TU Delft

Delft University of Technology

Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning

Imhoff, R. O.; Brauer, C. C.; van Heeringen, K. J.; Uijlenhoet, R.; Weerts, A. H.

DOI
10.1029/2021WR031591

Publication date
2022

Document Version
Final published version

Published in
Water Resources Research

Citation (APA)
Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., & Weerts, A. H. (2022). Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning. *Water Resources Research*, *58*(3), Article e2021WR031591. https://doi.org/10.1029/2021WR031591

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Check for updates

AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031591

Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning

R. O. Imhoff[1,2], C. C. Brauer[1], K. J. van Heeringen[2], R. Uijlenhoet[3], and A. H. Weerts[1,2]

[1]Hydrology and Quantitative Water Management Group, Wageningen University & Research, Wageningen, The Netherlands, [2]Department of Operational Water Management & Early Warning, Unit of Inland Water Systems, Deltares, Delft, The Netherlands, [3]Department of Water Management, Delft University of Technology, Delft, The Netherlands

Key Points:

- The potential of nowcasting for discharge forecasting was evaluated with 659 rain events spread over 12 lowland catchments
- Discharge forecast errors depend on initial catchment wetness and increase with both rainfall intensity and spatial variability
- Discharge exceedance thresholds can be on average forecast between 119 and 223 min earlier than with no rainfall forecast at all

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

R. O. Imhoff,
Ruben.Imhoff@deltares.nl

Citation:

Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., & Weerts, A. H. (2022). Large-sample evaluation of radar rainfall nowcasting for flood early warning. *Water Resources Research*, *58*, e2021WR031591. https://doi.org/10.1029/2021WR031591

Received 9 NOV 2021
Accepted 6 MAR 2022

Author Contributions:

Conceptualization: R. O. Imhoff, C. C. Brauer, K. J. van Heeringen, R. Uijlenhoet, A. H. Weerts
Formal analysis: R. O. Imhoff
Funding acquisition: R. Uijlenhoet, A. H. Weerts
Investigation: R. O. Imhoff
Methodology: R. O. Imhoff, C. C. Brauer, K. J. van Heeringen, R. Uijlenhoet, A. H. Weerts
Resources: R. Uijlenhoet, A. H. Weerts
Software: C. C. Brauer, K. J. van Heeringen
Supervision: C. C. Brauer, K. J. van Heeringen, R. Uijlenhoet, A. H. Weerts

© 2022. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract To assess the potential of radar rainfall nowcasting for early warning, nowcasts for 659 events were used to construct discharge forecasts for 12 Dutch catchments. Four open-source nowcasting algorithms were tested: Rainymotion Sparse (RM-S), Rainymotion DenseRotation (RM-DR), Pysteps deterministic (PS-D), and probabilistic (PS-P) with 20 ensemble members. As benchmark, Eulerian Persistence (EP) and zero precipitation input (ZP) were used. For every 5-min step in the available nowcasts, a discharge forecast with a 12-hr forecast horizon was constructed. Simulations using the observed radar rainfall were used as reference. Rainfall and discharge forecast errors were found to increase with both increasing rainfall intensity and spatial variability. For the discharge forecasts, this relationship depends on the initial conditions, as the forecast error increases more quickly with rainfall intensity when the groundwater table is shallow. Overall, discharge forecasts using RM-DR, PS-D, and PS-P outperform the other methods. Threshold exceedance forecasts were assessed by using the maximum event discharge as threshold. Compared to benchmark ZP, an exceedance is, on average, forecast 223 (EP), 196 (RM-S), 213 (RM-DR), 119 (PS-D), and 143 min (PS-P) in advance. The EP results are counterbalanced by both a high false alarm ratio (FAR) and inconsistent forecasts. Contrarily, PS-D and PS-P produce lower FAR and inconsistency index values than all other methods. All methods advance short-term discharge forecasting compared to no rainfall forecasts at all, though all have shortcomings. As forecast rainfall volumes are a crucial factor in discharge forecasts, a future focus on improving this aspect in nowcasting is recommended.

Plain Language Summary Flood warnings in quickly responding catchments are still challenging, as the timing and location of the rainfall forecast should be as accurate as possible. Radar rainfall nowcasting, a technique to statistically extrapolate the most recent rainfall observations, can potentially improve the rainfall forecasts up to several hours ahead. To evaluate the benefits and possible pitfalls of radar rainfall nowcasting for hydrological forecasting, we tested four different nowcasting algorithms for 659 rainfall events in the Netherlands. We used these nowcasts for the construction of discharge forecasts for 12 Dutch lowland catchments. Based on the evaluation of the large sample of discharge forecasts for these catchments, we conclude that all nowcasting methods advance short-term discharge forecasting compared to no rainfall forecasts at all. Nevertheless, all techniques have shortcomings. Rainfall and discharge forecast errors increase with increasing rainfall intensity and spatial variability, though discharge forecast errors also strongly depend on the initial catchment wetness. Moreover, forecast rainfall volumes have shown to be a crucial factor in the quality of the discharge forecast. Hence, rainfall nowcasting can be a valuable addition to hydrological forecasting systems. Yet we recommend a future focus on improving area-averaged rainfall volumes in nowcasting algorithms to further advance hydrological forecasting.

1. Introduction

The livability and economy of many areas worldwide are endangered by floods (Ceola et al., 2014; European Environment Agency, 2004; Jongman et al., 2012; Merz et al., 2010; Ward et al., 2013). Especially pluvial floods that occur on a short timescale, typically in small, urban, mountainous, and polder catchments, are difficult to predict (e.g., Cox et al., 2002; Ferraris et al., 2002). Consequently, this makes an adequate and timely anticipation by water authorities challenging. Such floods, generally caused by intense precipitation events, are expected to become more severe and occur more frequently in a changing climate (Arnell & Gosling, 2016; Hirabayashi et al., 2013; Klein Tank et al., 2014). Risk and damage can be reduced when a well-established flood early warning system (e.g., Delft-FEWS, Werner et al., 2013) is in place, which can make it possible to act timely

Validation: R. O. Imhoff
Visualization: R. O. Imhoff, C. C. Brauer
Writing – original draft: R. O. Imhoff
Writing – review & editing: C. C. Brauer, K. J. van Heeringen, R. Uijlenhoet, A. H. Weerts

(Pappenberger et al., 2015; UNISDR, 2002). In less extreme situations water managers and indirectly citizens can also benefit from improved hydrological predictions and early warnings on the short term for, for example, real-time control of the water system.

Flood early warning systems are only beneficial if the underlying hydrological forecasts are accurate, timely, and reliable. Uncertainty in the hydrological forecast originates from either the used hydrological model, for example, as a result of model structure, initial conditions, setup or calibration procedures (Beven, 1993; Clark et al., 2017; Melsen et al., 2016), or the precipitation forcing. Regarding the forcing, particularly the rainfall forecast, a phenomenon which is highly variable in space and time, is uncertain and significantly influences the forecast quality (e.g., Moulin et al., 2009; Sampson et al., 2014). Hence, improving rainfall forecasts on the short term is expected to result in better hydrological predictions.

Most early warning systems, if present at all, use short-range (12–72 hr) numerical weather prediction (NWP) model output as quantitative precipitation forecast (QPF). On the short term (up to approximately 6 hr ahead), the QPFs of the NWP models are often not sufficiently accurate for reliable early warnings. This is due to either one or all of the following reasons: (a) a too coarse temporal resolution, (b) a too low update frequency, or (3) the mislocation of rainfall events (Berenguer et al., 2012; Lin et al., 2005; Pierce et al., 2012; Roberts & Lean, 2008). An example of a too low update frequency can be found in the Netherlands, where the NWP model HARMONIE (Bengtsson et al., 2017) currently has an update frequency of 6 hr and regularly arrives at the end users 4 hr after the issue time of the forecast. Within those 4 hr, let alone the 6-hr validity of the forecast after that, initial conditions may have changed significantly, especially during convective rainfall events, leading to forecast errors already at the start of the issue time of the hydrological forecast (Sun et al., 2014).

These issues can be tackled by taking advantage of the following simultaneous developments: (a) increasingly rapid update cycle NWP models, (b) nowcasting, possibly incorporating machine learning techniques, and (c) a blended system using the former two (e.g., Bowler et al., 2006; Germann & Zawadzki, 2002; Golding, 1998; Sun et al., 2014; Turner et al., 2004). In this study, we will only focus on nowcasting and its potential for hydrological forecasting. Nowcasting is the (statistical) process of extrapolating real-time remotely sensed quantitative precipitation estimates (QPEs) into the future. Generally, QPE from weather radars is used for this due to the high spatial and temporal resolution of current radar rainfall products (typically 1 km and 5 min; Serafin & Wilson, 2000; Overeem, Holleman, & Buishand, 2009). The skill of nowcasting depends on a variety of environmental characteristics, such as season, event duration, scale of the rainfall system, size of the target location, and location in the radar composite with regard to the storm direction. Maximum skillful lead times generally range from less than 30 min for convective storms, to approximately 2 hr for larger-scale and more persistent rainfall events, up to a maximum of 6 hr for persistent stratiform events on a continental scale (Ayzel et al., 2019; Berenguer et al., 2011, 2012; Foresti et al., 2016; Germann et al., 2006; Germann & Zawadzki, 2002; Imhoff et al., 2020; Liguori & Rico-Ramirez, 2012; Lin et al., 2005; Mejsnar et al., 2018).

Nowcasted rainfall has already been successfully used as input for various hydrological models and forecasting systems (Berenguer et al., 2005; Germann et al., 2009; Heuvelink et al., 2020; Liguori et al., 2012; Liguori & Rico-Ramirez, 2013; Moreno et al., 2013; Pierce et al., 2005; Poletti et al., 2019; Sharif et al., 2006; Vivoni et al., 2006, 2007). Berenguer et al. (2005) and Heuvelink et al. (2020) have found significant improvements in discharge forecasts, with a gain in anticipation time of 10–170 min, depending on the catchment and event type. However, Berenguer et al. (2005) have also concluded that despite the improvement in rainfall forecast with the S-PROG model (Seed, 2003) compared to simple Lagrangian persistence, there is little difference between both methods when their QPFs are used for hydrological forecasts, because S-PROG tends to underestimate the rainfall volumes. This stresses the importance of rainfall volume forecasts for hydrological applications. In addition, the interplay of catchment properties (initial conditions, response times, and management, etc.) with the storm characteristics determine the hydrological predictability (Moreno et al., 2013).

Despite the insights gained from the aforementioned studies, all these studies are based on relatively small sample sizes of one to six events. Vivoni et al. (2006), Poletti et al. (2019), and Heuvelink et al. (2020) even recommend an analysis with a larger sample of events to draw statistically meaningful conclusions. Hence, in this study we aim to evaluate the potential added value of radar rainfall nowcasting for flood early warning based on a large sample of events. In particular, we will focus on the dependence of the nowcasts and subsequent hydrological forecast skill on both storm and catchment characteristics.

In a previous study, we analyzed nowcasts for 1,500+ events spread over 12 catchments in the Netherlands to evaluate rainfall predictability (Imhoff et al., 2020). Four open-source nowcasting algorithms were used: two benchmarking advection algorithms from the Rainymotion library (Ayzel et al., 2019) and two from the Pysteps library (Pulkkinen et al., 2019). This study will build on that by using the nowcasts from this large sample of events and by applying them to the hydrological models used in the operational systems of the involved Dutch water authorities of these 12 catchments (sizes varying from 6.5 to 957 km^2). To the authors' knowledge, this is the first hydrological application and systematic evaluation of radar rainfall nowcasting with a combination of such a large sample of events and this variety of nowcasting algorithms.

The outline of this paper is as follows: Section 2 describes the study area, the available nowcasts, the underlying radar rainfall product, and the experimental and forecast verification setup. This is followed by the results in Section 3, the discussion in Section 4, and the conclusions in Section 5.

2. Materials and Methods

2.1. Study Area

The study area is the same as in Imhoff et al. (2020) and comprises 12 lowland catchments in the Netherlands (Figure 1 and Table 1). These catchments are a combination of polders and (partially) freely draining catchments. The selection of these 12 catchments was based on their location (spread over the country) and was achieved in close collaboration with the water authorities that were involved in this study. The 12 catchments vary in size, from only 6.5 km^2 for the Hupsel Brook catchment to 957 km^2 for the Regge catchment. Although only small variations occur in the mean annual rainfall amounts in the Netherlands, the water balance varies much between catchments (Table 1). This is a result of intensive regulation in the Netherlands, especially the polders. Such regulations can consist of weirs, pumps, and surface water supply to prevent flooding, to support agriculture, or to deal with salt water intrusion (i.e., flushing). In addition, groundwater flow across basin boundaries can be substantial, in particular upward seepage in low-lying polders.

Two typical regulated polders are Beemster and Gouwepolder (Table 1). The Beemster is a deep polder, which is mostly covered by grass fields. It has a rather constant upward seepage of brackish water that is constantly flushed during the drier summer half year by pumping water into and out of the polder area. The Gouwepolder is a small polder that is mostly used for arboriculture. Because of its land use, groundwater tables are artificially kept at 0.7 m below the surface level. The polder has a slightly higher elevation than the surrounding polders, which leads to a downward seepage of water toward the other polders. Surface water is supplied to compensate for this loss. From both polders, water is discharged through pumps, which have discrete on/off settings, leading to a discharge time series that follows this on/off behavior. Therefore, discharge of these polders can only be analyzed, and compared to the other catchments, when summed over multiple time steps in order to get a more hydrograph-like discharge pattern in comparison to the on-off behavior of the pumps on a continuous timescale.

Freely-draining catchments are the Hupsel Brook, Grote Waterleiding, Luntersebeek, Roggelsebeek, Reusel, Aa, and Regge, though surface water supply and surface water level management play a role in these basins as well. The remaining three catchments, notably Dwarsdiep, Linde, and Delfland, are partially freely draining. These three catchments consist of subcatchments that are freely draining and subcatchments that have pumps to regulate surface water supply and discharge. Delfland is a special case in the sense that the majority of the southwest of the catchment consists of greenhouses, leading to a fast response of runoff to rainfall (Table 1).

The large variety in catchment characteristics has a pronounced effect on the rainfall nowcast skill (e.g., their locations with respect to the radars and areas, see Sections 2.5.3 and 3.3.2 of Imhoff et al., 2020). This also holds for the effect on the resulting discharge forecast skill, but these catchment-specific characteristics are hard to isolate with regard to their effect on the discharge forecast skill. Therefore, we decided not to isolate the effects of catchment characteristics on the discharge forecast skill in this study, but rather focus on the overall (potential) skill of radar rainfall nowcasting for discharge forecasting for the presented wide variety of lowland catchments in the Netherlands.

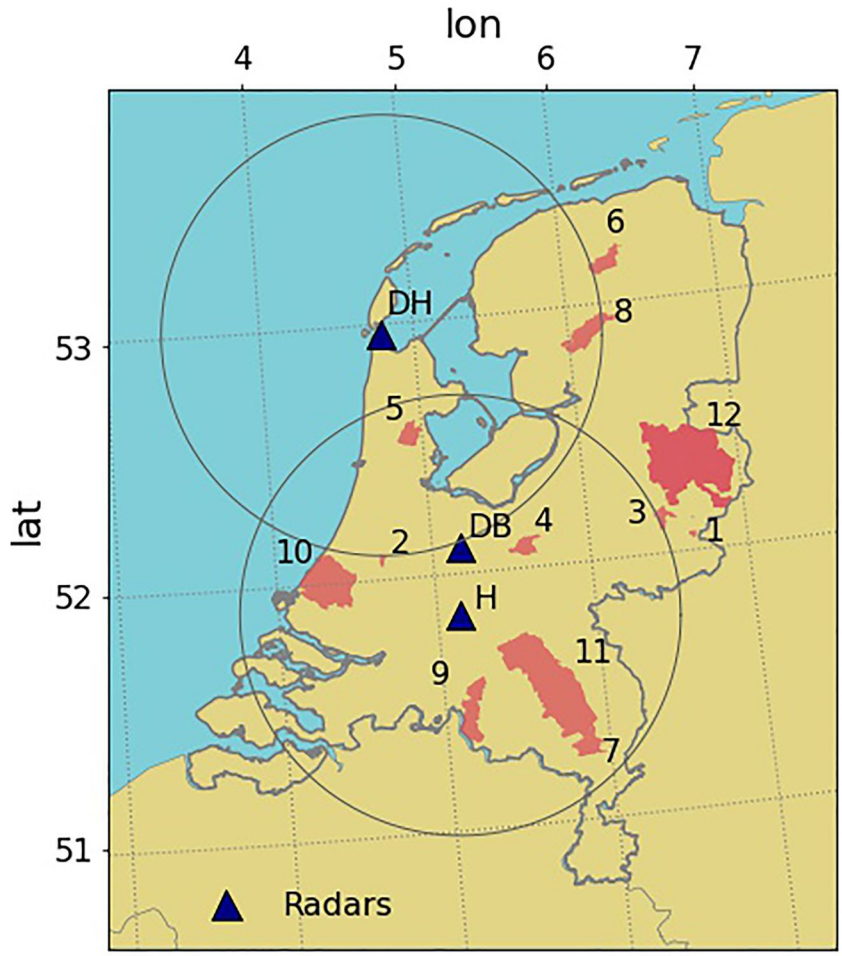

Figure 1. Overview of the 12 catchments in this study (red polygons). Catchment information is provided in Table 1. The three blue triangles indicate the locations of the three radars operated by the Royal Netherland Meteorological Institute: Den Helder (DH), De Bilt (DB; used until 2017), and Herwijnen (H; replaced the radar in DB). The two gray circles indicate a range of 100 km around radars DH and H, generally regarded as the distance up to which the radar quantitative precipitation estimate is expected to be reliable. The actual range of both radars is closer to 230 km, which is the domain used in this study (also shown in Figure 1 of Imhoff et al., 2020).

2.2. Nowcasts

The nowcasts used in this study were the same as the nowcasts constructed for the large sample of events in Imhoff et al. (2020). In this section, we briefly introduce the underlying radar rainfall product (Section 2.2.1), specifics about the available nowcasts (Section 2.2.2), and the set of algorithms used to construct these nowcasts (Section 2.2.3). For more information, we refer to Imhoff et al. (2020).

2.2.1. Bias-Adjusted Radar Rainfall Product

The radar rainfall product in this study is provided by the Royal Netherland Meteorological Institute (KNMI). It is the product KNMI has provided in real time during the study period (2008–2018). The product originates from two C-band weather radars (Figure 1; Beekhuis & Holleman, 2008). Between September 2016 and January 2017, the two single-polarized radars in De Bilt ("DB" in Figure 1) and Den Helder ("DH" in Figure 1) were replaced by two dual-polarized C-band radars in DH and Herwijnen ("H" in Figure 1), see also Beekhuis and Mathijssen (2018). Prior to the rainfall rate estimation, data are processed in a number of steps. Ground clutter is removed from the product with a Doppler-filter and subsequently, horizontal cross-sections are constructed at a nearly constant altitude of 1,500 m, which are called pseudo-constant plan position indicators (pseudo-CAPPIs). Because two radars are used in the composite, range-weighted compositing is used to combine the reflectivities

Table 1
Catchment Characteristics (for Locations, See Figure 1)

Number	Name	Size (km^2)	P	ET_{pot}	Q	f_{XG}	f_{XS}	Lag time (h)	Models used
			($mm\ yr^{-1}$)						
1	Hupsel Brook	6.5	820	587	410			4	WALRUS
2	Gouwepolder	10	888	613	526	−284	325	3	SOBEK RR
3	Grote Waterleiding	40	778	586	228	−73	40	8	WALRUS
4	Luntersebeek	63	871	591	193	−183		7	WALRUS
5	Beemster	71	904	608	707	194	111	8	SOBEK RR-CF
6	Dwarsdiep	83	826	574	536	160	55	11	WALRUS
7	Roggelsebeek	88	715	544	121	−73		9	WALRUS
8	Linde	150	869	591	286		31	8	SOBEK RR
9	Reusel	176	795	546	231		32	9	WALRUS
10	Delfland	379	944	575	430		32	5	SOBEK RR
11	Aa	836	759	542	287			11	WALRUS
12	Regge	957	774	583	277			12	WALRUS

Note. The numbers correspond to the numbers in this table. The water balance terms are determined over the studied period (2008–2018), with P precipitation, ET_{pot} the potential evapotranspiration, Q the observed discharge, f_{XG} the groundwater flux (negative values indicate water leaving the catchment and positive values indicate upward seepage), and f_{XS} the surface water supply. Q is based on the available discharge observations for the year 2015, which was the only year with discharge observations for all catchments. The indicated lag time is the average lag time over the studied period between the center of mass of the rainfall event and the first discharge peak following it, for events with rainfall intensities of 1.0 mm hr^{-1} or more. Individual events were selected with the R-package hydroEvents (Ladson et al., 2013; Wasko & Guo, 2021). The last column states the hydrological models used for these areas.

from both radars (Overeem, Holleman, & Buishand, 2009). Since 2013, non-meteorological echoes have been removed as an additional step with a cloud mask obtained from satellite imagery. Finally, rainfall intensities are estimated with a standard $Z - R$ relationship (Marshall et al., 1955). The radar rainfall product has a 5-min temporal and 1-km^2 spatial resolution.

As this product is not bias-corrected, underestimations of the true rainfall amount of 50% or more can be expected in the Netherlands (Hazenberg et al., 2014; Imhoff et al., 2021). This could lead to missed discharge responses of peaks in the hydrological forecast. Therefore, the radar rainfall product was bias adjusted with the CARROTS (Climatology-based Adjustments for Radar Rainfall in an OperaTional Setting) correction factors (Imhoff et al., 2021). These are fixed bias reduction factors, which vary per grid cell and day of the year, and were derived for the same radar rainfall product as the one used in this study. The factors are based on a 10-year historical data set of the radar rainfall product and a reference rainfall product. As such, the correction factors are available in real time and are independent of gauge availability, which suits this forecasting study. Normally, KNMI applies a mean field bias-adjustment procedure, but this has been shown by Imhoff et al. (2021) to be outperformed by the CARROTS correction for hydrological simulation for most of the 12 catchments in this study. Therefore, the radar rainfall product from the previous paragraph was corrected with the CARROTS factors.

As the computationally expensive nowcasts were already constructed with the unadjusted radar rainfall product in Imhoff et al. (2020), the bias adjustments were applied to the nowcasts as a post-processing step. Ideally, this is done prior to the calculation of the nowcasts, as the spatial correction factors do not advect along with the nowcasted rainfall fields, leading to correction factors that are not entirely representative of the error at that cell anymore once the radar rainfall fields are extrapolated to different grid cells.

2.2.2. Available Nowcasts

The nowcasts from Imhoff et al. (2020) were selected for the period 2008–2018 in a systematic manner (see also Figure 2 in Imhoff et al., 2020). Only the events with the largest rainfall accumulations were selected per catchment, season, and event duration. An "event" was defined as a period with one of the chosen durations, in contrast to the period from start to end of a rainy episode. This means that it does not have to rain continuously during an event defined in this manner, and that the actual rain storm could last longer than the event. Per catchment,

season, and duration, the largest eight rainfall sums were selected and used for nowcasting. Note that for a given catchment and duration, the events cannot have any overlap in time, but a 1-hr event can fall within the time span of a longer (e.g., 24-hr) event. Following this procedure for all combinations of catchments (12), seasons (4), and durations (4), this selection procedure resulted in $12 \times 4 \times 4 \times 8$ (highest rainfall accumulations) = 1,536 events. This procedure guaranteed an even distribution of the precipitation events over all seasons and chosen durations (1, 3, 6, and 24 hr) For more information on the event procedure, statistics, and model runs, see Sections 2.2.2 and 2.4 and Table 1 in Imhoff et al. (2020).

For this large sample of events, nowcasts were produced for each 5-min time step in the event duration and the 6 hr prior to it in order to have a nowcast available for every time step in the event. The nowcasts have a forecast horizon of 6 hr and a temporal resolution of 5 min (information about the event-averaged rainfall intensities and durations can be found in Figure S1 in Supporting Information S1). Summarizing, over 940,000 separate 5-min nowcasts were constructed for each nowcasting algorithm (see Section 2.2.3). Because the events of different durations can have an overlap, that is, an event with a 1-hr duration for catchment x in season y can fall within the time window of an event with a 24-hr window, the shortest durations from the events with an overlap in time were discarded from the analysis. This left 659 individual events for analysis in this study.

2.2.3. Nowcasting Algorithms

The nowcasts for the events described in Section 2.2.2 were constructed with four nowcasting algorithms: two from the Rainymotion library (Ayzel et al., 2019) and two from the Pysteps library (Pulkkinen et al., 2019). Both libraries are open source. The four methods are all field-based nowcasting methods, which use the rainfall intensity of the radar composite (on a Cartesian grid) to determine advection vectors for each tile (other algorithm options are object-oriented, analogue-based, or machine-learning methods, which are not part of this study). The models not only differ in the approach used to derive these advection factors, but also in their (post-)processing steps to construct deterministic or probabilistic nowcasts, and ways to take rainfall field development into account or not (see the next two paragraphs). Hence, this limits this analysis to only a focus on the benefit of field-based nowcasting for hydrological forecasting.

2.2.3.1. Rainymotion

Rainymotion (Ayzel et al., 2019) is a set of four models that make use of commonly used optical flow algorithms to determine the rainfall advection fields. The idea about the nowcasting library is that these four models can be an excellent benchmark for other, more advanced field-based rainfall nowcasting algorithms. Nowcasts were constructed with two of the models (Rainymotion v0.1 was used): *Sparse* (RM-S) and *DenseRotation* (RM-DR).

The RM-S model is the more basic of the two and only tracks the corners of precipitation fields, because these corners have sharp rainfall intensity gradients, which makes them easy to detect. The method identifies these corners with the Shi-Tomasi corner detector (Shi & Tomasi, 1994) from time step $t - 23$ to t. Subsequently, the Lucas-Kanade optical flow algorithm (Lucas et al., 1981) is used to track the identified features and the resulting motion is linearly extrapolated into the future. Finally, an affine transformation matrix is calculated for every lead time, which is used to extrapolate the radar image at time t by warping.

RM-DR uses a global optical flow algorithm (the Dense Inverse Search algorithm by Kroeger et al., 2016), which estimates a velocity vector for each grid cell in the radar composite by using the radar composite from time step $t - 1$ to t. The derived motion field is then extrapolated using a semi-Lagrangian advection scheme (Germann & Zawadzki, 2002). In contrast to more simple field-based methods using global optical flow algorithms, RM-DR allows for rotational movement of the rainfall fields. Finally, the forecast rainfall values are projected on the original grid with an inverse distance weighting interpolation method.

2.2.3.2. Pysteps

Pysteps (Pulkkinen et al., 2019) is a modular framework that allows users to use different parts and concepts of (probabilistic) nowcasting methods, and to develop them further. Originally, Pysteps was based on S-PROG (Seed, 2003) and STEPS (Bowler et al., 2006; Seed et al., 2013), but nowadays it also supports more recent approaches by for example, Nerini et al. (2017) and Pulkkinen et al. (2020). The main steps to construct a nowcast using Pysteps are described and visualized in Section 3.5 and Figure 4 in Pulkkinen et al. (2019).

In Imhoff et al. (2020), two setups of Pysteps (v0.2) were used to construct the nowcasts: one for deterministic nowcasts and one for probabilistic nowcasts. The deterministic setup, referred to as *Pysteps deterministic* (PS-D), is based on S-PROG (Seed, 2003), with the only difference being that eight instead of six cascade levels were used. The configuration was as follows: a Lucas-Kanade (Lucas & Kanade, 1981) optical flow algorithm using the radar composite from time $t - 3$ to t, a second-order auto-regressive model, the S-PROG masking method (0.1 mm hr^{-1} threshold), a backward semi-Lagrangian advection scheme (Germann & Zawadzki, 2002), and a method to match the forecast statistics with the latest observation (probability matching) that is based on the mean observed rainfall fields.

The probabilistic setup, referred to as *Pysteps probabilistic* (PS-P), is quite similar to PS-D, but follows all steps in the Pysteps workflow to obtain an ensemble nowcast. Most of the configuration was similar to PS-D, but the following modules were different from or absent in PS-D: use of the STEPS nowcasting method (Bowler et al., 2006) with a nonparametric noise method (Seed et al., 2013) to obtain 20 ensemble members, a lead time-dependent masking method, and a probability matching method that uses the cumulative distribution function.

2.3. Experimental and Forecast Verification Setup

2.3.1. Hydrological Model Setup

The employed hydrological models in this study (Table 1) are the models used operationally or for research purposes by the involved water authorities. SOBEK RR(-CF) (Prinsen et al., 2010; Stelling & Duinmeijer, 2003; Stelling & Verwey, 2006) is a semi-distributed model that can couple bucket-style rainfall-runoff (RR) modules for paved, unpaved, greenhouse, and waste water treatment plant areas to a hydraulic routing module (cf., for channel flow). SOBEK is often used in complex polder systems where many or all of these situations occur. The four catchments in this study that have a SOBEK model have the following number of sub-catchments: 1 for Beemster, 7 for Gouwepolder (areas ranging from 4.5 to 668 ha), 23 for Linde (areas ranging from 23 to 7,230 ha), and 25 for Delfland (areas ranging from 1.2 to 2,112 ha). Radar rainfall QPE and QPF were averaged over the sub-catchments as a preprocessing step prior to model simulation.

WALRUS (Brauer, Teuling, et al., 2014) is a lumped bucket-style RR model that is tailored to lowland catchments and accounts for typical lowland processes such as the coupling of groundwater and unsaturated zone, seepage and surface water supply, wetness-dependent preferential flow paths and groundwater-surface water feedbacks. Because the model is lumped, catchment-averaged radar QPE and QPF were used as forcing input.

As these models are used for operational or research purposes, most of the models were already calibrated (Brauer, Torfs, et al., 2014; Gerritsen, 2019; ; Heuvelink et al., 2020; Loos, 2015a, 2015b; Sun et al., 2020). The same setups were used in this study (Table S1 in the supplement). The catchments Roggelsebeek and Dwarsdiep were, however, not calibrated prior to this study. Therefore, WALRUS was calibrated for both catchments using a calibration method constisting of a Latin-Hypercube sampling method (LHS, used sample size was 2,500, McKay et al., 1979), followed by a Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) optimization with as starting point the 10 best parameter sets that followed from the LHS method. In this calibration approach, six WALRUS paramaters (c_W, c_G, c_Q, c_V, c_S and c_D, see the supplement for further information) were calibrated. The gauge-adjusted rainfall product provided by KNMI (referred to as R_A) was used as rainfall input for calibration. The R_A product is not available in real time (and was therefore not used for the forecasts) and can be considered the most accurate reference rainfall product in the Netherlands (Overeem, Buishand, & Holleman, 2009; Overeem et al., 2011; Overeem, Holleman, & Buishand, 2009). Calibration period was 2013–2014 for the Roggelsebeek and 2016–2017 for the Dwarsdiep catchment. The choice for this period was based on the discharge observation availability and quality for both catchments. According to water authority Limburg, the discharge measurements in the Roggelsebeek are not always reliable during summer in the period prior to 2019, due to plant growth in the section of the brook where the measurements are taken. So, this may have influenced the model calibration. The results of the calibration can be found in the supplementary material (Figures S2 and S3), while the overall model validation for all 12 catchments is discussed in Section 3.1. In addition, the WALRUS model for the Luntersebeek catchment was recalibrated for water board Vallei and Veluwe at the start of this study. Calibration procedure was similar to the procedure for the Roggelsebeek and Dwarsdiep and the year 2019 was used as calibration period. Results of the calibration can be found in Figure S4 in Supporting Information S1. Lastly, the calibration

procedure for the aforementioned three catchments resulted by no means in the optimal parameter set, due to for example, the equifinality issue (Beven, 1993). The same holds for the already calibrated parameters of the operational hydrological models. However, the effects of this are excluded from this study by comparing discharge forecasts with the hydrological model simulations using the "observed" CARROTS-corrected radar QPE instead of discharge observations, leaving out any model related errors (see Section 2.3.2).

2.3.2. Hydrological Forecasts

The event selection procedure resulted in events that were selected per catchment. Therefore, hydrological simulations for the events selected for a given catchment were only run for that catchment. For each 5-min issue time in the nowcasts for the x selected events, a hydrological forecast was made with a forecast horizon of 12 hr (the 6 hr forecast horizon of nowcast rainfall as input, followed by zero precipitation [ZP] for 6–12 hr in advance), which is more than the average response time of the catchments (Table 1, which indicates the average lag time over the studied period between the center of mass of the rainfall event and the first discharge peak following it, for events with rainfall intensities of 1.0 mm hr^{-1} or more). The model simulations were run with the nowcast rainfall inputs from the four algorithms (RM-S, RM-DR, PS-D, and PS-P) and the initial conditions were based on a continuous model simulation with the CARROTS-corrected radar QPE. In addition, two benchmark forecasting setups were considered: a hydrological simulation using *Eulerian Persistence* (referred to as EP) and a forecast without any precipitation input (referred to as ZP: *Zero Precipitation*). In the case of EP, which was also present in Imhoff et al. (2020), the rainfall intensity in the latest radar QPE ($t = 0$) was used as the forecast for the coming 12 hr. Hence, if it rained with 1.0 mm hr^{-1} at $t = 0$, it was assumed to keep on raining with 1.0 mm hr^{-1} for every time step in the subsequent 12 hr.

The hydrological model simulation quality is not only affected by the rainfall input, but also by the sources of error mentioned in the introduction: initial conditions, model setup, and calibration procedures (Beven, 1993; Clark et al., 2017; Melsen et al., 2016). This makes it impossible to differentiate between the effect of the rainfall forecast skill and the other sources of error, when we try to quantify its effects on the simulated discharge skill. In order to isolate the effect of the rainfall forecast skill on the simulated discharge, we chose not to use the observed discharge as reference. Instead, the reference for model simulation verification employed in this study was the hydrological model simulation with the "observed" radar rainfall after bias correction with the CARROTS factor (see Section 2.2.1). This choice discards any model and radar QPE errors from the subsequent analyses. Thus, the CARROTS-corrected QPE product is used as reference in evaluating the results. We come back to the effects of this decision in the Discussion (Section 4.4). On one occasion, in Figure 3 of the results, the hydrological model and the QPE-product are compared with the observed discharge and a model run with the R_A product provided by KNMI (Section 2.3.1).

Potential evapotranspiration (ET_{pot}) is, next to rainfall, also a required forcing variable for the used hydrological models. For both forecast and reference model runs, the gridded Makkink ET_{pot} product from KNMI (Hiemstra & Sluiter, 2011) was used as forcing. This means that the near real-time ET_{pot} product was also used for the forecasts instead of an ET_{pot} forecast, as this made it possible to analyze the effect of different rainfall inputs on the model runs in isolation. Besides, van Osnabrugge et al. (2019) showed that including ETpot forecast has little impact on the discharge forecasts, so we expect minor impacts from this decision.

2.3.3. Verification Metrics

In this section, the verification metrics that are used in Sections 2.3.4 and 2.3.5, are described. The implementation of these metrics per event and catchment is also described in these sections.

2.3.3.1. Mean Absolute Error

The mean absolute error (MAE) describes the average absolute error between forecast (F) and observation (O) at (lead) time t, averaged over the number of forecasts (N_f) with lead time t in the event. It is calculated as:

$$MAE = \frac{\sum_{i=1}^{N_f} |F_t - O_t|}{N_f} \quad (1)$$

2.3.3.2. Nash-Sutcliffe Efficiency

The Nash-Sutcliffe Efficiency (NSE) metric (Nash & Sutcliffe, 1970) is an often used metric in hydrology to match hydrographs with observed discharge time series. The NSE metric ranges from $-\infty$ to 1.0, with 1.0 representing a perfect agreement between observations and simulations and a value of 0.0 indicating that the simulation is not more skillful than the mean of the observations. The metric is formulated as:

$$NSE = 1 - \frac{\sum_{t=1}^{N_t} (F_t - O_t)^2}{\sum_{t=1}^{N_t} (O_t - \mu_o)^2}, \quad (2)$$

with N_t the number of time steps in the time series, or the number of forecasts with lead time t in the event, while μ_o is the mean observed discharge.

2.3.3.3. Kling-Gupta Efficiency

The Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009) builds upon the NSE metric and is formulated as:

$$KGE = 1 - \sqrt{(\rho - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad (3)$$

with

$$\alpha = \frac{\sigma_f}{\sigma_o}, \quad (4)$$

$$\beta = \frac{\mu_f}{\mu_o}. \quad (5)$$

Here, ρ is the Pearson correlation between observed and simulated discharge, α the flow variability error between observed and forecast discharge, and β the ratio of μ_f (mean simulated discharge) and μ_o. σ_f and σ_o are the standard deviations of, respectively, the forecast and observed discharge. Similar to the NSE metric, the KGE metric ranges from $-\infty$ to 1.0, with 1.0 representing a perfect agreement between observations and simulations.

2.3.3.4. False Alarm Ratio

The false alarm ratio (FAR) determines for a given threshold the ratio of false alarms over all forecasts (here, within an event) that predict an exceedance of the threshold (value between 0.0 and 1.0). It is formulated as:

$$FAR = \frac{FP}{TP + FP}, \quad (6)$$

with FP the number of false positives, that is, false alarms, in the forecasts and TP the number of true positives, that is, the hits. Hence, the FAR gives the forecaster an indication of the forecast reliability when the forecast predicts a threshold exceedance.

2.3.4. Dependency on the Rainfall Characteristics

The dependency of the forecast quality on the rainfall characteristics was tested with two methods. First, for three rainfall intensity classes (less than 2.0, between 2.0 and 5.0, and greater than or equal to 5.0 mm hr^{-1}), the NSE score as a function of forecast lead time was calculated for all six methods (intensities were based on the spread of rainfall intensities in the sample of events in order to have a representative amount of events left in all three classes). Input for the NSE calculation were the 1-hr discharge accumulations (rolling sum) of forecast and reference run. This was chosen, instead of the instantaneous discharges, to smooth the "pump on - pump off" behavior of the polders. Besides, a 1-hr accumulation is still less than the response time of all catchments (Table 1). The NSE was calculated following the method of Berenguer et al. (2005), who made a discharge time series per lead time, for example, all 1-hr lead time forecasts of all forecasts in the event are combined in one time series, and compared this to the reference discharge. The first lead time for which the NSE drops below a threshold of

0.9 is seen as the skillful lead time of the forecast. This 0.9 point is somewhat arbitrarily chosen, but it makes a comparison with Berenguer et al. (2005) and Heuvelink et al. (2020) possible, who followed the same approach.

Second, per catchment, the relationship between the mean event rainfall intensity and MAE of the forecast was evaluated per event. This was done for both the rainfall forecast (the nowcast) and the hydrological forecast, to be able to compare the error in the rainfall forecast with the error in the hydrological forecast. The mean event rainfall intensity was based on all rainy 5-min instances in the event. To reduce the dimensionality and in that way to be able to summarize the evaluation in one scatter plot, this was only calculated for RM-DR. To focus on the rainfall volume of the forecast, the 3-hr QPF sum (hence, the first 3 hours of the nowcast) was compared to the 3-hr sum of the reference rainfall. The first 3 hr instead of 6 hr of the rainfall nowcast were used, because the last 3 hr of the nowcast is seldom skillful (Imhoff et al., 2020). A similar approach was taken for the discharge forecast, though here the 12-hr forecast (the full forecast horizon) sum was used.

2.3.4.1. Spatial Variability of Rainfall

In addition to the mean event rainfall intensity, the MAEs of the forecasts were compared to the mean event spatial rainfall variability. The rainfall variability (I_σ) was calculated following the method of Lobligeois et al. (2014):

$$I_\sigma = \frac{\sum_{t=1}^{N_t} \sigma_t \cdot P_t}{\sum_{t=1}^{N_t} P_t}, \tag{7}$$

with

$$\sigma_t = \sqrt{\frac{\sum_{i=1}^{N_i} [P_i(t)]^2}{N} - \frac{\left[\sum_{i=1}^{N_i} P_i(t)\right]^2}{N^2}}. \tag{8}$$

Here, P_t is the catchment-averaged 5-min rainfall sum for time t, σ_t the standard deviation of the 5-min rainfall sum over the catchment area, $P_i(t)$ the 5-min rainfall sum at grid cell i and time t, and N_i the number of grid cells in the catchment area. In the calculation of I_σ, only the time steps where it rained on at least one grid cell were taken into account.

In the context of rainfall forecasting, we expect that spatial rainfall variability is more important for small catchments than for larger catchments. This is because having the exact location of the rainfall forecast wrong, which happens more easily when the variability of the rainfall event is higher (more small convective cells), has a larger impact on a small catchment, as it can lead to rainfall either hitting or missing the catchment in the forecast, whereas for larger catchments it merely determines where the rain falls in the catchment. The latter can have an effect on the simulated discharge at the outlet too, but can only be properly modeled with a distributed hydrological model, which is outside the scope of this paper.

2.3.5. Discharge Peak Forecast Verification

An important reason for implementing rainfall nowcasting in early warning systems is the (potential) ability to timely forecast peak discharges and threshold exceedances. In this section, we describe the verification procedure employed to assess the hydrological peak discharge forecast quality when nowcasting is used as rainfall forecast. Although a large sample analysis has as an advantage that it can provide robust statistics about the model simulation quality, it has as a disadvantage that it becomes nearly impossible to verify peak discharge or threshold exceedance forecast quality in an automated way. The reason for that is twofold: (a) for the larger catchments, the peak discharge may arrive later than the 12 hr forecast horizon, especially when there are more showers taking place after the event duration (which was merely a time window in which the maximum amount of rainfall fell, rather than the full period from start to end of a rainy episode), and (b) the use of a standard threshold as set by, for example, the water authority becomes difficult, as many events do not reach that threshold or already start above the threshold, leaving only a few forecasts for analysis.

For this reason, we have chosen to use a more pragmatic approach. For every event, the maximum discharge (also here based on the 1-hr accumulations, rolling sum, of the reference run) that occurred within the formulated event was regarded as the threshold for that event. In that way, a threshold was reached in every event. That, however, still leaves that forecasting such a peak discharge exactly right in magnitude and timing may be too

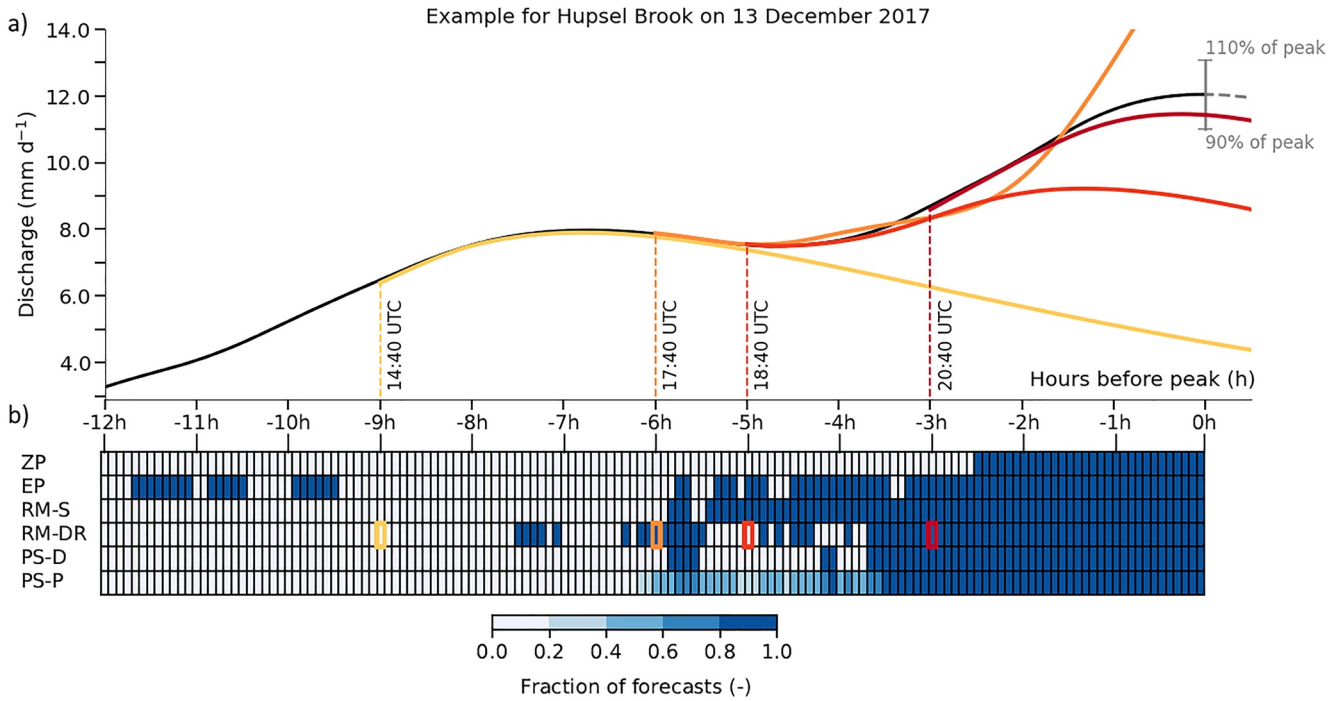

Figure 2. Example of discharge peak verification procedure using the discharge peak that occurred on 13 December 2017 in the Hupsel Brook catchment. (a) The forecast (using Rainymotion DenseRotation [RM-DR]; colors) and reference (black) discharge. Shown are four issue times of the forecasts: 14:40 UTC (yellow; 9 hr before peak), 17:40 UTC (orange; 6 hr before peak), 18:40 UTC (red; 5 hr before peak), and 20:40 UTC (burgundy; 3 hr before peak). The margin (between 90% and 110%) for a "correct" peak forecast, which is used for the peak forecast verification at the time of the peak, is indicated in the top right of the figure. Note that in contrast to the peak anticipation time and FAR, the forecast consistency only uses the 90% threshold exceedance at the time of the peak (see the forecast consistency paragraph in Section 2.3.5). (b) Visualization of the forecast consistency for the forecasts issued from 12 hr to 5 min prior to the peak. The table cell is colored dark blue when the forecast issued at that time exceeds the 90% peak value at the time of the observed discharge peak. In all other cases, the cell is colored white. PS-P gives a probability of exceeding this value. Therefore, the cells of the PS-P forecast are colored following the fraction of ensemble members that exceed this value (the darker blue, the more members). The four colored cell edges in the RM-DR row, indicate the four forecasts shown in (a).

strict a constraint for testing the forecast quality. Therefore, in this analysis, a peak magnitude error of 10% of the difference between the initial discharge at the start of the event and the highest discharge was allowed. The timing should, however, be right. Thus, if a maximum discharge occurred at 12:00 UTC, then this approach looks at all forecasts for this time and checks whether the forecast is within ±10% of the maximum discharge magnitude (see Figure 2a for an example of this magnitude range). As the forecast horizon was 12 hr in this study, it was possible to verify this from t_{-12} until t_0 hr, when the maximum discharge occurs. In the following three paragraphs, we will introduce the three aspects that were tested with this set rule for a "correct" forecast in mind. In the discussion, we come back to the choice of this magnitude error.

2.3.5.1. Peak Anticipation Time

An important water management question is how much time before a peak or threshold exceedance the forecast is able to capture the peak or threshold exceedance. A long time between forecast and occurrence allows water managers to take action (issue warnings or real-time control). Thus, what is the anticipation time (before a high discharge) a forecaster can expect given a nowcasting method used for hydrological forecasts? This was tested here: for every event, the first issue time (within all forecasts from t_{-12} until t_0 hr) for which the maximum discharge was forecast within the given magnitude range, given the constraint of a "correct" forecast in the aforementioned paragraph, was recorded. This was done for all methods and for all catchments. This method gives an overview of all the first issue times where the maximum discharge was forecast correctly. Figure 2 provides an example of this method for a peak discharge in the Hupsel Brook catchment on 13 December 2017. The first correct maximum discharge forecast for, for example, PS-D occurs around 6 hr prior to the maximum observed discharge in the event (indicated with the first dark blue cell in Figure 2b) and that time was recorded with this method. In the case of PS-P, every ensemble member was taken into account individually, giving for this event a range of anticipation times ranging from 3.3 to 6.2 hr prior to the maximum observed discharge. However, this

method does not yet take into account any overestimations or inconsistencies in the (subsequent) forecasts. This will be considered next.

2.3.5.2. False Alarm Ratio

Besides the timeliness of the forecast, we want to know how reliable the forecast is whenever it predicts a maximum discharge, since water managers lose their credibility when they issue too many warnings that turn out to be false alarms. Therefore, the FAR was calculated. Here, a "hit" was reached when the forecast discharge at the time of the observed maximum discharge fell within the predefined magnitude range. A "false alarm" was reached when the forecast discharge was more than 10% higher than the observed maximum discharge. In all other cases, the forecast was considered to have missed the maximum discharge. An example of a false alarm is present in Figure 2 for RM-DR. At 6 hr prior to the observed maximum discharge, the forecast (orange line in Figure 2a) strongly overestimates the observed maximum discharge. For RM-DR in this event, the FAR was found to be 0.15. It was even higher for EP (0.44). EP had the longest anticipation time for this event (Figure 2b), but this was caused by rainfall that fell early in the event and persisted in the EP forecast throughout the entire forecast, leading to overestimated discharges. However, a consistently "correct" peak forecast, was issued by EP only at 3:20 hr from the peak, something that will be considered in the subsequent paragraph. Finally, also here, all ensemble members of PS-P were taken into account individually.

2.3.5.3. Forecast Consistency

Finally, a forecast can also be inconsistent. It is damaging for water managers to issue, recall, and reissue warnings or flood prevention measures based on a previous erroneous forecast. Figure 2 gives an example of this for the RM-DR forecast. In this forecast, it happened frequently that a well forecast maximum discharge was succeeded by a forecast that underestimated the maximum discharge (more than 10% below the predefined magnitude range). For four issue times in Figure 2b for RM-DR, the forecast hydrographs are visualized in Figure 2a, illustrating how a "correct" forecast can be succeeded by a miss an hour later. Such inconsistencies lower the trust in the forecast. In this study, a consistency index was assigned per event based on the number of times a successful forecast was succeeded by a forecast that underestimated the maximum discharge by more than the predefined magnitude range. For this analysis, only the lower limit of the magnitude range was taken into account, so a false alarm succeeding a "correct" forecast was not taken into account. For PS-P, which gives a probability of exceeding the predefined threshold, as also visualized in Figure 2b, the inconsistency index was calculated as the fraction of previous forecasts minus the current fraction of forecasts that correctly predict a threshold exceedance, in case the current fraction is lower than the previous (an inconsistency in the forecast). In the example in Figure 2, the inconsistency index of RM-DR was 8. It was 0 for ZP, 7 for EP, 1 for RM-S, 2 for PS-D, and for 0.95 PS-P; with lower values for the inconsistency index indicating that the forecast is more consistent.

3. Results

In the following results, the quality of the hydrological models is assessed in Section 3.1, followed by an example forecast, focusing on one event, in Section 3.2. In the remaining part of the results (Sections 3.3 and 3.4), all events are jointly taken into account.

3.1. Hydrological Model Validation

The difference between the response of the two polders Gouwepolder and Beemster, and the freely draining catchments is directly visible in Figures 3b and 3e. Both polders react quickly to rainfall events with higher (specific) discharge peaks than in most other catchments, resulting from the erratic and rapidly responding pumping regime. In the Beemster, the effect of the surface water supply to flush salt water intrusion originating from seepage, is particularly visible during summer, when a relatively high baseflow persists. To a lesser extent, these characteristics of quick responses with a high specific discharge are also present for the partly freely draining catchments Dwarsdiep, Linde, and Delfland (Figures 3f, 3h and 3j).

The hydrological models for most catchments perform well (Figure 3). This specifically holds for the catchments Dwarsdiep, Reusel, Delfland, Aa, and Regge where the KGE is 0.7 or higher for both the model simulations with R_A (the reference radar rainfall product) and the simulations with R_C (the bias-corrected real-time available radar rainfall product). Typical for the Luntersebeek is that the simulation often misses the first discharge peak in fall

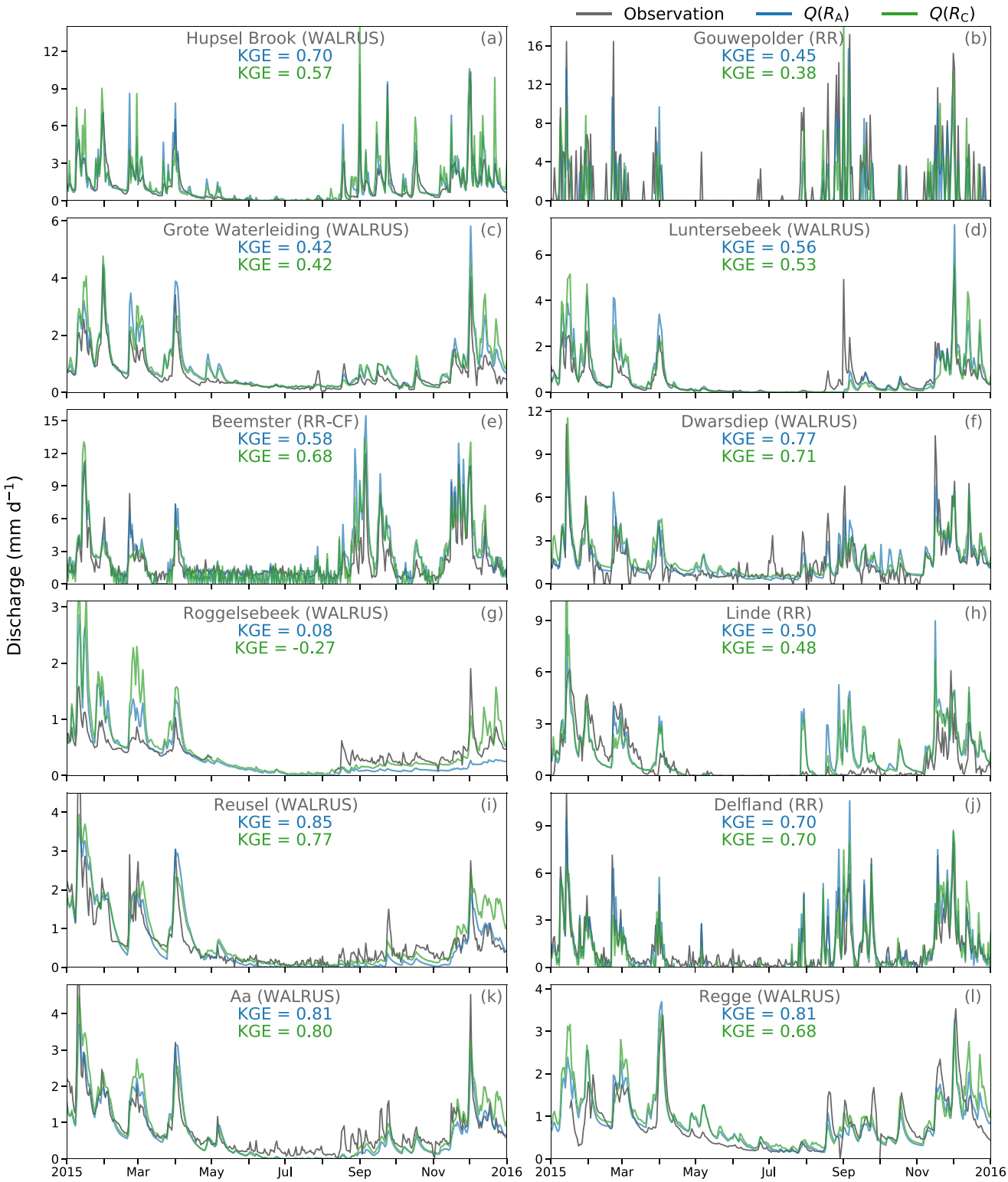

Figure 3. Simulated discharges for the 12 catchments as part of the hydrological model validation. The observed discharge is shown in gray, the model simulations with the gauge-adjusted radar rainfall product ($Q(R_A)$; the reference) in blue and the model simulations with the bias-corrected real-time available product ($Q(R_C)$) in green. Only the results for 2015 are shown, but the Kling-Gupta Efficiency (KGE)-values are based on the full period (2008–2018).

(Figure 3h). The opposite happens in the simulations for the Linde, where the discharge is overestimated during fall (Figure 3d). Although this overestimation is present every year, it is highest in 2015.

The simulations for the Roggelsebeek catchment lead to lower KGE values than for the other catchments, with a KGE of 0.08 for the model run with R_A as input and −0.27 for the run with R_C as input. This can be partly explained by the sometimes unreliable discharge measurements for this catchment, which have influenced the calibration procedure. Since the model run with R_C is taken as reference, the effect of the poor model results is not expected to majorly impact the results in the next sections.

The difference in the results between the simulations with R_A and R_C is small and often even absent. This indicates that the CARROTS-corrected radar rainfall product leads to sufficiently accurate simulation results that are similar to the results with the reference rainfall product (for the specifics of both products, see Sections 2.2.1 and 2.3.1). Hence, the majority of the discrepancy between model simulations and observations originates from the model setup (e.g., structure and parameterization) rather than the rainfall product used. In the following, the error resulting from the model setup, but also the remaining error in the rainfall product, is discarded, as the reference is the model run with R_C.

3.2. Example Forecast

Prior to analyzing all events, we zoom in on one event to highlight some typical differences between the tested models. Figure 4 shows both the rainfall (left column) and discharge forecasts (right column) for four issue times during an event that took place in the Hupsel Brook catchment on 22 October 2013. During this day, a frontal zone with convective activity passed in northeasterly direction over the country and hit the Hupsel Brook catchment at the end of the evening, resulting in approximately 11 mm of rainfall in just under an hour. Although the frontal-convective rainfall event itself is quite normal in the Netherlands during part of the year, the local occurrence of this short-duration high-intensity rainfall event, in combination with the quick discharge response, makes it a typical nowcasting challenge.

At the first issue time, rainfall occurs a little over 3 hours later (and the discharge peak approximately 8 hr later), which is forecast by none of the nowcasting models, except for several ensemble members of PS-P, which already indicate the possibility of a discharge peak a few hours later (Figures 4a and 4b). The nowcasts issued an hour later do forecast rainfall, but substantially less than the approximately 11 mm that would eventually fall between two and 3 hours later (Figures 4c and 4d). Besides, the timing of the rainfall in the RM-DR forecast is more than 3 hr off. For PS-P, more ensemble members forecast a substantial amount of rainfall and one member even overestimates the observed rainfall, which is visible in the discharge forecast as well. However, the ensemble median still highly underestimates the discharge peak, as do the other methods.

Most forecasts issued at 20:00 UTC (Figures 4e and 4f) capture the presence of rainfall starting an hour later. RM-DR overestimates this amount, whereas PS-D and the median of PS-P underestimate the amount. RM-S forecasts no rainfall at all at this point. This is also the case for EP and ZP, but that is expected due the absence of rainfall at the issue time of the forecasts. The last forecasts, issued at the start of the rainfall event (21:00 UTC, Figures 4g and 4h), capture the discharge peak quite well. RM-S, however, strongly overestimates the rainfall and discharge amounts, and has a timing error for the rainfall of approximately 2 hr. Both EP and ZP still show hardly any response (both rainfall and discharge forecast) at the start of the event, but do so for issue times just after this one, for example, half an hour later when it already rains (not shown here), sometimes also resulting in overestimations of the observed rainfall and reference discharge.

Typical for PS-D and to a lesser extent PS-P, is that they tend to underestimate the rainfall amount in their forecast, see for example, also Imhoff et al. (2020). Hence, although the forecast timing and spatial location of the rainfall fields are generally good, the rainfall amount is somewhat underestimated, leading to larger underestimations in discharge (Figures 4e–4h). The opposite holds for RM-S and RM-DR. Both models preserve the rainfall amount present in the latest observations, which reduces the underestimations, but also regularly result in overestimations. The effect of these model characteristics and their influence on the peak discharge forecasts, will be discussed in Section 3.4.

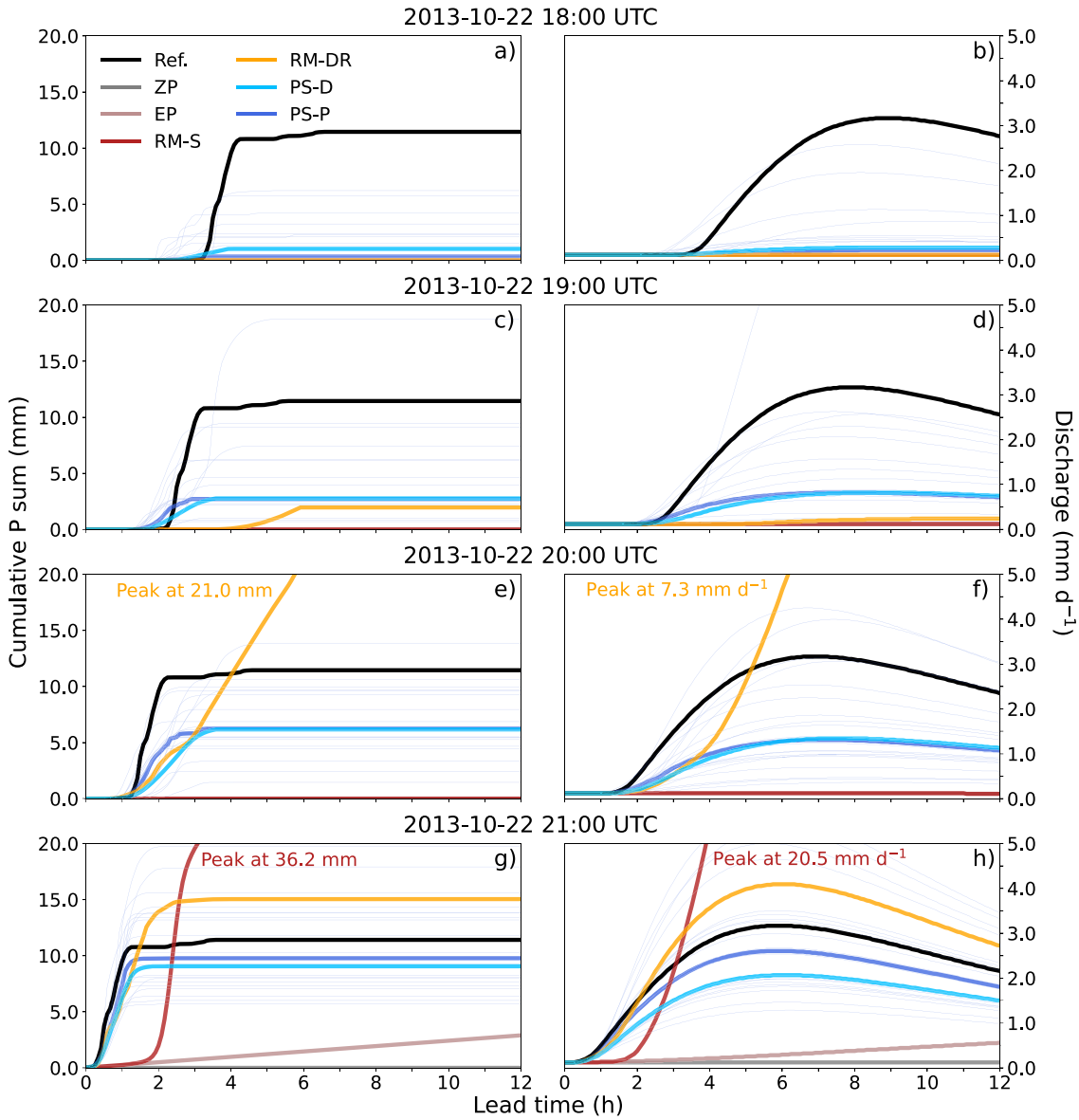

Figure 4. Rainfall and discharge forecasts for four issue times during the discharge peak that occurred in the Hupsel Brook catchment on 22 October 2013. Shown are the forecasts using rainfall inputs from all nowcasting algorithms and methods (colors), compared to the reference (black). The thin dark blue lines indicate the separate ensemble members of PS-P, whereas the thick dark blue line indicates the ensemble median.

3.3. Dependency on the Rainfall Characteristics

3.3.1. Rainfall Intensity

The skillfulness of the discharge forecast decreases with increasing rainfall intensity (Figures 5a–5c), which is similar to the effect on the rainfall forecast (e.g., Ayzel et al., 2019; Foresti et al., 2016; Pulkkinen et al., 2019). The higher rainfall intensities generally correspond to (more) convective systems, which have a lower predictability than stratiform systems with lower average rainfall intensities. With increasing rainfall intensity, the skillful lead time of the discharge forecasts (the first lead time for which the event-average NSE drops below a threshold of 0.9) decreases from on average 153 min for all methods together and a rainfall intensity of less than 2.0 mm hr^{-1}, to 110 min (between 2.0 and 5.0 mm hr^{-1}) and to 75 min for rainfall intensities of 5.0 mm hr^{-1} or more.

RM-DR, PS-D, and PS-P outperform, on average, the other methods, although especially the difference between RM-S and these three methods is not as substantial as in Figure 4 of Imhoff et al. (2020). The relative gain, the

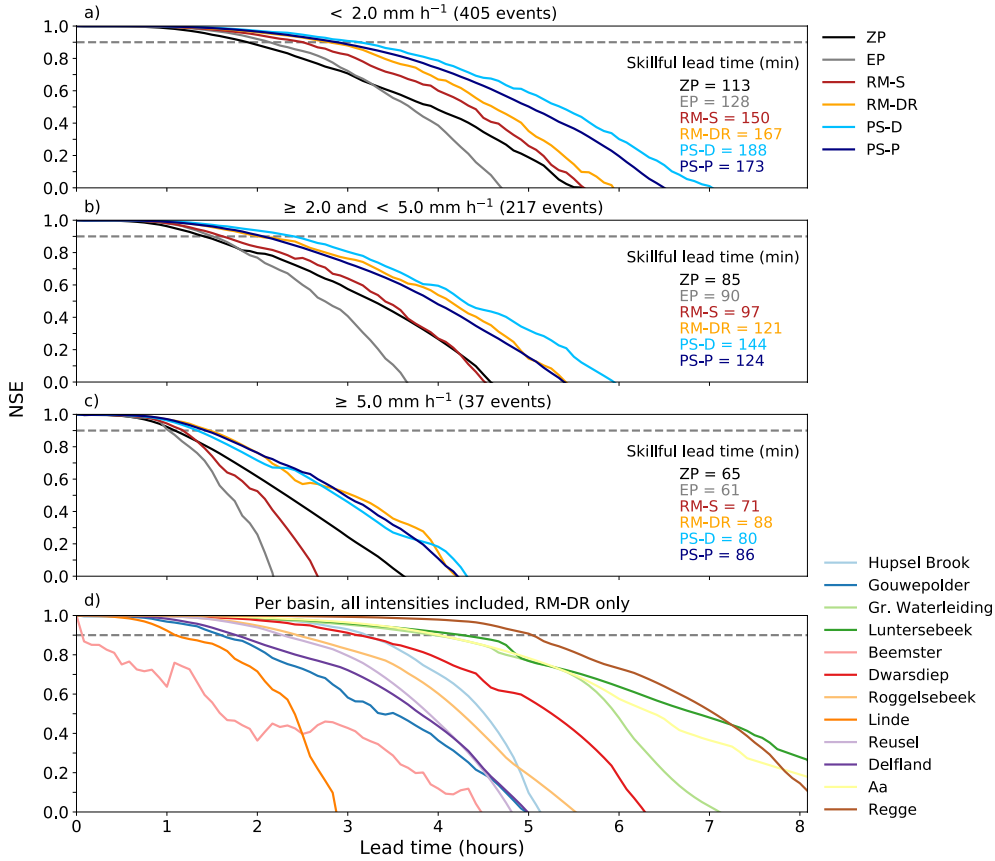

Figure 5. Event-average Nash-Sutcliffe Efficiency (NSE) for all studied methods as function of forecast lead time based on the 1-hr accumulated forecast discharge (rolling sum). (a–c) All catchments together for different rainfall thresholds: (a) <2.0 mm hr^{-1}, (b) ≥2.0 and <5.0 mm $hr^{-1,}$ and (c) ≥5.0 mm hr^{-1}. The indicated skillful lead time is based on the crossing point between an NSE of 0.9 and the event-average NSE for the given nowcasting method. For PS-P, the NSE was calculated for every ensemble member separately. (d) Similar to a–c, but per catchment (from small to large), for all events together and only Rainymotion DenseRotation (RM-DR) is shown.

difference between the skillful lead time of ZP and that of one of the other methods, gives an indication of the expected gain of a nowcasting method over having no rainfall forecast at all. For the lowest rainfall intensities (less than 2.0 mm hr^{-1}), the relative gain is 15 min for EP, 37 min for RM-S, 54 min for RM-DR, 75 min for PS-D, and 60 min for PS-P. This indicates that with PS-D, the discharge forecast is on average skillful 75 min further ahead than when no rainfall forecast (ZP) is used. However, for higher intensities, the relative gain is less. For intensities of 5.0 mm hr^{-1} or more, the relative gain reduces to no relative gain for EP, 6 min for RM-S, 23 min for RM-DR, 15 min for PS-D, and 21 min for PS-P.

In addition, Figure 5d shows per catchment the NSE per lead time for the discharge forecast, as based on the RM-DR forecasts and all events together. Skillful lead times can be substantially different between catchments, with for instance approximately an hour for the Linde, but more than 5 hours for the Regge catchment. Overall, the polders and some partly freely draining catchments (Gouwepolder, Beemster, Linde, and Delfland) show shorter skillful lead times than the other catchments. Besides that, the performance per catchment depends on multiple factors, such as location in the country (e.g., Figure 7 in Imhoff et al., 2020), the size of the catchment (e.g., Figure 6 in Imhoff et al., 2020), and catchment response time.

The decreasing skill with increasing rainfall intensity is also present in Figure 6 (and corresponding Table 2), which shows for RM-DR and per catchment the relationship between the mean event rainfall intensity and either the MAE of the rainfall forecast for the first 3 hours of the forecast (Figure 6a) or the MAE of the discharge forecast for the entire forecast horizon (12-hr sum, Figure 6b). For the rainfall nowcasts (Figure 6a), the relationship between the increase in rainfall intensity and MAE in the rainfall forecast is linear, with Pearson correlation

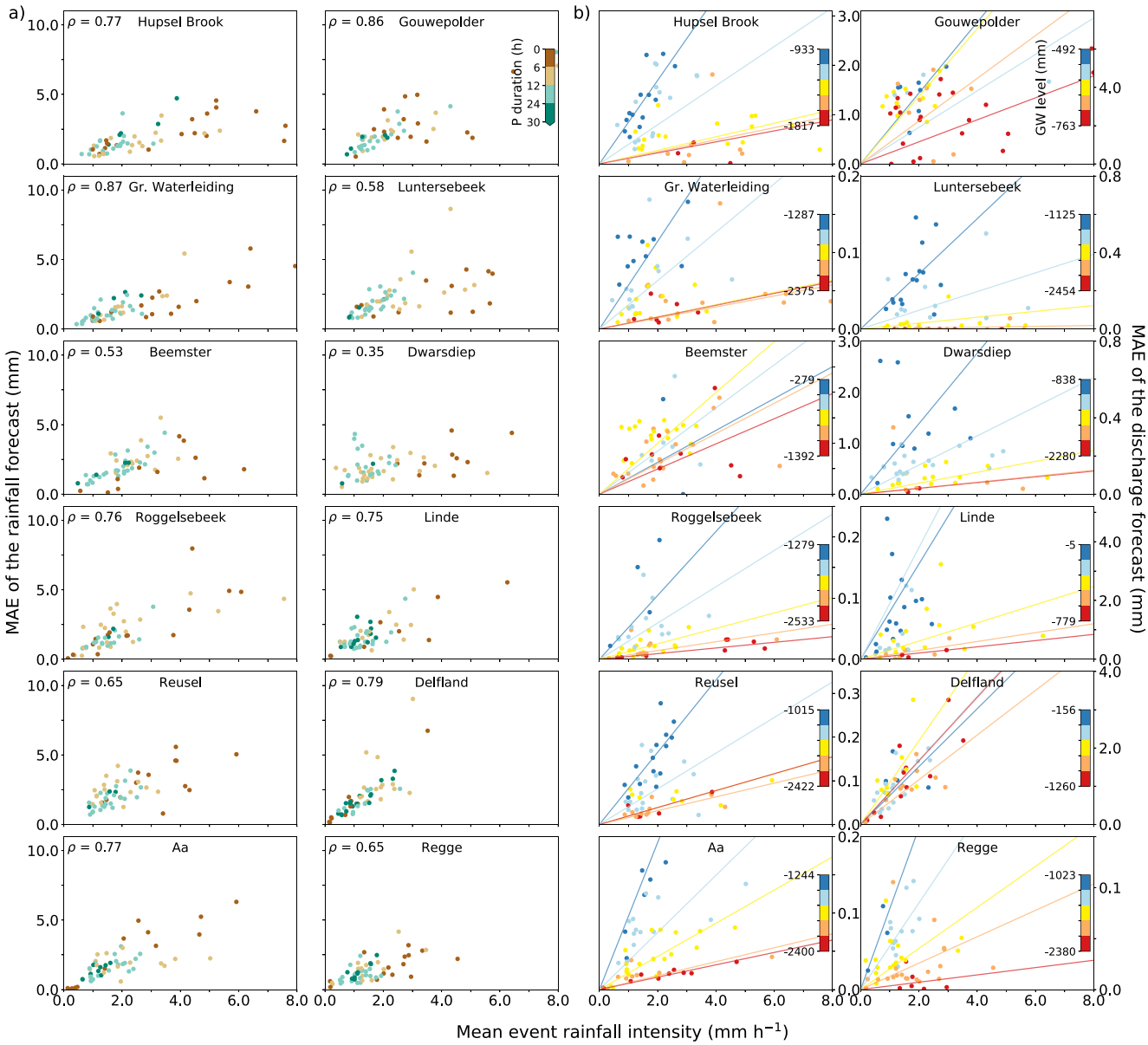

Figure 6. Relationship between the mean event rainfall intensity (based on all rainy 5-min instances) and the mean absolute error (MAE) of the: (a) rainfall forecast and (b) the discharge forecast per catchment. The rainfall MAE was calculated for the 3-hr sum (the first 3 hours of the nowcast) of the forecast and reference rainfall. The 12-hr sum (the entire forecast horizon) was used for the discharge MAE. The colors of the dots in (a) indicate the rainfall duration during the event and in (b) of the simulated groundwater table depth (in mm) at the start of the event. The groundwater depths are subdivided in classes of equal size (representing 20% of the values between minimum and maximum simulated groundwater level), but only the shallowest and deepest groundwater depths are indicated. Per groundwater class, a linear regression line with the same color is adjusted to the corresponding points. The statistics of these lines can be found in Table 2.

coefficients ranging from 0.35 (Dwarsdiep) to 0.87 (Grote Waterleiding). Generally, the higher rainfall intensities and the corresponding higher MAE of the rainfall forecasts occur for events with the shorter (less than 6-hr) durations. An exception to this is visible for the Hupsel Brook catchment, where the highest MAE occurred during a longer lasting event on 26 August 2010. This was an extreme event that led to 160 mm of rainfall in 24 hr (Brauer et al., 2011).

Although the error in the discharge forecast increases with increasing rainfall intensity as well (see also Figures 5a–5c), the relationship between the MAE of the discharge forecast and the mean event rainfall intensity is at first sight not linear and not as clear as for the rainfall forecast (Figure 6b). However, one should realize that the hydrological response and predictability depend also on the initial conditions. In Figure 6b, the data points

Table 2
Statistics of the Linear Regression Lines (Forced Through Zero) That Were Adjusted to the Data Points of Figure 6b, Which Corresponds to the Mean Event Rainfall Intensity Analysis

	Class 1		Class 2		Class 3		Class 4		Class 5	
Catchment	Slope	ρ	Slope	ρ	Slope	ρ	Slope	ρ	Slope	ρ
Hupsel Brook	0.85	0.79	0.40	0.77	0.14	0.32	0.12	0.29	0.11	0.93
Gouwepolder	1.85	0.41	0.95	0.45	1.77	0.26	1.12	−0.20	0.58	0.39
Grote Waterleiding	0.06	0.46	0.03	0.41	0.01	0.73	0.01	0.05	0.01	0.61
Luntersebeek	0.14	0.45	0.05	0.17	0.02	0.63	0.00	0.40	–[a]	–[a]
Beemster	0.31	−0.99	0.42	0.10	0.50	0.13	0.30	0.47	0.25	0.20
Dwarsdiep	0.18	−0.22	0.08	−0.11	0.03	0.39	0.02	−0.37	–[a]	–[a]
Roggelsebeek	0.05	0.93	0.03	0.26	0.01	0.61	0.01	0.40	0.00	0.83
Linde	1.62	0.02	1.90	0.12	0.31	0.30	0.15	0.77	0.11	0.99
Reusel	0.08	0.73	0.04	0.17	0.02	0.28	0.02	0.63	0.02	0.82
Delfland	0.76	−0.47	0.84	0.80	1.10	0.57	0.58	0.72	0.83	0.86
Aa	0.10	0.82	0.04	0.65	0.02	0.75	0.01	0.73	0.01	0.90
Regge	0.08	0.80	0.04	0.55	0.02	0.28	0.01	0.03	0.00	−0.62

Note. Indicated are the slope of the line and the Pearson correlation coefficient per class, going from shallow groundwater table depths (class 1; the blue lines in Figure 6b) to the deep groundwater depths (class 5; the red lines in Figure 6b).
[a]No (summer) discharge simulated in this class.

are colored, per class, based on the simulated initial groundwater table depth at the start of the event. Per class, a regression line is adjusted to the corresponding points. The slope of this regression line is for most catchments steeper for shallower groundwater depths (Table 2). The conclusion is that catchments respond faster to rainfall when the initial conditions are wet and thus less water can be stored in the soil, leading to more quick runoff processes, a higher runoff ratio, and higher discharge peaks. During drier conditions, a rainfall event primarily fills up the available storage followed by a sometimes minor or even completely absent discharge response. Therefore, the MAE of the discharge forecast is not always directly related to the rainfall intensity for dry initial conditions, but more to the available storage capacity in the catchment.

Thus, the MAE of the discharge forecast increases with increasing mean event rainfall intensity, and more strongly so for moist initial conditions than for drier conditions. This effect of the initial conditions is particularly visible for the freely draining catchments. It is less clear for the Gouwepolder, Beemster, and Delfland. Delfland is an exception to this and shows an almost linear relationship, similar to the MAE of the rainfall forecast. We expect this to be a result of the large number of greenhouses and paved areas in this region, which leads to a quick response of the hydrological system to rainfall. The polders Gouwepolder and Beemster have regulated groundwater table depths, which implies less variation in the initial groundwater depths as compared to the other catchments, leading to an indistinct effect of the initial groundwater depths on the MAE of the discharge forecast.

3.3.2. Rainfall Variability in Space

Next to the mean rainfall intensity, the MAE of both the rainfall and discharge forecasts increase with increasing spatial rainfall variability (I_σ, Figure 7). Overall, the results in this figure are similar to those presented in Figure 6. Nonetheless, the catchment size appears to play a role here. In Figure 7a, a linear regression line is adjusted to the data points. The slope of these lines decreases with increasing catchment size (from left top to bottom right), from 0.27 for the Hupsel Brook catchment, to 0.06 for the Aa and Regge catchments, respectively (Table 3). This indicates that the spatial rainfall variability, which is generally higher for small-scale convective events, has more impact on the forecasts for smaller catchments than for larger ones. For a small catchment such as the Hupsel Brook, a high spatial variability of the rainfall fields implies that it becomes challenging to predict whether the rainfall will fall inside or outside the catchment. For larger catchments, this is less of a concern, as the rainfall probably falls somewhere in the catchment area, leading to less uncertainty in the forecast. A similar

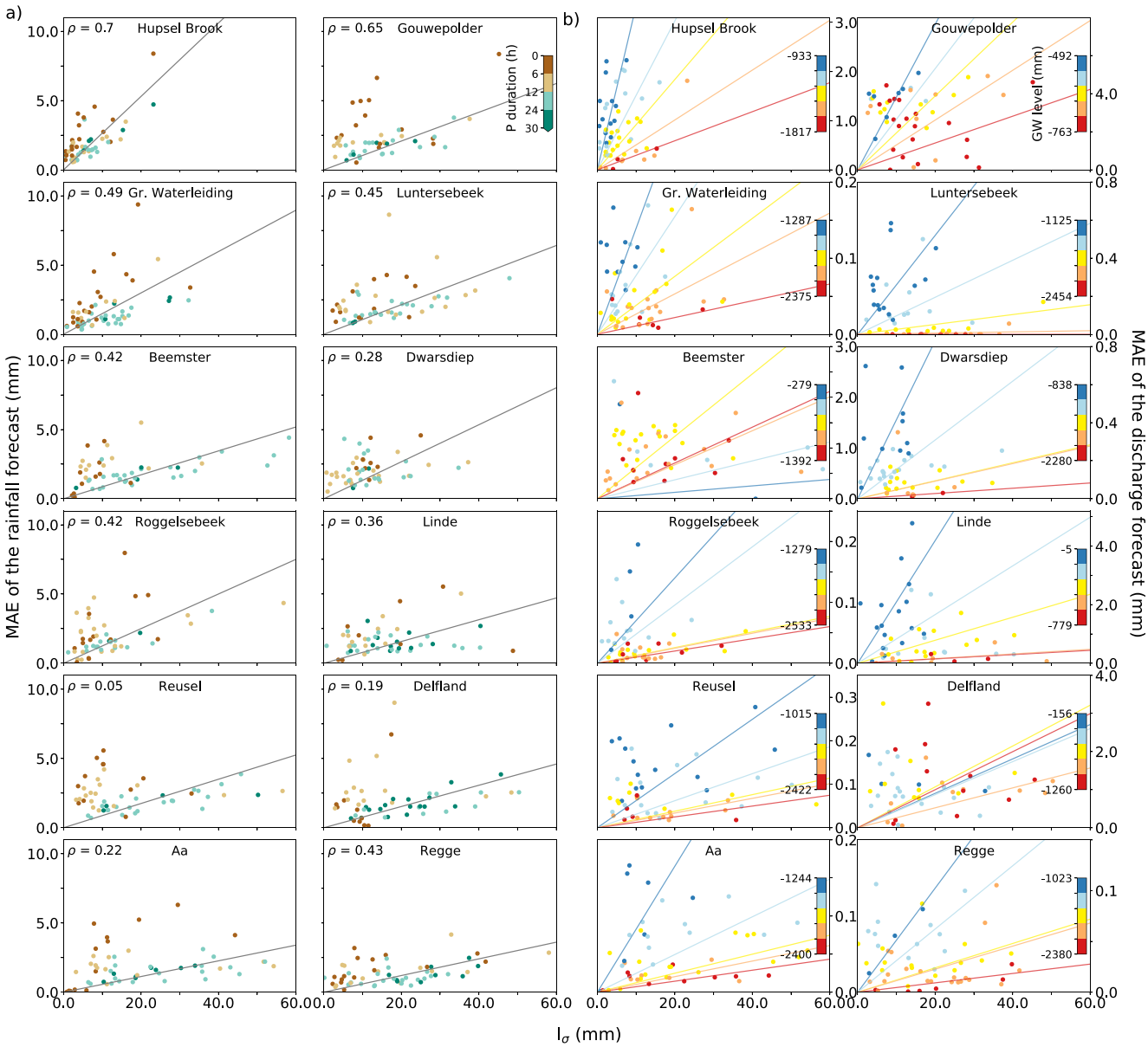

Figure 7. Relationship between the mean event rainfall variability (based on all rainy 5-min instances) and the mean absolute error (MAE) of the: (a) rainfall forecast and (b) the discharge forecast per catchment. The rainfall MAE was calculated for the 3-hr sum (the first 3 hours of the nowcast) of the forecast and reference rainfall. The 12-hr sum (the entire forecast horizon) was used for the discharge MAE. The gray lines in (a) show the results of a linear regression through all points per catchment. The colors of the dots in (a) indicate the rainfall duration during the event and in (b) of the groundwater table depth (in mm) at the start of the event. The groundwater depths are subdivided in classes of equal size (representing 20% of the values between minimum and maximum simulated groundwater level), but only the shallowest and deepest groundwater depths are indicated. Per groundwater class, a linear regression line with the same color is adjusted to the corresponding points. The statistics of these lines and the regression lines in (a) can be found in Table 3.

catchment size dependency is present in the MAE of the discharge forecast (Figure 7b), where the slopes of the individual regression lines per groundwater table class decrease with increasing catchment size.

Concluding, the skill of both the rainfall and discharge forecast decrease with increasing rainfall intensity. Although the increase in the forecast error is nearly linear for the rainfall forecast with increasing rainfall intensity, the error in the discharge forecast also depends on the initial conditions. The error is more pronounced for shallower initial groundwater table depths than for drier conditions, which is particularly prominent in the freely draining catchments in this study. Overall, the discharge forecasts based on rainfall forecasts from RM-DR, PS-D, and PS-P reach longer skillful lead times than the forecasts using the other (nowcasting) methods. In addition, the

Table 3
Statistics of the Linear Regression Lines (Forced Through Zero) That Were Adjusted to the Data Points of Figure 7, Which Corresponds to the Mean Event Rainfall Variability Dependence Analysis

	P forecast		Class 1		Class 2		Class 3		Class 4		Class 5	
Catchment	Slope	ρ	Slope	ρ	Slope	ρ	Slope	ρ	Slope	ρ	Slope	ρ
Hupsel Brook	0.27	0.70	0.33	0.34	0.15	0.77	0.09	0.34	0.05	0.85	0.03	0.99
Gouwepolder	0.10	0.65	0.38	0.60	0.27	0.10	0.19	0.30	0.13	0.229	0.07	0.28
Grote Waterleiding	0.15	0.49	0.01	−0.25	0.01	0.50	0.00	0.47	0.00	0.43	0.00	0.70
Luntersebeek	0.11	0.45	0.03	0.32	0.01	0.06	0.00	0.92	0.00	0.47	–[a]	–[a]
Beemster	0.09	0.42	0.01	−0.99	0.02	−0.51	0.06	−0.21	0.03	0.26	0.04	−0.03
Dwarsdiep	0.13	0.28	0.04	−0.11	0.02	0.05	0.00	0.08	0.00	−0.22	–[a]	–[a]
Roggelsebeek	0.13	0.42	0.01	0.90	0.00	0.03	0.00	0.65	0.00	−0.18	0.00	0.73
Linde	0.08	0.36	0.21	−0.01	0.08	−0.20	0.04	0.81	0.08	0.25	0.01	−1.00
Reusel	0.09	0.05	0.01	0.31	0.00	0.15	0.00	−0.27	0.00	0.34	0.00	−0.44
Delfland	0.08	0.19	0.05	−0.72	0.04	−0.26	0.05	−0.19	0.03	0.49	−0.09	0.08
Aa	0.06	0.22	0.01	0.79	0.00	0.62	0.00	0.58	0.00	0.30	0.00	−0.03
Regge	0.06	0.43	0.01	0.99	0.00	−0.08	0.00	0.18	0.00	0.02	0.00	0.85

Note. Indicated are the slope of the line and the Pearson correlation coefficient for the regression line Figure 7a and per class in Figure 7b, going from shallow groundwater table depths (class 1; the blue lines in Figure 7b) to the deep groundwater depths (class 5; the red lines in Figure 7b).
[a]No (summer) discharge simulated in this class.

spatial rainfall variability plays a role in both the rainfall and discharge forecasts, as well. In smaller catchments, the impact of the spatial rainfall variability on the forecast quality is higher.

3.4. Discharge Peak Forecast Verification

3.4.1. Peak Anticipation Time

Figure 8a shows per catchment and method how many hours before the highest discharge during the events, a "correct" forecast, that is, a hit (having a maximum peak magnitude error of ±10% with respect to the reference), is issued. On average for all catchments (see the right column in the figure), the first issue time a "correct" forecast takes place is 129 min prior to the highest discharge for ZP, 352 min for EP, 325 min for RM-S, 342 min for RM-DR, 248 min for PS-D, and 272 min for PS-P. This means that when using RM-DR, a peak discharge can be forecast almost 6 hr prior to the peak occurrence, given the allowed magnitude error of 10%. Although the allowed magnitude error is arbitrary, it allows for comparing the methods with a benchmark, which is ZP here. The timeliness of ZP gives an indication of the response time of the catchment during that event and represents the "forecast" without rainfall input. By using the other methods, the highest discharge in the event can be, on average, forecast 223 (EP), 196 (RM-S), 213 (RM-DR), 119 (PS-D), and 143 min (PS-P) earlier than with ZP. For RM-DR for instance, this indicates that the average gain of using this nowcasting method is that peak discharges can be forecast more than 3 hr earlier than without a rainfall forecasting method.

The gain reached with EP is the highest of all tested methods, which is remarkable. This is mainly caused by the method used, which only focuses on the first issue time a "correct" forecast is issued. If it rains (intensively) at the start of the event, EP may end up issuing a forecast within the 10% magnitude error for the time of the highest discharge, but this does not mean that this forecast is maintained during the subsequent issue times. Hence, this says nothing about the reliability of the forecast, which we will elaborate on in the following paragraphs. In addition, it is notable that the timeliness of PS-D and PS-P is substantially less than that of EP, RM-S, and RM-DR. An explanation for this is that PS-D and, to a lesser extent, PS-P dissipate the smaller scale rainfall fields, that is, these fields get shorter lifetimes in the algorithm. As a result of that, the nowcasts, particularly PS-D, often end up with lower rainfall volumes due to an excess of smoothing in the forecasts (see also Imhoff et al., 2020). Although PS-D and PS-P can give a better representation of the evolution and location of the rainfall fields than the other methods, the forecast rainfall volume is generally crucial in the magnitude of the peak discharge.

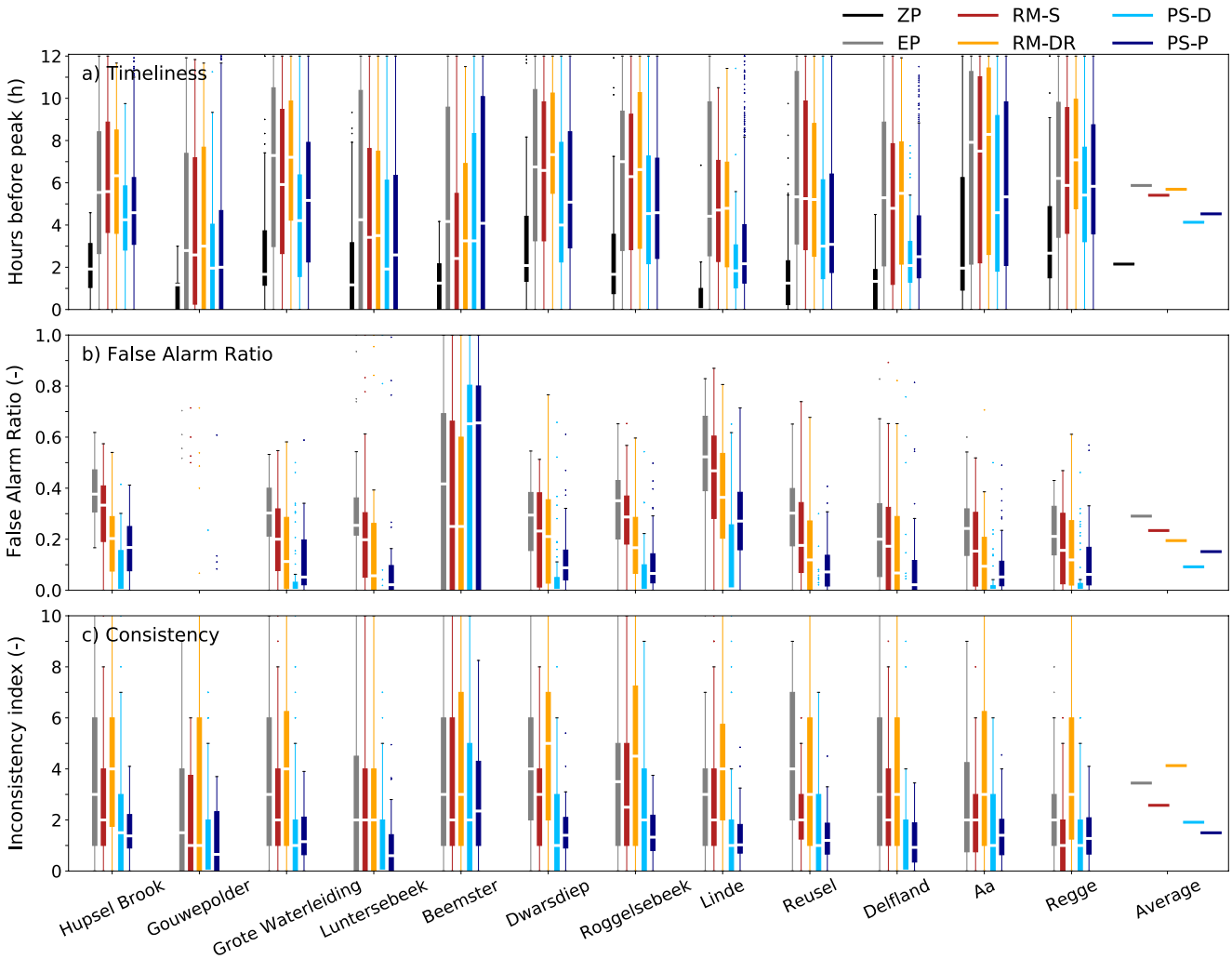

Figure 8. Discharge peak forecast verification per catchment for all methods in this study. All events were taken into account, but only the longest duration was selected when there is an overlap between events of different durations (1, 3, 6, or 24 hr). Panel (a) shows the timeliness of the peak forecast, (b) the False Alarm Ratio, and (c) the forecast consistency. 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, the whiskers indicating 1.5 × IQR, and the outliers indicated by dots. The catchment-average value per method is indicated in the right column of every panel. For PS-P, the metrics were calculated for every ensemble member individually.

3.4.2. False Alarm Ratio

For practical applications, a forecast is only reliable and useful if it is consistent and not too many false alarms are issued during the hours preceding the event. A false alarm in this study was defined as a forecast discharge above the 10% magnitude error of the highest discharge (Section 2.3.5). On average for all 12 catchments and the considered events, the FAR is 0.29 for EP, 0.23 for RM-S, 0.19 for RM-DR, 0.09 for PS-D, and 0.16 for PS-P (Figure 8b). This indicates that, despite the good score on timeliness of EP, 30% of the time EP forecasts a threshold exceedance, it issues a false alarm. An advantage of the underestimations of PS-D and to a lesser extent PS-P in combination with the absence of any rainfall field development in the other methods, is that the FAR of these two nowcasting methods is substantially lower than the FAR of EP, RM-S, and to a lesser extent RM-DR.

The FAR of ZP is zero, and therefore not shown, due to the absence of rainfall in the forecast, inevitably leading to underestimation of the discharge peak. In addition, the median FAR of all methods is zero for the Gouwepolder. In this polder, the pumps have a maximum capacity that is reached during intense rainfall events. So, although the timing of turning the pumps on or off can be wrong, the maximum capacity will not be overestimated, which highly reduces the number of false alarms (note that hydrological forecasting using nowcasting can potentially result in an optimized real-time control of this system).

3.4.3. Consistency

Finally, averaged over all catchments and events, the inconsistency index is 2.33 for EP, 1.49 for RM-S, 3.48 for RM-DR, 1.43 for PS-D, and 1.39 for PS-P (Figure 8c). Similar to the FAR, the inconsistency index for ZP is zero and is therefore not shown. The advantage of the ensemble forecast of PS-P becomes apparent here, as it associates an uncertainty with a forecast (see also Figure 2b) instead of a binary output indicating that the threshold is going to be exceeded ("1") or not ("0"). Overall, the discharge forecasts using RM-S, PS-D, and PS-P are significantly more consistent than those using EP and RM-DR (also visible in Figure 2).

Concluding, for the timeliness of a peak discharge forecast, it is advantageous to make use of a volume preserving nowcasting method. EP, RM-S, and RM-DR clearly outperform the other methods here. Nevertheless, both PS-D and PS-P show low FAR and inconsistency index scores, which gives a forecaster trust in the model outcome when a threshold exceedance is forecast. The good performance of EP in terms of timeliness is counterbalanced by both a high FAR and inconsistent forecasts. The timeliness of RM-S is somewhat counterbalanced by its high FAR of 0.23, whereas for RM-DR the inconsistency index was the highest of all tested methods.

4. Discussion

4.1. Relation to Other Studies

In line with the accompanying study by Imhoff et al. (2020), this analysis using 659 individual events yields a statistical foundation to test the hypotheses concerning dependencies of the performance of various nowcasting methods for hydrological forecasting on rainfall and catchment characteristics. Based on the NSE metric, Berenguer et al. (2005) and Heuvelink et al. (2020) found a gain in anticipation time of the discharge forecast of 10–170 min compared to a ZP forecast. The results of this study fall in between the minimum and maximum found in those studies, with on average a maximum gain of 75 min (for PS-D) for the lowest rainfall intensities and an average maximum gain of 23 min (for RM-DR) for rainfall intensities of 5.0 mm hr^{-1} or more. Heuvelink et al. (2020) studied three catchments that were also present in this study, namely the Regge, Grote Waterleiding, and Hupsel Brook. They found a relative gain between 15 and 40 min for the Regge and Grote Waterleiding catchments and up to 60 min for the Hupsel Brook catchment. The relative gain in this study, based on a comparison between RM-DR (an algorithm similar to the approach in Heuvelink et al., 2020) and ZP, is generally higher, with 51 min for Hupsel Brook, 102 min for Grote Waterleiding, and 89 min for the Regge.

The use of the NSE metric is not ideal to analyze forecasts for separate events, as this metric was originally developed for longer discharge time series. However, the use of the metric does allow for a comparison with previous studies, for example, by Berenguer et al. (2005) and Heuvelink et al. (2020). Although the NSE threshold of 0.9, used to define the maximum skillful lead time in the aforementioned (and this) studies, is somewhat arbitrary, it allows for comparing the different nowcasting methods with each other and to calculate the gain in anticipation time with regard to a benchmark (ZP here). This analysis has made it clear that RM-DR, PS-D, and PS-P, which are the more advanced nowcasting methods, outperform the other tested methods. In addition, the analysis, which relates an increase in mean event rainfall intensity to an increase in the MAE of both the rainfall and discharge forecasts, is in agreement with the results from the analysis using the NSE metric where the nowcast skill also decreases with increasing rainfall intensity.

Furthermore, Berenguer et al. (2005) found no significant improvements in the discharge forecasts when rainfall forecasts from S-PROG were compared to Lagrangian persistence, even though this improvement was present in their verification of the rainfall forecasts. A similar conclusion can be drawn from this study when comparing PS-D (similar to S-PROG) and RM-S or RM-DR (similar to Lagrangian persistence). We link this lack of improvement to the underestimations of the forecast rainfall volumes in case of PS-D and to a lesser extent also for PS-P, as became clearly noticeable in the discharge peak forecast verification section (Section 3.4 and Figure 8).

4.2. Discharge Peak Verification

An important reason for implementing rainfall nowcasting in early warning systems is the (potential) ability to timely forecast peak discharges and threshold exceedances. As mentioned in Section 2.3.5, the disadvantage of this large-sample analysis compared to focusing on a few (extreme) events, is that it becomes challenging to set a threshold or peak discharge as constraint for the discharge peak forecast verification. For this reason, we

chose a more pragmatic approach, where for every event the maximum discharge amount that occurred within the considered event was regarded as the threshold for that event. This made sure that a threshold was reached in every event. In addition, to keep the approach straightforward and systematic, we decided to use a fixed forecast horizon of 12 hr, which was generally within the range of response times of the 12 catchments considered (Table 1). However, for the larger catchments with slower response times, such as the Regge and Aa, the highest discharge during an event was not always the discharge peak, as this peak sometimes occurred later. For events where it kept raining after the end of the defined event, this was possible too. Hence, with the chosen approach, we have not always tested the ability of the forecasting system to forecast the discharge peak well, but rather to forecast a high discharge within an artificially set margin (a threshold exceedance), given the time of occurrence. Nevertheless, we expect that a focus on the discharge peak alone would give similar results.

The choice to allow a peak magnitude error of ±10% of the difference between the initial discharge at the start of the event and the highest discharge during the event was subjective. It merely allowed for a relative comparison among the studied methods, rather than that it provided hard numbers about for instance the peak anticipation time. In practice, the allowed error in peak magnitude and timing would depend on the catchment of interest and limits set by the water authorities. To estimate the sensitivity of this choice, that is, allowing a higher or lower magnitude or timing error, we tested the same approach with a magnitude error of 25%, and with a timing error of 30 min before and after the maximum observed discharge in combination with the 10% magnitude error. The results can be found in supplementary Figures S6 and S7. The relative differences, that is, the timeliness compared to benchmark ZP, are small, with for example, for RM-DR relative gains of 213 min (method in this study), 205 min (25% magnitude error allowed), and 230 min (10% magnitude and 30-min timing error allowed) and for PS-P relative gains of 137 min (method in this study), 145 min (25% magnitude error allowed), and 145 min (10% magnitude and 30-min timing error allowed). However, the absolute values can differ substantially and increase for the two other tested constraints in Figures S6 and S7, which indicates that the results from this analysis should be interpreted in a relative sense, that is, compared to a benchmark or to the other nowcasting methods.

4.3. Transferability of Results to Other Regions

This study focused on the Netherlands, with its typical lowland catchments and polder systems. Although we expect that the results in this study are to a certain extent transferable to other lowland regions with a temperate climate and similar radar products, we expect that the results do not hold for mountainous regions, although the error propagation from rainfall nowcast into discharge forecast will be, in principle, comparable. In mountainous regions, orography influences the spatial errors in the radar composite and with that the nowcasts (Anagnostou et al., 2010; Borga, 2002; Gabella et al., 2000). In addition, growth and decay processes, the pitfall of the tested nowcasting methods in this study, dominate over advection in these regions (Foresti et al., 2019; Foresti & Seed, 2015), leading to a different nowcasting and subsequent discharge forecasting skill. In addition, we cannot conclude if these results also hold for urban areas, even though the smallest catchments in this study have the size of urban areas. Nevertheless, promising results have been reported in discharge forecasting studies using nowcasting for mountainous regions (e.g., Berenguer et al., 2005; Germann et al., 2009; Moreno et al., 2013; Poletti et al., 2019) and for urban areas (e.g., Liguori et al., 2012; Sharif et al., 2006).

It is also notable that the employed hydrological models were lumped and semi-distributed, which makes the model results hardly sensitive, or not sensitive at all, to the location of (the forecast) rainfall in the catchment. The location of the rainfall in the catchment, that is, upstream or near the outlet, can influence the catchment response, especially in larger catchments with more heterogeneous terrain or in mountainous catchments, where this effect becomes more pronounced than in the lowland catchments of this study. However, to test this, a fully distributed or semi-distributed (containing a sufficient number of sub catchments to capture the catchment heterogeneity) hydrological model should be used, similar to the setup in Lobligeois et al. (2014), which is a recommendation for future analyses.

4.4. Actual Error With Observations

Finally, the reference in this study was a model run with the observed CARROTS-corrected radar data (Imhoff et al., 2021). This approach discarded any hydrological model and radar QPE errors, as actual discharge

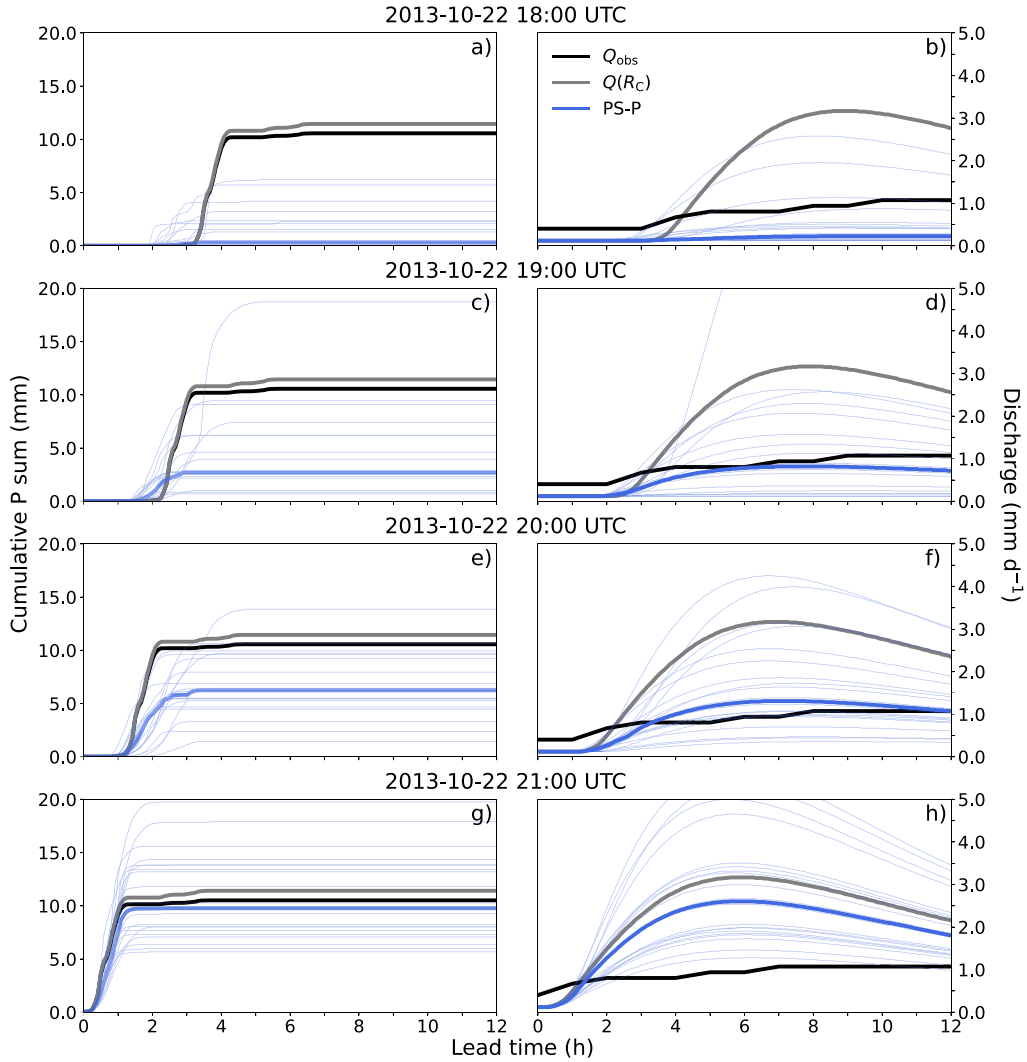

Figure 9. Rainfall and discharge forecasts for four issue times during the discharge peak that occurred in the Hupsel Brook catchment on 22 October 2013. Shown are the rainfall and discharge observations (Q_{obs}; black), the CARROTS-corrected quantitative precipitation estimate ($Q(R_C)$; used as reference in this study) and subsequent hydrological simulation (gray), and the forecasts using rainfall inputs from PS-P (blue). The thin dark blue lines indicate the separate ensemble members of PS-P and the thick dark blue line indicates the ensemble median. Figure 4 provides the results for this event for the other nowcasting methods considered in this study.

observations were not used. However, in a real-life operational setting, these model and QPE errors are present and limit the skill of the nowcast and subsequent discharge forecasts as well. To illustrate this effect, the same forecast as in Figure 4 (only PS-P shown this time for clarity) is compared to the observed discharge in Figure 9. Although the difference between the CARROTS-corrected QPE and observed rainfall was minimal (Figures 9a, 9c and 9e and 9g), the difference between the "reference" discharge and the actually observed discharge is large. Assuming that the discharge observations are correct, this difference is caused by the hydrological model, as the rainfall estimates are close to the observations. We expect that this discrepancy between hydrological model simulation and observation mainly originates from a combination of erroneous initial conditions, model structural errors, and estimation errors in seepage and surface water inlet fluxes. Hence, an adequate rainfall forecasting system is only part of the forecasting chain to obtain improved short-term predictions. Improving model structure, obtaining more accurate estimates for external fluxes, and state updating techniques (i.e., data assimilation) would also improve discharge forecasts.

5. Conclusion and Future Perspectives

Rainfall nowcasting holds a large potential for short-term discharge forecasting and corresponding early warnings. However, a systematic evaluation of this potential, possible pitfalls, and improvements of rainfall nowcasting for hydrological forecasting based on a large-sample (statistical) analysis was not present up to now. In this study, nowcasts for 659 individual events from an earlier study focusing on rainfall predictability by Imhoff et al. (2020) were used to construct discharge forecasts. The events were systematically selected for all four seasons, four durations (1, 3, 6, and 24 hr), and 12 catchments in the Netherlands, with sizes varying from 6.5 to 957 km^2. Four open-source nowcasting algorithms were tested: RM-S, RM-DR, PS-D (similar to S-PROG), and PS-P with 20 ensemble members. In addition, two benchmark forecasting setups were considered: a hydrological simulation using *Eulerian Persistence* (EP) and a forecast without any precipitation input (ZP). For every 5-min time step in the considered events, a discharge forecast with a 12-hr forecast horizon was issued using the available nowcasts for that issue time as forcing for the operational hydrological models of the involved water authorities of the 12 catchments. The reference for verification in this study was the hydrological model simulation with the "observed" radar rainfall in order to discard any model errors and radar QPE or other forcing-related errors.

The rainfall event characteristics are found to determine most of the forecast quality. With increasing rainfall intensity, the skill of both rainfall and discharge forecasts decreases. The error in the rainfall nowcasts increases nearly linearly with rainfall intensity. This relationship is not as clear for the subsequent discharge forecasts, but is found to depend on the initial groundwater table depths, especially for the freely draining catchments considered in this study. The discharge forecast error is generally more pronounced, and shows a steeper increase with rainfall intensity, for shallower initial groundwater depths than for drier initial conditions. Overall, the discharge forecasts with rainfall forecasts as input from RM-DR, PS-D, and PS-P reach longer skillful lead times than the forecasts using the other (nowcasting) methods.

In addition, the spatial rainfall variability plays a role in both the rainfall and discharge forecasts. The spatial variability of convective rainfall systems is generally higher than that of large-scale stratiform systems. Similar to the rainfall intensity, the errors in both the rainfall and discharge forecasts increase with increasing spatial variability of the rainfall events. Moreover, the impact of the spatial rainfall variability is larger for smaller catchments, especially on the rainfall forecast quality. For a small catchment, a high spatial variability of the rainfall fields makes it challenging to predict whether the rainfall will fall inside or outside the catchment, which also impacts the subsequent discharge forecast. For larger catchments, this is less of a concern, as the rainfall probably falls somewhere in the catchments area, leading to less uncertainty in the forecast of discharge at the catchment outlet. That is, provided that only lumped and semi-distributed hydrological models were used in this study.

From a water management perspective, it is crucial to forecast a threshold exceedance or the magnitude of a discharge peak well in advance. The potential of nowcasting for this purpose in hydrological forecasting systems was tested here by setting the highest discharge, per event, as threshold. A forecast was regarded as a hit when this threshold was forecast within a 10% margin above and below the threshold. Everything below this margin was regarded as a miss and everything above it as a false alarm. Although this margin was somewhat arbitrary, it allowed for comparing the methods with a benchmark, which was ZP in this study. Compared to ZP, the highest discharge in the event can be, on average, forecast 223 (EP), 196 (RM-S), 213 (RM-DR), 119 (PS-D), and 143 min (PS-P) earlier than with ZP. For instance for RM-DR, this indicates that the average time gain of using this nowcasting method is that peak discharges can be forecast more than 3 hr earlier than without a rainfall forecasting method. Overall, we found that for the timeliness of peak discharge forecasts, the forecast rainfall volume is a crucial factor, which makes it advantageous to use a volume-preserving nowcasting method.

However, timeliness is only part of the desired model behavior. For trust in the forecast, it is essential that the FAR is low and that the forecast is consistent, that is, subsequent forecasts do not often switch between threshold exceedance and no threshold exceedance. The high timeliness of EP is counteracted by both a high FAR (0.29) and inconsistent forecasts. To the contrary, PS-D and PS-P show both a low FAR and inconsistency index values, which gives a forecaster trust in the model outcome when a threshold exceedance is forecast. Moreover, the focus on the forecast consistency also reveals the advantages of an ensemble forecast (PS-P), as the indication of the uncertainty associated with the forecast leads to a lower inconsistency index.

Hence, all nowcasting methods have shown a benefit for short-term discharge forecasting compared to issuing no rainfall forecasts at all. However, the tested methods all have their shortcomings. For a water manager, it is

recommended to base the choice of a nowcasting method for hydrological predictions on the specific needs of the system. This requires a decision on, for instance, whether an alert is needed as early as possible, or if it is more crucial to have reliable and consistent forecasts. As forecast rainfall volumes have shown to be a crucial factor in the discharge forecasts, a future focus on improving this aspect in the nowcasting algorithms is recommended. Many volume-related errors originate from growth and dissipation processes of the rainfall fields over time, which are not or only stochastically (PS-P) modeled. Object-oriented (e.g., Dixon & Wiener, 1993; Han et al., 2009) nowcasting methods, methods that in some way take into account the rate of growth and dissipation (e.g., Pulkkinen et al., 2020) or methods that also use other predictors for storm initiation and dissipation (e.g., Mueller et al., 2003) have not been included in this study, but could provide a step toward reducing the sharp increase in rainfall and discharge forecast error with increasing rainfall intensity and spatial variability, especially for convective storms. Besides, machine learning initiatives (e.g., Ravuri et al., 2021), possibly in combination with current nowcasting methods, could advance nowcasting methods in this direction too.

Data Availability Statement

Gridded Makkink daily ET_{pot} values are available via https://dataplatform.knmi.nl/dataset/ev24-2, the archived gauge-adjusted radar QPEs via https://dataplatform.knmi.nl/dataset/rad-nl25-rac-mfbs-em-5min-2-0, and the unadjusted radar QPEs via https://doi.org/10.4121/uuid:05a7abc4-8f74-43f4-b8b1-7ed7f5629a01. The nowcasts for the large sample of events, using RM-S, RM-DR, PS-D, and PS-P, are available via https://doi.org/10.4121/16615900.

Acknowledgments
This study has been supported by funding from the DAISY2-project, supported by the European Regional Development Fund (grant no. PROJ-00581) and Deltares' Strategic Research Program. Catchment information, model setups, and the operational Delft-FEWS systems have been provided by the Dutch water authorities involved: Hoogheemraadschap Delfland, Hoogheemraadschap Hollands Noorderkwartier, Hoogheemraadschap Rijnland, Waterschap Aa en Maas, Waterschap De Dommel, Wetterskip Fryslân, Waterschap Limburg, Waterschap Noorderzijlvest, Waterschap Rijn en IJssel, Waterschap Vallei en Veluwe, and Waterschap Vechtstromen. We are thankful for their interest in our work and their help with the setup of the operational systems.

References

Anagnostou, M. N., Kalogiros, J., Anagnostou, E. N., Tarolli, M., Papadopoulos, A., & Borga, M. (2010). Performance evaluation of high-resolution rainfall estimation by X-band dual-polarization radar for flash flood applications in mountainous basins. *Journal of Hydrology*, *394*(1), 4–16. https://doi.org/10.1016/j.jhydrol.2010.06.026

Arnell, N. W., & Gosling, S. N. (2016). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, *134*(3), 387–401. https://doi.org/10.1007/s10584-014-1084-5

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. Retrieved from https://cdn.knmi.nl/system/data_center_publications/files/000/068/061/original/erad2008drup_0120.pdf?1495621011

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 University & Research. https://doi.org/10.18174/454537

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., De Rooy, W., et al. (2017). The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Monthly Weather Review*, *145*(5), 1919–1935. https://doi.org/10.1175/MWR-D-16-0417.1

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

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*, 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Borga, M. (2002). Accuracy of radar rainfall estimates for streamflow simulation. *Journal of Hydrology*, *267*(1), 26–39. https://doi.org/10.1016/S0022-1694(02)00137-3

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

Brauer, C. C., Teuling, A. J., Overeem, A., van der Velde, Y., Hazenberg, P., Warmerdam, P. M. M., & Uijlenhoet, R. (2011). Anatomy of extraordinary rainfall and flash flood in a Dutch lowland catchment. *Hydrology and Earth System Sciences*, *15*(6), 1991–2005. https://doi.org/10.5194/hess-15-1991-2011

Brauer, C. C., Teuling, A. J., Torfs, P. J. J. F., & Uijlenhoet, R. (2014). The Wageningen Lowland Runoff Simulator (WALRUS): A lumped rainfall–runoff model for catchments with shallow groundwater. *Geoscientific Model Development*, *7*(5), 2313–2332. https://doi.org/10.5194/gmd-7-2313-2014

Brauer, C. C., Torfs, P. J. J. F., Teuling, A. J., & Uijlenhoet, R. (2014). The Wageningen Lowland Runoff Simulator (WALRUS): Application to the Hupsel Brook catchment and the Cabauw polder. *Hydrology and Earth System Sciences*, *18*(10), 4007–4028. https://doi.org/10.5194/hess-18-4007-2014

Ceola, S., Laio, F., & Montanari, A. (2014). Satellite nighttime lights reveal increasing human exposure to floods worldwide. *Geophysical Research Letters*, *41*(20), 7184–7190. https://doi.org/10.1002/2014GL061859

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., et al. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Cox, D., Hunt, J., Mason, P., Wheater, H., Wolf, P., Gupta, H., et al. (2002). The challenge of predicting flash floods from thunderstorm rainfall. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, *360*(1796), 1363–1371. https://doi.org/10.1098/rsta.2002.1015

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

European Environment Agency. (2004). *Mapping the impacts of recent natural disasters and technological accidents in Europe* (Vol. 35). In D. Saurí i Pujol & E. T. C. for Terrestrial Environment (Eds.). Office for Official Publications of the European Communities.

Ferraris, L., Rudari, R., & Siccardi, F. (2002). The uncertainty in the prediction of flash floods in the northern Mediterranean environment. *Journal of Hydrometeorology*, *3*(6), 714–727. https://doi.org/10.1175/1525-7541(2002)003<0714:TUITPO>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., Nerini, D., Beusch, L., & Germann, U. (2019). Using a 10-year radar archive for nowcasting precipitation growth and decay: A probabilistic machine learning approach. *Weather and Forecasting*, *34*(5), 1547–1569. https://doi.org/10.1175/WAF-D-18-0206.1

Gabella, M., Joss, J., & Perona, G. (2000). Optimizing quantitative precipitation estimates using a noncoherent and a coherent radar operating on the same area. *Journal of Geophysical Research: Atmospheres*, *105*(D2), 2237–2245. https://doi.org/10.1029/1999JD900420

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., & 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

Gerritsen, T. (2019). *Hydrological intercomparison of rain gauge, weather radar and satellite observations*. Unpublished master's thesis. Wageningen University & Research.

Golding, B. W. (1998). Nimrod: A system for generating automated very short range forecasts. *Meteorological Applications*, *5*(1), 1–16. https://doi.org/10.1017/S1350482798000577

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Han, L., Fu, S., Zhao, L., Zheng, Y., Wang, H., & Lin, Y. (2009). 3D convective storm identification, tracking, and forecasting — An enhanced TITAN algorithm. *Journal of Atmospheric and Oceanic Technology*, *26*(4), 719–732. https://doi.org/10.1175/2008JTECHA1084.1

Hazenberg, P., Leijnse, H., & Uijlenhoet, R. (2014). The impact of reflectivity correction and accounting for raindrop size distribution variability to improve precipitation estimation by weather radar for an extreme low-land mesoscale convective system. *Journal of Hydrology*, *519*, 3410–3425. https://doi.org/10.1016/j.jhydrol.2014.09.057

Heuvelink, D., Berenguer, M., Brauer, C. C., & Uijlenhoet, R. (2020). Hydrological application of radar rainfall nowcasting in The Netherlands. *Environment International*, *136*, 105431. https://doi.org/10.1016/j.envint.2019.105431

Hiemstra, P., & Sluiter, R. (2011). *Interpolation of Makkink evaporation in The Netherlands*. Tech. Rep. No. TR-327. De Bilt, Royal Netherlands Meteorological Institute.thesis. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.705.9459&rep=rep1&type=pdf

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, *3*(9), 816–821. https://doi.org/10.1038/nclimate1911

Imhoff, R. O., Brauer, C., van Heeringen, K.-J., Leijnse, H., Overeem, A., Weerts, A., & Uijlenhoet, R. (2021). A climatological benchmark for operational radar rainfall bias reduction. *Hydrology and Earth System Sciences*, *25*(7), 4061–4080. https://doi.org/10.5194/hess-25-4061-2021

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*(8), e2019WR026723. https://doi.org/10.1029/2019WR026723

Jongman, B., Ward, P. J., & Aerts, J. C. J. H. (2012). Global exposure to river and coastal flooding: Long term trends and changes. *Global Environmental Change*, *22*(4), 823–835. https://doi.org/10.1016/j.gloenvcha.2012.07.004

Klein Tank, A., Beersma, J., Bessembinder, J., van den Hurk, B., & Lenderink, G. (2014). *KNMI '14: Climate scenarios for The Netherlands : A guide for professionals inclimate adaption*. KNMI. Retrieved from https://edepot.wur.nl/328690

Kroeger, T., Timofte, R., Dai, D., & Van Gool, L. (2016). Fast optical flow using dense inverse search. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *European Conference on Computer Vision* (Vol. 9908, pp. 471–488). Springer. https://doi.org/10.1007/978-3-319-46493-0_29

Ladson, A. R., Brown, R., Neal, B., & Nathan, R. (2013). A standard approach to baseflow separation using the Lyne and Hollick filter. *Australian Journal of Water Resources*, *17*(1), 25–34. https://doi.org/10.7158/13241583.2013.11465417

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, *2*(2), 164–168.

Liguori, S., Rico-Ramirez, M., Schellart, A., & Saul, A. (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

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. https://doi.org/10.1002/hyp.8415

Liguori, S., & Rico-Ramirez, M. A. (2013). A practical approach to the assessment of probabilistic flow predictions. *Hydrological Processes*, *27*(1), 18–32. https://doi.org/10.1002/hyp.9468

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*(14), L14801. https://doi.org/10.1029/2005GL023451

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., & Loumagne, C. (2014). When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. *Hydrology and Earth System Sciences*, *18*(2), 575–594. https://doi.org/10.5194/hess-18-575-2014

Loos, R. (2015a). *Making WALRUS applicable for large catchments: A case study in the reusel catchment*. Unpublished Master's Thesis. Wageningen University & Research.

Loos, R. (2015b). *MSc internship report at water authority Vechtstromen: Development of WALRUS models for FEWS Vecht*. Tech. Rep. Wageningen University & Research.

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 674–679). San Fransisco, CA: Morgan Kaufmann Publishers Inc.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441.

Marshall, J. S., Hitschfeld, W., & Gunn, K. L. S. (1955). Advances in radar weather. In H. E. Lansberg (Ed.), *Advances in geophysics* (Vol. 2, pp. 1–56). Academic Press Inc.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245. https://doi.org/10.2307/1268522

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

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Uijlenhoet, R., Mizukami, N., & Clark, M. P. (2016). HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models. *Hydrology and Earth System Sciences*, *20*(3), 1069–1079. https://doi.org/10.5194/hess-20-1069-2016

Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "Assessment of economic flood damage". *Natural Hazards and Earth System Sciences*, *10*(8), 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010

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

Moulin, L., Gaume, E., & Obled, C. (2009). Uncertainties on mean areal precipitation: Assessment and impact on streamflow simulations. *Hydrology and Earth System Sciences*, *13*(2), 99–114. https://doi.org/10.5194/hess-13-99-2009

Mueller, C., Saxen, T., Roberts, R., Wilson, J., Betancourt, T., Dettling, S., & Yee, J. (2003). NCAR auto-nowcast system. *Weather and Forecasting*, *18*(4), 545–561. https://doi.org/10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nerini, D., Besic, N., Sideris, I., Germann, U., & Foresti, L. (2017). A non-stationary stochastic ensemble generator for radar rainfall fields based on the short-space Fourier transform. *Hydrology and Earth System Sciences*, *21*(6), 2777–2797. https://doi.org/10.5194/hess-21-2777-2017

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*(10). 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*(7), 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*(12). 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 J. Bech (Ed.), *Doppler radar observations - Weather radar, wind profiler, ionospheric radar, and other advanced applications*. InTech. Retrieved from http://www.intechopen.com/books/doppler-radar-observations-weather-radar-wind-profiler-ionospheric-radar-and-other-advanced-applications/nowcasting

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

Prinsen, G., Hakvoort, H., & Dahm, R. (2010). Neerslag-afvoermodellering met SOBEK-RR. *Stromingen*, *15*(4), 8–24.

Pulkkinen, S., Chandrasekar, V., von Lerber, A., & Harri, A.-M. (2020). Nowcasting of convective rainfall using volumetric radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, 1–15. https://doi.org/10.1109/TGRS.2020.2984594

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

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

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

Sampson, C. C., Fewtrell, T. J., O'Loughlin, F., Pappenberger, F., Bates, P. B., Freer, J. E., & Cloke, H. L. (2014). The impact of uncertain precipitation data on insurance loss estimates using a flood catastrophe model. *Hydrology and Earth System Sciences*, *18*(6), 2305–2324. https://doi.org/10.5194/hess-18-2305-2014

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*(10), 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

Sharif, H. O., Yates, D., Roberts, R., & Mueller, C. (2006). The use of an automated nowcasting system to forecast flash floods in an urban watershed. *Journal of Hydrometeorology*, *7*(1), 190–202. https://doi.org/10.1175/JHM482.1

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

Stelling, G. S., & Duinmeijer, S. P. A. (2003). A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International Journal for Numerical Methods in Fluids*, *43*(12), 1329–1354. https://doi.org/10.1002/fld.537

Stelling, G. S., & Verwey, A. (2006). Numerical flood simulation. In *Encyclopedia of hydrological sciences. Part 2: Hydroinformatics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa025a

Sun, J., Xue, M., Wilson, J. W., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., et al. (2014). Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, *95*(3), 409–426. https://doi.org/10.1175/BAMS-D-11-00263.1

Sun, Y., Bao, W., Valk, K., Brauer, C. C., Sumihar, J., & Weerts, A. H. (2020). Improving forecast skill of lowland hydrological models using ensemble Kalman filter and unscented Kalman filter. *Water Resources Research*, *56*(8), e2020WR027468. https://doi.org/10.1029/2020WR027468

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

UNISDR. (2002). *Guidelines for reducing flood losses*. Tech. Rep. United Nations International Strategy for Disaster Reduction (UNISDR). Retrieved from https://www.undrr.org/publication/guidelines-reducing-flood-losses

van Osnabrugge, B., Uijlenhoet, R., & Weerts, A. (2019). Contribution of potential evaporation forecasts to 10-day streamflow forecast skill for the Rhine River. *Hydrology and Earth System Sciences*, *23*(3), 1453–1467. https://doi.org/10.5194/hess-23-1453-2019

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

Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., Beek, R. v., Bierkens, M. F. P., et al. (2013). Assessing flood risk at the global scale: Model setup, results, and sensitivity. *Environmental Research Letters*, *8*(4), 044019. https://doi.org/10.1088/1748-9326/8/4/044019

Wasko, C., & Guo, D. (2021). Package hydroEvents. Retrieved from https://cran.rstudio.com/web/packages/hydroEvents/index.html

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