is seen for the Radar, however a lower HR is also seen indicating that the Radar does not perform better, but it does contain fewer false alarms due to its tendency to underestimate the events.
5. The HR and PSS is close to similar around the beginning of the available lead time, but faster decreasing for Radar.

According to several measures, Radar performs better than NWP for the first LTSs but its performance declines faster. Thus the processing time is of high importance as possibilities of decreasing the processing time would mean increases in forecast performances and larger improvement obtained from using this forecast type compared to using NWP. On the other hand, NWP seems to have a trend of reaching a maximum quality after some lead time before slowly decreasing, showing the effect of spin-up time.

Similar plots using the 30 min temporal resolution can be found in Appendix D.7, and showed similar trends to those observed with the 10 min resolution. Plots for the binary statistics achieved using a threshold of 3 mm/h for all three temporal resolutions can also be found in Appendix D.7. As previously seen the products perform worse for this resolution, with more random results, and the tendency of Radar to perform better for the first hours is not evident for this threshold.

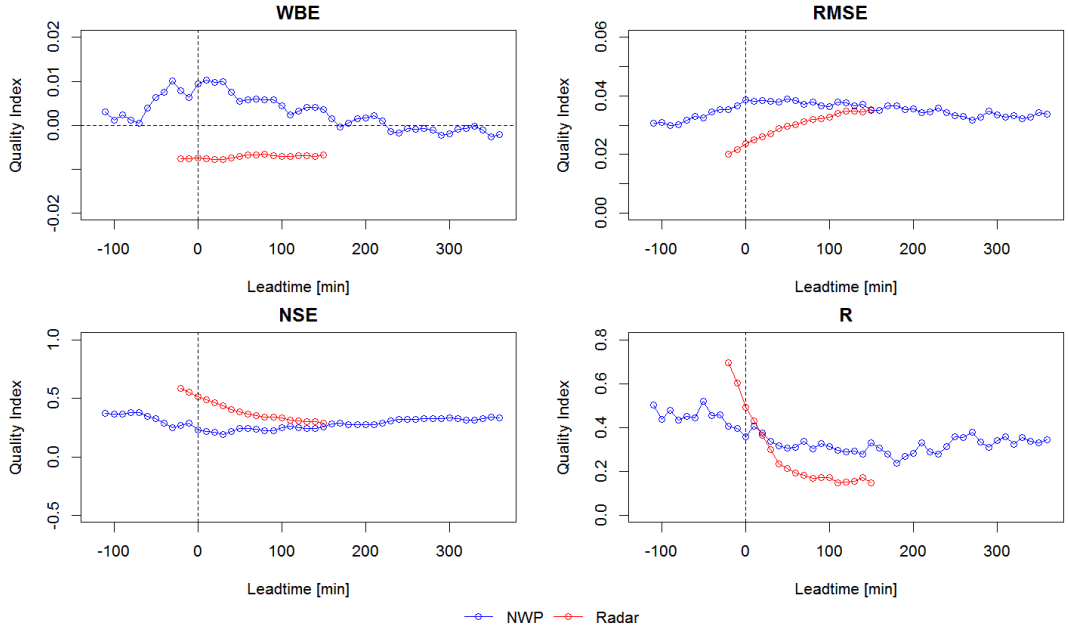


Figure 4.10: *Quality vs lead time for Radar and NWP for four different indices for a temporal resolution of 10 min.*

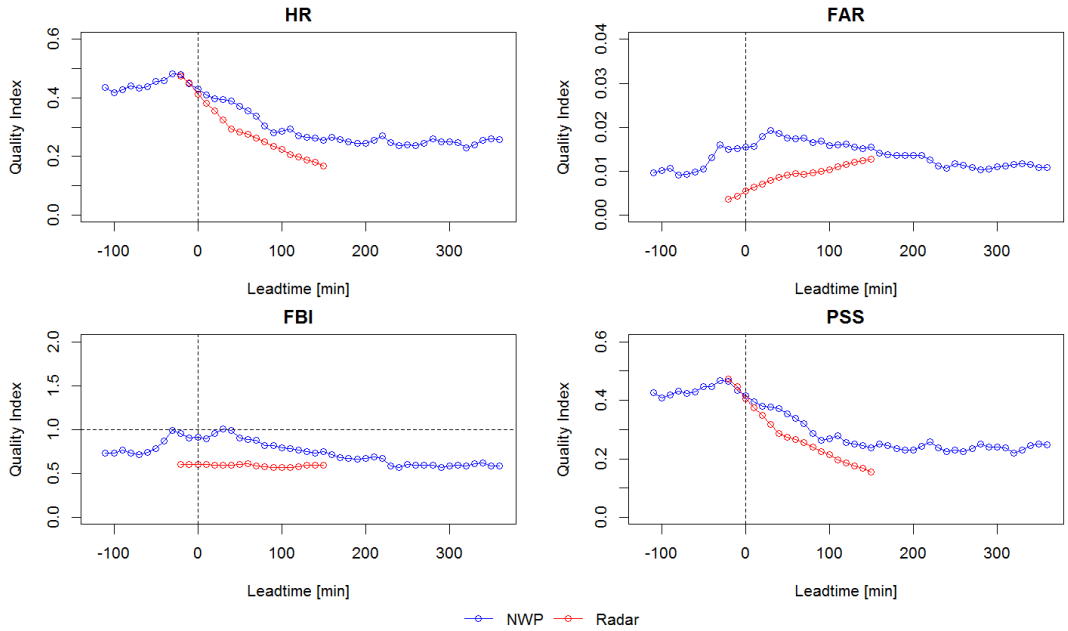


Figure 4.11: *Quality vs lead time for Radar and NWP for four different binary indices for a temporal resolution of 10 min.*

Figure 4.12 and 4.13 show how the tree forecast products perform over lead time for the temporal resolution of 1 hour and with a threshold of 1 mm/h for the binary statistics. It was assumed that the first 3 hours of the Ensemble data are not available for forecasting in this case.

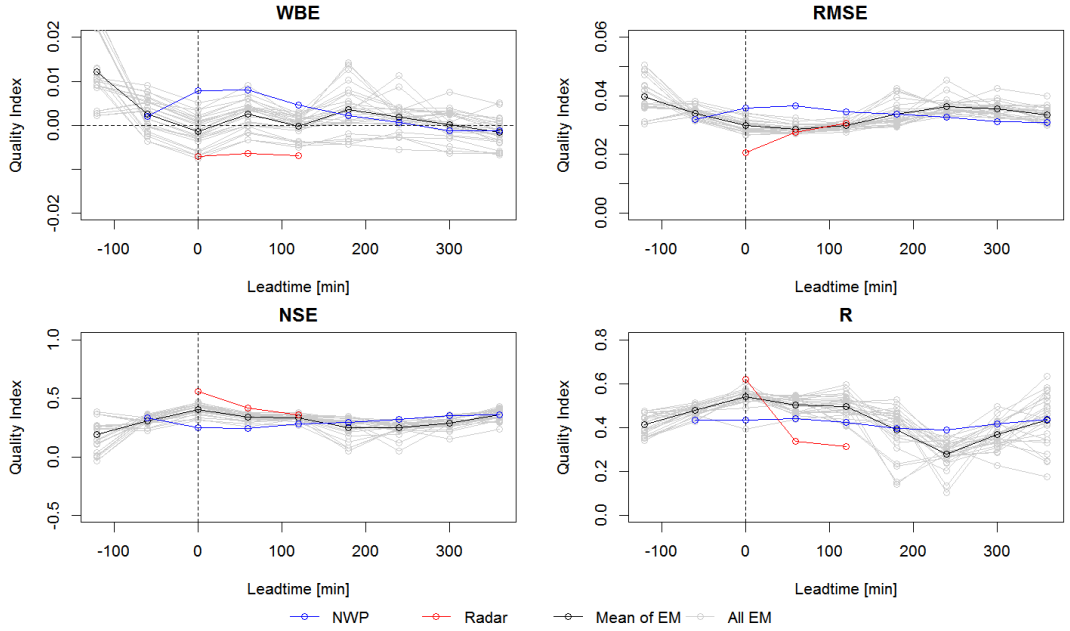


Figure 4.12: *Quality vs lead time for Radar, NWP and Ensemble (both mean of EM and all EMs) for four different indices for a temporal resolution of 1 h.*

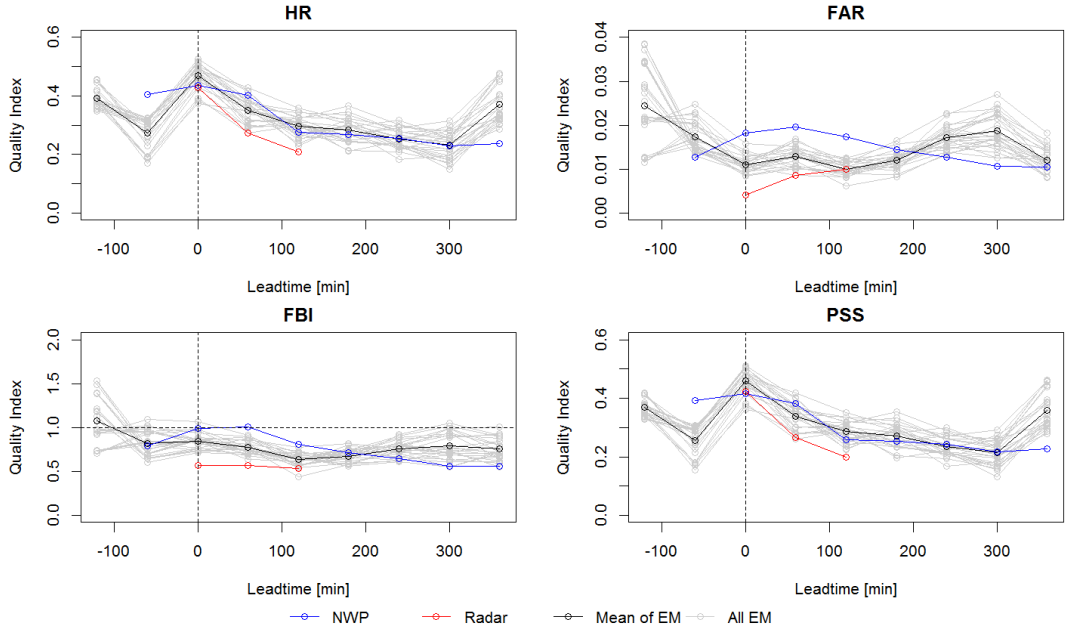


Figure 4.13: *Quality vs lead time for Radar, NWP and Ensemble (both mean of EM and all EMs) for four different binary indices for a temporal resolution of 1 h.*

In general, it is seen from the two figures that NWP and Ensemble perform similarly, however, according to some measures, Ensemble performs better than NWP especially for longer lead times. This could be due to Ensemble having a longer spin-up time than NWP. Radar performs better than NWP and Ensemble according to some measures (e.g. lower FAR and RMSE), however HR and PSS are lower for Radar. Again it is evident that if the processing time for Radar could be decreased, there is potential for further improvement in performance compared to NWP and Ensemble.

Besides this, it was noted that the NSE and PSS did not provide further information on the performance of the forecast products, as they resemble RMSE and HR respectively. However, NSE did show that the forecasts all perform better than the mean of the observations.

As it was shown that the weight of evidence approach might provide useful improvements to the quality of the prediction, this was investigated further, by plotting the results of this procedure as quality over lead time. A similar plot as Figure 4.13 was made using the weight of evidence results instead of the individual EM results, and can be seen in Figure 4.14. From this figure it became more clear that with a conservative approach using a weight of evidence of about 10 EMs or less, an improvement in prediction skill can be achieved at the cost of higher false alarm rates and higher bias for all LTSs.

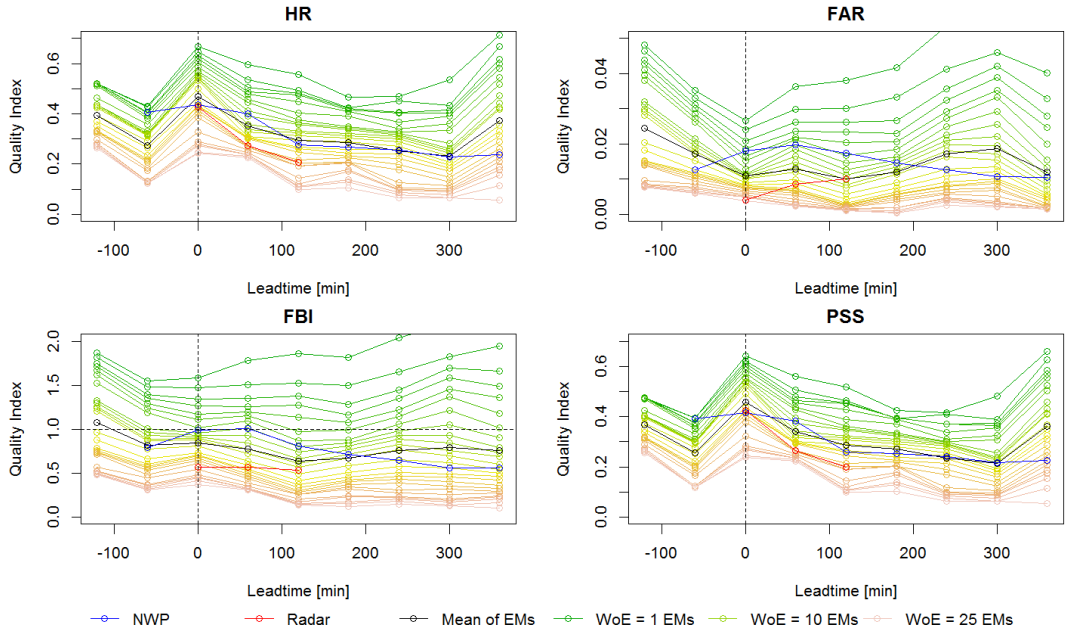


Figure 4.14: *Quality vs lead time for Radar, NWP, mean of the EMs and Ensemble results with the weight of evidence approach for four different binary indices for a temporal resolution of 1 h.*

From these investigation it was thus found that using the Ensemble might provide better predictions than Radar and NWP for some LTSs, even without exploiting opportunities related to having EMs but just focusing on their mean performance. If the weight of evidence approach is applied for Ensemble further improvement in the skill might be achieved when applying a conservative approach (less than 10 EMs predicts the event) and accepting the compromise between higher hit rates and false alarm rates.

4.3 Resolution Investigation

From the previous analyses it was seen that there is potential for Radar data to provide better forecasts than NWP and Ensemble on a short forecast horizon. On the other hand the use of Ensemble may also be beneficial, as some measures suggest a better performance than NWP on longer horizons, and the possibility of using a weight of evidence approach allows for even further improvements.

Now that the three products have been investigated individually and compared, the final section of this part of the study investigates the potentials of improving the performance of the three forecast products further by considering the data in different ways relevant for warning systems. Thus, a spatial and a temporal analysis was conducted. Both analyses highlight possible improvements from expanding the view of the data spatially and temporally and including more grid points or lead time in a warning system, while also considering the possible losses of these procedures. The results of the analyses are presented in the following sections.

4.3.1 Spatial Analysis

A spatial investigation of product performance was conducted for all three forecasts products for the 1 hour temporal resolution, and for Radar and NWP for the 10 min and 30 min resolution. Their performance over the available forecast lead time was investigated and compared by plotting ROC diagrams for the average performance of all RGs and over the available LTSs. Thus it should be noted that for NWP and Ensemble the plots show averages of six hours lead time, while for Radar the plots shows averages of two hours lead time. The results for the 10, 30 and 60 min resolutions can be seen in Figure 4.15, which shows the results of the spatial analysis, together with the results obtained in the ordinary binary analysis with spatial interpolation to the RGs.

The following conclusions can be drawn from these plots:

1. It is seen that only small differences can be observed for the three temporal resolutions for Radar and NWP.

2. It is seen that by increasing the included distance, higher hit rates and false alarm rates can be achieved for all products and all thresholds, however the higher the threshold the more mixed the results. In general the higher the threshold the lower the rates and the bigger the distance the higher the rates.

3. It is seen that for all three forecast products higher hit rates can be achieved from spatial analysis of the data for all distances and for both thresholds. The hit rate also increases more than the false alarm rate in all cases, indicating that the possible improvements obtained from increasing the distance are larger than the worsening caused by the inclusion of more data.

4. When comparing the three forecasts products for the 1h temporal resolution, they perform quite similar, however NWP performs the best for the lower thresholds. This comparison is more clear from Figure E.2 in Appendix E.1, where the results for the three products are plotted together.

A plot showing the performance of the individual EMs averaged over the first 9 hours of lead time and all RGs can be seen in Appendix E.1. No clear tendencies in performance was seen amongst the EMs, however it does seem like the first and last EM perform the best.

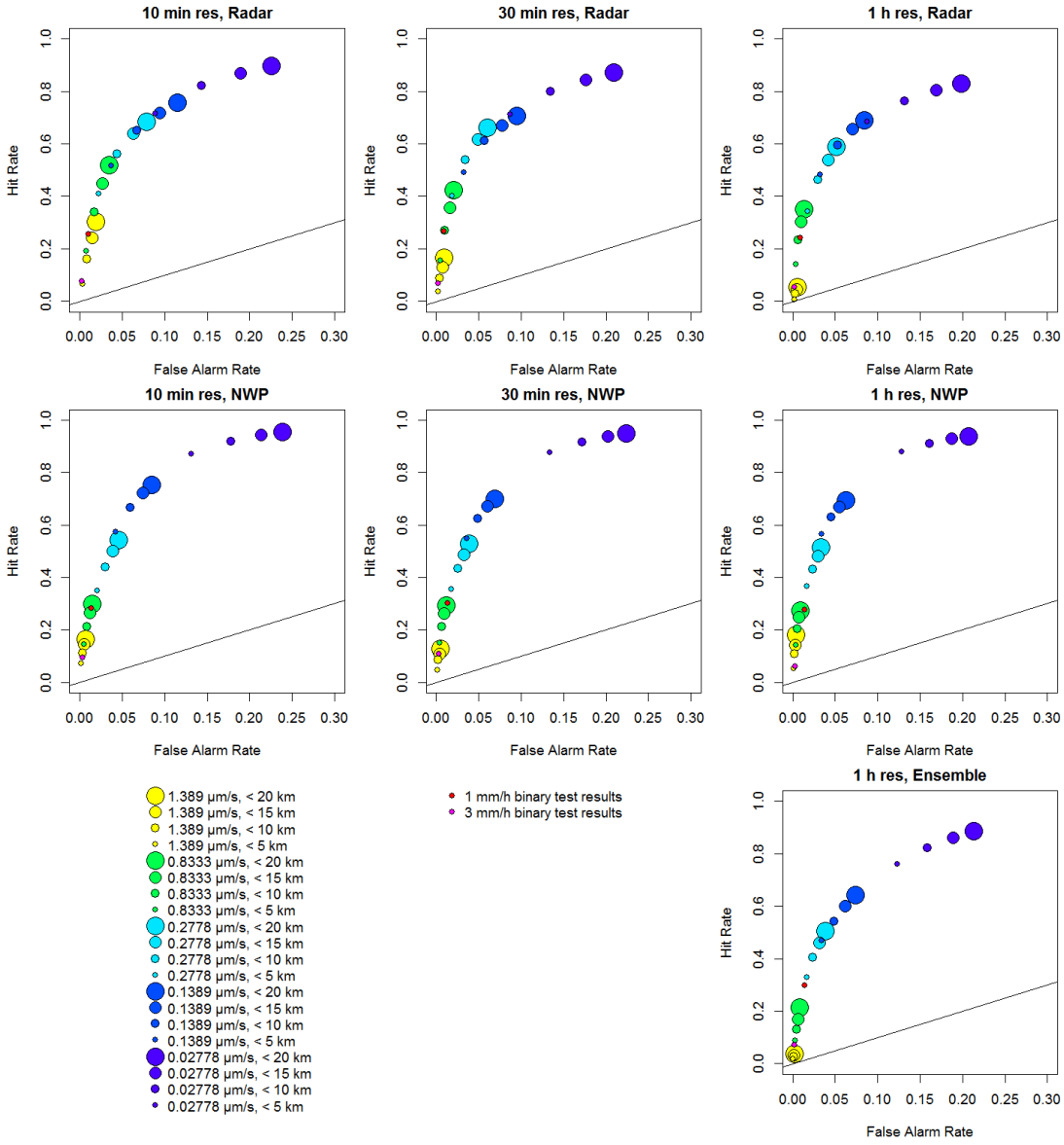


Figure 4.15: *ROC diagrams showing the spatial statistics for the three temporal resolutions for the mean of the available LTSs for all RG. In red and pink are the results obtained from the ordinary binary analysis. Note that the x-axis was stretched to improve visibility.*

The improvements seen using this method are related to the grids of the forecasts and the chosen distances. The larger the distance the higher the risk of ending up at the edge of the grid, and thus not necessarily obtaining improvement from increasing the distance. As NWP has the smallest grid the risk is highest for this product, however it does not seem to affect the results significantly.

All in all, the results of this section indicate that one can obtain higher hit rates at the expense of higher false alarm rates for the lower thresholds by increasing the distance included in the test. An improvement in performance can be obtained for all three products, indicating that in all cases, using forecasts from surrounding areas may improve warning systems. For this analysis the mean of the EMs was used, however as proven in the previous section the performance of the EMs can be improved by applying the weight of evidence approach. Combining these two methods could thus improve the performance of Ensemble even further.

4.3.2 Temporal analysis

The temporal investigation, which included stepwise increasing in the lead time in the test, was conducted in a similar way as the spatial investigation with the same thresholds, for the same temporal resolutions, but with a fixed distance of 5 km. The investigation was conducted using only the available lead time of each dataset. The results of this analysis can be seen in Figure 4.16 for all thresholds and the three temporal resolutions. In Appendix E.2 a plot with the results for all EMs and all thresholds can be seen. This shows that all EMs have similar results with no member-specific tendencies, however performance seem to be lower for the middle EMs.

The following conclusions can be drawn from Figure 4.16:

1. As 5 km distances are used, the results from this analysis is expected to show improvement compared to the results in the spatial analysis with the same distance. This is also evident when comparing with Figure 4.15. Thus by considering the temporal dimension a better forecast can be achieved.

2. As expected, an increase is seen in hit rate when more lead time is included, however this is linked to an increase in false alarm rate.

3. As with the spatial analysis, similar trends are seen for Radar and NWP for the three temporal resolutions, however the highest hit rates are obtained with the 10 min data. This is not an intuitive trend, as higher resolution means more data and less averaging, however the differences are very small and could be a coincidence.

4. In general the three forecast products perform similarly, however NWP performs best and Ensemble worst.

5. For the lower thresholds hit rates close to 1 can be achieved for all three forecast products, however higher false alarm rates are related to these. For the two highest thresholds, the trends are too vague as all values are very low and similar.

6. When comparing to the results from the ordinary threshold test for the thresholds 1 mm/h and 3 mm/h (indicated by red symbols on the figure), it is seen that the hit rate is improved in all cases, except for 3 mm/h for Ensemble, which might be due to shortage of data above 3 mm/h threshold for Ensemble.

7. It is also noticeable that for increasing thresholds steeper curves are observed. This means that depending on the threshold, the gain of including 1 more hour of the lead time will vary, and for higher intensities the timing of the rain prediction is less precise, resulting in larger improvements when more time is included. However, this could also be affected by the data scarcity above higher thresholds. When focusing on the 1 mm/h threshold, it seems that for all three products the biggest gain is obtained when including the first two hours instead of just the first hour.

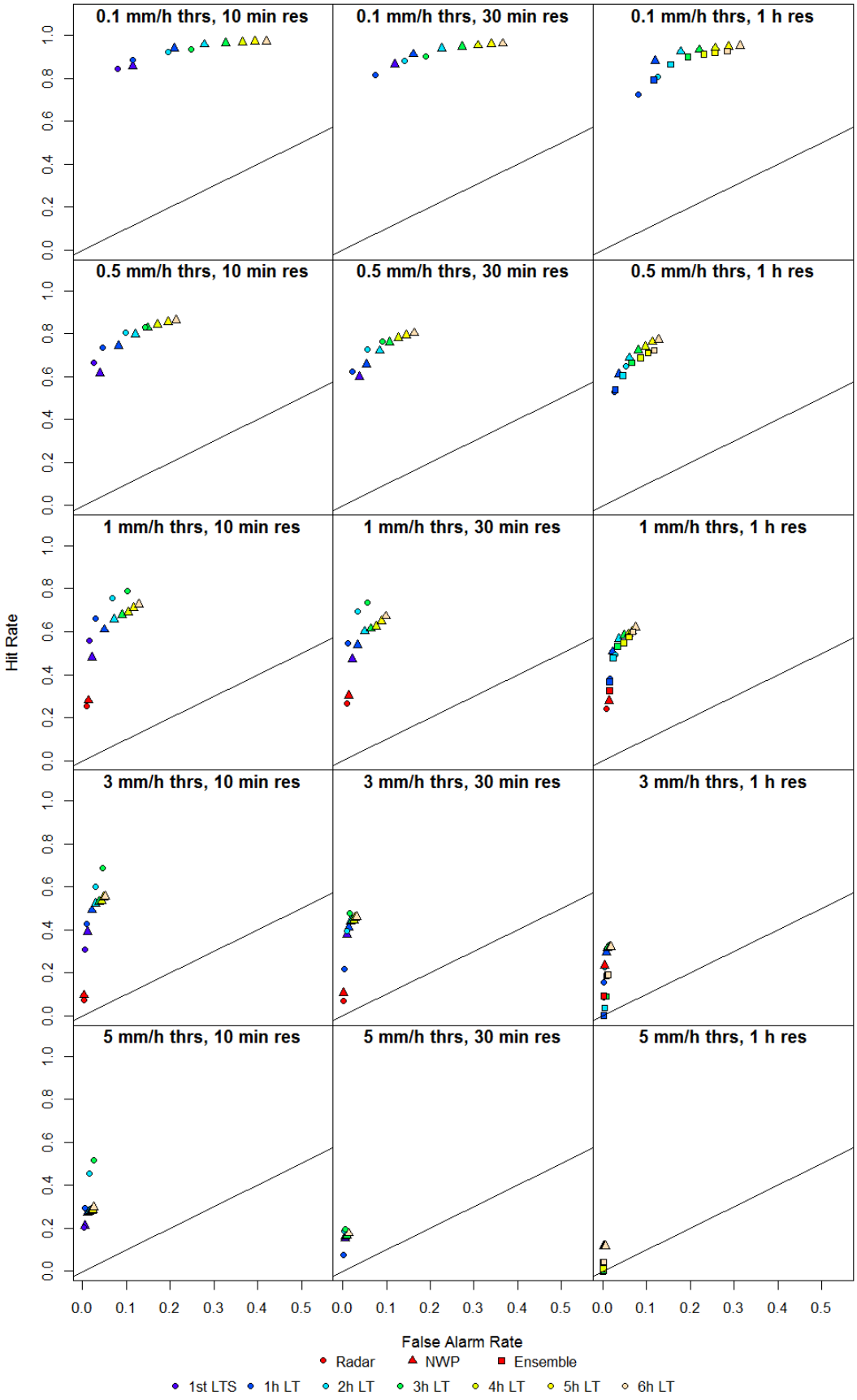


Figure 4.16: *ROC diagrams showing the temporal statistics for the three temporal resolutions for the mean of the RGs. In red are the results obtained from the ordinary binary analysis. Note that the x-axis was stretched to improve visibility.*

The overall conclusion from this investigation is, that considering more of the lead time, i.e. being more cautious on the timing of the events, can improve the hit rate but at the cost of more false alarms. Thus depending on the warning system, applying a conservative approach and including more hours of the lead time, can increase the performance of all three types of forecast products. In the case of flooding, a conservative approach is recommendable. However, the improvement achieved depends on the chosen threshold and amount of lead time included. As the applied method does not consider forecasts ahead of an event, further investigations are necessary to completely describe the benefits of expanding the temporal perspective on the data. Finally, combining the spatial and temporal approach could improve the performance of the forecasts products further and gives the possibility of choosing the thresholds that provides the optimal balance between hits and false alarms.

4.4 General Discussion

Several of the conducted analyses showed that the radar nowcast product has a tendency to underestimate the precipitation. As it is created using Marshall-Palmer coefficients that fit a certain rain drop type best, it can be expected that certain rain types are constantly underestimated. If that type of rain is highly represented in the data period, a bigger underestimation should be expected. The data used in this study covers a period from late summer to winter, and thus spring and summer are not represented. The drop size is known to be bigger in summer due to the higher temperatures, and smaller drop sizes may thus be overrepresented in the data. Besides this, radar stations can be sensitive to low lying clouds, which might occur more often in the fall and winter periods.

In general underestimation is commonly seen for radar data, and DMI uses different correction factors (e.g. Thorndahl et al. (2014)). From the binary analysis a FBI of about 0.55 was seen on average for the RG data and the threshold 1 mm/h, suggesting that for this data a correction factor of around 1.9 would be appropriate to correct for the underestimation. A correction factor of this size is also used by DMI for some radar observations (Vedel, 2016), suggesting that the found factor is not unusual and could be applied to the data. According to the accumulation plot (Figure 4.4) a factor of about 2 is appropriate, however this result is influenced by the large amount of missing data for Radar. According to the quantiles calculated in section 4.1.3 a factor of around 1.6 fit events larger than 0.1 mm/h. Thus it seems that a correction factor in the range of 1.6 to 2 is needed depending on the type of rain event. As larger events are most relevant to this study, this will be investigated further in the event analysis for individual events before choosing and applying a factor.

This study has investigated several methods of potentially improving the performance of the forecasts by considering them in different ways applicable to warning and control systems. As hit rates and false alarm rates are related, there is a trade-off between the two. The selected compromise depends on the purpose of the control or warning system and thus how conservative one wants to be. Furthermore, another important aspect related to warning or control systems is the resolution of the input data needed for the system to be useful. From the conducted analyses it was seen that the Ensemble has the potential of providing better forecasts, however in order to exploit this in a warning system one must be able to compromise on the resolution. Inter-hourly variations including short but high intensity events that can affect the sewer system might not be simulated and trigger the flow control or flood warning systems with an hourly resolution. Thus the question is, whether a high resolution prediction is more important than a more reliable input at lower resolution. So far, this study indicates that there might be a trade off between forecasts quality and both temporal and spatial resolution. The importance of the temporal resolution will therefore be investigated further in the second part of this study.

4.5 Further analysis

As always, some assumptions were made during the investigations conducted in this study. The main assumption of this study is the use of the RG data as reference data. Although these measurements might be close estimates of the truth, they are not the truth. This means that the use of other data as reference might provide very different results. A further investigation that could be relevant in this case is the comparison against radar observations. Radar observations are often used as reference data (e.g. (Liguori et al., 2012)), and provide other opportunities of comparisons with forecasts as radar observations are also gridded. This allows for different neighbourhood methods, e.g. the fraction skill score used by DMI for gridded comparisons, as described in (Korsholm et al., 2015) could be applied. Investigating the use of radar observations is highly relevant for flow and flood forecasting, and thus for the second part of this study, a Radar observation dataset is therefore generated and compared with the RGs for the selected events. Besides this, combining radar observations with measured RG data could provide an even better reference dataset, which might be closer to reality. This was however considered outside the scope of this project.

In general, many different tools are available to characterize model performance, as shown by (Bennett et al., 2013), and thus many further investigations are possible, also including investigation of the uncertainties involved. Forecast verification on its own is a wide topic, and many techniques have been developed to investigate forecast performances (Jolliffe and Stephenson, 2012). The verification measures used in this study are not among the most complex ones, however they are amongst the more common ones. Many more complex skill scores exists, e.g. using penalization or weighting of forecasts such as the Gandin and Murphy or the Gerrity equitable scores (Jolliffe and Stephenson, 2012). However to get a clear impression on the quality of datasets without introducing too many variables, simpler scores are sometimes more efficient. The use of more complex verification measures or uncertainty investigations was not considered relevant for this study, as the main interest lies in the use of the precipitation forecast in flood prediction. Thus a more natural extension of the conducted analyses is to relate the forecast products to runoff in high intensity cases. This is the focus of the second part of the study.

4.6 Conclusions of Part I

The first part of this study aimed at providing understanding of the strengths, weaknesses and differences between the three forecasts products: a Radar Nowcast (Radar), a deterministic Numerical Weather Prediction with radar data assimilation (NWP) and an Ensemble Numerical Weather Prediction system (Ensemble), on a longer time scale. The following main conclusions were found:

1. It was shown that the Radar has potential for providing more precise forecasts than the NWP and Ensemble on a short horizon, however the quality of this forecast product decreases fast with lead time, and thus the possible gain from using this product is highly dependent on the processing time, which is currently about 30 min but could possibly be decreased.

2. The Radar data was not corrected for the underestimations clearly documented in this study. However, applying a correction factor suitable to its performance during high intensity events would increase its performance further.

3. For NWP both over- and underprediction was observed depending on intensity and lead time. It was noted that the data assimilation performed in the first lead time hours is not sufficient to ensure generation of the right amounts of water throughout the entire forecast horizon. It should be noted, that for NWP decreasing the processing time is not as relevant as for Radar,

as it involves data assimilation and nudging, and because the product has a longer spin-up time.

4. NWP and Ensemble showed similar performance for the first 6 hours of available lead time, however it seems Ensemble has a longer spin-up time making it a better choice than NWP for lead times longer than 3-4 hours. Besides this it was seen that the possibility of using a weight of evidence approach for the Ensemble further improves its performance compared to NWP, when a conservative approach is wanted, which is often the case for flood warning systems.

5. Out of the indicators applied in the performance analysis, most information was obtained from the water balance error, bias, RMSE, R and hit rate. Due to the very low false alarm rates, no additional information was obtained from computing the Peirce Skill Score. The Nash-Sutcliffe Efficiency coefficient reflected RMSE, however it did prove that the forecast products all performed better than the mean of the observations.

6. Finally, further investigations showed that expanding the warning area both temporally and spatially can further improve the performance of all three forecast systems.

These results suggest that combining the three forecast products might provide the best result. If the processing time of Radar nowcast could be decreased, this could be combined with the NWP nowcast for the first approx. 2 hours of the forecast and improve the forecast performance. On the other hand, combining the NWP with the Ensemble after 3-4 hours could also improve the performance of the forecast. This also provides a possibility of extending the flood forecast beyond the current 6 hours of SURFF, as the Ensemble has a much longer forecast horizon. However using the Ensemble would, as mentioned in the discussion, mean a compromise on both the temporal and spatial resolution, which might not capture the needed details for flood modelling. This is investigated further in the second part of the study.

All in all the first part of this study showed potential for the three products and their possible combination, however as flood forecasting is the concern of this study, further analysis will be conducted to investigate the performance of the three forecast products during selected high intensity events and in relation to flood prediction.

CHAPTER 5

Methods for Part II - Event Analysis

5.1 Approach of Part II

Following the investigations in part one, a procedure for analysing and comparing the data for specific events was needed. Before further considerations, the following was noted on the choice of baseline scenario.

Areal coverage of a catchment by radar observations is often preferable over RG measurements as they are evenly distributed and often with a higher resolution than the RG network. A network of RGs does not necessarily catch the peak of events as they might fall in a location where there is no RG. Thus radar observations are better at describing the spatial and temporal distribution of events, and provide a better spatial picture of the movement and development of individual events. However rainfall depth may also vary within one radar cell of 500 m as proved by Jensen and Petersen (2005, as cited in (Schellart et al., 2012)) which found variations of up to a factor two in one radar cell. Villarini et al. (2010) investigated the accuracy of radar observations of an extreme flooding event in an urban catchment. They compared bias corrected radar rainfall with RG rainfall and concluded that radar data is useful in describing regional variations in extreme floods. Finally, Schellart et al. (2012) investigated the influence of errors and spatial variability related to using RG measurements and radar observations as input for runoff modelling of a small urban catchment. The two very different measurement methods also have very different inherent error characteristics and thus using the two resulted in very different peak flow and runoff volumes. Based on these reflections, the decision was made to use both the RG measurements and the Radar observations as baseline scenarios for the event analysis in this part of the study.

To obtain the best impression on both what was observed and what was forecasted for the four selected events, the following approach was applied in the event analysis:

1. The events were first described based on literature on the events.
2. An investigation of the measured data for the events based on RG and Radar was conducted.
3. An investigation of the situations forecasted by the three products was conducted, to document the prediction of the events for all relevant forecasts.
4. Baseline flood scenarios of the events were then modelled using the RG data as input to a flood model of Central Copenhagen.

5. Flood simulations based on the most promising forecasts were conducted, and the forecasted floods were compared to the baseline scenarios and differences and trends were discussed.

The Event Analysis is thus separated into two distinct parts: The precipitation analysis and the flooding analysis, both with subsections concerning the observed events and the forecasts performances. The methods used for this part of the study can therefore also be separated in two: 1) the visualization of the precipitation data for the events, and 2) the flood modelling, and will be described in the following sections.

5.2 Precipitation Data Visualization Methods

A major goal of the event analysis is to get an impression on the spatial and temporal distribution of the observed and predicted precipitation during the selected events. Visualizing the spatial distribution of the precipitation for each event at different time steps is therefore an important tool in this part. For the best visualization of both the temporal and spatial dimension, animations are sometimes used, and the reader is therefore in these cases referred to the electronic version of the report for the best display of the results. The applied visualization methods are described in the following sections.

5.2.1 Overview Plots & Animations of Observed Events

To get an impression on both the spatial and temporal extent of the four events the following variables are plotted for the spatially interpolated RG data and the Radar observations:

- **Average areal event intensity profiles** showing the intensities during the events spatially averaged over the grid.
- **Snapshot image of peak intensity time step** showing the spatial image with the maximum areal averaged intensity observed, thus showing the time step where most rain fell on the catchment.
- **Total event accumulation image** showing the spatial distribution of the total accumulated rain during the event.

Besides this, animations were made showing the spatial images of both the spatially interpolated RG and Radar data for each 10 min time step for each event.

5.2.2 Forecast Animations

The relevant forecasts for each event includes forecasts of some time before and after the events to account for temporal misplacements. Thus forecasts covering periods from approx. 9 hours before until 9 hours after each event were included in the analysis. For Ensemble a longer period of forecasts was considered due to the long forecast horizon.

To obtain an initial impression of the forecasts performances and to limit the number of forecasts included in the animations, areal average rainfall intensity plots were made for each forecast. To best visualize the forecasts at each time step, it was decided to combine all relevant forecasts in one animation showing a spatial image for all relevant forecasts at each time step. This way all forecasts can be viewed simultaneously for each event. To substitute the animations for the paper

version of this report, spatial images of accumulated forecasted rainfall during the events, were plotted for selected forecasts.

The figures and animations were used for the following:

- Compare forecast performances with the baseline scenarios.
- Compare performance of each forecast to identify trends for each forecast product.
- Compare the three different forecast products to identify differences in performance.
- For Ensemble: investigate the possible benefits of enlarging the grid.
- Finally, identify the most promising precipitation forecasts for the flood analysis.

5.3 Flood Simulation Methods

The Lynetten catchment is used to investigate the effects of the different precipitation products on flood modelling. To model floods using the MIKE FLOOD software three distinct parts are included: a rainfall-runoff model, a hydrodynamic network model and a 2D surface model. The simulations are conducted with the standard setup used in SURFF, which consists of the following processes: A simple distributed time-area surface runoff model (Model A, (DHI, 2014c)) is used to determine the runoff in each subcatchment based on a concentration time, a time-area curve and an impervious fraction, all defined at subcatchment scale. The hydrodynamic module uses this runoff model as input and solves the dynamic wave approximation of the Saint-Venant equations using the dimensions of the structures implemented in the model (DHI, 2014d). The 1D MIKE URBAN model is coupled to a 2D MIKE 21 model using the MIKE FLOOD tool in MIKE URBAN. This enables the simulation of the exchange of water between the network and the surface models (DHI, 2014b). For further details on MIKE URBAN and MIKE FLOOD, please refer to the documentation indices (DHI, 2014c) and (DHI, 2014b).

The rainfall-runoff model, Model A, assumes no runoff generation from the permeable areas. This means that the model typically underestimates the runoff generation during larger rain events, as it does not take into account the saturation of the soil. An alternative to this setup is the more detailed Model B, which amongst other differences uses Horton's equations to determine infiltration processes (DHI, 2014c). It is not the scope of this study to investigate different model setups, and as the same setup is used for all simulations it does not affect the conducted analyses. However, it is worth keeping in mind the possible underestimation of the true flooding.

SURFF uses A MIKE URBAN network model of the catchment which covers an area of 94 km^2 and consists of 2451 subcatchments with 5988 manholes, 206 basins, 139 outlets and a total pipe length of 505 km most of which is the combined sewer system. In the flood simulations only the central part of Copenhagen (approx. 15 km^2) is modelled in SURFF to reduce computation time. The areas included in the 1D and 2D models can be seen in Figure 5.1. The flooding is computed with a a multi cell solver using two different Digital Elevation Models (DEMs) of Central Copenhagen with resolutions of 16x16 m and 4x4 m, where the first is used for the numerical simulations and the second is used to distribute the flooding according to depth (DHI, 2014a). The time steps used for the flood simulations are all in the range of 2-3 seconds to ensure model stability. In all simulations two hours were added at the end of the events to ensure coverage of the peak of the flooding.

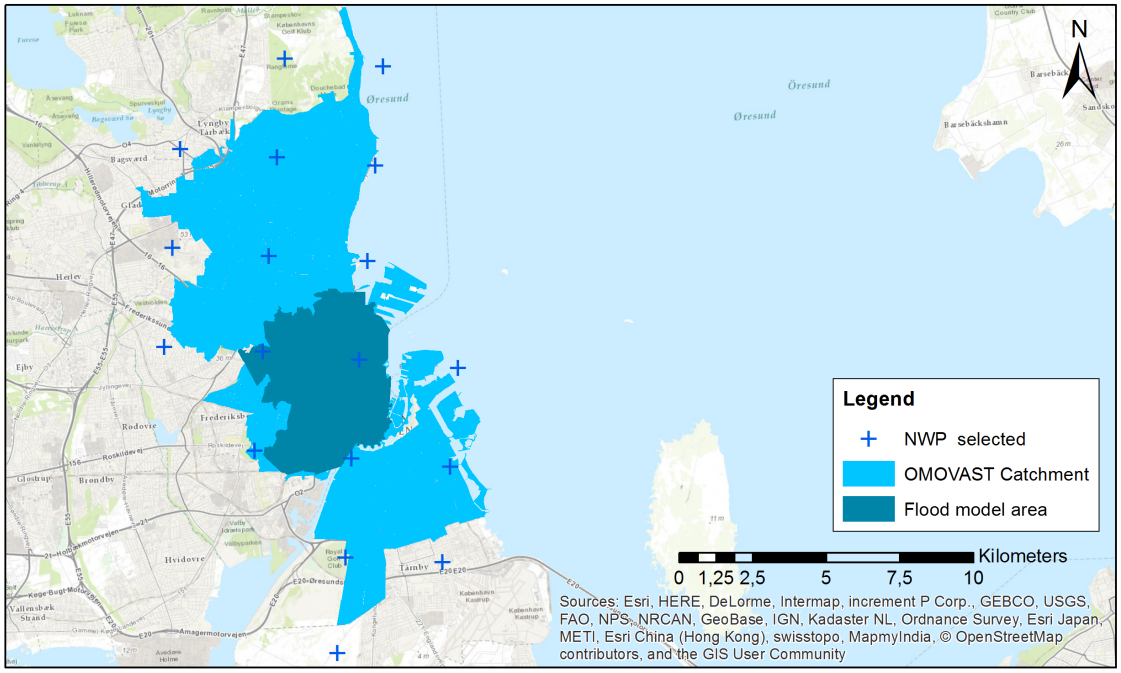


Figure 5.1: *Map of the catchment used in SURFF with the selected NWP grid centerpoints and the area included in the flood model marked in darker blue.*

5.3.1 Rainfall Input

Rain is applied in MIKE URBAN as a catchment boundary. If coordinates are specified, the rain will be applied using a geographical proximity principle according to the centerpoint of the catchments (DHI, 2014c). This means that only one rain time series is applied per catchment independent of its size. As the size of subcatchments varies from less than 200 m^2 to more than 2 km^2 the effect of the selected grid resolution varies over the catchment.

The SURFF model uses the NWP data that covers the catchment as rainfall input (18 data grid points, Figure 5.1). In the case of the NWP grid no subcatchment is larger than the grid cells, however to fully utilize information from higher resolution grids, like the Radar grid, an alternative method of applying the rain would be necessary. The NWP setup was chosen as the basis for all simulations in this study, and thus only the rainfall input values were changed, not the spatial resolution of the input. The RG, Radar and Ensemble data was thus interpolated to the NWP grid, as described in section 2.3, and the relevant 18 point were imported into the model. For the interpolation a cutoff distance of 2 km was used for Radar as for the analyses in Part I, however for ensemble the downscaling meant that a cutoff distance of 5 km was necessary to ensure at least one grid point contributing to each NWP grid point, and resulted in the number of contributing points ranging from 1 to 4. By equalizing the spatial resolution of the forecasts, the effect of using the different forecast products is investigated independently of the effects of changing spatial resolution in the model.

5.3.2 Hotstart

Hotstart of the network is possible in MIKE URBAN, and is relevant in this study when a forecast starts after an event. Hotstarting the system means including initial conditions of the various storages in the collection system, based on a result file from another simulation and a hotstart time (DHI, 2014c). For forecasts starting after an event, the conditions of the corresponding baseline scenario at the start of the forecast was loaded as hotstart of the system.

Hotstart of the water on the 2D surface is also possible with MIKE FLOOD (DHI, 2014b). In SURFF, hotstarts of both the drainage system and the surface water levels are conducted, however the latter was not included in this study. As far as possible forecasts before the occurrence of flooding was selected for simulation. However in some cases, flooding occurs in the baseline scenario prior to the forecast start, and thus the forecast simulations will result in less flooding than the baseline without necessarily being incorrect. The same problem occurs when a forecast does not cover the entire event. This is a drawback of the applied method and must be taken into account when analysing the flood simulations.

5.3.3 Simulated Scenarios

For this part of the study, a simple simulation approach was chosen to provide an indication of the potential of the forecast products in flood prediction. The baseline scenarios were based on the RGs interpolated to the NWP grid, and simulations were run for the four events using both the 10 min and the 60 min aggregation. Flood simulations were conducted for the best performing forecasts, and the baseline simulations were used as hotstart for forecasts starting after significant parts of the event. The selected forecasts, the start and end of the simulated period, as well as the use of hotstart can be seen from Table 5.1. Effective rain period refers to the period where most of the rain was seen. This was included to provide better indication of whether the forecasts cover the significant parts of the event. In most cases, the start of this period also indicates the time after which hotstart was applied. A list of all conducted simulations can be seen in Appendix H.1. Slightly different approaches were used for the three forecast products:

Radar: Due to the nature of the Radar data and its short forecast horizon, it was often the case that the late forecasts performed better than the earlier ones, and thus the simulations always included a hotstart. This also meant that when choosing radar forecasts after the major peak of the rain, the flooding simulated is largely affected by the baseline scenario used as hotstart, and thus does not only reflect the prediction of the forecasts. SURFF is only run every hour and thus the Radar forecasts provided between full hours are never used as model input. This was taken into account when selecting Radar forecasts for simulation.

NWP: In most cases the selected forecasts either started before the events or at least before the significant peaks of the events. However, as they are only 8 hours long, they did not always cover the entire events, and in these cases, less flooding than the baseline scenarios can be expected. Event 2 is a long event consisting of two major peaks with some hours between, and thus hotstart was necessary for the forecasts covering only the second peak. On the other hand the selected forecasts for the first peak did not cover the second peak and all selected forecasts can thus be expected to result in less flooding for this event.

Ensemble: As the Ensemble forecast have a longer horizons, all selected forecasts cover the entire peak of the events, and thus hotstarts were not necessary for this forecast product. For all forecasts, the period covering the event was extracted. In the few cases, where the forecast started after or ended before the event, only the available period was used.

If a forecast does not cover the entire event period it cannot be expected to predict the same amount of flooding as the baseline scenario. In general, simulations that do not cover the same periods are obviously incomparable. However, the conducted flood forecast investigation does provide an initial impression on the performance of the forecast products in flood prediction. This was the aim, as a simple, not overly time consuming, simulation approach was wanted. The prediction of flooding or not was considered the important parameter in these investigation, as this is the purpose of SURFF and the focus of this study. This means that the conducted simulations are

subjects to a less quantitative and more qualitative, visual investigation of the performance of the forecasts. Further investigations are therefore needed to allow quantitative comparisons of the products performance.

Table 5.1: *List of observed events and selected forecasts for flood simulation, with the start and end of their simulation period, and information on hotstart. The start and end of the effective rainfall period is also indicated for the four events.*

Event	Start	End	Hotstart
RG Event 1 Simulation	2011-07-02 08:00	2011-07-03 06:00	-
RG Event 1 Effective Rain Period	2011-07-02 15:00	2011-07-02 21:00	-
NWP FC 2011-07-02 11:00	2011-07-02 11:00	2011-07-02 19:00	no
NWP FC 2011-07-02 12:00	2011-07-02 12:00	2011-07-02 20:00	no
NWP FC 2011-07-02 13:00	2011-07-02 13:00	2011-07-02 21:00	no
Radar FC 2011-07-02 17:00	2011-07-02 17:00	2011-07-02 20:00	yes
Radar FC 2011-07-02 18:00	2011-07-02 18:00	2011-07-02 21:00	yes
Radar FC 2011-07-02 19:00	2011-07-02 19:00	2011-07-02 22:00	yes
Ensemble FC 2011-07-01 06:00	2011-07-02 08:00	2011-07-03 05:00	no
Ensemble FC 2011-07-01 12:00	2011-07-02 08:00	2011-07-03 06:00	no
Ensemble FC 2011-07-01 18:00	2011-07-02 08:00	2011-07-03 06:00	no
Ensemble FC 2011-07-02 00:00	2011-07-02 08:00	2011-07-03 06:00	no
Ensemble FC 2011-07-02 06:00	2011-07-02 08:00	2011-07-03 06:00	no
RG Event 2 Simulation	2014-08-30 06:00	2014-09-01 06:00	-
RG Event 2 Effective Rain Period	2014-08-30 15:00	2014-08-31 21:00	-
NWP FC 2014-08-31 00:00	2014-08-31 00:00	2014-08-31 08:00	no
NWP FC 2014-08-31 01:00	2014-08-31 01:00	2014-08-31 09:00	no
NWP FC 2014-08-31 05:00	2014-08-31 05:00	2014-08-31 13:00	yes
NWP FC 2014-08-31 06:00	2014-08-31 06:00	2014-08-31 14:00	yes
NWP FC 2014-08-31 07:00	2014-08-31 07:00	2014-08-31 15:00	yes
NWP FC 2014-08-31 08:00	2014-08-31 08:00	2014-08-31 16:00	yes
NWP FC 2014-08-31 09:00	2014-08-31 09:00	2014-08-31 17:00	yes
NWP FC 2014-08-31 10:00	2014-08-31 10:00	2014-08-31 18:00	yes
NWP FC 2014-08-31 11:00	2014-08-31 11:00	2014-08-31 19:00	yes
Ensemble FC 2014-08-29 00:00	2014-08-30 06:00	2014-08-31 05:00	no
Ensemble FC 2014-08-29 06:00	2014-08-30 06:00	2014-08-31 11:00	no
Ensemble FC 2014-08-29 12:00	2014-08-30 06:00	2014-08-31 17:00	no
Ensemble FC 2014-08-30 00:00	2014-08-30 06:00	2014-09-01 05:00	no
Ensemble FC 2014-08-30 06:00	2014-08-30 06:00	2014-09-01 06:00	no
Ensemble FC 2014-08-30 12:00	2014-08-30 12:00	2014-09-01 06:00	no
Ensemble FC 2014-08-30 18:00	2014-08-30 18:00	2014-09-01 06:00	no
Ensemble FC 2014-08-31 00:00	2014-08-31 00:00	2014-09-01 06:00	no
RG Event 3 Simulation	2015-09-03 21:00	2015-09-04 19:00	-
RG Event 3 Effective Rain Period	2015-09-04 05:00	2015-09-04 10:00	-
NWP FC 2015-09-04 03:00	2015-09-04 03:00	2015-09-04 11:00	no
Radar FC 2015-09-04 06:00	2015-09-04 06:00	2015-09-04 09:00	yes
Radar FC 2015-09-04 06:20	2015-09-04 06:20	2015-09-04 09:20	yes
Radar FC 2015-09-04 06:30	2015-09-04 06:30	2015-09-04 09:30	yes
Radar FC 2015-09-04 06:40	2015-09-04 06:40	2015-09-04 09:40	yes
Radar FC 2015-09-04 06:50	2015-09-04 06:50	2015-09-04 09:50	yes
Radar FC 2015-09-04 07:00	2015-09-04 07:00	2015-09-04 10:00	yes
Ensemble FC 2015-09-02 06:00	2015-09-03 21:00	2015-09-04 11:00	no
Ensemble FC 2015-09-02 12:00	2015-09-03 21:00	2015-09-04 19:00	no
RG Event 4 Simulation	2016-06-15 13:00	2016-06-16 23:00	-
RG Event 4 Effective Rain Period	2016-06-15 19:00	2016-06-16 15:00	-
NWP FC 2016-06-15 15:00	2016-06-15 15:00	2016-06-15 23:00	no
NWP FC 2016-06-16 06:00	2016-06-16 06:00	2016-06-16 14:00	yes
Ensemble FC 2016-06-14 00:00	2016-06-15 13:00	2016-06-16 05:00	no
Ensemble FC 2016-06-14 06:00	2016-06-15 13:00	2016-06-16 11:00	no
Ensemble FC 2016-06-14 12:00	2016-06-15 13:00	2016-06-16 17:00	no
Ensemble FC 2016-06-14 18:00	2016-06-15 13:00	2016-06-16 23:00	no
Ensemble FC 2016-06-15 00:00	2016-06-15 13:00	2016-06-16 23:00	no
Ensemble FC 2016-06-15 06:00	2016-06-15 13:00	2016-06-16 23:00	no
Ensemble FC 2016-06-15 12:00	2016-06-15 13:00	2016-06-16 23:00	no
Ensemble FC 2016-06-15 18:00	2016-06-15 18:00	2016-06-16 23:00	no

CHAPTER 6

Results of Part II - Event Analysis

6.1 Description of Events

As mentioned in the introduction, section 1, Copenhagen has experienced some extreme rain events over the last years. Three of the mentioned extreme events were selected for investigation in this part of the study along with one smaller, more recent event. A short description of the four events follows:

Event 1 is a cloudburst that occurred in the evening of Saturday the 2nd of July 2011 and was categorised as a more than 100 year event. The period up to the event had been dry except for one smaller rain event in the morning hours on Saturday. According to Vejen (2011) and Beredskabsstyrelsen (2012) convective clouds developed over Southern Sweden, moved slowly across Øresund in a south-western direction, where they increased in magnitude, and around 17:00 UTC in the evening the storm hit around Hellerup, moved slowly southwards and almost stagnated over Copenhagen, where it dropped large amounts of water as well as hail and lightning. All rain fell within a short period of around 3 hours. The largest precipitation amount was measured in the botanical garden in Central Copenhagen at around 135 mm in 24 hours and the highest 10 min intensity measured in 55 years was observed in Ishøj: 30 mm in 10 min. Thus the event resulted in large amounts of flooding in the Greater Copenhagen area including damages to infrastructure and properties (Beredskabsstyrelsen, 2012).

Event 2 occurred in the night between Saturday the 30th and Sunday 31st of August 2014. Small rain events were observed Saturday afternoon and evening after a longer dry period. The main event moved in from the South and hit Copenhagen with heavy showers and thunder caused by convective weather during the night and early morning hours. It continued to rain all Sunday with several peaks in intensity. The event was categorised as a more than 20 year event where it hit the worst. The event resulted in flooding several places in Copenhagen. The largest amount of rain fell over Østerbro, where a total of 135 mm was measured (Hansen and Pedersen, 2014).

Event 3 occurred in the morning of Friday the 4th of September 2015. It was a short, intense event that lasted only around 6 hours. The period up to the event had been dry. The event developed quickly and was a convective event caused by a low pressure over the North Sea, that moved in over Copenhagen from the south and continued north. Both rain and hail was observed during the event. The highest amount of rainfall was measured in Hellerup just north of Copenhagen as 44 mm (Siewertsen, 2015). The event was also categorised as a more than 20 year event in the worst places.

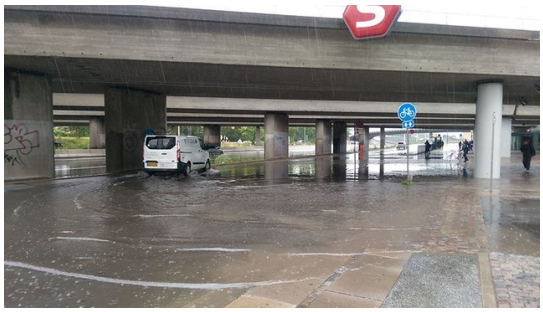
Event 4 started in the evening on Wednesday the 15th of June 2016 and continued until midday the next day. It was a longer rain event building up to high intensities and reaching its peak intensity around 7:00 UTC in the morning. The largest rain amount measured during the event was 55 mm in 14 hours in Skovshoved just north of Copenhagen (Siewertsen, 2016). The period up to the event had been relatively dry except for a little rain on Tuesday, however the event was part of a series of cloudbursts hitting several places in Denmark during Wednesday and Thursday. The rain intensities that hit Copenhagen were high enough to be considered cloudbursts in two places: Ved Lygten in northwest Copenhagen and at the faculty of Science, Copenhagen University in Frederiksberg (Siewertsen, 2016). The event was not as large as the other three, and can be categorised as around a 5 year event (Thomsen, 2016).

Photographs of some of the flooding observed during the four events can be seen in Figure 6.1.

(a) 02-07-2011, Lyngbyvej at DMI, Copenhagen (Andersen, 2011)

(b) 31-08-2014, Lyngbyvej at DMI, Copenhagen (Elkjær, 2014)

(c) 04-09-2015, Tuborgvej, Copenhagen (Høgsholt, 2015)



(d) 16-06-2016, Lyngbyvej at DMI, Copenhagen (Siewertsen, 2016)

Figure 6.1: *Photographs of flooding caused by the four events all taken around the same area in Copenhagen.*

6.2 Observed Events

The accumulated precipitation recorded during the four events by the 37 RGs can be seen in Table F.1 in Appendix F.1. A summary of the events as they were recorded by the RG and the Radar data is shown in Table 6.1. The values for the Radar data after correction with the factors identified in section 2.5 was also included to provide an impression on the effects of correcting.

From the table, the following was noted: Event 1 and 3 were short, high intensity events while Event 2 was a longer event with high intensities and Event 4 was a long event with lower intensities. Large differences are seen between Radar and RGs even after the Radar has been corrected based on the RG data. It seems that the corrected Radar data overestimates both the accumulated rain and the peak intensities. This is caused by the fact that while some single radar cells may measure very high values, on average, correction is needed. It should be noted that the original Radar data

was used in the following precipitation investigations, and the corrected data was only used for the flood simulations.

Table 6.1: *Overview of the four selected rain events: Cumulative rainfall and peak rain intensities based on RGs and Radar data before and after correction for underestimation (Radar and Radar corr.). Peak intensities are based on 10 min data.*

			Max cum. rainfall [mm]			Max peak intensity [$\mu m/s$]		
Event	Start [UTC]	Duration [h]	RGs	Radar	Radar corr.	RGs	Radar	Radar corr.
1	2011-07-02 15:00	6	119	103	149	49	57	83
2	2014-08-30 14:00	31	129	-	-	29	-	-
3	2015-09-04 05:00	6	44	28	55	26	22	43
4	2016-06-15 19:00	19	58	40	74	14	32	58

Event areal average profiles, snapshot of areal peak intensity and accumulated rainfall can be seen in Figure 6.2 - 6.8 for the RGs spatially interpolated to the Radar grid and the generated Radar observations. Similar figures with the RGs interpolated to the NWP and the Ensemble grid can be found in Appendix F.3. Animations of the observed events with the two data types can be seen in Appendix G. From the figures and animations the following was noted for the four events:

Event 1 consists of one extreme peak lasting only around 3 hours from 17.00-20.00 UTC. The two measurement types agree on the peak intensity time, but not location or magnitude. They agree on the total rainfall amount but not completely on the location, especially the northernmost high rainfall amount is not seen in the Radar observations.

Event 2 has two major peaks from 23.00 on Saturday to 13.00 on Sunday, with a small break in between at around 5:00-9:00 UTC. According to the RG the majority of the rain fell along the coast, however the highest intensity was observed around Central Copenhagen. No Radar data was available for comparison for this event.

Event 3 consists of one major peak at around 5:00-9:00 with a smaller second peak at around 10.00-11.00. There is a high agreement between the two datasets on both peak intensity time and location, however less on the intensity magnitude and total rainfall amount, as well as the spatial extent of the event. The highest rainfall amount are seen in outer Østerbro and Hellerup area in both datasets.

Event 4 consists of one long peak from midnight to around 8:00 UTC with smaller peaks before and after. The shown peak intensity for radar is 40 min after the one for the RGs, however from the animations a second peak in intensity can be observed in the Radar data at 06.30, but at a different location. In general, large differences are seen between the two datasets. From the animations it is also evident that Radar was affected by noise in this period.

In general, the animations (Appendix G) support the observed temporal correlation between the two datasets, especially for Event 3, often with no or a few time steps between similar spatial images. All in all, the two datasets seem to agree on temporal and to some extent spatial distribution of the events, but disagree on the magnitude of the intensities and in some cases also on the rainfall amount. Based on these observations it was decided to continue to use the RG data as baseline in the further investigations.

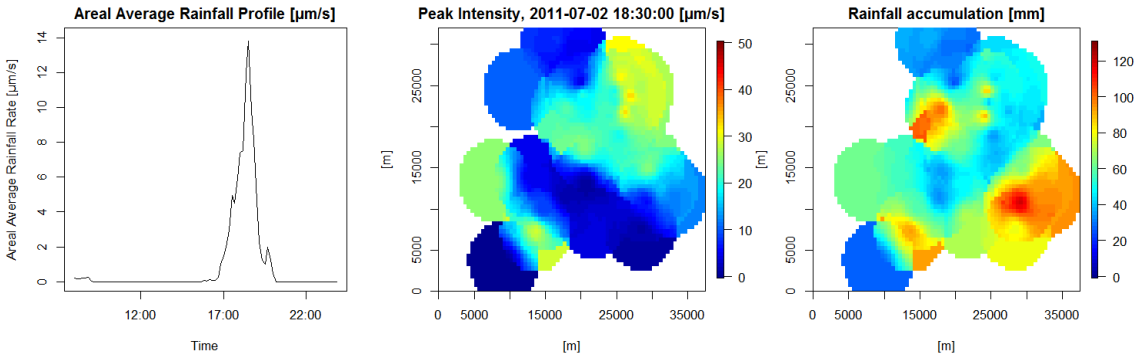


Figure 6.2: *Event 1 described by the 10 min RG data interpolated to the Radar grid. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

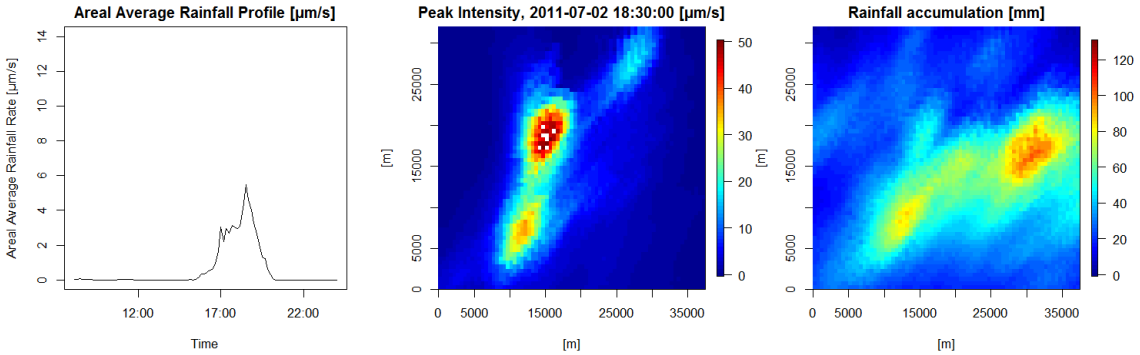


Figure 6.3: *Event 1 described by the 10 min Radar data. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

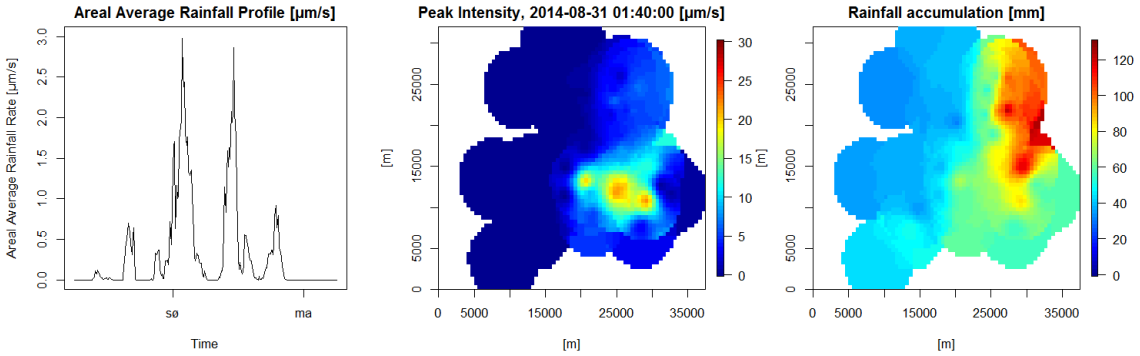


Figure 6.4: *Event 2 described by the 10 min RG data interpolated to the Radar grid. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

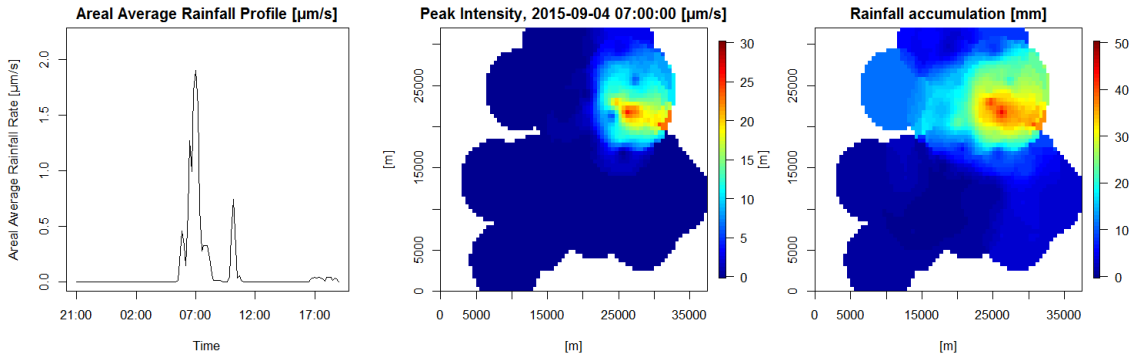


Figure 6.5: *Event 3 described by the 10 min RG data interpolated to the Radar grid. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

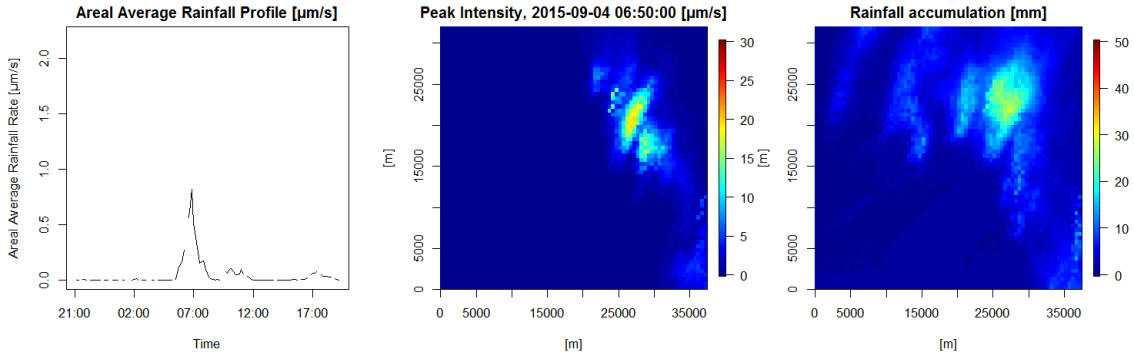


Figure 6.6: *Event 3 described by the 10 min Radar data. The areal average event intensity profile (left), snapshot image during the peak intensity period op the event (middle) and total event accumulation (right).*

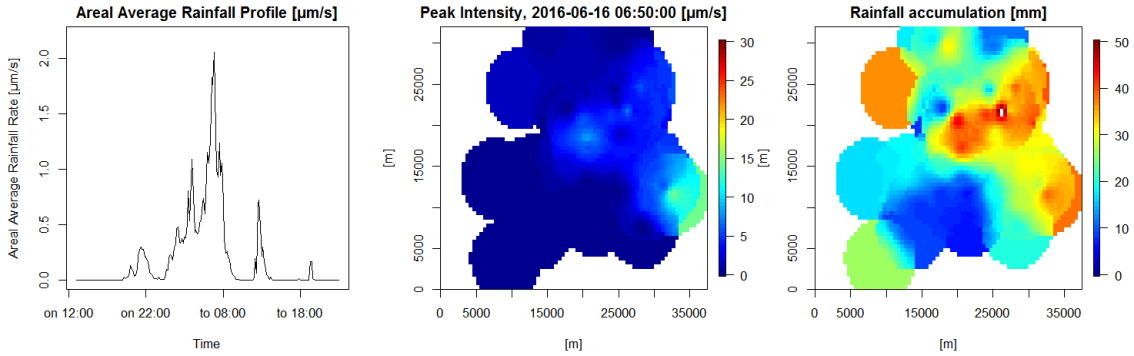


Figure 6.7: *Event 4 described by the 10 min RG data interpolated to the Radar grid. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

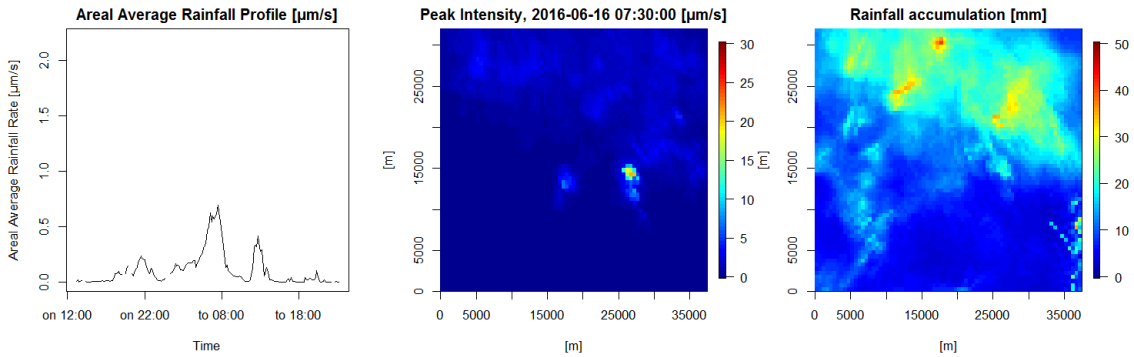


Figure 6.8: *Event 4 described by the 10 min Radar data. The areal average event intensity profile (left), snapshot image at peak intensity (middle) and total event accumulation (right).*

6.3 Forecasted Events

An initial impression on forecast performance was obtained from areal average profiles. For each of the forecasts covering the most relevant period of the events, a plot was made showing the areal average for all LTS of the forecast and the RG areal average spatially interpolated to the respective grids. For Radar and NWP this was done for the 10 min temporal resolution and for Ensemble the 60 min temporal resolution. For Ensemble, one plot with all 25 EMs and one plot with the maximum of the EMs at each time step was created for each event. The plots can be seen in Figure F.11 - F.25 in Appendix F.4.

From the area average profiles, it was seen that for all four events all three forecast products predict something, however with very varying precision. These figures were then used to identify the best performing forecasts. Based on the initial impressions, the range of relevant forecasts was narrowed down for further visualizations. Remembering that the investigated plots are areal averages, they do not reveal anything about the spatial distribution of the forecasts nor about the maximum predicted intensities. To investigate the spatial precision, animations were therefore made showing spatial images of each of the relevant forecasts at each time step. These animations can be seen in Appendix G. For Ensemble the maximum of the EMs was used in this case. The performance of each EM during a forecast can be seen in the animations in Appendix G.5 for one selected forecast for each of the four events. As previously concluded, no clear trends in performance amongst the members was seen.

As a paper compatible alternative to the animations, figures showing accumulated rainfall for selected forecasts, can be found in Appendix F.5. It should be noted that the accumulations were done on the entire forecast lead time, thus also including the periods which are not available for the user. This was chosen to fully utilize the provided information. For Ensemble the max of the forecasts for each LTS was used, and the forecast were only accumulated over the period covering the events, to avoid including rain from other events. These accumulation plots can be compared with the accumulation plots in Figure 6.2, 6.5 and 6.7 for Radar and in Figure F.2 - F.9 for NWP and Ensemble. The areal average profiles, animations and accumulation plots revealed the following main trends for the three forecast products:

Radar: A clear tendency of decreasing rain amounts with LTS is seen for all four events from the areal average profiles. This reflects the inability of the extrapolation to predict increases in precipitation, a trend that might be caused by several factors: A too fast extrapolation of the current situation or a sightly wrong direction of the movement which removes the event too fast from the system. In case of stagnating events, Radar may have difficulties seeing the event, which may help explain differences seen for Event 1. These troubles mean that the first LTS might be correct but as the forecasts almost never predict increasing rain, poor performance is seen during the build-up of the events, while a better performance is seen during the decrease of the event. High intensities are predicted at some time steps but often in a more narrow spatial extent than the other forecasts (perhaps influenced by the higher spatial resolution). Sometimes the highest intensities resemble noise more than rain, and especially for Event 4, some abstract behaviour is seen. Accumulating the Radar data results in clear tracks of the event peaks across the grid. This is due to the fact that the forecast model is an extrapolation of the current conditions, meaning that it is not as prone to simulate changes as the other model but instead resembles the current condition moving with a certain speed. This combined with the high resolution of the radar data provides a clearer impression on the direction and speed of the events. For instance, the data revealed that the main peak in Event 1 was forecasted to move with a speed of approx 11 m/s. The accumulation of the Radar forecasts does not cover the entire events horizon in most cases, and thus the accumulated

amounts cannot be compared directly to the RG amounts. For Event 3 it was noted that the highest accumulated rain amounts were located outside the catchment.

NWP: The best performing NWP forecasts agree with the observations on the peak locations for Event 1, 2 and 4, but they all underestimate the rainfall amounts. However, not all NWP forecasts cover the entire events, and thus results in less accumulation, making the amounts incomparable. Especially Event 3 is predicted poorly by the NWP forecasts, as only two forecast predicts high intensities: 2015-09-04 03:00 and 06:00.

Ensemble: From the areal average profiles it is seen that the variance in the number of EMs predicting the events is large. Event 3 was missed by almost all EMs, until a certain forecast, while almost all EMs forecasted rain in all forecasts covering Event 2. The Ensemble forecasts covers the entire event in most cases and very high rainfall amounts and intensities were predicted for all events, however always with the highest accumulated values outside the Greater Copenhagen area (e.g. Event 2, Figure G.12 and F.30). In the area of interest, Ensemble often estimates rainfall amounts of a similar magnitude as observed by the RGs, however the forecasted peak locations vary and often do not agree with the observed situation. The possible gains from increasing the grid size will be investigated further below.

An overview of maximum intensities and accumulated rainfall observed by the RGs and forecasted by the relevant forecasts for the four events can be seen in Table 6.2-6.4. It should be noted that in this case the maximum approach was not applied on Ensemble, meaning that all EMs were considered individually. The forecast that predicted the maximum intensity was not necessarily the one that predicted the maximum rainfall amount. A list of the forecasts that predicted the shown values can be seen in Appendix F.6. From this it is evident that the very high values seen for Radar for Event 4 is most likely caused by noise, as it is from a forecast prior to the effective rainfall period. The other forecasts all cover parts of the effective rainfall period.

Table 6.2: *Maximum measured/forecasted peak intensities (Peak int.), peak times, and maximum accumulated rainfall (Max acc.) for the 10 min RG data spatially interpolated to the Radar grid and the max of all relevant Radar forecast for the four events.*

	RG Radar Grid			Radar Max FC		
Event	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$
1	49.00	2011-07-02 18:50:00	119.65	58.66	2011-07-02 17:50:00	58.78
2	29.08	2014-08-31 01:50:00	123.60			
3	25.93	2015-09-04 07:00:00	44.13	108.84	2015-09-04 07:50:00	83.03
4	14.33	2016-06-16 06:50:00	57.81	277.37	2016-06-15 17:10:00	166.51

Table 6.3: *Maximum measured/forecasted peak intensities (Peak int.), peak times, and maximum accumulated rainfall (Max acc.) for the 10 min RG data spatially interpolated to the NWP grid and the max of all relevant NWP forecast for the four events.*

	RG NWP Grid			NWP Max FC		
Event	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$
1	49.00	2011-07-02 18:50:00	106.97	34.67	2011-07-02 19:20:00	61.44
2	19.52	2014-08-31 01:40:00	123.60	16.73	2014-08-31 10:40:00	53.32
3	23.00	2015-09-04 07:00:00	36.52	10.10	2015-09-04 07:20:00	22.78
4	14.33	2016-06-16 06:50:00	41.09	10.55	2016-06-16 06:50:00	16.23

Table 6.4: *Maximum measured/forecasted peak intensities (Peak int.), peak times, and maximum accumulated rainfall (Max acc.) for the 60 min RG data spatially interpolated to the Copenhagen Ensemble grid and the max of all relevant Ensemble forecast for the four events.*

	RG Ensemble Grid			Ensemble Max FC		
Event	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$	Peak int. $[\mu m/s]$	Peak time [UTC]	Acc. $[mm]$
1	15.89	2011-07-02 18:00	98.68	26.83	2011-07-02 19:00	129.99
2	7.86	2014-08-31 02:00	101.54	17.09	2014-08-31 23:00	104.77
3	5.09	2015-09-04 07:00	31.62	10.99	2015-09-04 12:00	73.92
4	4.68	2016-06-16 07:00	36.26	4.96	2016-06-16 17:00	50.73

The maximum values support the conclusions drawn from the accumulation plots and animations: NWP underestimates the intensities and rainfall amounts for all 4 events. Radar largely overestimates the peak intensities but underestimates the rainfall amounts, caused by the narrow intensity peaks in the data and the short forecast horizon respectively. For Ensemble only the data for the Copenhagen area was included, however it still overestimates peak intensities, except for Event 4. It also overestimates rainfall amounts, however comparing with the RGs interpolated to the Radar grid the amounts seem more realistic. This indicates that the spatial interpolation of the RGs to the Ensemble grid may magnify this impression due to the smoothing of the data.

On the timing of the peaks of each rain event, it was noted that the agreement is quite high between RG and NWP, while Ensemble is more varying. Radar is clearly affected by the noise in Event 4, but otherwise a good agreement is seen. Event 2 is clearly affected by the multiple peak situation and thus the peak times cannot be used to conclude on temporal agreement between the datasets for this event.

Increasing the distance on the Ensemble data allows to investigate the spatial offsets of the predicted rain events. It was previously noted that higher intensities occurred outside the area of interest. This either indicates that the events were more intense outside the area of interest, or that the forecasts overpredicted the precipitation and with spatial offsets. In the case of the former increasing the considered area might lead to the wrong conclusion about the risk of flooding in Copenhagen. From Table 6.4, it was seen that staying within the grid of the area of interest, already provided prediction of intensities similar or larger than observed by the RG. Expanding the grid increased the maximum intensities to 37.55, 20.55, 21.18, and 7.36 for the four events respectively, wrongfully increasing the risk of flooding in all cases. However, one must also take into account that RGs do have a tendency to underestimate rain, which might be magnified by the spatial interpolation. From the animations and accumulation plots it was seen that the increased area of the Ensemble data provides a clearer impression on the direction and speed of the events. This is also an important factor when forecasting events, as expanding the grid might provide an earlier warning of what is coming. Finally, from thorough investigation of the spatial images in the animations, no clear cases of spatial offset in event peaks were noted, thereby supporting these conclusions.

6.4 Baseline Flood

MIKE FLOOD simulations were run for the four events using the 10 min and 60 min RG data spatially interpolated to the NWP grid as precipitation input. Maps showing the extent of the flooding in the four cases can be found in Appendix H.2, Figures H.1 - H.7. Information on the flood area and volumes for the four events can be found in Appendix H.1. Two locations were noted to often result in flooding: The top left corner at the sea at Svanemøllen and in the southern

corner of Slotsholmen. Both places, the water accumulates and runs out into the sea, which might also be the case in reality.

As noted in the previous sections, Event 1 was the most extreme event and is thus also expected to result in most flooding, which is clearly seen in the figures. Event 3 and 4 do not result in a lot of flooding. For Event 4 the two previously mentioned locations are the only noteworthy flooded areas. Other locations were flooded that day, however the most noteworthy locations were outside the area of the flood model (e.g. Lindevang Station (Rasmussen, 2016)). It is also noteworthy that one of the two measured cloudbursts occurred outside the catchment, in Frederiksberg (Siewertsen, 2016), which might help explain the low amounts of flooding seen.

From the figures in Appendix H.2 less flooding is seen with the higher resolution for all four events. The effects of the temporal resolution on the forecast will be investigated in section 6.6.

6.5 Predicted Flood

Figure 6.9 - 6.13 show the simulated flooding of Central Copenhagen caused by the selected forecasts from the three forecast products for the four events, as well as the baseline scenarios for each event. The figures show the maximum surface water level during the events, thus they do not inform about the timing of the flooding. Only forecasts resulting in visible flooding were included in the figures. The flooded area and volume of water were also calculated for the forecast simulations and can be seen in the overview tables in Appendix H.1. The start of flooding in the baseline scenario was noted in the figure text for comparison with the forecast starts. No Radar simulations were run for Event 4. The Radar forecasts all predict the event poorly, and it was estimated that none of the forecasts would result in flooding. Simulations that did not cover the effective rainfall period are considered incomparable with the other simulations regarding flooding amounts. Thus, only comparison of forecasts within each product is made. The following trends were observed for the four events.

Event 1, Figure 6.9: All simulations resulted in some extent of flooding. The following was noted for the three forecast products:

- **Radar**: All three selected Radar forecasts predict very little flooding. Two out of three start after flooding was observed in the baseline scenario, and due to their short time horizon, none of them cover the entire effective rainfall period of the event. It should be noted that the last Radar forecast simulation at 19:00 is some time after the peak of the observed rain event and the beginning of flooding in the baseline scenario, therefore the observed flooding is, to some extent, affected by the hotstart of the network.
- **NWP**: All three NWP simulations start before flooding was observed in the baseline scenario and cover the peak of the event but not the entire effective rainfall period. They all largely underestimate the flooding compared to the baseline scenario, which can to some extent be explained by the missing tail of the event, however the best performing forecast seem to be the first one: 2011-07-02 11:00.
- **Ensemble**: The last selected forecast from 2011-07-02 at 06:00 resulted in an unstable model, due to too large water amounts, and the simulated flooding was thus not included. All selected forecasts covered the entire effective rainfall period and starts before the baseline flooding. Judging by the flood area, the Ensemble forecast from 2011-07-01 at 06.00 clearly performs the best, which is noteworthy as it is about 1.5 days ahead of the event. The forecast

from 2011-07-01 at 12:00 predicts much less flooding than the others. This could not be foreseen from the areal average and accumulation plots (Figure F.12 and F.28). The best performing Ensemble forecasts predict flooding in similar amounts as the baseline scenario and covers most of the flooded locations.

Event 2, Figure 6.10 and 6.11: All investigated forecasts predicted flooding but with very varying extent. The following was noted for the two forecast products:

- **NWP**: As this is a longer event, poorer performance can be expected from NWP, and the selected forecasts all start after flooding was observed in the baseline scenario. None of the simulated flooding covers all the areas flooded in the baseline scenario. The two NWP forecasts that covered the first peak of the event resulted in some flooding, but as they do not cover the second peak, the forecasted flooding is much less than the baseline scenario. Of the remaining seven selected NWP forecasts, six resulted in visual flooding. The forecast from 2014-08-31 at 08:00 only resulted in very little flooding in the southern corner of Slotsholmen and is therefore not shown. As they all cover the second peak, hotstart was used for these simulations, however as the hotstart does not include the surface water, the flooding caused by the first major peak is not included. Thus all six result in less flooding than the baseline scenario. The NWP forecast from 2014-08-31 at 06:00 and at 09:00 performs best at replicating the baseline flooding.
- **Ensemble**: The results of the eight Ensemble forecasts are similar, however, the last three of the selected forecasts simulate the most flooding and are thus closer to the baseline scenario. All forecast start before flooding was observed in the baseline scenario, except the last one which is only half an hour after. All forecast covers the two major peaks of the event, except for the first forecast, which ends before the second peak. It seems that while the NWP replicates the flooding in the northern part of the catchment better, Ensemble replicates the flooding in the southern end better. The flooding by the sea is overestimated by the Ensemble forecasts while they fail to replicate the smaller flooded areas on land, which could be due to the coarser temporal resolution.

Event 3, Figure 6.12: All investigated forecasts predicted flooding except one Radar forecast. The following was noted for the three forecast products:

- **Radar**: All radar forecast start around the time of flooding in the baseline scenario. The forecast from 06:00 did not predict any flooding and is therefore not shown. As the forecast at 7:00 also showed poor performance it was decided to include the forecasts in-between the two. Only one Radar forecasts predicts flooding similar to the baseline scenario, forecast 2015-09-04 at 06:20, which is also the only forecast that estimates an increase in precipitation (Figure F.17).
- **NWP**: The investigated NWP forecast starts before the baseline flooding and covers the effective rainfall period. However, it fails to replicate most of the flooding and only simulates some flooding on Slotsholmen. As this was the forecast with the highest accumulated rain, no flooding is expected from the other NWP forecasts.
- **Ensemble**: The two Ensemble forecasts covers the entire effective rainfall period and starts before the baseline flooding. They both overestimate the flooding, especially in the southern part of the catchment, where the baseline scenario did not predict flooding. However, the forecasts predict all significant flooding locations.

Event 4, Figure 6.13: All investigated forecasts predicted flooding except one NWP forecast. The following was noted for the two investigated forecast products:

- **NWP**: As this event was long but not very intense, poorer performance is expected from the NWP forecasts as none of them cover the entire effective rainfall period. The first NWP forecast predicts a high intensity rainfall peak prior to the actual event, however not enough to cause flooding. The second included forecast does not cover the build-up of the event, nor the start of baseline flooding, but it did predict the highest intensity of all forecasts, and this showed to be enough to cause flooding. However, it only catches the northern part of the flooding, but quite similar to the baseline scenario.
- **Ensemble**: The Ensemble forecasts catch the flooding at sea but none on land. The two middle forecast out of the four selected, predict closest to the baseline scenario. In general, the simulations for this event are more difficult to evaluate based on flooding, due to the low amounts.

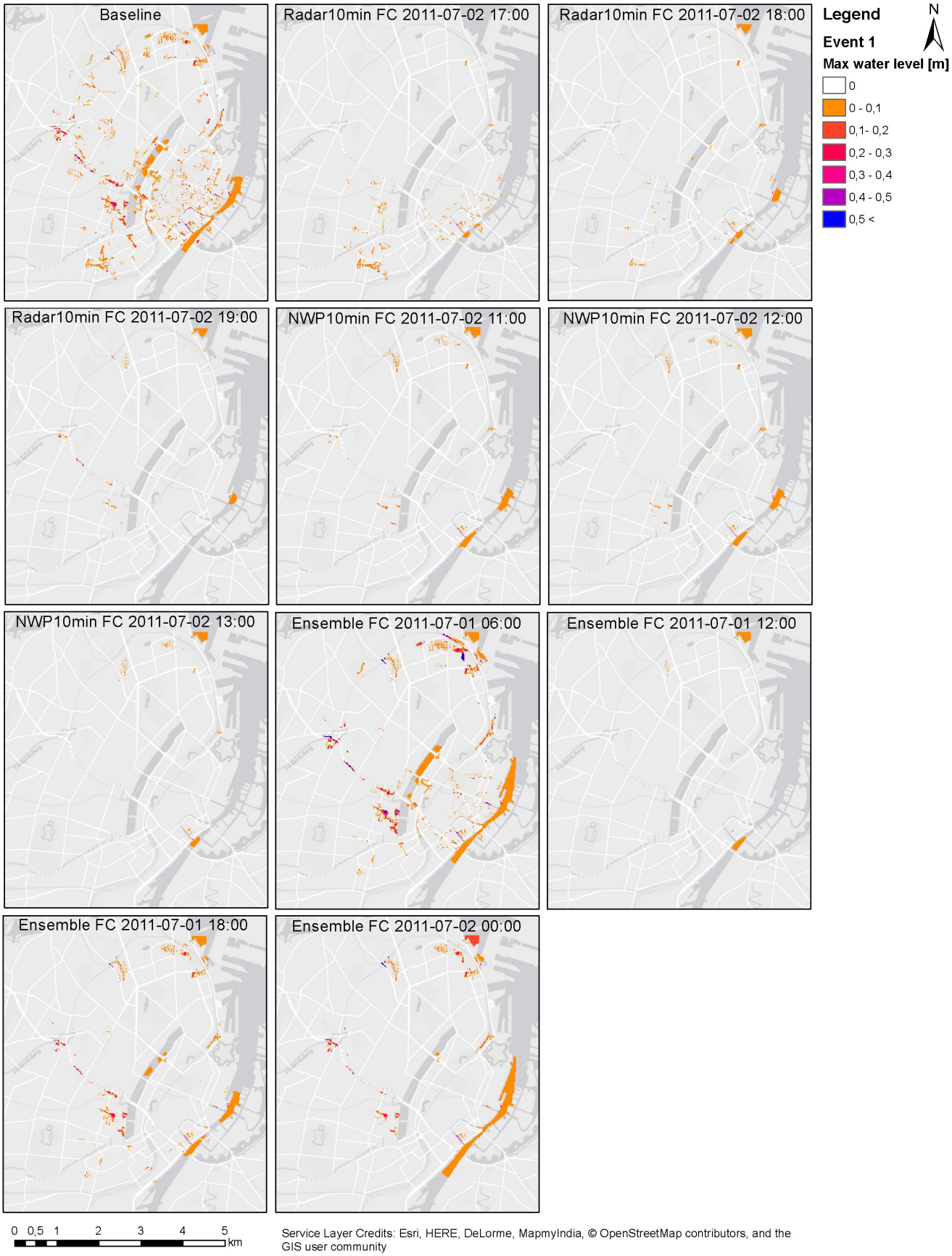


Figure 6.9: *Simulated flooding showed as maximum water level [m] simulated using 3 NWP forecast, 3 Radar forecast and 4 Ensemble forecast for Event 1, as well as the baseline scenario (top, left). Baseline flooding started at around 17:30 on 2011-07-02.*

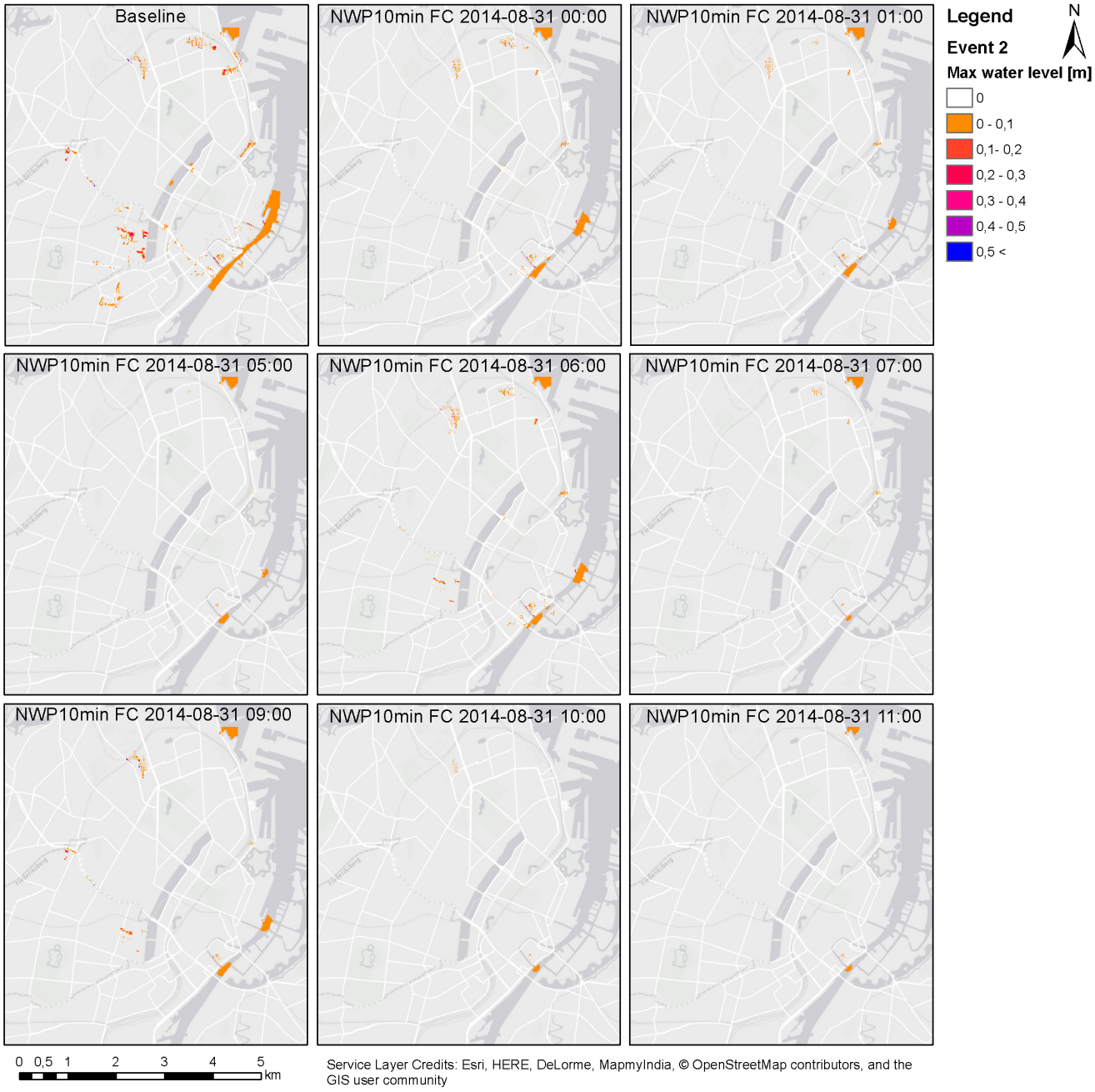


Figure 6.10: *Simulated flooding showed as maximum water level [m] simulated using 8 NWP forecast for Event 2, as well as the baseline scenario (top, left). Baseline flooding started at around 23:30 on 2014-08-30.*

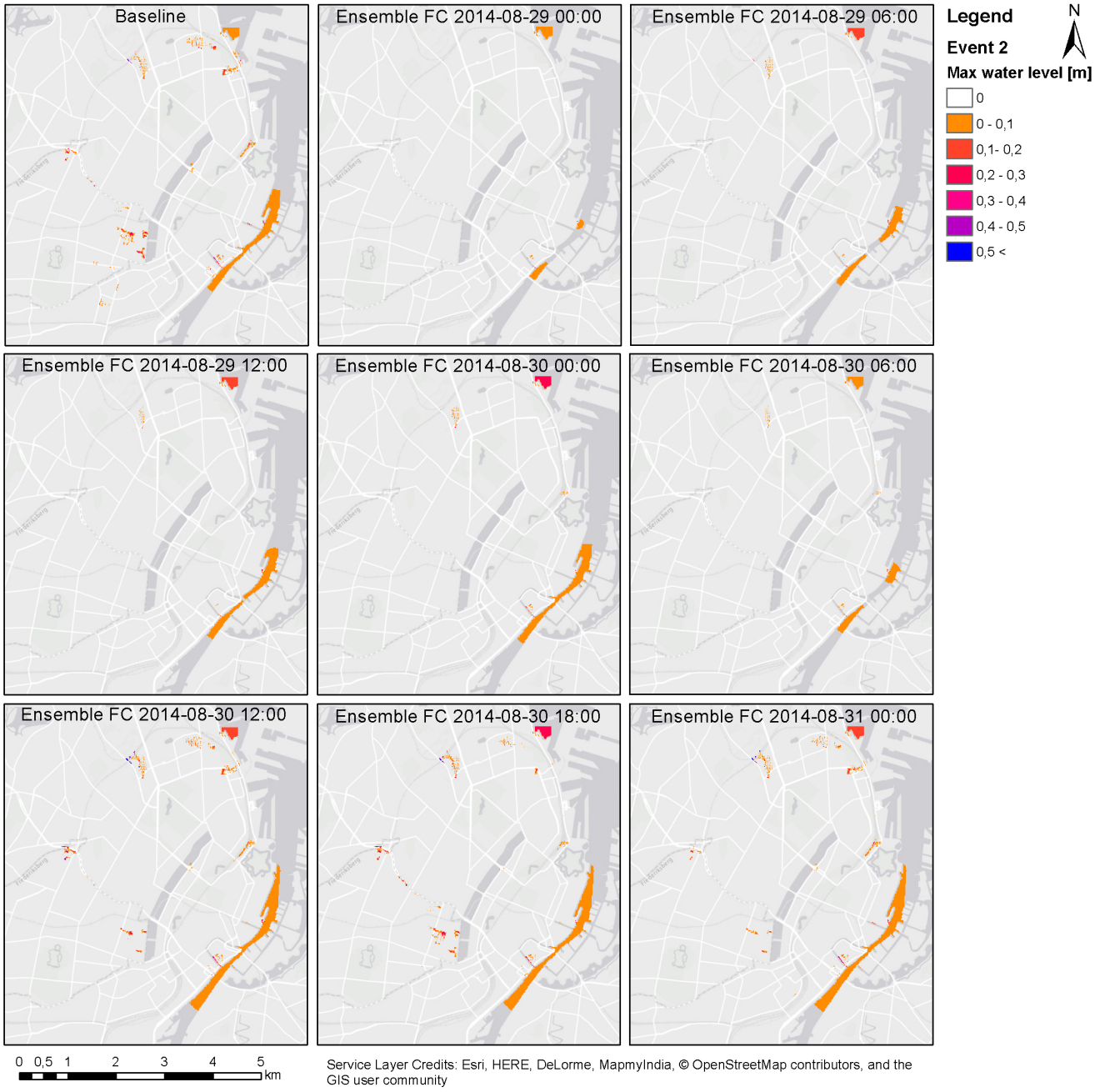


Figure 6.11: *Simulated flooding showed as maximum water level [m] simulated using 8 Ensemble forecast for Event 2, as well as the baseline scenario (top, left). Baseline flooding started at around 23:30 on 2014-08-30.*

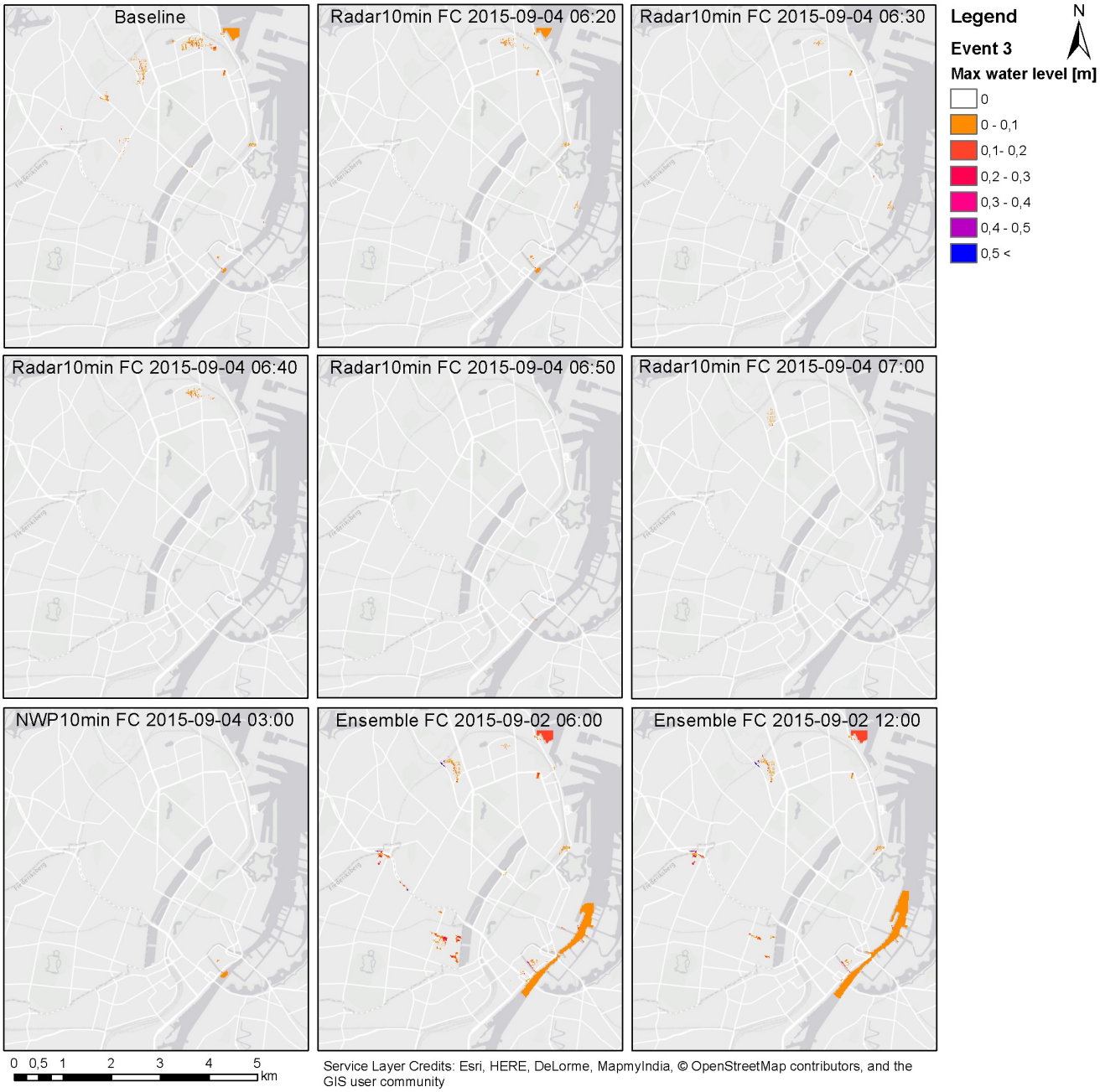


Figure 6.12: *Simulated flooding showed as maximum water level [m] simulated using 1 NWP forecast, 5 Radar forecast and 2 Ensemble F for Event 3, as well as the baseline scenario (top, left). Baseline flooding started at around 06:50 on 2015-09-04.*

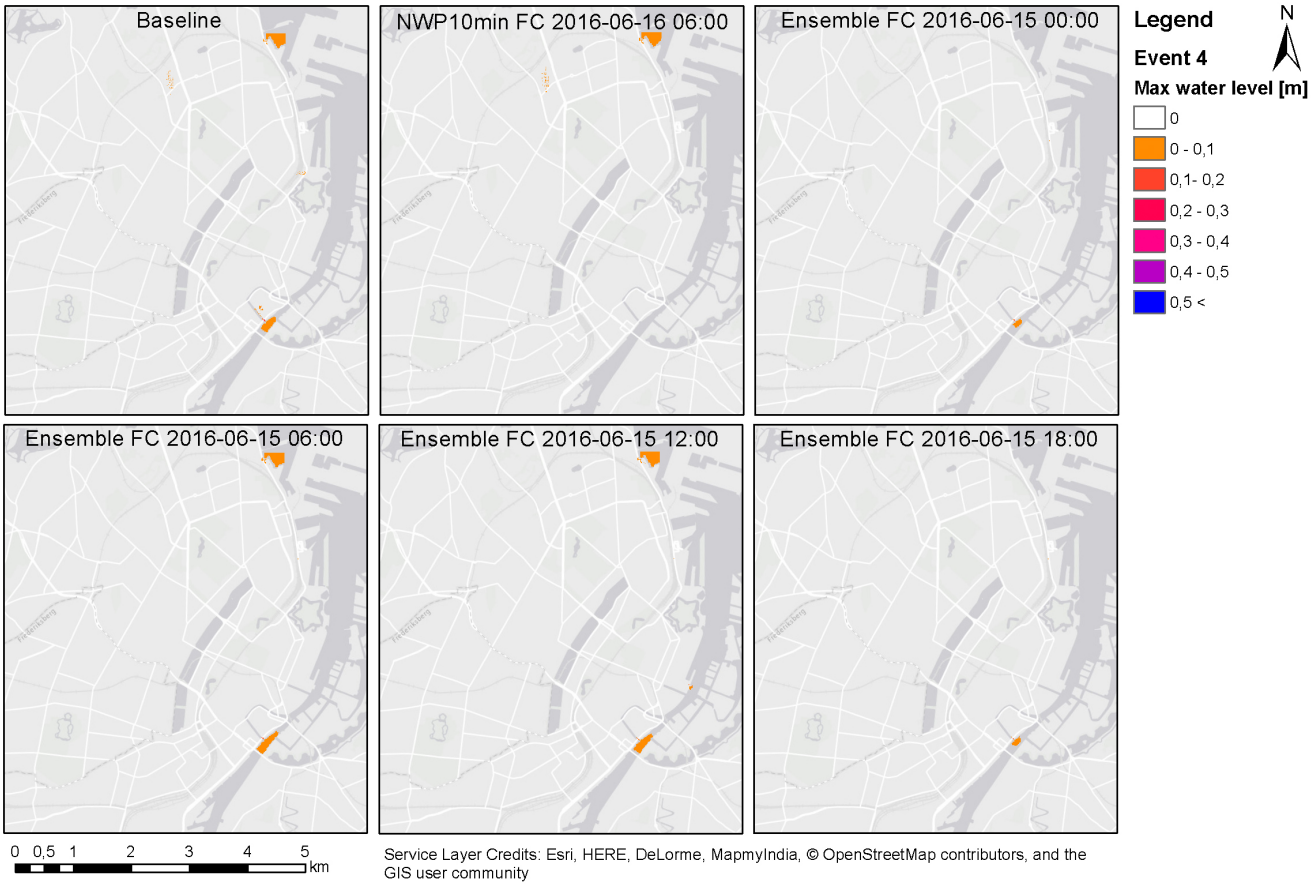


Figure 6.13: *Simulated flooding showed as maximum water level [m] simulated using 1 NWP forecast and 4 Ensemble forecast for Event 4, as well as the baseline scenario (top, left). Baseline flooding started at around 03:50 on 2016-06-16.*

For all four events good performance is seen for the Ensemble predictions, although with a tendency to overestimate, possibly caused by the maximum approach. As noted for the precipitation forecasts, the Ensemble does in some cases predict the events better before they get too close to present time, which might be related to the longer spin-up time of the Ensemble product. This trend is not as clear for the flood investigations, as forecasts closer to events were not selected for flood simulation, based on their poorer performance. However for Event 1, 2 and 4 the best performing simulation is still not the last of the ones investigated. Finally it should be noted, that the good performance might be linked to the overestimation of rain amounts and intensities which results in more visible flooding in more locations.

The overall impression taken from the flood maps is that the flood forecast is more dependent on the amounts of water fallen on the catchment, than on the spatial differences in the forecasts. This makes sense as the drainage system is connected and water from a larger area contributes to flooding in Central Copenhagen. However some differences in flooding locations were seen between the different forecast products. This also shows the drawbacks of the applied method: the simulated amounts are not the same when the simulated period is not the same. To proper evaluate the performance of the shorter NWP and Radar forecasts an alternative approach is needed.

6.6 Effects of Temporal Resolutions

As noted, Ensemble performs better than NWP for some events, however if Ensemble is used one must compromise on the temporal (and spatial) resolution. The effects of a coarser resolution was already commented on in section 6.4. A simple investigation was thus conducted on the temporal resolution. Simulations were run for the selected NWP forecasts for Event 1 with a resolution of 60 min, and the flood maps can be seen in Appendix H.3. A summary of the resulting floods can be seen in Table 6.5, for the four baseline scenarios and the three tested forecasts.

Table 6.5: *Overview of the flood simulations for the four baseline scenarios and the three investigated NWP forecasts for Event 1, with the flood start, maximum observed water depth, flooded area and flood volume for the two investigated temporal resolutions (10 min and 60 min) and the percent deviation between the two resolutions.*

	10 min				60 min				Percent Deviation		
	Flood Start [UTC]	Depth [m]	Area $[m^2]$	Volume $[m^3]$	Flood Start [UTC]	Depth [m]	Area $[m^2]$	Volume $[m^3]$	Depth [m]	Area $[m^2]$	Volume $[m^3]$
RG Event 1	17:30	1.57	2,372,784	97,318	17:20	1.47	1,143,632	56,047	6	52	42
NWP 11:00	15:20	1.25	356,112	13,218	15:40	1.18	306,544	10,715	6	14	19
NWP 12:00	16:20	1.03	357,440	13,015	16:10	1.03	286,608	11,546	0	20	11
NWP 13:00	16:10	0.54	182,576	4,840	16:10	0.32	121,584	3,788	41	33	22
RG Event 2	23:30	1.35	927,984	41,085	23:20	1.26	746,336	34,555	7	20	16
RG Event 3	06:50	1.03	252,496	6,639	06:30	0.76	67,424	1,522	26	73	77
RG Event 4	03:50	0.26	122,848	3,970	03:40	0.23	108,624	3,986	12	12	0

From Table 6.5 and focusing on Event 1 it is seen that in general, the forecasts are off on the timing of the flooding breakthrough. The tendency for generally lower values is clear both for observations and forecasts, however to very varying extents. The tendency for earlier breakthroughs is not as clear for the three forecasts as for the observed events. These two tendencies can be expected when decreasing the temporal resolution, as the averaging flattens the peaks. The offset on the timing by the forecasts compared to the RG scenario for both resolutions is much larger than the differences caused by the change in temporal resolutions, possibly related to the quality of the forecasts. This indicates that the timing of the forecasts is not significantly affected by the change in resolution. From the baseline scenarios of all four events, it would seem that there is a difference in effect depending on the length and intensity of the event: The two longer events have much lower deviation between the two resolutions, and the less intense event (Event 4) shows the least effect from changing the resolution.

Investigating three forecasts of one event is not enough to quantitatively describe the effects of the chosen temporal resolution. However, it does give an indication on which effects can be expected from decreasing the temporal resolution of the precipitation input. From these investigations it is clear, that a small change in flood breakthrough can be expected, however more importantly a large underestimation of flood extent might occur.

6.7 Discussion

As concluded previously, the Radar data has a tendency to underestimate the rain and could benefit from a correction factor. As the data was used for flood simulations in this part of the study, it was decided to correct for underestimation using a simple procedure. Thus a static, event based correction factor was applied, although many alternative methods exist. Borup et al. (2016) found that dynamic adjustments of radar data based on an aggregation period of the past 10-20 min of RG data resulted in the best performance when used to model combined sewer overflow compared to

static adjustments. However, as they also note, adjustment of the past 10-20 min are not relevant in relation to nowcasting as it does not support a long time horizon, and thus a static correction might still be the best solution for forecasts. A need for a more complex adjustment in this study is however supported by the conducted event analysis. It showed that the peak intensity area forecasted by the Radar was much more narrow than what was seen for the other precipitation products. This would not be improved by the chosen correction method, but could be countered by merging with other precipitation products.

Besides the Radar data's tendency to underestimate, it was noted from the event analysis that the Radar extrapolation almost exclusively yielded rain intensities that decreased over lead time. This trend expresses a problem with the extrapolation model, which might be overestimating the movements of the events, or simply has difficulties simulating build-up of events, as it extrapolates from the current conditions. This makes it difficult to use for flood forecasting, as it performs poorer during event build-up, where the prediction of flooding is most relevant. The short forecast horizon also does not enhance the performance of Radar in the conducted analysis, and an alternative simulation approach is needed to account for this. However, in real time application the short forecast horizon is countered by the frequent forecast generation.

Modelling involves many uncertainties, of which one type relates to the uncertainties of the input data. As mentioned in the discussion of Part I, one of the major assumptions in this study is the use of RG data as reference. For the event analysis, comparisons were done between Radar observations and the RG to investigate similarities. As they were found quite similar in some cases, and considering the underestimation of the Radar data, it was decided to use the RG data to simulate baseline scenarios for the flood investigations.

The results of the 2D model simulations are also uncertain on its own due to simplifications in model setup and the chosen model parameters. It is therefore difficult to assess how realistic the modelled flooding is. One way to validate the model results is to compare with photographs of the flooding at selected location, as was done in Brødbæk et al. (2015) for the event in 2011, where simulation results based on RG input were found to match photographed flooding well. In this study, however, the baseline scenarios are mainly used to assess the performance of the three forecast products, and as the same model setup is used in all simulations along with the same baseline scenario, the comparisons are relative to the model.

One major issue when investigating extreme precipitation events and consequent floods is their rare occurrence. Obtaining data for enough events to be able to conduct quantitative analyses and conclude on significant trends in behaviour and performance for the forecast products is a difficult task. A good example is Liguori et al. (2012), who investigated the performance of ensemble and deterministic forecasts for three different events coupled to a flow model for a small urban catchment. They did identify some trends in the performance, however they stress that three events are only enough to support potential applications of the forecast products for flood warnings in the catchment, not enough to define an improvement of one forecast type over the other. The same conclusion applies in this study. The conducted event analysis is a qualitative analysis focusing on differences between the forecast products, their performance and possible effects of the chosen temporal resolution. It should also be noted that for the four events, 46 different forecasts have been investigated with flood simulations, thereby extracting as much information as possible, to identify trends in the performance of the forecast products.

Finally, many assumptions and simplifications where made in the conducted flood forecast analysis. This can therefore be considered the main point in need of improvement. The used approach has its shortcomings as it favours Ensemble in several ways: It is often the only forecast covering the

entire event period, and the use of the maximum of the ensemble members further increases the predicted flooding. Alternative simulation approaches using shorter periods could provide better results for NWP and Radar, and is necessary to conduct a direct and more quantitative comparison of the three forecast products.

6.8 Further Analysis

Flood simulations with MIKE FLOOD provides much more information than what was assessed here. In this study the maximum flooded area and volume was used to asses the forecasts, however the use of indicators, like RMSE, to quantify the differences between the baseline scenarios and the forecasts on other parameters obtained from the simulation would be a good alternative to the visualization conducted here. As an example, another important aspect is the timing of the forecasted flooding, which could also be investigated from the simulations.

From the perspective of SURFF, the aim is to increase its ability to predict flooding, not necessarily in the correct amounts and extents, but more importantly in the prediction of the risk of flooding in a certain area. This is also why a qualitative assessment of flood maps was chosen in this study. However, to quantify the performance of different forecasts a threshold approach, similar to the one conducted in Part I, could be applied on the flood maps. For example, this could be implemented by separating the model catchment into areas of interest (e.g. based on the Copenhagen Cloudburst Branches, (The City of Copenhagen, 2015)), and doing contingency tables for each forecast product, by counting the number of times a forecast predicts flooding in the selected areas.

An important aspect of the conducted analyses in this study is the differences in both spatial and temporal resolution of the forecasts. This was investigated to some extent both in the general performance analysis and the event analysis. A thorough approach using different combinations of temporal and spatial resolutions could be applied for selected events, similar to what was conducted by Ochoa-Rodriguez et al. (2015). They used 15 different resolution combinations and found amongst other things that variations in temporal resolution had a higher impact than variations in spatial resolution, and that there was a strong link between the two.

6.9 Conclusion

The second part of this study aimed at investigating the performance of the three forecast products for four selected events and their resulting flood simulations: a Radar Nowcast (Radar), a deterministic Numerical Weather Prediction with radar data assimilation (NWP) and an Ensemble Numerical Weather Prediction system (Ensemble) for the events 1) 2nd of July 2011, 2) 30th-31st of August 2014, 3) 4th of September 2015 and 4) 15-16th of June 2016. The behaviour of the measured and forecasted precipitation was analysed and the data was used as input for flood simulations to provide an impression of the products performance in flood prediction.

The following important conclusions were found from the event precipitation forecast analysis:

1. RGs and Radar agree on temporal distribution, and for Event 3 also spatial distribution, of the four events, however not on rainfall amounts and intensities. A simple correction of the radar data does not improve these features satisfyingly, as it results in overprediction of both maximum observed intensity and rainfall amount.

2. Visual investigation of the Radar forecasts showed that the high intensity areas are more narrow and not less intense than the other precipitation products, meaning that with or

without correction some grid cells might overestimate rain amounts and intensities. However correction is still needed on average, and a more complex correction method could be more appropriate. The investigations also showed that Radar has difficulties during the build-up of events, which makes it less useful for flood prediction, as this is an important period.

3. For all four events a tendency to underestimate the precipitation intensities and amounts was seen for NWP, while Ensemble always overestimated with varying extent, possibly caused by the use of the maximum of the ensemble members in the analysis. However, in the best performing forecasts for Event 1, 2 and 4, NWP was better at predicting the right spatial distribution of the events. The event investigation clearly documented a long spin-up time for Ensemble, as the best performance was always seen for forecasts no to close to the actual event.

4. An interesting observation was that earlier Ensemble forecasts seemed to perform better than later ones. This indicates, that considering the Ensemble forecasts 1-2 days ahead might be relevant in flood forecasting.

The following important conclusions were found from the event flood forecast analysis:

1. From the flood simulations it was seen that in general Ensemble performed well in relation to flood extent and amounts. However it should be noted, that applied method of simulating the entire events, favours Ensemble due to its longer forecast horizon. The approach of using the maximum of the EMs increases the forecasted flooding. Indeed, overestimation was seen for some forecasts.

2. The applied simulation method has its limitations in analysing the performance of Radar and NWP, as events are often longer than the forecast horizon of these products. To get around this problem additional simulations should be conducted using shorter periods.

3. Considering spatial displacement of rainfall forecasts by investigating a larger area for the Ensemble data did not result in any clear benefits in forecasting the events other than a better visual impression of the development and movement of the events.

4. Decreasing the temporal resolution of the rainfall inputs decreases the simulated flooding, however with very varying differences depending on the event. This might to some extent be counteracted by the overestimation seen for Ensemble, and further analysis is needed to completely exclude drawbacks from using a lower temporal resolution.

All in all similar trends were found for the precipitation forecasts and the resulting flood simulations. This means that even though the link from rainfall to runoff is complex, conclusions on their relevance for flood prediction can still be drawn from the behaviour of the precipitation forecast. On the other hand the flood simulations also showed that even though there are clear spatial differences in the forecasted precipitation, similar flooding patterns can be predicted. Ensemble provided the best flood predictions and no predominant losses were identified from the decrease in temporal resolution. However, one must note, that the methods applied in this study in several ways favours Ensemble, and further analysis is needed to support the observed trends. Finally, the use of Ensemble, as investigated here, might also lead to significant false alarm rates, which cannot be assessed from an event analysis, but could be investigated using similar approaches as in Part I.

CHAPTER 7

Overall Conclusions and Final Remarks

This study has investigated the potential of three different precipitation forecast products for use in flood prediction: a Radar Nowcast product (Radar), a Numerical Weather Prediction with data assimilation (NWP) and an Ensemble Numerical Weather Prediction (Ensemble), by comparing against rain gauge measurements (RG). The study builds on the OMOVAST project, and thus the study area was Greater Copenhagen. The study was separated in two parts. In Part I, the performance of the three products for this study area was compared on different spatial and temporal scales in a general performance analysis considering data from a six month long period. In Part II, the potential for using the precipitation products in flood predictions was assessed by investigating their performance in predicting four high intensity events and by simple evaluation of flood simulation results for Central Copenhagen with the forecast products as precipitation input. The selected events were: 1) 2nd of July 2011, 2) 30th-31st of August 2014, 3) 4th of September 2015 and 4) 15-16th of June 2016. From these two parts an impression of the performance of the forecast product in predicting rainfall and as input in flood prediction has been obtained. Building on the individual conclusions in section 4.6 and 6.9 some overall observations will be highlighted in the following:

1. Potential was seen for the Radar product, especially in the first lead time hour, however difficulties of the extrapolation method to simulate weather development was evident. Combined with the short lead time, the current form of this product becomes less desirable for use in flood prediction, at least on its own. On the other hand, the high spatial and temporal resolutions are advantages of this product, which, however, did not provide major improvements to the flood simulations conducted in this study.

2. Underestimation of the precipitation by the Radar product was clearly documented in the first part of the study. However a simple correction method on event basis, proved not to have the wanted effects, due to a low spatial extent of peak intensities. A more complex correction method is suggested, however due to the forecast difficulties seen for of the product, in general, further development of the product is suggested.

3. Part I concluded that the performance of NWP and Ensemble were similar, however Part II showed that in extreme events NWP has a tendency to underestimate the rainfall amounts and intensities, while Ensemble has a tendency to overestimate, although influenced by the fact that the maximum of the ensemble members was applied in the analysis. This did however

highlight the importance of focusing on performance during extreme events, as it might be different from everyday performance.

4. As observed in Part I, it is relevant to consider the first lead time hour of Radar in flood prediction especially in light of possible benefits from its high resolution. Considering Ensemble from around hour 3-4 could improve the prediction, and also allows extending the forecast horizon much further into the future. From Part II Ensemble proved to be superior, however favoured by the applied approach. Simple investigation of the effect of the lower temporal resolution highlighted a possible reduction of flood extent from upscaling, which was to some extent counteracted by the tendency to overestimate for Ensemble. Thus it seems relevant to consider Ensemble for flood forecasting especially if longer forecast horizons are wanted. However, as an event analysis can only shed light on the hit rate, further analysis is needed to ensure that the applied approach does not result in significant false alarm rates, and thus worse prediction skill. Besides this, further analyses of the effects of decreasing the temporal and spatial resolution is also recommended prior to any upscaling, as well as quantitative comparisons with the performance of Radar and NWP.

From the above observations it is clear that considering all three forecast products in combination seems to be the optimal solution. This is of course also the most comprehensive solution. Radar observations are already assimilated into the NWP model (Korsholm et al., 2015) and the assimilation of Radar Nowcast into the model is also possible, and has already been tested with good results (Jensen et al., 2015). The UK Met office has, in collaboration with the Australian Bureau of Meteorology, developed a forecasting scheme combining an extrapolation radar forecast with a high-resolution NWP rainfall forecast, which is referred to as STEPS (Liguori et al., 2012). STEPS combines the two datasets and produces both a deterministic forecast and ensemble forecasts. Inspiration on combining different forecasts products could be drawn from this setup and its uses.

All three investigated precipitation forecast products showed potential as inputs for flood forecasting. This study indicated that flood prediction could benefit in several ways from the inclusion of Ensemble as input, however further investigations are needed. It also appears from the results of the study that there is good potential in considering more than one precipitation forecast in flood predictions.

References

Andersen, M. M. (2011). Skybrud over København - 3. udgave. http://www.dmi.dk/nyheder/arkiv/nyheder-2011/07/skybrud-over-koebenhavn-anden-udgave/ [accessed: 2016-07-14].

Atger, F. (2001). Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics*, 8(6):401–417.

Barros, V., Field, C., Dokken, D., Mastrandrea, M., Mach, K., Bilir, T., Chatterjee, M., Ebi, K., Estrada, Y., Genova, R., Girma, B., Kissel, E., Levy, A., MacCracken, S., Mastrandrea, P., and White, L., editors (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge United Kingdom and New York, NY, USA, 1 edition.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling and Software*, 40:1–20.

Beredskabsstyrelsen (2012). Redegørelse vedrørende skybruddet i Storkøbenhavn lørdag den 2 . juli 2011. Technical report, Beredskabsstyrelsen - Institut for Beredskabsevaluering, København.

Borup, M., Grum, M., Linde, J. J., and Mikkelsen, P. S. (2016). Dynamic gauge adjustment of high-resolution X-band radar data for convective rain storms: model-based evaluation against measured combined sewer overflow. *Journal of Hydrology*, 539:687–699.

Brødbæk, D., Courdent, V., Löwe, R., Meneses, E. J., and Petersen, S. O. (2015). OMOVAST - Operativ Model til Varsling og Styring. Technical report, Miljøministeriet, Naturstyrelsen, København.

Cappelen, J. and Scharling, M. (2010). Mere - og mere intens - regn over Danmark. http://www.dmi.dk/nyheder/arkiv/nyheder-2010/mere-og-mere-intens-regn-over-danmark/ [accessed: 2016-07-14].

DHI (2014a). Mike 21 Flow Model - Hydrodynamic Module. Technical report, DHI, Hørsholm.

DHI (2014b). MIKE Flood - User Manual. Technical report, DHI, Hørsholm.

DHI (2014c). MIKE URBAN Collection System User Guide. Technical report, DHI, Hørsholm.

DHI (2014d). MOUSE - Pipe Flow Reference Manual. Technical report, DHI, Hørsholm.

Elkjær, K. (2014). DMI overså skybrud: Derfor udsendte vi ikke varsel. http://www.bt.dk/vejret/dmi-oversaa-skybrud-derfor-udsendte-vi-ikke-varsel [accessed: 2016-07-14].

Feddersen, H. (2009). A Short-Range Limited Area Ensemble Prediction System. Technical report, DMI, Copenhagen.

Hansen, N. and Pedersen, P. S. (2014). Voldsomt skybrud over København. http://www.dmi.dk/nyheder/arkiv/nyheder-2014/08a/voldsomt-skybrud-over-koebenhavn/ [accessed: 2016-05-17].

Høgsholt, D. (2015). Se de vilde billeder: Uvejr med skybrud og hagl ramte København. http://vejr.tv2.dk/2015-09-04-se-de-vilde-billeder-uvejr-med-skybrud-og-hagl-ramte-koebenhavn [accessed: 2016-07-14].

Jensen, D. G. (2015). *Combining weather radar nowcasts and numerical weather prediction models to estimate short-term quantitative precipitation and uncertainty*. Ph.d. thesis, Aalborg University, Denmark.

Jensen, D. G., Petersen, C., and Rasmussen, M. R. (2015). Assimilation of radar-based nowcast into a HIRLAM NWP model. *Meteorological Applications*, 494(November 2014):485–494.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast Verification - A practitioner's Guide in Atmospheric Science*. John Wiley & Sons, Ltd., Chichester, UK, 2nd edition.

Korsholm, U. S., Petersen, C., Sass, B. H., Nielsen, N. W., Jensen, D. G., Olsen, B. T., Gill, R., and Vedel, H. (2015). A new approach for assimilation of 2D radar precipitation in a high-resolution NWP model. *Meteorological Applications*, 22(1):48–59.

Krüger (2015). Operativ Model til Varsling og Styring.

Liguori, S., Rico-Ramirez, M. a., Schellart, a. N. a., and Saul, a. J. (2012). Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. *Atmospheric Research*, 103:80–95.

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290.

Nielsen, J. E., Thorndahl, S., and Rasmussen, M. R. (2014). A numerical method to generate high temporal resolution precipitation time series by combining weather radar measurements with a nowcast model. *Atmospheric Research*, 138:1–12.

Ochoa-Rodriguez, S., Wang, L.-P., Gires, A., Pina, R. D., Reinoso-Rondinel, R., Bruni, G., Ichiba, A., Gaitan, S., Cristiano, E., van Assel, J., Kroll, S., Murlà-Tuyls, D., Tisserand, B., Schertzer, D., Tchiguirinskaia, I., Onof, C., Willems, P., and ten Veldhuis, M.-C. (2015). Impact of spatial and temporal resolution of rainfall inputs on urban hydrodynamic modelling outputs: A multi-catchment investigation. *Journal of Hydrology*, 531:389–407.

Olsen, B. T., Korsholm, U. S., Petersen, C., Nielsen, N. W., Sass, B. H., and Vedel, H. (2015). On the performance of the new NWP nowcasting system at the Danish Meteorological Institute during a heavy rain period. *Meteorology and Atmospheric Physics*, 127(5):519–535.

Rasmussen, C. (2016). BILLEDSERIE Gummibåde og druknede biler i hovedstaden. http://www.dr.dk/nyheder/vejret/billedserie-gummibaade-og-druknede-biler-i-hovedstaden [accessed: 2016-07-14].

Schellart, A., Liguori, S., Krämer, S., Saul, A., and Rico-Ramirez, M. (2012). Analysis of different quantitative precipitation forecast methods for runoff and flow prediction in a small urban area. *IAHS-AISH Publication*, 351(August 2014):614–619.

Schilling, W. (1991). Rainfall data for urban hydrology: what do we need? *Atmospheric Research*, 27:5–21.

Schilling, W. and Fuchs, L. (1986). Errors in stormwater modeling — a quantitative assessment. *Journal of Hydraulic Engineering*, 112(2):111–123.

Siewertsen, B. (2015). Hold fast et regnvejr. http://www.dmi.dk/nyheder/arkiv/nyheder-2015/09/hold-fast-et-regnvejr/ [accessed: 2016-07-14].

Siewertsen, B. (2016). Stadig masser af regn torsdag. http://www.dmi.dk/nyheder/arkiv/nyheder-2016/juni/stadig-masser-af-regn-torsdag/ [accessed: 2016-07-14].

The City of Copenhagen (2015). Climate Change Adaptation and Investment Statement - Part 2. Technical Report october, The City of Copenhagen, Copenhagen.

Thomsen, R. S. (2011). Teknisk rapport 11-03 Drift af Spildevandskomitéens Regnmålersystem Årsnotat 2010. Technical report, DMI, København.

Thomsen, R. S. (2012). Teknisk rapport 12-03 Drift af Spildevandskomitéens Regnmålersystem Årsnotat 2011. Technical report, DMI, København.

Thomsen, R. S. (2015). Teknisk rapport 15-03 Drift af Spildevandskomitéens Regnmålersystem Årsnotat 2014. Technical report, DMI, København.

Thomsen, R. S. (2016). DMI Report 16-03 Drift af Spildevandskomitéens Regnmålersystem Årsnotat 2015. Technical report, DMI, København.

Thorndahl, S., Bovith, T., Rasmussen, M. R., and Gill, R. S. (2012). On comparing NWP and radar nowcast models for forecasting of urban runoff. In Moore, R., Cole, S., and Illingworth, A., editors, *Weather Radar and Hydrology*, pages 620–625. IAHS Press, Exeter.

Thorndahl, S., Nielsen, J. E., and Rasmussen, M. R. (2014). Bias adjustment and advection interpolation of long-term high resolution radar rainfall series. *Journal of Hydrology*, 508:214–226.

Unden, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J., Cats, G., Cuxart, J., Eerola, K., Fortelius, C., Garcia-Moya, J. A., Jones, C., Lenderlink, G., McDonald, A., McGrath, R., Navascues, B., Woetman Nielsen, N., Odegaard, V., Rodriguez, E., Rummukainen, M., Room, R., Sattler, K., Sass, B. H., Savijarvi, O., Wichers Schreur, B., Sigg, R., The, H., and Tijm, A. (2002). HIRLAM-5 Scientific Documentation. *Environmental Geology*, 43(1):144.

Vedel, H. (2016). Personal communication, DMI.

Vejen, F. (2011). Tropisk styrtregn over København den 2. juli 2011. *Vejret (Dansk Meteorologisk Selskab)*, 3(128):1–11.

Villarini, G., Smith, J. A., Lynn Baeck, M., Sturdevant-Rees, P., and Krajewski, W. F. (2010). Radar analyses of extreme rainfall and flooding in urban drainage basins. *Journal of Hydrology*, 381(3-4):266–286.