NOWCASTING OF EXTREME RAINFALL IN DUTCH CITIES

Guo-Shiuan Lin

Thesis

submitted in partial fulfilment of the requirements for the

Master's degree of Applied Earth Science

at Delft University of Technology

Track: Environmental Engineering

Specialization: Environmental Sciences

Thesis committee

Prof. Remko Uijlenhoet	Delft University of Technology
Ruben Imhoff	Deltares
Dr. Marc Schleiss	Delft University of Technology

Abstract

Extreme rainfall brings substantial threats to lives, infrastructure, and the economy in cities. Radar rainfall nowcasting was proven able to provide forecasts up to 2 to 3 hours in advance on a catchment scale. However, an extensive evaluation of nowcasting skills for urban areas has not been performed yet. In this study, we selected 80 extreme events that occurred in 5 main Dutch cities (Amsterdam, The Hague, Groningen, Maastricht, and Eindhoven) from 2008 to 2021. We evaluated the performance of probabilistic nowcasts with 20 ensemble members applying short-term ensemble prediction system (STEPS) from Pysteps for these cities, focusing on analyzing the dependence on rainfall characteristics and city sizes. Nowcasts in Eindhoven (96 km^2) and Maastricht (67 km^2) had higher errors because the rainfall intensity of their events was higher. Besides, nowcasts at small areas showed higher error, especially when the size was below 100 km^2. We found that forecast errors were higher and the forecast was less reliable for the 1-h event durations than for 24-h durations. Despite these differences, skillful lead times measured by Pearson correlation in all the cities were about 20 to 24 minutes for both the 1-hour and 24-hour events. CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting) adjusted the bias in real-time QPE and QPF, but QPF still reduced with increasing lead time. Also, CARROTS did not adjust the rainfall spatial distribution much, so the skillful lead time did not change much. The skillful lead time in this study was shorter than the counterparts on the catchment scale because small areas are more sensitive to the displacement of forecast rainfall. Still, such lead time is similar to the findings in other research on short convective rainfall over small areas. Future research could try to apply machine learning, 3-dimensional nowcasting, or blending numerical weather prediction in the nowcasting process to better forecast the growth and decay of rainfall at a longer lead time.

Keywords: nowcasting, extreme precipitation, urban flooding

Table of Contents

1. Introduction

Growing evidence shows that extreme precipitation events with stronger intensity and higher frequency are expected to occur along with global warming (Min, et al., 2011; Myhre, et al., 2019; Shongwe, et al., 2011; Donat, et al., 2016; Sun & Ao, 2013). IPCC already observed that there is a significant increase of annual maximum daily precipitation in Europe and some regions in the world since 1950s (Masson-Delmotte, et al., 2021). Furthermore, extreme rainfall is expected to increase flooding in Europe and around the globe (Tabari, 2020; Madsen, et al., 2014; Ralph, et al., 2014; Field, et al., 2012). Flooding can harm the environment, economic, and human life by destroying ecosystems, inundating infrastructures, and disrupting socioeconomic networks, etc. (Merz, et al., 2010; Allaire, 2018; Kasmalkar, et al., 2020). Therefore, accurate weather forecasts that can estimate extreme rainfall events are important for the environment and human society (Alfieri, et al., 2012).

Currently, numerical weather prediction (NWP) models are widely used as the operational weather forecast algorithms around the world, including the Netherlands. NWP can give relatively reliable forecasts with larger grid spaces (> 10 km) up to about one week ahead (Kain, et al., 2008; Liguori, et al., 2012). For this reason, NWP is used for early warning for large-scale weather events such as cyclones that can be predicted from several days to one week in advance. However, due to its coarse spatial resolution and low update frequency, NWP is unable to forecast rainfall accurately with shorter lead times (< 6 hours) (Davolio, et al., 2015; Silvestro, et al., 2016). Thus, their ability to predict rapid rainfall-induced events like flash floods and pluvial floods is limited. This inability is exacerbated in areas with a fast hydrological response such as small catchments (in the magnitude of 10 km^2) and urban areas (Berne, et al., 2004).

Urban areas have faster hydrological responses and thus shorter anticipation times for flooding events than rural areas due to their higher fraction of impervious surface (Tingsanchali, 2012; Sharif, et al., 2006; Chen & Chandrasekar, 2012; Berne, et al., 2004). Besides, urban areas are usually characterized by having more complex terrains than rural areas, which causes higher small-scale rainfall variability (Cristiano, et al., 2017; Maier, et al., 2020; Schellart, et al., 2014). Also, when the catchment size goes down from rural catchments to small cities, the spatially or temporally misplaced forecast rainfall causes higher overall forecast error. With the reasons above, previous studies found that increasing forecast resolution can increase forecast accuracy in urban areas (Kotroni & Lagouvardos, 2004; Rafieeinasab, et al., 2015).

Simultaneously, urban areas have a higher population density, economic productivity, and infrastructure density than rural areas (UN, 2018). Therefore, flooding in urban areas often leads to substantial economic loss and safety concerns. For instance, the record-breaking flooding in July 2021 in Germany, Belgium, and Netherlands claimed more than 200 lives and caused devastating damages to residential structures and critical infrastructure like roads and bridges (Dewan, 2021; Koks, et al., 2021). In the Netherlands, it caused an estimated economic loss of 400 million euros and thousands of people had to be evacuated (Sharma, 2021). The flooding in Paris in June 2016 claimed several human lives and more than 20,000 were evacuated (Chrisafis, 2016). Thus, forecasting technique that can capture the rapid and intensive rainfall events on a smaller spatial scale is necessary to improve the early warning system in cities.

Rainfall nowcasting, the process of statistically extrapolating the real-time quantitative precipitation estimates (QPEs) from weather radar(s), has the potential to provide forecasts up to 3 hours in advance (Shehu & Haberlandt, 2022; Jensen, et al., 2015). So, it is also called “very short term” forecasting (Sun, et al., 2014). Forecasts within such lead time can complement the gap at shorter lead times in NWP models. Nowcasting forecasts rainfall by estimating the advection of operational

radar QPE under the assumptions that rainfall intensity and motion stay stationary. Probabilistic nowcasting further adds stochastic noises to the forecasts to accout for uncertainties. Current operational radar rainfall products have improved significantly to a high temporal (5 minutes) and spatial resolution (1 km) (Overeem, et al., 2009). By using the operational QPE as input, the resulting nowcasts have the same high spatial and temporal resolution. Therefore, the high-resolution rainfall nowcasts have the potential to model the rapid development of localized and short-lived rainfall.

Under the background of rapid development in high-resolution operational radar rainfall products and the urgent need for predicting rapid extreme rainfall events, extensive research focusing on nowcasting performance in urban areas in Europe has been conducted in the past decade. For instance, Foresti et al. (2016) verified that nowcasting could estimate sewage overflows for four extreme events with lead times ranging from 30 to 90 minutes in Ghent and Leuven in Belgium. Shehu & Haberlandt (2021) showed that nowcasting can forecast urban floods with a lead time up to one hour based on study of 110 events in Hannover in Northern Germany. Similarly, nowcasting was also shown to have the ability to forecast urban flooding in the small town of Lystrup in Denmark with about 30 minutes lead time based on two events (Thorndahl, et al., 2016). Nowcasting was shown to forecast skillful rainfall volumes and sewage flows up to a lead time of 90 minutes in the city of Linz in Austria based on the analysis on five extreme summer rainfall events (Achleitner, et al., 2009). Although several analyses of nowcasting performance in hydrological applications have been conducted, most of them focused on few events or a single catchment. In other words, a comprehensive analysis on the applicability of nowcasting to extreme rainfall in cities has not been conducted, so its utility remains to be assessed quantitatively.

Imhoff et al. (2020) conducted the first large-sample analysis of over 1,500 rainfall events to systematically determine the skill of nowcasting for twelve catchments (from 6.5 to 957 km^2) in the Netherlands. They showed that nowcasting skills depend on precipitation event duration, season, catchment size and location with respect to the movement of rainfall field. The uncertainty in rainfall forecasts was found to be larger for smaller catchments, summer convective rainfall, and upwind regions of rainfall movement. Therefore, the usefulness of nowcasting depends on local conditions. The areas in the mentioned study focused on catchment scales, so its results may not be directly applicable to Dutch cities. In other words, there has not been a quantitative analysis of nowcasting capability on extreme precipitation events in Dutch cities.

Following the previous work, the aim of this research is to extend the analysis in Imhoff et al. (2020) from rural catchments to urban areas, focusing on five of the major urbanized areas in the Netherlands (Amsterdam, The Hague, Groningen, Maastricht, and Eindhoven). 80 events are selected between 2008 and 2021, and their corresponding nowcasting performance is assessed in a systematic manner. The research questions are:

- What is the performance (in terms of accuracy of forecast rainfall intensity, skillful lead time, and reliability) of nowcasting during extreme rain over major cities in the Netherlands?
- What are the main factors (e.g., precipitation intensity, duration, city size and location) that influence the nowcasting performance?
- Can operational radar rainfall nowcasting be improved with bias-corrected radar QPE?
    1. To what extent do the uncorrected and bias-corrected QPE differ from the reference?
    2. What is the effect of the differences in the QPE products on the nowcasting performance, in terms of forecast rainfall volumes and spatial correspondence?

By quantifying the usefulness of nowcasts to estimating the quick development of extreme rainfall in the five major cities in the Netherlands, results of this research will provide insight to very-short-term early warning in the Netherlands.

2. Material and methods

We introduce the study area in Section 2.1 and radar rainfall dataset in Section 2.2. The event selection method, rainfall analysis method, nowcasting models, and experimental setup are detailed in Section 2.3. Finally, various verification metrics for nowcast performance are explained in Section 2.4.

2.1. Study area

To focus on investigating nowcasts for urban areas, we selected five urban municipalities with different sizes and locations spread over the Netherlands, namely the municipalities Amsterdam, The Hague, Eindhoven, Groningen, and Maastricht, as shown in Figure 1. The selection criteria were based on the goal to assess the influence of city location on applying nowcasting to major Dutch cities with a high population. Only five cities were studied because their locations are close to the other big cities nearby, so the respective resulting nowcasting performance is expected to be representative of the nearby cities. A detailed description of the cities is listed in Table 1. The municipal data and their boundaries were determined by the district and neighborhood map 2020 version 2 (Statistics Netherlands and the Land Registry, 2020).

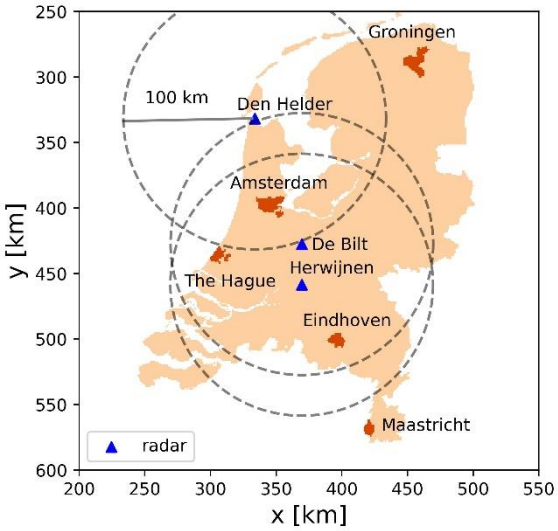


Figure 1. Locations of the five cities and KNMI radars in the study. The radar at De Bilt was replaced by the radar at Herwijnen in 2017. The QPE composites were based on the radar data from Den Helder and De Bilt (before 2017) or Herwijnen (after 2017). A range of 100 km around each radar station is shown. The x- and y-values are the coordinates in the KNMI radar projection.

Table 1. Details of the studied cities (Regionale kerncijfers Nederland, 2021)

Municipality	Population	Municipality Size (km^2)	Number of radar grid cells (1 km^2each)	Urban surface ratio	Location in the Netherlands
Amsterdam	873,338	165.5	211	0.86	West
The Hague	548,320	82.45	96	0.89	West
Eindhoven	235,691	87.66	96	0.90	South
Groningen	233,273	185.6	213	0.54	North
Maastricht	120,227	55.99	67	0.70	Southeast

Previous research shows that the skillful lead time of nowcasts depends on catchment size because smaller catchments are more sensitive to the displacement of forecast rainfall (Imhoff, et al., 2020; Pulkkinen, et al., 2019; Heuvelink, et al., 2020). Thus, to measure the effect of city size on nowcasting

performance without the dependence on location, three defined square subareas of 100, 64, and 16 km^2 were constructed in all the cities. In addition, three more subareas of sizes 4, 400, and 900 km2 were further defined in and around Groningen to study nowcast dependence on city sizes across a larger magnitude. It was only done for Groningen because Groningen was the largest city in the study. Nowcasting results of these subareas were compared to determine nowcasting dependence on area size. Maps of the square areas in each city are shown in Figure 2.

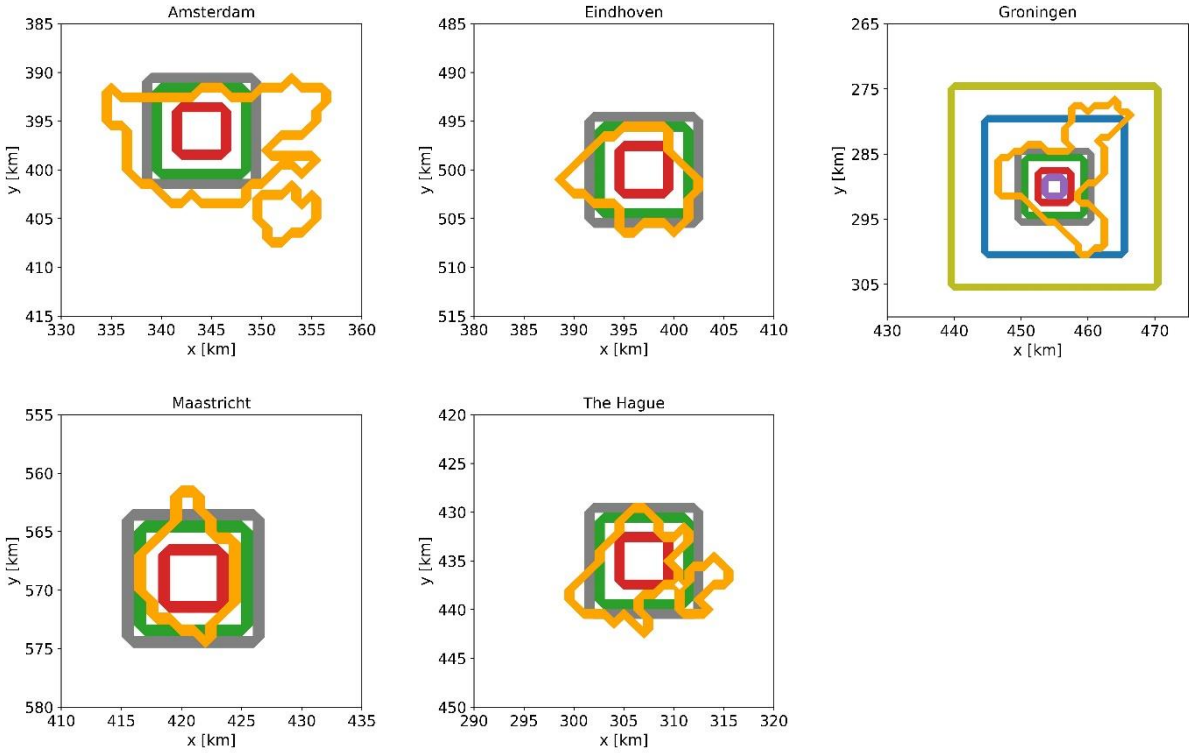


Figure 2. Map of the three square subareas of 100, 64, and 16 km2 in each city. Orange lines are the city boundaries. In and around Groningen three more subareas of size 4, 400, and 900 km2 are further defined to study nowcast dependence on city sizes across a larger magnitude.

2.2. Radar rainfall products

High spatial resolution and update frequency are the main advantages of nowcasting, so, two high-resolution radar rainfall datasets from the Royal Netherlands Meteorological Institute (KNMI) were used in the study. Both were quantitative precipitation estimates (QPEs) converted from radar reflectivity during rainfall events using Equation 1 (Marshall, et al., 1955). Z is the radar reflectivity in dBZ and R is the rainfall intensity in mm/hr.

$$Z = 200R^{1.6}$$

Equation 1

The first dataset was the 5-minute precipitation accumulations from the climatological gauge-adjusted QPE (Overeem, et al., 2009; KNMI, 2022). The radar rainfall dataset has a temporal resolution of 5 minutes and a spatial resolution of 1 kilometer. The radar rainfall data has been adjusted with measurements from validated KNMI rain gauge networks consisting of 31 automatic and 325 manual rain gauges (the exact amount of rain gauges varied slightly with time). An hourly mean-field bias (MFB) adjustment using the 31 automatic gauges and a daily spatial adjustment using the 325 manual rain gauges are performed. Because of these processings, the dataset is only updated monthly. From 2008 to January 2017, the data came from two single-polarized C-band radars in De Bilt (52.103 degrees N, 5.179 degrees E) and Den Helder (at 52.955 degrees N, 4.79 degrees E). From February 2017 onwards, the radar in De Bilt was replaced by the radar at Herwijnen

(5.1381 degrees N, 51.8369 degrees E) and the radar in Den Helder was replaced by a new radar at the same location. Both new radars are dual-polarized C-band radars. The domain of the radar rainfall data covers the entire land surface of the Netherlands. The dataset was used as a reference ("real" rainfall) and is referred to as climatological rainfall later in the report. For a more complete description of the rainfall radar products and adjustment methods, readers are referred to Overeem et al. (2009).

As the climatological radar dataset is not available in real time, it cannot be used as the input data for nowcasting in practice. Thus, the second rainfall radar product, the 5-minute non MFB-corrected real-time precipitation accumulations over the Netherlands (KNMI, 2022) is used. The dataset has the same temporal and spatial resolution as the previous dataset, but it is available in real-time. Thus, it is the QPE that is available for weather forecasts prior to the rainfall events. The dataset is referred to as real-time rainfall later.

However, it should be noted that the deviation between the real-time QPE and the actual rainfall can be large, especially during extreme precipitation events (Hazenberg, et al., 2011; Van de Beek, et al., 2016; Schleiss, et al., 2020). The reason is that the fixed Z-R relationship in Equation 1 does not hold for heavy rainfall events in which raindrop size distribution changes drastically (Schleiss, et al., 2020). Also, the cities that are further away from the radars generally have worse quality of real-time QPE (usually with larger underestimation) (Imhoff, et al., 2021). Therefore, a set of correction factors, CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting), was proposed by Imhoff et al. (2021). CARROTS has the same dimension as the real-time rainfall radar with a correction factor for each grid cell and each day of a year. CARROTS was calculated based on historical datasets of 5-min real-time and climatological radar rainfall between 2009 and 2018 in the Netherlands. Equation 2 describes the relationship between CARROTS-adjusted real-time rainfall and real-time rainfall, in which $f_{CARROTS}$ is the CARROTS factor for that day of a year.

$$QPE_{real-time} \times f_{CARROTS}(day\ of\ a\ year) = QPE_{CARROTS-adjusted}$$

Equation 2

CARROTS factors vary between 0.7 and 4.6 depending on the location in the Netherlands and day of the year. The CARROTS-adjusted real-time rainfall is referred to as CARROTS rainfall later.

Real-time QPE was used for nowcast inputs and verification from Section 3.3 to Section 3.5, so the evaluation of nowcasting performance is not affected by QPE errors. Climatological and CARROTS QPEs were only run with selected events and analyzed in Section 3.2 and Section 3.6 for comparison. Climatological QPE is used as observation to verify the "real" skill of nowcasts in an operational setting based on different input QPEs in Section 3.6.

2.3. Methods

2.3.1. Extreme event selection

To evaluate nowcasting performance in each city, extreme events were selected for each city by using RadarTools (RIONED, 2020). RadarTools employs the climatological rainfall radar products from KNMI. Based on the rainfall data, RadarTools has a built-in application "Radar Count" to list the most extreme rainfall within each municipality in the Netherlands from 2008 onwards. The list of each city consists of the extreme events given that the rainfall at any grid cell ($1km \times 1km$) within the city boundary is higher than a certain threshold corresponding to the user-defined period. For instance, the threshold for a 60-minute event is that the precipitation sum in 60 minutes in any grid cell within the city boundary is higher than 15 mm. For 24-hour events, the threshold is 30 mm. Figure 3 shows the interface of the Radar Count application and the list of the extreme 1-h events in Amsterdam.

The time in Figure 3 indicates the onset of events in UTC. For more detailed description of the extreme event selection criteria on RadarTools, readers are referred to (RIONED, 2020).

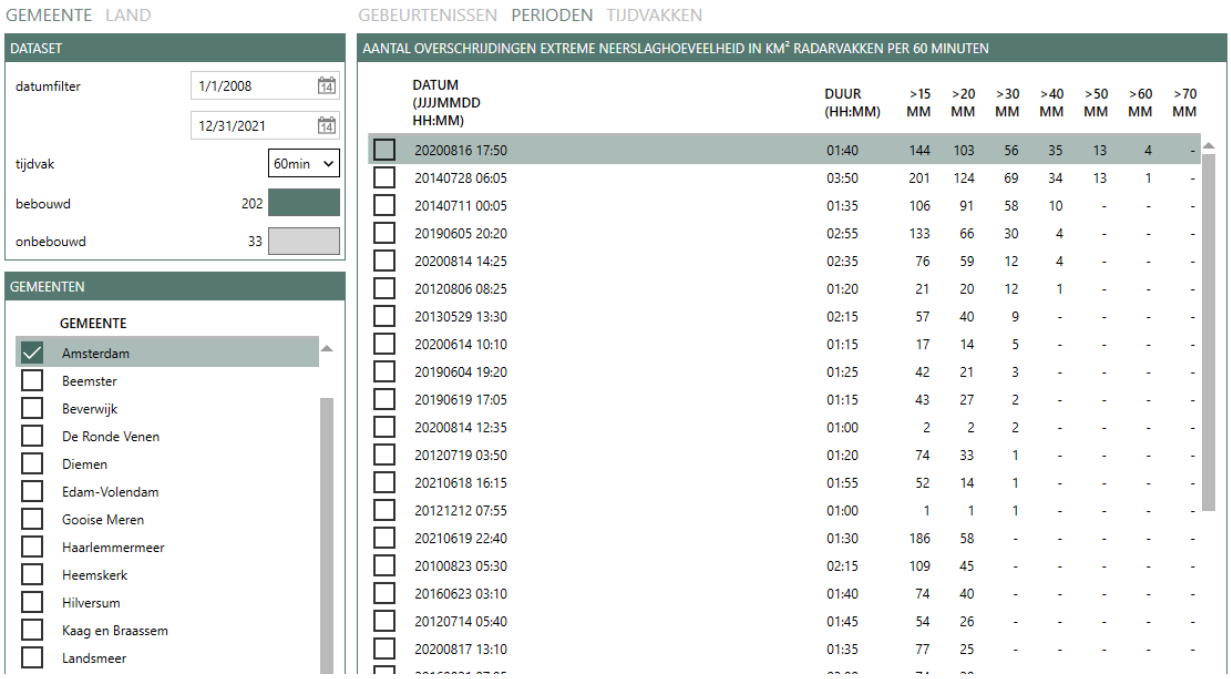


Figure 3. Interface of RadarTools with the 1-h extreme events in Amsterdam as an example. The events are ranked by the numbers of the grid cells of the highest rainfall accumulation.

Figure 3 shows that the events are ranked by the number of grid cells of the highest precipitation sum during the period in the municipality. As the research focuses on the analysis of nowcasting performance of the extreme events, only the top eight events were selected for each municipality for the 1-h period. When selecting the 24-h extreme events, the 24-h events that covered the period of the previously selected 1-h events were excluded per city to ensure independence of the events. Finally, the selection created a list of eight event start times for each city for the period of 1 and 24 hours. Selected event times are shown in Supplementary Table 1 and Supplementary Table 2.

Two durations (1 hour and 24 hours) were selected because nowcasting skills were found to vary depending on event duration (Imhoff, et al., 2020; Turner, et al., 2004). In general, the maximum skillful lead time is shorter for convective rainfall, which typically lasts shorter (Liguori & Rico-Ramirez, 2012; Mejsnar, et al., 2018), and longer for stratiform rainfall, which usually lasts longer (Berenguer, et al., 2011; Olsson, et al., 2014). This selection procedure led to 5 (cities) × 2 (durations) × 8 (events) = 80 events. The procedure is illustrated in Figure 4.

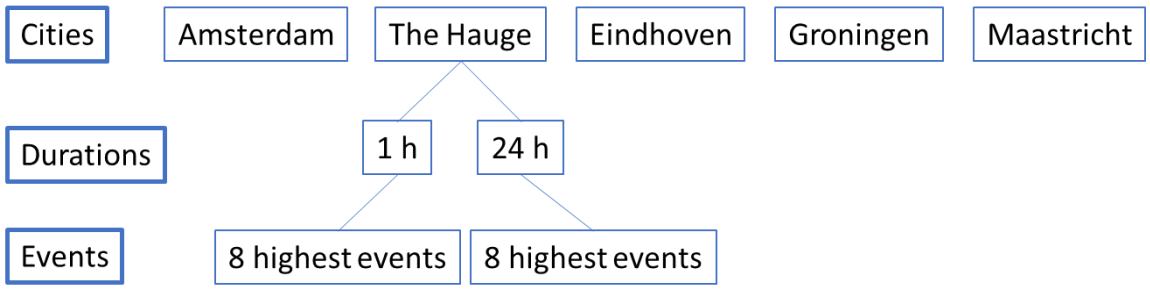


Figure 4. Precipitation event selection flowchart. 80 extreme events are selected based on the flowchart.

2.3.1. Rainfall characteristic analysis

Nowcast performance varies with rainfall characteristics (Imhoff, et al., 2020; Heuvelink, et al., 2020), so it is important to analyze rainfall characteristics. We conducted the analysis by using the real-time QPE because it is the input for nowcasting in most parts of this research. Two important parameters to quantify rainfall characteristics are calculated, area-mean rainfall (Equation 3) and fraction of cells covered by rainfall (Equation 4). Within area mean rainfall, the means with and without considering dry cells (rainfall < 0.1 mm/hr) are calculated separately.

$$area\ mean\ rainfall = \frac{QPE\ sum\ over\ the\ area}{number\ of\ grid\ cells\ in\ the\ area}$$

Equation 3

$$fraction\ of\ cells\ covered\ by\ rainfall = \frac{number\ of\ wet\ grid\ cells}{total\ number\ of\ grid\ cells}$$

Equation 4

Note that we did not apply advection correction (Anagnostou & Krajewski, 1999) when calculating rainfall accumulation. This was because the extreme events in the study were often caused by isolated convective cells that develop in small areas in a short time, particularly for the 1-h events that clustered in summer months (Supplementary Figure 10). Thus, using advection correction may lead to wrong estimation of storm velocity (Anagnostou & Krajewski, 1999).

2.3.3. Nowcasting algorithms

Pysteps (Pulkkinen, et al., 2019), an open-source Python nowcasting framework, was chosen to implement in this study for the reasons below. First, it is free to install and provides several nowcasting algorithms (e.g., short-term ensemble prediction system, ANVIL, and LINDA). Second, Pysteps provides multiple verification modules to derive and visualize e.g., Pearson correlation coefficient, continuous rank probability score, and receiver operating characteristic curve (ROC) which are useful for analyzing the accuracy and consistency of nowcasts. Third, Pysteps can provide probabilistic nowcasts (or ensemble nowcasts) by adding stochastic perturbations that are correlated, in space and time, to the observed rainfall fields. Given the higher difficulty in forecasting extreme rainfall, probabilistic forecasts have higher chance to capture, to a certain extent, the quick growth and dissipation rates of extreme events. Lastly, the resulting nowcasts can be compared with the results in (Imhoff, et al., 2020), which used Pysteps as one of the main algorithms too.

Pysteps includes multiple nowcasting methods in different steps, e.g., optical flow methods, advection methods, and noise methods. In the study, we use the STEPS (short-term ensemble prediction system) (Seed, 2003; Bowler, et al., 2006; Seed, et al., 2013)with the characteristics below.

1. Semi-Lagrangian advection method to extrapolate future rainfall field
2. Lucas-Kanade optical flow method that uses the most recent three radar rainfall fields to track its local features
3. An autoregressive model of order 2
4. Eight cascade levels to decompose the rainfall field with decreasing spatial scale
5. A non-parametric noise method to add stochastic perturbations to the results to represent uncertainty during rainfall evolution
6. For one nowcast at each issue time, 20 ensemble forecasts are produced. Each ensemble member represents a possible future state given uncertainties of e.g., initial state and advection field of formulations.

The STEPS setup was chosen because it showed the longest skillful lead time among the nowcasting algorithms tested in Imhoff et al. (2020). The 20 ensemble members comprised a probabilistic nowcast that showed a spread of possible rainfall fields in the future. Although an ensemble number as high as 96 was shown to have better performance in forecasting extreme precipitation (Pulkkinen, et al., 2019), 20 ensemble members were chosen because it struck a balance between computational efficiency and stochastic forecast in the need of this study.

The nowcasting results were compared with the rainfall field given by Eulerian persistence. Eulerian persistence or "no nowcasting", uses the most recent QPE as forecast for future time steps. It was used as a baseline to evaluate the additional values of nowcasting such as reduction in error and gained lead time.

2.3.4. Experimental setup and high-performance computation

To evaluate nowcast skill with lead time, nowcasts were issued from 4 hours before the start of the events with a 10-minute time step, forecast resolution of 5 minutes, and forecast horizon of 4 hours. The last issue time of nowcast was 10 minutes before the end of the event. The experimental setup for a 1-h event is illustrated in Figure 5. Forecasts using Eulerian persistence had the same issue time and horizon as the nowcasts. For instance, the Eulerian persistence issued at 13:50 uses the most recent QPE at 13:45 as the rainfall forecast for the next 4 hours until 17:50. Finally, only the nowcasts that fall within the event duration are analyzed, which was further explained in Section 2.4.

The maximum lead time of 4 hours was chosen because Imhoff et al. (2020) showed that on average nowcasting is no longer skillful beyond a 4-hour lead time for 24-hour extreme rainfall events. It was expected that the skillful lead time (defined in Sections 2.4.3 and 2.4.4) would be shorter for shorter events. Due to the smaller sizes of the cities in this study, it was believed that the skillful lead time would be shorter than 4 hours.

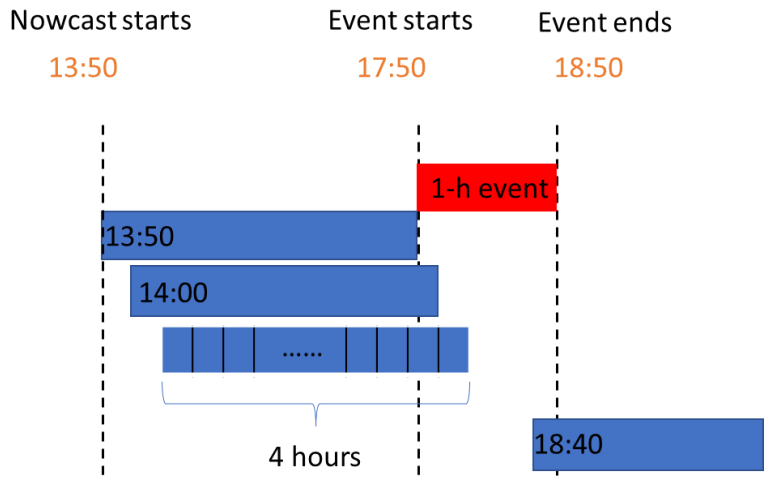


Figure 5. Illustration of the issue time and forecast horizon for the 1-h event starts at 17:50. The blue bars show forecast horizon of 4 hours with 5-minute resolution at different issue times. The red bar shows the event duration. Only the nowcasts that fall within the event duration are analyzed.

With the foregoing method, there were $5 \times 8 \times 30 = 1{,}200$ nowcasts for 1-h events and $5 \times 8 \times 168 = 6{,}720$ for 24-h events. Furthermore, each of the nowcasts was comprised of 48 time steps (5-minute resolution within the 4-hour forecast horizon) and 20 ensemble members. Due to the high computational demands and memory requirements, the nowcasts were computed on the supercomputer at the Delft University of Technology, Delft Blue (Delft High Performance Computing Centre (DHPC), 2022), which launched in April 2022. Delft Blue consists of more than 11,000 CPU cores spread over 220 compute nodes and incorporates a high-speed parallel storage subsystem.

Each compute node consists of Intel Xeon Gold 6248R of 48 cores. The compute nodes are built with the latest Intel Cascade Lake refresh processors offering high performance and power efficiency. Running 1 probabilistic nowcast with 48 time steps and 20 ensemble members for the whole radar domain in the Netherlands took around 3 minutes on 1 core. By parallel computation on Delft Blue, the total computation time for all the nowcasts significantly reduces.

2.4. Verification method and metrics

After explaining the event selection and nowcasting method, this section outlines the verification method and metrics for the output nowcasts. The verification was only performed for the nowcasts during the period of the rainfall events. Using Figure 5 as an example, the rainfall event lasts from 17:50 to 18:50. Nowcasts at each issue time composed of nowcasts of the next 48 timesteps (because each time step is 5 minutes, 48 timesteps equal a forecast horizon of the next 4 hours). Thus, some nowcasts are not within the event period. For example, the first nowcast time of the nowcast issued at 13:50 is 13:55 and the fourth nowcast time is 14:10. Such nowcasts were excluded in the verification because they are not within the event period (17:50 to 18:50). Besides, all metrics only verified the time steps at which observed area-mean rainfall higher than 0.1 mm/hr, except the receiver operating characteristic curve because it calculates hit rate and false-alarm rates with a set of thresholds. As a result, although many nowcasts were produced, only those within the event period and where rainfall is observed were verified.

2.4.1. Continuous rank probability score (CRPS)

Estimating rainfall intensity and location accurately is an important aspect of a rainfall forecast. Errors in the rainfall forecast can lead to wrong decision-making in operational water management such as underestimating flooding risks and mistaking inundated areas. Continuous rank probability score (CRPS) is one of the most widely used accuracy metrics for probabilistic precipitation forecasts (Imhoff, et al., 2020; Ravuri, et al., 2021; Shehu & Haberlandt, 2022; Löwe, et al., 2014).

For each event, the CRPS was used to quantify the average error between the nowcasting results and radar rainfall observation. Because the forecasts consisted of ensembles, CRPS was calculated for the entire forecast distribution. CRPS was formulated as

$$CRPS = \frac{1}{N_f}\sum_{i=1}^{N_f}\int_{-\infty}^{\infty}\left(P_{F_i}(x) - P_{O_i}(x)\right)^2 dx$$

Equation 5

N_f is the number of forecasts per lead time or city depending on the desired variable to be analyzed, P_{F_i} is the exceedance probability of nowcast rainfall, and P_{O_i} is counterpart for the observed rainfall. A smaller value of CRPS indicates that the difference between nowcast and observation is smaller. For deterministic nowcasts and Eulerian persistence, CRPS reduces to mean absolute error (MAE), so the comparison between probabilistic and deterministic nowcasts is feasible.

Previous studies show that forecast error is higher in smaller catchments and for more intense rainfall, so the analysis of CRPS was done with different rainfall and urban characteristics. The influence of rainfall intensity on CRPS was analyzed for all events. The dependence of nowcasts on city size was analyzed by comparing the CRPS of the subareas in each city.

2.4.2. Bias

Another way to measure forecast accuracy is by comparing total forecast rainfall volume with observation over the whole city. Wrong nowcast precipitation volume can further lead to wrong estimation of hydrological responses. For instance, sewage overflow and flooding area can be

wrongly modeled. Therefore, knowing the relative difference between nowcast and observed rainfall is important for hydrological application. Bias was used to quantify the ratio between nowcast and observed rainfall volume with the equation below.

$$bias = \frac{F_i}{O_i}$$

Equation 6

F_i and O_i are forecast and observed rainfall accumulation during the events. When the value is closer to 1, it means the nowcast is more accurate. Smaller values of the bias indicate a larger underestimation of the nowcast and vice versa. Because it verifies rainfall over the whole area, rainfall spatial variability was not considered, unlike CRPS. Therefore, bias gave a general indication of nowcasting performance over the considered area. For the same reasons as in 2.4.1, bias was analyzed with both urban and rainfall characteristics.

On the other hand, bias also exists between real-time QPE and climatological QPE as described in Section 2.2. So, we also define QPE bias (Equation 7), which is the ratio between the area-mean real-time or CARROTS QPE and the climatological QPE within the city boundary. This metric was used to measure the accuracy of real-time QPE.

$$QPE\ bias = \frac{realtime\ QPE}{climatological\ QPE}$$

Equation 7

2.4.3. Fraction skill score

Besides accuracy, skillful lead time was also estimated using the Fraction skill Score (FSS) and Pearson correlation. Skillful lead time is an important index to measure how much time in advance nowcasts can forecast the events with desired accuracy. Fraction skill Score (FSS) is a recently developed metric used for verifying rainfall forecasts (N.M. Roberts, 2008). It quickly gains popularity in rainfall forecast research (Skok, 2015; Mittermaier, et al., 2013; Zhao & Zhang, 2018; Rezacova, et al., 2009).

To compute FSS, first, rainfall fields of both nowcast and observation were converted to a binary field according to a pre-defined threshold. In the study it was set at 1 mm/hr. Grid cells with rainfall higher than the threshold became 1 and those lower than the threshold became 0. Second, a square area surrounding the grid cell was created with the defined length scale. Three length scales tested in the study were 1, 5, 10, and 15 km. Numbers of grid cells within the square area that were higher than the thresholds were accumulated for the nowcast rainfall and the observed rainfall separately. Third, FSS was calculated based on the mean square error (MSE) between the two numbers. For a more detailed description of the metric, readers are referred to N.M Roberts (2008).

FSS ranges between 0 and 1 with 1 meaning perfect nowcasts and 0 meaning the largest possible MSE that can be obtained from the nowcast and observed fractions. Skillful lead time of nowcasts given by FSS is derived by Equation 8 and Equation 9.

$$FSS_{skillful} \geq FSS_{random} = 0.5 + \frac{f_0}{2}$$

Equation 8

$$f_0 = \frac{N_{>threshold}}{N_{total}}$$

Equation 9

FSS_{random} is the random forecast skill. f_0 is the number of grid cells in the observation that are higher than the threshold divided by the number of total grid cells in the considered city within the forecast horizon.

FSS is generally higher when the length scale is larger because the verification becomes less sensitive to the displacement of rainfall. Also, by setting the same length scale, FSS becomes a useful tool to assess nowcasts at different cities at the same spatial resolution. Thus, FSS is used to compare skillful lead time of nowcasts in different cities.

2.4.4. Pearson correlation coefficient

Pearson correlation provides another method to measure skillful lead time besides FSS using the spatial correlation between nowcast and observation. Skillful lead time derived by Pearson correlation takes rainfall spatial variability at the finest scale (1 km^2) into account, while FSS upscales the rainfall field to the defined sizes.

For every event, we tested different lead times of the nowcasts and calculated the Pearson's correlation coefficient per grid cell with Equation 10.

$$\rho = \frac{1}{N_f} \sum_{i=i}^{N_f} \frac{(F_i - \mu_F)(O_i - \mu_o)}{\sigma_F \sigma_o}$$

Equation 10

F_i and O_i denote the average nowcast and observed rainfall in the city. N_f is the number of forecasts at a given lead time in the event considered, μ is the mean rainfall, and σ is the standard deviation. Pearson correlation coefficient was calculated for the ensemble means of each nowcast. Because Pearson's correlation coefficient was calculated per grid cell, it produced a two-dimensional field. The average of Pearson's correlation coefficient from all the grid cells was calculated to examine the skillful lead time of the nowcasts. $1/e$ $(\cong 0.37)$ was used to determine the decorrelation time between nowcast and observed rainfall. Once the average Pearson's correlation coefficient drops below 0.37, the forecast is not considered skillful anymore (Choi & Kim, 2022; Imhoff, et al., 2020; Mejsnar, et al., 2018; Germann, et al., 2006).

As Imhoff et al. (2020) found that maximum skillful lead time varies with event durations and catchment sizes, analyzing skillful lead time per event duration was the main goal in this part. Skillful lead times of the subareas of various sizes were also compared. By using Pearson correlation, a systematic comparison of skillful lead time regarding different event durations and area sizes was performed.

2.4.5. Receiver operating characteristic (ROC) curve

Unreliable forecasts can compromise the lead time of forecasts and lead to wrong decision-making and early warning responses. So, after knowing the accuracy and skillful lead time of nowcast, we quantified its reliability at a fixed lead time of 30 minutes by using receiver operating characteristic (ROC) curves. ROC curve plots hit rate on the y axis and false alarm rate on the x axis. Hit rate and false alarm rate were calculated over a defined threshold. In this study, we used three thresholds (0.1, 1, and 5 mm/hr) to assess nowcast reliability on forecasting different rainfall intensities.

Hit rate and the false alarm rate were computed by applying Equation 11 and Equation 12.

$$\text{Hit rate (HR)} = \frac{TP}{FN+TP}$$

Equation 11

FN: misses, or false negative, observed rainfall is over the threshold but it is not forecast

TP: hits, or true positives, both forecast and observation are over the rainfall threshold

$$\text{False alarm rate (FAR)} = \frac{FP}{FP+TN}$$

Equation 12

FP: false positive forecasts, forecasts exceed the threshold but observation does not

TN: true negative forecast, both forecast and observation do not exceed the threshold

The *FN*, *TP*, *FP*, and *TN* were calculated per grid cell. The ratio between HR and FAR is an indicator to nowcasting reliability: higher ratio means higher reliability. If the ratio *HR* to *FAR* is below 1, the nowcast is considered to have no skill. As previous study already found that reliability of nowcasts reduces with lead time (Imhoff, et al., 2020), this study focused on the reliability of nowcasts according to different cities, rainfall durations, and thresholds (0.1, 1, and 5 mm/hr). Therefore, with accuracy metrics in Section 2.4.1 and 2.4.2, skillful lead time in Section 2.4.3 and 2.4.4, and the reliability quantification in this section, nowcasting capability was comprehensively identified.

3. Results

We start with analyzing the characteristics of the rainfall events in Section 3.1. Then, observed and nowcast rainfall of two events are shown as examples to show the skill and challenges of nowcasting in Section 3.2. Following that, performance of the nowcasts is analyzed in four main aspects. First, the performance in the five cities is compared in Section 3.3. Second, dependence of nowcasting performance on rainfall characteristics, including intensity and duration are assessed in Section 3.4. Third, nowcasting skills at subareas of different sizes are compared to study dependence of nowcasting skills on area sizes in Section 3.5. Finally, different QPEs and their corresponding nowcasts are compared in Section 3.6.

3.1. Rainfall characteristics

Real-time observed QPEs of all the selected events in each city are analyzed prior to nowcasting. The reults are shown in Figure 6. Average rainfall intensity for 1-h events is 4.9 mm/hr (considering all cells) and 7.3 mm/hr (only considering wet cells); for 24-h events is 0.7 mm/hr (17.3 mm/day) (considering all cells) and 2.1 mm/hr (49.8 mm/day) (only considering wet cells) mm/day. To provide a rough impression of the extremeness, the rainfall intensity of 1-year return period in the Netherlands is 11.2 mm/hr for 1-h duration and 32.7 mm/day for 24-h duration for a radar area of approximately 6 km^2 (using GEV-fitted parameters in Table 3 in Overeem et al., (2009)). However, rainfall intensity of the same return period in the five cities should be lower due to the areal reduction factor of around 0.8 to 0.95 depending on the city sizes and event durations (Overeem, et al., 2010). Therefore, the intensities of the selected rainfall are slightly lower the events of 1-year return period.

For the 1-h events, panel (a) shows that Eindhoven and Maastricht have higher rainfall than the other three cities before setting the threshold. After setting the threshold of 0.1 mm/hr, Eindhoven still has the highest rainfall, and rainfall in the other cities becomes similar. For the 24-h events, area-mean rainfall reduces to under 2 mm/hr in all cities. Also, the area-mean rainfall in each city is similar. If only considering grid cells above the threshold, Groningen has the highest mean rainfall.

The fraction of cells covered by rainfall is plotted in the panels (e) and (f) in Figure 6. It shows that Eindhoven and Maastricht have higher fractions. This is most likely due to their smaller area. Panel (e) shows that Amsterdam has the lowest wet fraction among all the cities which means the 1-h events in Amsterdam are more spatially concentrated. Comparing to the 1-h events, the wet fraction of 24-h events is smaller, and Groningen has the smallest fraction. Among the 24-h events, there are no events that constantly cover every pixel in the whole city. For more complete visualization of the rainfall events, the rainfall accumulation of each event in each city is shown in Supplementary Figure 1 and Supplementary Figure 2.

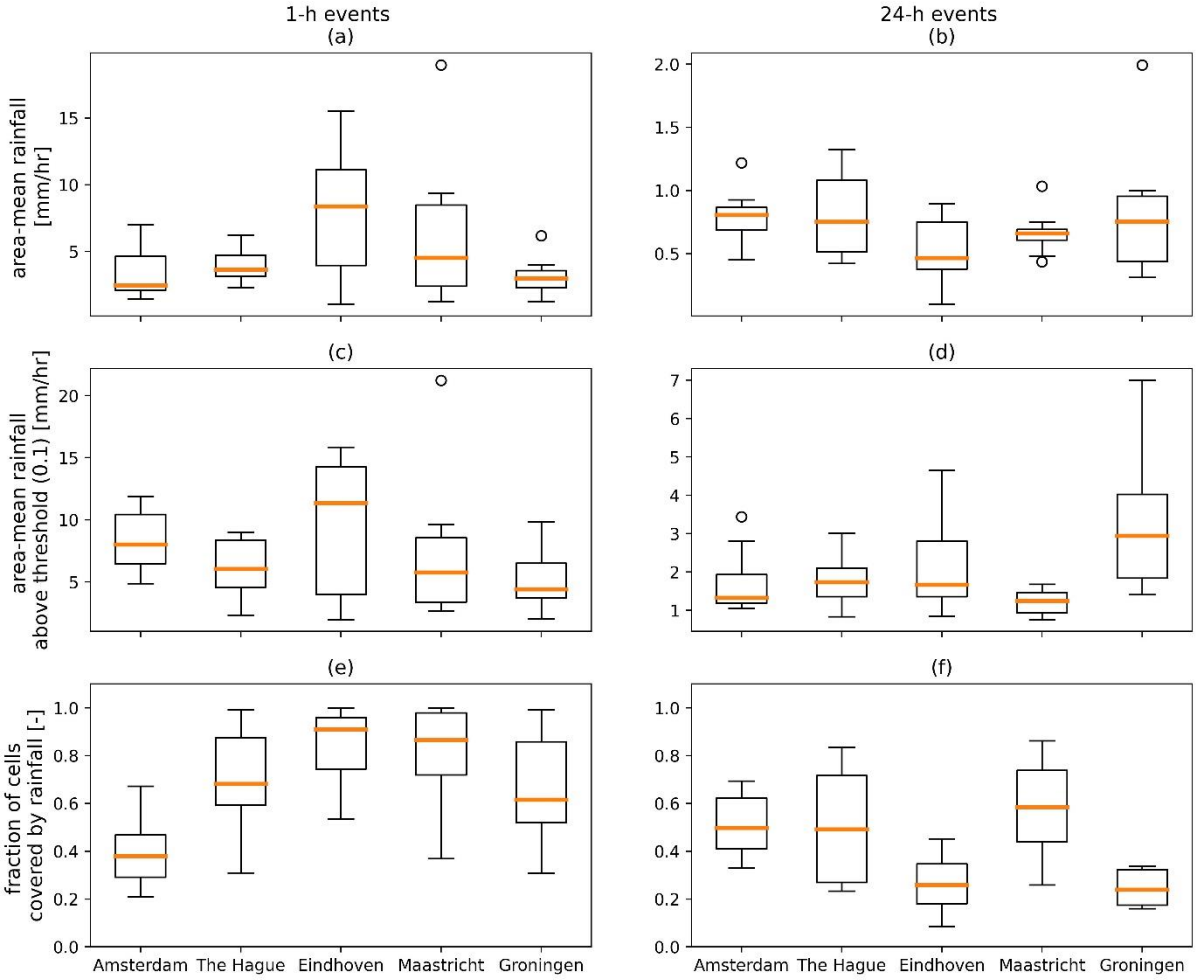


Figure 6. Rainfall characteristics of all the 1-h and 24-h rainfall events in each city. Figure (a) and (b) shows the area mean rainfall. Figure (c) and (d) shows that area mean rainfall excluding the cells that do not observe rainfall (rainfall intensity < 0.1 mm/hr). Figure (e) and (f) shows the fraction of grid cells that observe rainfall higher than 0.1 mm/hr during the events. The four black lines in each column show quartiles of the 8 events in each city. The orange lines are the medians.

3.2. Case study

Before analyzing all events, we show one 24-h event and 1-h event as example to demonstrate the difference in the radar rainfall products and challenges of nowcasting in small cities. For the 24-h event, we focus on the rainfall event in Maastricht on 13th July 2021 14:00. The event was during the most severe flooding in Europe in recent years that occurred between 12 and 25 of July in countries including mainly Germany, Belgium, and south Netherlands, as described in Chapter 1.

The climatological, CARROTS, and real-time rainfall observation and nowcasts with 30-, 60-, and 90-minute lead times using the real-time rainfall are shown in the upper 2 rows in Figure 7. First, a large bias of QPE between climatological and real-time rainfall is clearly present. In Maastricht, the real-time rainfall volume is only 21% of the climatological rainfall. This underestimation significantly decreases the accuracy of nowcasts which are run with real-time rainfall as input. Therefore, as the figure shows, none of the three nowcasts are able to estimate the high rainfall that is present only in the climatological QPE. To adjust this bias, CARROTS is used. The ratio between CARROTS rainfall and climatological rainfall in Maastricht becomes 43%. Although underestimation still exists, nowcasts run with CARROTS QPE should show better correspondence to climatological QPE than nowcasts run with real-time QPE.

Comparing the three nowcasts run with different lead times, it shows that as lead time increases, nowcast rainfall intensities reduce. The small orange area of the most intensive rainfall shown in

Figure 7 in the real-time rainfall observation dissipates already when the lead time is as short as 30 minutes. So, the most intensive rainfall is not forecast by any other nowcasts at longer lead times. This inevitably causes large underestimation of rainfall. As shown in Figure 8, the real-time rainfall intensity in Maastricht is 15.8 mm/day, while nowcasts with 30-min lead time give an ensemble mean rainfall intensity around 11 mm/day and nowcasts of 90-min lead time only estimate around 8.2 mm/day.

By using the 20 ensemble members of the nowcasts, the probability of exceedance can be plotted. The bottom 2 rows in Figure 7 show the rainfall exceedance probability maps according to thresholds of 5 and 10 mm/day. In the climatological, CARROTS, and real-time rainfall observation maps, areas where rainfall is higher than the threshold are given a probability of 1 while the other regions are given 0. When the rainfall intensity is 5 mm/day, the probability maps of nowcasts are quite similar to the observation at the three lead times. Yet, when the threshold increases to 10 mm/day, probability maps at the three lead times show large deviation among themselves and to the observed exceedance probability. The area of precipitation higher than the threshold is highly underestimated already when the lead time is as short as 30 minutes, and it almost disappears when the lead time is 90 minutes. This implies that it is more difficult to nowcast high rainfall.

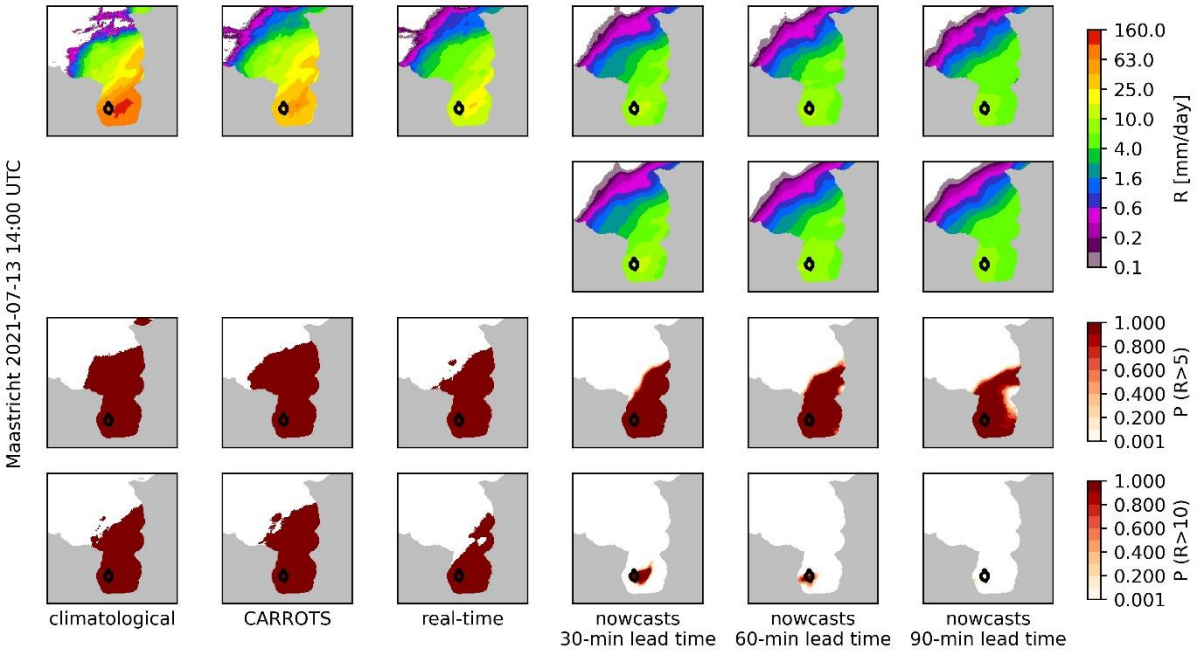


Figure 7. Climatological, CARROTS-adjusted, and real-time rainfall measurement and nowcasts with 30-, 60-, and 90-minute lead time of the extreme event starting from 14:00 13th July 2021 to 14:00 14th July 2021 in Maastricht. Upper 2 rows: rainfall intensity. The first row shows the nowcasts at three lead times using individual ensemble member no.1. The second row shows the nowcasts as ensemble mean. Bottom 2 rows: rainfall exceedance probability with threshold of 5 and 10 mm/day respectively. The black borders indicate Maastricht.

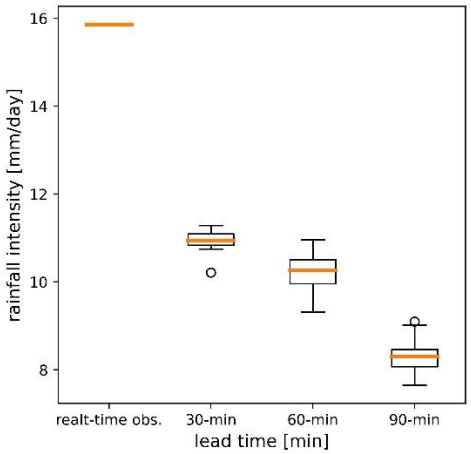


Figure 8. Nowcast rainfall intensity in Maastricht versus lead time for the 24-h extreme event starting at 14:00 13th July 2021. The spread is from the 20 ensemble members of the nowcasts. The orange lines show the medians of the ensembles.

For the second example case, the 1-h event starts at 01:55 23rd June 2016 in The Hague is selected. The heavy storm brought record rainfall and strong wind to the central-western part of the Netherlands and caused more than 20 million euros of property damage (DutchNews.nl, 2016).

The climatological, CARROTS, and real-time rainfall measurement and nowcasts with 30-, 60-, and 90-minute lead times are shown in Figure 9. The QPEs are slightly better than in the previous case. Bias between real-time and CARROTS QPE to climatological QPE in the Hague rise to 0.31 and 0.55. The upper-right and center-left areas of high rainfall intensity are also correctly identified. However, in this case, nowcasting performance strongly depends on the size and location of the target area. Although the nowcasts forecast well the general rainfall pattern, rainfall in the Hague is largely missed because it is on the edge of the rainfall field. Thus, the rainfall intensity is severely underestimated, as shown clearly in the probability maps. The underestimation would be smaller for a city that is closer to the upper-right and center-left regions, or if the city size is large enough to reduce the sensitivity caused by displacements of rainfall.

From the above examples, we can observe three challenges related to radar-based nowcasting techniques. First, the underestimation of real-time QPE during extreme events. As later shown in Figure 21, the bias between real-time and climatological QPE is around 0.3 to 0.8 for each city. Second, the tendency to underestimate small areas of intensive rainfall already at short lead time and worsen as lead time increases. The main cause for underestimation is that radar-based nowcasts cannot predict the growth of new rainfall (Shehu & Haberlandt, 2021; Pulkkinen, et al., 2019; Bowler, et al., 2006; James, et al., 2018) because it does not consider information of atmospheric stability such as convective available potential energy (CAPE) and moisture convergence which can lead to convective rainfall (Steinheimer & Haiden, 2007; Sun, et al., 2014). Third, the size and location of the cities are important elements in the verification of nowcasts. The smaller is the area, the more important to forecast the rain at the right location and time.

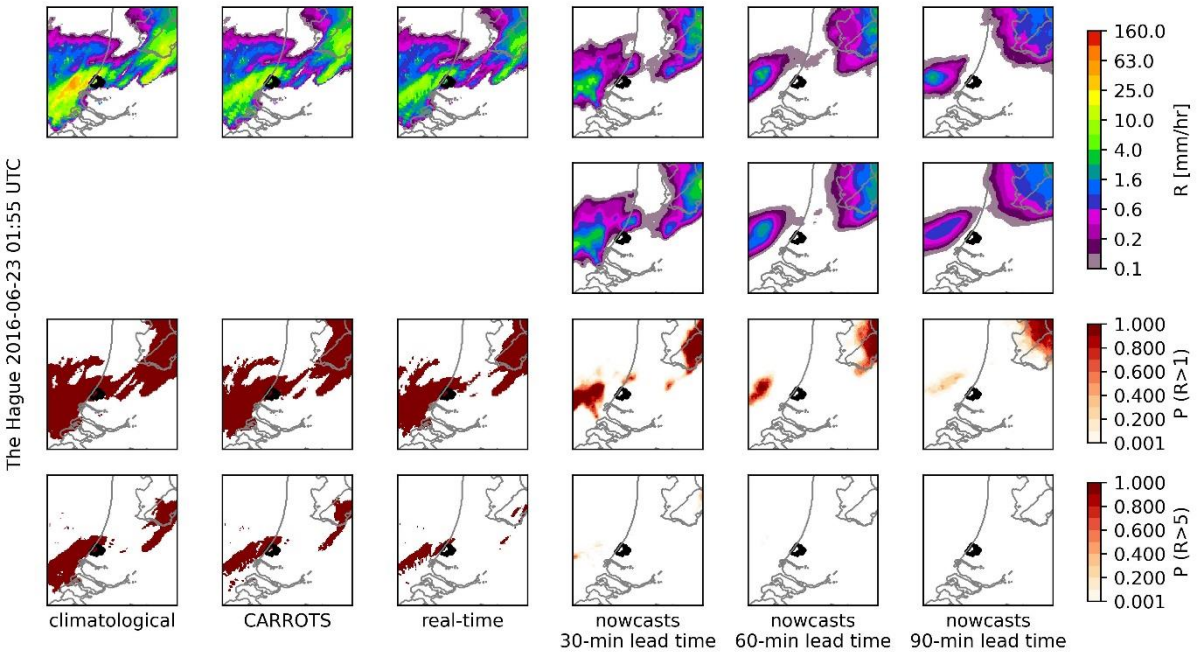


Figure 9. Climatological, CARROTS-adjusted, and real-time rainfall measurement and nowcasts with 30-, 60-, and 90-minute lead time of the extreme event start from 01:55 23rd June 2016 to 02:55 23rd June 2016 in the Hague. Upper 2 rows: rainfall intensity. The first row shows the nowcasts at three lead times using individual ensemble member no.1. The second row shows the nowcasts as ensemble mean. Bottom 2 rows: rainfall exceedance probability with threshold of 5 and 10 mm/day respectively. The black borders indicate the Hague and the gray boarders show the Dutch coastline.

3.3. Nowcast performance in the five cities

With the rainfall analysis and visualization of two events, we proceeded to the analysis of all 80 events. In this section, nowcast skill in different cities was compared and related to the average rainfall characteristics in each city.

3.3.1. Nowcast accuracy

We started with measuring nowcast accuracy. The average CRPS of the eight 1-h events in each city is shown in Figure 10 (a). Lower forecast accuracy is found in Maastricht and Eindhoven which have higher CRPS by a factor of 2 to 3 than the other cities. The cause is that Eindhoven and Maastricht have higher mean rainfall of their eight 1-h events: 8.1 mm/hr and 6.4 mm/hr, which are twice hihger than in the other cities (3.1 mm/hr in Groningen, 3.4 mm/hr in Amsterdam, and 3.9 mm/hr in the Hague). Forecast error is higher when rainfall intensity is higher. This causality is further confirmed by the strong positive correlation shown in Figure 11 (a) which plots the CRPS of each 1-h event along the mean rainfall intensity. The rainfall events with higher intensity usually also have higher standard deviation. In addition, a strong positive correlation is also found between CRPS and spatial standard deviation of rainfall, as shown in Figure 11 (b). Convective and small-area rainfall tends to have a larger standard deviation than stratiform and large-area rainfall. Therefore, the higher CRPS in Maastricht and Eindhoven might also be due to their more convective, heterogeneous, and dynamic rainfall. We also compared the difference between CRPS of nowcasts and MAE of Eulerian persistence in the cities, but the results are quite alike between the cities, as shown in Supplementary Figure 3. In other words, the additional gains in accuracy are quite similar in the cities.

This strong dependence of CRPS on rainfall intensity has to be decoupled to enable the analysis of other influences on nowcast accuracy like city sizes. So, CRPS is normalized by the standard deviation of the rainfall during the events to exclude the dependence on rainfall intensity. The results are

shown in Supplementary Figure 5. The clear relationship between CPRS and rainfall intensity disappears. However, Supplementary Figure 6 shows that normalized CRPS still shows differences among the cities. It presents that larger cities like Amsterdam and Groningen have the lowest normalized CRPS in both 1-h and 24-h events. Thus, CRPS might depend on area sizes, which we further studied in Section 3.5.

After comparing CRPS in the cities, we analyze its variation with lead time. For 1-h events, CRPS in all cities increase between lead time from 0 to 60 minutes. After 60 minutes, the curves of CRPS plateau out (except for Maastricht, whose CRPS keeps increasing), which is caused by reaching the maximum possible error between nowcasts and observation. Comparing average CRPS of nowcasts and average MAE of Eulerian persistence, nowcasts do not show additional skills after about 60-minute lead time.

CRPS reduces significantly in the 24-h events compared to the 1-h events, as shown in Figure 10 (b). Noted that the scale of the y axis is smaller in (b) than in (a). This is a result of the less intensive rainfall of 24-h events, as shown in Figure 6. Eulerian persistence shows similar skill as for the 1-h events, with the MAE reaching maximum when the lead time is about 20 minutes. Yet, nowcasts of 24-h events show much better accuracy because the average CRPS is always lower than the MAE of Eulerian persistence throughout the lead times from 0 to 4 hours. Same results are found when comparing the CRPS of nowcasts and MAE of Eulerian persistence in each city, as shown in Supplementary Figure 4.

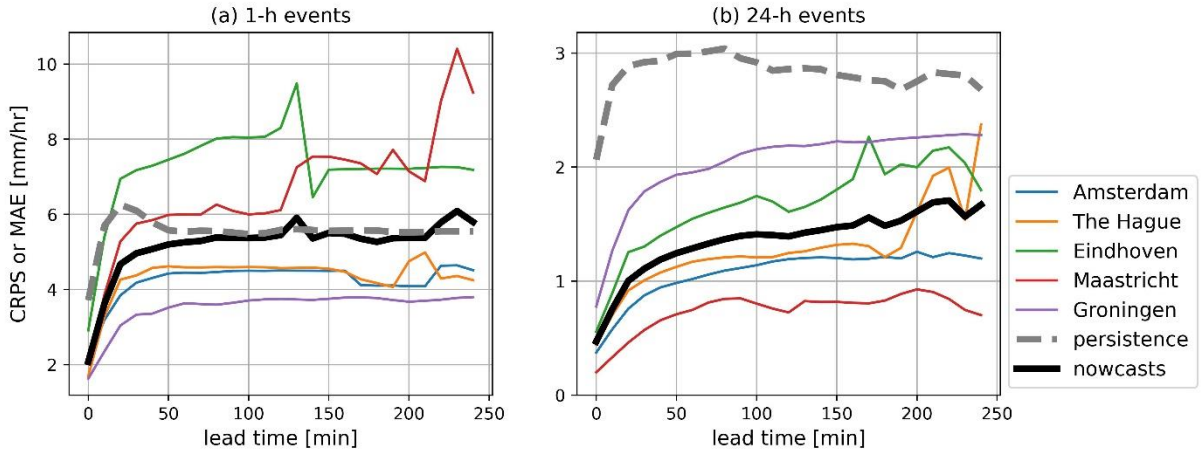


Figure 10. Average CRPS for each city for the 1-h events (left) and 24-h events (right). Note the different scales of y axis in the two plots. Shown CRPS is resampled to 10 minutes. The black curves show the average CRPS given by nowcasts in all cities. The dashed gray lines show the average MAE of Eulerian persistence from all cities.

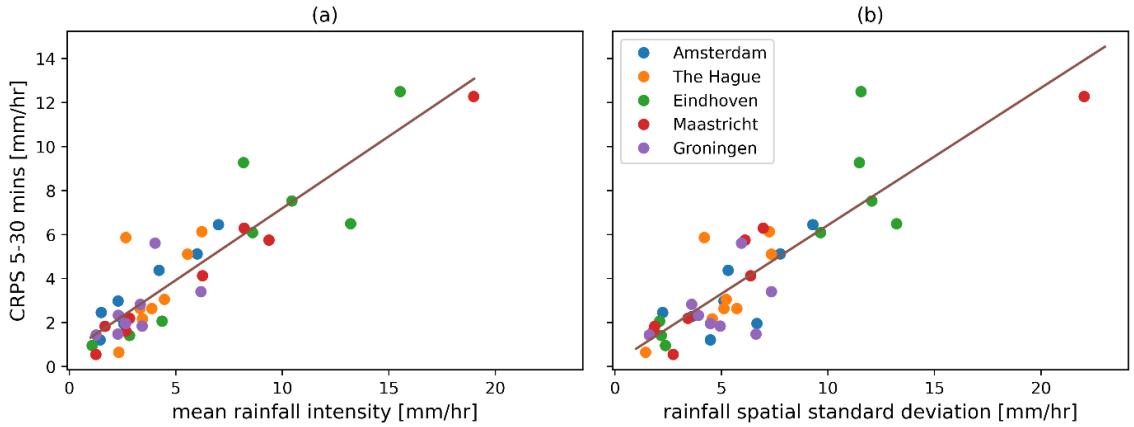


Figure 11. CRPS (average between 5- to 30-minute lead time) against mean rainfall intensity and rainfall spatial standard deviation in the five cities during the 1-h events. Each dot is one rainfall event.

3.3.2. Nowcast skillful lead time

The CRPS shows that nowcasts generally have higher skill than Eulerian persistence within the four-hour forecast horizon. Following this finding, we further used FSS to measure skillful lead times in the cities. FSS at four length scales is calculated for the 1-h events. Results are shown in Figure 12.

Figure 12 shows that FSS is larger when the length scale is longer, as expected because larger areas are less sensitive to displacement of rainfall. Notably, the increase in FSS is higher from 1 to 5 km than from 5 to 10 km. A length scale of 10 km already leads to an area of 100 km^2 that is larger than Maastricht, the Hague, and Eindhoven. Further upscaling to 15 km does not enhance FSS much because 225 km^2 is larger than all the cities. Thus, about 5 km can be considered as near the optimal spatial resolution for nowcasting extreme rainfall in the cities.

FSS among the cities does not show a clear distinction as the CRPS does, because the difference in skillful lead time is small (at the same length scale). The skillful lead times (within in which FSS is higher than FSS_{random}) in the cities at all length scales vary slightly between 12 and 26 minutes. The shortest skillful lead time is found in Amsterdam when the length scale is 1 km, and the longest one is in Eindhoven when the length scale is 15 km. FSS at a length scale of 5 km in each city is averaged and plotted in the bottom-right figure. It shows that the average skillful lead time is around 18 minutes. Skillful lead time derived by FSS shows that nowcast performance in forecasting rainfall higher than 1 mm/hr is limited to 20 minutes, although nowcasts can improve the CRPS of Eulerian persistence up to 1-h lead time (Figure 10 (a)). In Section 3.4.2, we further applied Pearson correlation to measure skillful lead time without considering the threshold (1 mm/hr) and length scale that are necessary for the calculation of FSS.

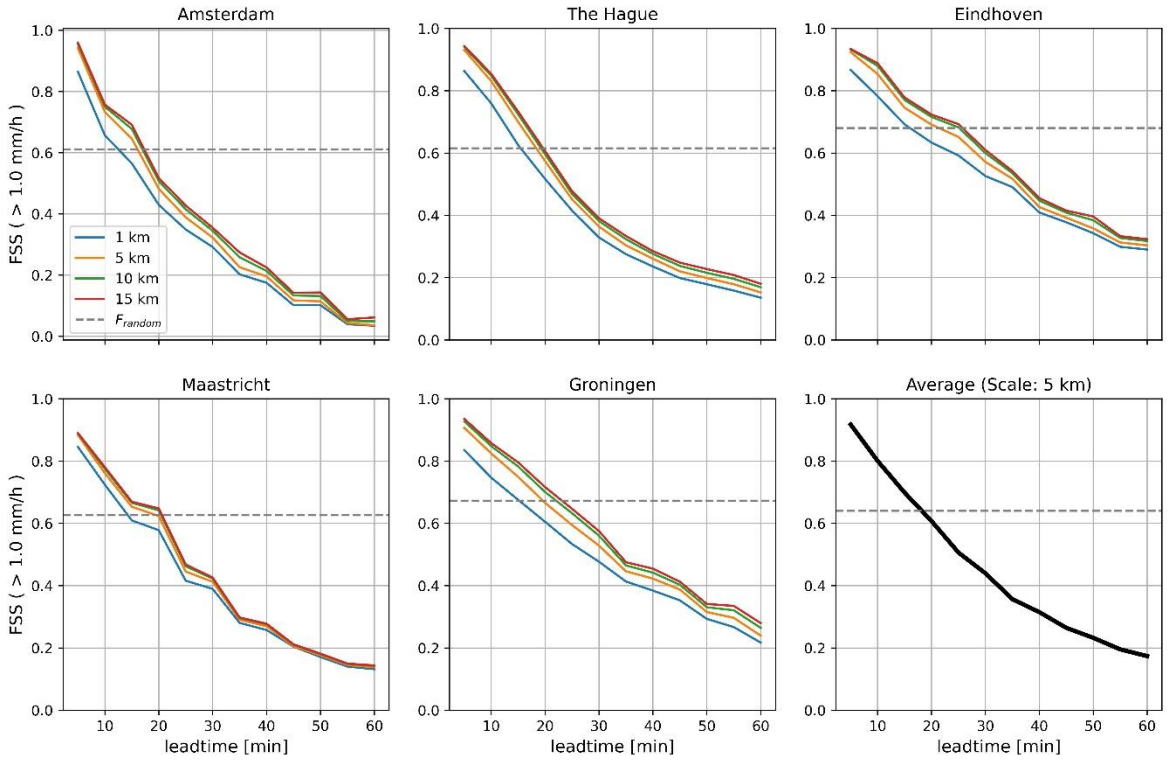


Figure 12. Fraction skill score for the 1-h events in each city. The bottom-right figure shows the average FSS of all the five cities at a length scale of 5 km. The gray dashed line shows the random fraction skill score in each city. FSS lower than the gray lines means that nowcasts are not skillful.

3.3.3. Nowcast reliability

After measuring the accuracy and skillful lead time of nowcasts, we continue to quantify the reliability. The ROC curve of each city for the 1-h and 24-h events is shown in Figure 13. First, we compared nowcast reliability for the 1-h events. When the threshold is 0.1 mm/hr, Maastricht and Eindhoven have higher reliability than the other cities. This is probably due to the higher rainfall intensity and spatial coverage in their eight selected events (as shown in Supplementary Figure 1), so it becomes easier for nowcasts to forecast rainfall higher than 0.1 mm/hr correctly. For bigger cities like Amsterdam and Groningen, parts of the cities are not covered by the rainfall events, so wrong forecast rainfall in those areas increases the false-alarm rate. Despite the difference in reliability on forecasting low rainfall intensity, as the threshold increases to 1 and 5 mm/hr, nowcast reliability becomes quite similar across the cities. Especially at 5 mm/hr, all nowcasts show very low reliability that is close to no skill. A similar situation is observed in the 24-h events.

Section 3.3.1 and this section show that the same nowcasts can show good skill in one verification metric but perform poorly in another one. Figure 11 shows that when the rainfall intensity is higher, nowcasts have higher error. Yet, we found that nowcasts show higher reliability (at least when the threshold is 0.1 mm/hr) in the cities with hihger observed rainfall intensity. Therefore, our results show that it is of importance to evaluate nowcasts from various aspects. The value of selecting a wide range of verification metrics and multi-metric analysis is further discussed in Section 4.3.

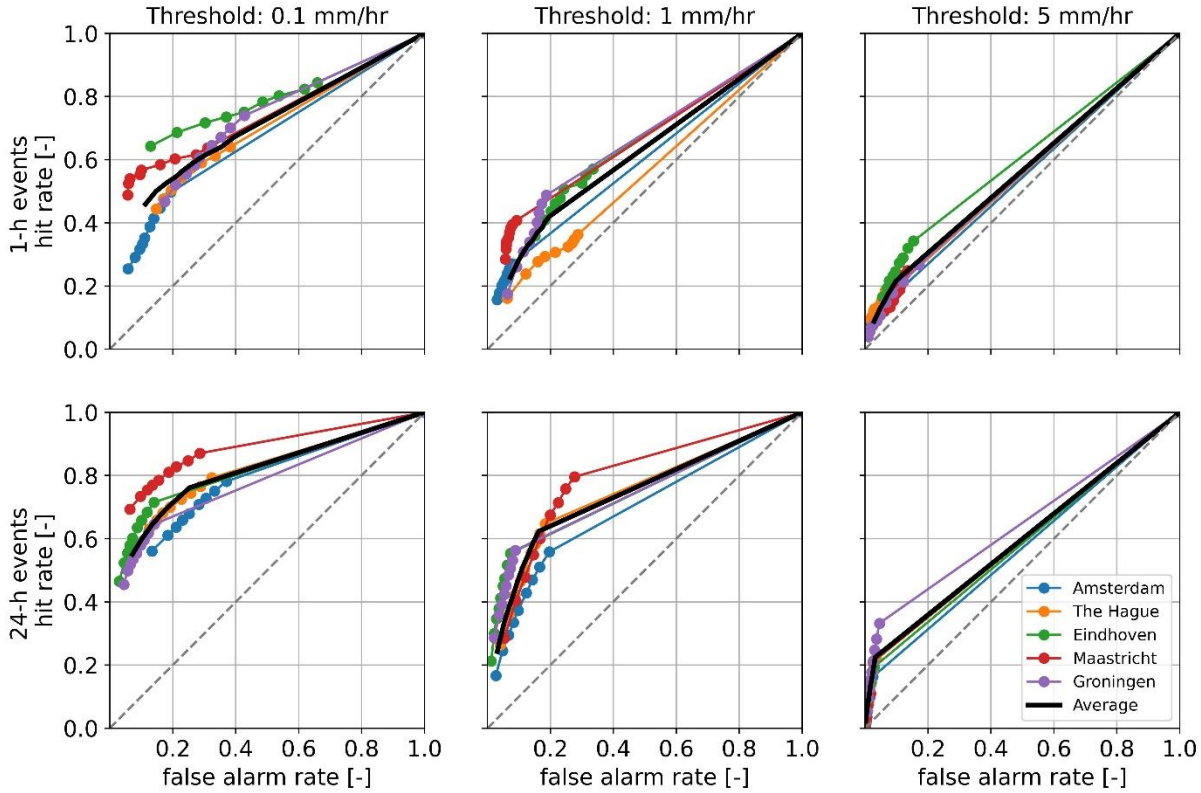


Figure 13. ROC curve of the 1-h (upper) and 24-h (below) events in each city at three thresholds (0.1, 1, and 5 mm/hr). Hit rate and false alarm rates are calculated at 30-minute lead time. The gray line means hit rate equals false alarm rate, below which nowcasts are not skillful. Higher above the line means higher reliability. There are nine dots on each line indicating the forecast probability of exceeding the threshold form 0.1 (rightmost) to 0.9 (leftmost).

To summarize the comparison of nowcast performance in the five cities, first, cities with higher rainfall intensity and spatial variation have lower nowcast accuracy. Second, reliability for forecasting low rainfall (0.1 to 1 mm/hr) is generally higher in Eindhoven and Maastricht probably because their higher rainfall intensity and spatial coverage cause higher hit rate. Nevertheless, reliability for

estimating higher rainfall intensity (5 mm/hr) is equivalently low in the cities. Finally, although difference is discovered in nowcast accuracy and reliability, skillful lead times are similar in the cities.

3.4. Dependence on rainfall characteristics

Section 3.3 shows that CRPS has a strong correlation to rainfall intensity and spatial variation. Following this finding, we further analyzed nowcast accuracy, skillful lead time, and reliability regarding different rainfall durations in the coming sections.

3.4.1. Nowcast accuracy

Average biases between nowcast rainfall and real-time QPE for 1-h and 24-h events are shown in Figure 14. In general, bias decreases with increasing lead time, because the longer ahead of the events, the more likely nowcasts underestimate or are incapable of predicting the events. For the 1-h events, biases of area-mean rainfall drop below 0.8 within 30-minute lead time for all cities. The largest underestimation happens in Amsterdam, where the bias of area-mean rainfall drops below 0.5 around lead time of 20 minutes. For the other cities, it takes longer, around 60 to 120 minutes, before the biases fall below 0.5. Comparing the average bias of nowcast and Eulerian persistence for the 1-h events, nowcasts reduce the overestimation of Eulerian persistence at short lead time. However, nowcasts show similar accuracy as Eulerian persistence at longer lead time above 50 minutes.

Despite the large underestimation of the 1-h events, nowcasts largely improve for the 24-h events. Figure 14 shows that nowcasts can estimate the total rainfall volume accurately (error within 20%) with a lead time up to 150 minutes for the 24-h events. Such performance also surpasses Eulerian persistence, which constantly overestimates rainfall intensity for the 24-h events. On the other hand, nowcasts sometimes still overestimate rainfall intensity for the 24-h events. Take the peak in Eindhoven for example, rainfall volume over the city is 1.4 times higher than the real-time QPE. The high fluctuation in the bias of rainfall is suspected to be caused by smaller observed rainfall intensity of the 24-h events than the 1-h events. Thus, small changes in their nowcast rainfall volumes can result in large changes in their bias.

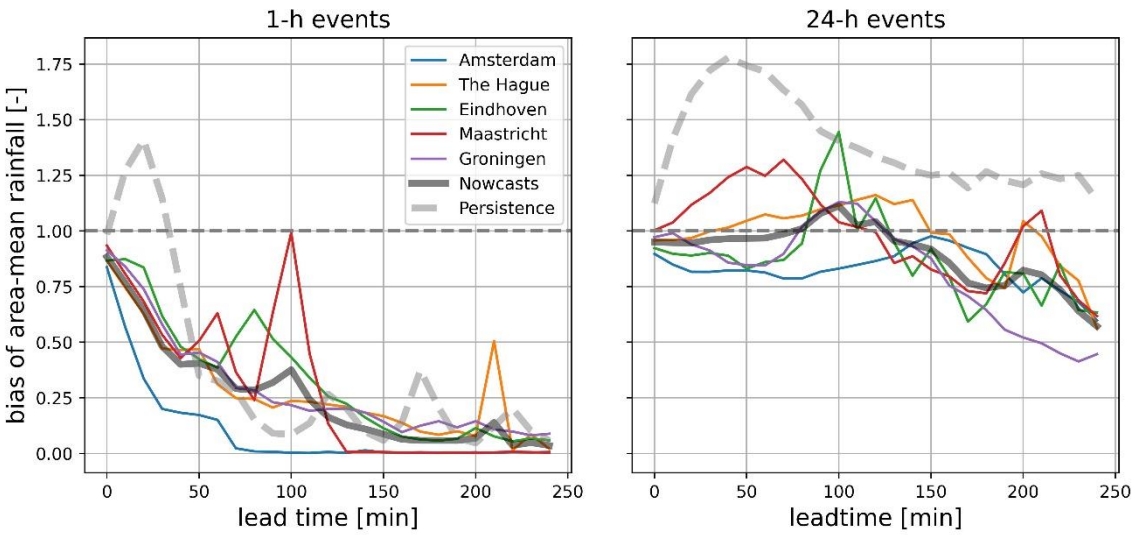


Figure 14. Bias of nowcasts in the cities. The gray horizontal dashed line y = 1 marks the same rainfall intensity between nowcast rainfall and real-time QPE. Bias is the ratio between nowcasts and real-time QPE. Bias larger than 1 means nowcasts overestimate rainfall, vice versa. The solid gray line and the dashed gray line are the average bias of nowcasts and Eulerian persistence respectively.

3.4.2. Nowcast skillful lead time

Figure 12 shows that the average skillful lead time of nowcasts for 1-h events at a length scale of 5 km is about 20 minutes. To compare the skillful lead time of 1-h and 24-h events, FSS of 24-h events is also calculated. However, the skillful lead time at the same length scale and same threshold as the 1-h events only slightly improves a few minutes, as shown in Supplementary Figure 7. Therefore, we use another method that applies Pearson correlation and decorrelation distance to determine the skillful lead time of forecasts in this section.

The average Pearson correlation coefficient between nowcast and observed rainfall in each city for the 1-h and 24-h events is shown in Figure 15. The difference between the cities is quite small. The cities have similar skillful lead times between 20 and 30 minutes for both event durations. The average skillful lead times for both event durations are very close: 20 minutes (1-h events) and 24 minutes (24-h events). Therefore, similar to the results from FSS, nowcast skillful lead time seems to be quite independent of the tested event duration. One more apparent difference is that the curves of the 24-h events are smoother than the curves of the 1-h events because they are averaged from more nowcasts.

Although the skillful lead time seems quite limited for large-scale early warning operation such as evacuating people and controlling water infrastructure, it is two folds the skillful lead time given by Eulerian persistence. The skillful lead time is compared with the other previous research and its application is discussed in Section 4.1.

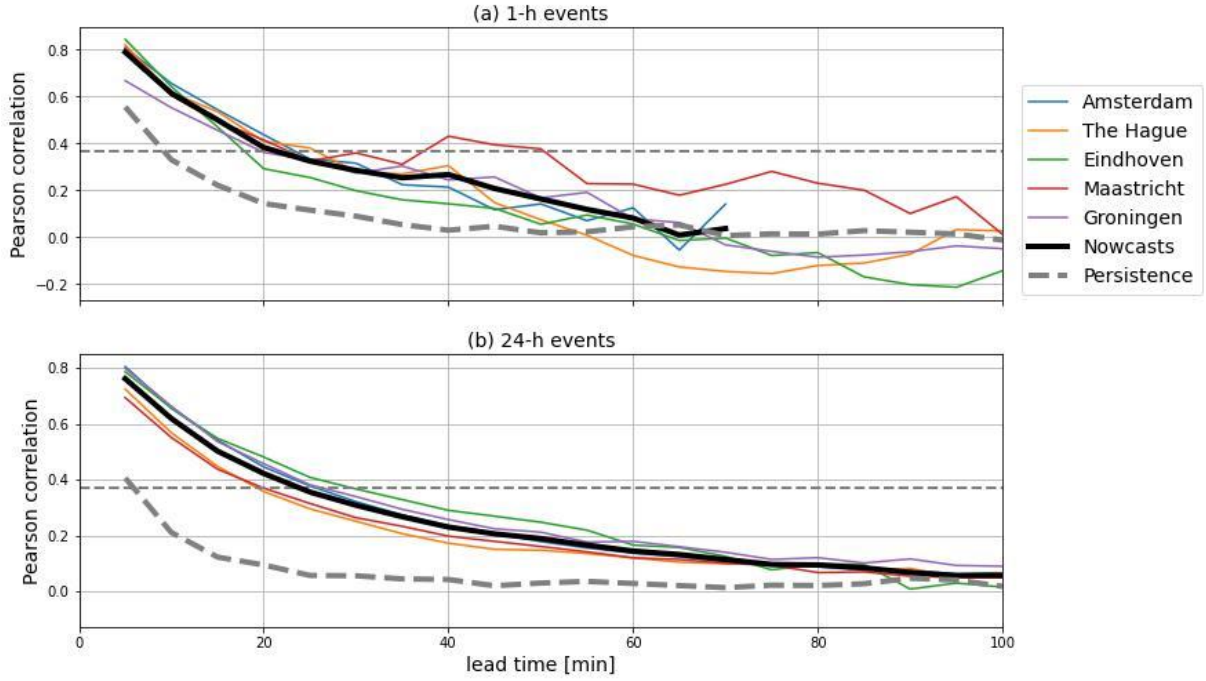


Figure 15. Pearson correlation coefficient averaged by the eight 1-h or 24-h events in each city. The gray dashed line marks a correlation of 0.37 (1/e) below which nowcasts are considered as no skills. The solid black line and the dashed gray line are the average Pearson correlation of nowcasts and Eulerian persistence respectively.

3.4.3. Nowcast reliability

Following the finding in Section 3.3.3 that nowcast reliability varies per city, we further compared the reliability for events of different durations. Figure 13 presents that the ROC curves from the 24-h events show higher reliability than the 1-h events at all three thresholds. This could be explained by the fact that the growth of short-lived rainfall is harder to be forecast by nowcasts. Localized convective rainfall can develop in tens of minutes (Mejsnar, et al., 2018; Germann, et al., 2006). Thus, very likely the small-scale and short-lived heavy rainfall is not present in the real-time QPE before the

nowcast issue time, so nowcast is not able to forecast it, as previously explained in Section 3.2. Conversely, more persistent events that have higher autocorrelation allow the nowcasting algorithm to forecast rainfall more correctly from previous time steps. Despite the generally higher reliability of nowcasts for the 24-h events, the reliability for forecasting rainfall higher than 5 mm/hr still drops to very close to random forecast. Therefore, estimating high rainfall intensity is still a challenge.

Summarizing the results from Section 3.4.1 to Section 3.4.3, we found that nowcasts improve the estimation of total rainfall volume for 24-h events compared to Eulerian persistence, whereas the improvement is not clear in the 1-h events. In addition, the reliability of nowcasts is lower for the 1-h events. Skillful lead time does not differ between 1-h and 24-h event duration in the selected cities.

3.5. Dependence on city sizes

After analyzing nowcast skills in different cities and for different rainfall characteristics, we analyzed its variation with area size. Nowcast accuracy and skillful lead time in the defined subareas of various sizes in and around the five cities are quantified in this section.

3.5.1. Nowcast accuracy

Similar to the structures in the previous sections, we started with comparing nowcast accuracy (CRPS) in the subareas. The correlation between area size and CRPS is confirmed in Figure 16 and Figure 17. They show a clear trend that smaller areas tend to have larger CRPS, meaning lower accuracy. Figure 16 shows that when enlarging the considered area from 16 to 100 km^2, CRPS reduces in almost all five cities. This nonlinear relationship between size and CRPS becomes clearer in Figure 17 which plots the size of subareas from 4 to 900 km^2 in and around Groningen. It shows that for both event duration, CRPS in Groningen reduces as the area expands, even across a larger magnitude of area size. Especially when the size reduces from 100 to 4 km^2, the error enlarges rapidly. Thus, the spatial scale under 100 km^2 might be a critical obstacle for the STEPS method used in this study.

As mentioned in Section 2.1, this result is most likely because smaller areas are more sensitive to the displacement of forecast rainfall, so the error is higher. The case study in Figure 9 serves as a good example to illustrate the difficulty to forecast rainfall with the right location and intensity for small areas. Thus, improving the accuracy of nowcasts at a finer spatial resolution is necessary to improve nowcast accuracy in small cities.

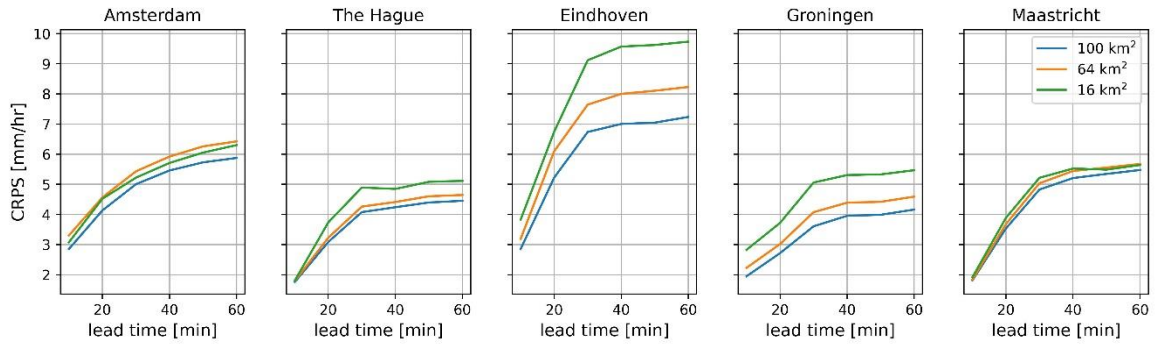


Figure 16. Average CRPS of the 1-h events in the subareas in Amsterdam, the Hague, and Eindhoven against lead time of nowcasts

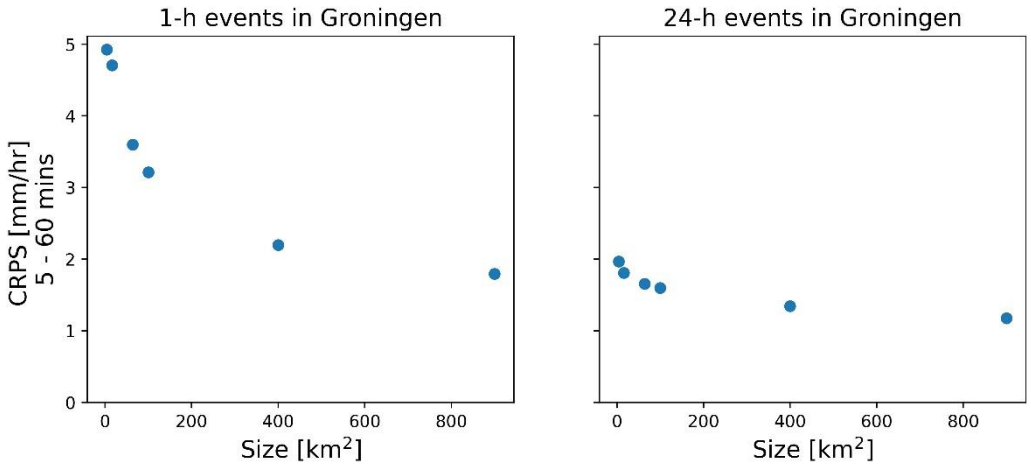


Figure 17. Average CRPS (lead time of between 5 to 60 minutes) of the eight 1-h and 24-h events in Groningen against verification areas from 4 to 900 km²

3.5.2. Nowcast skillful lead time

Skillful lead time across different area sizes is also evaluated. As previously shown in Figure 15, Pearson correlation and skillful lead time are very similar (around 20 minutes) in the five cities with radar areas from 67 to 211 km^2. Similar results are found in the subareas (16 to 100 km^2) in the cities, as shown in Supplementary Figure 8. Thus, we test the Pearson correlation across a larger range of areas (4 to 900 km^2).

Figure 18 shows the Pearson correlation between nowcasts and observation for the 1-h and 24-h events in the subareas of 4 to 900 km^2 in Groningen. We found that skillful lead time does change with the size of the considered area, but with less sensitivity (compared to the sensitivity of CRPS, bias, and reliability shown previously). That is, Pearson correlation only changes when the difference in size is significant. Roughly three pairs of similar Pearson correlation coefficients are shown in both plots. Namely, the Pearson correlation between 4 and 16 km^2, 64 and 100 km^2, and 400 and 900 km^2 are close and there is visible difference between the three pairs. Larger subareas (e.g., 400 and 900 km^2) have higher Pearson correlation and thus longer lead time than smaller subareas for both event durations. In addition, the improvement of lead time with increasing size is slightly larger for the 24-h events than for the 1-h events. The skillful lead time in the subarea of 900 km^2 prolongs to 30 minutes, which nearly doubles the skillful lead time in the subarea of 4 km^2. However, the skillful lead time for 900 km^2 for 1-h events is still limited to 20 minutes although it is better than the counterpart for smaller areas. Even though additional 15 minutes in lead time for a 24-h event is not such a pronounced improvement for real-world application, this analysis still shows the relationship between size of the verification area and skillful lead time.

In summary, we found that nowcast in larger areas show lower CRPS and higher spatial correlation to the observed rainfall. Thus, improving spatial accuracy to a finer scale is essential to apply nowcast in cities under 100 km^2.

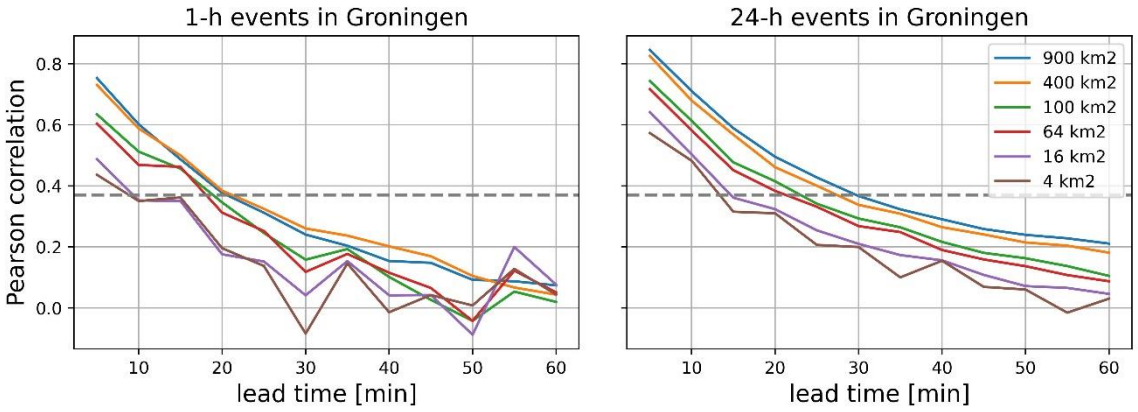


Figure 18. Pearson correlation between nowcasts and real-time QPE for the 1-h and 24-h events in Groningen

3.6. QPE error and nowcast dependence on the QPE products

So far, the nowcasts in the study were run with real-time QPE as input. Besides, the verification was also performed with the same dataset. However, as explained in Section 2.2, the real-time QPE can deviate much from actual rainfall under certain circumstances. Therefore, a good verification result with the real-time radar rainfall observation (e.g., low CRPS, high maximum skillful lead time, etc.) does not guarantee that the nowcasts are capable of forecasting actual rainfall well. So, we evaluated the difference among the three QPEs and their respective nowcasts in the section. First, two cases of the worst performing real-time QPE accuracy were shown. Second, the three radar products during the events that occurred before 2020 were analyzed. Third, CARROTS and climatological QPE were also run as input to nowcasts and compared with the nowcasts run with real-time QPE. All the output nowcasts were verified with the climatological QPE to evaluate their "real" skills, where we assume that the climatological product is the true rainfall. Note that only the 1-h extreme events before 2020 (23 events) were run with climatological and CARROTS QPE because the format of the climatological radar products changed at the end of 2019.

3.6.1. Two cases with the largest QPE bias

To show the large difference between the QPE products, we selected two events with the strongest QPE bias. The two largest deviations of QPE between climatological and real-time QPE occurred in Maastricht during the 1-h event started at 15:20, July 14, 2010, and in Eindhoven during the 1-h event started at 08:45, July 27, 2013. The comparison between the accumulated rainfall using the three radar products during the events is shown in Figure 19. Similar to the case study in Section 3.2, real-time and CARROTS QPEs are much lower than the climatological QPE. The real-time and CARROTS QPE in the event in Maastricht only estimate 12% and 27% of the total climatological rainfall accumulation. This can be caused by the large distance between Maastricht to the weather radars (as shown in Figure 1). Although Eindhoven is closer to the radars, its real-time and CARROTS QPEs also show bias of 28% and 51% during the event probably due to the rapid changes in raindrop size distribution in heavy convective rainfall (Schleiss, et al., 2020).

For the same cases, results of nowcast 1-h rainfall accumulation issued at the onset of the events are shown in Figure 20. The figure clearly shows that the spatial extent and intensity of extreme rainfall given by the nowcasts of climatological rainfall are substantially larger than those from the other two datasets. Although CARROTS and real-time observation capture quite well the rainfall with moderate intensity, it misses the area where the highest rainfall is observed by only the climatological QPE, because our nowcasting setup only advects rainfall shown in the input QPE. Thus, different radar QPEs lead to different nowcast rainfall intensities because of the QPE bias.

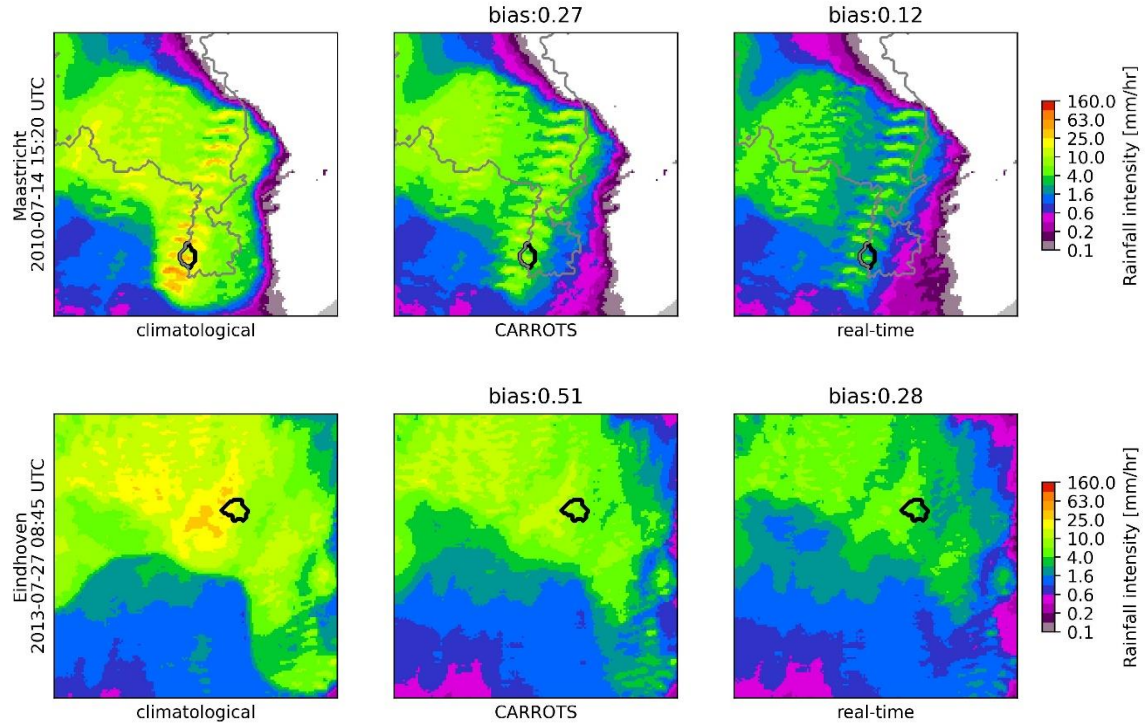


Figure 19. Accumulated QPEs during the 2 most QPE-biased events in Maastricht and Eindhoven. Bias is the ratio between accumulated CARROTS or real-time QPE and the climatological QPE within the city border. Lower bias means larger underestimation. The black borders indicate the areas of Maastricht (upper) and Eindhoven (below). The gray border shows the radar domain within the Netherlands. The shift of rainfall pattern is visible in Maastricht, so applying advection correction may adjust the QPE accumulation, but it is not expected to adjust much the QPE bias.

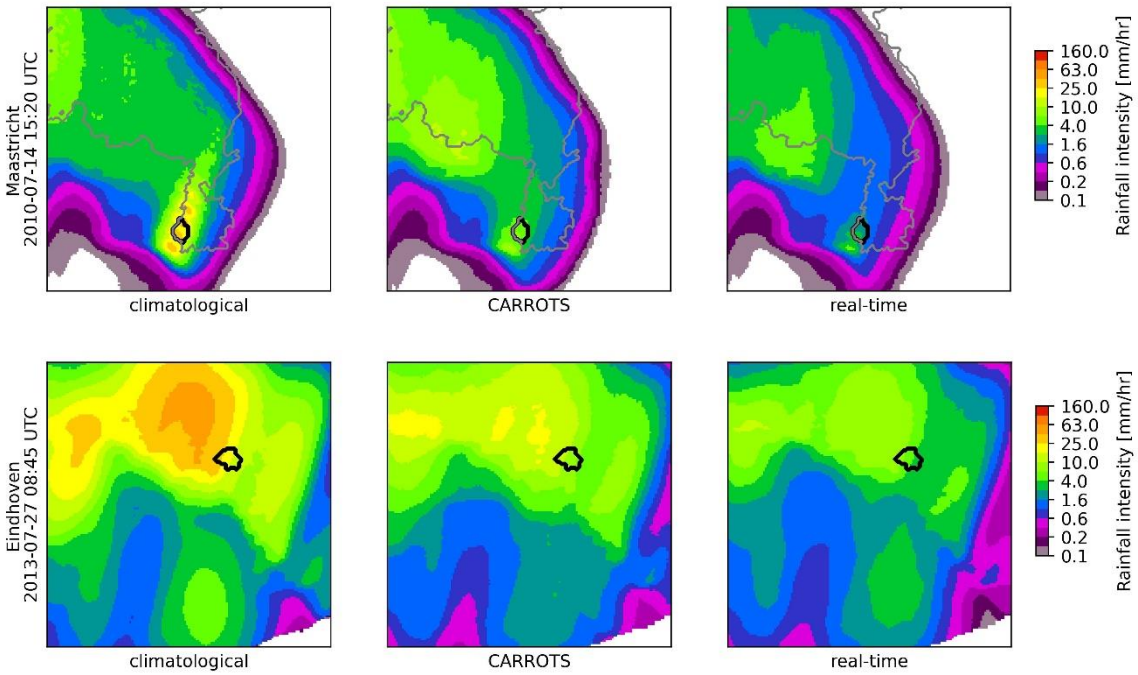


Figure 20. Nowcast 1-h rainfall intensity around Maastricht and Eindhoven issued at onset of the events during the 2 most QPE-biased events. Nowcasts are run with three different QPE products as input. The black borders indicate the areas of Maastricht (upper) and Eindhoven (below). The gray border shows the radar domain within the Netherlands. Shown is the ensemble mean of 20 ensemble members.

3.6.2. Comparison between the QPEs

After showing the two individual events, the average QPE bias during the 1-h events before 2020 in each city is shown in Figure 21. It shows that real-time and CARROTS QPEs are generally lower than climatological QPE. The underestimation differs per city. The smallest QPE bias (largest underestimation) exists in Maastricht. Thus, the accuracy of real-time nowcasts in Maastricht is compromised the most. Eindhoven, despite having the highest extreme rainfall intensity among the cities, which is a common cause for underestimation of QPE, its bias is not that high. This is probably due to its short distance to the KNMI radars, so better QPE quality is assured. On the other hand, the QPE bias of Groningen is the lowest although it is further from the KNMI radars. These results are similar to the finding in Figure 3 in Imhoff et al. (2021). Noted that 24-h events have higher QPE bias in all cities because the rainfall intensity is lower so the ratio between real-time and climatological QPE becomes larger, as shown in Supplementary Figure 9.

Figure 21 also proves that CARROTS improves the real-time QPE in terms of rainfall volume. The bias of real-time QPE improves from as low as 0.31 (69% relative error) to between 0.64 and 1.30 (<36% relative error) after adjusting with CARROTS. In certain situations, the bias even becomes very close to 1, like for the 24-h events in Eindhoven. Adjusting the bias to closer to 1 shows that CARROTS rainfall become closer to climatological rainfall. Overall, CARROTS QPE is on average 77% (1-h events) and 92% (24-h events) hihger than the real-time QPE. Hence, Using CARROTS rainfall to run nowcasts should yield better results than using just real-time rainfall.

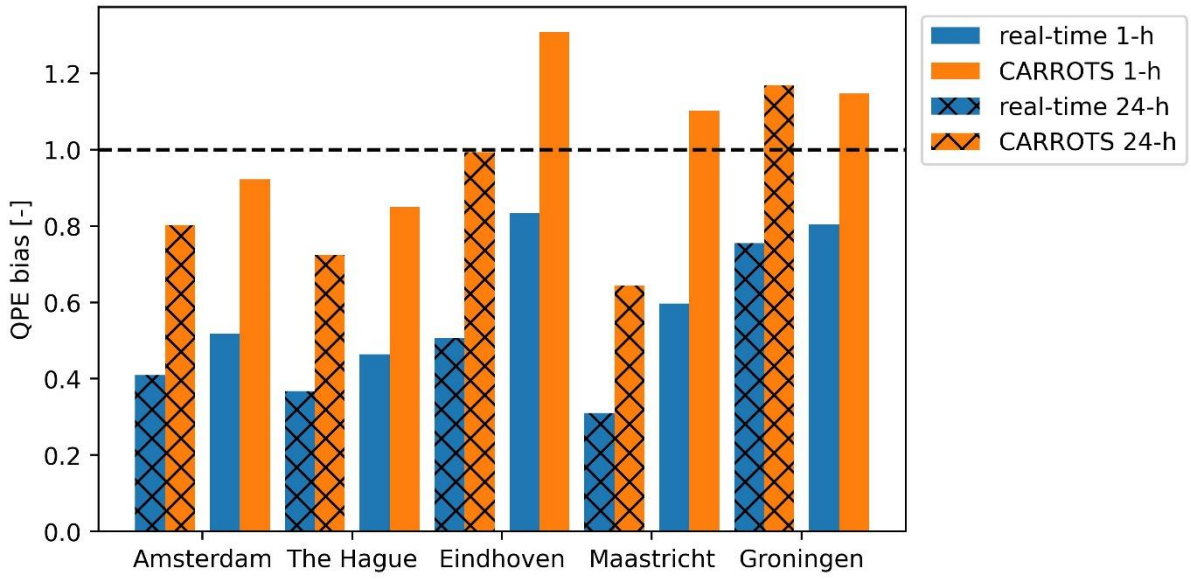


Figure 21. Average QPE bias between climatological and the real-time radar measurement before and after adjusted with CARROTS in the cities. Only considering the 23 1-h events happening before 2019. Value below 1 means real-time or CARROTS QPE is lower than climatological QPE, vice versa.

3.6.3. Nowcasts with different QPEs

Due to the difference in QPEs, nowcast results with different QPE products are expected to show difference as well. Figure 22 shows the bias of area-mean rainfall between nowcasts using three different radar products and climatological observation in each city. It shows that nowcasts using real-time QPE give the largest underestimation in all the cities. This is a direct consequence of the bias existing in the real-time QPE, as shown in Figure 21. Thus, a good verification performance using real-time QPE in the section from 3.2 to 3.4 should be judged meticulously.

CARROTS increase the nowcast rainfall volume by 70% compared to nowcasts given by real-time QPE and show a bias closer to 1. The bias given by CARROTS and climatological QPE is similar, showing

that CARROTS adjust the rainfall quite correctly in the cities. Yet, the CARROTS nowcasts sometimes overestimate the rainfall, particularly in Eindhoven and Maastricht at short lead times. This implies that the CARROTS factors are slightly too high during these events. Such phenomenon is also observed in Figure 21 which shows that CARROTS QPE is higher than climatological QPE (QPE bias larger than 1) in some cities.

Despite this improvement, the underestimation of rainfall still enlarges as lead time increases regardless of the QPEs. In all the cities, the decline in bias is sharper at the beginning indicating that underestimation of rainfall volume already occurs when lead time is short. The short skillful lead time is also reflected in the Pearson correlation between nowcast and climatological observed rainfall. As shown in Figure 23, the maximum skillful lead time for 1-h events is around 20 minutes no matter which product is used as input. This reason could be that the rainfall spatial fields in the observation are similar, as can be seen in the 2 events in Figure 19. So, although CARROTS adjusts the nowcast rainfall volume, the nowcast rainfall still shows low spatial correspondence to the climatological QPE for lead times longer than 20 minutes. Thus, from the perspective of timely warning, skillful lead time cannot be prolonged by using different QPEs in the Netherlands.

In summary, CARROTS can adjust the underestimation in the real-time QPE and the nowcasts to a certain degree. Yet, although CARROTS increase both the QPE and nowcast rainfall volume by more than 70%, underestimation of nowcast rainfall still enlarges as lead time increases. In addition, CARROTS do not improve the spatial correlation of nowcasts with climatological observation. Hence, improving nowcasting models is required to forecast extreme rainfall volume in small cities accurately with longer lead time. Possible solutions are further discussed in Section 4.5.1.

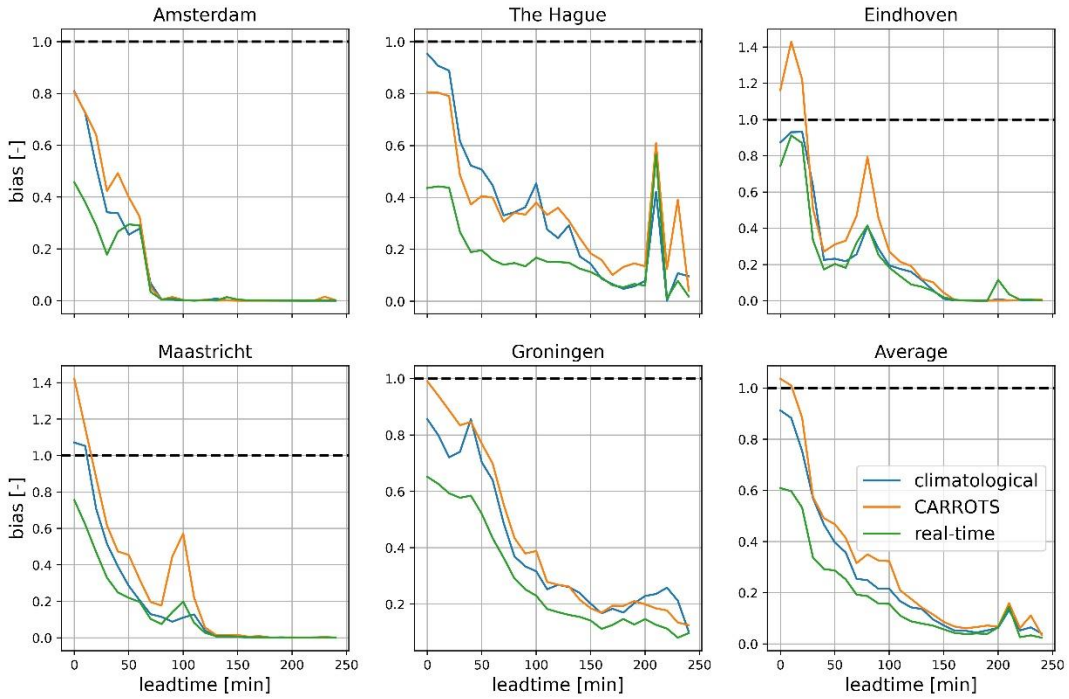


Figure 22. Bias of the nowcasts run with climatological, CARROTS, and real-time radar rainfall from the 23 1-h events before 2020. The bias is resampled to 10 minutes. The bias is calculated between nowcast and observed climatological QPE. Bias equals 1 meaning nowcast rainfall volume equals climatological observations. Bias higher than 1 indicating nowcast overestimates rainfall and vice versa.

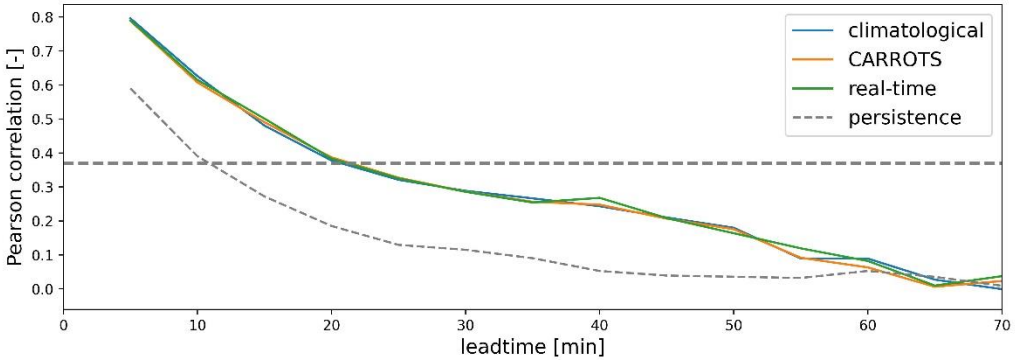


Figure 23. Pearson correlation of the nowcasts run with climatological, CARROTS, and real-time rainfall verified with the climatological QPE for the 23 1-h events before 2020.

4. Discussion

Nowcast skills were systematically assessed for different cities, rainfall characteristics, area sizes, and QPE inputs from Section 3.3 to Section 3.6. We established the link of nowcast performance with the abovementioned variables. In this section, we first compare the results in the study to previous research (Section 4.1) and evaluate the chosen event selection criteria and verification metrics in Sections 4.2 and 4.3 respectively. Then, we explain the application of the results for operational radar rainfall nowcasting in Section 4.4 and summarized future improvement possibilities in Section 4.5.

4.1. Relation to previous work

We compared our results to previous studies on nowcast performance in Europe and 2 studies in the Netherlands. First, a large-sample analysis of nowcast performance in 12 catchments ranging from 6.5 to 957 km^2 in the Netherlands (Imhoff, et al., 2020). Second, a study on the hydrological application of probabilistic nowcast in three neighboring catchments of 6.5, 40, and 957 km^2 in East Netherlands (Heuvelink, et al., 2020).

Some similar results found in the previous study are also observed in our research. One such finding is the dependence of CRPS on rainfall intensity. CRPS is found to have a strong positive correlation to rainfall intensity in this study and Imhoff et al. (2020) identified that CRPS is higher in summers when average rainfall intensity is the highest among the seasons. The positive correlation between the mean absolute error (MAE) of deterministic nowcasts and rainfall intensity is found in a follow-up study (Imhoff, et al., 2022). Another similarity is that the underestimation of rainfall given by nowcasts enlarges as lead time increases. Heuvelink et al. (2020) also showed that the relative error of rainfall volume between nowcasts and observation rises to more than 50% after lead time of 2 hours for events lasting for several days. Our study further shows that the underestimation is more severe for the 1-h events because the relative error is up to 50% already at a lead time of 30 minutes, as shown in Figure 14 (b).

Similar skillful lead time for heavy and short events are found in our research too. Skillful lead time determined by Pearson correlation is around 20 to 30 minutes in this study, as shown in Figure 15. Such lead time is close to the finding in Imhoff et al. (2020) from the analysis of 384 1-h events in the Netherlands. Similarly, nowcasts for convective rainfall over 5.0 mm/hr are only reliable until 30-min lead time in Belgium (Foresti, et al., 2016) and skillful lead times of both convective and stratiform rainfall are shorter than 20 minutes in Germany (Shehu & Haberlandt, 2021). Therefore, radar-based nowcasts seem to show a general incapaibility to forecast intensive evnets that occur in a short time.

On the other hand, the skillful lead time for the 24-h events, which is also around 20 to 30 minutes in this study, is much shorter than in the previous research. For events longer than 24-h hours, skillful lead times are 90 to 120 minutes in Imhoff et al. (2020) and 25–170 minutes in Heuvelink et al. (2020). In addition, the skillful lead time in this study does not vary much with area sizes, which is the behavior shown in both Imhoff et al. (2020) and Heuvelink et al. (2020).

There are two main reasons causing the shorter skillful lead time. First, this study selects extreme events only based on the highest rainfall accumulation in any grid cell in the city while half of the events in Imhoff et al. (2020) are chosen based on catchment-average rainfall amounts. So, the events in this study might have higher spatial variability or parts of the cities do not observe rainfall during the events (as shown in Supplementary Figure 1) which induces a lower Pearson correlation. Second, the cities in the study on average are smaller than the catchments in their studies. As shown in Section 3.2 and Section 3.5, verification of nowcasts in smaller areas is more sensitive to displacement of forecast rainfall, so their Pearson correlation is lower. These reasons lead to the shorter skillful lead time in this study.

4.2. Event selection

Different event selection methods can lead to different nowcast verification results. Each selection criterion has its limitations. As described in Section 4.1 and shown in Supplementary Figure 1, the 1-h events in this study generally only have high precipitation in certain parts of the cities. Thus, the total rainfall amount in the whole city is less extreme. From the perspective of sending early warnings for the municipalities, extreme events that cause high rainfall amount to the whole municipalities, instead of only certain regions, might be the events that should be further studied. Thus, future study can choose extreme events according to total rainfall accumulation and analyze their nowcast performance.

Also, the study does not choose events based on high hydrological response in the cities. Choosing events that cause high hydrological peaks is the method used in Heuvelink et al. (2020). Such selection can apply more directly to study the performance of nowcasts on estimating hydrological response. Since the hydrological response is usually much quicker in the city than in the rural catchments studied in Heuvelink et al. (2020), this will be an important area to continue this study on.

Another limitation of the research is that 74 events out of the 80 selected events happened between May and September, which are summer months in the Netherlands. Monthly distribution of the events is shown in Supplementary Figure 10. This selection might restrict the applicability of the research given that real-time QPE in winters in the Netherlands often underestimates stratiform rainfall due to sampling over the melting layer (Imhoff, et al., 2021). Such phenomenon might be another uncertain factor for nowcasting in winters.

4.3. Performance metrics

The study comprises the metrics from various perspectives to evaluate nowcast skill. We used bias to compare observed and nowcast rainfall volume, calculated CRPS to show the accuracy, applied Pearson correlation and fraction skill score to measure skillful lead time, and plotted ROC curves to quantify reliability. Therefore, the study offers a comprehensive evaluation of nowcast performance of extreme rainfall events.

However, there are some limitations to each metric. For instance, FSS uses 1 mm/hr as an arbitrary threshold, so other thresholds may lead to different results. Setting the threshold is to assess nowcast skills in forcasting rainfall above certain intensity, so it can be adjusted based on the aim of research. On the other hand, Pearson correlation does not inclue thresholds in verification, so the skillful lead time derived from it is for the whole range of rainfall intensity. In the case of evaluating nowcast skill for high rainfall intensity, FSS might be a better option. Besides, although this study mostly applied Pearson correlation coefficient to calculate the skillful lead time, modified Pearson correlation coefficient and Spearman correlation coefficient could also be used (Germann & Zawadzki, 2002; Mejsnar, et al., 2018). Different choice of the spatial correlation coefficient would cause different skillful lead time (Mejsnar, et al., 2018). Therefore, the skillful lead time derived in the study should not be considered as the only standard.

An important finding from the various metrics is that good performance in one metric does not guarantee (usually not) equivalently good performance in another metric. As shown in Figure 10 and Figure 15, Amsterdam and Groningen show better results than the other three cities in CRPS of 1-h events; yet their maximum skillful lead time is the same as the others. Also, even though Maastricht and Eindhoven have higher CRPS (worse performance) (Figure 10), their bias is closer to 1 (better performance) (Figure 14). Another good result is shown in the higher hit rate of Eindhoven and Maastricht than the other cities at a low threshold, as shown in Figure 13. All these results are mainly due to the hihger rainfall intensities of the rainfall events in these two cities.

Overall, we presented the value of conducting analysis with multiple performance metrics. When the goal is to forecast heavy rainfall, the metrics that incorporates threholds might be better choices. When applying nowcasts to early warning, authorities should evaluate nowcast skills by different metrics and make decisions based on the needs of the municipality.

4.4. Operational radar rainfall nowcasting

Despite the shorter lead time identified in the study, nowcasts can still provide effective early warning insights for cities in the Netherlands. For instance, nowcasts of short lead time can be used to forecast small-scale hazards like pluvial flooding, flash floods, and landslides that are induced by local rainfall extremes (Alfieri, et al., 2012). The short lead time might be too late to evacuate people, but it can inform people to adopt the shelter-in-place strategy (Haynes, et al., 2009). Nowcasting can identify areas of hazardous rainfall intensity and warn the community by disseminating real-time warning information on webpages, mobile applications, and social media (Acosta-Coll, et al., 2018). The warning message can be sent to the local authorities to better prepare for emergency actions. Therefore, nowcasting can still be a timely and useful early warning tool for Dutch cities.

From Section 3.2 to Section 3.6, we summarized several challenges that the nowcasting model struggles with. Similar limitations of operational rainfall nowcasting are also found in other regions in the world. For instance, Short-range Warning of Intense Rainstorms in Localized Systems (SWIRLS), the nowcasting system in Hong Kong also fails to forecast well the growth and decay of rainfall at longer lead time (Woo & Wong, 2017). Likewise, PIAF, the nowcasting system in France (Moisselin, et al., 2019), shows that the quality of nowcasts in France deteriorates quickly and becomes lower than NWP within a lead time of 75 to 90 minutes due to the same reason (Lovat, et al., 2022). Radar nowcasts in the Czech Republic show that local convective storms are poorly forecast and the skillful lead time varies up to 40% depending on precipitation characteristics (Mejsnar, et al., 2018).

To reduce QPE error in operational radar rainfall nowcasting, CARROTS are applied. We found that it corrects the bias in QPE and nowcast rainfall volume, but it does not change much the spatial distribution of forecast rainfall. We found that the average CARROTS factor during the 1-h events before 2020 are 1.73 [-], but the average spatial standard deviation within the cities is only 0.04 [-] (2% of the average CARROTS factor). Thus, although the CARROTS factors are at high resolution of 1 km^2, they are basically the same for the whole cities (radar domains from 67 to 211 km^2 in this study). As a result, the rainfall intensity in a city is adjusted by almost the same CARROTS factor, so the rainfall spatial structure remains mostly unchanged.

Given the common challenges concluded in the above studies, and the fact that even using the best QPE product available in the Netherlands (the climatological QPE) does not extend the skillful lead time, more improvements should be emphasized on the STEPS procedures. Thus, future possibilities to improve nowcasting technique is detailed in the next section.

4.5. Future perspectives

Some innovative nowcasting procedures outlined below have been proposed to try to push the limits of operational rainfall nowcasting. However, it is still not clear whether they will lead to dramatic improvements during strong and convective events which are intrinsically chaotic. In addition, more analysis that can supplement this research is proposed.

4.5.1. Recent development in nowcasting techniques

To counter the difficulty of tracking small rainfall fields, some research proposed new tracking methods. For instance, nowcasts using a multi-scale tracking scheme showed more consistent rainfall distribution with the observation regarding location and intensity of heavy precipitation up to 1-h lead time (Wang, et al., 2013). NowPrecip developed by MeteoSwiss combines optical flow and

localized architecture to address distinct rainfall patterns in complex terrains, such as alpine regions (Sideris, et al., 2020). Such techniques might adjust the insufficiency in the Lucas-Kanade optical flow and Semi-Lagrangian advection method used in our study.

Due to the difficulty to parameterize the non-linear growth and decay of extreme rainfall, more recent research starts to integrate machine learning techniques into extreme precipitation nowcasting. For example, generative neural networks trained with three years of radar data outperform traditional optical flow-based nowcast (M. & L., 2020). Another machine learning model MSDM combines optical flow, random forest, and convolutional neural network and shows improved nowcast performance in various perspectives (Li, et al., 2021). More recently, nowcasts using a deep generative model manages to show accurate nowcasts up to a lead time of 90 minutes and with higher rainfall spatial precision (Ravuri, et al., 2021). Such development can improve both rainfall bias and spatial accuracy. Lastly, NowCasting-Nets produces nowcasts by convolutional deep neural networks which outcompete forecasts given by classical machine-learning approaches, optical-flow method, and Eulerian persistence (Ehsani, et al., 2022). With these findings, machine learning methods are proved to be valuable tools in extreme rainfall nowcasting and should be considered in future work on enhancing nowcasting skills in Dutch cities.

Blending numerical weather prediction (NWP) with nowcasting techniques is also proven to improve nowcasting skills. One study combines high-resolution NWP ensembles with extrapolation-based nowcasting to provide probabilistic precipitation forecasts that outperform traditional nowcasts at all lead times (0-8 hr) (Kober, et al., 2012). Previously mentioned Nowprecip also incorporates local rainfall growth and decay given by NWP into the computation of nowcasts to capture rainfall formed by orographic lift (Sideris, et al., 2020). With the significantly improved efficiency in computation and resolution of NWP in the last decade, blending NWP in the nowcasting process has a strong potential to advance nowcast skills (Sun, et al., 2014; Poletti, et al., 2019). Thus, using the Pysteps procedure that blends QPE extrapolation and NWP might be a good next step for advancing nowcasting skills. (Bowler, et al., 2006; Seed, et al., 2013)

Another potential method to better simulate rainfall dynamics is 3-dimensional nowcasting with the use of vertically profiling radars. 3-dimensional weather radar can measure scan temperature, humidity, and wind in 3-D (Steinheimer & Haiden, 2007) and identify extreme convective events like thunderstorms with a lifetime below 60 minutes (Yoshikawa, et al., 2012; Ushio, et al., 2015). Because it better captures the vertical motion of rainfall convection, research showed that extrapolating 3-D radar reflectivity outperformed 2-D nowcasting (Otsuka, et al., 2016; Sun, et al., 2022). Furthermore, combing a convolutional neural network and 3-D radar reflectivity may further improve nowcast reliability (Kim, et al., 2021).

Finally, the research often used the mean of ensemble members and the mean of nowcasts at different issue times as final products to verify the observation. Such methods are easy to implement but tend to miss the extreme and rapid development of short-lived rainfall. Thus, some studies suggest combining ensembles using weighted average method and giving more recent nowcasts higher weights. (W. & C., 2019). So, future study can test the performance of integrating weights to ensembles and nowcasts into STEPS.

4.5.2. Analysis of urban feature on nowcast skills

The study shows that nowcast performance depends both on rainfall and urban characteristics, so it is sometimes a complicated task to determine its dependence on one of the two factors without taking the other one into account. For instance, nowcast reliability depends on both cities and rainfall characteristics, as shown in Figure 13, which makes it hard to determine the contribution of each factor. Normalizing nowcast error with rainfall intensity is a proven method to detangle the effects of

rainfall characteristics, as shown in Supplementary Figure 5 and Supplementary Figure 6. To control the city sizes and shapes, defining subareas of the same shape and size for each city is a good approach. Apart from these methods, future research can choose similar rainfall events for the cities and compare nowcast performance among them. In this way, the dependence on rainfall characteristics is controlled and the influence of urban features can be better determined.

This research discusses nowcast skills in different cities and areas of different sizes. Another potential urban feature that may change nowcast skills is surface roughness. The five cities in the study have different topography and land use which might impose additional uncertainty to nowcasts. Urban roughness can create flow anomalies by changing the intensity and position of moisture convergence (Yang, et al., 2021). Urban heat and aerosols induced by cities can also modify rainfall in and around the cities (Sarangi, et al., 2018; Schmid & Niyogi, 2017; Shepherd, 2005). A study analyzed over 1000 research on the effect of urbanization on rainfall and found that precipitation is 16% higher above a city (Liu & Niyogi, 2019). Furthermore, city shapes are also found to have notable influences on rainfall accumulation (Zhang, et al., 2022). Such phenomenon adds higher difficulty to precipitation nowcasting around urban areas than in rural regions. Therefore, future analysis can study the dependence of nowcast skills on different topography and urbanized ratios of the areas.

5. Conclusion

This research used short-term ensemble prediction system (STEPS) from Pysteps and 80 extreme rainfall events between 2008 and 2021 to analyze probabilistic nowcasting skills for Dutch cities. These 80 events consisted of the most extreme events in 5 main Dutch cities (Amsterdam, the Hague, Groningen, Maastricht, and Eindhoven) for 2 durations (1 hour and 24 hours). In addition, subareas of various sizes from 4 to 900 km^2 were defined in each city, and three QPE products (real-time, CARROTS, and climatological) were used. Nowcasting skills were verified with diverse metrics including continuous ranked probabilistic score, Pearson correlation, fraction skill score, bias, and receiver operating characteristic curve. Dependence of nowcast skills with respect to the cities, rainfall characteristics, area sizes, and QPE products were studied.

Nowcast accuracy and reliability were different in the five cities mainly due to the different rainfall characteristic and size per city. A strong positive correlation was found between rainfall intensity and continuous ranked probabilistic scores. 1-h events in Eindhoven and Maastricht had higher continuous rank probabilistic scores (indicating higher errors) because the rainfall intensity of their events was higher and the cities were smaller. Yet, the reliability of nowcasts at the low threshold (0.1 mm/hr) was higher in Eindhoven and Maastricht.

Nowcasting performance also depended on rainfall durations. Nowcast rainfall underestimation was more severe for 1-h events than for 24-h events. Nowcasts estimated the total rainfall volume accurately (error within 20%) with a lead time up to 2.5 hours for the 24-h events, but only 30 minutes for the 1-h events. Also, the reliability of nowcasts was higher for 24-h events than for 1-h events. Reliability is also higher when the threshold of interest is lower. When the threshold is 5 mm/hr, nowcasts show reliability that is close to random forecast for both event durations.

Nowcast error is larger and skillful lead time is lower for smaller areas because they are more sensitive to the misplaced forecast rainfall. Particularly, nowcast errors increase quickly as the area shrinks to below 100 km^2. Besides, results from analysis with the fraction skill score suggest that skillful lead time for an area of 1 km^2 is much shorter than for an area of 25 km^2. These findings imply the limitation on spatial accuracy for the employed STEPS method.

Different quantitative precipitation estimates (QPEs) in the Netherlands were compared and run with STEPS. We found that real-time QPE seriously underestimates the climatological QPE (best reference in the Netherlands) by 17% to 69% depending on the city and rainfall event. CARROTS reduced the error in QPE to within 36%. Thus, using CARROTS-adjusted QPE to run nowcasts also reduced the rainfall underestimation of nowcasts by 70% during a lead time of 4 hours. Yet, underestimation of rainfall still enlarged at longer lead times and the average skillful lead time for the 1-h events was limited to around 20 minutes regardless of the QPEs used.

This research shows that radar-based nowcasting technique can provide skillful forecasts at a very short lead time up to 20 to 24 minutes (indicated by Pearson correlation for 1-h and 24-h events respectively) in Dutch cities. Although the event selection method and choice of metric could alter the lead time, such lead time is a common obstacle that many operational nowcasting models are challenged with when forecasting extreme rainfall in small areas. To advance nowcasting skills, it is recommended to employ machine learning, 3-D nowcasting, or blending nowcasts with numerical weather prediction (NWP) to better model the growth and dissipation of convective rainfall.

References

Achleitner, S., Fach, S., Einfalt, T. & Rauch, W., 2009. Nowcasting of rainfall and of combined sewage flow in urban drainage systems. *Water Science and Technology,* 59(6), pp. 1145-1151.

Acosta-Coll, M., Ballester-Merelo, F. & Martínez-Peiró, M., 2018. Early warning system for detection of urban pluvial flooding hazard levels in an ungauged basin. *Natural hazards,* Volume 92(2), pp. 1237-1265.

Alfieri, L. et al., 2012. Operational early warning systems for water-related hazards in Europe. *Environmental Science & Policy,* Volume 21, pp. 35-49.

Allaire, M., 2018. Socio-economic impacts of flooding: A review of the empirical literatur. *Water Security,* pp. 18-26.

Anagnostou, E. N. & Krajewski, W. F., 1999. Real-time radar rainfall estimation. Part I: Algorithm formulation. *Journal of Atmospheric and Oceanic Technology,* Volume 16(2), pp. 189-197.

Anagnostou, E. N. & Krajewski, W. F., 1999. Real-time radar rainfall estimation. Part II: Case study. *Journal of Atmospheric and Oceanic Technology,* Volume 16(2), pp. 198-205.

Berenguer, M., Sempere-Torres, D. & Pegram, G. G., 2011. SBMcast–An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *Journal of Hydrology,* 404(3-4), pp. 226-240.

Berne, A., Delrieu, G., Creutin, J. D. & Obled, C., 2004. Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology,* 299(3-4), pp. 166-179.

Bowler, N. E., Pierce, C. E. & Seed, A. W., 2006. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society,* Volume 132(620), pp. 2127-2155.

Chen, H. & Chandrasekar, V., 2012. *High resolution rainfall mapping in the Dallas-Fort Worth urban demonstration network.* s.l., IEEE International Geoscience and Remote Sensing Symposium.

Choi, S. & Kim, Y., 2022. Rad-cGAN v1. 0: Radar-based precipitation nowcasting model with conditional Generative Adversarial Networks for multiple dam domains. *Geoscientific Model Development,* Volume 15(15), pp. 5967-5985.

Chrisafis, A., 2016. *The Guardian.* [Online] Available at: https://www.theguardian.com/world/2016/jun/03/paris-river-seine-floods [Accessed 7 3 2022].

Cristiano, E., ten Veldhuis, M. C. & Van De Giesen, N., 2017. Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas–a review. *Hydrology and Earth System Sciences,* 21(7), pp. 3859-3878.

Davolio, S., Silvestro, F. & Malguzzi, P., 2015. Effects of increasing horizontal resolution in a convection-permitting model on flood forecasting: The 2011 dramatic events in Liguria, Italy. *Hydrometeorolgy,* Volume 16, p. 1843–1856.

Dewan, A., 2021. *CNN.* [Online] Available at: https://edition.cnn.com/2021/08/23/europe/germany-floods-belgium-climate-change-intl/index.html [Accessed 21 08 2022].

Donat, M. G. et al., 2016. More extreme precipitation in the world's dry and wet regions. *Nature Climate Change, 6(5),* pp. 508-513.

DutchNews.nl, 2016. *DutchNews.nl.* [Online] Available at: https://www.dutchnews.nl/news/2016/06/heavy-storms-over-holland-bring-flash-floods-tear-down-trees/ [Accessed 28 7 2022].

Ehsani, M. R. et al., 2022. NowCasting-Nets: Representation Learning to Mitigate Latency Gap of Satellite Precipitation Products Using Convolutional and Recurrent Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing,* Volume 60, pp. 1-21.

Field, C. B., Barros, V., Stocker, T. F. & Dahe, Q., 2012. *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change.* s.l.:Cambridge University Press.

Foresti, L., Reyniers, M., Seed, A. & Delobbe, L., 2016. Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium. *Hydrology and Earth System Sciences,* 21(1), pp. 505-527.

Germann, U. & Zawadzki, I., 2002. Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Monthly Weather Review,* Volume 130(12), pp. 2859-2873.

Germann, U., Zawadzki, I. & Turner, B., 2006. Predictability of precipitation from continental radar images. Part IV: Limits to prediction. *Journal of the Atmospheric Sciences,* Volume 63(8), pp. 2092-2108.

Haynes, K. et al., 2009. 'Shelter-in-place' vs. evacuation in flash floods. *Environmental Hazards,* Volume 8(4), pp. 291-303.

Hazenberg, P., Leijnse, H. & Uijlenhoet, R., 2011. Radar rainfall estimation of stratiform winter precipitation in the Belgian Ardennes. *Water Resources Research,* Volume W02507, pp. 47,.

Heuvelink, D., Berenguer, M., Brauer, C. C. & Uijlenhoet, R., 2020. Hydrological application of radar rainfall nowcasting in the Netherlands. *Environment international,* Volume 105431, p. 136.

Imhoff, R. et al., 2021. A climatological benchmark for operational radar rainfall bias reduction. *Hydrology and Earth System Sciences,* Volume 25(7), pp. 4061-4080.

Imhoff, R. O. et al., 2020. Spatial and temporal evaluation of radar rainfall nowcasting techniques on 1,533 events. *Water Resources Research,* p. 56(8).

Imhoff, R. O. et al., 2022. Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning. *Water Resources Research,* Volume 58(3).

James, P. M., Reichert, B. K. & Heizenreder, D., 2018. NowCastMIX: Automatic integrated warnings for severe convection on nowcasting time scales at the German Weather Service. *Weather and Forecasting,* Volume 33(5), pp. 1413-1433.

Jensen, D. G., Petersen, C. & Rasmussen, M. R., 2015. Assimilation of radar-based nowcast into a HIRLAM NWP model. *Meteorol. Appl.,* Volume 22, p. 485–494.

Kain, J. S. et al., 2008. Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting,* 23(5), pp. 931-952.

Kasmalkar, I. G. et al., 2020. When floods hit the road: Resilience to flood-related traffic disruption in the San Francisco Bay Area and beyond. *Science advances,* 6(32), p. 2423.

Kim, D. K. et al., 2021. Improving precipitation nowcasting using a three-dimensional convolutional neural network model from Multi Parameter Phased Array Weather Radar observations. *Atmospheric Research,* Volume 262, p. 105774.

KNMI, 2022. *Precipitation - 5 minute precipitation accumulations from climatological gauge-adjusted radar dataset for The Netherlands (1 km, extended mask) in KNMI HDF5 format.* [Online] Available at: https://dataplatform.knmi.nl/dataset/rad-nl25-rac-mfbs-em-5min-2-0 [Accessed 01 05 2022].

KNMI, 2022. *Precipitation - radar/gauge 5 minute real-time accumulations over the Netherlands.* [Online] Available at: https://dataplatform.knmi.nl/dataset/nl-rdr-data-rtcor-5m-1-0 [Accessed 01 05 2022].

Kober, K., Craig, G. C., Keil, C. & Dörnbrack, A., 2012. Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society,* Volume 138(664), pp. 755-768.

Koks, E., Van Ginkel, K., Van Marle, M. & Lemnitzer, A., 2021. Brief Communication: Critical Infrastructure impacts of the 2021 mid-July western European flood event. *Natural Hazards and Earth System Sciences Discussions,* pp. 1-11.

Kotroni, V. & Lagouvardos, K., 2004. Evaluation of MM5 high-resolution real-time forecasts over the urban area of Athens, Greece. *Journal of Applied Meteorology,* 43(11), pp. 1666-1678.

Li, D., Liu, Y. & Chen, C., 2021. MSDM v1. 0: A machine learning model for precipitation nowcasting over eastern China using multisource data. *Geoscientific Model Development,* Volume 14(6), pp. 4019-4034.

Liguori, S. & Rico-Ramirez, M. A., 2012. Quantitative assessment of short-term rainfall forecasts from radar nowcasts and MM5 forecasts. *Hydrological Processes,* 26(25), p. 3842–3857.

Liguori, S., Rico-Ramirez, M. A., Schellart, A. N. A. & Saul, A. J., 2012. Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. *Atmospheric Research,* Volume 103, pp. 80-95.

Liu, J. & Niyogi, D., 2019. Meta-analysis of urbanization impact on rainfall modification. *Scientific reports,* Volume 9(1), pp. 1-14.

Lovat, A., Vincendon, B. & Ducrocq, V., 2022. Hydrometeorological evaluation of two nowcasting systems for Mediterranean heavy precipitation events with operational considerations. *Hydrology and Earth System Sciences,* Volume 26(10), pp. 2697-2714.

Löwe, R. et al., 2014. Probabilistic online runoff forecasting for urban catchments using inputs from rain gauges as well as statically and dynamically adjusted weather radar. *Journal of Hydrology,* Issue 512, pp. 397-407.

M., M. & L., M., 2020. Performance Comparison between Deep Learning and Optical Flow-Based Techniques for Nowcast Precipitation from Radar Images. *Forecasting,* Volume 2(2), pp. 194-210.

Madsen, H. et al., 2014. Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. *Journal of Hydrology,* pp. 519, 3634-3650.

Maier, R. et al., 2020. Spatial Rainfall Variability in Urban Environments—High-Density Precipitation Measurements on a City-Scale. *Water,* 12(4), p. 1157.

Marshall, J., Hitschfeld, W. & Gunn, K., 1955. Advances in radar weather. *Advances in geophysics,* pp. 1-56.

Masson-Delmotte, V. et al., 2021. *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change,* Cambridge, United Kingdom: Cambridge University Press.

Mejsnar, J., Sokol, Z. & Minářová, J., 2018. Limits of precipitation nowcasting by extrapolation of radar reflectivity for warm season in Central Europe. *Atmospheric Research,* Volume 213, pp. 288-301.

Merz, B., Kreibich, H., Schwarze, R. & Thieken, A., 2010. Review article "Assessment of economic flood damage". *Natural Hazards and Earth System Sciences,* pp. 1697-1724.

Min, S. K., Zhang, X., Zwiers, F. W. & Hegerl, G. C., 2011. Human contribution to more-intense precipitation extremes. *Nature, 470(7334),* pp. 378-381.

Mittermaier, M., Roberts, N. & Thompson, S. A., 2013. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorological Applications,* Volume 20(2), pp. 176-186.

Moisselin, J.-M.et al., 2019. *Seamless approach for precipitations within the 0–3 hours forecast-interval.* Madrid, Spain, European Nowcasting Conference.

Myhre, G. et al., 2019. Frequency of extreme precipitation increases extensively with event rareness under global warming. *Nature,* pp. 9(1), 1-10.

N.M. Roberts, H. L., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review,* Volume 136(1), pp. 78-97.

Olsson, J., Simonsson, L. & Ridal, M., 2014. Rainfall nowcasting: Predictability of short-term extremes in Sweden.. *Urban Water Journal,* 11(7), pp. 605-615.

Otsuka, S. et al., 2016. Precipitation nowcasting with three-dimensional space–time extrapolation of dense and frequent phased-array weather radar observations. *Weather and Forecasting, ,* Volume 31(1), pp. 329-340.

Overeem, A., Buishand, T. A. & Holleman, I., 2009. Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar. *Water Resources Research,* Volume 45.

Overeem, A., Buishand, T. A., Holleman, I. & Uijlenhoet, R., 2010. Extreme value modeling of areal rainfall from weather radar. *Water Resources Research,* Volume 46(9).

Overeem, A., Holleman, I. & Buishand, A., 2009. Derivation of a 10-year radar-based climatology of rainfall. *Journal of Applied Meteorology and Climatology, 48,* p. 1448–1463.

Poletti, M. L. et al., 2019. Using nowcasting technique and data assimilation in a meteorological model to improve very short range hydrological forecasts.. *Hydrology and Earth System Sciences,* Volume 23(9), p. 3823–3841.

Pulkkinen, S. et al., 2019. Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0).. *Geosci. Model Dev,* 12(10), p. 4185–4219.

Rafieeinasab, A. et al., 2015. Toward high-resolution flash flood prediction in large urban areas–Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic model. *Journal of Hydrology,* Volume 531, pp. 370-388.

Ralph, F. M. et al., 2014. A vision for future observations for western US extreme precipitation and flooding. *Journal of Contemporary Water Research & Education,* pp. 153(1), 16-32.

Ravuri, S. et al., 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature,* Issue 597(7878), pp. 672-677.

Ravuri, S. et al., 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature,* Volume 597(7878), pp. 672-677.

Regionale kerncijfers Nederland, 2021. *Regionale kerncijfers Nederland, CBS Statline.* [Online] Available at: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?dl=5A35F [Accessed 18 3 2022].

Rezacova, D., Zacharov, P. & Sokol, Z., 2009. Uncertainty in the area-related QPF for heavy convective precipitation. *Atmospheric research,* Volume 93(1-3), pp. 238-246.

RIONED, 2020. *RadarTools,* Netherlands: RIONED.

Sarangi, C. et al., 2018. Aerosol and urban land use effect on rainfall around cities in Indo-Gangetic Basin from observations and cloud resolving model simulations. *Journal of Geophysical Research: Atmospheres,* Volume 123(7), pp. 3645-3667.

Schellart, A. et al., 2014. Comparing quantitative precipitation forecast methods for prediction of sewer flows in a small urban area. *Hydrological Sciences Journal,* 59(7), pp. 1418-1436.

Schleiss, M. et al., 2020. The accuracy of weather radar in heavy rain: a comparative study for Denmark, the Netherlands, Finland and Sweden.. *Hydrology and Earth System Sciences,* Volume 24(6), pp. 3157-3188.

Schmid, P. E. & Niyogi, D., 2017. Modeling urban precipitation modification by spatially heterogeneous aerosols. *Journal of Applied Meteorology and Climatology,* Volume 56(8), p. 2141–2153.

Seed, A. W., 2003. A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology,* Volume 42(3), p. 381–388.

Seed, A. W., Pierce, C. E. & Norman, K., 2013. Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resources Research,* Volume 49(10), pp. 6624-6641.

Sharif, H. O., Yates, D., Roberts, R. & Mueller, C., 2006. The use of an automated nowcasting system to forecast flash floods in an urban watershed.. *Journal of Hydrometeorology,* 7(1), pp. 190-202.

Sharma, P., 2021. *Holland Times.* [Online] Available at: https://www.hollandtimes.nl/2021-edition-7-september/flood-disaster-in-limburg/ [Accessed 7 3 2022].

Shehu, B. & Haberlandt, U., 2021. Relevance of merging radar and rainfall gauge data for rainfall nowcasting in urban hydrology. *Journal of Hydrology,* Volume 594, p. 125931.

Shehu, B. & Haberlandt, U., 2022. Improving radar-based rainfall nowcasting by a nearest-neighbour approach–Part 1: Storm characteristics.. *Hydrology and Earth System Sciences,* Issue 26(6), pp. 1631-1658.

Shepherd, J. M., 2005. A review of current investigations of urban-induced rainfall and recommendations for the future.. *Earth Interactions,* Volume 9(12), pp. 1-27.

Shongwe, M. E., van Oldenborgh, G. J., Van den Hurk, B. & van Aalst, M., 2011. Projected changes in mean and extreme precipitation in Africa under global warming. Part II: East Africa. *Journal of climate,* Volume 24(14), pp. 3718-3733.

Sideris, I. V., Foresti, L., Nerini, D. & Germann, U., 2020. NowPrecip: Localized precipitation nowcasting in the complex terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society,* Volume 146(729), pp. 1768-1800.

Silvestro, F. et al., 2016. The flash flood of the Bisagno Creek on 9th October 2014: an "unfortunate" combination of spatial and temporal scales. *Journal of Hydrology,* Volume 541, pp. 50-62.

Skok, G., 2015. Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteorological Applications,* Volume 22(3), pp. 477-484.

Statistics Netherlands and the Land Registry, 2020. *District and neighborhood map.* s.l., s.n.

Steinheimer, M. & Haiden, T., 2007. Improved nowcasting of precipitation based on convective analysis fields. *Advances in Geosciences,* Volume 10, pp. 125-131.

Steinheimer, M. & Haiden, T., 2007. Improved nowcasting of precipitation based on convective analysis fields. *Advances in Geosciences,* Volume 10, pp. 125-131.

Sun, J. & Ao, J., 2013. Changes in precipitation and extreme precipitation in a warming environment in China. *Chinese Science Bulletin,* 58(12), pp. 1395-1401.

Sun, J. X. M. et al., 2014. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society,* Volume 95 (3), pp. 409-426.

Sun, J. et al., 2014. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society,* Volume 95(3), pp. 409-426.

Sun, N., Zhou, Z., Li, Q. & Jing, J., 2022. Three-Dimensional Gridded Radar Echo Extrapolation for Convective Storm Nowcasting Based on 3D-ConvLSTM Model. *Remote Sensing,* Volume 14(17), p. 4256.

Tabari, H., 2020. Climate change impact on flood and extreme precipitation increases with water availability. *Nature,* pp. 10(1), 1-10.

Thorndahl, S., Nielsen, J. E. & Jensen, D. G., 2016. Urban pluvial flood prediction: a case study evaluating radar rainfall nowcasts and numerical weather prediction models as model inputs. *Water Science and Technology,* 74(11), pp. 2599-2610.

Tingsanchali, T., 2012. Urban flood disaster management. *Procedia engineering,* Volume 32, pp. 25-37.

Turner, B. J., Zawadzki, I. & Germann, U., 2004. Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE). *Journal of Applied Meteorology and Climatology,* 43(2), pp. 231-248.

UN, 2018. *Revision of world urbanization prospects,* s.l.: UN.

Ushio, T., Wu, T. & Yoshida, S., 2015. Review of recent progress in lightning and thunderstorm detection techniques in Asia. *Atmospheric Research,* Volume 154, pp. 89-102.

Van de Beek, C. Z., Leijnse, H., Hazenberg, P. & Uijlenhoet, R., 2016. Close-range radar rainfall estimation and error analysis. *Atmospheric Measurement Techniques,* 9(8), p. 3837–3850.

W., N. & C., Y., 2019. Optimize Short-Term Rainfall Forecast with Combination of Ensemble Precipitation Nowcasts by Lagrangian Extrapolation.. *Water,* Volume 11(9), p. 1752.

Wang, G., Wong, W., Liu, L. & Wang, H., 2013. Application of multi-scale tracking radar echoes scheme in quantitative precipitation nowcasting. *Advances in Atmospheric Sciences,* Volume 30(2), pp. 448-460.

Woo, W. C. & Wong, W. K., 2017. Operational application of optical flow techniques to radar-based rainfall nowcasting. *Atmosphere,* Volume 8(3), p. 48.

Yang, L., Ni, G., Tian, F. & Niyogi, D., 2021. Urbanization exacerbated rainfall over European suburbs under a warming climate. *Geophysical Research Letters,* Volume 48(21).

Yoshikawa, E. et al., 2012. MMSE beam forming on fast-scanning phased array weather radar. *IEEE Transactions on Geoscience and remote sensing,* Volume 51(5), pp. 3077-3088.

Zhang, W., Yang, J., Yang, L. & Niyogi, D., 2022. Impacts of city shape on rainfall in inland and coastal environments. *Earth's Future,* p. 10.

Zhao, B. & Zhang, B., 2018. Assessing hourly precipitation forecast skill with the fractions skill score. *Journal of Meteorological Research,* Volume 32(1), pp. 135-145.

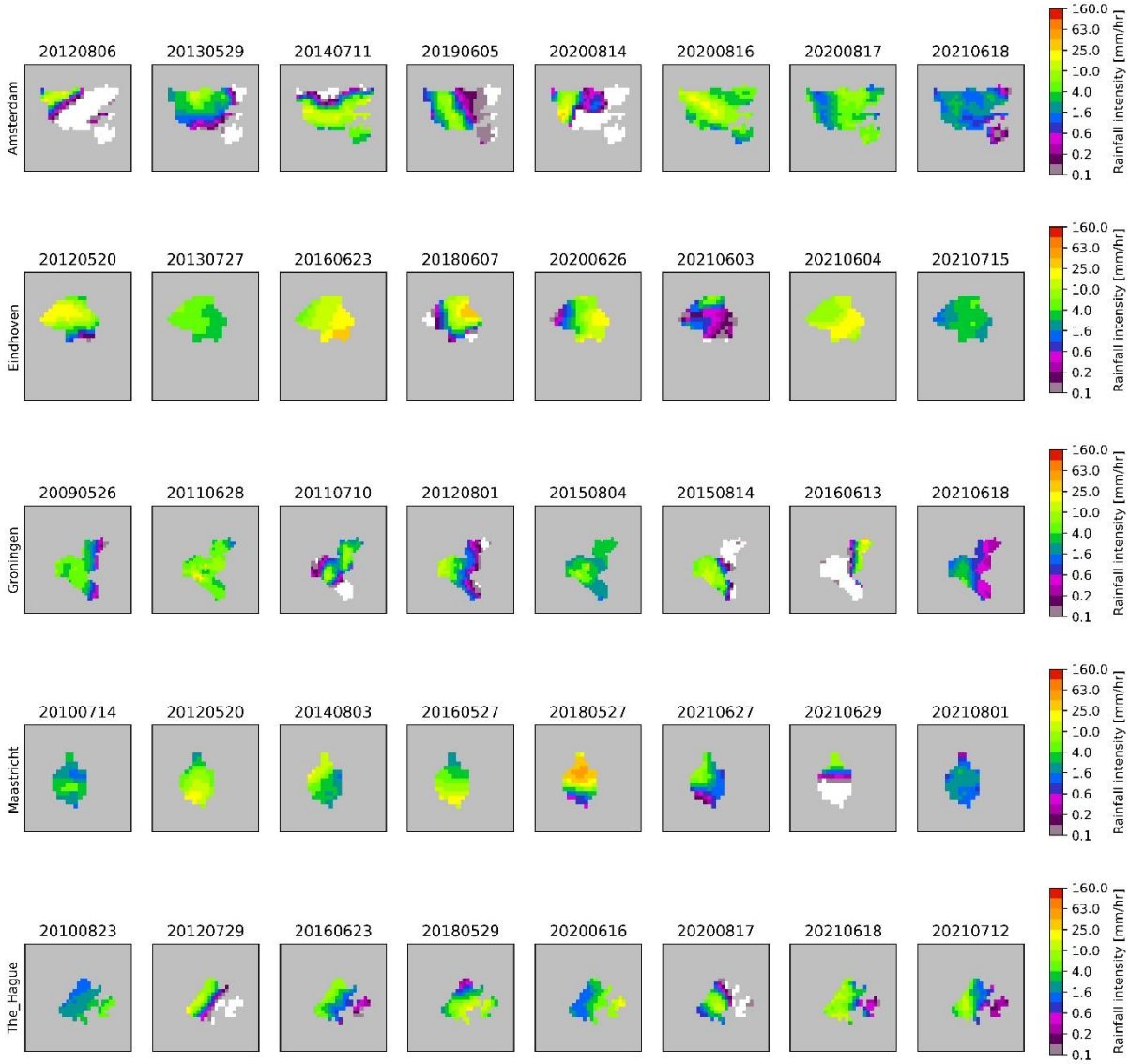
Appendix

Supplementary Table 1. List of the onset times of the 1-h events in each city

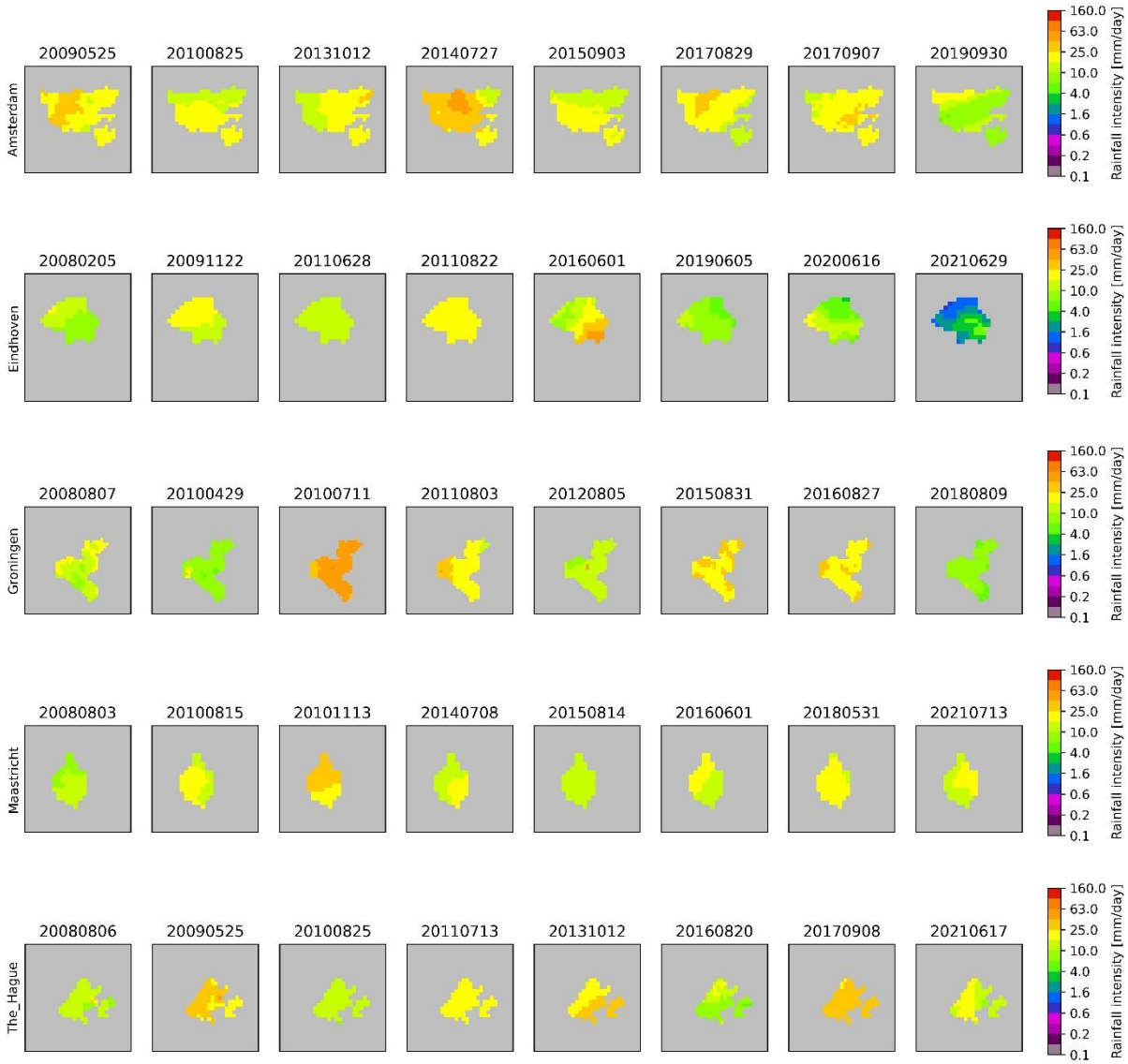
Groningen	Amsterdam	The Hague	Eindhoven	Maastricht
2021-06-18-18:05	2021-06-18-16:15	2021-07-12-14:45	2021-07-15-01:45	2021-08-01-12:10
2016-06-13-13:55	2020-08-17-13:10	2021-06-18-14:00	2021-06-04-16:20	2021-06-29-16:45
2015-08-14-19:35	2020-08-16-17:50	2020-08-17-11:30	2021-06-03-13:40	2021-06-27-12:35
2015-08-04-04:45	2020-08-14-14:25	2020-06-16-13:05	2020-06-26-19:30	2018-05-27-11:05
2012-08-01-23:30	2019-06-05-20:20	2018-05-29-15:00	2018-06-07-15:00	2016-05-27-16:15
2011-07-10-14:30	2014-07-11-00:05	2016-06-23-01:55	2016-06-23-18:30	2014-08-03-16:35
2011-06-08-21:00	2013-05-29-13:30	2012-07-29-19:55	2013-07-27-08:45	2012-05-20-15:00
2009-05-26-01:50	2012-08-06-08:25	2010-08-23-04:35	2012-05-20-16:55	2010-07-14-15:20

Supplementary Table 2. List of the onset times of the 24-h events in each city

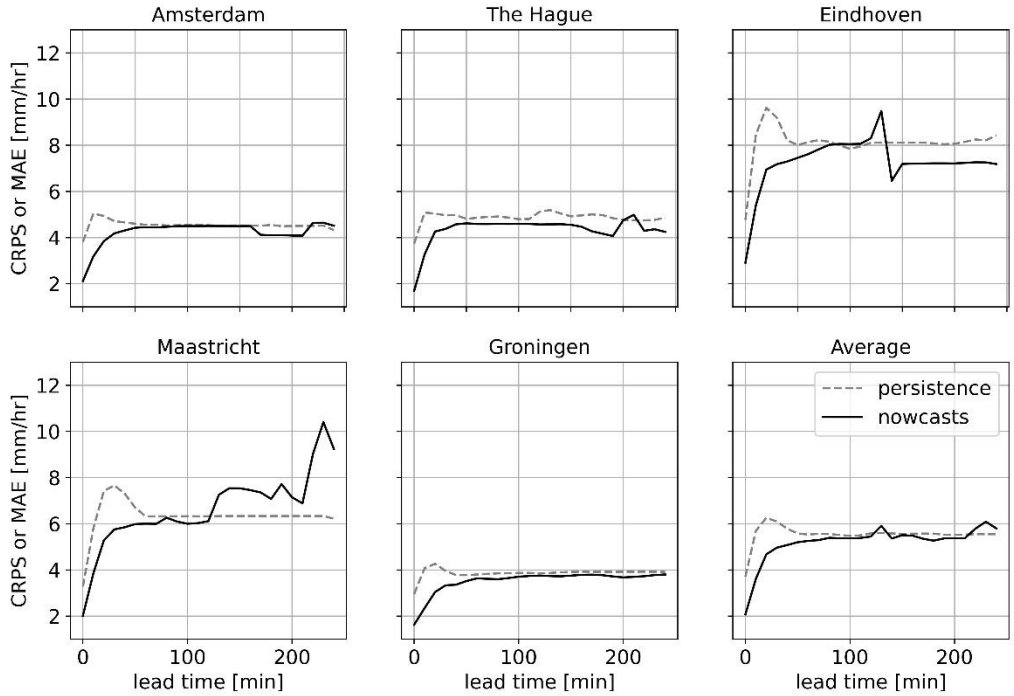
Groningen	Amsterdam	The Hague	Eindhoven	Maastricht
2018-08-09-13:00	2019-09-30-21:00	2021-06-17-16:00	2020-06-26-05:00	2021-07-13-14:00
2016-08-27-13:00	2017-09-07-19:00	2017-09-08-04:00	2020-06-16-17:00	2018-05-31-17:00
2015-08-31-01:00	2017-08-29-18:00	2016-08-20-10:00	2019-06-05-02:00	2016-06-01-02:00
2012-08-05-12:00	2015-09-03-20:00	2013-10-12-18:00	2016-06-01-01:00	2015-08-14-21:00
2011-08-03-00:00	2014-07-27-12:00	2011-07-13-18:00	2011-08-22-12:00	2014-07-08-20:00
2010-07-11-21:00	2013-10-12-17:00	2010-08-25-12:00	2011-06-28-02:00	2010-11-13-02:00
2010-04-29-07:00	2010-08-25-13:00	2009-05-25-15:00	2009-11-22-23:00	2010-08-15-01:00
2008-08-07-10:00	2009-05-25-18:00	2008-08-06-23:00	2008-02-05-05:00	2008-08-03-02:00



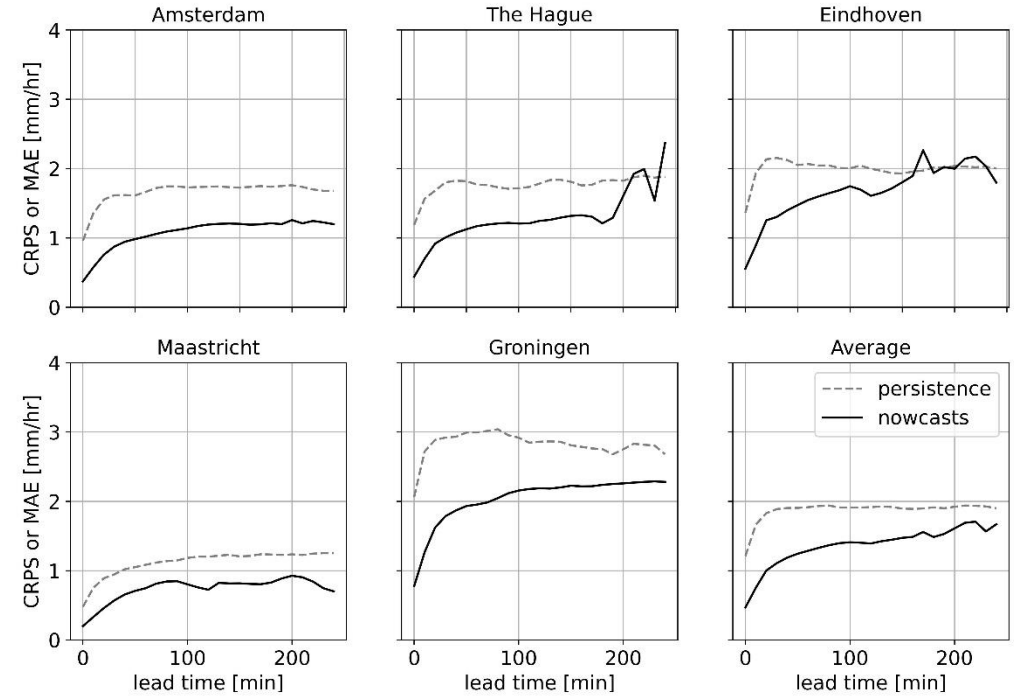
Supplementary Figure 1. Rainfall intensity of all 1-h events in each city



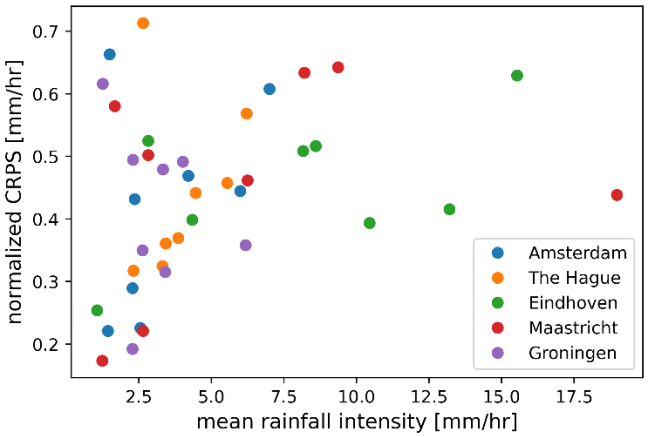
Supplementary Figure 2. Rainfall intensity of all 24-h events in each city



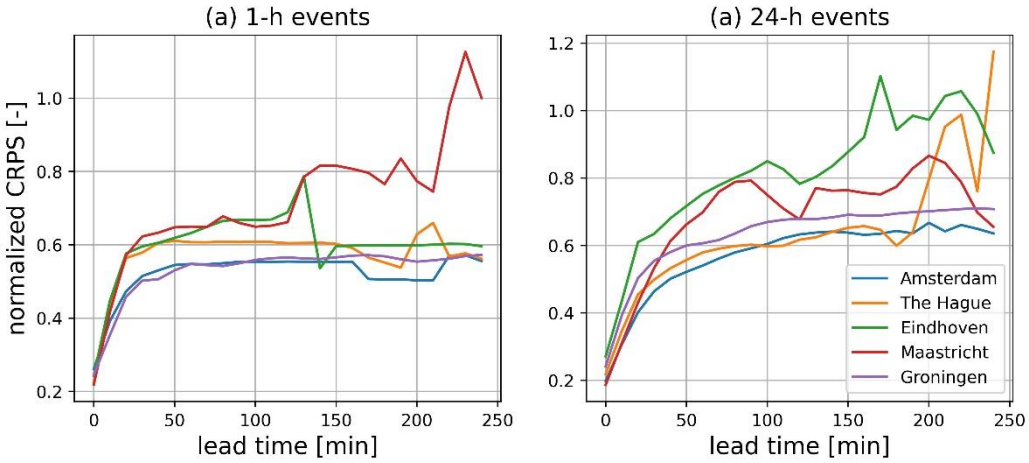
Supplementary Figure 3. CRPS of nowcasts and MAE of Eulerian persistence in each city (averaged over the 1-h events)



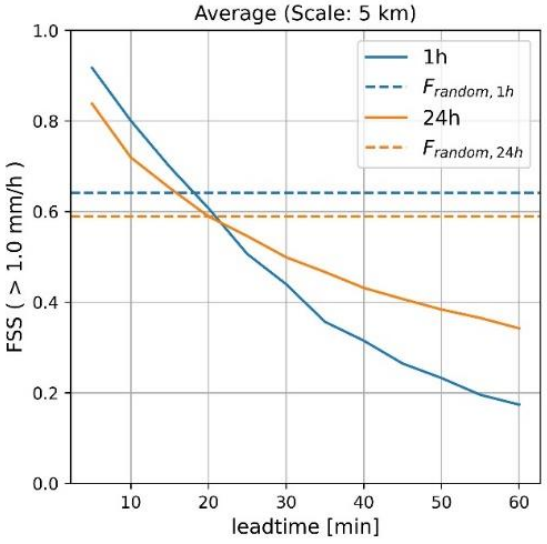
Supplementary Figure 4. CRPS of nowcasts and MAE of Eulerian persistence in each city (averaged over the 24-h events)



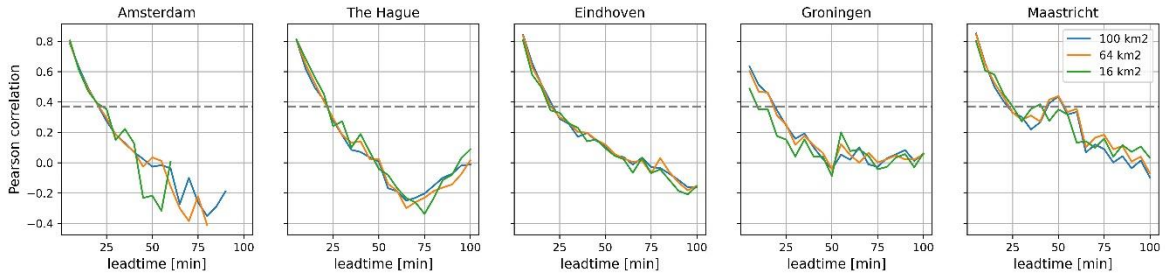
Supplementary Figure 5. Normalized CRPS (average between 5- to 30-minute lead time) against mean rainfall intensity in the five cities during the 1-h events



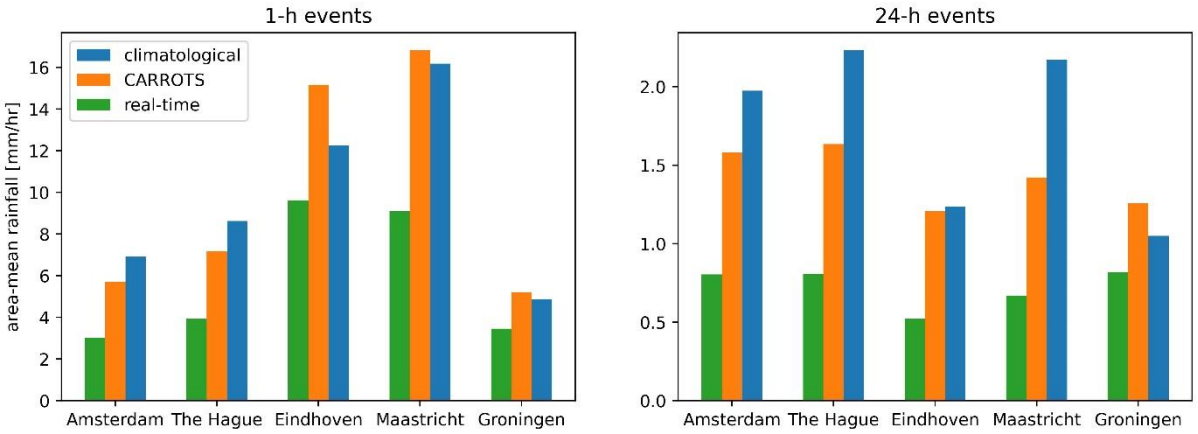
Supplementary Figure 6. Average CRPS for each city after normalized with standard deviation of the rainfall events for the 1-h events (left) and 24-h events (right)



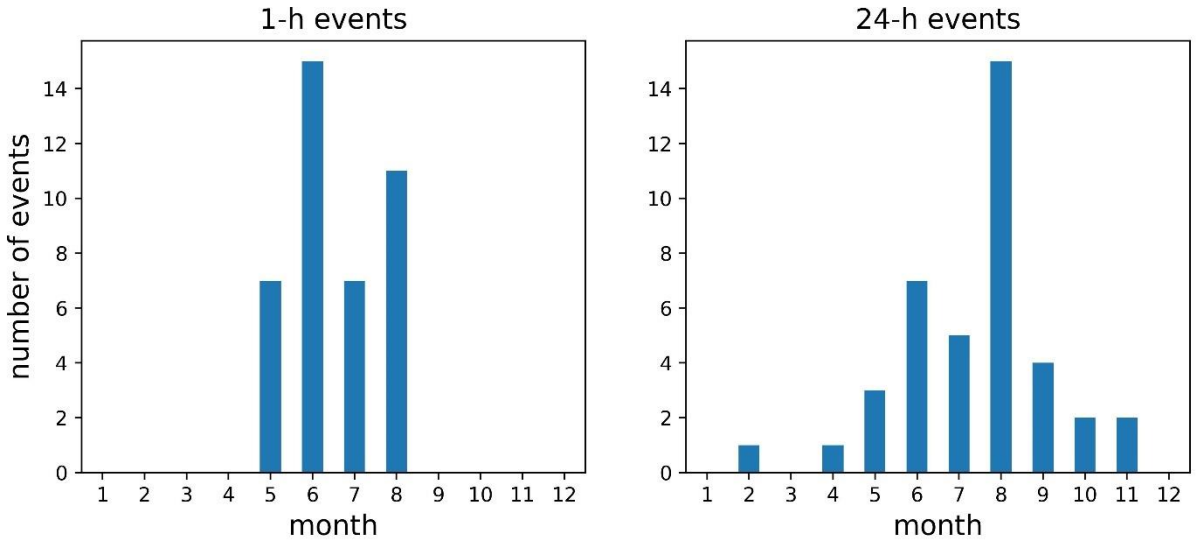
Supplementary Figure 7. Average FSS of all 1-h and 24-h events at a length scale of 5 km. The dashed lines show the random fraction skill scores. FSS lower than the dashed lines means that nowcasts are not skillful.



Supplementary Figure 8. Pearson correlation of 1-h events in the subareas in each city against lead time.



Supplementary Figure 9. Area-mean rainfall using climatological, CARROTS, and real-time QPE. Only considering the 23 1-h events happening before 2019.



Supplementary Figure 10. Monthly distribution of the extreme events.