Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026723

Spatial and Temporal Evaluation of Radar Rainfall Nowcasting Techniques on 1,533 Events

R. O. Imhoff[1,2], C. C. Brauer[1], A. Overeem[1,3], A. H. Weerts[1,2], and R. Uijlenhoet[1]

[1]Hydrology and Quantitative Water Management Group, Wageningen University & Research, Wageningen, The Netherlands, [2]Operational Water Management, Department of Inland Water Systems, Deltares, The Netherlands, [3]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

Key Points:
- A large-sample analysis of 1,533 events is used to determine the skill of radar rainfall nowcasts
- Evaluation takes place with four nowcasting algorithms from the Pysteps and Rainymotion libraries
- Nowcast skill is found to depend on event duration, season, catchment size, and location

Supporting Information:
- Supporting Information S1

Correspondence to:
R. O. Imhoff,
ruben.imhoff@wur.nl

Citation:
Imhoff, R. O., Brauer, C. C., Overeem, A., Weerts, A. H., & Uijlenhoet, R. (2020). Spatial and temporal evaluation of radar rainfall nowcasting techniques on 1,533 events. *Water Resources Research*, *56*, e2019WR026723. https://doi.org/10.1029/2019WR026723

Received 8 NOV 2019
Accepted 23 JUN 2020
Accepted article online 29 JUN 2020

Abstract Radar rainfall nowcasting, the process of statistically extrapolating the most recent rainfall observation, is increasingly used for very short range rainfall forecasting (less than 6 hr ahead). We performed a large-sample analysis of 1,533 events, systematically selected for 4 event durations and 12 lowland catchments (6.5–957 km^2), to determine the predictive skill of nowcasting. Four algorithms are tested and compared with Eulerian Persistence: Rainymotion Sparse, Rainymotion DenseRotation, Pysteps deterministic, and Pysteps probabilistic with 20 ensemble members. We focus on the dependency of nowcast skill on event duration, season, catchment size, and location. Maximum skillful lead times increase for longer event durations, due to the more persistent character of these events. For all four event durations, Pysteps deterministic attains the longest average decorrelation times, with 25 min for 1-hr durations, 40 min for 3 hr, 56 min for 6 hr, and 116 min for 24 hr. During winter, with more persistent stratiform precipitation, we find three times lower mean absolute errors than for convective summer precipitation. Higher skill is also found after spatially upscaling the forecast. Catchment location matters too: Given the prevailing storm movement, two times higher skillful lead times are found downwind than upwind toward the edge of the domain. In most cases, Pysteps algorithms outperform the Rainymotion benchmark algorithms. We speculate that most errors originate from growth and dissipation processes which are not or only partially (stochastically) accounted for.

Plain Language Summary Early warning systems are a key instrument for timely responses to flood risk. These warnings depend on accurate weather forecasts. Most numerical weather prediction models have trouble to accurately forecast the timing and especially the location of rainfall on time scales of less than 6 hr. This time frame is important for decisions in small, mountainous, polder, and urban basins, where river discharge responds quickly to rainfall. Radar rainfall nowcasting, the process of statistically extrapolating the most recent rainfall observations, has the potential to provide forecasts up to 6 hr. We considered 1,533 rainfall events to evaluate the quality of five nowcasting methods. Nowcasts are better for longer-lasting rain storms with relatively low intensity, which typically occur in winter, than for short, intensive storms, which typically occur in summer. Forecasts are useful up to 2 hr ahead for events of 1 day, while this is only 25 min for 1-hr events. Moreover, nowcasts are better for larger basins and in the downwind direction with respect to the prevailing storm movement. Hence, nowcasts are useful, but improvements are needed to be able to forecast longer ahead.

1. Introduction

The frequency and severity of intense precipitation events are likely to increase in a changing climate (with, e.g., a 12% increase in high-intensity precipitation per degree of warming in the Netherlands), which can lead to more severe floods and present a danger to livability and economy (e.g., IPCC, 2012, 2013, 2014; KNMI, 2015). Well-established early warning systems (e.g., Delft-FEWS Werner et al., 2013) make it possible to act accordingly and in time, expectedly resulting in a lower risk and less damage (Pappenberger et al., 2015). Most early warning systems, if present at all, use a combination of short-range (12–72 hr) and medium-range (up to 10 days) numerical weather prediction (NWP) in combination with hydrological and hydraulic models to predict river discharges and water levels. However, the quantitative precipitation forecasts (QPFs) provided by the employed NWP systems are often not sufficient for reliable early warnings on the short term, that is, up to 6 hr, due to (1) a too coarse temporal resolution and a too low update

frequency (i.e., the lead time becomes already quite long, making the forecast less reliable) and (2) the mislocation of rainfall events (e.g., Berenguer et al., 2012; Pierce et al., 2012).

In addition to the increasing availability of NWP models that focus on short-term precipitation forecasting (<12 hr ahead), there has been a significant improvement of the spatial and temporal resolution of radar rainfall products over the last decades, typically to 1 km and 5 min (e.g., Overeem, Holleman, & Buishand, 2009; Serafin & Wilson, 2000). These radar products have high potential for very short term rainfall forecasts (Germann & Zawadzki, 2002, 2004; Turner et al., 2004) and can therefore be a valuable addition to early warning systems. Very short term forecasting with QPE from, for example, operational weather radars is called nowcasting. Essentially, nowcasting is the process of extrapolating real-time remotely sensed observations (often radar) by estimating the advection of the precipitation fields. Increasingly, the spatial and temporal properties of these fields and the statistical properties of the available QPE are taken into account as well. However, in the current nowcasting models, physical processes governing the growth and dissipation of precipitation cells are not accounted for.

Nowcasts can be applied up to several hours ahead (Germann et al., 2006; Lin et al., 2005) and approximately 30 min for convective cells (e.g., Ayzel et al., 2019; Foresti et al., 2016; Liguori & Rico-Ramirez, 2012). In this time frame, it is thought to fill the gap for very short term forecasts up to 3 hr ahead or even 6 hr on a continental scale (e.g., Berenguer & Sempere Torres, 2013), after which short-range and mid-range NWP models should take over.

Nowcasts can be made in a deterministic sense, with, for example, TITAN (Dixon & Wiener, 1993), S-PROG (Seed, 2003), and Com-SWIRLS (Wong et al., 2016), or in a probabilistic sense by accounting for uncertainty in precipitation forecasts by means of ensembles. Examples of probabilistic algorithms are STEPS (Bowler et al., 2006; Seed, 2003; Seed et al., 2013), SBMcast (Berenguer et al., 2011), the stochastic- and analogue-based models by Atencia and Zawadzki (2014, 2015), ENS (Sokol et al., 2017), and Pysteps (Pulkkinen et al., 2019). The ensemble QPF can be directly applied to hydrological ensemble forecasts (e.g., Berenguer et al., 2005; Heuvelink et al., 2020; Vivoni et al., 2006).

As operational nowcasting for hydrological purposes is still in an early stage of development, advice is needed on the skill of radar nowcasting in general and differences between the performance of algorithms in particular. Most studies so far have focused on the development of nowcasting algorithms in combination with a quantification of the rainfall prediction quality and errors in either deterministic or probabilistic nowcasting algorithms (e.g., Foresti et al., 2016; Germann & Zawadzki, 2002, 2004; Germann et al., 2006; Lin et al., 2005; Turner et al., 2004). The results generally follow from studies with analyses based on relatively small samples of 2–15 precipitation events. The studies by Berenguer and Sempere Torres (2013), Foresti and Seed (2015), and Mejsnar et al. (2018) are exceptions to this. Foresti and Seed (2015) use a data set of 20 months of operational nowcasts in order to analyze the spatial distribution of radar rainfall nowcasting errors in a mountainous region in south-east Australia.

In order to draw statistically meaningful conclusions about the rainfall forecasting skill in a lowland area with a temperate climate such as the Netherlands, a study with a large number of precipitation events should take place. Accordingly, the objective of this study is to quantify the skill of radar rainfall nowcasting algorithms for the short-term predictability of rainfall for different catchments in the Netherlands. Earlier studies suggest that forecast skill and uncertainty of nowcasting algorithms depend on factors such as climate, geography, and orography (Foresti & Seed, 2015; Foresti et al., 2016; Germann et al., 2009), with a higher variability in forecast errors for smaller regions (Vivoni et al., 2007). Therefore, a particular focus will be on the dependency of the forecast skill on event type and duration, seasonality, catchment size, and location for 12 catchments. The objective excludes blending with NWP, which will be the next stage in improving the short-term predictability of rainfall for lead times of more than 3 hr.

In this study, 1,533 events spread over 12 catchments with sizes varying from 6.5 to 957 km^2 are analyzed. For this analysis, four open-source (Python) nowcasting algorithms are used: two benchmarking advection algorithms of Rainymotion (Ayzel et al., 2019) and deterministic and probabilistic versions of Pysteps (Pulkkinen et al., 2019). To the authors' knowledge, this is the first radar rainfall nowcasting study with a combination of this variety of algorithms and such a large sample of events.

The outline of this paper is as follows: Section 2 contains descriptions of the study area, radar data, event selection, nowcasting algorithms, verification metrics, and experimental setup. This is followed by the results (section 3), discussion (section 4), and conclusions (section 5).

2. Materials and Methods

2.1. Study Area

In this study, we analyze rainfall over 12 Dutch catchments and polder areas with different sizes and locations (Figure 1). We focus on catchments instead of the entire radar domain, because we want to assess the usefulness of nowcasting for the involved water authorities. Incidentally, this highly reduces the storage requirement for this large-sample analysis (compared to nowcasts for the full domain). The catchments were chosen in close collaboration with involved water authorities and are spread over the country, as we expected a dependency of the nowcast skill on the location with regard to the prevailing storm movement. With south-westerlies as the predominant wind direction in the Netherlands, catchments in the northeastern part of the country are expected to have skillful rainfall predictions up to longer lead times than catchments in the southwest.

2.2. Data and Event Selection

2.2.1. Radar Rainfall Product

The Royal Netherlands Meteorological Institute (KNMI) operates two C-band weather radars, indicated together with the composite extent for the employed data sets in Figure 1. The two radars in De Bilt and Den Helder were replaced by two new radars between September 2016 and January 2017. The current operational radars at Den Helder and Herwijnen are dual-polarized, which was not the case before September 2016. See Beekhuis and Holleman (2008) and Beekhuis and Mathijssen (2018) for more information.

In the data processing, Doppler filtering is used, and since 2013 also a cloud-mask from satellite data is applied to remove non-meteorological echoes. Subsequently, horizontal cross-sections of reflectivity at constant altitude, called pseudo-constant plan position indicators (pseudo-CAPPI), are constructed from the volumetric reflectivity data per radar. For the final composite, 1,500 m pseudo-CAPPIs are employed, and the reflectivities from both radars are combined using range-weighted compositing (Overeem, Holleman, & Buishand, 2009). Rainfall is then estimated with a fixed Z-R relationship (Marshall et al., 1955):

$$Z_{\mathrm{h}} = 200R^{1.6}, \tag{1}$$

with Z_{h} the reflectivity at horizontal polarization ($\mathrm{mm^6\,m^{-3}}$) and R the rainfall rate ($\mathrm{mm\,hr^{-1}}$). The conversion from $\mathrm{mm^6\,m^{-3}}$ to dBZ, often the unit of the measured reflectivities, is performed with $10 \times \log_{10}(Z_h)$. Reflectivities below 7 $\mathrm{dBZ_h}$ ($\approx$0.1 $\mathrm{mm\,hr^{-1}}$) are ignored to prevent noise accumulation, and reflectivities above 55 $\mathrm{dBZ_h}$ ($\approx$100 $\mathrm{mm\,hr^{-1}}$) are fixed at 55 $\mathrm{dBZ_h}$ to suppress hail or strong residual clutter-induced reflection. Isolated pixels with reflectivities above 7 $\mathrm{dBZ_h}$ are discarded. The resulting radar rainfall composite has a 1-$\mathrm{km^2}$ spatial and 5-min temporal resolution and has been archived since 2008.

In this study, the quantitative precipitation estimation (QPE) is assumed to be the true rainfall intensity for the verification of the rainfall forecasts, that is, the output of the nowcasting algorithms. Provided that radar QPE comes with substantial uncertainty and bias (e.g., Foresti & Seed, 2015; Germann et al., 2006; Hazenberg et al., 2011; Van de Beek et al., 2016), a good verification result in this study does not necessarily mean that the true rainfall amounts are well predicted by the algorithms.

Note that KNMI produces another, more accurate radar product which includes corrections with rain gauge data. This gauge-adjusted data set has the same coverage, spatial and temporal resolution as the operational data (used for nowcasting). However, the QPE is adjusted with two rain gauge networks (consisting of 31 automatic and 325 manual gauges), and it is therefore considered as an accurate rainfall reference product. As this product is not available in real time, we have only used this "reference" data set for the event selection. For more details regarding this data set, we refer to Beekhuis and Holleman (2008), Overeem, Holleman, and Buishand (2009), Overeem, Buishand, and Holleman (2009), Overeem et al. (2011), and Beekhuis and Mathijssen (2018).

Number	Catchment name	Area (km^2)
1	Regge	957
2	Aa	836
3	Delfland	379
4	Reusel	176
5	Linde	150
6	Rijnland	89
7	Roggelsebeek	88
8	Dwarsdiep	83
9	Beemster	71
10	Luntersebeek	63
11	Grote Waterleiding	40
12	Hupsel Brook	6.5

Figure 1. Map of the Netherlands with the 12 catchments (green areas) and 3 radars (red triangles). The large circle indicates the extent of the radar rainfall composite. The catchment sizes are indicated in the table next to the figure.

2.2.2. Event Selection Procedure

A large number of events were selected systematically from the 5-min gauge-adjusted radar rainfall composites for the period 2008–2018. Only large rainfall accumulations were selected, as these are most interesting to study for both the assessment of precipitation predictability and the hydrological application. The gauge-adjusted data set is employed for event selection instead of the operational product employed for the nowcasting analysis (section 2.2.1), because this yields the events with the actual highest rainfall volumes.

In the Dutch climate, the highest rainfall rates originate from convective precipitation events during summer and early autumn. Since one of the aims of this study is to find the seasonal dependence of the nowcasting skill, events in other seasons, which include stratiform events, should be present as well (see Table 1 for seasonal statistics of the Dutch weather). The adopted event selection procedure guarantees that an even spread of strong precipitation events is obtained over all seasons per event duration.

The events are selected as follows (Figure 2): Per catchment and for each season, eight events are selected per event duration (1, 3, 6, and 24 hr). Note that an "event" is not defined by the start and end of rainfall, but instead periods with a certain duration are used in which it does not have to rain continuously. Hence, the full duration is considered as an event. With this description of an event, the highest rainfall sums per duration, catchment, and season are ranked from high to low, in which the next "event" cannot occur within the time span of a previously selected "event" with a higher rainfall sum. The eight events, for that duration, consist of the events with the four highest catchment-averaged rainfall sums and the four highest rainfall sums for any grid cell in the catchment. If one of the four events of the grid maxima is the same event as already present in the four maxima from the catchment-averaged list, the next maximum in the list of grid maxima is used to avoid overlapping events. Summed over all durations (4), seasons (4), and catchments (12), this selection procedure leads to $4 \times 4 \times 12 \times 8 = 1{,}536$ events (see Table 1 for the statistics of these events).

Table 1
Rainfall and Wind Statistics for the Selected Events and Seasonal Averages for KNMI Station De Bilt

	Mean rainfall intensity (mm hr^{-1})				
Aggregation interval	DJF	MAM	JJA	SON	Daily mean wind direction
1 hr	8.5	17.1	27.1	15.4	223° (SW)
3 hr	4.6	7.7	12.1	7.5	227° (SW)
6 hr	3.0	4.5	7.0	4.6	228° (SW)
24 hr	1.2	1.5	2.2	1.7	222° (SW)
Climatology	0.09	0.08	0.10	0.11	SW

Note. For the event statistics, rainfall intensities are catchment averages over the duration and all events in that duration. The mean daily wind direction for the events is obtained from measured wind directions during these events at KNMI station De Bilt. For the climatology, the seasonal accumulations for station De Bilt are averaged over the period 1981–2010 (KNMI, 2011). Mean climatological wind directions have been determined over periods with rainfall by Buishand and Velds (1980).

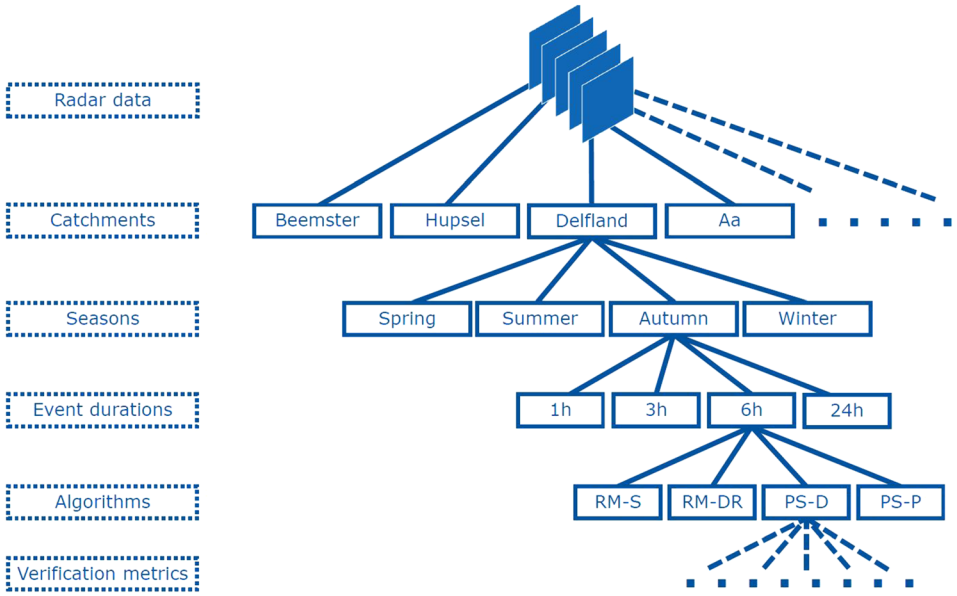

Figure 2. Schematization of the employed event selection procedure. Per catchment, season, and event duration, eight events are selected.

2.3. Nowcasting Algorithms

The following paragraphs give a brief description of the main characteristics of the employed algorithms and the way they are set up in this study. For brevity, the algorithm names are abbreviated from here on (Table 2). The following algorithms were chosen because they are open source and because they allow for a comparison between different methods, namely, a global optical flow method compared to a corner detection method, and purely advection-based nowcasting compared to methods that incorporate the spatial and temporal scales of rainfall for rainfall field evolution (either with or without uncertainties taken into account).

2.3.1. Rainymotion

Rainymotion (Ayzel et al., 2019) was introduced as a benchmark to test and develop other nowcasting algorithms and as such to replace the commonly used benchmark Eulerian Persistence, which is the procedure of using the most recent QPE as forecast. It is a set of four models based on widely used optical flow algorithms to determine advection of rainfall fields. Two models (from Rainymotion v0.1) are used in this study and briefly described below.

The first model, called *Sparse* (RM-S), tracks the corners of precipitation fields (which are scaled to brightness integer values ranging from 0 to 255), as these locations have sharp rainfall intensity gradients which are relatively easy to find and track. The Sparse method identifies these corners from time $t-23$ (e.g., 5-min steps) to t with the Shi-Tomasi corner detector (Shi & Tomasi, 1994). With the Lucas-Kanade optical flow algorithm (window size is 20×20 cells) (Lucas & Kanade, 1981), the identified features are tracked. The obtained motion is then linearly extrapolated to the future. Subsequently, a transformation matrix is calculated per lead time and is used to extrapolate the most recent radar image by warping using an affine transformation matrix (Ayzel et al., 2019).

The second model, *DenseRotation* (RM-DR), uses a global optical flow algorithm to estimate a velocity for each grid cell (with rainfall scaled to brightness integers) in the composite. The default method (also used here) for this is the Dense Inverse Search algorithm introduced by Kroeger et al. (2016). It uses the QPE from time $t-1$ to t. Rainymotion offers the opportunity to change this optical flow algorithm to a variety of other algorithms. The velocity field is extrapolated with the semi-Lagrangian advection scheme as introduced by Germann and Zawadzki (2002). A forward semi-Lagrangian advection scheme is used here. This methodology allows for rotational movement, which is not the case with, for example, a constant-vector advection scheme. After these steps, the resulting pixel values are interpolated with Inverse Distance Weighting to

Table 2
Overview of the Radar Rainfall Nowcasting Methods Used in This Study

Name	Abbreviation	Reference
Eulerian Persistence	EP	—
Rainymotion Sparse	RM-S	Ayzel et al. (2019)
Rainymotion DenseRotation	RM-DR	Ayzel et al. (2019)
Pysteps deterministic (S-PROG)	PS-D	Seed (2003) and Pulkkinen et al. (2019)
Pysteps probabilistic	PS-P	Bowler et al. (2006), Seed et al. (2013), and Pulkkinen et al. (2019)

project them on the original grid. This is different than in Germann and Zawadzki (2002), who used a bilinear interpolation technique.

2.3.2. Pysteps

Pysteps (Pulkkinen et al., 2019) is a modular framework for the development of nowcasting methods. With a wide variety of configurations, it is a platform for deterministic and probabilistic nowcasting applications. The core of Pysteps is based on S-PROG (Seed, 2003) and STEPS (Bowler et al., 2006; Seed et al., 2013). The main steps toward an ensemble nowcast in Pysteps are as follows:

1. Read radar composites and determine the motion field.
2. Use an advection method for the extrapolation of the radar images into the future. Generally, a backward semi-Lagrangian method, allowing for rotational movement, is used (Germann & Zawadzki, 2002).
3. Apply cascade decomposition, generally fast Fourier transform (FFT), to decompose the rainfall field into a multiplicative cascade (Seed, 2003). The levels of this cascade represent different spatial scales, which are assumed to represent different precipitation lifetimes (Germann et al., 2006; Seed, 2003; Venugopal et al., 1999).
4. Estimate the autoregressive (AR) model parameters and apply the AR model in time. Together with the cascade model used in space, this methodology handles the temporal evolution of and correlation within the precipitation structure.
5. Add stochastic perturbations to the AR models and to the advection field as a way to take into account uncertainty in rainfall intensities and the motion field for the ensemble forecast.
6. Perform the actual extrapolation after recomposing the cascade with the iterated AR model and the stochastic perturbations. This will give the raw nowcast ensemble.
7. Apply post-processing operations to ensure that the nowcast has the same statistical properties as the latest available observations.

Pysteps (v0.2 was used) is flexible in the choice for, for example, optical flow methods, advection methods, noise methods, and spatial/temporal decomposition methods. In this study, two setups of Pysteps are used: one for deterministic nowcasts and one for probabilistic nowcasts. The deterministic setup, from here on referred to as *Pysteps deterministic* (PS-D), resembles the S-PROG algorithm (Seed, 2003) and has the following configuration: a Lucas-Kanade optical flow method using the QPE from time $t-3$ to t (Lucas & Kanade, 1981), a backward semi-Lagrangian advection method (Germann & Zawadzki, 2002), an AR model of order 2, the S-PROG masking method (threshold is 0.1 mm hr^{-1}), a probability matching method to match the forecast statistics with the observations based on the mean observed rainfall fields, and eight cascade levels (instead of six in the original S-PROG). PS-D follows most of the aforementioned seven steps, except for Step 5 (stochastic perturbations).

The probabilistic setup, from here on referred to as *Pysteps probabilistic* (PS-P), follows the aforementioned seven steps, with the following configuration: a Lucas-Kanade optical flow method using the QPE from time $t-3$ to t (Lucas & Kanade, 1981), a backward semi-Lagrangian advection method (Germann & Zawadzki, 2002), the STEPS nowcasting method (Bowler et al., 2006), a non-parametric noise method (Seed et al., 2013), FFT for the spatial decomposition with eight cascade levels, an AR model of order 2, a lead time-dependent masking method, the cumulative distribution function (cdf) used as probability matching method, and 20 ensemble members. For both PS-D and PS-P, the rainfall fields are transformed to dB prior to nowcasting.

Pulkkinen et al. (2019) found that the optimum ensemble size for Pysteps depends on the rainfall intensity threshold that is assessed. For low-intensity thresholds, there was only a marginal improvement between 24

and 48 ensemble members. This indicates that for these thresholds, the chosen ensemble size of 20 is probably sufficient. However, for higher thresholds (e.g., 5 mm hr^{-1}), there was even significant improvement in model performance when increasing the ensemble size to as much as 96 members (Figures 13 and 14 in Pulkkinen et al., 2019). Hence, when the nowcasting algorithm is used to forecast high-intensity rainfall events, a larger ensemble size is desirable. The downside of this choice would be that this might not be computationally feasible in an operational setting when a new set of nowcasts has to be made every 5 min, which is why an ensemble size of 20 members was chosen.

2.4. Verification Metrics

2.4.1. Pearson's Correlation

For every event per lead time t, Pearson's correlation coefficient (ρ) is calculated as

$$\rho = \frac{1}{N_f} \sum_{i=1}^{N_f} \frac{(F_i - \mu_F)(O_i - \mu_O)}{\sigma_F \sigma_O}, \tag{2}$$

where F_i and O_i are the forecast and observed rainfall amounts at a given grid cell, N_f corresponds to the number of forecasts with lead time t in the event, μ is the mean of the forecasts (μ_F) and observations (μ_O), and σ is the standard deviation of the forecasts (σ_F) and observations (σ_O) at a given grid cell. If this is calculated in a distributed manner, that is, per grid cell, it will result in a two-dimensional field with the correlation per grid cell. These numbers are then averaged over all grid cells, to obtain one averaged correlation per event.

As it is useful for an end-user to have an idea of the maximum lead time for which a forecast is still skillful, the 1/e-line ($\rho \approx 0.37$) is used as threshold (e.g., Berenguer et al., 2011; Germann & Zawadzki, 2002). Once the correlation drops below this line, generally referred to as the decorrelation time, the forecast is no longer seen as skillful. The lead time at which this occurs is the so-called decorrelation time of the forecast.

2.4.2. MAE and CRPS

For deterministic runs and per event, the mean absolute error (MAE) is calculated per lead time as

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N_f} |F_i - O_i|}{N_f}. \tag{3}$$

In the case of a probabilistic forecast, for example, for PS-P, the entire forecast distribution is available for comparison with the observations. In order to do this, the cdfs of forecast and observation are used. While the cdf of the observation is a single step-function, that is, there is only one value, the cdf of the probabilistic forecast is a curve. The area between these two cdfs is a measure for the continuous rank probability score (CRPS), which is formulated as

$$\mathrm{CRPS} = \frac{1}{N_f} \sum_{i=1}^{N_f} \int_{-\infty}^{+\infty} (P_{F_i}(x) - P_{O_i}(x))^2 \mathrm{d}\, x. \tag{4}$$

Here, $P_{F_i}(x)$ and $P_{O_i}(x)$ are the forecast and observed non-exceedance probability, for the ith forecast with lead time t. x is the forecast/observed rainfall sum, which is approximated numerically as interval with a step dx that is variable and depends on the rainfall sum per ensemble member. This decomposition to a stepwise function is explained in Hersbach (2000).

The advantage of using the CRPS is that it reduces to the MAE for deterministic forecasts, which enables the comparison between the MAE of the deterministic forecasts and the CRPS of the probabilistic forecasts.

2.4.3. Brier Score

When a forecast gives a 20% probability of rainfall, then ideally, it rains in 20% of the cases for which this forecast is issued. This gives a reliable forecast, whereas unreliable forecasts significantly deviate from this optimum. With a reliability diagram, this characteristic is tested by counting the number of observations that actually exceed a given threshold per forecast probability.

Simultaneously, this approach can be used to obtain an indication of the ensemble skill, as compared to a benchmark. Below the climatological frequency of exceeding a given rainfall threshold, the forecast is

unable to distinguish situations with different frequencies of occurrence: the point of no resolution. Additionally, there is a point of no skill, where the probabilistic forecast is not able to predict better than a reference (e.g., the climatology) whether an event will occur or not. This is tested with the Brier Skill Score (BSS), which is based on the Brier Score (BS) (Jolliffe & Stephenson, 2012):

$$\mathrm{BS} = \frac{1}{N_f} \sum_{i=1}^{N_f} (P_{F_i} - P_{O_i})^2, \tag{5}$$

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{6}$$

The BS is similar to the CRPS (Equation 4), but with the difference that the CRPS is the BS integrated over all thresholds. Thus, Equation 5 verifies whether for forecast i, a predefined threshold is exceeded by forecast P_{F_i} (given as a probability between 0 and 1) and by observation P_{O_i} (0 or 1). To put this in perspective, a reference ($\mathrm{BS}_{\mathrm{ref}}$) is used to determine the BSS. This reference can be the climatology, but also persistence, deterministic forecasts or probabilistic forecasts.

2.4.4. Fractions Skill Score

The Fractions Skill Score (FSS) is a spatial verification score which uses a fractions-based BS (see section 2.4.3) over successively larger cell lengths (Roberts & Lean, 2008). By increasing the length scale n (e.g., in km), the area used for verification increases, generally leading to a higher FSS value. n can increase up to $2N-1$, with N the longest length scale in the extent. The FSS ranges from 0 to 1, with 1 corresponding to a perfect forecast. With this metric, a minimum length scale to reach a required skill can be found, which is the target upscaling resolution of the data. For a predefined threshold and one forecast, it is calculated as

$$\mathrm{FSS}(n) = 1 - \frac{\mathrm{MSE}(n)}{\mathrm{MSE}_{\mathrm{ref}}(n)}, \tag{7}$$

where $\mathrm{MSE}(n)$ is the mean squared error between observed and forecast fractions for length scale n. $\mathrm{MSE}_{\mathrm{ref}}(n)$ is defined as a reference MSE for length scale n, which is the largest MSE that can be obtained from the observed and forecast fractions. It is formulated as

$$\mathrm{MSE}_{\mathrm{ref}}(n) = \frac{1}{N_x N_y}\left[\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} O_{i,j}^2(n) + \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} F_{i,j}^2(n)\right], \tag{8}$$

where N_x and N_y are the number of columns (x) and rows (y) in the radar composite, respectively. $O_{i,j}^2(n)$ and $F_{i,j}^2(n)$ are the observed and forecast fractions, per grid cell, of surrounding points up to length scale n that exceed a given rainfall intensity threshold. $O_{i,j}^2(n)$ is calculated as

$$O_{i,j}^2(n) = \frac{1}{n^2} \sum_{k=1}^{n}\sum_{l=1}^{n} I_O\left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2}\right]. \tag{9}$$

Here, I_O is the binary field of exceedances of a given rainfall intensity threshold for the observations. k and l are integer values ranging from 1 to length scale n, used to count the threshold exceedances for every cell within the window around cell (i, j). The equation for $F_{i,j}^2(n)$ is the same, with the only difference that not I_O, but I_F is used. See also Figure 2 in Roberts and Lean (2008) for a schematic example of the method. Generally, the FSS value above which the forecast is considered useful lies in between the perfect and the random forecast skill ($= 0.5 + \frac{f_0}{2}$). The random forecast skill, indicated with f_o, is defined as the domain average observed rainfall fraction above the threshold (Mittermaier & Roberts, 2010; Roberts & Lean, 2008).

2.4.5. Receiver Operating Characteristic

The receiver operating characteristic (ROC) curve analyses the predictive ability of exceeding a certain threshold with probabilistic forecasts. In essence, the curve plots the hit rate (HR) versus the false alarm rate (FAR) for predefined probability thresholds. HR is calculated as

$$HR = \frac{TP}{TP + FN}, \tag{10}$$

where TP is the number of true positives, the "hits": Both the forecast and observation exceed the threshold. FN is the number of false negatives, the "misses": The observation exceeds the threshold, but the forecast does not. The FAR is calculated as

$$FAR = \frac{FP}{FP + TN}, \tag{11}$$

where FP is the number of false positives, the "false alarms": The forecast exceeds the threshold, but the observation does not. TN is the number of true negatives: Neither the forecast nor the observation exceeds the threshold.

On or below the 1:1 line between HR and FAR, the forecast is not better than a random forecast (no skill). A higher skill leads to a larger area under the curve, which has a maximum value of 1.0 and a minimum target value of 0.5 (the 1:1 line).

2.5. Experimental and Forecast Verification Setup

The nowcasts for the 1,536 events were run (equally spread) on two high-performance clusters with Intel Xeon processors with 2.2 GHz and 8 GB memory per core and Intel Xeon processors with 3.6 GHz and 8 GB memory per core. Run times on these clusters for 5-min forecasts with a 6-hr lead time were on average (for one core): 40 s for RM-S, 130 s for RM-DR, 40 s for PS-D, and 1,250 s for PS-P (for 20 ensemble members). Hence, the run time for one PS-P nowcast took longer than the update frequency. From an operational perspective, this would require either reducing the forecast horizon (e.g., forecasts with a 3-hr lead time) or reducing the update frequency. Another option would be to run PS-P on multiple cores.

During 3 out of the 1,536 events, one algorithm did not finish successfully, leaving 1,533 events available for analysis. In these three cases (two failures for PS-D and one for PS-P), the initialization of the output file failed. After analyzing the log files, we concluded that this error was likely caused by the high-performance cluster rather than caused by the algorithm.

The nowcasts were produced with a lead time of 6 hr and a temporal resolution of 5 min (hence 72 lead times). Nowcasts were already initiated for the 6 hr prior to the onset of each event in order to have a 6-hr forecast for every time step within the event. Note that only forecasts for times within the actual event duration were analyzed. The durations are thus the time windows during which the nowcasts were analyzed, while nowcasts are made for a longer time frame around the events. In total, over 940,000 separate 5-min forecasts have been analyzed per algorithm. In the following paragraphs, the verification procedures for these analyses are briefly introduced.

2.5.1. Event Type and Duration Dependency

The events selected for the four durations contain different types of rainfall, generally from convective and small-scale for the shortest event durations (1 hr) to more stratiform, larger-scale systems for the longer event durations (e.g., 24 hr). As a result of the differences in rainfall types for these event durations, the mean decorrelation distance increases with almost a factor 2.5 from 1 to 24 hr in the Netherlands (Van de Beek et al., 2012). Based on this, we expected that this would lead to a difference in predictive skill of the nowcasts for the event durations. For the purpose of finding, per algorithm, the dependency of forecasting skill on the event type and duration, Pearson's correlation coefficient was calculated per lead time t and for every event (averaged over all grid cells within a catchment). As PS-P is a probabilistic run, Pearson's correlation was calculated for every ensemble member separately. As such, all separate model runs within the ensemble were taken into account. The average skillful lead time per algorithm and event duration could be estimated from the 1/e-line.

2.5.2. Seasonal Dependency

Rainfall characteristics and seasonal differences vary considerably between and within event durations (Van de Beek et al., 2012). Whereas winters in the Netherlands generally have widespread frontal, stratiform rainfall fields of low to intermediate intensity, summer rainfall also consists of more localized convective showers with higher rainfall intensities. We expected that this also impacts the nowcasts for the 12 catchments. To verify the nowcasts for the different seasons, we focused on one event duration: 6 hr. Within

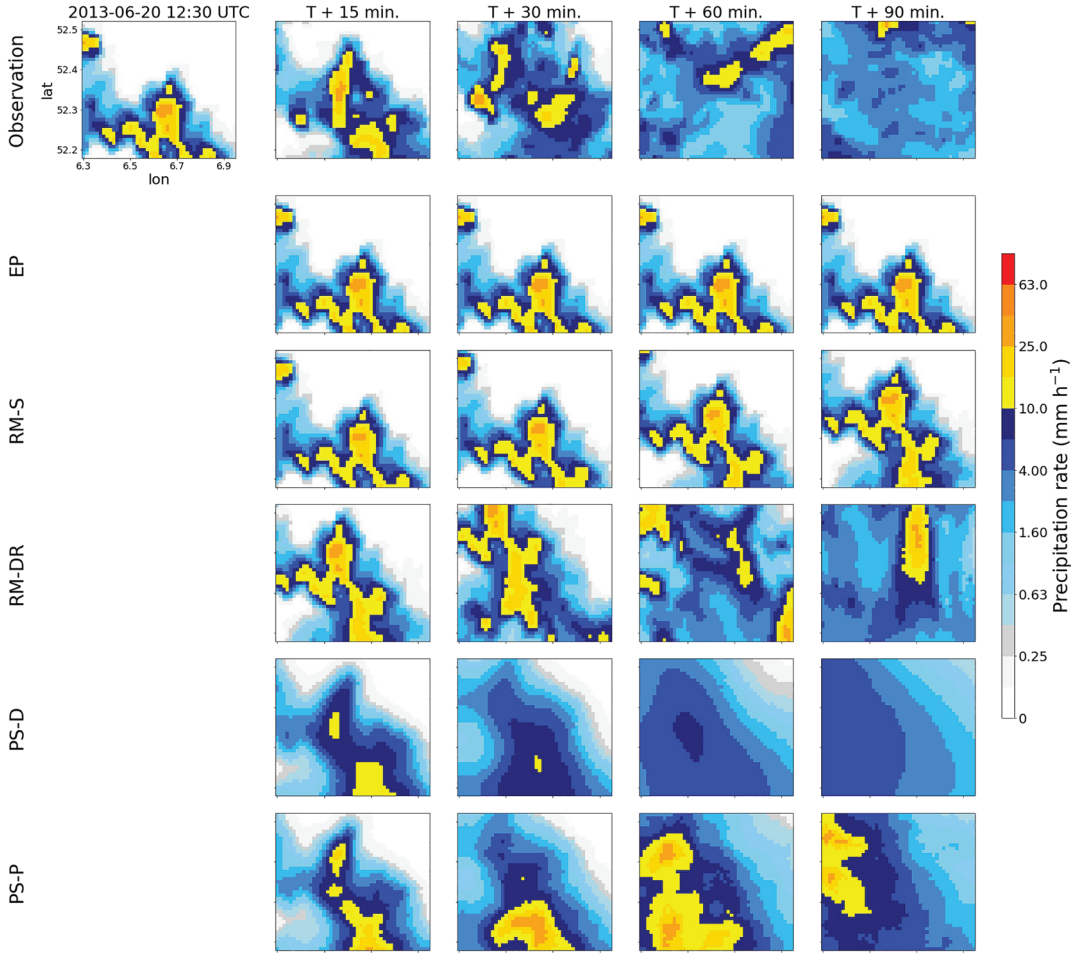

Figure 3. Example of a set of nowcasts for the Regge. The illustrated event took place on 20-06-2013 and resulted in an average of 29.4 mm over the area in 3 hr (between 12:05 and 15:05 UTC), with local maxima around 45 mm. For PS-P, only ensemble member 10 is shown.

this interval, the MAE and CRPS were calculated per catchment to estimate the error between forecasts and observations per event in a season.

2.5.3. Dependency on Catchment Size and Location

Vivoni et al. (2006) found that flood forecast skill increases with increasing catchment area. This may also be the case for the precipitation predictability of nowcasts. Within the field of NWP, it is common practice to upscale forecasts to a coarser resolution, as this gives a better representation of the rainfall fields when forecast rainfall fields are mislocated (e.g., Mittermaier, 2006). It is possible that the spatial resolution necessary for a minimum forecast skill is larger than the smallest catchments in this study (Figure 1). Hence, it is useful to find a minimum scale on which forecasts are still skillful in order to draw conclusions about the dependency of nowcast skill on catchment size.

For this analysis, the FSS was estimated for the two largest catchments (Aa and Regge), with a maximum length scale of 49 km (given a rectangular box around the catchments). The FSS was calculated for every odd number from 1 to 97 km for events with a 6-hr event duration. At 97 km ($2N-1$, with N the longest length scale in the Aa catchment, because of its elongated shape), the skill approaches an asymptote where FSS $= 1$, when the forecast is unbiased; that is, the fraction of observed rainfall exceeding the threshold over the entire domain is the same as the fraction of forecast rainfall exceeding this threshold. If not, asymptotic behavior will take place at a value lower than 1 (Mittermaier & Roberts, 2010; Roberts & Lean, 2008).

In addition to the catchment area, the location with regard to the radar location(s) and storm movement may influence the nowcast skill. To determine whether or not this relationship between location and skill is

present, the maximum skillful lead time (similar to section 2.5.1) was used for the 6-hr event duration. Since differences in catchment size would affect the results, the correlation and maximum skillful lead time are calculated for 5×5 cells in the center of the catchment, as this fits in the output extent of all 12 catchments. For the Hupsel Brook catchment (6.5 km^2), cells surrounding the catchment are used as well. Similar to section 2.5.1, both metrics are calculated for each ensemble member in the nowcasts of PS-P.

2.5.4. Ensemble Forecast Verification

Ensemble predictions are used to account for the uncertainty in predictions. As such, the ensemble spread gives the forecaster an indication of the uncertainty in the forecast. The ensemble mean often shows a better skill than a purely deterministic forecast (e.g., Richardson, 2000). An ensemble, however, is only useful when the ensemble has a representative spread, ideally with a minimal bias. In addition, an end-user needs to know how trustworthy a resulting forecast probability to exceed a certain rainfall threshold is. For this purpose, the reliability diagram and ROC curve were employed in this study for events with a 6-hr duration. Only PS-P is used for the ensemble forecast verification, since this is the only probabilistic nowcasting algorithm in this study.

3. Results

An example nowcast for an event (from the 24-hr duration) in the Regge catchment at 12:30 UTC on 20 June 2013 is shown in Figure 3. Although the event in Figure 3 is just one event out of the large sample, it gives a good example of the difficulty of forecasting convective precipitation affected by storm movement, growth, and dissipation, and merging and splitting of the precipitation systems. All algorithms have difficulties capturing these processes well. Whereas RM-DR, PS-D, and PS-P (ensemble member 10 is shown) seem to capture the movement to a certain extent, RM-S has the right direction, but almost no movement. Naturally, there is no movement for EP either. PS-D has, for this particular case, the disadvantage that there is too much dissipation, leading to the loss of the high-intensity rainfall centers while a mean large-scale field of rainfall persists. This is likely due to the short lifetime and small extent of the rain structures, which are decomposed into more quickly dissipating fields in PS-D. Between RM-DR and PS-P, which both capture the high-intensity rainfall cells, the main differences are the size and location of the rainfall systems. Based on visual inspection of this example, RM-DR approximates the observations for longer lead times best.

In the remainder of this section, the full sample of events is used for the verification of the forecasting skill of these nowcasting algorithms, but only the results for the 6-hr event duration are shown (except for section 3.1, as this section focuses on the different durations). Note that only the forecasts for times within the predefined events are analyzed here.

3.1. Event Type and Duration Dependency

With increasing event duration, the decorrelation time increases (Figure 4). Maximum skillful lead times, seen as the mean of the intersections between the 1/e-line and the mean correlation of an event, increase from 25 min for 1-hr durations, to 40 min for 3-hr durations, 56 min for 6-hr durations, and 116 min for 24-hr durations. In all cases, PS-D attains the longest skillful lead times.

The type of rainfall system determines the difference between these event durations. Whereas the shortest durations generally consist of short-lifetime high-intensity convective precipitation events, the longer durations consist of larger, more persistent systems that generally have a higher predictability.

The correlation varies between events, and this variability decreases with increasing event duration, as indicated with the colored error bars in Figure 4. This indicates that small-scale systems with shorter lifetimes vary more between events, leading to more variability in the nowcasting results. This sometimes leads to significant negative correlations for EP and RM-S, meaning that forecast and observed rainfall in cells have the opposite tendency, for example, decreasing rainfall amounts in the forecast while the observations have increasing amounts.

Although the correlations of PS-D and PS-P are quite similar for the shorter event durations, the attained correlations for the 6- and 24-hr durations are 15–25% lower for PS-P than for PS-D. Note that comparing a deterministic with a probabilistic run is not entirely fair, because the main advantages of a probabilistic run are not weighed. Between those two algorithms and RM-DR, the difference is considerable, with maximum skillful lead times that are generally 35–60% lower than the Pysteps algorithms. This suggests that

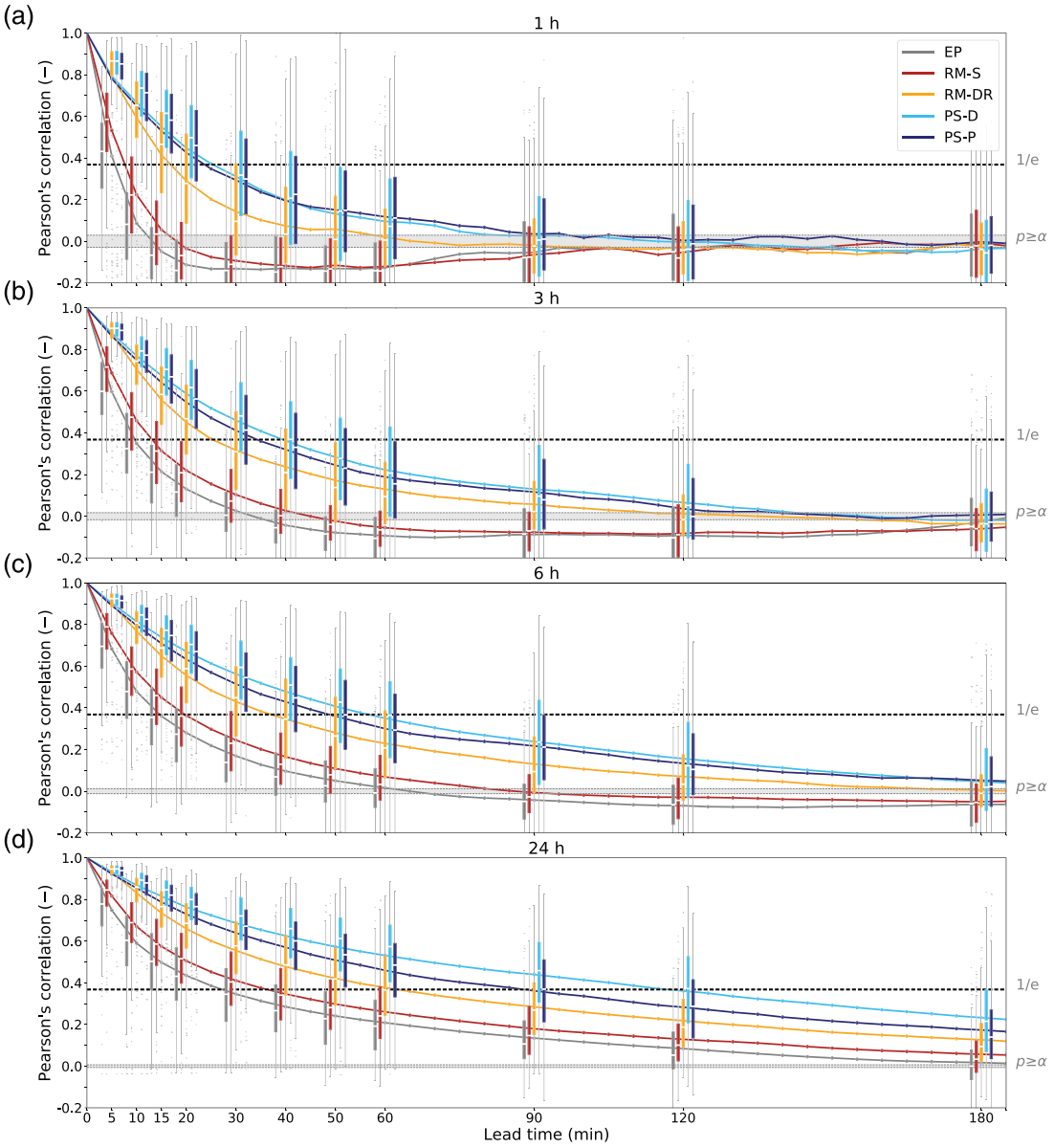

Figure 4. Pearson's correlation as a function of lead time (5-min steps), averaged over all cells within the catchment and events (in that order), for event durations of 1-hr (a), 3-hr (b), 6-hr (c), and 24-hr (d). The dotted line indicates a correlation of 1/e, the minimum correlation for a skillful nowcast. The boxes indicate the variability in results per event, with the median in white, the interquartile (25th to 75th percentile) range (IQR) in colored boxes, 1.5 × IQR starting outside the boxes in gray bars, and the outliers in gray dots. The horizontal gray band around a correlation of 0.0 indicates correlations that do not differ significantly from 0.0, based on a two-tailed *T*-test with $\alpha = 5\%$.

taking spatial and temporal scales into account, as done in the Pysteps algorithms, adds value to only rotation-permitting advection. However, we have run Pysteps with only advection (similar to RM-DR), and it performs slightly better than RM-DR according to some metrics (see Figures S1 and S2 in the supporting information), indicating that spatial and temporal scale decomposition in Pysteps is not the only explanation for this difference.

Compared to the other three algorithms, RM-S attains lower correlations, with values closer to EP than to the other algorithms. Compared to RM-DR, maximum skillful lead times are generally a factor 2 smaller. Based on these results, it seems that using a corner tracking method for the optical flow leads to lower correlations than using global optical flow algorithms. It was expected that EP performs worse than the other

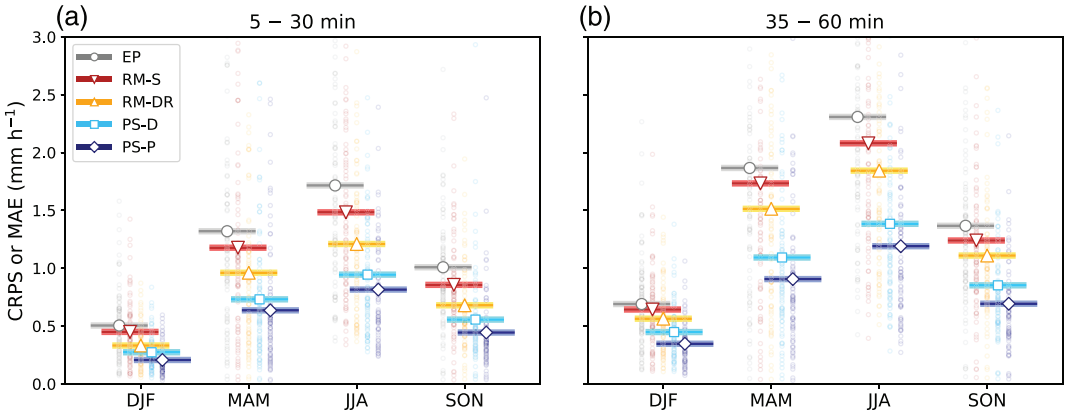

Figure 5. CRPS and MAE per season for all events and catchments for the 6-hr event duration, averaged over lead times of 5–30 min (a) and 35–60 min (b). The MAE is shown for all deterministic runs and the CRPS for PS-P. The thick lines with a marker indicate the mean CRPS or MAE for all runs and catchments in that season. The scattered points are the mean CRPS or MAE per event.

nowcasting algorithms. However, skillful lead times still reach 25 min for events with long durations. For the event duration of, for example, 1 hr, on the other hand, skillful lead times are generally close to 5 min.

3.2. Seasonal Dependency

There are considerable differences in forecast errors between the seasons (Figure 5). The forecast errors are lowest during winter with event- and catchment-averaged MAE and CRPS values between 0.2 and 0.5 mm hr^{-1} (Figure 5a). Summers have the highest forecast errors with MAE and CRPS values on average between 0.8 and 1.7 mm hr^{-1}. This difference is caused by the variation in precipitation types between seasons in the Netherlands, leading to higher rainfall intensities during summer (Table 1), and an increase in the spatial and temporal variability of the rainfall fields. Generally, frontal systems cause the rainfall in the Dutch winter, whereas scattered convective rain showers are more dominant during summer, especially for situations with high rainfall sums. With more persistent rainfall fields in frontal systems than in convective systems, the predictability of these systems is higher, undoubtedly leading to lower forecast errors.

Spring has relatively high errors as well, with MAE and CRPS values on average only 22% lower than during summer, caused by the increasing contribution of convective showers during this season. MAE and CRPS during fall are in between winter and summer, when high rainfall sums are often caused by storms that originate from frontal zones with additional convective input from the relatively warm seawater.

For longer lead times, that is, 35–60 min (Figure 5b), the relative difference between the seasons remains the same. However, the errors increase for longer lead times with approximately 45% for all seasons, which is caused by the decreasing skill of the nowcasting algorithms for longer lead times (see, e.g., Figure 4).

The difference between the algorithms is consistent over all seasons and lead times, with EP having the highest MAE values and PS-P always having the lowest CRPS values. This difference between highest (EP) and lowest (PS-P) errors is generally a factor 2 or more. It is remarkable that the performance of PS-P is better than PS-D here, while this was the opposite in Figure 4. This may be caused by the bias insensitivity of the correlation metric, which is accounted for by the MAE and CRPS (for a quantification of the biases, see Figure S5 in the supporting information).

3.3. Dependency on Catchment Size and Location

3.3.1. Catchment Size

Corresponding to Roberts and Lean (2008) and Mittermaier and Roberts (2010), the FSS increases with increasing length scale, that is, after upscaling of the model simulations to a coarser resolution (Figure 6). Additionally, FSS decreases with increasing lead time. These relationships together give an indication of the minimum length scale required to reach a skillful forecast, that is, FSS $\geq 0.5 + \frac{f_0}{2}$, for lead time t.

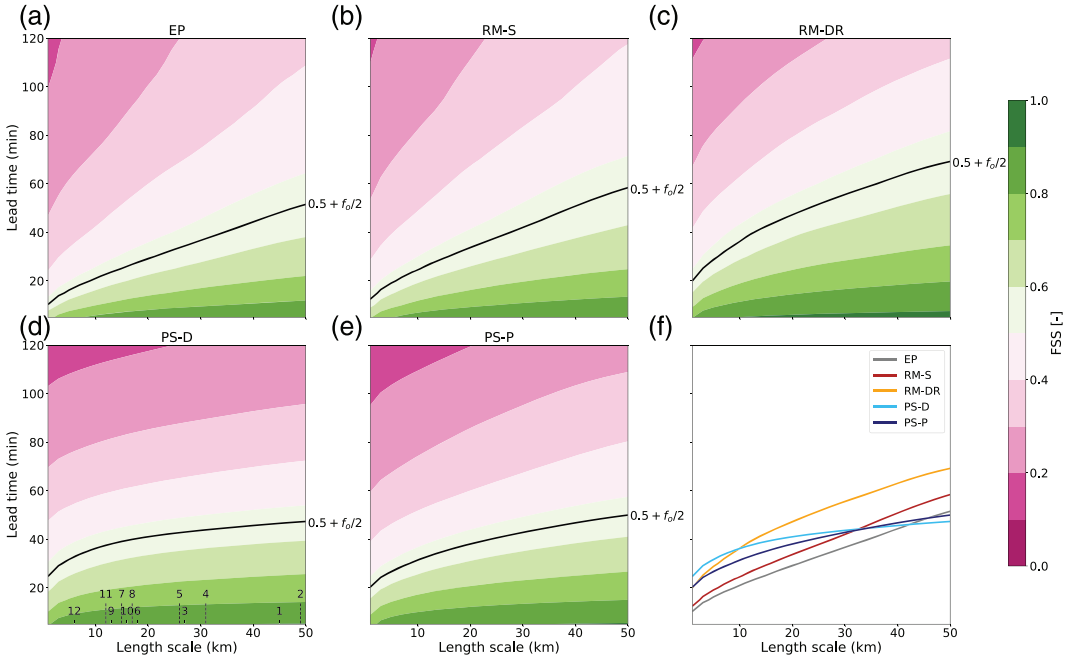

Figure 6. Fractions Skill Score (FSS) as a function of lead time and catchment length scale (mean of all events). Plots are made for the 6-hr durations and for a threshold of 1.0 mm hr^{-1} for the five algorithms (a–e), based on the nowcasts for the catchments Aa and Regge. The black contour line (at FSS $= 0.5 + \frac{f_0}{2}$) indicates the minimum FSS to derive a skillful spatial scale. In panel (f), the contour lines of the algorithms are combined to facilitate comparison. In (d), the longest length scale present in all catchments is indicated with the catchment number (same as in Figure 1).

With the FSS metric in Figure 6, we focus on the absolute differences, that is, biases, between forecast and observations, and, due to the upscaling procedure, the FSS is less sensitive to spatial differences caused by the mislocation of forecast rainfall fields. It returns the minimum length scale for upscaling in order to reach a required skill (e.g., FSS $\geq 0.5 + \frac{f_0}{2}$), which has a hydrological relevance as this directly links to catchment sizes. Therefore, it is of interest to find out whether an FSS $\geq 0.5 + \frac{f_0}{2}$ is actually achievable for the range of catchment sizes in this study given a desired skillful forecast horizon (i.e., lead time t).

For instance, for a skillful forecast horizon of 60 min, Figure 6c shows that the event-averaged minimum length scale, indicated with a black line at FSS $= 0.5 + \frac{f_0}{2}$, is approximately 36 km for RM-DR. Hence, to still have a skillful nowcast for this lead time, the nowcast has to be upscaled to 36×36 km^2. Upscaling to this length scale is only possible for the two largest catchments in this study, although the total area is already larger than the catchment areas (Figure 1). For the other catchments, this means that the upscaling requirement is considerably larger than taking the catchment-averaged rainfall, thus making it (on average) impossible to reach skillful forecasts of 60 min for those catchments. Note that an FSS of $0.5 + \frac{f_0}{2}$ (for a skillful forecast horizon of 60 min) is not attained with the other algorithms for any upscaling length scale up to 50 km.

Table 3 indicates the lead time for which an FSS of at least $0.5 + \frac{f_0}{2}$ is attained for a set of length scales (as shown in Figure 6). After upscaling to 10×10 km^2, which is already larger than the catchment size of seven of the studied catchments, the maximum skillful lead times range from 21 min for EP to 37 min for RM-DR and PS-D. Above a length scale of 10 km, RM-DR outperforms all other methods. This is clearly a different perspective than found in Figures 4 and 5 and caused by the stronger bias in the nowcasts of the Pysteps algorithms (not shown here, see Figure S5).

Table 3
Indication of the Maximum Lead Time for Which an FSS of at Least $0.5+\frac{f_0}{2}$ Can Be Attained for a Set of Length Scales, That Is, Upscaling Resolutions

	Max. skillful lead time (min)					
Algorithm	1 km	10 km	20 km	30 km	40 km	50 km
EP	8	21	30	37	43	51
RM-S	10	24	34	43	51	58
RM-DR	20	37	45	55	62	70
PS-D	25	37	41	43	45	48
PS-P	20	35	38	43	46	50

Despite the increase of FSS with increasing length scale, a maximum FSS of 1.0 will not be attained when the forecast is biased (Mittermaier & Roberts, 2010; Roberts & Lean, 2008). All algorithms underestimate the rainfall in the forecast rainfall fields for a threshold of 1.0 mm hr^{-1} (as used in Figure 6), especially in the presence of growth and dissipation processes during the event. However, the underestimations of the rainfall volumes for lead times exceeding 20 min are considerably higher for PS-D (see Figure S5). To a lesser extent, this is also the case for PS-P. This effect is partly caused by the dissipation of the smaller-scale rainfall fields; that is, these fields have a shorter lifetime in the Pysteps algorithms. Especially PS-D tends to end up with lower rainfall volumes due to an excess of smoothing in the forecasts. In addition, both PS-D and PS-P use probability matching, which fixes the number of wet pixels in the forecast to the number of wet pixels in the latest available observation. The rainfall pixels that have left the domain are subtracted from this number.

Because of this bias, PS-D generally has the highest FSS values on a length scale smaller than 5 km, due to the smallest displacement error in the forecasts, but for larger length scales, RM-DR starts to outperform the Pysteps algorithms. RM-DR has a smaller bias and therefore a steeper increase in the FSS with increasing length scale (due to the effective correction for mislocation with increasing length scale).

For a length scale of 30 km, which corresponds to approximately upscaling to the largest catchment in this study (Regge), a maximum skillful lead time of 55 min can be attained with the best performing algorithm (RM-DR), whereas this is approximately 30 min when the nowcast is upscaled to the area of the Hupsel Brook catchment (6.5 km^2; the best performing algorithm is PS-D in this case). Hence, higher skill can be reached for the larger catchments in this study when the forecasts are upscaled.

3.3.2. Catchment Location

The catchment location with regard to the prevailing wind direction and the proximity to the upwind edge of the radar domain matters in most cases (Figure 7). For the 6-hr event duration, the prevailing wind direction is southwest (Figure 7a). This directly affects the average maximum skillful lead time of the four algorithms considered, with mean skillful lead times increasing from 20–30 min to more than 45 min in the downwind direction (sizes of the circles in Figure 7b), and for the northwest (Beemster) and southeast (Roggelsebeek) of the country (Figure 7b). The catchments located upwind are closer to the edge of the radar domain. This means that some rainfall fields are not yet present in the radar mosaic when the nowcast is issued. The available rainfall fields generally also have biased rainfall amounts, as the QPE quality deteriorates toward the edge of the domain. This inevitably leads to less skill of the nowcast with increasing lead times.

The four quarters in the circles in Figure 7b indicate the maximum skillful lead time for the algorithms RM-S, RM-DR, PS-D, and PS-P, based on the mean of all events. Per algorithm, a similar tendency of increasing skill in the downwind direction is present. For none of the catchments, however, RM-S has a skillful forecast beyond 30 min. On the other hand, the nowcasts of PS-D and PS-P are in most cases skillful up to more than 50 min toward the (north)east of the Netherlands. The average maximum skillful lead times are higher for PS-D than for PS-P (based on 5×5 center cells instead of the entire catchment), which corresponds to Figure 4.

3.4. Ensemble Forecast Verification

The reliability diagram in Figure 8a illustrates that all four 30-min intervals, up to a lead time of 120 min, have a positive BSS. This means that compared to the climatological frequency of exceeding the threshold of 1.0 mm hr^{-1}, all forecasts with PS-P have skill up to at least 2 hr ahead. However, note that the sharpness of the forecasts, the tendency to forecast with probabilities near 0% or 100%, decreases with increasing lead time (Figure 8b). This is especially the case for forecasts with high probabilities of exceeding the threshold, as the number of forecasts with a probability close to or at 100% reduces.

For probabilities less than 50%, the forecast probability is smaller than the observed frequency. Contrarily, the observed frequency is generally smaller than the forecast probability for probabilities exceeding 50% (70% for 5–30 min). This particular shape of the curve reveals that the ensemble is under-dispersive, meaning that the observed rainfall amount falls outside the ensemble spread of PS-P in many forecasts (see also

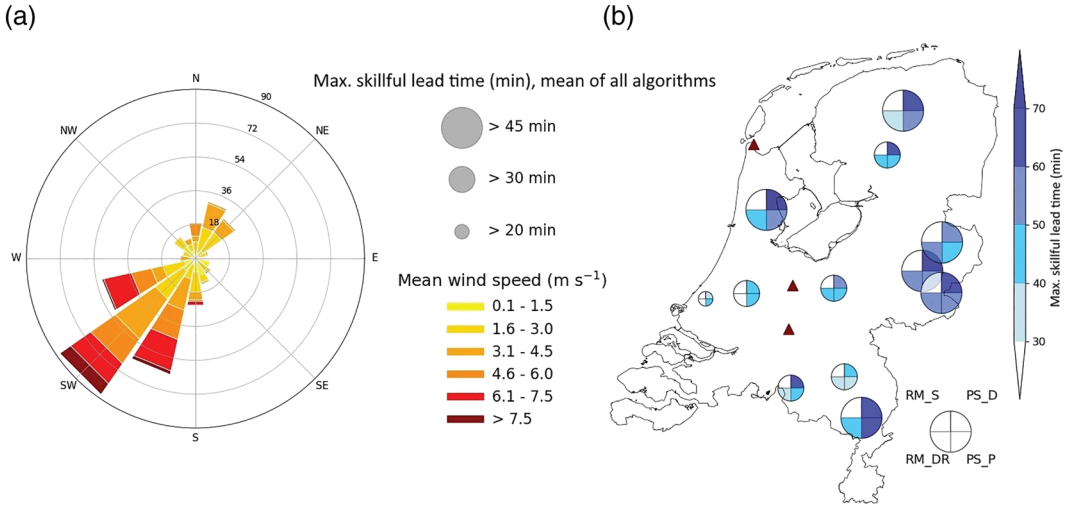

Figure 7. (a) Wind rose indicating the most frequent wind directions at KNMI station De Bilt during the events with the 6-hr duration. The length of the bars is an indication of the number of events with that wind direction. The hue is an indication of the mean wind speed. (b) The mean maximum skillful lead time of all 6-hr events for the 5×5 center cells per catchment (size of the circles indicating the average of the four algorithms) and per algorithm (hue in the quarters). EP is left out of this analysis. The red triangles indicate the locations of the radars.

supporting information Figure S6). To overcome this, the ensemble should be either wider, that is, the standard deviation of the probabilistic forecast should be larger by including more members or more noise per member, or, in case of a systematic bias, the error between ensemble mean and the observation should reduce.

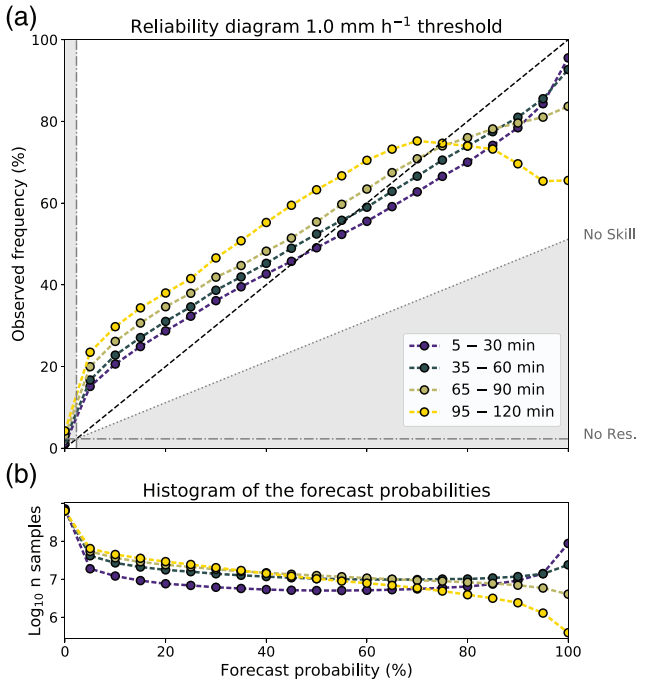

Figure 8. (a) Reliability diagram of exceeding a threshold of 1.0 mm hr^{-1} for events within the 6-hr event duration for PS-P. The reference is the climatological frequency of exceeding this threshold in the studied events. The mean probabilities for lead times of 30 min are shown. (b) Histogram indicating the number of times a probability was forecast by an ensemble member.

The ROC curve in Figure 9 also indicates skillful probabilistic forecasts, with an area under the curve ranging from 0.95 (5–30 min) to 0.81 (95–120 min) for the analyzed events, indicating a good discrimination skill of the ensemble for cells that exceed the threshold of 1.0 mm hr^{-1}.

The largest distance between the HR and FAR, seen as the optimal forecast of exceeding the threshold, lies around a forecast probability (the circles in the graph) of 10% to 20%. Although that is an unconfident forecast, the FAR of these forecasts is generally low (smaller 0.2), and HR is always larger than 0.6 (often larger than 0.8). Forecasts with higher probabilities, that is, more confidence, have a lower HR and FAR. Toward these higher probabilities, the difference between the shorter lead times (5–30 min) and the longest (95–120 min) is that the HR exceeds 0.5 for all forecast probabilities (the dots in Figure 9) for shorter lead times. However, for the longest lead times (95–120 min), the HR reduces to 0 for the highest forecast probabilities. Hence, whereas a confident forecast with a forecast probability of 100% can still be useful for a forecaster for lead times between 5 and 30 min, it becomes worthless toward 2 hr ahead, as the HR is almost zero. Note, however, that hardly any nowcast has a probability larger than 50% for longer lead times.

4. Discussion

4.1. Relation to Previous Work

With 1,536 events in this study (of which 1,533 are analyzed), a statistical foundation is available which allows for testing the hypothesized dependencies of nowcast skill on event duration, season, catchment location, and catchment size. In this section, we explore how the findings in this study relate to other studies.

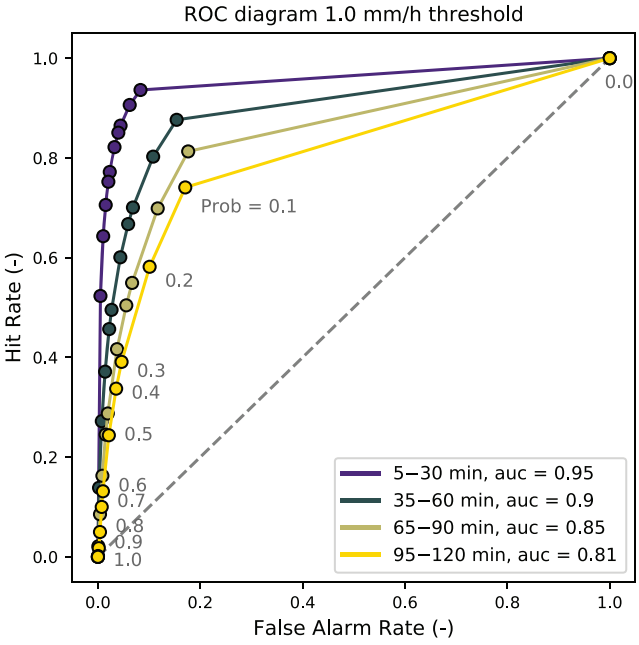

Figure 9. ROC curve of exceeding a threshold of 1.0 mm hr^{-1} with the nowcasts of PS-P for the events within the 6-hr event duration. The numbers indicate the forecast probabilities of exceeding this threshold (-). auc indicates the area under the curve: an indication of the skill of the probabilistic forecast. Shown are the mean rates for lead times of half an hour.

The maximum skillful lead times found for the event durations of 1 and 3 hr (25 and 40 min) show similarities with the approximately 30-min applicability of nowcasts found for convective systems in, for example, Liguori and Rico-Ramirez (2012), Foresti et al. (2016), Mejsnar et al. (2018), and Ayzel et al. (2019). For longer event durations, maximum skillful lead times are somewhat lower than previously found for large-scale persistent systems in the continental United States (skillful lead times ranged from approximately 4 to 8 hr Germann & Zawadzki, 2002) and for stratiform events in Barcelona (approximately 3 to 4 hr Berenguer et al., 2011). In the first case, the radar mosaic extent is much larger (continental United States) than in this study, inevitably leading to longer skillful lead times. However, the convective events in Berenguer et al. (2011) resulted in skillful lead times of 30–90 min, which is quite similar to the results in this study. Also note that the maximum skillful lead times in the aforementioned papers are based on Zawadzki's (1973) correlation, which is the correlation without subtraction of the mean (instead of Pearson's correlation as used in this study) and yields on average 13% higher correlations (Mejsnar et al., 2018).

In addition, the maximum skillful lead times are based on Pearson's correlation and the intersection with the 1/e-line in this study. While this approach makes a comparison with the aforementioned literature possible, it should be noted that these maximum skillful lead times depend on the chosen metric (and the somewhat arbitrary 1/e-line). Both the analysis with the FSS (Figure 6) and the use of the Critical Success Index (CSI; Figures S3 and S4 in the supporting information) lead to different maximum skillful lead times and smaller differences between the algorithms, although the ranking between the algorithms remains the same when the CSI is considered. Hence, with the employed metric (Pearson's correlation), we are only able to provide an indication of the differences between the event durations. The actual skillful lead times are indicative and depend on the focus of the reader.

The probabilistic runs with Pysteps have resulted in lower skillful lead times than with the deterministic runs. We have to note that comparing a probabilistic run with a deterministic one is not a fair comparison, because it neglects the major advantage of probabilistic forecasts, that is, the uncertainty estimate. When the ensemble mean is used (not shown), the probabilistic runs give similar skillful lead times and even higher skillful lead times for durations of 6 hr or more. This effect of using the ensemble mean is in agreement with, for example, Richardson (2000) and is in the advantage of using probabilistic forecasts, as they also contain information about the uncertainty of the forecast.

On a seasonal scale, we find more skill during winter than during summer, which is expected seeing the increasing decorrelation distance from summer to winter in Van de Beek et al. (2012). For regions with a temperate climate and a similar difference between winter and summer precipitation types, we expect similar results.

With regard to the catchment size dependency, the nowcasts have to be upscaled to better represent the rainfall fields, as is the case for high-resolution NWP forecasts (e.g., Mittermaier, 2006). This means that the smallest catchments in this study can become smaller than the cell size of the upscaled rainfall fields, while upscaling is still possible for larger catchments. Similar behavior was found for flood forecast skill with increasing catchment area by Vivoni et al. (2006). Earlier studies have also suggested that forecast skill and uncertainty of nowcasting algorithms depend on location (Foresti et al., 2016; Germann et al., 2009). We find this in this study too, with increasing forecast skill in the downwind direction of the operational radars with south-westerlies as the main wind direction. Hence, whereas the application of nowcasting in flood forecasting is likely to be beneficial (e.g., Berenguer et al., 2005; Liguori et al., 2012; Moreno et al., 2013; Pierce et al., 2005; Poletti et al., 2019; Vivoni et al., 2006, 2007), the catchment properties will influence the eventual skill.

4.2. The Catchment Perspective and Resulting Event Selection

The focus on catchments instead of the full radar domain is interesting from a hydrological perspective as the statistics are directly tailored to the involved catchments, including the dependency on their sizes and locations. Additionally, the event selection procedure and the resulting rainfall forecasts can be directly applied in a follow-up hydrological analysis for the same basins.

The chosen approach, however, limits the analysis of the size and location dependency to a more exploratory phase, solely indicating the presence of relations between catchment size and location, and forecast skill (section 3.3). It is recommended to continue with the focus on these relationships, as was also done by Foresti and Seed (2015) for a mountainous area near Melbourne in Australia. This requires using the full radar domain to find the statistics and identify these relationships on a larger domain. Understanding these relationships on this domain will make it possible to correct the nowcasts in real time (via, e.g., bias corrections or machine learning techniques) and to better take uncertainties into account. Note that such a procedure would change the event selection procedure to, for example, a national level and it would substantially increase the storage requirements for this number of events.

The systematic event selection procedure ensures reproducibility, and it allows for an equal number of events in all event durations and seasons. However, within the selected event durations, continuous rainfall was not a requirement. This means that not the full nowcasting time is used for forecasting and analysis of periods with rainfall, although this has the advantage that it allows for testing whether false positives occur (rain forecast, but not observed). Ideally, only the actual event, that is, from the start until the end of rainfall, is part of the nowcast and thus analyzed. This, however, also has as disadvantage that the classification in durations (1 hr, 3 hr, etc.) becomes less clear.

Moreover, the choice to select the events based on both catchment-averaged maxima and grid cell maxima has merely to do with the subsequent step in this project: the hydrological application of these nowcasts. The involved water authorities that manage the studied catchments have different hydrological models and water management systems, which require either lumped or gridded rainfall input. It would have been possible to conduct this study with events based on either catchment-averaged or pixel maxima as input.

4.3. Transferability of Results to Other Regions

Although this study focuses on the Netherlands, the results should be transferable to other regions with a temperate climate and with similar radar products. It is noteworthy that the Netherlands is a lowland country and that the results from this study will likely not hold for mountainous regions. In mountainous regions, growth and decay processes dominate over the advection of rainfall fields (e.g., Foresti & Seed, 2015; Foresti et al., 2018). Hence, larger errors are expected for nowcasts in these regions, which affects the skillfulness of the forecasts.

4.4. Dependency on Radar QPE Product

The QPE product in this study consists of two radars with a radial extent of approximately 320 km (of which only the first 200 km is used in the composite). At this moment, an improved operational product is available (but not yet archived for a longer period), which also includes two Belgian and one German radar located relatively close to the Dutch border. The expectation is that this will increase skillful lead times, especially for catchments that are located further away in the upstream direction of the radar. In the Netherlands, that is most often toward the southwest (see section 3.3.2), but the expected results are of course non-exclusive to this region. Hence, the location dependency is expected to change due to this improvement.

A second potential improvement is that this product has an automatic bias correction based on measured precipitation amounts from 32 KNMI automatic rain gauges at WMO weather stations. This will not make any difference for the results of this study, because the QPE is used as reference. Nevertheless, for a hydrological application, obtaining the true precipitation volumes does matter. Most radar products underestimate the precipitation volumes, so we expect that a bias-corrected QPE product leads to larger discharge volumes and therefore better hydrological simulations than with the QPE in this study.

4.4.1. Three-Dimensional Data Input

All algorithms in this study make use of two-dimensional rainfall fields. There are also nowcasting algorithms that make use of the entire volumetric radar scan. TITAN is an example of such an algorithm

(Dixon & Wiener, 1993). The advantage of three-dimensional data is that also the heterogeneity in the vertical direction, that is, on different elevations, can be used. This would allow for physically based corrections for, for example, the vertical profile of reflectivity. In most cases, however, the volumetric data are not or only marginally corrected for the often substantial errors. From that perspective, the post-processed two-dimensional fields have an advantage, too. It is also noteworthy that nowcasting with two-dimensional fields comes with lower computational requirements. Ideally, corrections (e.g., clutter- and bias-correction) already take place on the original volumetric radar scans. This would allow for a better use of all information present in the radar scans, and it would allow for a fair comparison between centroid tracking algorithms such as TITAN and the cross-correlation algorithms that are used in this study.

5. Conclusion and Future Perspectives

In this study, the skill of radar rainfall nowcasting in predicting rainfall up to 6 hr ahead was tested with a large-sample analysis. In total, 1,536 events were run (of which 1,533 successfully completed and thus were analyzed) spread over 4 event durations (1, 3, 6, and 24 hr) and 4 seasons for 12 lowland catchments in the Netherlands, a country with a temperate maritime climate. Four algorithms were tested and compared to Eulerian Persistence (EP), which is the "poor man's" approach of using the most recent radar QPE as forecast. The tested algorithms were Rainymotion Sparse (RM-S), Rainymotion DenseRotation (RM-DR), Pysteps deterministic (PS-D; similar to S-PROG), and Pysteps probabilistic (PS-P) with 20 ensemble members. Model performance was assessed by a verification with the radar QPE, which was assumed to be the observed rainfall amount. The focus in this study was on finding the relationship between nowcast skill and dependencies on event duration, season, catchment size, and location with regard to the radar location and prevailing wind direction. In addition, the ensemble forecasts with PS-P were analyzed.

Pearson's correlation is used to study the maximum skillful lead time up to which the forecast is still seen as useful. This average maximum skillful lead time increases with increasing event duration (in an absolute sense), with 25 min for events with a 1-hr duration, 40 min for 3-hr, 56 min for 6-hr, and 116 min for 24-hr event durations. The reason for this increase in maximum skillful lead time is the increasing persistence, that is, the increasing spatial extent and temporal scale of the rainfall fields, of events with longer durations. These maxima are in all cases found for PS-D, although PS-P shows similar performance for the 1-hr event duration. For longer event durations, the average maximum skillful lead times of PS-P are generally 15–25% lower. Compared to RM-DR, which still outperforms RM-S by a factor 2 and EP by more than a factor 2, the average maximum skillful lead time of forecasts with Pysteps algorithms is generally 35–60% higher. Given these maximum skillful lead times, improvements such as blending with NWP (for lead times shorter than 3 hr) are clearly necessary to bridge the gap with the 3- to 6-hr skillful lead time desired for these very short range forecasts.

Both the event duration and the season are found to affect the skill of the nowcasts. During winter, when more persistent frontal, stratiform rainfall is present, average MAEs and CRPSs are a factor 3 lower than during summers, with generally more convective rainfall with higher intensities. The rainfall predictability during spring, when the number of convective showers increases, is relatively low, with MAE and CRPS values closer to summer (a 22% difference) than to winter. Forecast errors during autumn are more in between winter and summer and thus lower (by 26%) than during spring. This is due to more persistent autumn storms originating from frontal zones with additional convection due to the relatively warm seawater. The nowcast results indicate a consistent performance difference between the algorithms, with from high to low performance the following ranking: PS-P, PS-D, RM-DR, RM-S, and EP.

Although PS-P and PS-D have shown the longest skillful lead times and the lowest forecasts errors over the seasons, most forecasts have to be upscaled for optimal use, which affects the minimal spatial scale on which the forecasts can be properly used. For all algorithms in this study, the forecast generally has to be upscaled in order to reach an FSS of at least $0.5+\frac{f_0}{2}$ (with f_0 the random forecast skill), the minimal FSS for a skillful forecast. The maximum skillful lead time that we have found after upscaling to a cell size comparable to the catchment area of the smallest catchment (Hupsel Brook, 6.5 km^2) is 30 min, while this is 55 min after upscaling to a cell size comparable to the largest catchment (Regge, 957 km^2). Thus, if upscaling is possible, higher skill can be attained for larger catchments than for smaller ones. For upscaling resolutions of more

than $10 \times 10\,km^2$, RM-DR has outperformed all other algorithms. This effect results from the stronger bias present in the Pysteps forecasts for increasing lead times, which has a pronounced influence on the FSS. It is noteworthy that all algorithms have a bias toward lower rainfall volumes, which is not necessarily higher for Pysteps than for the other algorithms for small thresholds. However, for a threshold of 1.0 mm hr^{-1} or higher, especially PS-D has (for lead times of ≥20 min) a considerably stronger underestimation of the rainfall volumes than the other algorithms.

Besides the catchment area, the catchment location with regard to the proximity to the upwind edge of the radar domain and the prevailing wind direction (SW) also matters. The prevailing south-westerlies affect the mean skillful lead times of the nowcasts with skillful lead times of 20–30 min in the southwest of the Netherlands to more than 45 min in the (north)east. For water managers in the southwest of the country, it is therefore recommended to work with a radar mosaic that incorporates the radar in Jabbeke, in the northwest of Belgium (e.g., used in Foresti et al., 2016). Note that with respect to the catchment size and location dependency, this study is limited to a focus on catchments and polders. A more complete statistical analysis of these spatial dependencies requires the usage of the full radar domain in the analysis.

As for the ensemble predictions, PS-P has been the only probabilistic nowcasting algorithm in this study. The ensemble of this algorithm turns out to be reliable up to at least 120 min ahead for rainfall amounts of ≥1.0 mm hr^{-1}. However, for all tested 30-min intervals the ensemble is under-dispersive, which indicates that the ensemble spread should be wider if the error between observation and ensemble does not change. After 60 min, the ensemble looses its sharpness regarding the higher probabilities: Forecast probabilities of exceeding 1.0 mm hr^{-1} are rarely (close to) 100%. Moreover, optimal forecasts, that is, with the largest HR to FAR ratios, are found around forecast probabilities of 10% to 20%.

This study has shown that there is a clear advantage in using a global optical flow algorithm (RM-DR) over a corner detecting method (RM-S). In most cases, PS-D and PS-P are able to outperform the "benchmark" algorithms RM-S and RM-DR. Most errors present in the nowcasts are a result of growth and dissipation processes, which are not or only stochastically (e.g., PS-D and PS-P) taken into account in the algorithms. Although PS-P makes a good step toward accounting for many of the uncertainties in the current nowcasting procedures, there is still much to gain with the ensemble. An increasing focus on nowcast uncertainties is therefore recommended in order to further improve probabilistic radar rainfall nowcasts.

Acknowledgments

This study was supported by funding from the DAISY2-project, supported by the European Regional Development Fund (Grant PROJ-00581). We would like to thank Hidde Leijnse (KNMI) and the Rainymotion and Pysteps community (in particular Daniele Nerini and Georgy Ayzel) for their valuable input. Catchment data, information, and feedback were provided by a group of potential end-users of the nowcasting products, consisting of the national meteorological institute, several Dutch water authorities, and consultancy firms: KNMI, Rijkswaterstaat, Hoogheemraadschap Delfland, Hoogheemraadschap Rijnland, Waterschap Limburg, Hoogheemraadschap Hollands Noorderkwartier, Waterschap Aa en Maas, Waterschap De Dommel, Wetterskip Fryslân, Waterschap Noorderzijlvest, Waterschap Rijn en IJssel, Waterschap Vallei en Veluwe, Waterschap Vechtstromen, Royal Haskoning DHV, and Hydrologic. Finally, we would like to thank Jonathan J. Gourley (A.E.), Maik Heistermann, Michael Dixon, and one anonymous reviewer for their constructive feedback and interest in our work.

Data Availability Statement

The used radar data in this study are available via https://dataplatform.knmi.nl/catalog/datasets/index.html?x-dataset=rad_nl25_rac_mfbs_em_5min&x-dataset-version=2.0 (gauge-adjusted QPE archive) and https://doi.org/10.4121/uuid:05a7abc4-8f74-43f4-b8b1- 7ed7f5629a01(unadjusted, operational, radar data). The daily quality controlled observations of KNMI can be obtained online (via https://dataplatform.knmi.nl/catalog/datasets/index.html?x-dataset=etmaalgegevensKNMIstations&x-dataset-version=1). Rainymotion and Pysteps are available online (at https://www.doi.org/10.5281/zenodo.2561582 and https://www.doi.org/10.5281/zenodo.2631910). Model configurations and run scripts (Python) can be found online (at https://www.doi.org/10.5281/zenodo.3826582).

References

Atencia, A., & Zawadzki, I. (2014). A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Monthly Weather Review*, *142*(11), 4036–4052. https://doi.org/10.1175/MWR-D-13-00117.1

Atencia, A., & Zawadzki, I. (2015). A comparison of two techniques for generating nowcasting ensembles. Part II: Analogs selection and comparison of techniques. *Monthly Weather Review*, *143*(7), 2890–2908. https://doi.org/10.1175/MWR-D-14-00342.1

Ayzel, G., Heistermann, M., & Winterrath, T. (2019). Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0.1). *Geoscientific Model Development*, *12*(4), 1387–1402. https://doi.org/10.5194/gmd-12-1387-2019

Beekhuis, H., & Holleman, I. (2008). From pulse to product, highlights of the digital-IF upgrade of the Dutch national radar network. In *Proceedings of the Fifth European Conference on Radar in Meteorology and Hydrology (ERAD 2008)*, Helsinki, Finland.

Beekhuis, H., & Mathijssen, T. (2018). From pulse to product, highlights of the upgrade project of the Dutch national weather radar network. In L. de Vos, H. Leijnse, & R. Uijlenhoet (Eds.), *10th European Conference on Radar in Meteorology and Hydrology (ERAD 2018): 1–6 July 2018, Ede-Wageningen, The Netherlands* (pp. 960–965). Wageningen, the Netherlands: Wageningen University & Research. https://doi.org/10.18174/454537

Berenguer, M., Corral, C., Sánchez-Diezma, R., & Sempere-Torres, D. (2005). Hydrological validation of a radar-based nowcasting technique. *Journal of Hydrometeorology*, *6*(4), 532–549. https://doi.org/10.1175/JHM433.1

Berenguer, M., & SempereTorres, D. (2013). Radar-based rainfall nowcasting at European scale: Long-term evaluation and performance assessment, *36th Conference on Radar Meteorology* (pp. 15B.3–1–15B.3–7). Breckenridge, CO: American Meteorological Society.

Berenguer, M., Sempere-Torres, D., & Pegram, G. G. S. (2011). SBMcast—An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *Journal of Hydrology*, *404*(3-4), 226–240. https://doi.org/10.1016/j.jhydrol.2011.04.033

Berenguer, M., Surcel, M., Zawadzki, I., Xue, M., & Kong, F. (2012). The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. Part II: Intercomparison among numerical models and with nowcasting. *Monthly Weather Review*, *140*(8), 2689–2705. https://doi.org/10.1175/MWR-D-11-00181.1

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*, *132*(620), 2127–2155. https://doi.org/10.1256/qj.04.100

Buishand, T. A., & Velds, C. A. (1980). *Klimaat van Nederland 1: Neerslag en verdamping*. De Bilt, The Netherlands: Royal Netherlands Meteorological Institute (KNMI).

Dixon, M., & Wiener, G. (1993). TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, *10*(6), 785–797. https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2

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*, *20*(1), 505–527. https://doi.org/10.5194/hess-20-505-2016

Foresti, L., & Seed, A. (2015). On the spatial distribution of rainfall nowcasting errors due to orographic forcing. *Meteorological Applications*, *22*(1), 60–74. https://doi.org/10.1002/met.1440

Foresti, L., Sideris, I. V., Panziera, L., Nerini, D., & Germann, U. (2018). A 10-year radar-based analysis of orographic precipitation growth and decay patterns over the Swiss Alpine region. *Quarterly Journal of the Royal Meteorological Society*, *144*(716), 2277–2301. https://doi.org/10.1002/qj.3364

Germann, U., Berenguer, M., Sempere-Torres, D., & Zappa, M. (2009). REAL—Ensemble radar precipitation estimation for hydrology in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, *135*(639), 445–456. https://doi.org/10.1002/qj.375

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*, *130*(12), 2859–2873. https://doi.org/10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2

Germann, U., & Zawadzki, I. (2004). Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *Journal of Applied Meteorology*, *43*(1), 74–89. https://doi.org/10.1175/1520-0450(2004)043<0074:SDOTPO>2.0.CO;2

Germann, U., Zawadzki, I., & Turner, B. (2006). Predictability of precipitation from continental radar images. Part IV: Limits to prediction. *Journal of the Atmospheric Sciences*, *63*(8), 2092–2108. https://doi.org/10.1175/JAS3735.1

Hazenberg, P., Leijnse, H., & Uijlenhoet, R. (2011). Radar rainfall estimation of stratiform winter precipitation in the Belgian Ardennes. *Water Resources Research*, *47*, W02507. https://doi.org/10.1029/2010WR009068

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Heuvelink, D., Berenguer, M., Brauer, C. C., & Uijlenhoet, R. (2020). Hydrological application of radar rainfall nowcasting in the Netherlands. *Environment International*, *136*, 105,431.

IPCC (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. In C. B. Field et al. (Eds.). Cambridge, United Kingdom and New York, NY: Cambridge University Press.

IPCC (2013). *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. In T. F. Stocker et al. (Eds.). Cambridge, United Kingdom and New York, NY: Cambridge University Press.

IPCC (2014). *Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Climate Change 2014: Impacts, Adaptation and Vulnerability*. In C. B. Field et al. (Eds.). Cambridge, United Kingdom and New York, NY: Cambridge University Press.

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Chichester, West Sussex, United Kingdom: John Wiley & Sons.

KNMI (2011). Klimaatatlas: Langjarige gemiddelden 1981–2010. http://www.klimaatatlas.nl/

KNMI (2015). *KNMI'14: Climate scenarios for the Netherlands; A guide for professionals in climate adaption*. De Bilt, The Netherlands: KNMI.

Kroeger, T., Timofte, R., Dai, D., & VanGool, L. (2016). Fast optical flow using dense inverse search. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *European Conference on Computer Vision, Lecture Notes in Computer Science* (Vol. 9908, pp. 471–488). Amsterdam, The Netherlands: Springer.

Liguori, S., & Rico-Ramirez, M. A. (2012). Quantitative assessment of short-term rainfall forecasts from radar nowcasts and MM5 forecasts. *Hydrological Processes*, *26*(25), 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*, *103*, 80–95. https://doi.org/10.1016/j.atmosres.2011.05.004

Lin, C., Vasić, S., Kilambi, A., Turner, B., & Zawadzki, I. (2005). Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, *32*, L14801. https://doi.org/10.1029/2005GL023451

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence* (pp. 121–130). San Fransisco, CA. Morgan Kaufmann Publishers Inc.

Marshall, J. S., Hitschfeld, W., & Gunn, K. L. S. (1955). Advances in radar weather. In *Advances in geophysics* (Vol. 2, pp. 1–56). New York, NY: Academic Press, Inc.

Mejsnar, J., Sokol, Z., & Minářová, J. (2018). Limits of precipitation nowcasting by extrapolation of radar reflectivity for warm season in Central Europe. *Atmospheric Research*, *213*, 288–301. https://doi.org/10.1016/j.atmosres.2018.06.005

Mittermaier, M. P. (2006). Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmospheric Science Letters*, *7*(2), 36–42. https://doi.org/10.1002/asl.127

Mittermaier, M. P., & Roberts, N. (2010). Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Weather and Forecasting*, *25*(1), 343–354. https://doi.org/10.1175/2009WAF2222260.1

Moreno, H. A., Vivoni, E. R., & Gochis, D. J. (2013). Limits to flood forecasting in the Colorado Front Range for two summer convection periods using radar nowcasting and a distributed hydrologic model. *Journal of Hydrometeorology*, *14*(4), 1075–1097. https://doi.org/10.1175/JHM-D-12-0129.1

Overeem, A., Buishand, T. A., & Holleman, I. (2009). Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar. *Water Resources Research*, *45*, W10424. https://doi.org/10.1029/2009WR007869

Overeem, A., Holleman, I., & Buishand, A. (2009). Derivation of a 10-year radar-based climatology of rainfall. *Journal of Applied Meteorology and Climatology*, *48*, 1448–1463. https://doi.org/10.1175/2009JAMC1954.1

Overeem, A., Leijnse, H., & Uijlenhoet, R. (2011). Measuring urban rainfall using microwave links from commercial cellular communication networks. *Water Resources Research*, *47*, W12505. https://doi.org/10.1029/2010WR010350

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., & Thielen, J. (2015). The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, *51*, 278–291. https://doi.org/10.1016/j.envsci.2015.04.016

Pierce, C., Bowler, N., Seed, A., Jones, A., Jones, D., & Moore, R. (2005). Use of a stochastic precipitation nowcast scheme for fluvial flood forecasting and warning. *Atmospheric Science Letters*, *6*(1), 78–83. https://doi.org/10.1002/asl.102

Pierce, C., Seed, A., Ballard, S., Simonin, D., & Li, Z. (2012). Nowcasting. In *Doppler radar observations-weather radar, wind profiler, ionospheric radar, and other advanced applications*. London, UK: IntechOpen. https://doi.org/10.5772/39054

Poletti, M. L., Silvestro, F., Davolio, S., Pignone, F., & Rebora, N. (2019). Using nowcasting technique and data assimilation in a meteorological model to improve very short range hydrological forecasts. *Hydrology and Earth System Sciences*, *23*(9), 3823–3841. https://doi.org/10.5194/hess-23-3823-2019

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667. https://doi.org/10.1002/qj.49712656313

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Seed, A. W. (2003). A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology*, *42*(3), 381–388. https://doi.org/10.1175/1520-0450(2003)042<0381:ADASSA>2.0.CO;2

Seed, A. W., Pierce, C. E., & Norman, K. (2013). Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resources Research*, *49*, 6624–6641. https://doi.org/10.1002/wrcr.20536

Serafin, R. J., & Wilson, J. W. (2000). Operational weather radar in the United States: Progress and opportunity. *Bulletin of the American Meteorological Society*, *81*(3), 501–518. https://doi.org/10.1175/1520-0477(2000)081<0501:OWRITU>2.3.CO;2

Shi, J., & Tomasi, C. (1994). Good features to track, *IEEE Conference on Computer Vision and Pattern Recognition*(pp. 593–600). Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPR.1994.323794

Sokol, Z., Mejsnar, J., Pop, L., & Bližňák, V. (2017). Probabilistic precipitation nowcasting based on an extrapolation of radar reflectivity and an ensemble approach. *Atmospheric research*, *194*, 245–257. https://doi.org/10.1016/j.atmosres.2017.05.003

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*, *43*(2), 231–248. https://doi.org/10.1175/1520-0450(2004)043<0231:POPFCR>2.0.CO;2

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), 3837–3850. https://doi.org/10.5194/amt-9-3837-2016

Van de Beek, C. Z., Leijnse, H., Torfs, P. J. J. F., & Uijlenhoet, R. (2012). Seasonal semi-variance of Dutch rainfall at hourly to daily scales. *Advances in Water Resources*, *45*, 76–85. https://doi.org/10.1016/j.advwatres.2012.03.023

Venugopal, V., Foufoula-Georgiou, E., & Sapozhnikov, V. (1999). Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysical Research*, *104*(D24), 31,599–31,610. https://doi.org/10.1029/1999JD900437

Vivoni, E. R., Entekhabi, D., Bras, R. L., Ivanov, V. Y., Van Horne, M. P., Grassotti, C., & Hoffman, R. N. (2006). Extending the predictability of hydrometeorological flood events using radar rainfall nowcasting. *Journal of Hydrometeorology*, *7*(4), 660–677. https://doi.org/10.1175/JHM514.1

Vivoni, E. R., Entekhabi, D., & Hoffman, R. N. (2007). Error propagation of radar rainfall nowcasting fields through a fully distributed flood forecasting model. *Journal of Applied Meteorology and Climatology*, *46*(6), 932–940. https://doi.org/10.1175/JAM2506.1

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., & Heynert, K. (2013). The Delft-FEWS flow forecasting system. *Environmental Modelling & Software*, *40*, 65–77. https://doi.org/10.1016/j.envsoft.2012.07.010

Wong, W. K., Cheng, V. T. L., & Woo, W. C. (2016). Community SWIRLS Nowcasting System (Com-SWIRLS). In *WMO WWRP 4th International Symposium on Nowcasting and Very-short-range Forecast 2016 (WSN2016)*, Hong Kong.

Zawadzki, I. I. (1973). Statistical properties of precipitation patterns. *Journal of Applied Meteorology*, *12*(3), 459–472. https://doi.org/10.1175/1520-0450(1973)012<0459:SPOPP>2.0.CO;2