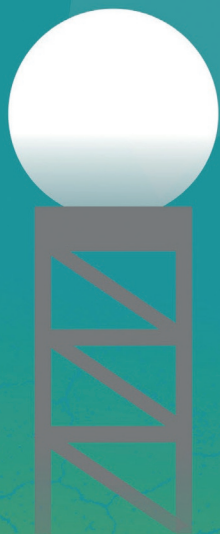


RAINFALL NOWCASTING FOR FLOOD EARLY WARNING



RUBEN O. IMHOFF

Propositions

- 1 | Weather radar data from neighbouring countries are essential for a well-performing Dutch rainfall nowcasting product.
(this thesis)
- 2 | Adequate bias adjustment is indispensable when using radar rainfall data for operational water management.
(this thesis)
- 3 | The communication of scientific results to users is more important than the result itself.
- 4 | Training of PhD candidates should focus more on a non-academic career, as most PhDs leave academia.
- 5 | If we can learn one thing from the COVID pandemic, it is that nobody acts on forecasts anyway.
- 6 | It is a fact that science is not an opinion.

Propositions belonging to the thesis, entitled:

Rainfall nowcasting for flood early warning

Ruben Imhoff

Wageningen, 1 November 2022

Rainfall nowcasting for flood early warning

Ruben Olaf Imhoff

Thesis committee

Promotors

Prof. Dr A. H. Weerts
Special professor, Hydrological Predictability
Wageningen University & Research

Prof. Dr R. Uijlenhoet
Professor of Hydrology & Water Resources
Delft University of Technology

Co-promotor

Dr C. C. Brauer
Lecturer and researcher, Hydrology and Quantitative Water Management Group
Wageningen University & Research

Other members

Prof. Dr J. Vila-Guerau de Arellano, Wageningen University & Research
Prof. Dr P. Willems, KU Leuven, Belgium
Dr T. Winterrath, Deutscher Wetterdienst, Offenbach am Main, Germany
Dr U. Germann, MeteoSwiss, Locarno-Monti, Switzerland

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE).

Rainfall nowcasting for flood early warning

Ruben Olaf Imhoff

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on 1 November 2022
at 1:30 pm in the Omnia Auditorium.

R. O. Imhoff
Rainfall nowcasting for flood early warning
ix+245 pages.

PhD thesis, Wageningen University, Wageningen, The Netherlands (2022)
With references, with summaries in English and Dutch

ISBN 978-94-6447-309-4
DOI 10.18174/573867
© 2022 R. O. Imhoff

Nomenclature

List of abbreviations

ACCORD	A Consortium for Convection-scale Modeling Research and Development
ALADIN	Aire Limitée Adaptation Dynamique Développement International (International development for limited-area dynamical adaptation)
ALARO	ALADIN and AROME combined model
ANC	Auto-nowcast system
AR model	Autoregressive model
AROME	Application de la Recherche à l'Opérationnel à Meso-Echelle (Applications of Research to Operations at Mesoscale)
BS	Brier score
BSS	Brier skill score
CAPPI	Constant plan position indicator
CARROTS	Climatology-based Adjustments for Radar Rainfall in an Operational Setting
CDF	Cumulative distribution function
CML	Commercial microwave link
CPU	Central processing unit
CRPS	Continuous ranked probability score
CSI	Critical success index
DJF	December-January-February (meteorological winter)
ECMWF	European Centre for Medium-Range Weather Forecasts
EP	Eulerian persistence
FAR	False alarm rate
FFT	Fast Fourier transform
FN	False negatives
FP	False positives
FSE	Fractional standard error
FSS	Fractions skill score
HR	Hit rate
HRRR	High-resolution rapid refresh
IQR	Interquartile range
JJA	June-July-August (meteorological summer)
KGE	Kling-Gupta efficiency
KNMI	Koninklijk Nederlands Meteorologisch Instituut (Royal Netherlands Meteorological Institute)
LAM	Limited area model
LHS	Latin-Hypercube sampling
LPBM	Lagrangian probability matching scheme
MAE	Mean absolute error
MAM	March-April-May (meteorological spring)

MFB	Mean field bias
NCEP	National Centers for Environmental Prediction
NSE	Nash-Sutcliffe efficiency
NWP	Numerical weather prediction
QPE	Quantitative precipitation estimate
QPF	Quantitative precipitation forecast
PS-A	Pysteps advection
PS-D	Pysteps deterministic
PS-P	Pysteps probabilistic
RMI	Royal Meteorological Institute of Belgium
RM-DR	Rainymotion DenseRotation
RM-S	Rainymotion Sparse
RMSE	Root mean square error
ROC	Receiver operating characteristic
SSFT	Short-space Fourier transform
SOBEK RR	SOBEK rainfall-runoff model
SOBEK RR-CF	SOBEK rainfall-runoff channel-flow model
SON	September-October-November (meteorological autumn)
S-PROG	Spectral Prognosis
STEPS	Short-Term Ensemble Prediction System
TITAN	Thunderstorm Identification, Tracking, Analysis and Nowcasting
TN	True negatives
TP	True positives
TPU	Tensor processing unit
USD	United States Dollars
UTC	Universal time coordinated
VIL	Vertical integrated liquid content
VPR	Vertical profile of reflectivity
WALRUS	Wageningen Lowland Runoff Simulator
WMO	World Meteorological Organization
ZP	Zero precipitation

List of symbols

ϵ_k	Perturbation field at spatial scale level k (-)
λ_k^{ext}	Ratio of the explained to the unexplained variance of the extrapolation component at scale level k (-)
λ_k^{nwp}	Ratio of the explained to the unexplained variance of the NWP component at scale level k (-)
μ	Mean of variable (L T ⁻¹)
ρ	Pearson's correlation (-)
σ	Standard deviation of variable (L T ⁻¹)
a_S	Surface water area fraction in WALRUS (-)
B	Deterministic categorical bias (-)
c_D	Channel depth parameter of WALRUS (L)
c_G	Groundwater reservoir constant in WALRUS (L T)
c_Q	Quickflow reservoir constant in WALRUS (T)
c_S	Surface water parameter of WALRUS (L T ⁻¹)
c_V	Vadose zone relaxation time parameter of WALRUS (T)
c_W	Wetness index parameter of WALRUS (L)
ET_{pot}	Potential evapotranspiration (L T ⁻¹)
F_{clim}	Climatological adjustment factor (-)
F_{MFB}	MFB adjustment factor (-)
F_{MFBS}	Adjustment factor to spatially adjust R_{MFB} (-)
F_S	Spatial adjustment factor (-)
f_{XG}	Groundwater flux (L T ⁻¹)
f_{XS}	Surface water supply (L T ⁻¹)
G	Rain gauge rainfall (L T ⁻¹)
I_σ	Rainfall variability following Lobligois et al. (2014) (L T ⁻¹)
k	Spatial scale level (-)
P	Precipitation (L T ⁻¹)
q^{nwp}	Regression coefficient for the NWP skill per lead time (-)
Q	Discharge (L T ⁻¹)
$Q(R_A)$	Simulated discharge with rainfall input from R_A (L T ⁻¹)
$Q(R_C)$	Simulated discharge with rainfall input from R_C (L T ⁻¹)
$Q(R_{\text{MFB}})$	Simulated discharge with rainfall input from R_{MFB} (L T ⁻¹)
$Q(R_U)$	Simulated discharge with rainfall input from R_U (L T ⁻¹)
R	Rainfall rate (L T ⁻¹)
R_A	Gauge-adjusted radar QPE (L T ⁻¹)
R_C	CARROTS-corrected radar QPE (L T ⁻¹)
R_{MFB}	MFB-adjusted radar QPE (L T ⁻¹)
R_U	Unadjusted radar QPE (L T ⁻¹)
v	Velocity field (L T ⁻¹)
w_k^ϵ	Blending weight for the noise component at spatial scale level k (-)
w_k^{ext}	Blending weight for the extrapolation component at spatial scale level k (-)
w_k^{nwp}	Blending weight for the NWP component at spatial scale level k (-)

Nomenclature

$\gamma_k^{\text{blended}}$	Blended cascade at spatial scale level k (-)
γ_k^{ϵ}	Noise cascade at spatial scale level k (-)
γ_k^{ext}	Extrapolation cascade at spatial scale level k (-)
γ_k^{nwp}	NWP cascade at spatial scale level k (-)
Z_h	Reflectivity ($\text{mm}^6 \text{mm}^{-3}$)

Contents

Nomenclature	v
1 Introduction	1
2 Description of data, study areas and methods	15
3 Operational radar rainfall bias reduction	35
4 Comparing and evaluating radar rainfall nowcasting techniques	57
5 Nowcasting with opportunistic rainfall data	81
6 Blending nowcasts and numerical weather prediction to extend skillful lead times	95
7 Evaluation of rainfall nowcasting for flood early warning	127
8 Synthesis	157
9 Appendices	175
Bibliography	213
Statement of authorship contribution	229
Summary	231
Samenvatting	235
Acknowledgements	239
List of publications	241
Graduate school certificate	243



1

Introduction

“Pourquoi les météorologistes ont-ils tant de peine à prédire le temps avec quelques certitudes?”

—Henri Poincaré, *Science et Méthode* (1909)

1.1 | The motivation to improve rainfall forecasts for operational flood forecasting

LIVING with water in flood-prone areas is a constant balance between the benefits of the water availability, trading routes and fertile soils, and the occasionally disastrous effects of flooding and droughts on the livability and economy of these areas (UNISDR, 2002; European Environment Agency, 2004; Merz et al., 2010; Jongman et al., 2012; Ward et al., 2013; Ceola et al., 2014). Just a year ago, in July 2021, severe flash flooding took place in Belgium, Germany and the Netherlands (the Netherlands and Belgium are also the focus area of this thesis). The cause for these floods was a persistent mesoscale low-pressure system in Northwestern and Central Europe during this period, which locally resulted in extreme rainfall amounts. The floods caused over 240 casualties, of which most in Belgium and Germany, and led to more than 25 billion USD in economic and infrastructural damages (AON, 2021; Koks et al., 2021; Kreienkamp et al., 2021). These recent floods are illustrative for the risk of (flash) flooding, even in countries that, in theory, have the means to implement effective adaptation to flood risk (Jongman, 2018).

From a (flash) flood perspective, the disadvantages of living with water, particularly risk and damage, can be reduced when timely action on a forecast flood is taken. This is reachable when flood early warning systems are in place (UNISDR, 2002; Pappenberger et al., 2015). With most people living in flood-prone areas worldwide, it is not surprising that well-established flood forecasting systems have a significant monetary and humanitarian benefit. Pappenberger et al. (2015) estimated, based on an analysis for the European Flood Awareness System, that every 1 Euro invested in flood forecasting systems, has a return in the order of 400 Euros. For Bangladesh, this is even a ratio of 559:1 (Subbiah et al., 2008). Hallegatte (2012) provides a more conservative estimate, but still estimates a global average return of in between 4 and 36 Euros for every 1 Euro invested in (hydrometeorological) early warning systems.

An important component of the flood forecasts in these early warning systems is the precipitation forecasts that feed the underlying hydrological models of the water system. Hence, to enable flood early warning, accurate and timely precipitation forecasts are a prerequisite (Ingram et al., 2002; Thorndahl et al., 2013; Pappenberger et al., 2015). The degree to which a correct precipitation forecast can be issued, the so-called **predictability** of the system, depends on the precipitation system and is especially challenging for short-lived high-intensity rainfall cells that can cause flash floods in fast responding small, urban, or mountainous catchments and inundations in polder catchments (e.g. Cox et al., 2002; Ferraris et al., 2002). In a changing climate, both the frequency and severity of these intense rainfall events and subsequent probability of severe (local flash) flooding are expected to increase (IPCC, 2011, 2013, 2014, 2021; Willems & Olsson, 2012; Hirabayashi et al., 2013; Arnell & Gosling, 2016; Kreienkamp et al., 2021). This increase will generate more flood damages as well (Pielke & Downton, 2000; Botzen & Van Den Bergh, 2008; Arnbjerg-Nielsen & Fleischer, 2009; Bouwer et al., 2010; Mirza, 2011;

Frame et al., 2020). Davenport et al. (2021) estimated that over one-third of the cumulative flood damages in the US between 1988 and 2017 have already been caused by climate change-induced precipitation changes. In addition, our flood exposure increases, too, due to a rapid urbanization degree, growing population and increasing wealth (Kundzewicz et al., 2014). Hence, it is essential to further improve rainfall forecasting to enable, or keep enabling, (flash) flood early warning.

Yet, why is it that rainfall and subsequent (flash) flood forecasting are particularly challenging in fast responding, small catchments, even though their forecasting systems only require rainfall forecasts for a few hours ahead? To answer this question, Section 1.2 starts with a brief overview of the history of weather forecasting, followed by the current state of the art in forecasting, its pitfalls and alternative forecasting methods in Section 1.3. This is followed by the problems and open questions that are addressed in this thesis (Section 1.4 and 1.5), and the outline of this thesis (Section 1.6).

1.2 | Weather forecasting in a historical context

Weather has always played an important role in our societies (Nebeker, 1995; Inness & Dorling, 2012). It determined and still determines our agricultural yield and water availability, which has a direct impact on our society, already since the commencement of agriculture (and before that). In later stages, the weather also dictated the fate of large war and trade shipping fleets, up to the point where our present-day society completely relies on the weather for, for instance, food, travel, energy, construction and leisure (Frisinger, 2018). Being able to predict what the weather is going to do in the coming hours, days, weeks, seasons or even years is extremely valuable, as it allows us to prepare for what is going to happen to increase well-being by either optimizing and maximizing gains (e.g. crop yield) or to mitigate the effect(s) of disruptive events (Inness & Dorling, 2012). Our current society can still, globally, benefit by over 160 billion USD a year from improved weather forecasts (Kull et al., 2021), and that is when it is expressed only in economic terms.

If forecasting were an easy task, however, you would probably not be reading this thesis now. The Babylonians already tried to make weather forecasts by combining their observations of the sky with astronomy, later followed by a more philosophical view on what was called *μετέωρος* (meteorology) and weather forecasting by the Greek (Nebeker, 1995; Taub, 2003; Brutsaert, 2013; Frisinger, 2018). For the centuries after, weather forecasts consisted of communicating a sparse set of observations (Inness & Dorling, 2012). If experienced ‘forecasters’ received observations from locations upwind, they would be able to make a prediction about the weather at their location tomorrow (a simple advection-based view on forecasting, so to say – keep that idea in mind for the remainder of this thesis, where the focus will be on nowcasting). As the meteorological observation techniques evolved rapidly in the 17th and 18th century, achieved by the chemists and physicists of that time, weather forecasting got a stronger observation-based foundation, albeit still heavily reliant on the forecaster’s experience (Nebeker, 1995; Lynch, 2008; Frisinger, 2018).

It took the (scientific) forecasting world until the 19th and early 20th century, the period of the development of thermodynamics, to come up with a set of fundamental physical principles (equations) to, in theory, analytically solve the dynamics of the atmosphere (e.g. in the works

of the American meteorologist Abbe, 1901). This was followed by Vilhelm Bjerkness, who, together with his son and colleagues at the so-called 'Norwegian school' in Bergen (Norway), continued with this perspective and introduced differential equations to describe the evolution of a set of atmospheric variables (Nebeker, 1995; Friedman, 1993). Forecasting had to take place in two steps: an initialisation step, where the initial state of the atmosphere is determined with observations, and a prognostic step in which the equations are used to forecast the future state of the atmosphere (Lynch, 2008). The main issue at that time was that these equations could neither be solved analytically nor numerically.

Lewis Fry Richardson was the first to take this a step further by producing the first numerical weather forecast in which he tried to forecast the atmospheric pressure change six hours later (Nebeker, 1995; Inness & Dorling, 2012). Although this first try was not that successful, it laid the foundation for future numerical weather prediction approaches. In 1922, exactly one century ago to date, Richardson formulated a method to numerically solve the equations that describe atmospheric flow. His idea was to divide the globe into 'cells' with specific dynamic variables at the centre of each of these cells. Richardson imagined, he called it a 'fantasy', a giant forecasting building (Figure 1.1) where every worker would solve the aforementioned equations, pass this on to their neighbours and use the neighbours' information again as initialisation for the calculation of the next time step (Richardson, 1922).

Although Richardson's forecasting imagination may sound surrealistic, his ideas were actually not that far from current reality. If we replace the workers in Richardson (1922) by computer cores, the 'cells' by numerical grid cells and the forecasting building by a super computer facility, we have our current numerical weather prediction (NWP) setup as used by many meteorological offices worldwide (for instance, ECMWF in Europe and NCEP in the USA). At the time of writing this thesis, it is exactly 100 years after Richardson's formulation and we still use the forecasting factory idea. The scientific field has managed to significantly improve the NWP model forecasts, however, by taking into account more and more processes and by running on finer and finer spatial and temporal resolutions (Nebeker, 1995; Simmons & Hollingsworth, 2002).

Nevertheless, a perfect forecast is something that we will never accomplish. This 'problem' was already studied in the early 20th century by Henry Poincaré (see, for example, Poincaré, 1952) and later on further described by Edward Lorenz. Both Poincaré and Lorenz found and describe that small differences in the initial conditions of the aforementioned differential equations can lead to immense, unpredictable differences in the final forecast output. These findings were the starting point of the so-called chaos theory (Lorenz, 1963, 1993). Important in the chaos theory is the scaling issue, both in space and in time (Feigenbaum, 1983; Schertzer & Lovejoy, 1987; Lovejoy, 2019). Atmospheric processes, but for instance also hydrological processes, take place on spatial scales from less than a mm to approximately 10,000 km. The smaller scales, both in space and time, eventually determine what happens on cascades of larger scales, but also the other way around. The space and time scales are coupled, meaning that large spatial scales (in the order of 1–10,000 km) correspond to larger temporal scales (weeks, months, years), while smaller spatial scales (in the order of mm to m) correspond to smaller temporal scales (less than a second to hours; Orlanski, 1975). Although we are only interested in the statistics of the smaller scales for accurate weather predictions on intermediate to large scales, we still need observations and understanding of the processes on all scales to reach that (Lovejoy, 2019).

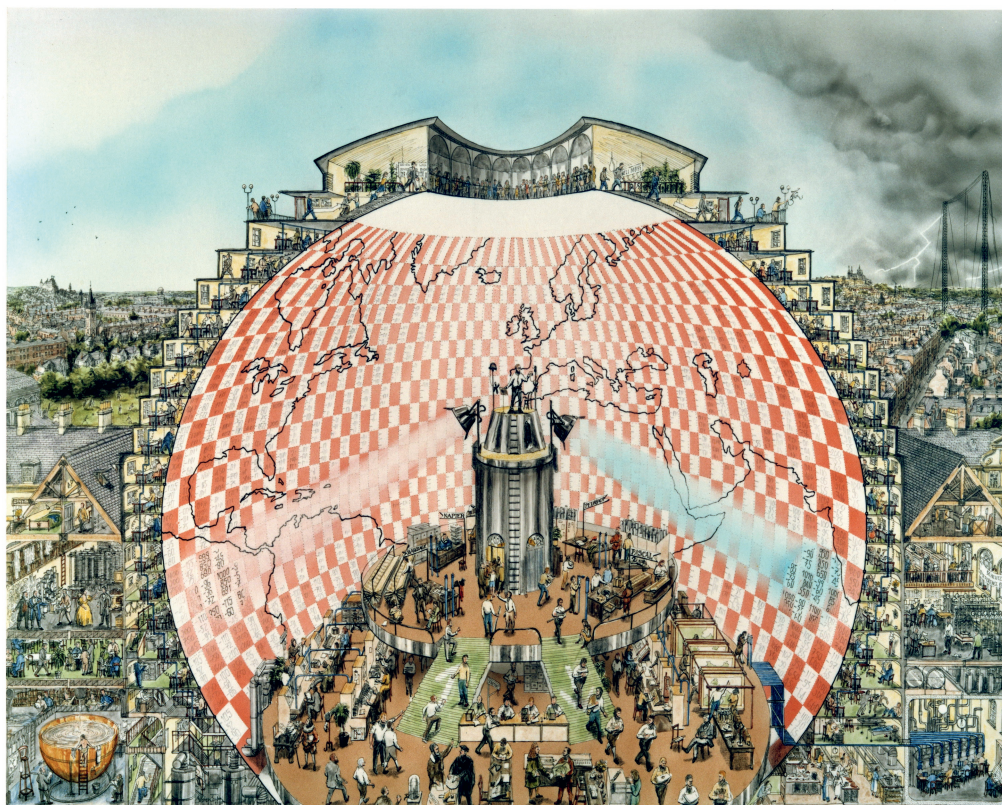


Figure 1.1 | Artist's impression of Richardson's 'Weather Forecasting Factory' ©Stephen Conlin, 1986. Based on the work of L.F. Richardson, and on advice from Prof. John Byrne, Trinity College Dublin. All Rights Reserved.

The initialisation of our NWP models is based on observations, but we do not have observations everywhere at all times and at all spatial scales; let alone our lack of understanding of some of the processes and their interactions (or even the unknown unknowns about these processes) on different spatial and temporal scales. And even if we would do so, a measurement is never perfect. Any inaccuracy in an observation and thus initialisation can already lead to a different outcome, something which increases the further we forecast ahead (Lorenz, 1963, 1993). Hence, forecasting the weather remains challenging and, although NWP models are constantly improving (Inness & Dorling, 2012), we will never be able to produce a perfect forecast.

This holds for any atmospheric variable, but the more a variable fluctuates in space and time, the more difficult it becomes to forecast it well (Orlanski, 1975). Rainfall, and precipitation in general, is highly variable in both space and time. At the same time, we need accurate and timely rainfall forecasts for (flash) flood early warning, as described in Section 1.1. Coming back to the aforementioned scaling issue, flood early warning for large catchments requires accurate forecasts on a different spatio-temporal scale (in the order of 10,000 km and days to weeks) than (flash) flood early warning for small mountainous or urban catchments (in the order of a km or less, and less than an hour to hours; Gupta & Waymire, 1979; Berndtsson &

Niemczynowicz, 1988; Willems & Vrac, 2011). The focus in this thesis is on small and quickly-responding catchments, which require accurate rainfall forecasts on a high spatial resolution and up to only several hours ahead to enable (flash) flood early warning. It may be clear by now that fulfilling this aim is not straightforward, when taking into account the aforementioned challenges of forecasting the weather, particularly on this spatio-temporal scale. The next section will describe the current state of the art in forecasting precipitation, and in particular rainfall in this thesis, including its weaknesses and the knowledge gap that we try to address in this thesis.

1.3 | Present-day rainfall forecasting techniques

1.3.1 | Numerical weather prediction

A century after the work of Richardson (1922), NWP systems still try to numerically solve the set of non-linear equations by subdividing the world or smaller regions into grid cells, both in the horizontal and vertical. Atmospheric processes that take place on smaller scales than this grid are parameterised. Ever since the introduction of the computer, this numerical way of solving the state of the atmosphere has been a constant balance between advances in our scientific knowledge and computational power (Nebeker, 1995; Inness & Dorling, 2012). Therefore, we can partially attribute advances in weather forecasting to past and future computer innovations. This trade-off between understanding of the atmosphere and computational power has also determined our current forecasting setup. In this setup, coarser spatial and temporal scale global weather forecasts are run by for instance the European Centre for Medium-range Weather Forecasts (ECMWF, Europe) and the National Centers for Environmental Prediction (NCEP, USA), for the entire world and up to 10 days into the future (so-called medium range). These forecasts are then used by local meteorological offices as boundary conditions to run so-called Limited Area Models (LAM) for smaller regions on shorter ranges (12–72 h ahead), which makes it feasible to run the models on a finer spatial and temporal resolution, and to take more processes into account (e.g. Tang et al., 2013). The scales on which this is done, are increasingly getting finer, for instance nowadays 1.3 km and 5 min for the Belgian NWP that was used in Chapter 6 of this thesis (Bubnová et al., 1995; Termonia et al., 2018). For a more elaborate background on NWP processes, its evolution and numerics, see Coiffier (2011).

The observation-based initialisation step is still a major part of the NWP forecasting chain and uses a wealth of information sources, from radio soundings and field observations to remotely sensed observations from for instance commercial planes and satellites. Subsequently, data assimilation techniques are used to steer imperfections in the initial observations, due to either the lack of observation accuracy or the lack of a high observation density in both space and time, in the right direction (e.g. Navon, 2009). Hence, current NWP is an interplay between observations, simulation and assimilation.

Despite the aforementioned approaches and improvements to NWP models over time, the rainfall forecasts of NWP models are still not sufficiently accurate for reliable early warnings on the short term (defined as up to six hours into the future), especially when we zoom in on small catchments or urban areas (Bowler et al., 2006; Pierce et al., 2012). The reason for this is that the NWP models are run on either a too coarse temporal resolution or with a too low update frequency to capture the timing of rainfall in these small areas, or they fail to capture the location of (heavy intensity) rainfall cells (Lin et al., 2005; Roberts & Lean, 2008; Berenguer et al.,

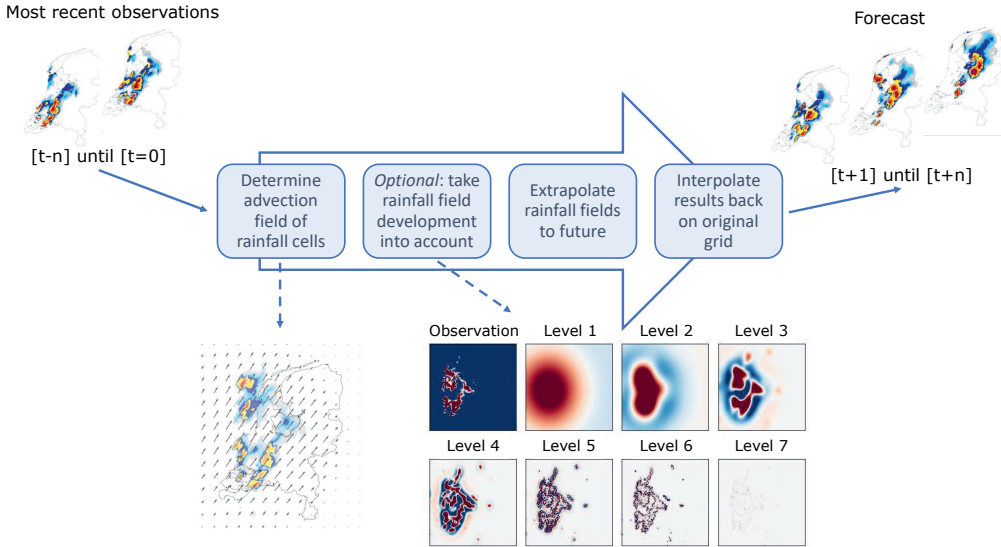


Figure 1.2 | Schematic overview of the construction of a nowcast from the most recent (radar) observations. Four steps are illustrated: (1) determination of the motion of the observed rainfall fields; (2) the option to take rainfall field development into account (not present in plain persistence-based nowcasting), shown is the method from S-PROG and STEPS (Seed, 2003; Bowler et al., 2006; Seed et al., 2013), which assigns different lifetimes to distinct spatial scales; (3) the extrapolation of the rainfall fields to the future, taking steps (1) and (2) into account; and (4) the interpolation of the results back onto the original regular grid for application and visualization purposes.

2012; Pierce et al., 2012). This concerns the scaling issue, mentioned in Section 1.2. The effects of the too low update frequency, which is generally in the order of 3–6 h (for instance, 6 h in the Netherlands and Belgium; Bubnová et al., 1995; Bengtsson et al., 2017; Termonia et al., 2018; de Rooy et al., 2022), are enhanced by the time it takes between the start of the model run and the arrival time of the results at the end users, which can add an additional 2–4 h. Consequently, within those 2–4 h (and during the 6-h validity of the forecast after that) the observation-based initial conditions of the NWP simulations have changed considerably, especially during convective rainfall events. This could cause forecast errors already at the start of the issue time of both the rainfall forecast and the subsequent hydrological forecast (Sun et al., 2014).

Concluding, despite the past and current advances in NWP models, current rainfall forecasts from NWP models still have trouble providing the accuracy and timeliness that is warranted for (flash) flood early warning. Fortunately, there are multiple ways forward. One is the further improvement of NWP models over time and the advancement of rapid update cycle NWP models (e.g. Golding, 1998; Sun et al., 2014). These advancements will not be discussed in this thesis, but will be briefly mentioned in Section 8.2.2. Instead, in this thesis the focus is on another, more statistics and observation-based option to improve short-term rainfall forecasting, as compared to current NWP models. This is so-called nowcasting of rainfall, which will be further described in the next section.

1.3.2 | Nowcasting

The term nowcasting originates from the 1970's and indicates a spectrum of (remotely sensed) observation-based forecasting techniques (Browning, 1980). Nowcasting of rainfall is in principal a (set of) statistical process(es) to extrapolate real-time remotely-sensed quantitative precipitation estimates (QPEs) into the future (Pierce et al., 2012). Although nowcasting methods come in different shapes and sizes, see also the subsequent paragraphs of this section, most nowcasting methods follow the basic steps indicated in Figure 1.2, albeit with different approaches. These steps consist of: (1) the determination of the advection field of the most recent observed rainfall fields, (2) an optional step to take the field development statistically into account, (3) an extrapolation step (taking into account steps 1 and 2) and (4) a final interpolation to get the output forecast on the same regular grid as the input observations. Step 2, the field development, is modelled differently per concept, if present at all. Shown in Figure 1.2 is the method from S-PROG and STEPS (Seed, 2003; Bowler et al., 2006; Seed et al., 2013), which decomposes the rainfall field into a multiplicative cascade of different spatial scales. As mentioned in Section 1.2, these different spatial scales also correspond to different temporal scales (Feigenbaum, 1983; Schertzer & Lovejoy, 1987; Lovejoy, 2019). The S-PROG and STEPS approaches assign different lifetimes to these spatial scales, as a first step to model the development of these spatial scales differently. Other approaches take, for instance, growth and dissipation of rainfall cells into account (e.g. Dixon & Wiener, 1993; Han et al., 2009), or use different information sources to estimate this rainfall field development (e.g. Mueller et al., 2003).

Nowcasting has gained popularity over the past decades, predominantly because nowcasting techniques allow us to take full advantage of present-day high spatial and temporal resolution remotely sensed data (e.g., typically around 1 km and 5 min for weather radars; Serafin & Wilson, 2000; Overeem et al., 2009b). By doing so, forecasts with the same update frequency can be constructed. As the system is also much faster than current NWP models, it is feasible to have these frequent model runs. This strongly reduces the initialisation, latency and update frequency problem of NWP, as was described in Section 1.3.1.

The rainfall nowcasting approach ensures that rainfall fields are located at the correct location at the start of the forecast, which results in more skill at the start time of the forecast than for the NWP rainfall forecasts, which suffer from initialisation issues and a low update frequency. However, as rainfall nowcasts are statistics and extrapolation-based, in contrast to the physics-based NWP forecasting approach, their skill quickly reduces with increasing lead time, due to growth, dissipation, merging and splitting evolution of the rainfall cells (see also Figure 1.3, and Browning, 1980; Germann et al., 2006; Pierce et al., 2012). Previous studies have indicated that maximum skillful lead times of radar-based rainfall nowcasting generally range from less than 30 min for convective rainfall cells to a maximum of 6 h for persistent stratiform events on continental scales (Germann & Zawadzki, 2002; Germann et al., 2006; Lin et al., 2005; Berenguer et al., 2011, 2012; Liguori & Rico-Ramirez, 2012; Foresti et al., 2016; Mejsnar et al., 2018; Ayzel et al., 2019b).

The strength of a nowcasting algorithm in forecasting stratiform or convective rainfall, and with that the aforementioned statistics, depends on the type of nowcasting algorithm that is used. A large number of nowcasting algorithms is available nowadays and we can categorise them in the following five groups: field-based nowcasting methods (e.g. Bowler et al., 2006; Seed, 2003; Seed et al., 2013; Berenguer et al., 2011; Sokol et al., 2017; Ayzel et al., 2019b), object-oriented

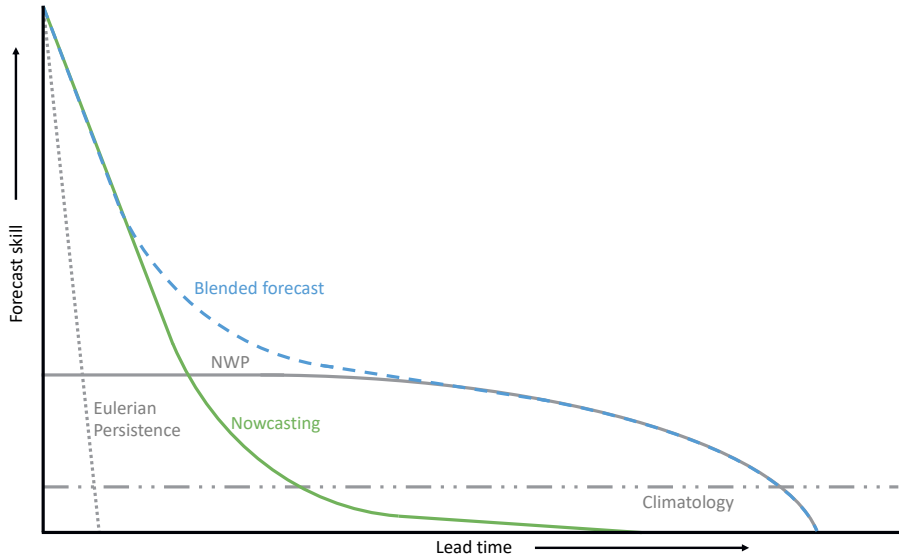


Figure 1.3 | Schematic overview of the rainfall forecast skill as a function of lead time for different forecasting systems: Eulerian Persistence (the method of using the last observation as forecast; dotted grey), the climatology (dash-dotted grey), NWP (grey), nowcasting (green) and blended forecasting (dashed blue). Adjusted after Lin et al. (2005) and Germann et al. (2006). Note that both axes do not contain numbers; the exact skill and lead time are case-specific and depend on variables such as location, (catchment) size, rainfall type and initial conditions. The quantification of this dependence is one of the aims of this thesis.

methods (e.g. Dixon & Wiener, 1993; Han et al., 2009), analogue-based methods (e.g. Atencia & Zawadzki, 2014, 2015; Zou et al., 2020), conceptual reservoir-based models as introduced in the 1980's (Georgakakos & Bras, 1984a,b; Andrieu et al., 2003) and machine-learning methods (e.g. Foresti et al., 2019; Ravuri et al., 2021; Shehu & Haberlandt, 2022). This thesis will only focus on field-based nowcasting methods, which essentially advect the latest rainfall observation with a vector field that is determined with optical flow methods. In addition, these methods can add statistical pre- and post-processing steps to account for e.g. growth and dissipation of rainfall cells. A more elaborate introduction of these field-based methods will follow in Chapter 2.

A more recent advancement in rainfall nowcasting is to account for uncertainty, increasing rapidly with increasing lead time, by means of ensemble nowcasts. Examples of probabilistic nowcasting algorithms are STEPS and the more recent open-source version pysteps (Bowler et al., 2006; Seed et al., 2013; Pulkkinen et al., 2019), SBMcast (Berenguer et al., 2011) and ENS (Sokol et al., 2017). The advantage is that the ensemble forecast can be applied in the discharge forecast to construct probabilistic (flash) flood forecasts. In that way, it can give a rough estimate of the uncertainty of the forecast (Berenguer et al., 2005; Vivoni et al., 2006; Heuvelink et al., 2020).

1.3.3 | Blended rainfall forecasts

Despite the continuous improvements in NWP forecasting models and the emergence of nowcasting techniques for short-term rainfall forecasting, rainfall nowcasting techniques still lose skill too quickly and the NWP rainfall forecasts do not provide sufficient skill on the short term

to provide skillful rainfall forecasts for the entire forecast horizon for flash flood early warning (a horizon of approximately 6 h). A way to extend the skillful lead times of nowcasting is by optimally combining the nowcasts and NWP rainfall forecasts, so-called blending (e.g. Golding, 1998; Bowler et al., 2006; Atencia et al., 2010; Kober et al., 2012, 2014; Bailey et al., 2014; Nerini et al., 2019; Yoon, 2019; Radhakrishnan & Chandrasekar, 2020), or with data assimilation techniques, which update the NWP with radar or nowcast information and the other way around (Golding, 1998; Koopmans et al., 2018; Honda et al., 2022).

With blending, it is, in principle, possible to obtain the best out of both worlds, i.e. rainfall nowcasting and NWP (Figure 1.3). The challenge for a well-performing blending technique is to find the optimal combination between the NWP forecast and rainfall nowcasts for each lead time in the forecast. This can become challenging when the forecasts are quite different. Therefore, state-of-the-art blending techniques try to tackle this with e.g. Bayesian methods (Nerini et al., 2019), the adjustment of the rainfall field mislocation in the NWP forecast at the start of the forecast (Atencia et al., 2010), adjustable weights in space and time, and ensemble forecasts (Bowler et al., 2006; Seed et al., 2013).

1.4 | The missing link in short-term operational flood forecasting

In operational flood forecasting systems, the rainfall forecasts of Section 1.3 are used as starting point of the forecasting chain. Hydrological models then use this rainfall information to simulate the processes that occur in a catchment and determine the fluxes out of the catchment: discharge (our main focus for flood forecasting), evapotranspiration and, in some cases, infiltration to deeper groundwater. There are nowadays many hydrological models, with concepts ranging from conceptual and lumped to fully-distributed physics-based models, and from top-down modelling approaches to bottom-up data-driven modelling approaches (Beven, 1989; Blöschl & Sivapalan, 1995; Sivapalan et al., 2003). These hydrological modelling concepts vary in their level of complexity and have different skill for specific regions. All of them, however, have substantial uncertainties in their initial conditions, parameter value estimations or even conceptualisation of the physical processes in a catchment (Beven, 1993; Melsen et al., 2016; Clark et al., 2017). And that, on top of the mentioned rainfall forecast uncertainties (Section 1.3) and other forcing uncertainties. Catchments can act as low-pass filters for the aforementioned uncertainties, which can filter out (high-frequency) systematic random errors, but systematic errors persist or can even be enhanced in catchments and their hydrological models (due to non-linear processes; Brauer et al., 2016). Hence, the rainfall forecasting chain, which is already uncertain by itself (Section 1.3), gets an additional level of uncertainty due to the hydrological modelling errors and uncertainties when we move to discharge forecasts in a flood early warning system. Data assimilation techniques are generally used to steer the forecasting system into the right direction, if necessary (see also Section 1.3.1). Thus, just like the NWP modelling chain described in Section 1.3.1, the hydrological modelling chain also consists of an initialisation, simulation and assimilation step.

Despite the need for more accurate rainfall forecasts in space and time, as described in Section 1.1, radar rainfall data are still rarely used in operational flood forecasting systems, and therefore, neither are radar-based forecasting techniques such as nowcasting. Water managers are hesitant to use this information and for a reason: (real-time) radar data come with substantial systematic biases, caused by sources of error related to the radar reflectivity measurements,

errors in the conversion from radar reflectivity to QPE and spatio-temporal sampling errors (Austin, 1987; Joss & Lee, 1995; Gabella et al., 2000; Sharif et al., 2002; Uijlenhoet & Berne, 2008; Zawadzki, 2018; Ochoa-Rodriguez et al., 2019). Alternative rainfall data is also an option for nowcasting, and has already been applied with satellite data (Hill et al., 2020; Kumar et al., 2020), but the use of other, opportunistic, sensors, such as rainfall estimates from commercial microwave links (e.g. Messer et al., 2006; Leijnse et al., 2007; Zinevich et al., 2009; Chwala et al., 2012; Rayitsfeld et al., 2012; Overeem et al., 2013; Doumounia et al., 2014) and personal weather stations (Vos et al., 2019) is still unexplored.

Another reason for the limited use of radar rainfall nowcasting techniques in (flash) flood forecasting systems is the early stage of development of this field for flood forecasting. The strengths and weaknesses of the method are, therefore, largely unknown to water managers. Hence, although radar rainfall nowcasting certainly has potential for short-term forecasting, as described in Section 1.3, the skill of radar rainfall nowcasting for (flash) flood forecasting is still largely uncharted territory.

Previous studies, so far, have predominantly focused on the development of nowcasting algorithms in combination with a quantification of the rainfall forecast quality and errors, often based on only a relatively small number of rainfall events (in the order of 2–15 events, see Germann & Zawadzki, 2002, 2004; Germann et al., 2006; Turner et al., 2004; Lin et al., 2005; Foresti et al., 2016). This practically means that the extent to which the skill decreases with lead time and the transition point between nowcasting and NWP in Figure 1.3 is not well defined for rainfall forecasting at the catchment scale, notably with respect to its usage in (flash) flood early warning systems for quickly responding catchments, polders and urban areas. This makes it currently impossible to draw any firm conclusion on the added value of a (blended) short-term rainfall forecasting system (Section 1.3.3) for flood early warning. Hence, this calls for a more systematic assessment of (radar) rainfall nowcasting for flood early warning and to identify possible improvements. Together, they can serve as a basis for advice to water managers about strengths and weaknesses of the method.

1.5 | Positioning of this thesis and research questions

Summarizing Section 1.4, it is no wonder that, for example in the Netherlands, water authorities are not using nowcasting-based rainfall forecasting in their operational flood forecasting systems, due to (1) the considerable errors in the radar QPE product(s) and the subsequent discharge forecasting chain, (2) the unknowns about the rainfall forecasting skill of such a system at the catchment scale and (3) the unknowns about the strengths and weaknesses of such a system for (flash) flood early warning. Therefore, the aim of this thesis is *to identify if and how operational flood forecasting can be improved with (radar) rainfall nowcasting-based techniques*. The main focus area of this thesis is the Netherlands and Belgium, with a focus on lowland catchments. These lowland catchments generally have shallow groundwater tables, are heavily managed (especially the Dutch polder systems) and, therefore, respond quickly to short-lasting (intense) rainfall events, potentially causing local (flash) flooding or inundations of polders. An extensive part of both countries consists of flat terrain, which is, together with the temperate maritime climate in this region, theoretically a favourable location for radar-based forecasting techniques.

To meet the objective, this thesis focuses on a large part of the short-term forecasting chain,

from operational rainfall data to short-term discharge forecasts, and tries to find answers to the questions within the following themes:

Radar rainfall bias reduction

- How can real-time radar rainfall bias adjustments be improved?

Short-term rainfall forecasting

- To what extent does the rainfall nowcasting skill at the catchment scale depend on the factors event type, event duration, seasonality, catchment location and catchment size?
- What is the added value of rainfall nowcasting techniques that statistically take into account rainfall field development, compared to persistence-based techniques?
- How can alternative rainfall information sources be used for rainfall nowcasting?
- To what extent can blending between nowcasts and NWP extend the skillful lead time of rainfall forecast?

Short-term flood forecasting

- What are the strengths and limitations of rainfall nowcasting for discharge forecasting and how does this relate to the rainfall forecasting skill of the previous theme?
- What is the added value of rainfall nowcasting techniques that statistically take into account rainfall field development for (peak) discharge forecasting, compared to persistence-based techniques?

Each of the chapters in this thesis addresses one or multiple of the listed questions. The outline of this thesis is described in the next section.

1.6 | Thesis outline

Chapter 2 introduces the study area (quickly responding and lowland catchment in the Netherlands and Belgium), the studied catchments, along with the radar data and NWP forecasts provided by the meteorological organisations of both countries. In addition, this chapter describes the nowcasting and verification methods that have been used throughout the other chapters of this thesis.

The nowcasting-based forecasting chain starts with the quality of the radar QPE product, as systematic biases in the rainfall volume highly impact the discharge simulations for a catchment. As these products come with considerable errors, a well-performing bias-adjustment method is required. This is possible, but in an operational setting generally limited by the number of available rain gauges. In Chapter 3, a simple and real-time applicable radar QPE adjustment method is introduced, which can make it more feasible to use bias-adjusted radar QPE in an operational setting.

In order to draw statistically meaningful conclusions about the rainfall forecasting skill with nowcasting in the study area, Chapter 4 introduces a large-sample analysis to quantify the skill of radar rainfall nowcasting algorithms for the short-term predictability of rainfall for twelve catchments in the Netherlands. A particular focus is on the dependence of the forecast skill and forecast uncertainties on the factors seasonality, catchment location, catchment size, event type and duration.

For the possible implementation of nowcasting-based (blended) rainfall forecasting for flood early warning, we generally assume the use of radar QPE. Unfortunately, not all continents are covered by weather radar observations (Saltikoff et al., 2019; WMO, 2020), which means that the radar-based results and implementations described in this thesis are limited to some regions in the world, for example the study area of this thesis. Chapter 5 explores the opportunities and limitations of a promising alternative rainfall information source that can be used for rainfall nowcasting: signal level data from commercial microwave links in cellular telecommunication networks, which has a coverage of roughly four million links worldwide (Ericsson, 2016).

Besides alternative rainfall information sources, we can also take advantage of blending (ensemble) nowcasts and present NWP forecasts. Therefore, Chapter 6 describes the implementation of the well-known STEPS blending approach (Bowler et al., 2006; Seed et al., 2013) in the open-source pysteps framework (Pulkkinen et al., 2019), and highlights added functionalities to this approach. We hypothesize that such a blended system can provide better rainfall forecasts than nowcasting or NWP alone, and test this both at the country level and at the catchment scale.

Chapter 7 builds on Chapter 4 by using the rainfall forecasts from the large sample of events as input for the hydrological models used in the operational systems of the involved Dutch water authorities of these 12 catchments. In this way, the aim of this chapter is to quantify the skill and limitations of radar rainfall nowcasting for discharge forecasting in small and medium-sized (relatively quickly responding) catchments.

The concluding Chapter 8 provides a synthesis of the main findings of this thesis and it provides an outlook on future opportunities and pathways of the nowcasting-based forecasting field for flood early warning for scientists and practitioners.



2

Description of data, study areas and methods

2.1 | Introduction

SEVERAL rainfall products, catchments, nowcasting methods, hydrological models and verification metrics are used multiple times in Chapters 3–7. Therefore, I bundle and give a more extensive description of these aspects in this chapter. Section 2.2 describes the operational and reference rainfall products, consisting of rain gauge observations, radar observations, NWP forecasts and rainfall estimates from commercial microwave link data. The radar rainfall products are described for both the Dutch and Belgian radar composites, with a particular focus on the pre- and post-processing correction steps that have been applied. In Chapters 3–6, the focus in the analyses is on multiple spatial scales: from the 1×1 km² grid cell size, to the catchment scale and the entire radar domain (and country) scale. Chapter 7 only focuses on the catchment scale. The studied catchments in this thesis comprise of twelve Dutch lowland catchments and four catchments that are partly or completely in Belgium (described in Section 2.3). Sections 2.4–2.6 introduce the hydrological models, nowcasting methods and verification metrics that are used throughout this thesis. Specific applications of these methods are further described in the subsequent chapters.

2.2 | Rainfall products

In this section, the Dutch and Belgian rainfall products and their processing steps are described. All products are listed in Table 2.1 and subsequent sections will describe the specifics and processing steps of these datasets.

2.2.1 | The Netherlands

2.2.1.1 | Rain gauges

The Royal Netherlands Meteorological Institute (KNMI) operates a relatively dense rain gauge network, with approximately one station per 100 km² (Figure 2.1). Despite this density of the network, only 32 out of 351 rain gauges operate automatically every 10 min (note that the number of gauges varies slightly over time). The remaining 319 manual rain gauges report with a latency of one to several days. Hence, only the automatic rain gauges (approximately one station per 1,300 km²) have a sufficiently high temporal frequency to be used for real-time applications. Such applications can be, for instance, radar rainfall adjustments in an operational setting when we need to make full use of the update frequency of state-of-the-art radar rainfall products (e.g. for nowcasting on a 5-min time step).

In addition, the rain gauge density, both for the automatic and manual rain gauges, varies in space. Table 2.2 indicates the number of automatic and manual gauges that are located within the catchment boundaries of the twelve lowland catchments of Section 2.3.1. This number clearly varies as a result of catchment size and, to a smaller extent, as a result of the spatial heterogeneity of the rain gauge distribution over the Netherlands.

2.2.1.2 | Radar product

KNMI operates two C-band weather radars (Figure 2.2). Between September 2016 and January 2017, both radars were replaced by dual-polarization radars (the radars were single-polarized before 2016). During this process, the radar in De Bilt (Figure 2.2) was replaced by a new one in Herwijnen (Figure 2.2). The radar renewals and relocation have had a limited impact on the

Table 2.1 | Overview of the rainfall products in this thesis. NL indicates the Netherlands and BE indicates Belgium.

Dataset	Abbreviation	Spatio-temporal resolution (km ²)	Temporal resolution (min)	Real-time available	Country
Rain gauges	G	1 per 100 km ² , ^a	60 ^b	No ^a	NL
Unadjusted radar QPE	R_U	1	5	Yes	NL
MFB-adjusted radar QPE	R_{MFB}	1	5	Yes	NL/BE
CARROTS-corrected radar QPE ^c	R_C	1	5	Yes	NL
Gauge-adjusted radar QPE	R_A	1	5	No	NL
Numerical weather prediction QPF	NWP	1.3	5	Yes	BE
Commercial microwave link QPE	CML	1	15	No ^d	NL

^a1 per 1,300 km² when only automatic gauges are considered, which are available in real time.

^bFor the period 2008–2018; publicly available on 10-min temporal resolution nowadays.

^cWill be introduced in Chapter 3.

^dIn principle possible to obtain in real time.

QPE product, mainly because the operational products are not yet (fully) using the additional information from the dual-polarization (Beekhuis & Holleman, 2008; Beekhuis & Mathijssen, 2018).

In the data processing, the radar product is Doppler filtered for ground clutter. This resulting product is then used to construct horizontal cross-sections at a nearly constant altitude of 1,500 m, so-called pseudo-constant plan position indicators (pseudo-CAPPIs). Subsequently, range-weighted compositing is used to combine the reflectivities from both radars into one product (Overeem et al., 2009b). Since 2013, non-meteorological echoes are also removed from the radar product as an additional step with a cloud-mask obtained from satellite imagery. As a final step, rainfall rates (the QPE) are estimated with a fixed $Z - R$ relationship (Marshall et al., 1955):

$$Z_h = 200R^{1.6}. \quad (2.1)$$

In this equation, Z_h is the reflectivity at horizontal polarization (mm⁶ m⁻³, but generally given in dBZ, according to $10 \times \log_{10}[Z_h]$) and R is the rainfall rate (mm h⁻¹). The final product is called the unadjusted radar QPE (R_U) in this chapter.

KNMI also provides adjusted radar rainfall products, based on the aforementioned product, but adjusted with quality controlled observations from both 32 automatic hourly and 319 manual daily rain gauges (Overeem et al., 2009a,b, 2011). The adjustment methods to construct this product are described in Section 2.2.1.3 and 2.2.1.4. This adjusted radar rainfall product is considered the most accurate reference rainfall product in the Netherlands. It is not available in real time, but generally becomes available within one month.

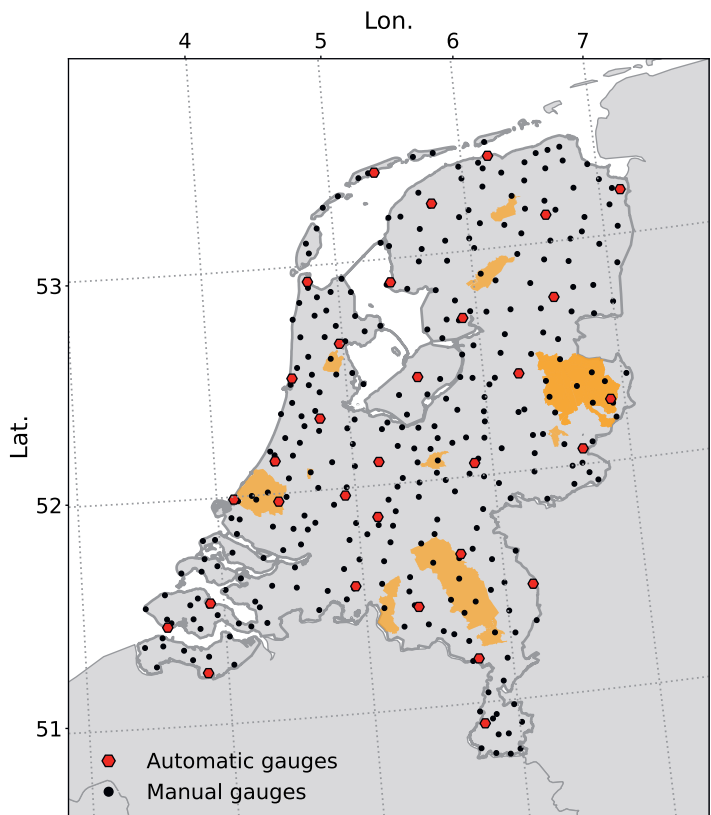


Figure 2.1 | Overview of the locations of the 32 automatic and 319 manual Dutch rain gauges currently operated by KNMI. The orange areas in the background are the twelve lowland catchments, described in Section 2.3.1. Note that the number of rain gauges changed slightly from 2009 until present.

Table 2.2 | The number of automatic and manual rain gauges located in the catchments of Section 2.3.1. Gouwepolder is part of Rijnland and used for discharge simulations in Chapters 3 and 4. The

Name	Size (km ²)	Indicated as in Chapter 4	No. of gauges	
			automatic	manual
Hupsel Brook	6.5	12	1	1
Grote Waterleiding	40	11	0	0
Luntersebeek	63	10	0	1
Beemster	71	9	0	1
Dwarsdiep	83	8	0	0
Roggelsebeek	88	7	0	1
Rijnland (Gouwepolder)	89 (10)	6	0	1
Linde	150	5	0	1
Reusel	176	4	0	1
Delfland	379	3	1	4
Aa	836	2	0	5
Regge	957	1	1	8

2.2.1.3 | Mean field bias correction

The radar rainfall product of Section 2.2.1.2 can be corrected in real time for systematic biases with a mean field bias (MFB) adjustment method and this method is subsequently also used as first step in the construction of the reference rainfall product (see Section 2.2.1.4). The MFB adjustment method is the adjustment technique that is currently operationally used by KNMI in the Netherlands. We will elaborate on the use, strengths and pitfalls of this method in Chapter 3 of this thesis. The MFB adjustment method provides a spatially uniform multiplicative adjustment factor that is applied to the operational radar rainfall product described in Section 2.2.1.2 (from here onwards, we will refer to this product as the unadjusted QPE product: R_U). The MFB-adjustment factor (F_{MFB}) is calculated as (Holleman, 2007; Overeem et al., 2009b):

$$F_{\text{MFB}} = \frac{\sum_{n=1}^N G(i_n, j_n)}{\sum_{n=1}^N R_U(i_n, j_n)}, \quad (2.2)$$

with $G(i_n, j_n)$ the hourly rainfall sum for gauge n at location (i_n, j_n) and $R_U(i_n, j_n)$ the unadjusted hourly rainfall sum for the corresponding radar grid cell. The calculation of F_{MFB} only takes place when both the rainfall sum of all rain gauges together and the hourly sum of all corresponding radar grid cells is at least 1.0 mm. In all other cases, F_{MFB} equals 1.0.

The MFB-adjustment factors are determined from the 1-h accumulations of both R_U and the 32 automatic rain gauges, as only the automatic gauges are operationally available every hour (Holleman, 2007; Overeem et al., 2009b).

2.2.1.4 | Spatial adjustments for the reference product

The adjustment procedure to derive the gauge-adjusted radar rainfall product (the reference product, from here onwards referred to as R_A) consists of three steps: (1) mean field bias correction (see Section 2.2.1.3), (2) derivation of a daily spatial adjustment factor per grid cell, and (3) spatial adjustment of the hourly or higher frequency MFB-adjusted rainfall fields (step 1) using the spatial adjustment from step 2.

A spatial adjustment factor (step 2) is derived per grid cell as follows (for a more elaborate description, see Section 3 in Overeem et al., 2009a,b):

$$F_S(i, j) = \frac{\sum_{n=1}^N w_n(i, j) \cdot G(i_n, j_n)}{\sum_{n=1}^N w_n(i, j) \cdot R_U(i_n, j_n)}, \quad (2.3)$$

with N the number of radar-gauge pairs, $G(i_n, j_n)$ the daily rainfall sum for manual rain gauge n at location (i_n, j_n) and $R_U(i_n, j_n)$ the unadjusted daily rainfall sum for the corresponding radar grid cell. $w_n(i, j)$ is a weight for gauge n , based on the following function:

$$w_n(i, j) = e^{-\frac{d_n^2(i, j)}{\sigma^2}}. \quad (2.4)$$

Here, $d_n^2(i, j)$ is the squared distance between gauge n and the grid cell for which the factor is derived. σ determines the smoothness of the adjustment factor field. It was set to 12 km by Overeem et al. (2009a,b), based on the average gauge spacing in the Netherlands.

Finally, to spatially adjust the hourly MFB-adjusted rainfall fields (step 3), two more steps are followed. First, the hourly MFB-adjusted rainfall fields (see Section 2.2.1.3 for the MFB-adjustment method) are accumulated to daily sums. For each grid cell, a new adjustment field is then determined:

$$F_{MFBs}(i, j) = \frac{R_S(i, j)}{R_{MFB}(i, j)}, \quad (2.5)$$

with $R_S(i, j)$ the spatially-adjusted daily sum for grid cell (i, j) obtained using Eq. 2.3, and $R_{MFB}(i, j)$ the MFB-adjusted daily sum for grid cell (i, j) . Second, the 1-h or higher frequency MFB-adjusted rainfall fields are multiplied with adjustment factor $F_{MFBs}(i, j)$.

2.2.1.5 | Opportunistic rainfall estimates from commercial microwave link data

Signal level data from commercial microwave links (CMLs) can also be used for rainfall estimations. CMLs are near-ground radio connections used in cellular telecommunication networks. As these links operate at frequencies where raindrops significantly absorb and scatter radio waves (Hogg, 1968; Atlas & Ulbrich, 1977; Olsen et al., 1978), rainfall attenuates the signals between the transmitting and receiving CML antennas. Although this is a nuisance from the telecommunication perspective, rain-induced attenuation can be used to estimate path-averaged rainfall intensities (e.g. Messer et al., 2006; Leijnse et al., 2007; Zinevich et al., 2009; Overeem et al., 2011; Chwala et al., 2012; Rayitsfeld et al., 2012; Doumounia et al., 2014; Gosset et al., 2015; Uijlenhoet et al., 2018; Chwala & Kunstmann, 2019).

Data from 1,751 CMLs covering the Netherlands (for an overview of these links, see Figure 5.2a) were provided by T-Mobile NL. A CML is the link along one path, often having two sub-links for communication in both directions. Minimum and maximum received powers over 15-min intervals were obtained from on average 2,400 sub-links with a precision of 1 dB, based on 10-Hz sampling. The links measured 10–30 m above the ground, used microwave frequencies ranging from ~13 to 40 GHz (majority 37–40 GHz), and had an average length of 3.1 km. The transmitted signal levels, which were not available, were constant for each link.

The CML QPE is the same as the validation dataset in Overeem et al. (2013), who retrieved rainfall maps with RAINLINK on a 1-km² spatial and 15-min temporal resolution for twelve days from June, August and September 2011. RAINLINK parameters were based on an independent calibration dataset with events in June and July 2009 and 2011 (Overeem et al., 2011, 2013, 2016a). Overeem et al. (2013) and Rios Gaona et al. (2015) provide more information on the characteristics and QPE performance of the 12-day validation dataset.

First, RAINLINK performs several corrections to the data and determines the reference attenuation belonging to dry weather (Overeem et al., 2011, 2013, 2016a). Subsequently, it computes minimum and maximum specific attenuation (dB km⁻¹) from the decrease in received signals with respect to the reference, correcting for wet antennas. Rainfall intensity is estimated from specific attenuation using a power-law relationship (Atlas & Ulbrich, 1977; Olsen et al., 1978),



Figure 2.2 | Map of the Netherlands with the twelve lowland catchments (green areas) and three radars (red triangles). Gouwepolder (see Tables 2.2 and 2.3) is part of Rijmland and was used for the discharge simulations in Chapters 3 and 7. The domain of the radar composite is indicated with the large circle, with grey areas indicating areas outside this domain. Catchment sizes and information are provided in Tables 2.2 and 2.3.

with coefficients derived from climatological rain drop-size distributions and computed electromagnetic scattering by raindrops (Leijnse et al., 2008). Fifteen-minute average rainfall intensities are then computed using a weighted average of minimum and maximum intensities. Finally, path-averaged rainfall intensities are spatially interpolated to obtain rainfall maps using ordinary kriging with a variogram model based on rain gauge data (for parameter values, see Van de Beek et al., 2012; Overeem et al., 2016a).

2.2.2 | Belgium

2.2.2.1 | Radar data

The Royal Meteorological Institute of Belgium (RMI) provides radar rainfall estimates from a composite of five C-band weather radars (Jabbeke, Helchteren, Avesnois, Wideumont and Neuheilenbach, see Figure 2.3). Two of the radars are dual-polarized (Helchteren and Jabbeke). RMI processes the radar data in four steps to move from reflectivity measurements per radar to domain-wide rainfall estimates:

1. Non-meteorological echoes are removed with Doppler filtering and clutter is identified and removed based on the vertical profile of reflectivity (VPR), image texture, a satellite cloud mask, and information from the dual polarization of the two dual-polarized radars.
2. Per radar, the reflectivity at the ground is estimated with an averaged VPR to extrapolate non-convective rainfall to the ground level. Missing data is interpolated and convective precipitation is identified, which is followed by a conversion from reflectivity into rain rate using the Marshall-Palmer relationship (Marshall et al., 1955), where stratiform rainfall, convective rainfall and hail are treated differently (similar to Wagner et al., 2012):

$$R_{i,j}(t) = \begin{cases} 10^{\frac{\text{dBZ}_{i,j}(t)-23}{16}} & \text{if dBZ} \leq 44 \\ 10^{\frac{\text{dBZ}_{i,j}(t)-19}{19}} & \text{otherwise} \end{cases} \quad (2.6)$$

In this approach, RMI uses a maximum reflectivity of 55 dBZ (approximately 88 mm h⁻¹) to deal with hail and a minimum reflectivity of 7 dBZ as a rain-no rain threshold to filter out artifacts.

3. Once the rainfall rates are estimated per radar, they are adjusted with an MFB factor that is based on the sums of the radar rainfall and rain gauge measurements (from 152 gauges) over the past hour (see Section 2.2.1.3 for the MFB approach).
4. The resulting rainfall rates of all five radars are combined with so-called range-weighted compositing, which uses weights based on the distance of the grid cell to the radars. During the months with predominantly convective precipitation (May through August), this range-weighted combination takes place with only the three closest radars to the grid cell. Finally, the composited rain rates are MFB adjusted (the whole domain at once, which is different from step 3), using the same hourly rain gauge sums as in step 3.

More information about the Belgian QPE product is provided by Goudenhoofdt & Delobbe (2016).

2.2.2.2 | NWP rainfall forecasts

RMI also provides high-frequency NWP rainfall forecasts with the ALARO configuration of the NWP system developed by ACCORD (A Consortium for Convection-scale Modeling Research and Development), formerly known as the ALADIN system (Bubnová et al., 1995; Termonia et al., 2018). The physics parameterisations of the ALARO model include the multi-scale precipitation and cloud scheme ‘Modular Multiscale Microphysics and Transport’ (3MT; Gerard et al., 2009). The ALARO model is run operationally at 1.3 km resolution 4 times a day on a 548 × 548 domain covering Belgium, the Netherlands and Luxembourg (Benelux), with a lead time of up to 48 h. The forecast is available at the latest 4 hours after the analysis time. As the ALARO NWP model is maintained and co-developed at RMI, the code was adapted to produce high-frequency precipitation output at every time step, which was then accumulated to obtain a high operational temporal resolution of 5 min. This high-frequency precipitation output is produced over a smaller sub-domain which covers the Belgian radar composite and the Benelux. The high temporal frequency of this product is beneficial for blending procedures on a high temporal resolution and is therefore used in Chapter 6 of this thesis.

Table 2.3 | Characteristics of the catchments in Figure 2.2. The water balance terms are determined over the period 2008 – 2018, with P precipitation, ET_{pot} the potential evapotranspiration, Q the observed discharge, f_{xG} the ground-water 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 (this 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 h^{-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 (described in Section 2.4). Note that Gouwepolder is part of the Rijnland area (Figure 2.2)

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

2.3 | Study areas

2.3.1 | A variety of twelve Dutch lowland catchments

The Netherlands has a maritime temperate climate, with precipitation that predominantly falls as rainfall (even in winter). Therefore, rainfall is generally mentioned in this thesis, instead of precipitation. The study area in Chapters 3, 4 and 7 consists of 12 lowland catchments in the Netherlands (Figure 2.2). These catchments are a combination of polders and (partially) freely-draining catchments. We selected these 12 catchments based on their location (spread over the country) and in close collaboration with the water authorities that were involved in the studies of Chapters 3, 4 and 7. 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 and management style varies much between these catchments (Table 2.3). This is a result of intensive water level regulation in the Netherlands, especially in the Dutch 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 catchment boundaries can be substantial, in particular upward seepage in low-lying polders.

Two typical regulated polders in the study area are Beemster and Gouwepolder (Table 2.3). The Beemster is a deep polder in the province of Noord-Holland, which is mostly covered by grass fields (and which is famous for its cheese from the grass-fed cows). It has a rather constant upward seepage of brackish water that is constantly flushed during the summer half year by pumping water into and out of the polder area. The Gouwepolder, in the province of Zuid-Holland, is a small polder that is mostly used for arboriculture. Because of its land

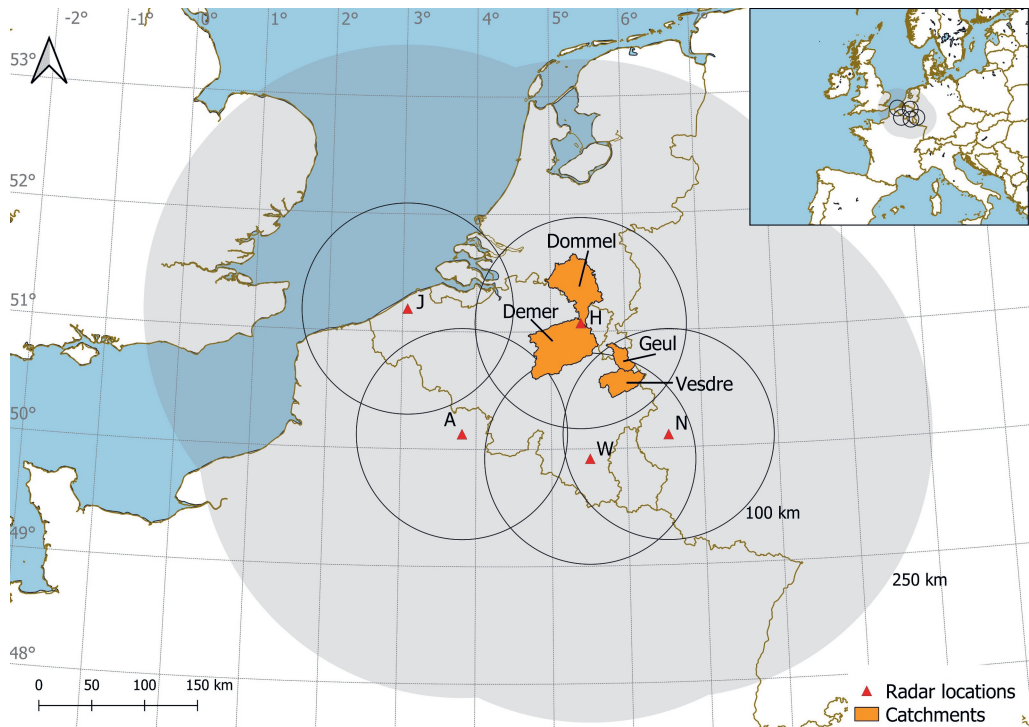


Figure 2.3 | Map of the Belgian study area with the four Belgian and Dutch catchments indicated in orange and the five radar locations as red triangles ('J' is Jabbeke, 'H' is Helchteren, 'A' is Avesnois, 'W' is Wideumont and 'N' is Neuheilenbach). The circles indicate a distance of 100 km from the radar and the grey shaded area indicates the maximum range of the radars based on a 250 km distance from the nearest radar.

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 downward seepage of water towards 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 behaviour.

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 catchments 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 most of the southwest of the catchment (located in the Westland region) consists of greenhouses. The presence of these greenhouses leads to a fast response of runoff to rainfall (Table 2.3) and therefore the region is heavily managed to prevent flooding.

2.3.2 | Four catchments in Belgium

The study areas to test and evaluate the blended forecasts that will be introduced in Chapter 6, are located in Belgium and the south of the Netherlands (Figure 2.3). The reason to focus on different catchments in this chapter than the aforementioned twelve lowland catchments is twofold: (1) the blending study is conducted with the Belgian radar and NWP products, and these catchments are located more centrally in the domain of these products, and (2) with the blending approach we aim to extend the skillful lead time of the short-term rainfall forecasts, which can also be beneficial for somewhat larger catchments, which have a slower response time. Four catchments are selected (indicated in orange in Figure 2.3):

- Vesdre (685 km²), which is a steep and quickly responding catchment in the Belgian Ardennes. It is a tributary of the Ourthe catchment, which flows into the River Meuse near the Belgian city Liège. The hydrological response of the Vesdre (with a mean discharge of 525 mm y⁻¹) is furthermore heavily influenced by two reservoirs in the area (Eupen and La Gileppe; Bruwier et al., 2015).
- Demer (2268 km²), which is a tributary of the River Dijle and eventually flows into the River Scheldt. The mean discharge of the Demer is 193 mm y⁻¹.
- Geul (323 km²), which has its origins in the foothills of the Belgian Ardennes and flows via the hills of southern Limburg in the Netherlands to the River Meuse. The river is, just like the Vesdre, quickly responding. It has a mean discharge of 391 mm y⁻¹.
- Dommel (1691 km²), which is a relatively flat catchment with predominantly sandy (permeable) soils and which has its origin in the north of Belgium and flows through the south of the Netherlands to the River Meuse near the city of 's-Hertogenbosch. The mean discharge of the Dommel is 261 mm y⁻¹.

The hydrological response of these catchments is outside the scope of this study, the focus in Chapter 6 is mainly on the rainfall forecasts, but is subject of further study.

2.4 | Hydrological models

The employed hydrological models in this thesis (used in Chapters 3 and 7) are the models used operationally or for research purposes by the involved Dutch water authorities of the twelve Dutch lowland catchments from Section 2.3.1 (Table 2.3). The hydrological models, SOBEK-RR(-CF) and WALRUS, are tailored to lowland catchment systems. SOBEK RR(-CF) (Stelling & Duinmeijer, 2003; Stelling & Verwey, 2006; Prinsen et al., 2010) is a semi-distributed model that can couple bucket-style rainfall-runoff modules (RR) 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. Four of the twelve catchments (Beemster, Gouwpolder, Linde and Delfland) have a SOBEK model. They have the following number of sub-catchments: 1 for Beemster, 7 for Gouwpolder (areas ranging from 4.5 to 668 ha), 23 for Linde (areas ranging from 23 to 7230 ha) and 25 for Delfland (areas ranging from 1.2 to 2112 ha). Radar rainfall QPE and QPF were spatially averaged over the sub-catchments as a pre-processing step prior to model simulation for these catchments.

WALRUS (Brauer et al., 2014a) is a lumped conceptual rainfall-runoff model that is tailored to lowland catchments and accounts for typical lowland processes such as the coupling of ground-water 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 prior to this thesis (Brauer et al., 2014b; Loos, 2015a,b; Gerritsen, 2019; Heuvelink et al., 2020; Sun et al., 2020). The same setups were used in this thesis (described in Table E.1 of Appendix E). The catchments Roggelsebeek and Dwarsdiep were, however, not calibrated prior to this thesis. Therefore, WALRUS was calibrated for both catchments using a calibration method, which consists 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 parameters (c_W , c_G , c_Q , c_V , c_S and c_D , see Appendix E for further information) were calibrated. The gauge-adjusted rainfall product provided by KNMI (Section 2.2.1) was used as rainfall input for calibration. The 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. This may have influenced the model calibration. The results of the calibration can be found in Appendix E (Figures E.2 and E.3), while the overall model validation for all twelve catchments is discussed later in Section 7.3.1.

In addition, the WALRUS model for the Luntersebeek catchment was recalibrated for water board Vallei and Veluwe. 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 Appendix Figure E.4. The calibration procedure for the aforementioned three catchments did not result in perfect discharge simulations, as both performance metrics based on discharge and realism of internal model fluxes were taken into account. Hence, the optimal parameter set will likely not be attained due to e.g. 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 the chapters by comparing discharge forecasts with the hydrological model simulations using the ‘observed’ radar QPE instead of discharge observations, leaving out any model related errors.

2.5 | Nowcasting methods

The section gives a brief description of the main characteristics of the nowcasting algorithms that are used in Chapters 4–7. The setup of the methods is described in more detail in the individual chapters. The algorithms were chosen because they are open source and because they allow for a comparison between different methods, namely a global optical flow method compared to a corner detection method, and purely advection-based nowcasting compared to methods that incorporate the spatial and temporal scales of rainfall for rainfall field evolution (either with or without uncertainties taken into account).

2.5.1 | Rainymotion

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

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

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

2.5.2 | Pysteps

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

1. Read radar composites and determine the motion field.
2. Use an advection method for the extrapolation of the radar images into the future. Generally, a backward semi-Lagrangian method, allowing for rotational movement, is used (Germann & Zawadzki, 2002).
3. Apply cascade decomposition, generally Fast Fourier Transform (FFT), to decompose the rainfall field into a multiplicative cascade (Seed, 2003). The levels of this cascade represent different spatial scales, which are assumed to represent different precipitation lifetimes (Venugopal et al., 1999; Seed, 2003; Germann et al., 2006).

4. Estimate the autoregressive (AR) model parameters and apply the AR model in time. Together with the cascade model used in space, this methodology handles the temporal evolution of and correlation within the precipitation structure.
5. Add stochastic perturbations to the AR-models and to the advection field as a way to take into account uncertainty in rainfall intensities and the motion field for the ensemble forecast.
6. Perform the actual extrapolation after recomposing the cascade with the iterated AR-model and the stochastic perturbations. This will give the raw nowcast ensemble.
7. Apply post-processing operations to ensure that the nowcast has the same statistical properties as the latest available observations.

Pysteps is flexible in the choice for e.g. optical flow methods, advection methods, noise methods and spatial/temporal decomposition methods. Throughout this thesis, two setups of pysteps are used: one for deterministic nowcasts and one for probabilistic nowcasts. The deterministic setup, from here-on referred to as *pysteps deterministic*, resembles the S-PROG algorithm (Seed, 2003) and has the following configuration: a Lucas-Kanade optical flow method (using the QPE from time $t - 3$ to t) (Lucas et al., 1981), a backward semi-Lagrangian advection method (Germann & Zawadzki, 2002), an AR-model of order 2, the S-PROG masking method (threshold is 0.1 mm h^{-1}), a probability matching method to match the forecast statistics with the observations based on the mean observed rainfall fields, and 8 cascade levels (instead of 6 in the original S-PROG). Pysteps deterministic follows most of the aforementioned seven steps, except for step 5 (stochastic perturbations).

The probabilistic setup follows the aforementioned 7 steps, with the following configuration: a Lucas-Kanade optical flow method using the QPE from time $t - 3$ to t (Lucas et al., 1981), a backward semi-Lagrangian advection method (Germann & Zawadzki, 2002), the STEPS nowcasting method (Bowler et al., 2006), a non-parametric noise method (Seed et al., 2013), FFT for the spatial decomposition with 8 cascade levels, an AR-model of order 2, a lead time-dependent masking method, the cumulative distribution function used as probability matching method, and 20 ensemble members. For both methods, the rainfall fields are transformed to dBR prior to nowcasting.

Pulkkinen et al. (2019) found that the optimum ensemble size for pysteps depends on the rainfall intensity threshold that is assessed. For low intensity thresholds, there was only a marginal improvement between 24 and 48 ensemble members. This indicates that for these thresholds, the chosen ensemble size of 20 is probably sufficient. However, for higher thresholds (e.g. 5 mm h^{-1}), there was even significant improvement in model performance when increasing the ensemble size to as much as 96 members (Figures 13 and 14 in Pulkkinen et al., 2019). Hence, when the nowcasting algorithm is used to forecast high-intensity rainfall events, a larger ensemble size is desirable. The downside of this choice would be that this might not be computationally feasible in an operational setting when a new set of nowcasts has to be made every 5 min, which is why an ensemble size of 20 members was chosen in most chapters of this thesis.

2.6 | Verification metrics

Throughout all chapters of this thesis, a large number of verification metrics are used. This section introduces and describes these verification metrics. Their application will be described in the specific chapters.

2.6.1 | Pearson's correlation

Pearson's correlation is a measure for the strength of the linear relationship between two variables; here, forecast and observation. Pearson's correlation coefficient (ρ) is calculated, for instance for every lead time in the forecast, as:

$$\rho = \frac{1}{N_f} \sum_{i=1}^{N_f} \frac{(F_i - \mu_F)(O_i - \mu_O)}{\sigma_F \sigma_O}, \quad (2.7)$$

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

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

2.6.2 | Root mean square error

The root mean square error is the standard deviation of the forecast errors and gives extra weight to larger outliers. For deterministic runs and per forecast event, it is formulated as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_f} (F_i - O_i)^2}{N_f}}. \quad (2.8)$$

2.6.3 | Mean absolute error

The mean absolute error (MAE) gives the average of the absolute error between forecast and observation. For deterministic runs and per forecast event, the MAE is calculated as:

$$\text{MAE} = \frac{\sum_{i=1}^{N_f} |F_i - O_i|}{N_f}. \quad (2.9)$$

2.6.4 | Ensemble verification scores

2.6.4.1 | Continuous ranked probability score

In the case of a probabilistic forecast, e.g. with pysteps probabilistic, the entire forecast distribution is available for comparison with the observations. In order to do this, the cumulative

distribution functions (cdf) of forecast and observation are used. While the cdf of the observation is a single step-function, i.e. there is only one value, the cdf of the probabilistic forecast is a curve. The area between these two cdfs is a measure for the continuous ranked probability score (CRPS), which is formulated as:

$$CRPS = \frac{1}{N_f} \sum_{i=1}^{N_f} \int_{-\infty}^{+\infty} \left(P_{F_i}(x) - P_{O_i}(x) \right)^2 dx . \quad (2.10)$$

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

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

2.6.4.2 | Brier Score

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

Simultaneously, this approach can be used to obtain an indication of the ensemble skill, as compared to a benchmark. Below the climatological frequency of exceeding a given rainfall threshold, the forecast is unable to distinguish situations with different frequencies of occurrence: the point of no resolution. Additionally, there is a point of no skill, where the probabilistic forecast is not able to predict better than a reference (e.g. the climatology) whether an event will occur or not. This is tested with the Brier Skill Score (BSS), which is based on the Brier Score (BS); (Jolliffe & Stephenson, 2012):

$$BS = \frac{1}{N_f} \sum_{i=1}^{N_f} (P_{F_i} - P_{O_i})^2 , \quad (2.11)$$

$$BSS = 1 - \frac{BS}{BS_{\text{ref}}} . \quad (2.12)$$

The BS is similar to the CRPS (Eq. 2.10), but with the difference that the CRPS is the BS integrated over all thresholds. Thus, equation 2.11 verifies whether for forecast i , a predefined threshold is exceeded by forecast P_{F_i} (given as a probability between 0 and 1) and by observation P_{O_i} (0 or 1). To put this in perspective, a reference (BS_{ref}) is used to determine the BSS. This reference can be the climatology, but also persistence, deterministic forecasts or probabilistic forecasts.

2.6.4.3 | Receiver Operating Characteristic

The Receiver Operating Characteristic (ROC) curve analyses the predictive ability of exceeding a certain threshold with probabilistic forecasts. In essence, the curve plots the Hit Rate (HR) versus the False Alarm Rate (FAR) for predefined probability thresholds. HR is calculated as:

$$HR = \frac{TP}{TP + FN} , \quad (2.13)$$

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

$$FAR = \frac{FP}{FP + TN} , \quad (2.14)$$

where FP is the number of false positives, the 'false alarms': the forecast exceeds the threshold, but the observation does not. TN is the number of true negatives: neither the forecast nor the observation exceed the threshold.

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

2.6.5 | Spatial scores

2.6.5.1 | Fractions Skill Score

The Fractions Skill Score (FSS) is a spatial verification score which uses a fractions-based Brier score (see 2.6.4.2) over successively larger cell lengths (Roberts & Lean, 2008). By increasing the length scale n (e.g. in km), the area used for verification increases, generally leading to a higher FSS value. n can increase up to $2N - 1$, with N the longest length scale in the extent. The FSS ranges from 0 to 1, with 1 corresponding to a perfect forecast. With this metric, a minimum length scale to reach a required skill can be found, which is the target upscaling resolution of the data. For a predefined threshold and one forecast, it is calculated as:

$$FSS(n) = 1 - \frac{MSE(n)}{MSE_{ref}(n)} , \quad (2.15)$$

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

$$MSE_{ref}(n) = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O_{ij}^2(n) + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F_{ij}^2(n) \right] , \quad (2.16)$$

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

$$O_{ij}^2(n) = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n I_O \left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2} \right]. \quad (2.17)$$

Here, I_O is the binary field of exceedances of a given rainfall intensity threshold for the observations. k and l are integer values ranging from 1 to length scale n , used to count the threshold exceedances for every cell within the window around cell (i, j) . The equation for $F_{ij}^2(n)$ is the same, with the only difference that not I_O , but I_F is used. See also Figure 2 in Roberts & Lean (2008) for a schematic example of the method.

Generally, the FSS value above which the forecast is considered useful, lies in between the perfect and the random forecast skill ($= 0.5 + \frac{\theta_0}{2}$). The random forecast skill, indicated with f_o , is defined as the domain average observed rainfall fraction above the threshold (Roberts & Lean, 2008; Mittermaier & Roberts, 2010).

2.6.6 | Categorical scores

2.6.6.1 | Critical success index

The Critical success index (CSI; Schaefer, 1990) gives the fraction of forecasts that correctly exceed a predefined threshold over the total number of hits, misses and false alarms in all forecasts. For a perfect forecast, CSI equals 1. CSI is a categorical score, which uses the categorical information that is also used in Equations 2.13 and 2.14. It is formulated as:

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \quad (2.18)$$

2.6.7 | Hydrological verification scores

2.6.7.1 | 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:

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{N_f} (F_i - O_i)^2}{\sum_{t=1}^{N_f} (O_i - \mu_o)^2}, \quad (2.19)$$

with N_f the total number of time steps (t) in the time series, or the number of forecasts with lead time t in the event, and μ_o is the mean observed discharge.

2.6.7.2 | Kling-Gupta efficiency

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

$$\text{KGE} = 1 - \sqrt{(\rho - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \quad (2.20)$$

with

$$\alpha = \frac{\sigma_f}{\sigma_o}, \quad (2.21)$$

$$\beta = \frac{\mu_f}{\mu_o}. \quad (2.22)$$

Here, ρ is Pearson's correlation (Equation 2.7) 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.



3

Operational radar rainfall bias correction

This chapter was originally published as:

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

SIGNIFICANT biases in real-time radar quantitative precipitation estimations (QPE) limit its use in hydro-meteorological forecasting systems. Here, we introduce CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting), a set of fixed bias reduction factors, which vary per grid cell and day of the year. The factors are based on a historical set of 10 years of 5-min radar and reference rainfall data for the Netherlands. CARROTS is both operationally available and independent of real-time rain gauge availability, and can thereby provide an alternative to current QPE adjustment practice. In addition, it can be used as benchmark for QPE algorithm development. We tested this method on the resulting rainfall estimates and discharge simulations for twelve Dutch catchments and polders. We validated the results against the operational mean field bias (MFB) adjusted rainfall estimates and a reference dataset. This reference consists of the radar QPE, that combines an hourly MFB adjustment and a daily spatial adjustment using observations from 32 automatic and 319 manual rain gauges. Only the automatic gauges of this network are available in real-time for the MFB adjustment. The resulting climatological correction factors show clear spatial and temporal patterns. Factors are higher far from the radars and higher from December through March than in other seasons, which is likely a result of sampling above the melting layer during the winter months. The MFB-adjusted QPE outperforms the CARROTS-corrected QPE when the country-average rainfall estimates are compared to the reference. However, annual rainfall sums from CARROTS are comparable to the reference and outperform the MFB-adjusted rainfall estimates for catchments far from the radars, where the MFB-adjusted QPE generally underestimates the rainfall amounts. This difference is absent for catchments closer to the radars. QPE underestimations are amplified when used in the hydrological model simulations. Discharge simulations using the QPE from CARROTS outperform those with the MFB-adjusted product for all catchments. Moreover, the proposed factor derivation method is robust. It is hardly sensitive to leaving individual years out of the historical set and to the moving window length, given window sizes of more than a week.

“Le jour vient où une seule carotte, fraîchement observée, déclenchera une révolution.”

—Paul Cézanne

3.1 | Introduction

Radar rainfall estimates are essential for hydro-meteorological forecasting systems. In these systems, the data are used to force hydrological models (e.g., Borga, 2002; Thorndahl et al., 2017), to initialize Numerical Weather Prediction models (e.g., Haase et al., 2000; Rogers et al., 2000) or as input data for rainfall nowcasting techniques (e.g., Ebert et al., 2004; Wilson et al., 2010; Foresti et al., 2016; Heuvelink et al., 2020, and Chapter 4 of this thesis). A major disadvantage of radar quantitative precipitation estimations (QPE) are the considerable biases with respect to the true rainfall, caused by three main groups of errors: (1) sources of errors related to the reflectivity measurements, (2) sources of errors in the conversion from reflectivity to rainfall rate and (3) spatio-temporal sampling errors (Austin, 1987; Joss & Lee, 1995; Creutin et al., 1997; Gabella et al., 2000; Sharif et al., 2002; Uijlenhoet & Berne, 2008; Ochoa-Rodriguez et al., 2019, and Chapter 5 of this thesis). These biases can amplify when used in hydrological models (Borga, 2002; Borga et al., 2006; Brauer et al., 2016). Hence, radar QPE requires corrections before operational use in hydro-meteorological (forecasting) models.

A large number of correction methods is already available. These methods range from corrections prior to the rainfall estimations, e.g. corrections for physical phenomena such as ground clutter, attenuation, the vertical profile of reflectivity and variations in raindrop size distribution (e.g., Joss & Pittini, 1991; Germann & Joss, 2002; Berenguer et al., 2006; Cho et al., 2006; Uijlenhoet & Berne, 2008; Kirstetter et al., 2010; Qi et al., 2013; Hazenberg et al., 2013, 2014), to statistical post-processing steps for bias removal in the radar QPE using rain gauge data. These post-processing methods either merge rain gauge and radar QPE from the same interval or base correction factors on the total precipitation in both products over a past period, such as a number of rainy days (e.g. seven days in Park et al., 2019). An often used method is the mean field bias (MFB) correction method, which determines a spatially-averaged correction factor from the ratio between rain gauge observations and the radar QPE of the superimposed grid cells at the locations of these gauges (Smith & Krajewski, 1991; Seo et al., 1999). This method, which is used operationally in the Netherlands and many other countries (Holleman, 2007; Harrison et al., 2009; Thorndahl et al., 2014; Goudenhoofdt & Delobbe, 2016), does not account for any spatial variability in the radar QPE bias, even though the bias is known to increase with increasing distance from the radar (Koistinen & Puhakka, 1981; Joss & Lee, 1995; Koistinen et al., 1999; Gabella et al., 2000; Michelson & Koistinen, 2000; Seo et al., 2000).

It is possible to account for this spatial variability with geostatistical techniques (e.g. ordinary kriging, kriging with external drift or co-kriging, Krajewski, 1987; Creutin et al., 1988; Wackernagel, 2003; Schuurmans et al., 2007; Goudenhoofdt & Delobbe, 2009; Sideris et al., 2014) or Bayesian merging methods (Todini, 2001). Although these methods substantially improve the QPE in the spatial domain, all gauge-based radar QPE adjustment methods are limited by the timely availability of sufficient, and ideally quality-controlled, rain gauge observations (for an overview of methods and their limitations, see Ochoa-Rodriguez et al., 2019). The gauge networks operated by the Royal Netherlands Meteorological Institute (KNMI) are an example of this issue. Although there is approximately one station per 100 km², only 32 out of 351 rain gauges operate automatically. The remaining 319 manual rain gauges report just once a day. Thus, only the automatic rain gauges are used for the MFB adjustment that takes place every hour in real-time (Holleman, 2007), and since recently even every five minutes.

In addition, two potential operational (forecasting) issues need to be considered when using these more advanced geostatistical and Bayesian merging methods: (1) the methods are computationally expensive, especially methods such as co-kriging and Bayesian merging that integrate radar and rain gauges (Ochoa-Rodriguez et al., 2019), and (2) when the adjustment method changes the spatial structure of the original radar rainfall fields (kriging and Bayesian methods), this may impact the continuity of the rainfall fields over time and thereby also the radar rainfall nowcasts (Ochoa-Rodriguez et al., 2013; Na & Yoo, 2018). In case the nowcasts suffer from errors due to these adjustments, this suggests that adjustment methods should be applied to the nowcasts as a post-processing step. To do this, the forecaster would need to estimate the future (bias) correction factors (a method for this using MFB adjustment is described in Seo et al., 1999) or simply assume that the latest correction factors are exemplary for the coming hours.

Hence, operational hydro-meteorological forecasting calls for a radar rainfall adjustment approach that (1) takes the spatial variability in radar QPE errors into account and (2) is available in real time so that it can be used operationally for radar-based rainfall forecasts, such as nowcasting. Here, we present CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting): a set of gridded climatological adjustment factors for every day of the year, based on a historical set of 10 years of 5-min radar and reference rainfall data for the Netherlands. When sufficient rain gauges are operationally available, which would allow for a robust application of more advanced geostatistical and Bayesian merging methods, CARROTS can serve as a benchmark for testing these and other more sophisticated adjustment techniques.

3.2 | Data and methods

3.2.1 | Radar rainfall estimates

Rainfall estimates from the gauge-adjusted ('reference') and unadjusted radar datasets with a 1-km² spatial and 5-min temporal resolution, as described in Section 2.2.1, were used for the period 2009–2018. In the remainder of this chapter, we refer to the gauge-adjusted radar QPE as R_A and to the unadjusted radar QPE as R_U . The year 2008 is actually the first year in the KNMI archive of both data sets, but it was left out of the analysis here. R_U for this year showed a significantly different behaviour than the other years, especially during the first half year in which the product rarely underestimated and frequently even overestimated the rainfall sums. The reason for this behaviour is not yet fully understood. KNMI (2009) reported that spring was exceptionally dry in the north of the country and that the months January and May were among the warmest on record. On some days with overestimations, clear bright band effects were visible in the radar mosaic, which may have contributed to the systematic differences.

3.2.2 | Bias correction factors

Figure 3.1 indicates the need for correction of the real-time available radar rainfall product. R_U systematically underestimates the true rainfall amounts, averaged for the land surface area of the Netherlands, by 55%. This bias is not uniform in space, as will be highlighted in Section 3.3, and in time with higher underestimations during winter (on average 65%) than during the other seasons (50–55 %). In the following two subsections, the operationally used MFB-adjustment method and the in this chapter proposed CARROTS method will be introduced.

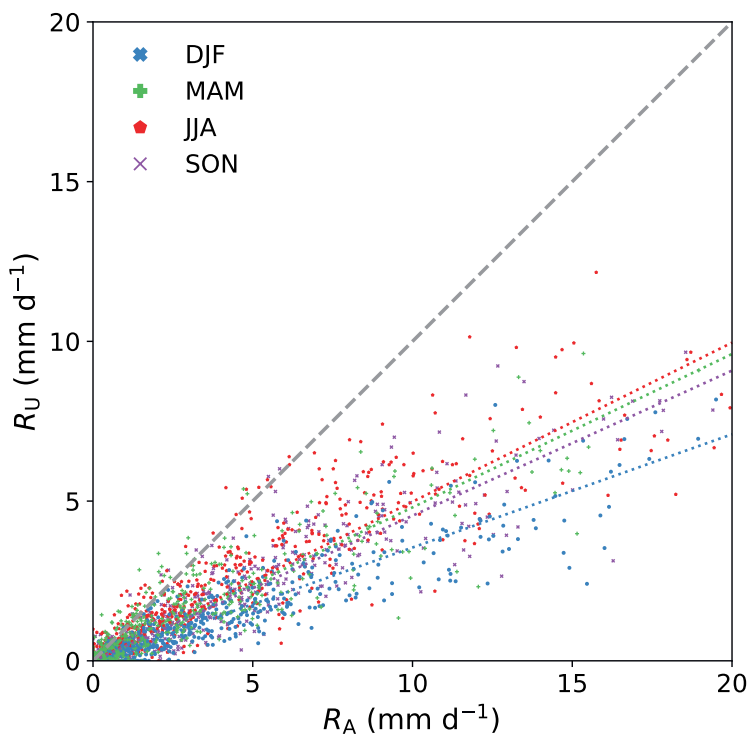


Figure 3.1 | The systematic discrepancy between the reference rainfall (R_A) and the unadjusted radar QPE (R_U). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are indicated in Table 3.1. The colored dashed lines are a linear fit, forced through the origin, per season between R_A and R_U .

Table 3.1 | Statistics of Figure 3.1. Indicated are the sample size, the slope of a linear fit between the two rainfall products (R_A and R_U ; the colored dashed lines in Figure 3.1) for all observations and the Pearson correlation coefficient. This is indicated per season (DJF is winter, MAM is spring, JJA is summer and SON is autumn) and for all seasons together (Total).

Season	Sample size	Slope	Pearson correlation	R^2
DJF	902	0.35	0.90	0.81
MAM	920	0.48	0.89	0.79
JJA	920	0.50	0.89	0.79
SON	910	0.45	0.92	0.85
Total	3652	0.45	0.89	0.79

3.2.2.1 | Mean field bias adjustment

The mean field bias (MFB) adjustment method is the operational adjustment technique in the Netherlands and it was used in this chapter for comparison with the proposed climatological bias reduction method (Section 3.2.2.2). This method provides a spatially uniform multiplicative adjustment factor that is applied to R_U . The derivation of the adjustment factor (F_{MFB}) is described in Section 2.2.1.3 of Chapter 2.

In this chapter, the MFB-adjustment factors were determined from the 1-h accumulations of both R_U and the 32 automatic rain gauges, as only the automatic gauges were operationally available every hour (Holleman, 2007; Overeem et al., 2009b). The adjustment factors at the temporal resolution of the radar QPE (5 min) were assumed to equal the 1-h adjustment factors for a given hour.

Moreover, this analysis took place with archived datasets, which were validated and consisted of quality-controlled rain gauge observations. It is noteworthy that the same quality control is absent and that missing data occurs in real-time, which can lead to deteriorating results when the MFB adjustment is applied in an operational test case.

3.2.2.2 | CARROTS method

To derive the climatological bias correction factors for the CARROTS method, both R_U and R_A were used for the years 2009–2018. The use of the reference data for this method was possible, because the CARROTS method did not require a real-time availability of the data. The bias correction factors were determined per grid cell in the radar domain according to the following three steps:

1. For every day in the period 2009–2018, all 5-min rainfall sums (both R_U and R_A) within a moving window of 31 days (the day of interest plus the fifteen days before and after it) were summed. The purpose of the moving window was to smooth the systematic day-to-day variability of the estimated rainfall in the 10-year data. Sections 3.2.4 and 3.3.4 describe a leave-on-year-out validation of the method and they describe the sensitivity of the method to the moving window size.
2. For every day of the year, the 31-day sums around that day were averaged over the ten years. Thus, the value for e.g. 16 January consisted of the average 31-day sum for the period 1 to 31 January over the ten years.
3. Finally, gridded climatological adjustment factors (F_{clim}) were calculated per day of the year as:

$$F_{\text{clim}}(i, j) = \frac{R_A(i, j)}{R_U(i, j)}, \quad (3.1)$$

with $R_A(i, j)$ the reference rainfall sum and $R_U(i, j)$ the unadjusted (operational) radar rainfall sum at grid cell (i, j) for the ten years.

3.2.2.3 | Spatial adjustments for the reference product

The adjustment procedure to derive R_A consists of three steps: (1) mean field bias correction (one adjustment factor for the whole country which varies per hour, see Section 3.2.2.1 and

Section 2.2.1.3), (2) derivation of a daily spatial adjustment factor per grid cell, and (3) spatial adjustment of the hourly or higher frequency MFB-adjusted rainfall fields (step 1) using the spatial adjustment from step 2. The full procedure to derive R_A is described in more detail in Section 2.2.1.4 of Chapter 2 and in Section 3 of Overeem et al. (2009a,b).

3.2.3 | Hydrometeorological application

Both bias adjustment methods were applied to the ten years (2009–2018) of R_U . In order to provide a hydro-meteorological testbed, both the CARROTS and MFB-adjusted QPE products (from here-on referred to as R_C and R_{MFB} , respectively) were validated against the reference rainfall. First, this was done at country level. The estimated daily rainfall sums for all grid cells within the land surface area of the Netherlands were compared to the reference in a similar way as the comparison in Figure 3.1. To subdivide these results per year and season, an additional hourly rainfall sum validation was performed as well. The results of this analysis can be found in Appendix A and the analysis was done as follows: for every rainy hour (when the sum of at least one grid cell was larger than 0.0 mm), we computed the Root Mean Square Error (RMSE) by squaring the differences between the three QPE products (R_U , R_C and R_{MFB}) on the one hand and the reference on the other, and taking the average of these squared differences over all grid cells within the land surface area of the Netherlands. Subsequently, the RMSE was averaged over all rainy hours in that season and year. Finally, the seasonal mean RMSE was divided by the average hourly rainfall rate for that season and year, resulting in the Fractional Standard Error (FSE) score. The FSE score was calculated for every season in the ten years to be able to compare the seasonal performance of the hourly rainfall estimates of R_U , R_C and R_{MFB} .

Second, the annual rainfall sums for twelve lowland catchments (a combination of catchments and polders) in the Netherlands (Section 2.3.1) were compared with the reference. In addition, R_C and R_{MFB} were used as input for the rainfall-runoff models of the twelve catchments. Most of the involved water authorities use these (lowland) rainfall-runoff models either operationally or for research purposes, often embedded in a Delft-FEWS system, which is a data-integration platform, used world-wide by many hydrological forecasting agencies and water management organizations, that brings data handling and model integration together for operational forecasting (Werner et al., 2013). Catchment characteristics, the hydrological models for the twelve catchments and their calibration are described in more detail in Section 2.4 of Chapter 2. All twelve model setups were run with a 5-min time step for the period 2009–2018.

The resulting discharge simulations were validated for the same period and 5-min timestep using the Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009), described in Section 2.6.7.2. In this chapter, the discharge simulated with R_A as input was regarded as the observation. Note that this validation method was not a leave-one-out or split-sample validation, as the full 10-year dataset was used for R_A , the CARROTS- and MFB-adjustment derivation, and shorter periods in those 10 years were used for hydrological model calibration. However, the sensitivity of the CARROTS factor was tested by leaving individual years out of the derivation period (Section 3.2.4).

3.2.4 | Sensitivity analysis

As mentioned in Section 3.2.2.2, the purpose of the 31-day moving window in the factor derivation of CARROTS was to smooth the day-to-day variability of rainfall. To test the sensitivity of the method to the employed moving window size, the adjustment factors were re-derived for a range of moving window sizes (1 day, 1 week, 2 weeks, 6 weeks and 2 months). The derived factors were then compared to the original factor in this chapter, which was based on a moving window size of 31 days, and used to derive adjusted QPE products. Subsequently, these QPE products served as input for one of the 12 catchments, namely the WALRUS model for the Aa catchment (Figure 2.2), to test the effect on the simulated discharges (see Section 3.3.4 and Figure 3.7 for the results). The Aa catchment was chosen because the unadjusted QPE product (R_U) for this catchment has one of the highest biases of the twelve studied catchments (see Section 3.3 and Figure 3.3).

Besides the moving window choice, the length of the radar rainfall archive (ten years) was finite. To test whether or not this archive length was sufficient for reaching a stable factor derivation, individual years in the ten-year archive were left out of the CARROTS method. Hence, the adjustment factors were recalculated ten times in a leave-one-year-out method, applied to R_U and used as input for the WALRUS simulations for the Aa catchment. See Section 3.3.4 and Figure 3.3 for the results.

3.3 | Results

3.3.1 | Seasonal and spatial variability

The adjustment factors from CARROTS present the spatial variability in the radar QPE errors, with generally higher adjustment factors towards the edges of the radar domain (Figure 3.2). This difference is most pronounced from December through March, with more than two times higher factors in the south and east of the country than in the central and northwestern parts (Figure 3.2a, b and l). Figure 3.2 demonstrates a clear annual cycle of the adjustment factors, with higher adjustment factors from December through March than in the other months. Figure 3.3a shows similar results for the catchment-averaged adjustment factors, with factors ranging from 2.1 for the Beemster polder to 3.2 for the Hupsel Brook catchment in January, whereas adjustment factors range from 1.3 for the Grote Waterleiding catchment to 1.6 for the Roggelsebeek catchment in June.

An explanation for these higher adjustment factors from December through March is that radar QPE often severely underestimates the rainfall amounts for stratiform systems, which regularly occur during the Dutch winter. This especially holds when the QPE is constructed from reflectivities sampled above the melting layer (Fabry et al., 1992; Kitchen & Jackson, 1993; Germann & Joss, 2002; Bellon et al., 2005; Hazenberg et al., 2013). This seems to be the case here as well. A simple first-order estimation of the 0°C isotherm level, using a constant wet adiabatic lapse rate of 5.5 K km^{-1} with ground temperature data for all rainy hours in the ten years (Figure 3.3b), indicates that the 1500 m pseudo-CAPPI is generally above the 0°C isotherm level from December through March. This coincides with the months with higher adjustment factors (Figure 3.3c) and could thus explain the winter effect on the adjustment factors. This effect is presumably even stronger further away from the radars, because the QPE product consists of samples at

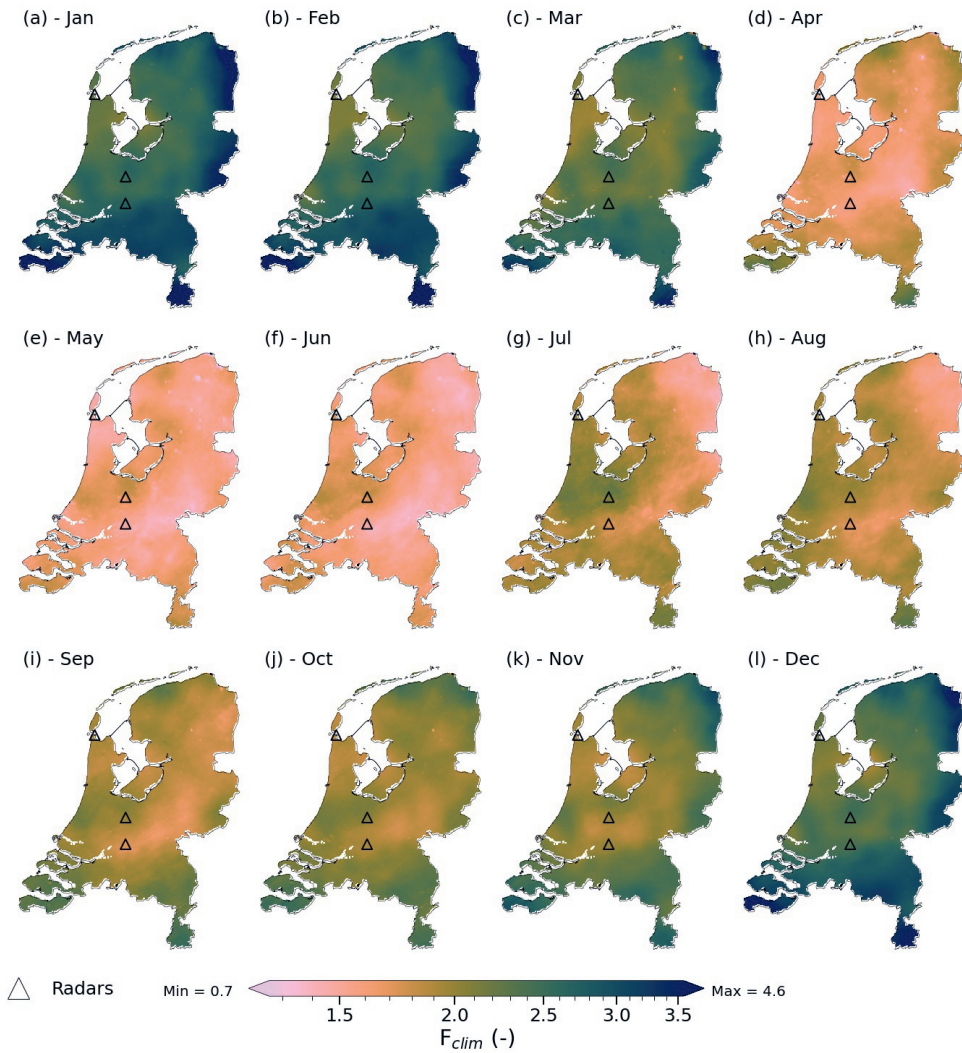


Figure 3.2 | Spatial variability of the CARROTS factors, as derived from the archived radar and reference data for the period 2009–2018. Shown are monthly averages of the daily factors.

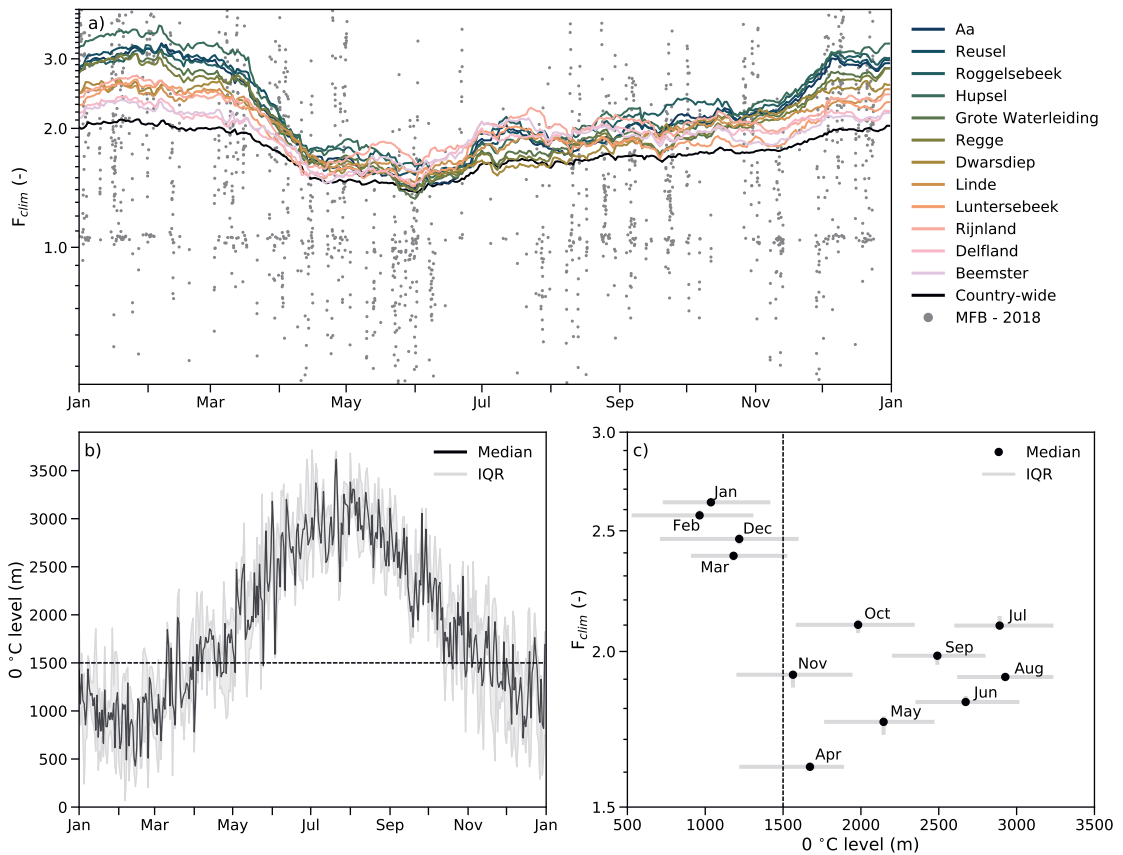


Figure 3.3 | Seasonal dependency of the CARROTS factors and comparison with the operational MFB-adjustment factor. (a) Temporal variability of the climatological daily adjustment factors for the twelve catchments (colours, catchment-averaged), the country-average (black line) and of the country-wide hourly MFB factor for the (example) year 2018 (grey dots, some also fall outside the indicated range). (b) Estimate of the height of the 0°C isotherm at KNMI station De Bilt for all rainy hours in the ten year period, based on a constant wet adiabatic lapse rate of 5.5 K km⁻¹. (c) Dependency of the monthly adjustment factor on the estimated 0°C isotherm level for KNMI station De Bilt and the superimposed grid cell of this station. Depending on the location in the radar composite, the minimum CARROTS factor can take place in a different month, but is always between April and June. Note that for this analysis, the adjustment factor was based on only the rainfall sums within that month, the effective adjustment factor for that month, which roughly coincides with the factor for the 15th of the month in the CARROTS method. The grey bars indicate the interquartile range (IQR) for that month, based on the spread in hourly 0°C isotherm level estimates (the horizontal bars) and the sensitivity to leaving out individuals years in the ten-year period for the factor derivation (vertical bars).

Table 3.2 | Statistics of Figure 3.4. Indicated are the sample size, the Pearson correlation and the slope of a linear fit between the reference and the two adjusted radar QPE products (R_{MFB} and R_{C} ; the colored dashed lines in Figure 3.4). This is indicated per season and for all seasons together (Total).

Season	Sample size	Slope		Pearson correlation		R^2	
		R_{MFB}	R_{C}	R_{MFB}	R_{C}	R_{MFB}	R_{C}
DJF	902	0.87	0.95	0.99	0.92	0.98	0.85
MAM	920	0.90	0.86	0.99	0.92	0.98	0.85
JJA	920	0.92	0.90	0.99	0.91	0.98	0.83
SON	910	0.90	0.94	0.99	0.93	0.98	0.86
Total	3652	0.90	0.92	0.99	0.92	0.98	0.85

even higher altitudes than 1500 m for locations at more than 120 km from the radars. Besides, an additional dependency of the monthly factor on the time of year that cannot be explained by temperature, seems to be present with lower adjustment factors during spring and early summer and higher factors for the subsequent period (Figure 3.3c).

3.3.2 | Evaluation of the rainfall sums

The MFB-adjusted QPE (R_{MFB}) significantly reduces the systematic bias of R_{U} (Figure 3.1), from a 55% underestimation on average for the Netherlands to 10% (Fig 3.4a and Tab 3.2). However, the remaining bias in R_{MFB} is generally caused by a systematic underestimation of the reference rainfall. The overall underestimation is less for R_{C} (8%, Figure 3.4b), but results from estimation errors associated with either under- or overestimates of the reference rainfall. The spread in Figure 3.4b is significantly wider than in Figure 3.4a, indicating that the country-wide QPE error of R_{C} is often higher than for R_{MFB} . The yearly FSE in Table A.1 clearly indicates this too, with a systematically higher FSE for R_{C} than for R_{MFB} .

An advantage of the MFB adjustment is that it corrects for the circumstances during that specific day and thus also for instances with overestimations (Figure 3.3a). On a country-wide level, this is clearly advantageous, also compared to CARROTS (Figure 3.4). The negative effect of the spatial uniformity of the factor, however, becomes apparent in Figure 3.5, which compares the annual precipitation sums of the two adjusted radar rainfall products with the reference and R_{U} for the twelve catchments. For all catchments, both adjusted products manage to significantly increase the QPE towards the reference. However, for nine out of twelve catchments, R_{C} outperforms R_{MFB} (Figure 3.5e). Exceptions are Beemster, Luntersebeek and Dwarsdiep, where the performance of both products is similar. Differences between the performance of R_{C} and R_{MFB} become most apparent for catchments that are located closer to the edges of the radar domain. For instance, R_{MFB} for the Aa and Regge catchments, which are located in the far south and east of the country, still underestimates the annual reference rainfall sums with on average 20% for the Aa (mean annual R_{MFB} is 610 mm and mean annual R_{A} = 761 mm) and 13% for the Regge (mean annual R_{MFB} is 673 mm and mean annual R_{A} = 776 mm), while this is on average only 5% (both under- and overestimations occur) for R_{C} (Figure 3.5b and c).

The MFB-adjusted QPE performs better for the Beemster polder, Dwarsdiep polder (Figure 3.5d) and Luntersebeek catchment (Figure 3.5a) due to their location in the radar mosaic. The Luntersebeek catchment (central Netherlands, Figure 2.2) is located closer to both radars. There, R_{MFB} generally performs better and sometimes even overestimates the true rainfall, which is

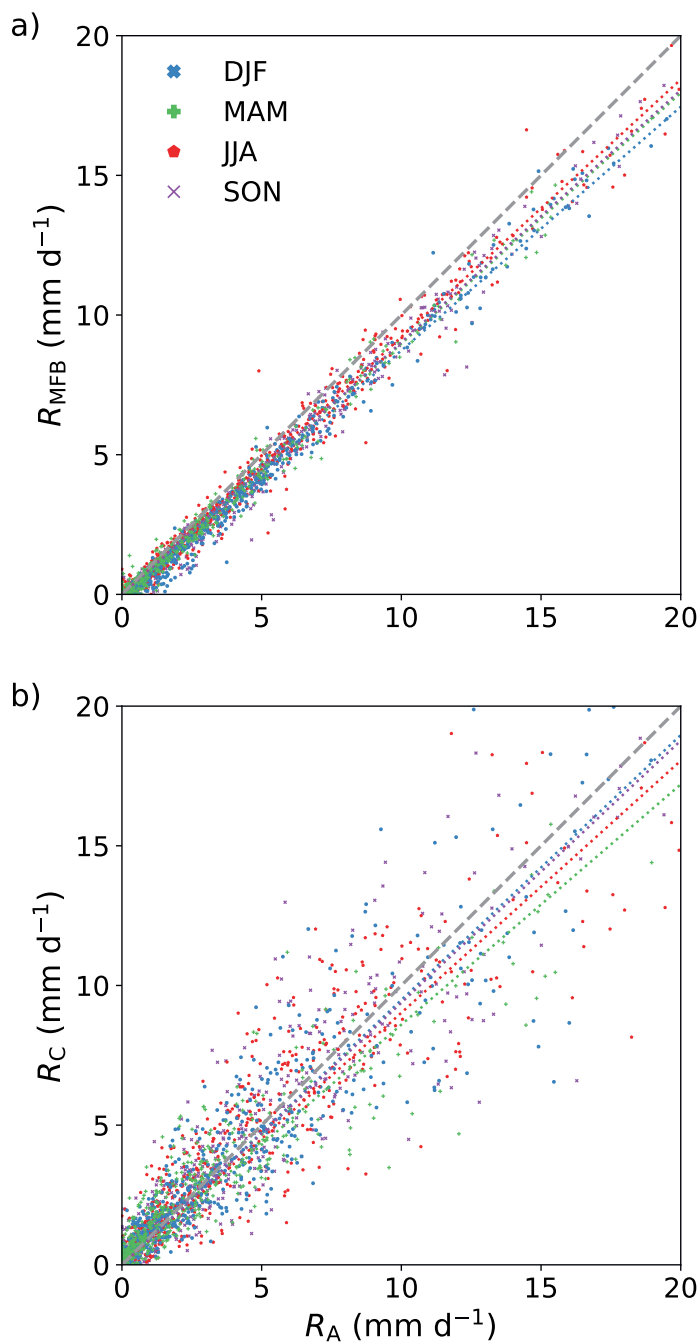


Figure 3.4 | Comparison between the reference rainfall (R_A) and the two adjusted radar QPE products: (a) R_{MFB} and (b) R_C). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are indicated in Table 3.2. The colored dashed lines are a linear fit, forced through the origin, per season between the reference and the two QPE products.

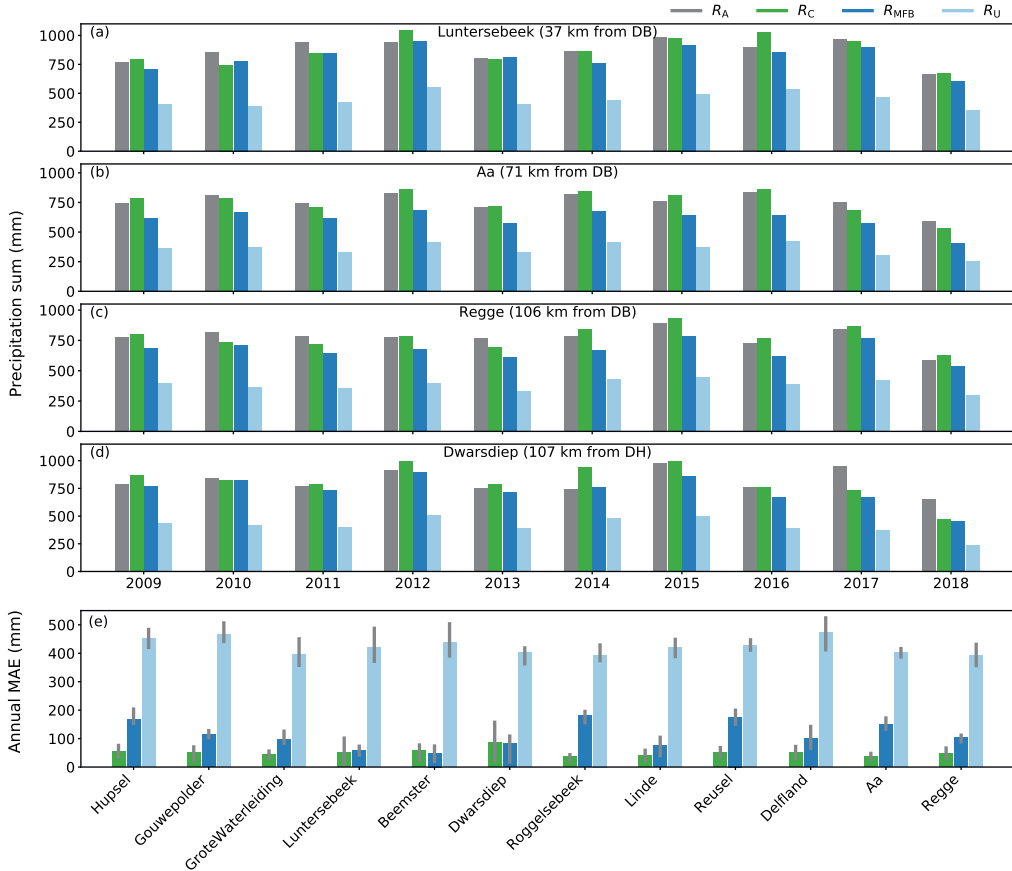


Figure 3.5 | Effect of the adjustment factors on the catchment-averaged annual rainfall sums. (a – d) The results for a sample of four catchments that are spread over the country (and thus the radar domain): (a) Luntersebek, (b) Aa, (c) Regge and (d) Dwarsdiep. Shown are R_A (grey), the estimated rainfall sum after correction with the CARROTS factors (R_C ; green), the estimated rainfall sum after correction with the MFB-adjustment factors (R_{MFB} ; dark blue) and the rainfall sum with the unadjusted radar rainfall estimates (R_U ; light blue). The distance between the catchment centre and the closest radar in the domain is given in the title of subfigures a–d (DH is Den Helder and DB is De Bilt). The radar in Herwijnen, which replaced the radar in De Bilt in January 2017, is not included here, because this radar was operational for the shortest time in this analysis. (e) the mean absolute error of the annual precipitation sum between the QPE products and the reference rainfall sum (R_A). The vertical grey lines, per bar, indicate the IQR of the mean absolute error (MAE) based on the ten years.

consistent with Holleman (2007). The performance of R_{MFB} for the Dwarsdiep catchment is similar to its performance for the Linde catchment (both in the north of the country), but R_{C} shows more variability in the error from year to year for the Dwarsdiep catchment (Figure 3.5d), leading to a better relative performance of R_{MFB} . The CARROTS QPE tends to overestimate the rainfall amount of the three aforementioned catchments (Beemster, Dwarsdiep and Luntersebeek) for some years (e.g. by 16% for the Luntersebeek in 2016). Overall, the performance of R_{C} and R_{MFB} are not that different for these three catchments, with on average just a lower MAE for R_{MFB} than for R_{C} for the Luntersebeek catchment and Dwarsdiep polder (Figure 3.5e).

Summarizing, the CARROTS factors have a clear annual cycle, with generally higher adjustment factors further away from the radars (Section 3.3.1). On average for the Netherlands, the MFB-adjusted QPE outperforms the CARROTS-corrected QPE. However, the spatial variability in the CARROTS factors, in contrast to the uniform MFB adjustment, results in estimated annual rainfall sums for the twelve hydrological catchments that are generally closer to the reference (for nine out of twelve catchments) than with the MFB-adjusted QPE, especially for the east and south of the country. This effect is expected to become more pronounced when the adjusted QPE products are used for discharge simulations.

3.3.3 | Effect on simulated discharges

The severe underestimations of R_{U} have a considerable effect on the discharge simulations for the twelve catchments (Figure 3.6). This leads to hardly any discharge response and thus negative KGE values for most catchments as compared to discharge simulations with the reference rainfall data. The effect is most pronounced for the freely draining catchments in the east and south of the country. These catchments are more driven by groundwater flow than the polders in the west of the country. Groundwater flow gets hardly replenished, because of similar estimated annual evapotranspiration and R_{U} sums, resulting in too low baseflows. The polders, especially Delfland and Beemster, are an exception to this, because they are less driven by groundwater-fed baseflow and more by direct runoff from greenhouses or upward seepage flows, which makes them more responsive to individual rainfall events leading to higher KGE values (with R_{U} as input) compared to the other catchments.

The model runs using R_{MFB} as input significantly improve the simulated discharges, compared to the runs with R_{U} . Nevertheless, the model runs still strongly underestimate the simulated discharges compared to those from the reference runs for the catchments in the south and east of the country (Figure 3.6a–f). This is particularly noticeable for the catchments Reusel (KGE = 0.26) and Roggelsebeek (KGE = 0.04). The spatial uniformity of the MFB factors is identified as the cause of these effects, because the MFB method can not correct for the sources of errors leading to the biased QPE in space. This already led to clear underestimations in the annual rainfall sums for these regions (Figure 3.5).

The CARROTS QPE outperforms R_{MFB} , when this product is used as input for the twelve rainfall-runoff models. This is not exclusively the case for the six catchments in the east and south of the country (Figure 3.6a–f), but also for the other polder and catchment areas. Only for the Beemster polder the difference is minimal. The Beemster is mostly fed by upward seepage, leading to a more predictable baseflow for all models runs. In addition, the catchment is located close to an automatic weather station and is located in between both operational radars, which makes

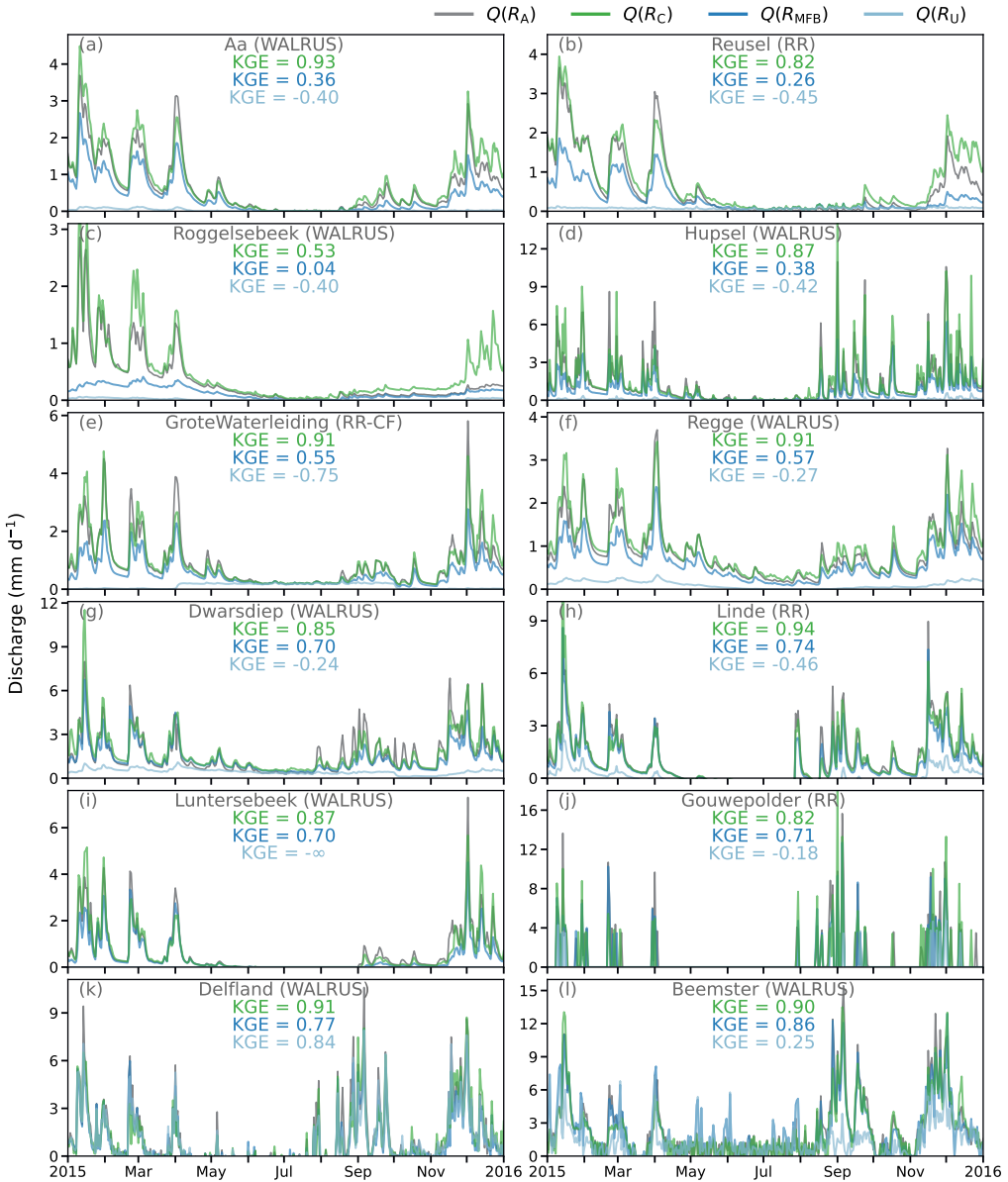


Figure 3.6 | Differences in simulated discharges for the twelve catchments (a-l) as a result of the differences between rainfall estimates. The models are run for the period 2009–2018 with the following rainfall products as input: the reference (R_A ; grey), the QPE corrected with the CARROTS factors (R_C ; green), the MFB-adjusted QPE (R_{MFB} ; dark blue) and the unadjusted radar rainfall estimates (R_U ; light blue). Only the simulated discharges for 2015 are shown here for clarity; the KGE is based on all years.

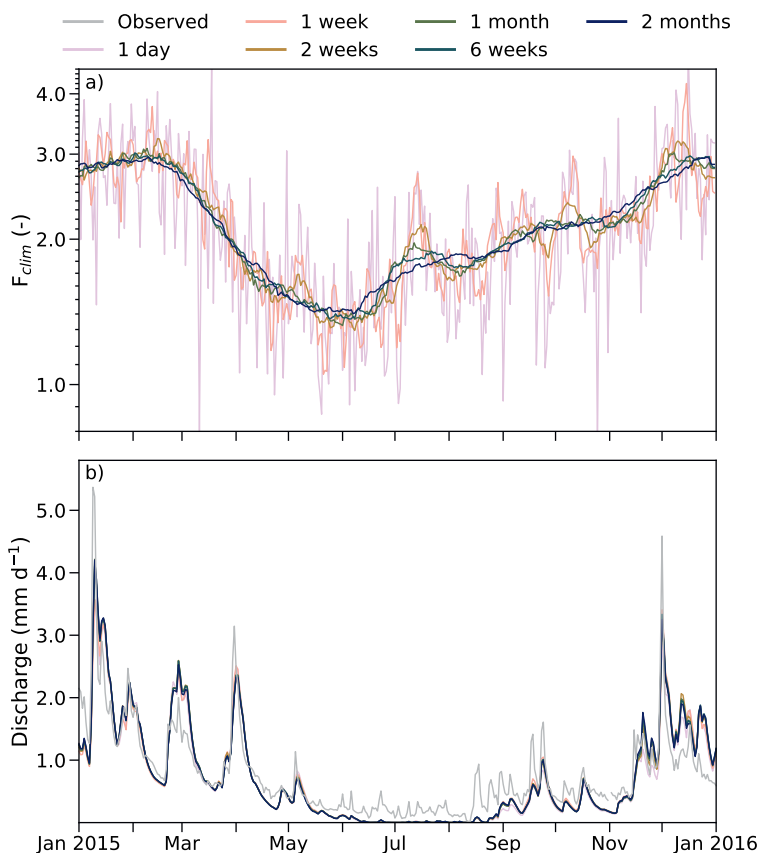


Figure 3.7 | Sensitivity of the CARROTS factor derivation to the moving window size. (a) The adjustment factors for the Aa catchment for six different moving window sizes. The moving window size of 31 days was used in the methodology of this Chapter. (b) The effect of the six moving window sizes in (a) on the simulated discharges for the Aa. Similar to Figure 3.6, the CARROTS factors were derived and discharge was simulated for the full period (2009–2018), but only 2015 is shown here. The grey line indicates the observed discharge.

the MFB adjustment more beneficial for this region. The difference in performance between the hydrological model simulations is small, with a KGE of 0.90 (using R_C) versus 0.86 for R_{MFB} , as compared to the reference run.

3.3.4 | Sensitivity analysis

The use of a different moving window size hardly influences the CARROTS factors for moving window sizes of two weeks or longer, but this does not hold for moving window sizes of a day or, to a lesser extent, one week (Figure 3.7a). The factor derived with a moving window size of one day fluctuates heavily from day to day. This suggests that the adjustment factor is still quite sensitive to individual events in the 10-year period, when a moving window size of seven days or less is used. Moving window sizes of more than a month (6 weeks and 2 months were tested here), lead to similar CARROTS factors as with a 1-month (31-day) moving window size, but

somewhat more smoothed. A similar effect likely takes place for a seasonal (3-month) moving window. For larger moving window sizes (half a year to a year, for instance), we expect that the seasonality in the factor is lost and that an average correction factor remains.

In contrast to this, the differences between these six sets of CARROTS factors (Figure 3.7a) lead to minimal variations in the simulated discharges for the Aa catchment, when these factors are used to adjust the input QPE (Figure 3.7b). Differences in timing and magnitude ($0.2\text{--}0.3\text{ mm d}^{-1}$) are visible during peaks and recessions, for instance in early April. However, these are small compared to the differences between the model runs with R_C and R_{MFB} (Figure 3.6). However, the use of a window size of 1 day or, to a lesser extent, of a week clearly leads to more fluctuations in the CARROTS factor (Figure 3.7a) and can therefore influence the rainfall estimation for individual events (and the factor will also be influenced by these individual events). For quickly responding catchments and urban catchments, this could still lead to different results. Concluding, a 31-day smoothing of the climatological adjustment factor is warranted.

In addition, leaving individual years out of the ten-year archive has a limited impact on the CARROTS factors (see also the vertical bars in Figure 3.3c). Similar to the aforementioned results for the moving window size analysis, it leads to hardly any variations in the simulated discharges for the Aa catchment (not shown here). This suggests that the ten-year archive length was sufficiently long for the factor derivation.

3.4 | Discussion

In this chapter, we introduced the CARROTS method to derive adjustment factors that reduce the bias in radar rainfall estimates. We derived these factors using 10 years of 5-min radar and reference rainfall data for the Netherlands. The method and resulting QPE product outperformed the mean field bias (MFB) adjustment, that is used operationally in the Netherlands, for catchments in the east and south of the country. When the QPE products were used as input for hydrological model runs, the method outperformed the MFB-adjustment method for all catchments.

The main difference that distinguishes the CARROTS method from the MFB adjustment is the presence of a high-density network of (manual) rain gauges in the reference dataset, a dataset that is not available in real-time. This allows for spatial adjustments. Overeem et al. (2009b) demonstrate that this reference dataset mostly depends on the daily spatial adjustments from the manual rain gauges, while the higher-frequency MFB adjustment based on the automatic gauges plays a smaller role in the adjustments of this reference product. According to Saltikoff et al. (2019), at least 40 countries have an archive of historical radar data for a period of ten years or more. The proposed CARROTS method is potentially valuable for these countries, especially when the density of their network of automatic rain gauges is, similar to the Netherlands, significantly smaller than the total network of rain gauges. An additional advantage of the method is the real-time availability of the correction factors, which is independent of the timeliness of the rain gauge data.

MFB adjustment of radar rainfall fields is still the most frequently applied adjustment method (Holleman, 2007; Harrison et al., 2009; Thorndahl et al., 2014; Goudenhoofd & Delobbe, 2016). The results indicate that this choice may be reconsidered for hydrological applications in the

Netherlands, especially further away from the radar and in case a country-wide or large-region adjustment factor is applied. This could also hold for other regions, especially mountainous regions where the uniformity of the MFB-adjustment factor is likely not sufficient to correct for all orography-related errors (Borga et al., 2000; Gabella et al., 2000; Anagnostou et al., 2010). More regionalized MFB adjustments are possible, but depend on the density and availability of the automatic gauge stations.

However, the proposed CARROTS method has to be recalculated for every change in the radar setup, calibration, additional post-processing steps (e.g. VPR corrections, Hazenberg et al., 2013) or final composite generation algorithm. For instance, including a new radar in the composite would require a recalculation of the adjustment factors, thereby assuming the presence of an archive of the new composite product. This could potentially limit the usefulness of the proposed method. As mentioned in Section 3.2.1, the replacement of both Dutch radars by dual-polarization radars in combination with the replacement of the radar at location De Bilt to location Herwijnen (Figure 2.2) between September 2016 and January 2017 only had a limited impact on the operational products, and thereby on the CARROTS derivation. The operational products are not yet (fully) making use of the dual-polarization potential. We expect that the factors will have to be recalculated as soon as the additional information from the dual-polarization radars is used to improve the products or when e.g. the German and Belgian radars close the Dutch border are added to the composite.

That CARROTS is relatively insensitive to such minor changes in the composite or the year-to-year variability of rainfall, is likely a result of the ten-year archive that has been used. The sensitivity analysis in Section 3.3.4 has shown that leaving individual years out of the archive hardly influences the CARROTS factors. Nevertheless, based on the current analysis we cannot conclude what the minimum number of years in the archive has to be to obtain stable CARROTS factors that are similar to the factors derived in this Chapter. This is a recommendation for future research. In the case of a new radar QPE product, it is also recommended to recalculate the archive (if possible), to make sure new CARROTS factors can be derived.

Although the results are promising, this method is not expected and meant to outperform more advanced spatial QPE adjustment methods, such as geostatistical and Bayesian merging methods (for an overview of methods and their limitations, see Ochoa-Rodriguez et al., 2019). A major advantage of these methods is the real-time derivation of spatial adjustment factors, in contrast to the proposed method in this chapter, which was solely based on historical data. The MFB-adjustment factors can also be derived in near real-time, but are uniform in space, which can explain the worse performance as compared to the proposed method in this Chapter. A possible disadvantage of these real-time methods (MFB, geostatistical and Bayesian merging) is the dependency on the timely availability of rain gauge data, which is not the case for CARROTS. Altogether, we consider the proposed climatological radar rainfall adjustment method as a benchmark for the development and testing of operational radar QPE adjustment techniques.

Another possible option would be to combine the CARROTS method with the real-time application of the MFB adjustment, i.e. CARROTS is applied and the resulting QPE is then adjusted with real-time MFB-adjustment factors. This would allow for real-time temporal corrections of the QPE, without the need for a high density of rain gauges in real-time, while the corrections in space are based on the (historical) CARROTS factors.

As mentioned in the previous paragraph, the climatological adjustment factor is not calculated for the current meteorological conditions and resulting QPE errors, which could lead to considerable errors during extreme events. Nonetheless, this is also the case for the MFB-adjustment technique (Schleiss et al., 2020). The absolute errors for the 10 highest daily sums in this chapter for the Aa and Hupsel Brook catchments (one of the largest and the smallest catchments in the study) are similar for the MFB and climatological adjustment methods, with on average a 20% difference with the reference (this would have been 50 to 60 % without corrections). In most of these events, both R_C and R_{MFB} underestimated the true rainfall amount. However, for a small number of these top 10 events, the QPE products overestimated the true rainfall amount. This occurred more frequently with CARROTS (25% of the cases) than with the MFB adjustment (15% of the cases). Note that for individual events in these twenty extremes, the errors can still reach 48% for the QPE adjusted with CARROTS and 64% for the MFB-adjusted QPE. A way to better correct for biases during extreme events could be to derive either different Z-R relationships, depending on the type of rainfall, or dBZ-dependent correction factors, which could be derived in a similar way to the CARROTS derivation method. Whether this works or not for extreme events depends on the number of such events in the available historical dataset.

Finally, the CARROTS factors were derived with the reference rainfall data for the Netherlands. The same data was used as reference in this Chapter. Although the use of the same data as training and validation set is sub-optimal, leaving out individual years has had a limited impact on the estimated adjustment factors and the resulting QPE and discharge simulations (see also the vertical bars in Figure 3.3c). Note, however, that in catchments with a large number of manual rain gauges, but where automatic rain gauges are not nearby, the CARROTS results will likely be closer to the reference than the MFB-adjusted simulations. Although this is warranted for the CARROTS method, it can partly explain why the method works better for some catchments than others.

3.5 | Conclusions

A known issue of radar quantitative precipitation estimations (QPE) are the significant biases with respect to the true rainfall amounts. For this reason, radar QPE adjustments are needed for operational use in hydro-meteorological (forecasting) models. Current QPE adjustment methods depend on the timely availability of quality-controlled rain gauge observations from dense networks. This especially applies to methods that correct for the spatial variability in the QPE errors. To overcome this issue and to provide a benchmark for future QPE algorithm development, we have presented CARROTS (Climatology-based Adjustments for Radar Rainfall in an OperaTional Setting): a set of gridded climatological adjustment factors for every day of the year. The factors were based on a historical set of 10 years of 5-min radar rainfall data and a reference dataset for the Netherlands. The climatological adjustment factors were compared with the mean field bias (MFB) adjustment factors, which are used operationally in the Netherlands. For the period 2009–2018, daily and sub-daily rainfall estimates with both the MFB-adjusted and CARROTS-adjusted QPE were validated against the reference rainfall for the land surface area of the Netherlands. In order to provide a hydrometeorological testbed, both adjustment factors were also validated on the estimated annual rainfall sums and the effect of the adjusted QPE products on simulated discharges with the rainfall-runoff models for twelve Dutch lowland catchments.

The CARROTS factors show clear spatial and temporal patterns, with higher adjustment factors towards the edges of the radar domain. This is caused by larger QPE errors further away from the radars. The factors are also higher from December through March than in other seasons. This is likely a result of sampling above the melting layer during these months, which causes higher underestimations in the unadjusted radar rainfall product.

On average for the Netherlands, the MFB-adjusted QPE outperforms the CARROTS-corrected QPE. Although the MFB factors are based on the current over- or underestimations in the QPE, the factor is spatially uniform and does not correct for spatial errors. This directly impacts the adjusted QPE when the QPE products are tested for the twelve Dutch catchments. The MFB-adjusted QPE leads to annual rainfall sums that still underestimate those of the reference for the catchments in the east and south of the country (towards the edge of the radar domain). This bias is almost absent for the annual rainfall sums after correction with the CARROTS factors (up to 5% over- and underestimation for the same catchments). For catchments closer to radars, this effect decreases and both adjustment methods perform well.

The effects of both adjustment methods on the QPE is amplified when they are used as input for the rainfall-runoff models of the twelve studied catchments. The discharge simulations with the CARROTS QPE outperforms those using the MFB-adjusted QPE for all catchments. For hydrological applications in the Netherlands, these results indicate that the current operational use of a country-wide MFB adjustment may be reconsidered as it often performs worse than the proposed climatological adjustment factor, which can be seen as the minimum benchmark to outperform.

Despite the aforementioned results, the CARROTS method has two main limitations: (1) for every change in the radar setup, the radar calibration, post-processing algorithms or the final composite generation method, the adjustment factors have to be recalculated; (2) the factor is not calculated for the actual meteorological conditions and resulting QPE errors, which could lead to considerable errors during extreme events. Nonetheless, the latter is also the case for the MFB-adjustment technique (Schleiss et al., 2020), even though the MFB factors are derived in real-time.

The main advantage of the introduced method is the continuous availability of spatially distributed adjustment factors, due to the independence of timely rain gauge observations. This is beneficial for operational use. In addition, the CARROTS factors are shown to be robust, as the derivation is not found to be sensitive to leaving out individual years or the used moving window, especially when this window is longer than a week.

Finally, this method is not expected and meant to outperform more advanced spatial QPE adjustment methods (which require data from dense rain gauge networks for robust application), but it can serve as a benchmark for the development and testing of more advanced operational radar QPE adjustment techniques. QPE adjustment methods (including CARROTS) greatly benefit from a denser, frequently-available rain gauge network. From that perspective, crowd-sourced personal weather stations hold a promise for improving radar rainfall products, given their direct surface measurements and dense networks (Vos et al., 2019). This also holds for rain gauge observations from other governmental or third parties, e.g. the water authorities in the Netherlands. Hence, we think that this could further improve radar rainfall products in the near future.



4

Comparing and evaluating radar rainfall nowcasting techniques

This chapter was originally published as:

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. doi:10.1029/2019WR026723

RADAR rainfall nowcasting, the process of statistically extrapolating the most recent rainfall observation, is increasingly used for very-short-range rainfall forecasting (less than six hours ahead). We performed a large-sample analysis of 1,533 events, systematically selected for four event durations and twelve lowland catchments (6.5–957 km²), to determine the predictive skill of nowcasting. Four algorithms are tested and compared with Eulerian Persistence: rainymotion Sparse, rainymotion DenseRotation, pysteps deterministic and pysteps probabilistic with 20 ensemble members. We focus on the dependency of nowcast skill on: event duration, season, catchment size and location. Maximum skillful lead times increase for longer event durations, due to the more persistent character of these events. For all four event durations, pysteps deterministic attains the longest average decorrelation times, with 25 min for 1-h durations, 40 min for 3 h, 56 min for 6 h and 116 min for 24 h. During winter, with more persistent stratiform precipitation, we find three times lower mean absolute errors than for convective summer precipitation. Higher skill is also found after spatially upscaling the forecast. Catchment location matters too: given the prevailing storm movement, two times higher skillful lead times are found downwind than upwind towards the edge of the domain. In most cases, pysteps algorithms outperform the rainymotion benchmark algorithms. We speculate that most errors originate from growth and dissipation processes which are not or only partially (stochastically) accounted for.

“Someone told me long ago
There’s a calm before the storm
I know
It’s been coming for some time”

—Creedence Clearwater Revival, *Have You Ever Seen The Rain* (1971)

4.1 | Introduction

The frequency and severity of intense precipitation events are likely to increase in a changing climate (with e.g. a 12% increase in high-intensity precipitation per degree of warming in the Netherlands), which can lead to more severe floods and present a danger to livability and economy (e.g. IPCC, 2011, 2013, 2014; KNMI, 2015). Well-established early warning systems (e.g. Delft-FEWS; Werner et al., 2013) make it possible to act accordingly and in time, expectedly resulting in a lower risk and less damage (Pappenberger et al., 2015). Most early warning systems, if present at all, use a combination of short-range (12–72 h) and medium-range (up to 10 days) numerical weather prediction (NWP) in combination with hydrological and hydraulic models to predict river discharges and water levels. However, the quantitative precipitation forecasts (QPF) provided by the employed NWP systems are often not sufficient for reliable early warnings on the short-term, i.e. up to 6 hours, due to (1) a too coarse temporal resolution and a too low update frequency (i.e., the lead time becomes already quite long, making the forecast less reliable), and (2) the mislocation of rainfall events (e.g. Pierce et al., 2012; Berenguer et al., 2012).

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

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

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

As operational nowcasting for hydrological purposes is still in an early stage of development, advice is needed on the skill of radar nowcasting in general and differences between the performance of algorithms in particular. Most studies so far have focused on the development of nowcasting algorithms in combination with a quantification of the rainfall prediction quality and

errors in either deterministic or probabilistic nowcasting algorithms (e.g. Germann & Zawadzki, 2002, 2004; Turner et al., 2004; Lin et al., 2005; Germann et al., 2006; Foresti et al., 2016). The results generally follow from studies with analyses based on relatively small samples of 2-15 precipitation events. The studies by Berenguer & Sempere Torres (2013), Foresti & Seed (2015) and Mejsnar et al. (2018) are exceptions to this. Foresti & Seed (2015) use a dataset of 20 months of operational nowcasts in order to analyse the spatial distribution of radar rainfall nowcasting errors in a mountainous region in south-east Australia.

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

In this chapter, 1533 events spread over twelve catchments with sizes varying from 6.5 to 957 km², are analysed. For this analysis, four open-source (Python) nowcasting algorithms are used: two benchmarking advection algorithms of rainymotion (Ayzel et al., 2019b), and deterministic and probabilistic versions of pysteps (Pulkkinen et al., 2019). To the authors' knowledge, this is the first radar rainfall nowcasting study with a combination of this variety of algorithms and such a large sample of events.

The outline of this paper is as follows: section 2 contains descriptions of the study area, radar data, event selection, nowcasting algorithms, verification metrics and experimental setup. This is followed by the results (section 4.3), discussion (section 4.4) and conclusions (section 4.5).

4.2 | Materials and methods

4.2.1 | Study area

In this chapter, we analyse rainfall over twelve Dutch lowland catchments and polder areas with different sizes and locations (see Figure 2.2 and Section 2.3.1). We focus on catchments instead of the entire radar domain, because we want to assess the usefulness of nowcasting for the involved water authorities. Incidentally, this highly reduces the storage requirement for this large-sample analysis (compared to nowcasts for the full domain). The catchments were chosen in close collaboration with involved water authorities and are spread over the country, as we expected a dependency of the nowcast skill on the location with regard to the prevailing storm movement. With south-westerlies as the predominant wind direction in the Netherlands, catchments in the northeastern part of the country are expected to have skillful rainfall predictions up to longer lead times than catchments in the southwest. For more information regarding the twelve lowland catchments, we refer to Section 2.3.1.

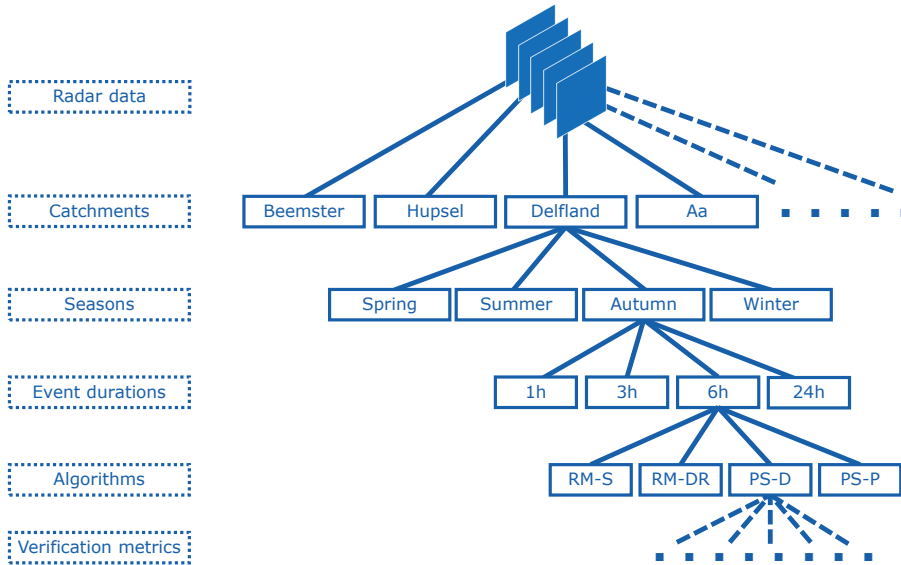


Figure 4.1 | Schematization of the employed event selection procedure. Per catchment, season and event duration, eight events are selected.

4.2.2 | Data and event selection

4.2.2.1 | Radar rainfall product

In this chapter, rainfall estimates from the gauge-adjusted ('reference') and unadjusted radar datasets with a 1-km² spatial and 5-min temporal resolution, as described in Section 2.2.1, were used. The real-time unadjusted quantitative precipitation estimation (QPE) (provided by KNMI) is assumed to be the true rainfall intensity for the verification of the rainfall forecasts, i.e. the output of the nowcasting algorithms. This is also the reason that the CARROTS correction factors of Chapter 3 were not used in this chapter. Provided that radar QPE comes with substantial uncertainty and bias (e.g. Germann et al., 2006; Hazenberg et al., 2011; Foresti & Seed, 2015; Van de Beek et al., 2016), a good verification result in this chapter does not necessarily mean that the true rainfall amounts are well predicted by the algorithms. The gauge-adjusted reference rainfall product is not available in real-time, so we have only used this 'reference' dataset for the event selection (Section 4.2.2.2).

4.2.2.2 | Event selection procedure

A large number of events was selected systematically from the 5-min gauge-adjusted radar rainfall composites for the period 2008–2018. Only large rainfall accumulations were selected, as these are most interesting to study for both the assessment of precipitation predictability and the hydrological application. The gauge-adjusted dataset is employed for event selection instead of the operational product employed for the nowcasting analysis (section 4.2.2.1), because this yields the events with the actual highest rainfall volumes.

In the Dutch climate, the highest rainfall rates originate from convective precipitation events during summer and early autumn. Since one of the aims of this chapter is to find the seasonal

Table 4.1 Rainfall and wind statistics for the selected events and seasonal averages for KNMI station De Bilt. For the event statistics, rainfall intensities are catchment averages over the duration and all events in that duration. The mean daily wind direction for the events is obtained from measured wind directions during these events at KNMI station De Bilt. For the climatology, the seasonal accumulations for station De Bilt are averaged over the period 1981–2010 (KNMI, 2011). Mean climatological wind directions have been determined over periods with rainfall by Buishand & Velds (1980).

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

dependence of the nowcasting skill, events in other seasons, which include stratiform events, should be present as well (see Table 4.1 for seasonal statistics of the Dutch weather). The adopted event selection procedure guarantees that an even spread of strong precipitation events is obtained over all seasons per event duration.

The events are selected as follows (Figure 4.1): per catchment and for each season, eight events are selected per event duration (1, 3, 6 and 24 hours). Note that an ‘event’ is not defined by the start and end of rainfall, but instead periods with a certain duration are used in which it does not have to rain continuously. Hence, the full duration is considered as an event. With this description of an event, the highest rainfall sums per duration, catchment and season are ranked from high to low, in which the next ‘event’ cannot occur within the time span of a previously selected ‘event’ with a higher rainfall sum. The eight events, for that duration, consist of the events with the four highest catchment-averaged rainfall sums and the four highest rainfall sums for any grid cell in the catchment. If one of the four events of the grid maxima is the same event as already present in the four maxima from the catchment-averaged list, the next maximum in the list of grid maxima is used to avoid overlapping events. Summed over all durations (4), seasons (4) and catchments (12), this selection procedure leads to $4 \times 4 \times 12 \times 8 = 1536$ events (see Table 4.1 for the statistics of these events).

4.2.3 | Nowcasting algorithms

The tested nowcasting methods in this chapter are Eulerian persistence, rainymotion Sparse, rainymotion DenseRotation, pysteps deterministic and pysteps probabilistic with 20 ensemble members. Rainymotion v0.1 and pysteps v0.2 were employed. The methods are described in more detail in Section 2.5 of Chapter 2 and we follow the model setup described in this section. For brevity, the algorithm names are abbreviated from here on (Table 4.2).

4.2.4 | Experimental and forecast verification setup

The nowcasts for the 1536 events were run (equally spread) on two high-performance clusters with Intel Xeon processors with 2.2 GHz and 8 GB memory per core, and Intel Xeon processors with 3.6 GHz and 8 GB memory per core. Run times on these clusters for 5-min forecasts with a 6-h lead time were on average (for one core): 40 s for RM-S, 130 s for RM-DR, 40 s for PS-D

Table 4.2 | Overview of the radar rainfall nowcasting methods used in this chapter.

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

and 1250 s for PS-P (for 20 ensemble members). Hence, the run time for one PS-P nowcast took longer than the update frequency. From an operational perspective, this would require either reducing the forecast horizon (e.g. forecasts with a 3-h lead time) or reducing the update frequency. Another option would be to run PS-P on multiple cores.

During 3 out of the 1536 events, one algorithm did not finish successfully, leaving 1533 events available for analysis. In these three cases (two failures for PS-D and one for PS-P), the initialization of the output file failed. After analysing the log files, we concluded that this error was likely caused by the high-performance cluster rather than caused by the algorithm.

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

4.2.4.1 | Event type and duration dependency

The events selected for the four durations contain different types of rainfall, generally from convective and small-scale for the shortest event durations (1-h) to more stratiform, larger-scale systems for the longer event durations (e.g. 24-h). As a result of the differences in rainfall types for these event durations, the mean decorrelation distance increases with almost a factor 2.5 from 1-h to 24-h in the Netherlands (Van de Beek et al., 2012). Based on this, we expected that this would lead to a difference in predictive skill of the nowcasts for the event durations. For the purpose of finding, per algorithm, the dependency of forecasting skill on the event type and duration, Pearson's correlation coefficient was calculated per lead time t and for every event (averaged over all grid cells within a catchment), as described in Equation 2.7 in Section 2.6. As PS-P is a probabilistic run, Pearson's correlation was calculated for every ensemble member separately. As such, all separate model runs within the ensemble were taken into account. The average skillful lead time per algorithm and event duration could be estimated from the 1/e-line.

4.2.4.2 | Seasonal dependency

Rainfall characteristics and seasonal differences vary considerably between and within event durations (Van de Beek et al., 2012). Whereas winters in the Netherlands generally have wide-spread frontal, stratiform rainfall fields of low to intermediate intensity, summer rainfall also

consists of more localized convective showers with higher rainfall intensities. We expected that this also impacts the nowcasts for the twelve catchments. To verify the nowcasts for the different seasons, we focused on one event duration: 6 h. Within this interval, the MAE and CRPS (described in Equations 2.9 and 2.10 of Section 2.6) were calculated per catchment to estimate the error between forecasts and observations per event in a season.

4.2.4.3 | Dependency on catchment size and location

Vivoni et al. (2006) found that flood forecast skill increases with increasing catchment area. This may also be the case for the precipitation predictability of nowcasts. Within the field of NWP, it is common practice to upscale forecasts to a coarser resolution, as this gives a better representation of the rainfall fields when forecast rainfall fields are mislocated (e.g. Mittermaier, 2006). It is possible that the spatial resolution necessary for a minimum forecast skill is larger than the smallest catchments in this chapter (Figure 2.2 and Table 2.2). Hence, it is useful to find a minimum scale on which forecasts are still skillful in order to draw conclusions about the dependency of nowcast skill on catchment size.

For this analysis, the FSS (described in Equation 2.15 of Section 2.6) was estimated for the two largest catchments (Aa and Regge), with a maximum length scale of 49 km (given a rectangular box around the catchments). The FSS was calculated for every odd number from 1 to 97 km for events with a 6-h event duration. At 97 km ($2N - 1$, with N the longest length scale in the Aa catchment, because of its elongated shape), the skill approaches an asymptote where $FSS = 1$, when the forecast is unbiased, i.e. the fraction of observed rainfall exceeding the threshold over the entire domain is the same as the fraction of forecast rainfall exceeding this threshold. If not, asymptotic behaviour will take place at a value lower than 1 (Roberts & Lean, 2008; Mittermaier & Roberts, 2010).

In addition to the catchment area, the location with regard to the radar location(s) and storm movement may influence the nowcast skill. To determine whether or not this relationship between location and skill is present, the maximum skillful lead time (similar to section 4.2.4.1) was used for the 6-h event duration. Since differences in catchment size would affect the results, the correlation and maximum skillful lead time are calculated for 5×5 cells in the centre of the catchment, as this fits in the output extent of all twelve catchments. For the Hupsel Brook catchment (6.5 km^2), cells surrounding the catchment are used as well. Similar to section 4.2.4.1, both metrics are calculated for each ensemble member in the nowcasts of PS-P.

4.2.4.4 | Ensemble forecast verification

Ensemble predictions are used to account for the uncertainty in predictions. As such, the ensemble spread gives the forecaster an indication of the uncertainty in the forecast. The ensemble mean often shows a better skill than a purely deterministic forecast (e.g. Richardson, 2000). An ensemble, however, is only useful when the ensemble has a representative spread, ideally with a minimal bias. In addition, an end-user needs to know how trustworthy a resulting forecast probability to exceed a certain rainfall threshold is. For this purpose, the reliability diagram and ROC curve (described in Sections 2.6.4.2 and 2.6.4.3) were employed in this chapter for events with a 6-h duration. Only PS-P is used for the ensemble forecast verification, since this is the only probabilistic nowcasting algorithm in this chapter.

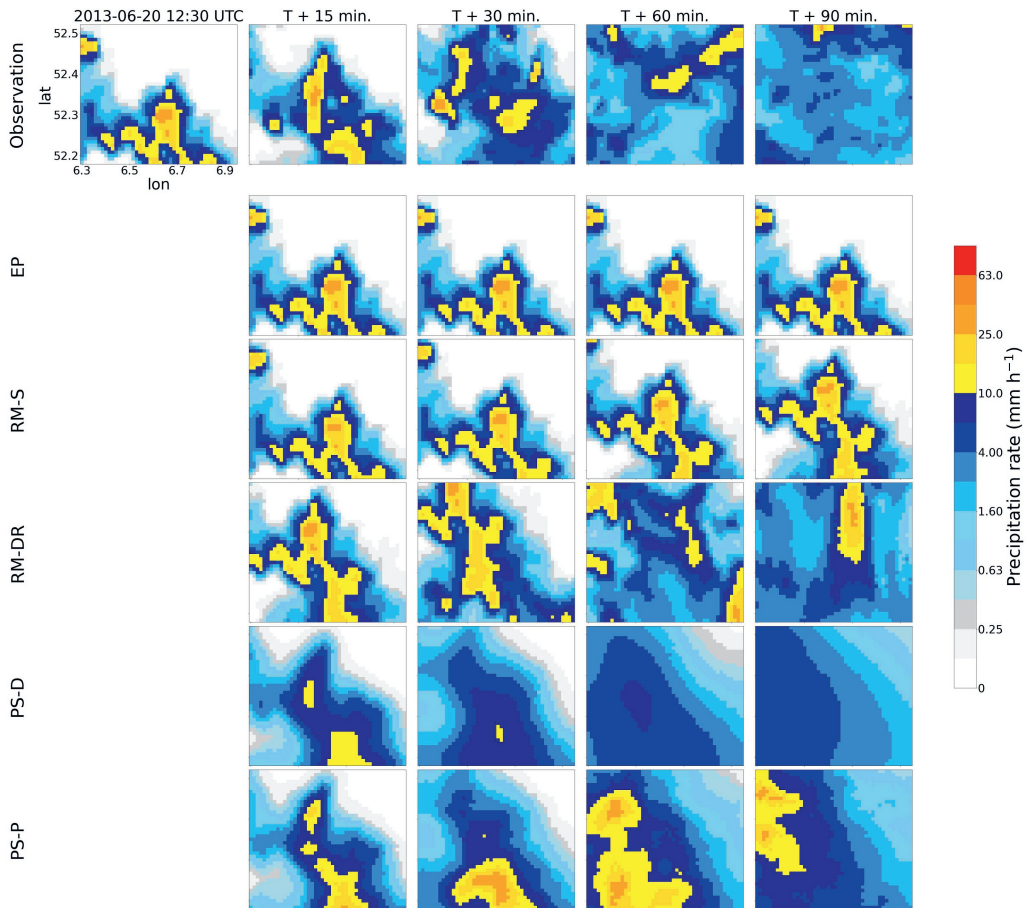


Figure 4.2 | Example of a set of nowcasts for the Regge. The illustrated event took place on 2013-06-20 and resulted in an average of 29.4 mm over the area in 3 hours (between 12:05 and 15:05 UTC), with local maxima around 45 mm. For PS-P, only ensemble member 10 is shown.

4.3 | Results

An example nowcast for an event (from the 24-h duration) in the Regge catchment at 12:30 UTC on 20 June, 2013, is shown in Figure 4.2. Although the event in Figure 4.2 is just one event out of the large sample, it gives a good example of the difficulty of forecasting convective precipitation affected by storm movement, growth and dissipation, and merging and splitting of the precipitation systems. All algorithms have difficulties capturing these processes well. Whereas RM-DR, PS-D and PS-P (ensemble member 10 is shown) seem to capture the movement to a certain extent, RM-S has the right direction, but almost no movement. Naturally, there is no movement for EP either. PS-D has, for this particular case, the disadvantage that there is too much dissipation, leading to the loss of the high-intensity rainfall centres while a mean large-scale field of rainfall persists. This is likely due to the short lifetime and small extent of the rain structures, which are decomposed into more quickly dissipating fields in PS-D. Between

RM-DR and PS-P, which both capture the high-intensity rainfall cells, the main differences are the size and location of the rainfall systems. Based on visual inspection of this example, RM-DR approximates the observations for longer lead times best.

In the remainder of this section, the full sample of events is used for the verification of the forecasting skill of these nowcasting algorithms, but only the results for the 6-h event duration are shown (except for section 4.3.1, as this section focuses on the different durations). Note that only the forecasts for times within the predefined events are analysed here.

4.3.1 | Event type and duration dependency

With increasing event duration, the decorrelation time increases (Figure 4.3). Maximum skillful lead times, seen as the mean of the intersections between the $1/e$ -line and the mean correlation of an event, increase from 25 min for 1-h durations, to 40 min for 3-h durations, 56 min for 6-h durations and 116 min for 24-h durations. In all cases, PS-D attains the longest skillful lead times.

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

The correlation varies between events and this variability decreases with increasing event duration, as indicated with the colored error bars in Figure 4.3. This indicates that small-scale systems with shorter lifetimes vary more between events, leading to more variability in the nowcasting results. This sometimes leads to significant negative correlations for EP and RM-S, meaning that forecast and observed rainfall in cells have the opposite tendency, e.g. decreasing rainfall amounts in the forecast while the observations have increasing amounts.

Although the correlations of PS-D and PS-P are quite similar for the shorter event durations, the attained correlations for the 6-h and 24-h durations are 15–25% lower for PS-P than for PS-D. Note that comparing a deterministic with a probabilistic run is not entirely fair, because the main advantages of a probabilistic run are not weighed. Between those two algorithms and RM-DR, the difference is considerable, with maximum skillful lead times that are generally 35–60% lower than the pysteps algorithms. This suggests that taking spatial and temporal scales into account, as done in the pysteps algorithms, adds value to only rotation-permitting advection. However, we have run pysteps with only advection (similar to RM-DR) and it performs slightly better than RM-DR according to some metrics (see Figures B.1 and B.2 in Appendix B), indicating that spatial and temporal scale decomposition in pysteps is not the only explanation for this difference.

Compared to the other three algorithms, RM-S attains lower correlations, with values closer to EP than to the other algorithms. Compared to RM-DR, maximum skillful lead times are generally a factor two smaller. Based on these results, it seems that using a corner tracking method for the optical flow, leads to lower correlations than using global optical flow algorithms. It was expected that EP performs worse than the other nowcasting algorithms. However, skillful lead times still reach 25 min for events with long durations. For the event duration of e.g. 1 h, on the other hand, skillful lead times are generally close to 5 min.

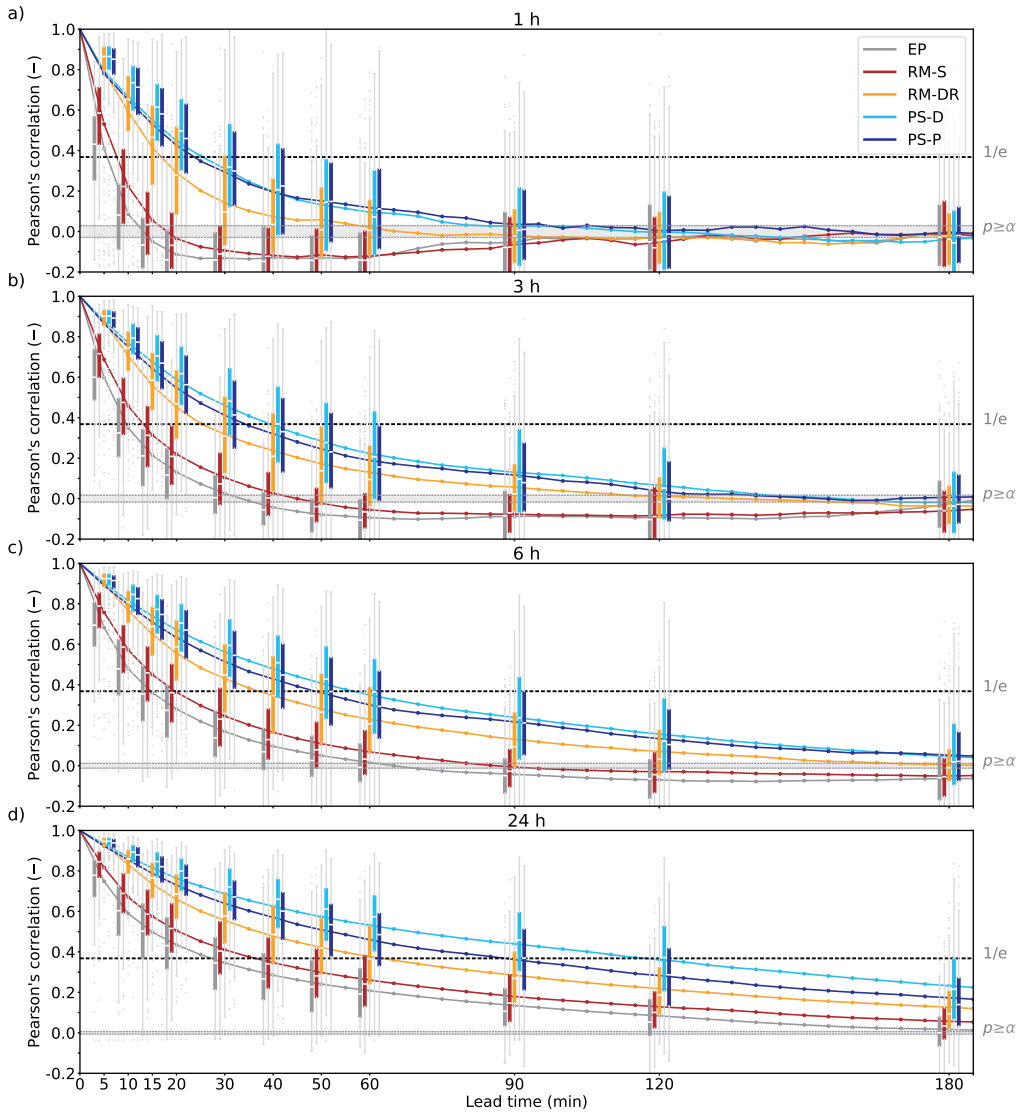


Figure 4.3 | Pearson's correlation as a function of lead time (5-min steps), averaged over all cells within the catchment and events (in that order), for event durations of 1-h (a), 3-h (b), 6-h (c) and 24-h (d). The dotted line indicates a correlation of $1/e$, the minimum correlation for a skillful nowcast. The boxes indicate the variability in results per event, with: the median in white, the interquartile (25th–75th percentile) range (IQR) in colored boxes, $1.5 \times \text{IQR}$ starting outside the boxes in grey bars and the outliers in grey dots. The horizontal grey band around a correlation of 0.0 indicates correlations that do not differ significantly from 0.0, based on a two-tailed T-test with $\alpha = 5\%$.

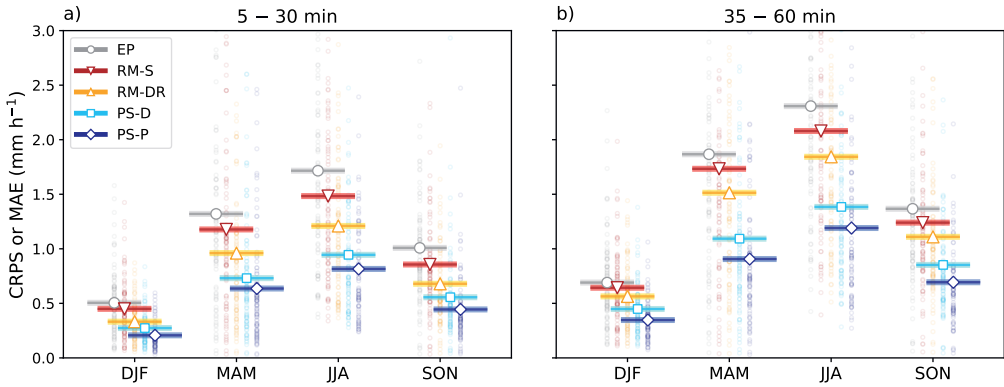


Figure 4.4 | CRPS and MAE per season for all events and catchments for the 6-h event duration, averaged over lead times of 5–30 min (a) and 35–60 min (b). The MAE is shown for all deterministic runs and the CRPS for PS-P. The thick lines with a marker indicate the mean CRPS or MAE for all runs and catchments in that season. The scattered points are the mean CRPS or MAE per event.

4.3.2 | Seasonal dependency

There are considerable differences in forecast errors between the seasons (Figure 4.4). The forecast errors are lowest during winter with event and catchment averaged MAE and CRPS values between 0.2 and 0.5 mm h⁻¹ (Figure 4.4a). Summers have the highest forecast errors with MAE and CRPS-values on average between 0.8 and 1.7 mm h⁻¹. This difference is caused by the variation in precipitation types between seasons in the Netherlands, leading to higher rainfall intensities during summer (Table 4.1), and an increase in the spatial and temporal variability of the rainfall fields. Generally, frontal systems cause the rainfall in the Dutch winter, whereas scattered convective rain showers are more dominant during summer, especially for situations with high rainfall sums. With more persistent rainfall fields in frontal systems than in convective systems, the predictability of these systems is higher, undoubtedly leading to lower forecast errors.

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

For longer lead times, i.e. 35–60 min (Figure 4.4b), the relative difference between the seasons remains the same. However, the errors increase for longer lead times with approximately 45% for all seasons, which is caused by the decreasing skill of the nowcasting algorithms for longer lead times (see e.g. Figure 4.3).

The difference between the algorithms is consistent over all seasons and lead times, with EP having the highest MAE values and PS-P always having the lowest CRPS values. This difference between highest (EP) and lowest (PS-P) error is generally a factor 2 or more. It is remarkable that the performance of PS-P is better than PS-D here, while this was the opposite in Figure 4.3. This may be caused by the bias insensitivity of the correlation metric, which is accounted for by

Table 4.3 | Indication of the maximum lead time for which an FSS of at least $0.5 + \frac{t_0}{2}$ can be attained for a set of length scales, i.e. upscaling resolutions.

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

the MAE and CRPS (for a quantification of the biases, see Figure B.2 in Appendix B).

4.3.3 | Dependency on catchment size and location

4.3.3.1 | Catchment size

Corresponding to Roberts & Lean (2008) and Mittermaier & Roberts (2010), the FSS increases with increasing length scale, i.e. after upscaling of the model simulations to a coarser resolution (Figure 4.5). Additionally, FSS decreases with increasing lead time. These relationships together give an indication of the minimum length scale required to reach a skillful forecast, i.e. $FSS \geq 0.5 + \frac{t_0}{2}$, for lead time t .

With the FSS metric in Figure 4.5, we focus on the absolute differences, i.e. biases, between forecast and observations and, due to the upscaling procedure, the FSS is less sensitive to spatial differences caused by the mislocation of forecast rainfall fields. It returns the minimum length scale for upscaling in order to reach a required skill (e.g. $FSS \geq 0.5 + \frac{t_0}{2}$), which has a hydrological relevance as this directly links to catchment sizes. Therefore, it is of interest to find out whether an $FSS \geq 0.5 + \frac{t_0}{2}$ is actually achievable for the range of catchment sizes in this chapter given a desired skillful forecast horizon (i.e. lead time t).

For instance, for a skillful forecast horizon of 60 min, Figure 4.5c shows that the event-averaged minimum length scale, indicated with a black line at $FSS = 0.5 + \frac{t_0}{2}$, is approximately 36 km for RM-DR. Hence, to still have a skillful nowcast for this lead time, the nowcast has to be upscaled to $36 \times 36 \text{ km}^2$. Upscaling to this length scale is only possible for the two largest catchments in this chapter, although the total area is already larger than the catchment areas (Figure 2.2 and Table 2.2). For the other catchments, this means that the upscaling requirement is considerably larger than taking the catchment-averaged rainfall, thus making it (on average) impossible to reach skillful forecasts of 60 min for those catchments. Note that an FSS of $0.5 + \frac{t_0}{2}$ (for a skillful forecast horizon of 60 min) is not attained with the other algorithms for any upscaling length scale up to 50 km.

Table 4.3 indicates the lead time for which an FSS of at least $0.5 + \frac{t_0}{2}$ is attained for a set of length scales (as shown in Figure 4.5). After upscaling to $10 \times 10 \text{ km}^2$, which is already larger than the catchment size of seven of the studied catchments, the maximum skillful lead times range from 21 min for EP to 37 min for RM-DR and PS-D. Above a length scale of 10 km, RM-DR outperforms all other methods. This is clearly a different perspective than found in Figures 4.3 and 4.4, and caused by the stronger bias in the nowcasts of the pysteps algorithms (not shown here, see Figure B.5 in Appendix B).

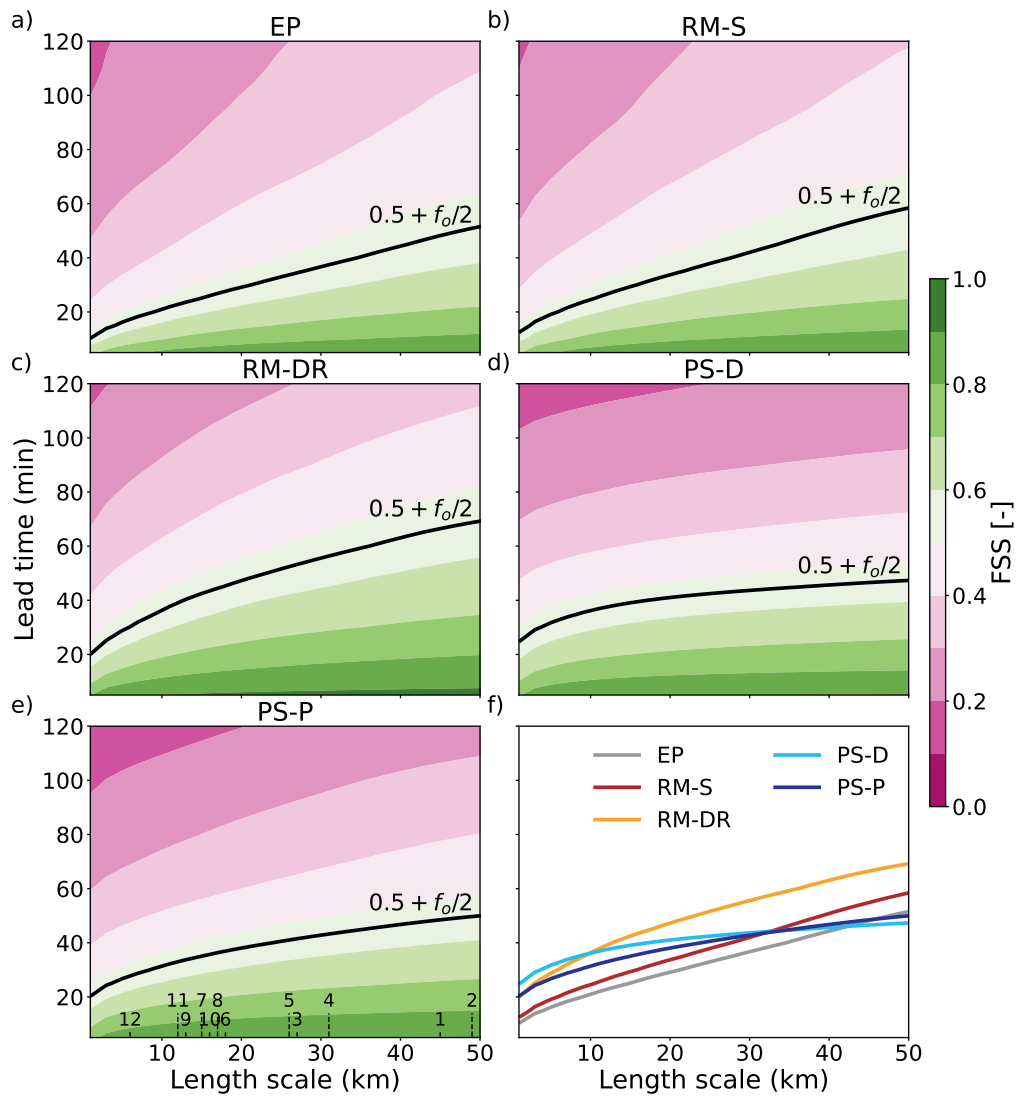


Figure 4.5 | Fractions Skill Score (FSS) as a function of lead time and catchment length scale (mean of all events). Plots are made for the 6-h durations and for a threshold of 1.0 mm h^{-1} for the 5 algorithms (a–e), based on the nowcasts for the catchments Aa and Regge. The black contour line (at $FSS = 0.5 + \frac{f_o}{2}$) indicates the minimum FSS to derive a skillful spatial scale. In panel (f), the contour lines of the algorithms are combined to facilitate comparison. In (d), the longest length scale present in all catchments is indicated with the catchment number (same as in Table 2.2).

Despite the increase of FSS with increasing length scale, a maximum FSS of 1.0 will not be attained when the forecast is biased (Roberts & Lean, 2008; Mittermaier & Roberts, 2010). All algorithms underestimate the rainfall in the forecast rainfall fields for a threshold of 1.0 mm h^{-1} (as used in Figure 4.5), especially in the presence of growth and dissipation processes during the event. However, the underestimations of the rainfall volumes for lead times exceeding 20 min are considerably higher for PS-D (see Figure B.5 in Appendix B). To a lesser extent, this is also the case for PS-P. This effect is partly caused by the dissipation of the smaller-scale rainfall fields, i.e. these fields have a shorter lifetime in the pysteps algorithms. Especially PS-D tends to end up with lower rainfall volumes due to an excess of smoothing in the forecasts. In addition, both PS-D and PS-P use probability matching, which fixes the number of wet pixels in the forecast to the number of wet pixels in the latest available observation. The rainfall pixels that have left the domain, are subtracted from this number.

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

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

4.3.3.2 | Catchment location

The catchment location with regard to the prevailing wind direction and the proximity to the upwind edge of the radar domain, matters in most cases (Figure 4.6). For the 6-h event duration, the prevailing wind direction is southwest (Figure 4.6a). This directly affects the average maximum skillful lead time of the four algorithms considered, with mean skillful lead times increasing from 20–30 min to more than 45 min in the downwind direction (sizes of the circles in Figure 4.6b), and for the northwest (Beemster) and southeast (Roggelsebeek) of the country (Figure 4.6b). The catchments located upwind are closer to the edge of the radar domain. This means that some rainfall fields are not yet present in the radar mosaic when the nowcast is issued. The available rainfall fields generally also have biased rainfall amounts, as the QPE quality deteriorates towards the edge of the domain. This inevitably leads to less skill of the nowcast with increasing lead times.

The four quarters in the circles in Figure 4.6b indicate the maximum skillful lead time for the algorithms RM-S, RM-DR, PS-D and PS-P, based on the mean of all events. Per algorithm, a similar tendency of increasing skill in the downwind direction is present. For none of the catchments, however, RM-S has a skillful forecast beyond 30 min. On the other hand, the nowcasts of PS-D and PS-P are in most cases skillful up to more than 50 min towards the (north)east of the Netherlands. The average maximum skillful lead times are higher for PS-D than for PS-P (based on 5×5 centre cells instead of the entire catchment), which corresponds to Figure 4.3.

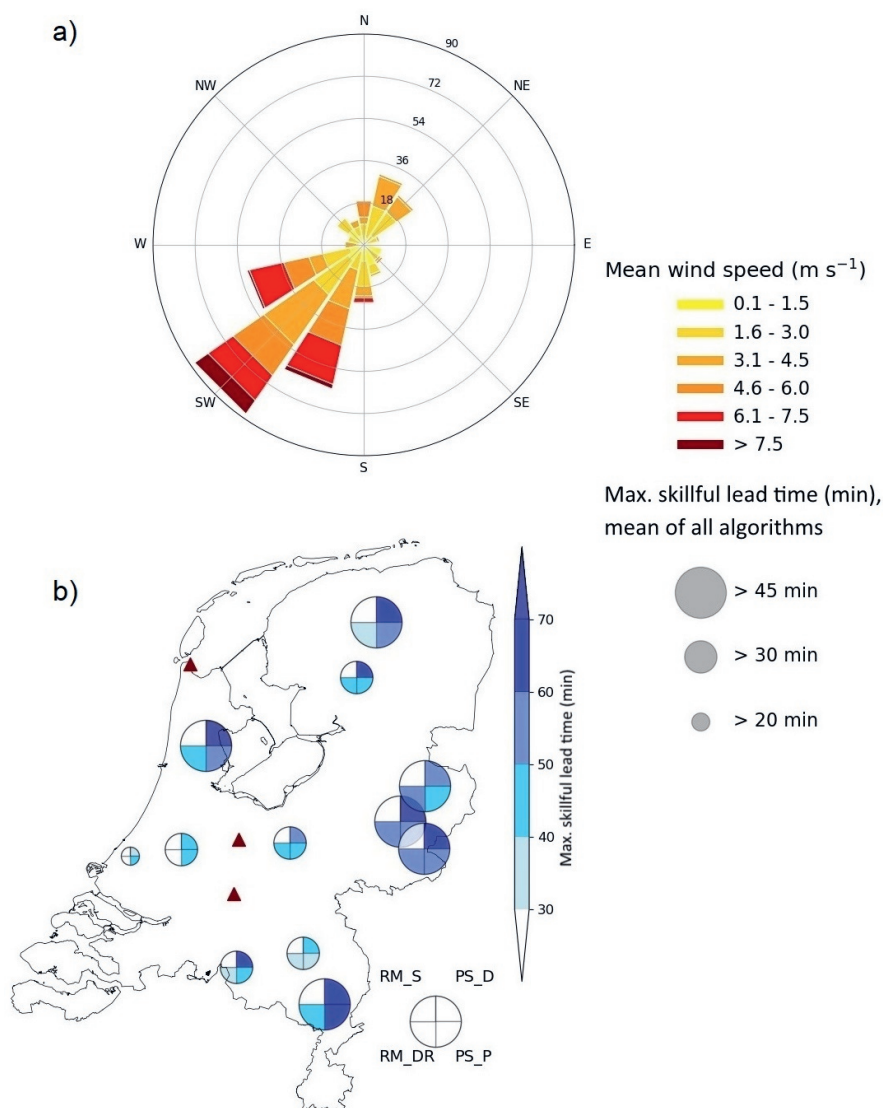


Figure 4.6 | a) Wind rose indicating the most frequent wind directions at KNMI station De Bilt during the events with the 6-h duration. The length of the bars is an indication of the number of events with that wind direction. The hue is an indication of the mean wind speed. b) The mean maximum skillful lead time of all 6-h events for the 5x5 centre cells per catchment (size of the circles indicating the average of the four algorithms) and per algorithm (hue in the quarters). EP is left out of this analysis. The red triangles indicate the locations of the radars.

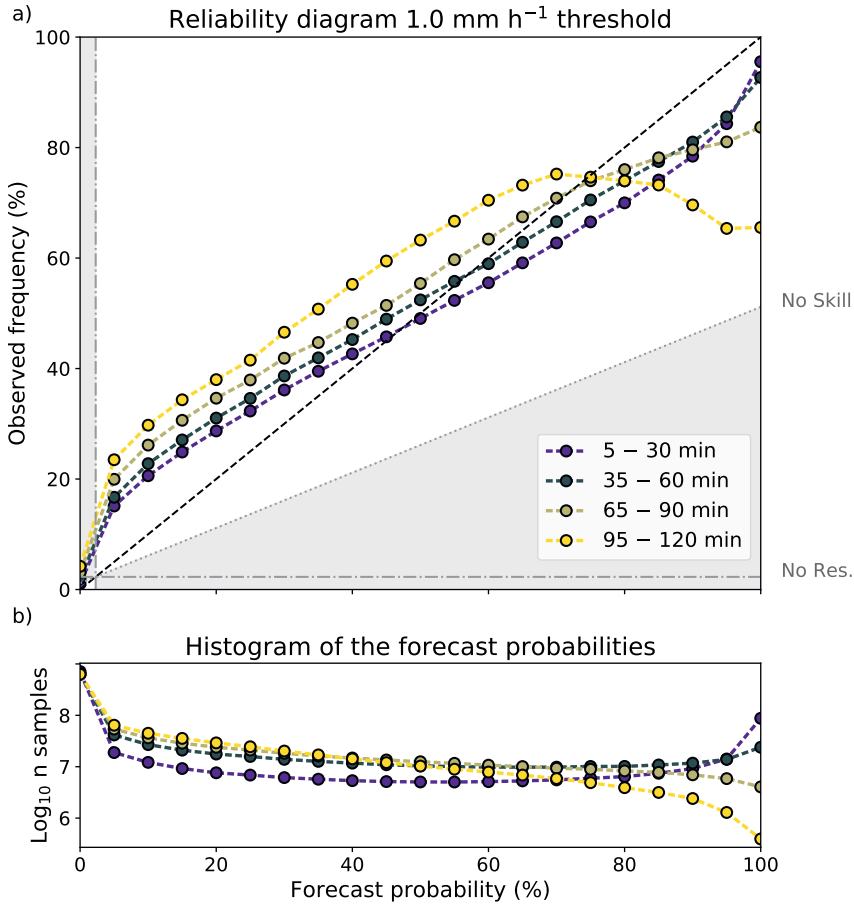


Figure 4.7 | a) Reliability diagram of exceeding a threshold of 1.0 mm h^{-1} for events within the 6-h event duration for PS-P. The reference is the climatological frequency of exceeding this threshold in the studied events. The mean probabilities for lead times of 30 min are shown. b) Histogram indicating the number of times a probability was forecast by an ensemble member.

4.3.4 | Ensemble forecast verification

The reliability diagram in Figure 4.7a illustrates that all four 30-min intervals, up to a lead time of 120 min, have a positive BSS. This means that compared to the climatological frequency of exceeding the threshold of 1.0 mm h^{-1} , all forecasts with PS-P have skill up to at least two hours ahead. However, note that the sharpness of the forecasts, the tendency to forecast with probabilities near 0 or 100%, decreases with increasing lead time (Figure 4.7b). This is especially the case for forecasts with high probabilities of exceeding the threshold, as the number of forecasts with a probability close to or at 100% reduces.

For probabilities less than 50%, the forecast probability is smaller than the observed frequency. Contrarily, the observed frequency is generally smaller than the forecast probability for probabilities exceeding 50% (70% for 5–30 min). This particular shape of the curve reveals that the

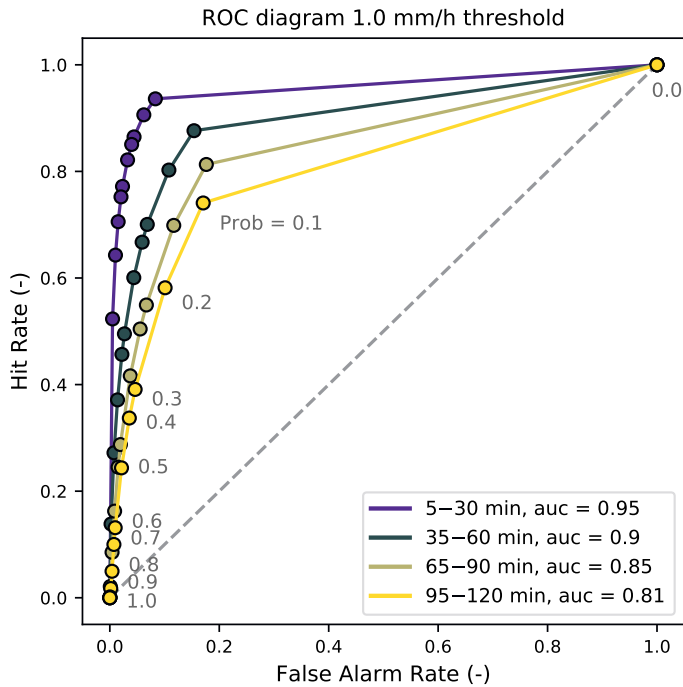


Figure 4.8 | ROC curve of exceeding a threshold of 1.0 mm h^{-1} with the nowcasts of PS-P for the events within the 6-h event duration. The numbers indicate the forecast probabilities of exceeding this threshold (-). auc indicates the area under the curve: an indication of the skill of the probabilistic forecast. Shown are the mean rates for lead times of half an hour.

ensemble is under-dispersive, meaning that the observed rainfall amount falls outside the ensemble spread of PS-P in many forecasts (see also Appendix Figure B.3). To overcome this, the ensemble should be either wider, i.e. the standard deviation of the probabilistic forecast should be larger by including more members or more noise per member, or, in case of a systematic bias, the error between ensemble mean and the observation should reduce.

The ROC curve in Figure 4.8 also indicates skillful probabilistic forecasts, with an area under the curve (auc) ranging from 0.95 (5–30 min) to 0.81 (95–120 min) for the analysed events, indicating a good discrimination skill of the ensemble for cells that exceed the threshold of 1.0 mm h^{-1} . The largest distance between the Hit Rate (HR) and False Alarm Rate (FAR), seen as the optimal forecast of exceeding the threshold, lies around a forecast probability (the circles in the graph) of 10 to 20%. Although that is an unconfident forecast, the FAR of these forecasts is generally low (smaller 0.2) and HR is always larger than 0.6 (often larger than 0.8). Forecasts with higher probabilities, i.e. more confidence, have a lower HR and FAR. Towards these higher probabilities, the difference between the shorter lead times (5–30 min) and the longest (95–120 min), is that the HR exceeds 0.5 for all forecast probabilities (the dots in Figure 4.8) for shorter lead times. However, for the longest lead times (95–120 min), the HR reduces to 0 for the highest forecast probabilities. Hence, whereas a confident forecast with a forecast probability of 100% can still be useful for a forecaster for lead times between 5 and 30 min, it becomes worthless towards

two hours ahead, as the Hit Rate is almost zero. Note, however, that hardly any nowcast has a probability larger than 50% for longer lead times.

4.4 | Discussion

4.4.1 | Relation to previous work

With 1536 events in this chapter (of which 1533 are analysed), a statistical foundation is available which allows for testing the hypothesized dependencies of nowcast skill on event duration, season, catchment location and catchment size. In this section, we explore how the findings in this chapter relate to other studies.

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

In addition, the maximum skillful lead times are based on Pearson's correlation and the intersection with the $1/e$ -line in this chapter. While this approach makes a comparison with the aforementioned literature possible, it should be noted that these maximum skillful lead times depend on the chosen metric (and the somewhat arbitrary $1/e$ -line). Both the analysis with the FSS (Figure 4.5) and the use of the Critical Success Index (CSI; Figures B.3 and B.4 in Appendix B) lead to different maximum skillful lead times and smaller differences between the algorithms, although the ranking between the algorithms remains the same when the CSI is considered. Hence, with the employed metric (Pearson's correlation) we are only able to provide an indication of the differences between the event durations. The actual skillful lead times are indicative and depend on the focus of the reader.

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

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

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

4.4.2 | The catchment perspective and resulting event selection

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

The chosen approach, however, limits the analysis of the size and location dependency to a more exploratory phase, solely indicating the presence of relations between catchment size and location, and forecast skill (section 4.3.3). It is recommended to continue with the focus on these relationships, as was also done by Foresti & Seed (2015) for a mountainous area near Melbourne in Australia. This requires using the full radar domain to find the statistics and identify these relationships on a larger domain. Understanding these relationships on this domain will make it possible to correct the nowcasts in real-time (via e.g. bias corrections or machine learning techniques) and to better take uncertainties into account. Note that such a procedure would change the event selection procedure to e.g. a national level and it would substantially increase the storage requirements for this number of events.

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

Moreover, the choice to select the events based on both catchment-averaged maxima and grid cell maxima has merely to do with the subsequent step in this project: the hydrological application

of these nowcasts. The involved water authorities that manage the studied catchments have different hydrological models and water management systems, which require either lumped or gridded rainfall input. It would have been possible to conduct this study with events based on either catchment-averaged or pixel maxima as input.

4.4.3 | Transferability of results to other regions

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

4.4.4 | Dependency on radar QPE product

The QPE product in this chapter consists of two radars with a radial extent of approximately 320 km (of which only the first 200 km is used in the composite). At this moment, an improved operational product is available (but not yet archived for a longer period), which also includes two Belgian and one German radar located relatively close to the Dutch border. The expectation is that this will increase skillful lead times, especially for catchments that are located further away in the upstream direction of the radar. In the Netherlands, that is most often towards the southwest (see section 4.3.3.2), but the expected results are of course non-exclusive to this region. Hence, the location dependency is expected to change due to this improvement.

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

4.4.4.1 | Three-dimensional data input

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

4.5 | Conclusion and future perspectives

In this chapter, the skill of radar rainfall nowcasting in predicting rainfall up to six hours ahead was tested with a large sample analysis. In total, 1536 events were run (of which 1533 successfully completed and thus were analysed) spread over four event durations (1, 3, 6 and 24 h) and four seasons for twelve lowland catchments in the Netherlands, a country with a temperate maritime climate. Four algorithms were tested and compared to Eulerian Persistence (EP), which is the “poor man’s” approach of using the most recent radar QPE as forecast. The tested algorithms were rainymotion Sparse (RM-S), rainymotion DenseRotation (RM-DR), pysteps deterministic (PS-D; similar to S-PROG) and pysteps probabilistic (PS-P) with 20 ensemble members. Model performance was assessed by a verification with the radar QPE, which was assumed to be the observed rainfall amount. The focus in this chapter was on finding the relationship between nowcast skill and dependencies on: event duration, season, catchment size and location with regard to the radar location and prevailing wind direction. In addition, the ensemble forecasts with PS-P were analysed.

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

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

Although PS-P and PS-D have shown the longest skillful lead times and the lowest forecasts errors over the seasons, most forecasts have to be upscaled for optimal use, which affects the minimal spatial scale on which the forecasts can be properly used. For all algorithms in this chapter, the forecast generally has to be upscaled in order to reach a fractions skill score (FSS) of at least $0.5 + \frac{\bar{\epsilon}_0}{2}$ (with $\bar{\epsilon}_0$ the random forecast skill), the minimal FSS for a skillful forecast. The

maximum skillful lead time that we have found after upscaling to a cell size comparable to the catchment area of the smallest catchment (Hupsel Brook, 6.5 km²) is 30 min, while this is 55 min after upscaling to a cell size comparable to the largest catchment (Regge, 957 km²). Thus, if upscaling is possible, higher skill can be attained for larger catchments than for smaller ones. For upscaling resolutions of more than 10×10 km², RM-DR has outperformed all other algorithms. This effect results from the stronger bias present in the pysteps forecasts for increasing lead times, which has a pronounced influence on the FSS. It is noteworthy that all algorithms have a bias towards lower rainfall volumes, which is not necessarily higher for pysteps than for the other algorithms for small thresholds. However, for a threshold of 1.0 mm h⁻¹ or higher, especially PS-D has (for lead times of ≥ 20 min) a considerably stronger underestimation of the rainfall volumes than the other algorithms.

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

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

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



5

Nowcasting with opportunistic rainfall data

This chapter was originally published as:

Imhoff, R. O., Overeem, A., Brauer, C. C., Leijnse, H., Weerts, A. H., & Uijlenhoet, R. (2020). Rainfall nowcasting using commercial microwave links. *Geophysical Research Letters*, 47, e2020GL089365. doi :10.1029/2020GL089365



ACCURATE and timely precipitation forecasts are crucial for early warning. Rainfall nowcasting, the process of statistically extrapolating recent rainfall observations, is increasingly used for short-term forecasting. Nowcasts are generally constructed with high-resolution radar observations. As a proof of concept, we construct nowcasts with country-wide rainfall maps estimated from signal level data of commercial microwave links (CMLs) for twelve summer days in the Netherlands. CML nowcasts compare well to radar rainfall nowcasts. Provided well-calibrated CML rainfall estimates are employed, CML nowcasts can outperform unadjusted real-time radar nowcasts for high rainfall rates, which are underestimated as compared to a reference. Care should be taken with the sensitivity of the advection field derivation to areas with low CML coverage, and the inherent measurement scale of CML data, which can be larger than the application scale. We see potential for rainfall nowcasting with CML data, for example in regions where weather radars are absent.

“Well, I’ve been afraid of changin’
‘Cause I’ve built my life around you
But time makes you bolder”

—Fleetwood Mac, *Landslide* (1975)

5.1 | Introduction

Accurate and timely precipitation forecasts are crucial for flood early warning, water management and agriculture (Ingram et al., 2002; Pappenberger et al., 2015; Thorndahl et al., 2017). The required forecast horizon depends on the application and can range from weeks to less than an hour ahead. For forecast horizons of three hours or less, rainfall nowcasts, the (statistical) process of extrapolating real-time remotely sensed quantitative precipitation estimates (QPE), are increasingly used (e.g. Ebert et al., 2004; Wilson et al., 2010; Liguori & Rico-Ramirez, 2012; Pierce et al., 2012; Foresti et al., 2016). Most nowcasts are made with radar-based QPE, but these and other high-resolution rainfall observations are not omnipresent (Lorenz & Kunstmann, 2012; Kidd et al., 2017; Saltikoff et al., 2019; WMO, 2020).

Alternative sources are needed to increase the coverage and spatio-temporal resolution of rainfall information, and to enable or improve high-resolution quantitative precipitation forecasting (QPF). A promising option is signal level data from the roughly four million commercial microwave links (CMLs) worldwide (Ericsson, 2016). These are near-ground radio connections used in cellular telecommunication networks. As these links operate at frequencies where raindrops significantly absorb and scatter radio waves (Hogg, 1968; Atlas & Ulbrich, 1977; Olsen et al., 1978), rainfall attenuates the signals between the transmitting and receiving CML antennas. Although this is a nuisance from the telecommunication perspective, rain-induced attenuation can be used to estimate path-averaged rainfall intensities (e.g. Messer et al., 2006; Leijnse et al., 2007; Zinevich et al., 2009; Overeem et al., 2011; Chwala et al., 2012; Rayitsfeld et al., 2012; Doumounia et al., 2014; Gosset et al., 2015; Uijlenhoet et al., 2018; Chwala & Kunstmann, 2019).

When two-dimensional rainfall fields are constructed from the CML data (Overeem et al., 2013, 2016b; Graf et al., 2020), (operational) nowcasting may become feasible. In this chapter, the opportunities and limitations of nowcasting rainfall with CML data are explored for the first time. Its potential is evaluated by comparing to gauge-adjusted radar rainfall data and nowcasts. Nowcasts are created with both CML and radar QPE by employing pysteps (Pulkkinen et al., 2019) in a probabilistic sense for twelve summer days. We focus on the Netherlands, as country-wide rainfall maps from CML data are already available and have been evaluated (Overeem et al., 2013). Hence, the Netherlands acts as a testbed for rainfall nowcasting with CML data.

5.2 | Data and Methods

5.2.1 | Commercial microwave link and radar rainfall estimates

Data from 1,751 CMLs covering the Netherlands for twelve days from June, August and September 2011 (Figure 5.2a), provided by T-Mobile NL, were used to construct rainfall estimates on a 1-km² spatial and 15-min temporal resolution. The data and methods are described in Section 2.2.1.5. In addition, rainfall estimates from the gauge-adjusted ('reference') and unadjusted radar datasets with a 1-km² spatial and 5-min temporal resolution, as described in Section 2.2.1, were used. In the remainder of this chapter, we refer to the gauge-adjusted radar QPE as ' R_A ' and to the unadjusted radar QPE as ' R_U '.

Nowcasts were made with all three datasets. The nowcasts with both R_A and R_U were produced at a 5-min temporal resolution. This is a finer resolution than for the CML nowcasts (15-min),

because this allowed to make full use of the data as it would have been operationally available. For a fair comparison, we evaluated the nowcasts at a 15-min interval by accumulating the nowcasts to 15-min rainfall sums. To correct for the resulting shift between rainfall patterns in consecutive maps, a significant effect when rainfall moves or develops quickly, advection correction (temporal interpolation) was applied following the method of Anagnostou & Krajewski (1999). In this procedure, the interval of the discrete temporal interpolation was 1 min and the advection vectors were estimated with the optical flow algorithm used in the nowcasts (section 5.2.2.1 and Lucas et al., 1981). The same method was applied to accumulate the QPE datasets to a 15-min resolution.

5.2.2 | Experimental setup

5.2.2.1 | Nowcasts for twelve summer days

Nowcasts were made with the rainfall estimates for the twelve summer days in 2011 (see section 5.2.1). Summer is an important period from an early warning perspective in the Netherlands, because rainfall intensities are relatively high. The rainfall estimation technique from CMLs also performs well during this period (Overeem et al., 2016b), which makes the selected period an ideal test case for summer or high intensity rainfall nowcasting.

Probabilistic nowcasts with 20 ensemble members and a forecast horizon of three hours were made with pysteps (v1.1.1; Pulkkinen et al., 2019) at the temporal resolution of the input data (5 min for radar and 15 min for CML). Pysteps and the used setup is described in more detail in Section 2.5.

5.2.2.2 | Verification approach

To assess the nowcast skill, R_A (the reference) was regarded as “truth”. Only pixels over land were considered when calculating the verification metrics, because both the CML and R_A data are limited to the land surface. Two verification methods were applied to every (15-min) forecast in the twelve events and to every separate ensemble member.

A spatial assessment took place with the CSI, as described in Section 2.6.6.1. We calculated the CSI with the forecast rainfall sum during the first hour after the nowcasts were issued (with advection correction applied for the accumulations) for two thresholds: 1.0 and 5.0 mm. Higher sums would be of interest for application in (sub-)tropical regions, but these were hardly present in the events (Appendix Figure C.1).

Overeem et al. (2016b) show that the CML QPE is generally better represented on a coarser resolution than 1 km² and we expect the same for the QPF. This holds for most forecasts, particularly due to the mislocation of rainfall fields at the highest resolution (e.g. Mittermaier, 2006). To assess and compare the nowcasting skill over successively larger aggregation scales, we used the Fractions Skill Score (FSS), which uses a fractions-based Brier score (Roberts & Lean, 2008; Jolliffe & Stephenson, 2012). The FSS is described in more detail in Equation 2.15 of Section 2.6. As for the CSI, the FSS was calculated for two intensity thresholds (1.0 and 5.0 mm h⁻¹) and three lead times (15, 30 and 60 min) with forecast rainfall amounts accumulated up to that lead time.

5.3 | Results

5.3.1 | Comparing nowcasts for one event

We initially focus on the squall line passing from southwest to northeast on 10 September 2011 (Figure 5.1), which, because of the convective character and high rainfall sum, is more exemplary of events for other climates than the Dutch temperate maritime climate. For this event, the CML QPE resembles the radar-based products (Figure 5.1) but despite similarities in the larger-scale rainfall fields, the CML QPE generally misses the fine-scale precipitation features present in the QPE of R_A (the reference) and R_U . This is a result of the lower and irregular link coverage (on average 0.22 km km^{-2} for the twelve events; Figure 5.2a), whereas the radars cover every 1 km^2 . The interpolation of the CML data to a 1-km grid results in a more smoothed display of specifically the convective cells. Rainfall volumes, however, are underestimated by the R_U QPE (on average 61% lower for these time intervals), whereas the volumes from the CML QPE are closer to the reference data for this event (4% lower).

These differences in the QPE affect the nowcasts indicated in Figure 5.2, which shows the fraction of ensemble members that exceed a threshold of (c) 1.0 and (d) 5.0 mm h^{-1} . Compared to R_A (blue contours) and for a rainfall intensity of 1.0 mm h^{-1} , the storm locations in the forecast rainfall fields from the R_U nowcasts are more accurate than those from the CML nowcasts up to at least 30 min ahead. After that forecast horizon, both nowcasts become more uncertain and more rain is missed, particularly in the newly formed rainfall cells in the southwest.

The perspective changes for a rainfall intensity of 5.0 mm h^{-1} . Due to underestimations in the R_U QPE, the nowcast misses most of the observed threshold exceedances and only captures parts of the rainfall fields during the first 15 min. The CML nowcast still captures the higher intensities with storms that are generally well-located up to at least 30 min ahead. For longer lead times, the forecast becomes more uncertain with false alarms in the west, but note that the R_U nowcast has no skill anymore at this point.

5.3.1.1 | Rainfall field advection

The nowcasts strongly depend on the rainfall field advection determined with the optical flow algorithm at the start of the nowcast (e.g. Pierce et al., 2012). The domain of the radar composite, with a radius of 200 km, provides rainfall estimates beyond the border of the Netherlands, resulting in a coherent motion field above and just outside the country (Figures 5.3b and e). As the CML data is limited to the Dutch land surface area, in combination with a lower and irregular data coverage, the derived motion fields show sharp transitions (Figures 5.3a and d). At 18:45 h, the advection vectors are smaller in the south(west) than as derived with the R_U QPE, which is partly explained by the absence of data over sea. In addition, the interpolation procedure in the CML QPE derivation, based on e.g. only eight links on the most southwesterly region Zeeuws-Vlaanderen, results in a stationary rainfall field where rainfall enters the southwest of the country. This pseudo-stationarity leads to smaller vectors in this region. A similar process takes place at 20:00 h with minimal advection in the north. This is caused by the rainfall field on one of the Wadden islands in the northwest, which advects in northeasterly direction, but is not observed southwest and northeast of the island due to the absence of links there. This apparently small difference affects the entire motion field estimates in the north of the domain.

To test whether the coarser temporal resolution of the CML QPE is a limiting factor too, motion

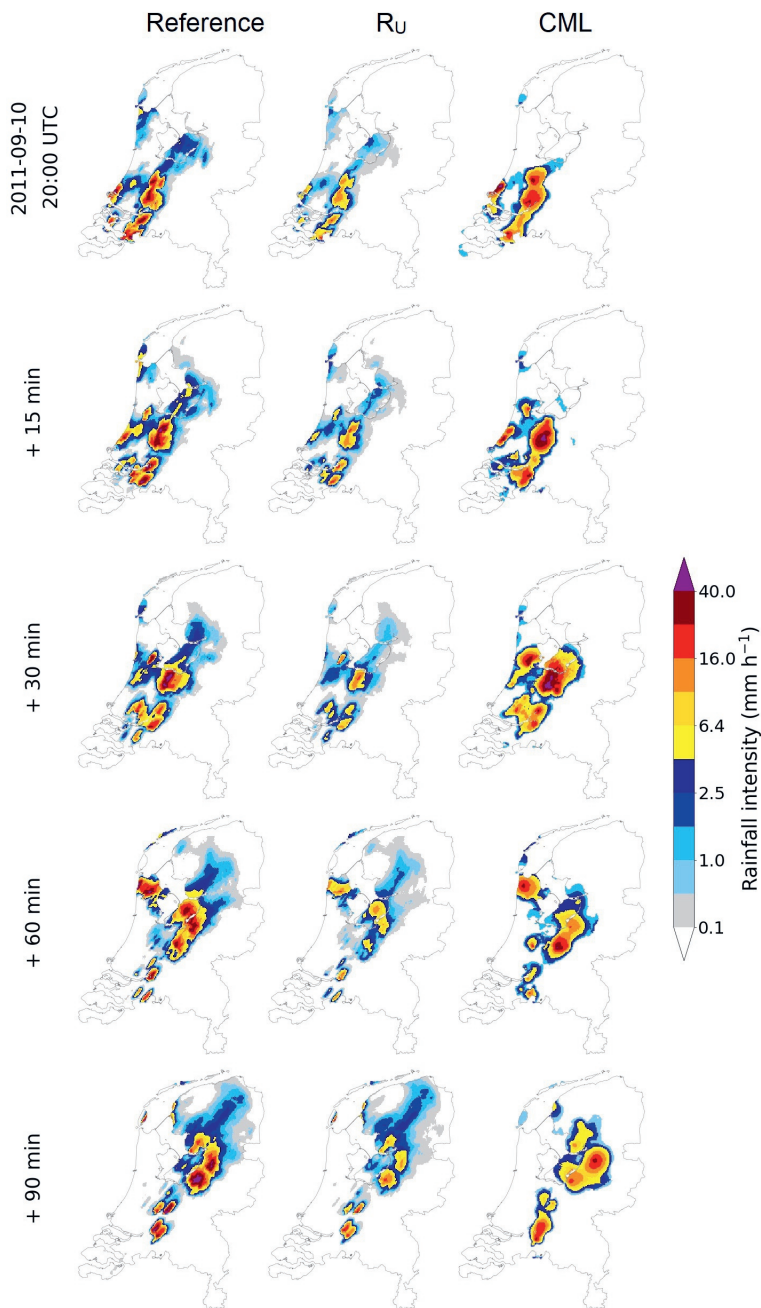


Figure 5.1 | Comparing CML estimates to radar for the Netherlands at 10 September 2011 20:00 UTC. Shown are the reference rainfall intensity (R_A), and the QPE of R_U and CML for five times.

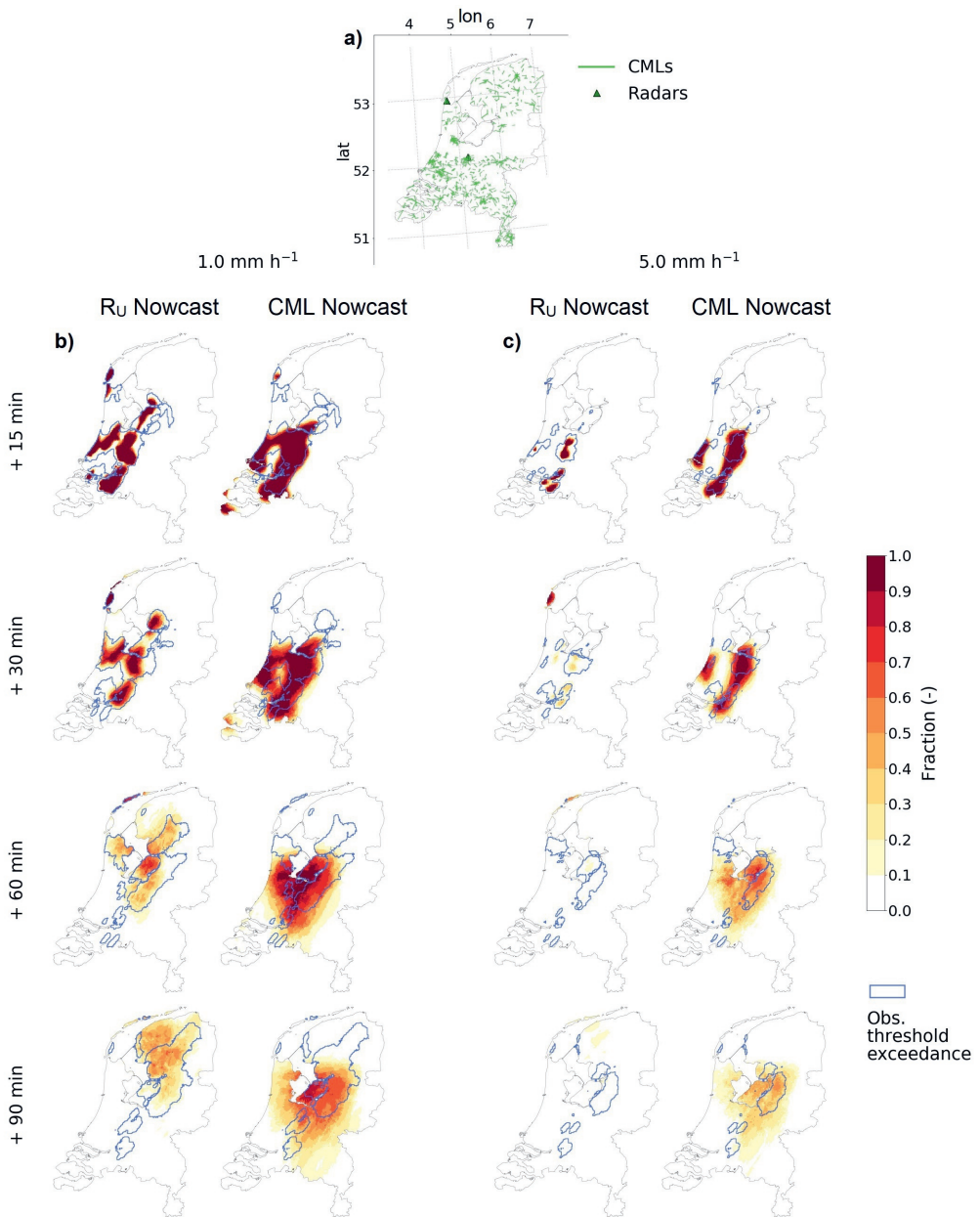


Figure 5.2 | Comparing CML nowcasts to radar rainfall nowcasts for the Netherlands at 10 September 2011 20:00 UTC, using the QPE of Figure 5.1. (a) Link coverage at the start of the nowcast and radar locations. (b–c) Probabilities of exceeding rainfall intensities of 1.0 (b) and 5.0 mm h⁻¹ (c) for four lead times, based on ensemble nowcasts created with R_U and CML QPE. Blue contour lines indicate regions where the threshold is exceeded in the reference.

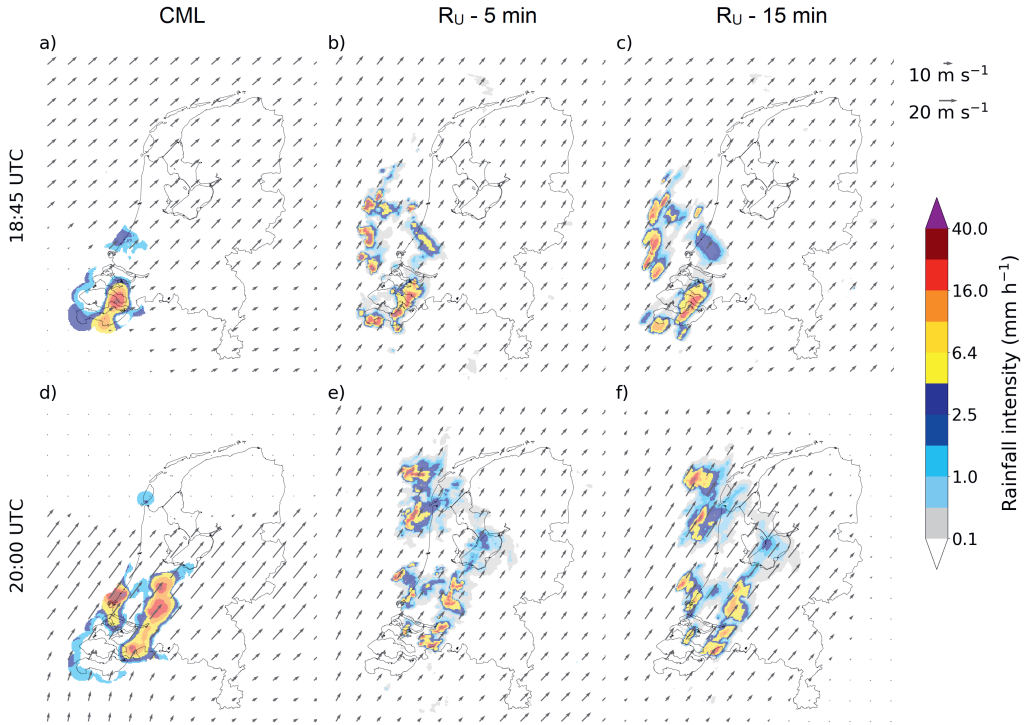


Figure 5.3 | Motion field estimates at the start of the nowcast ($t = 0$), derived with the Lucas-Kanade algorithm (QPE from $t - 3$ to t used; Lucas et al., 1981), during the event on 10 September 2011, with nowcasts starting at 18:45 UTC (a–c) or 20:00 (d–f), obtained from CML (a,d), R_U with 5-min resolution (b,e) and R_U QPE when only 15-min accumulations would have been available (c,f). In contrast to Figure 5.1, the displayed QPE is not clipped to the Dutch land surface.

fields are derived with the R_U QPE when only 15-min accumulations (and thus a 15-min interval) would have been available (Figures 5.3c and f). The differences between the optical flow fields in (b) and (c) are minimal. At 20:00 h, however, the motion in the central part of the Netherlands is $\sim 40\%$ stronger for the 15-min resolution (f) than for the 5-min resolution (e), and similar for CML (d). In addition, the observed near-absence of motion around the eastern borders (Figure 5.3f) is caused by the radial interpolation (applied in the motion field derivation) of the decreasing vector sizes in easterly directions.

Both the lower data coverage and the coarser temporal resolution of CML QPE impact the rainfall field advection. The lack of CML data over sea results in sharp transitions in the derived optical flow fields.

5.3.1.2 | Quality assessment

The average skill for all CML nowcasts (18:45 – 23:45 UTC) is comparable to the skill of the R_U nowcasts, given a 1.0 mm h^{-1} threshold and a 1-h accumulation time (Figure 5.4a and b). The rainfall fields moved in a narrow band from southwest to northeast, indicated by the blue contours (R_A). CSI scores outside these contoured regions are always lower than 0.2, as only

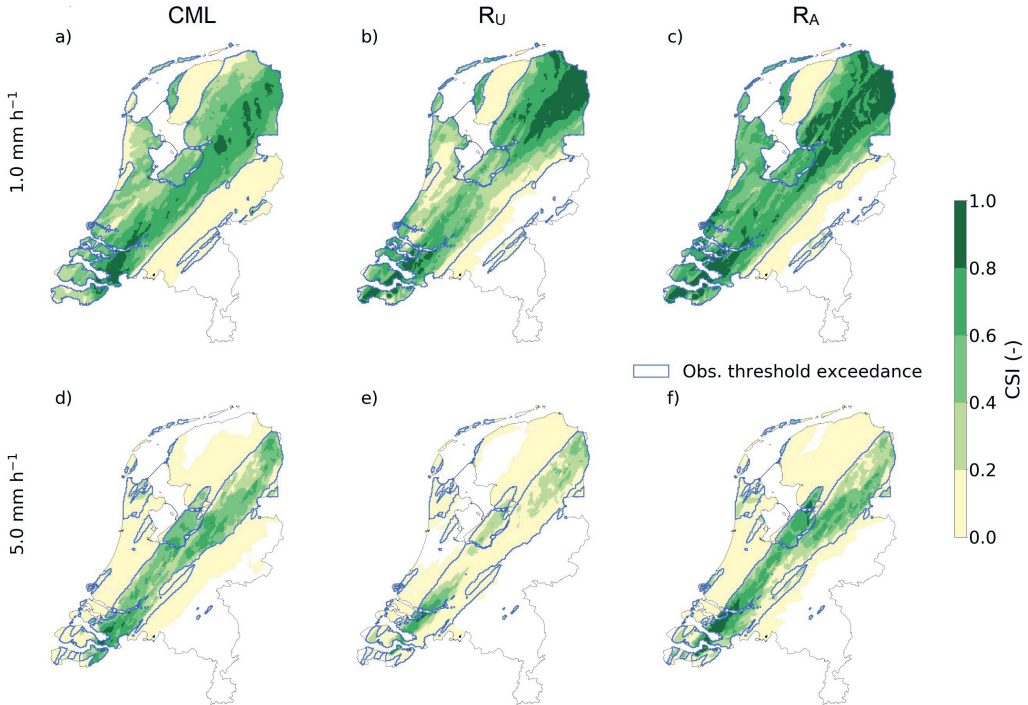


Figure 5.4 | Performance of CML and radar nowcasts. Critical success index (CSI) of forecast hourly rainfall sums for the land surface area of the Netherlands during the event of 10 September 2011, using thresholds of 1.0 (a–c) or 5.0 (d–f) mm h⁻¹, for the CML nowcasts (a,d), R_U nowcasts (b,e) and R_A nowcasts (the reference; c,f). All available nowcasts at a 15-min interval are taken into account. Blue contours indicate the threshold exceedances in the reference (R_A) QPE for all considered time intervals.

false alarms occur there. More false alarms are present in the CML nowcasts than in the other nowcasts, e.g. southeast of the rainfall band, due to more misplaced rainfall fields (Figure 5.1 and 5.2b) and differences in the estimated motion (Figure 5.3). Apart from the number of misses, CSI exceeds 0.4 in 68% (CML) and 61% (R_U) of the areas with observed threshold exceedances in the reference data. The forecasts in the northeast of the Netherlands are better for the R_U nowcasts (CSI above 0.8) than for the CML nowcasts (CSI is 0.6 – 0.8, and a minority over 0.8). Using the gauge-adjusted radar data (R_A) for the nowcasts (Figure 5.4c), leads to spatially similar nowcasts compared to the R_U nowcasts, but with much higher CSI values (generally above 0.6).

Corresponding to Figure 5.2c, the performance of the R_U nowcasts decreases significantly for a threshold of 5.0 mm h⁻¹, with CSI values below 0.4, except for small regions in the southwest and northeast (Figure 5.4e). The skill of the CML nowcasts (Figure 5.4d), however, remains comparable to that of the R_A nowcasts (Figure 5.4f). For both nowcasts, average CSI values exceed 0.6 in parts of the SW-NE oriented rainfall band, although both nowcasts have many false alarms outside this region and missed events in the northwest.

To conclude, the CML nowcasts are comparable to radar rainfall nowcasts. Whereas the R_U now-

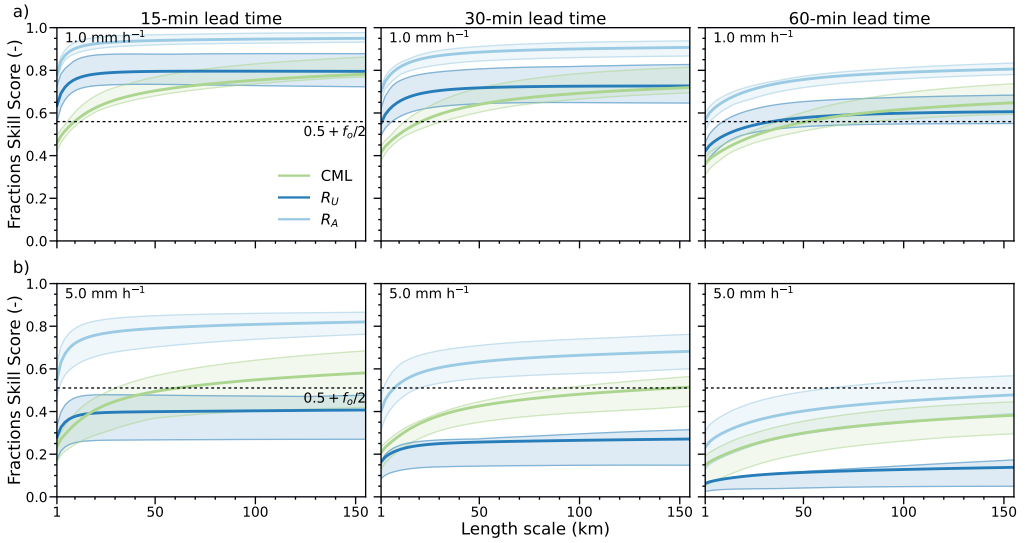


Figure 5.5 | Skill of the forecast rainfall accumulations at different spatial scales. Average FSS of all 15-min nowcasts for twelve events (thick lines) using three lead times (15, 30 and 60 min, with the rainfall accumulated over these intervals up to that lead time) from the rainfall nowcasts constructed with CML data (green), R_U (dark blue) and R_A (light blue). The FSS is calculated for two thresholds: 1.0 (a) and 5.0 mm h⁻¹ (b). Shaded areas indicate the interquartile ranges of the average FSS per event.

casts outperform the CML nowcasts for a rainfall intensity of 1.0 mm h⁻¹, the reverse holds for a 5.0 mm h⁻¹ intensity. Nowcasts for the other eleven events show similar behaviour (Appendix Figures C.2 – C.12).

5.3.2 | Nowcast verification on twelve summer events

On average, these differences remain when assessing the nowcasts for all twelve events with the FSS for three lead times (Figure 5.5; forecast rainfall amounts are accumulated up to that lead time). Analogous to the results for 10 September 2011, the R_U nowcasts outperform the CML nowcasts for a threshold of 1.0 mm h⁻¹ (Figure 5.5a), whereas the CML nowcasts generally have higher FSS values when a threshold of 5.0 mm h⁻¹ is considered (Figure 5.5b). In all cases, the FSS is highest for the 15-min lead times. The continuous ranked probability score (following Hersbach, 2000) for these nowcasts, also indicates that the maximum skillful lead time is around 15 min at the 1 km²-scale (Figure C.13 in Appendix C).

Both radar nowcasts are skillful at all spatial resolutions ($FSS > 0.5 + \frac{f_0}{2}$) for the 1.0 mm h⁻¹ threshold, when the lead time is 30 min or less. This is not the case for the CML nowcasts, with FSS values in between 0.36 and 0.46 at 1 km² for all three lead times. Upscaling the forecast leads to skillful nowcasts, e.g. for length scales of 11 km or more for the 15-min lead time. Although the radar rainfall nowcasts generally approximate an asymptotic FSS value, e.g. after a length scale of approximately 25 km for the 15-min lead time, the CML nowcast skill keeps improving with increasing resolutions, even beyond the indicated 150 km length scale. Beside the misplacement of forecast rainfall fields, as also present in the radar nowcasts, this

result can be explained by larger regions with a low density of CMLs and the resulting effective measurement scale of the CMLs, which is typically much coarser than the 1 km^2 grid the data is projected on (Overeem et al., 2016b).

For the 5.0 mm h^{-1} threshold, only the R_A and CML nowcasts are skillful, and the latter only after upscaling to length scales of 50 km or more. This is supposedly an unfeasibly large upscaling resolution for most applications in urban areas and small catchments. Such scales are, however, important for larger-scale hydrological and agricultural applications.

5.4 | Discussion and Conclusion

Nowcasts are increasingly used to forecast rainfall up to several hours ahead. The underlying data are commonly quantitative precipitation estimates (QPE) from operational radars. In this chapter, we demonstrate that nowcasts can be constructed with country-wide rainfall maps as estimated with received signal level data from commercial microwave links (CMLs) in the Netherlands. Probabilistic nowcasts were created with both CML and radar QPE using pysteps for twelve summer days in 2011. The CML nowcasts generally compare well to radar rainfall nowcasts. Compared to a reference gauge-adjusted radar dataset, we found that for low rainfall intensities, the radar rainfall nowcasts outperform the CML nowcasts due to a more coherent advection field and more detailed rainfall structures in the QPE. Most non-bias-corrected radar QPE products underestimate rainfall volumes, however, which results in many misses in the forecasts of high-intensity rainfall cells. The CML QPE and resulting nowcasts gave more accurate rainfall volumes, making this data source suitable for real-time nowcasting of high-intensity rainfall events. This result could be different in the absence of a well-calibrated CML rainfall estimation algorithm, e.g. due to a lack of a high-quality reference, or when real-time bias-adjusted radar QPE would be available.

We identify two limitations for nowcasting with CML data. First, the regionally low CML density and the inherent measurement scale of the CML data for longer links makes the 1 km^2 resolution on which the data is projected, not representative of the actual measurement scale. To attain skillful forecasts, forecasts have to be upscaled to resolutions of $11 \times 11 \text{ km}^2$ or more for most of the considered events. This may be unfeasibly large for some small scale (urban) applications. Second, the rainfall advection derivation is found to be sensitive to the more limited data coverage of the CML data. We expect that a blended system with advection estimations from for instance (satellite) remotely sensed motion fields can overcome this issue. A coarser derivation resolution than 1 km^2 also leads to have more CML information per vector, especially when this derivation scale depends on the local link density.

This chapter analysed summer rainfall in a temperate climate with a mixture of stratiform and convective rainfall. Despite some resemblance of the higher intensity showers with rainfall from e.g. (sub)tropical regions, we expect other nowcasting performances in those regions. A pilot study on CML rainfall estimation in Burkina Faso, Nigeria, São Paulo and Sri Lanka successfully demonstrated the potential of this technology (Doumounia et al., 2014; Rios Gaona et al., 2018; GSMA, 2019). Potentially, this makes rainfall nowcasting feasible for areas without radar coverage but with high CML density, like developing regions. A prerequisite for operational use is the real-time availability of CML data. Recent acquisition method developments by (Chwala et al., 2016) and the project (TEL4RAIN, 2020) are promising in this direction. Moreover, the

introduction of 5G cellular networks can lead to new opportunities as E-band CMLs are an essential part of these networks (Ericsson, 2019; Fencil et al., 2020). Provided that the nowcasts are skillful on the relevant spatial resolutions and lead times, CML nowcasts can provide an opportunity to reduce fatalities and economic loss, e.g. by improving hazardous weather and (flash-)flood early warning(s). To achieve this, the CML QPE and QPF need to be optimized and evaluated on datasets with a high density of CMLs covering entire (rainy) seasons. This is part of ongoing efforts.

A stylized illustration featuring a large, grey, fluffy cloud at the top. Below the cloud, several vertical, light blue-grey lines represent rain falling against a solid teal background. In the center of the image, a large, white, bold number '6' is superimposed over the rain and the lower part of the cloud.

6

Blending nowcasts and numerical weather prediction to extend skillful lead times

This chapter is submitted to Quarterly Journal of the Royal Meteorological Society as:

Imhoff, R. O., De Cruz, L., Dewettinck, W., Brauer, C. C., Uijlenhoet, R., van Heeringen, K.-J., Velasco-Forero, C., Nerini, D., Van Genderachter, M., & Weerts, A. H. (2022). Scale-dependent blending of ensemble rainfall nowcasts and NWP in the open-source pysteps library, *submitted to Quarterly Journal of the Royal Meteorological Society*

THE first few hours ahead in rainfall forecasting are crucial for (flash) flood early warning. As this time scale is not sufficiently well captured by the rainfall forecasts of numerical weather prediction (NWP) models, radar rainfall nowcasting can provide an alternative. Because this observation-based method quickly loses skill after the first two hours of the forecast, it needs to be combined with NWP forecasts to extend the skillful lead time of short-term rainfall forecasts. We implemented an adaptive scale-dependent ensemble blending method in the open-source `pysteps` library, based on the STEPS scheme. In this implementation, the extrapolation (ensemble) nowcast, (ensemble) NWP and noise components are blended with skill-dependent weights that vary per spatial scale level. To constrain the (dis)appearance of rain in the ensemble members to regions around the rainy areas, we have developed a Lagrangian blended probability matching scheme and incremental masking strategy. We describe the implementation details and evaluate the method using three heavy and extreme (July 2021) rainfall events in four Belgian and Dutch catchments. We benchmark the results of the 48-member blended forecasts against the Belgian NWP forecast, a 48-member nowcast and a simple 48-member linear blending approach. Both on the radar domain and catchment scale, the introduced blending approach predominantly performs similarly or better (in terms of event-averaged CRPS and CSI values) than the other three tested methods, although the difference, particularly with, the linear blending method reduces when we focus on catchment-average cumulative rainfall sums instead of instantaneous rainfall rates. By properly combining observations and NWP forecasts, blending methods such as these are a crucial component of seamless prediction systems.

“Aimer, ce n’est pas se regarder l’un l’autre,
c’est regarder ensemble dans la même direction.”

—Antoine de Saint-Exupéry, *Terre des hommes* (1939)

6.1 | Introduction

Intense precipitation events can lead to disruptive (pluvial) floods. The persistent mesoscale low-pressure system in Northwestern and Central Europe in July 2021, which locally resulted in extreme rainfall amounts that led to severe (flash) flooding, is an example of this. The floods caused over 240 casualties, of which most in Belgium and Germany, and led to more than 25 billion USD in economic and infrastructural damages (AON, 2021; Koks et al., 2021; Kreienkamp et al., 2021). The disruptive effects of (flash) flooding can be reduced when there is a timely anticipation of the approaching flood, which is possible when there is a well-established flood early warning system in place (UNISDR, 2002; Pappenberger et al., 2015). Such early warning systems are only beneficial if the underlying rainfall forecasts are reliable and rapidly available. However, intense rainfall events that occur at small spatio-temporal scales, are difficult to forecast. As this is the spatial and temporal scale at which flash floods take place, typically in small urban and mountainous catchments, improving short-term rainfall forecasting is a crucial step to ensure timely and adequate response to flood risk through early warning systems (e.g. Cox et al., 2002; Ferraris et al., 2002).

If a regional or national water management authority has an early warning system in place, the underlying precipitation forecasts will generally be based on short-range (12–72 h) numerical weather prediction (NWP) model forecasts. Although NWP models are continuously improving, issuing timely and reliable rainfall forecasts, along with the necessary assimilation steps in NWP models, at the short time scales of flash floods (of the order of 6 h) remains challenging. Because NWP models are computationally expensive, they are either run on a too coarse temporal resolution (e.g. hourly or coarser) or at a too low update frequency (e.g. every six hours) for usage in flash flood early warning systems. In addition, most operational NWP systems have a latency of several hours between model initialisation and delivery at the end user. Consequently, the timing and location of intense rainfall events are often missed (Lin et al., 2005; Roberts & Lean, 2008; Berenguer et al., 2012; Pierce et al., 2012).

One way to tackle this problem is the use of rainfall nowcasting techniques, which (statistically) extrapolate real-time remotely sensed quantitative precipitation estimates (QPEs) into the future. Rainfall nowcasting allows us to take advantage of the high spatial and temporal resolutions of remotely sensed data (for instance, 1 km² and 5 min for the QPE of current weather radars, Serafin & Wilson, 2000; Overeem et al., 2009b). In addition, its initial conditions are always equal to the most recent observations, which makes it useful for flood forecasting purposes (Berenguer et al., 2005; Pierce et al., 2005; Sharif et al., 2006; Vivoni et al., 2006, 2007; Germann et al., 2009; Liguori & Rico-Ramirez, 2012, 2013; Moreno et al., 2013; Poletti et al., 2019; Heuvelink et al., 2020; Imhoff et al., 2022).

At present, a large number of nowcasting algorithms is available, which can be categorised in field-based nowcasting methods (e.g., Bowler et al., 2006; Seed, 2003; Seed et al., 2013; Berenguer et al., 2011; Sokol et al., 2017; Ayzel et al., 2019b), object-oriented methods (e.g., Dixon & Wiener, 1993; Han et al., 2009), analogue-based methods (e.g., Atencia & Zawadzki, 2014, 2015; Zou et al., 2020) and machine-learning methods (e.g., Foresti et al., 2019; Ravuri et al., 2021). More recently, the nowcasting field has been progressing towards more community-driven, free and open-source software, with pysteps as an example of this (Pulkkinen et al., 2019). Since its release, the pysteps community has grown rapidly and the framework now includes more nowcasting

approaches, including those of Hering et al. (2006), Nerini et al. (2017a) and Pulkkinen et al. (2020, 2021).

One of pysteps' main features is an efficient Python implementation of the probabilistic field-based nowcasting scheme STEPS and its deterministic predecessor S-PROG (originally in C++; Bowler et al., 2006; Seed, 2003; Seed et al., 2013). This method considers the dynamical scaling of the rainfall predictability by decomposing rainfall fields into a multiplicative cascade, representing different spatial scales (see also: Lovejoy & Schertzer, 1995; Marsan et al., 1996; Harris et al., 1996; Foufoula-Georgiou, 1998; Seed et al., 1999). By applying different spatially and temporally correlated stochastic perturbations for each spatial scale to a deterministic extrapolation nowcast, pysteps generates an ensemble of rainfall forecasts. The result is that pysteps allows large-scale features to evolve more slowly than small-scale features, which ensures an appropriate representation of uncertainty associated with the growth and dissipation of rainfall.

Despite this representation of the uncertainty associated with growth and decay of rainfall, pysteps, and other nowcasting methods, quickly lose skill after the first two to three hours. This maximum skillful lead time of the forecast depends on the type and scale of the precipitation system, with only 30 min for small-scale convective rainfall events, 2 h for larger-scale, more persistent rainfall events and up to 6 h for continental-scale persistent stratiform events (Germann & Joss, 2002; Germann et al., 2006; Lin et al., 2005; Berenguer et al., 2011, 2012; Liguori & Rico-Ramirez, 2012; Foresti et al., 2016; Mejsnar et al., 2018; Ayzel et al., 2019b; Imhoff et al., 2020a).

To extend the skillful lead time to the time scale of flash floods and improve early warnings as a result, we have to bridge the gap between nowcasting and short-range NWP model forecasts. Alongside the recent developments of improving NWP and nowcasting techniques, it is necessary to combine the two products, so-called blending, in order to obtain seamless predictions (Sun et al., 2014). There is a plethora of blending techniques present (e.g. Golding, 1998; Bowler et al., 2006; Atencia et al., 2010; Kober et al., 2012, 2014; Bailey et al., 2014; Nerini et al., 2019; Yoon, 2019; Radhakrishnan & Chandrasekar, 2020; Vannitsem et al., 2021), but none are available in a widely-used open-source nowcasting framework. The pysteps initiative has demonstrated that such an open-source implementation can accelerate collaborations and future developments, which would justify a similar approach concerning blending of nowcasting and NWP.

Therefore, we have implemented an adaptive, scale-dependent blending in pysteps based on earlier work in the STEPS scheme (Bowler et al., 2006; Seed et al., 2013). In this blending implementation, the blending of the extrapolation nowcast, NWP and stochastic noise components is performed at different spatial scales using varying blending weights per cascade level. Here, we describe the implementation of the STEPS ensemble blending approach in the pysteps framework, together with some new functionalities. We test the method on three heavy rainfall events in 2021 that led to discharge peaks, in the case of the July 2021 event even to widespread disastrous flooding, in the Belgian and Dutch catchments Vesdre, Demer, Geul and Dommel, with a focus on both the national (the entire radar domain) and catchment scale. We benchmark the results against the Belgian NWP rainfall forecasts (which covers the entire study area and has a 5-min temporal resolution), ensemble nowcasts with pysteps and a simple linear blending between the former two.

6.2 | Scale-dependent blending framework

This section describes the implementation of the STEPS blending approach in the existing pysteps framework. The description is limited to the main procedures used to construct a blended STEPS forecast in pysteps and to functionalities that were added in this study. For more information regarding specific STEPS or pysteps functionalities, we refer to Bowler et al. (2006), Seed et al. (2013) and Pulkkinen et al. (2019).

Figure 6.1 gives an overview of the workflow to compute a blended precipitation forecast in pysteps. In principle, this implementation blends, per ensemble member, an extrapolation now-cast component with either a deterministic or an ensemble NWP forecast and a noise component. First, the extrapolation and NWP components are decomposed in multiplicative spatial cascades, of which each level captures the features at the corresponding spatial scale (Section 6.2.1). The blending takes place level-by-level, meaning that the weights of the different components are specific to each cascade level (Section 6.2.3.2). The scale-dependent blending weights are computed from the recent skill of the forecasts components, and converge to a climatological value (see Section 6.2.3.1), meaning that the blending weights vary both per spatial scale and in time (per issue time, but also over the forecast horizon). After the blending of the components per cascade level (Section 6.2.5), the cascade levels are recomposed to one blended forecast and some post-processing steps take place (Section 6.2.6), resulting in a blended ensemble forecast for that issue time.

6.2.1 | Constructing the cascades for the three components

Pysteps contains an import module that allows for importing radar composites from various meteorological organizations. This functionality has been extended with a separate import module for NWP forecasts. As these forecasts are generally on a different (coarser) spatial resolution and spatial projection than the radar data, a reprojection has to take place prior to the blending procedure. Therefore, we have implemented a reprojection module that reprojects the NWP forecasts on to the spatial projection of the radar data with an affine transformation and that, if necessary, downscales the forecast to the radar grid by means of a nearest neighbour approach (more interpolation methods are available).

Once all data is imported, pysteps thresholds and transforms the data. Pysteps contains multiple transformation methods, but here, we only focus on the dBR transform (Pulkkinen et al., 2019):

$$\text{dBR}_{i,j}(t) = \begin{cases} 10\log_{10} R_{i,j}(t) & \text{if } R \geq 0.1 \text{ mm h}^{-1} \\ -15 & \text{otherwise} \end{cases}, \quad (6.1)$$

with $R_{i,j}$ the precipitation rate in mm h^{-1} at grid cell i,j and time t . $\text{dBR}_{i,j}(t)$ is -15 for precipitation intensities of less than 0.1 mm h^{-1} (this is adjustable) to ensure a clear rain - no rain transition for the nowcasting step. The transformation ensures that the precipitation data has a near-Gaussian distribution, which is needed for the stochastic processes that assume Gaussianity. Finally, the transformed fields are used to determine the motion fields of the radar observations and the NWP forecast(s) with one of the optical flow methods in pysteps (see Pulkkinen et al., 2019). These motion fields are stored and later on used for the extrapolation of the cascades in Section 6.2.4.

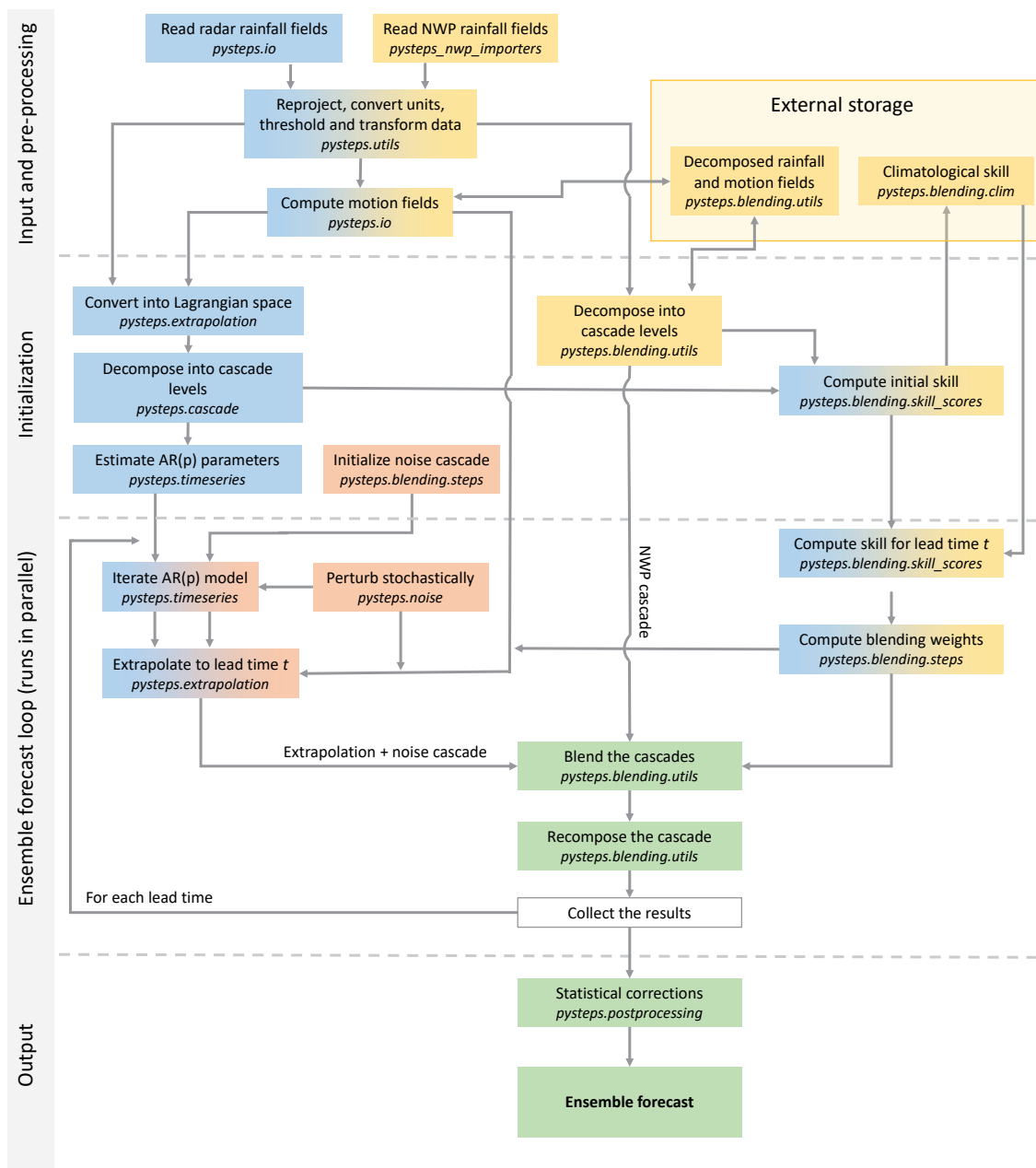


Figure 6.1 | Schematic overview of the workflow for computing blended precipitation forecasts using pysteps. For each chart element, the bottom row in italic describes the pysteps module used to execute the described task. Blue colors represent the elements that are part of the extrapolation cascade, red colors those of the noise cascade and yellow those of the NWP model cascade. An overlap in colors indicates that the process in the chart element is used for multiple cascades (for instance, ‘compute motion fields’ is performed for both the extrapolation and NWP model cascade). Finally, green colors indicate the merged cascades up to the final output.

A property of precipitation is that its lifetime exhibits power-law scaling with respect to spatial scale (e.g. Venugopal et al., 1999; Seed, 2003; Germann et al., 2006), which was used in S-PROG (Seed, 2003) and later in STEPS (Bowler et al., 2006; Seed et al., 2013) as main motivation to decompose precipitation fields into a multiplicative cascade. The levels in this cascade represent different spatial scales, which are treated differently in the nowcasting scheme. An advantage of the log-transformation in Equation 6.1 is that this decomposition into a multiplicative cascade becomes an additive cascade in the log-transformed space (Seed, 2003). With a fast Fourier Transform (FFT), the precipitation field is decomposed in such an additive cascade, which is then filtered with Gaussian weight functions (Pulkkinen et al., 2018) in k cascade levels representing different spatial scales, which are normalised afterwards:

$$\text{dBR}_{i,j}(t) = \sum_{k=1}^K \sigma_k(t) Y_{k,i,j}(t) + \mu_k(t) . \quad (6.2)$$

Here, σ_k is the standard deviation and $\mu_k(t)$ the mean of level k (out of K cascade levels). $Y_{k,i,j}(t)$ represents the spatial variability in the original precipitation field for grid cell i,j and is one of the (radar) extrapolation, NWP or noise components throughout the blending procedure. $Y_{k,i,j}(t)$ has zero mean and a standard deviation of 1.0. This decomposition is performed for both the radar data and NWP forecast(s). The noise component has the same dimensions and K cascade levels as the radar and NWP components, and will be represented by spatially and temporally correlated stochastic perturbations during the forecast (see Section 6.2.2).

Once the radar data has been decomposed, the derived motion fields are used to advect the past radar observations to the current time step in order to have all radar observations in Lagrangian coordinates. The decomposed NWP forecast and motion fields can be stored and reopened with pysteps, which saves calculation time when the NWP data has an update frequency that is lower than the update frequency of the blended nowcasts (for instance, a 6-h versus a 5-min update frequency).

After the aforementioned steps, three cascades are present which represent the extrapolation, NWP and noise components and which will be blended with a set of blending weights per cascade level k . We will elaborate on this in Section 6.2.5.

6.2.2 | Temporal evolution of the extrapolation and noise cascades

To estimate the change of the precipitation fields in the extrapolation and noise cascades over time, pysteps simulates the temporal evolution of these fields over time with a p^{th} -order (generally second-order) auto-regressive (AR(p)) process per cascade level. This AR(p) process injects a stochastic perturbation term which represents the uncertainty in growth and decay of the precipitation field over time. In pysteps, this stochastic term is added to the extrapolation component, avoiding the need of a separate noise component (Pulkkinen et al., 2019). The presence of a separate noise component in the blending approach (Bowler et al., 2006) makes the temporal evolution of the extrapolation and noise cascades two separate processes, where the extrapolation cascade regresses without added noise, as:

$$Y_{k,i,j}^{\text{ext}}(t + t_l) = \sum_{p=1}^{p_{\max}} \phi_{k,p} Y_{k,i,j}^{\text{ext}}(t + t_l - p\Delta t) , \quad (6.3)$$

and the noise cascade regresses with added noise, according to:

$$Y_{k,i,j}^{\epsilon}(t + t_l) = \sum_{p=1}^{p_{\max}} \phi_{k,p} Y_{k,i,j}^{\epsilon}(t + t_l - p\Delta t) + \phi_{k,0} \epsilon_{k,i,j}(t + t_l) . \quad (6.4)$$

In these equations, Y^{ext} and Y^{ϵ} represent the extrapolation and noise cascades at cascade level k and lead time $t + t_l$, p is the AR-order, Δt the internal time step (generally the time interval between two consecutive radar observations), $\phi_{k,p}$ are parameters that control the rate of temporal evolution at cascade level k and for order number p (determined from the initial radar observations, see Pulkkinen et al., 2019, for the derivation of these parameters), and $\epsilon_{k,i,j}(t)$ represents the perturbation field at cascade level k . This perturbation field is a correlated Gaussian random field, that is constructed using FFT filtering, which ensures that the noise field has the desired correlation structure (for more information and the available filtering methods, see Schertzer & Lovejoy, 1987; Pegram & Clothier, 2001; Bowler et al., 2006; Pulkkinen et al., 2019).

6.2.3 | Blending weights

6.2.3.1 | Initial skill and skill per lead time

STEPS bases the blending weights on the real-time skill (Pearson's correlation) of the extrapolation and NWP components. As the forecast lead time advances, the weights increase for the noise component, while they decrease for the extrapolation component. The NWP skill regresses towards climatological values during the forecast. This real-time skill-based blending procedure avoids the need for a parameterization of the blending process and weights determination, and ensures that the blending weights represent the real-time state of the components that are blended (Bowler et al., 2006).

The AR-(p) model (Section 6.2.2) determines the skill decrease of the extrapolation component per cascade level k as follows (Bowler et al., 2004):

$$\rho_k^{\text{ext}}(t + t_l) = \sum_{p=1}^p \phi_{k,p} \rho_k^{\text{ext}}(t + t_l - p\Delta t), \quad (6.5)$$

with t the issue time of the forecast, t_l the lead time and starting value

$$\rho_k^{\text{ext}}(t) = 1 . \quad (6.6)$$

Approximating the evolution by an AR(2)-process yields (Hamilton, 1994):

$$\rho_k^{\text{ext}}(t + \Delta t) = \frac{\phi_{k,1}}{1 - \phi_{k,2}} . \quad (6.7)$$

The NWP skill per lead time is based on an initial skill, which is the Pearson correlation at cascade level k between the most recent radar observation and the corresponding NWP forecast for that time ($\rho_k^{\text{nwp}}(t)$), and regresses toward a climatological value (Bowler et al., 2004):

$$\rho_k^{\text{nwp}}(t + \Delta t) = q_k^{\text{nwp}} \rho_k^{\text{nwp}}(t) + (1 - q_k^{\text{nwp}}) \overline{\rho_k^{\text{nwp}}}, \quad (6.8)$$

with

$$q_k^{\text{nwp}} = e^{-t_l/L_{1,k}} (2 - e^{-t_l/L_{2,k}}). \quad (6.9)$$

In these equations, $L_{1,k}$ and $L_{2,k}$ are coefficients that represent the decorrelation times of the NWP forecast skill estimates per cascade level. We have not adjusted these coefficients here, but estimates can be found in Table 4 in Bowler et al. (2004). $\overline{\rho_k^{\text{nwp}}}$ is the climatological skill value toward which Equation 6.8 regresses.

In Bowler et al. (2004, 2006), $\overline{\rho_k^{\text{nwp}}}$ are fixed values based on an analysis of NWP forecasts for the UK during April 2003. As the NWP forecast skill varies over time as a function of prevailing precipitation type (stratiform or convective, e.g. Mittermaier et al., 2013; Prakash et al., 2016), these fixed values are not always representative of the NWP skill over the forecast horizon. Therefore, we have implemented a module in pysteps that computes the climatological skill based on a multi-day moving window, which can be adjusted to the (operational) needs of the user. At every issue time (for instance, every 5 min), the current skill of the NWP forecast, as derived with the most recent radar observation, is stored (note that a negative correlation is regarded as zero) and at the end of the day a day-average skill is calculated. Subsequently, the climatological skill at a given issue time is the daily average skill over the number of (past) days in the moving window.

6.2.3.2 | Weights determination

The three components are blended per spatial cascade level, which is performed with a weighted sum of the three components (in log space). These weights vary over time, as a function of both the initial skill and skill per lead time of the NWP and extrapolation components. STEPS comes with two blending methods, introduced by Bowler et al. (2006) and Seed et al. (2013), which we have both added to pysteps to allow users to choose the ideal method for their case. Below, we describe the principle of both blending methods and show the difference in resulting weights for a test case on 2021-06-29 13:30 UTC in Figure 6.2. In Sections 6.3.3.3 and 6.4.3 we describe the evaluation and effect of both methods on the resulting rainfall forecast for this case.

Bowler et al. method The method introduced by Bowler et al. (2006) assumes that the sum of the squared weights equals one (implying that the sum of the weights can exceed one) and that the three cascades are uncorrelated. The weights depend on the current and expected skill of the extrapolation and NWP components (see Equations 6.5 and 6.8) and are calculated per lead time (t_l) as:

$$w_k^{\text{ext}}(t + t_l) = \rho_k^{\text{ext}}(t + t_l) \sqrt{\frac{\lambda_k^{\text{ext}}(t + t_l)}{\lambda_k^{\text{ext}}(t + t_l) + \lambda_k^{\text{nwp}}(t + t_l)}}, \quad (6.10)$$

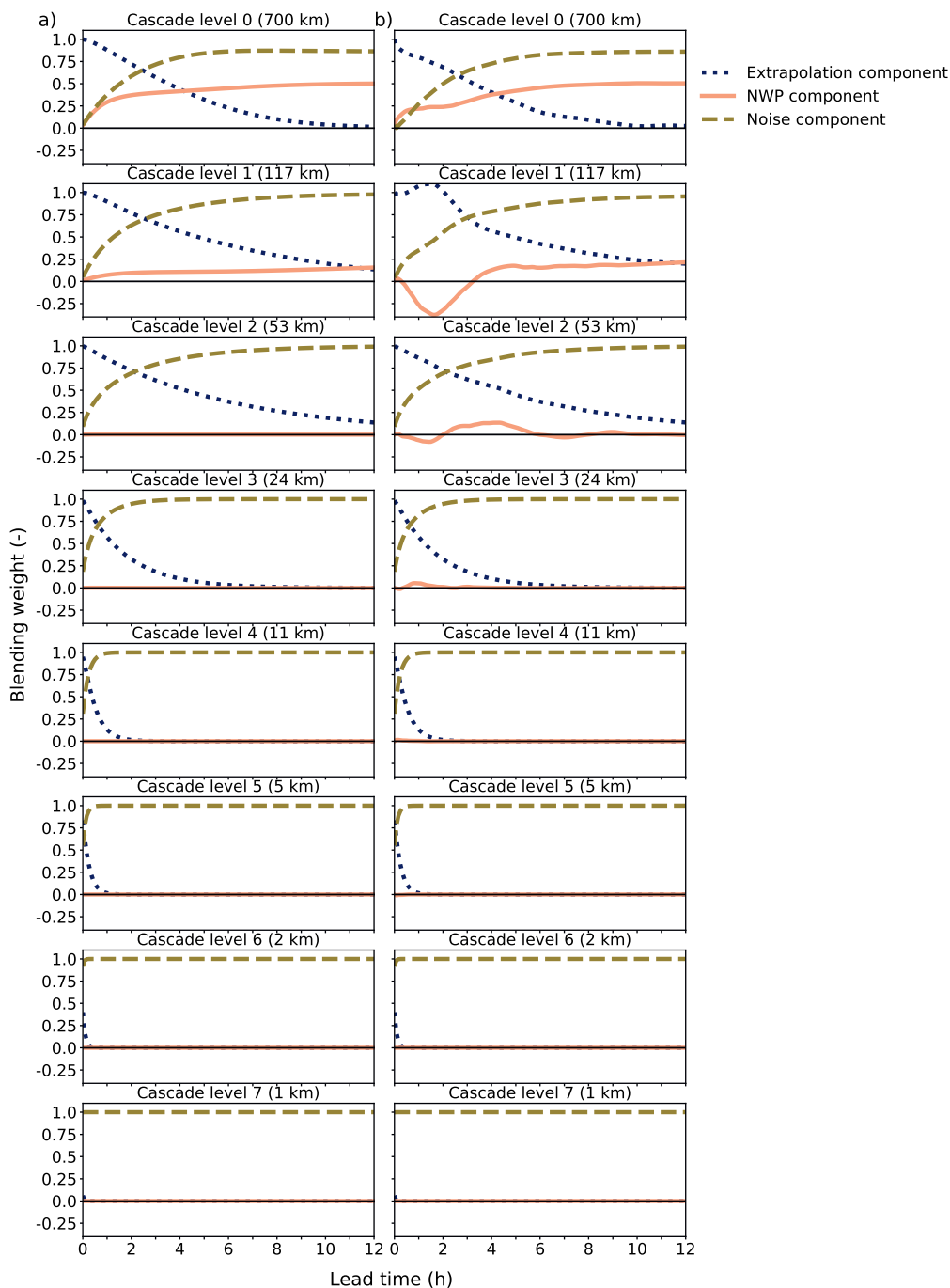


Figure 6.2 | Resulting blending weights per cascade level for the test case of 2021-06-29 13:30 UTC, when using the approaches of (a) Bowler et al. (2006) and (b) Seed et al. (2013). The blue lines correspond to the blending weights of the extrapolation component, the pink lines to the NWP component and the olive green lines to the noise component. The (approximate) corresponding spatial scale per cascade level is indicated above every sub-panel.

$$w_k^{\text{nwp}}(t + t_l) = \rho_k^{\text{nwp}}(t + t_l) \sqrt{\frac{\lambda_k^{\text{nwp}}(t + t_l)}{\lambda_k^{\text{ext}}(t + t_l) + \lambda_k^{\text{nwp}}(t + t_l)}} , \quad (6.11)$$

$$w_k^{\text{e}}(t + t_l) = \sqrt{1 - (w_k^{\text{ext}}(t + t_l))^2 - (w_k^{\text{nwp}}(t + t_l))^2} , \quad (6.12)$$

with $w_k^{\text{ext}}(t + t_l)$, $w_k^{\text{nwp}}(t + t_l)$ and $w_k^{\text{e}}(t + t_l)$ the weights for the extrapolation, NWP and noise cascade, respectively, at scale level k and time $t + t_l$. $\lambda_k^{\text{ext}}(t + t_l)$ and $\lambda_k^{\text{nwp}}(t + t_l)$ are the ratios of the explained to the unexplained variance of the extrapolation and NWP components, and are calculated as:

$$\lambda_k^{\text{ext}}(t + t_l) = \frac{(\rho_k^{\text{ext}}(t + t_l))^2}{1 - (\rho_k^{\text{ext}}(t + t_l))^2} , \quad (6.13)$$

$$\lambda_k^{\text{nwp}}(t + t_l) = \frac{(\rho_k^{\text{nwp}}(t + t_l))^2}{1 - (\rho_k^{\text{nwp}}(t + t_l))^2} . \quad (6.14)$$

Seed et al. method A disadvantage of the method by Bowler et al. (2006) is that the method assumes that all cascades are uncorrelated. To enable the blending of more than two forecasts, for example multiple NWP forecasts, Seed et al. (2013) introduced a covariance-based method to determine the blending weights. We have adapted the original formulation slightly to avoid zero-values in the denominator, and used the normalised covariance matrix, which consists of the cross-correlations between the extrapolation and model cascades (A. Seed, private communication, Dec. 2021). Considering these adjustments, the weights are determined as:

$$\vec{w}_k(t + t_l) = \begin{bmatrix} \rho_k^{1,1}(t + t_l) & \rho_k^{1,2}(t + t_l) & \dots \\ \rho_k^{2,1}(t + t_l) & \rho_k^{2,2}(t + t_l) & \dots \\ \dots & \dots & \dots \end{bmatrix}^{-1} \cdot \begin{bmatrix} \rho_k^1(t + t_l) \\ \rho_k^2(t + t_l) \\ \dots \end{bmatrix} , \quad (6.15)$$

with $\rho_k^{1,2}(t + t_l)$ the cross-correlation between model 1 and 2, for example the extrapolation and NWP cascade, on scale level k and for time $t + t_l$. $\rho_k^{1,1}(t + t_l)$ is the cross-correlation of model 1 with itself, which should equal 1.0. If more than two models (i.e., more than just one extrapolation and one NWP component) will be blended, the matrix increases from 2×2 to $n_{\text{models}} \times n_{\text{models}}$, as indicated with the dots in Equation 6.15. $\rho_k^1(t + t_l)$ is the skill of model component 1, for example the extrapolation cascade $\rho_k^{\text{ext}}(t + t_l)$, on cascade level k and for time $t + t_l$.

Subsequently, the noise weight is calculated as

$$w_k^{\text{e}}(t + t_l) = \sqrt{1 - \vec{w}_k(t + t_l) \cdot \begin{bmatrix} \rho_k^1(t + t_l) \\ \rho_k^2(t + t_l) \\ \dots \end{bmatrix}} . \quad (6.16)$$

To prevent taking the square root of a negative number, the noise weight is set to 0.0 when

$\tilde{w}_k(t + t_l) \cdot \begin{bmatrix} \rho_k^1(t + t_l) \\ \rho_k^2(t + t_l) \\ \dots \end{bmatrix}$ exceeds 1.0. In addition, the sum of the weights can exceed 1.0 (this also

holds for the Bowler et al., 2006, method) and the weights per component can become smaller than 0. This is normal behaviour for covariance-based weight determination methods and is meant to adjust the forecast to values outside the range of the model components when all components conditionally under- or overestimate the true value (Radchenko et al., 2021).

6.2.4 | Advection of the extrapolation and noise cascades

Before we can blend the different cascades, the extrapolation and noise cascades have to be extrapolated from the issue time (the most recent observation) to time $t + t_l$. Both cascades are in Lagrangian space, because this allows for the temporal evolution of the cascades through an AR(p) process (Section 6.2.2). In pysteps, this extrapolation step is one of the last steps in the forecasting framework, but since the NWP forecast for time $t + t_l$ is not in Lagrangian space, the extrapolation and noise cascade first have to be advected prior to the blending step. This change primarily affects the post-processing steps that have been implemented in pysteps (Pulkkinen et al., 2019), which are therefore adjusted in Section 6.2.6.

The advection of the extrapolation and noise cascades takes place with the motion fields that have been derived at the start of the framework (Section 6.2.1). It is possible that the derived motion fields change over the course of the forecast horizon. The NWP forecast can give useful information about this and, therefore, also the derived motion fields are blended (Section 6.2.4.1). Finally, parts of the noise cascade can advect out of the domain in the downwind direction, while no new noise is advected into the domain from the upwind direction. To prevent this loss of noise, the noise that advects out of the domain on one or multiple side(s) is allowed to move into the domain on the opposite side(s) again (so-called mirroring).

6.2.4.1 | Blending the velocity fields

The motion fields for time $t + t_l$ are blended using the 2nd cascade level weights as follows (Bowler et al., 2006):

$$\mathbf{v}(t + t_l) = w_2^{\text{ext}*}(t + t_l)\mathbf{v}^{\text{ext}}(t + t_l) + w_2^{\text{nwp}*}(t + t_l)\mathbf{v}^{\text{nwp}}(t + t_l), \quad (6.17)$$

with \mathbf{v}^{ext} and \mathbf{v}^{nwp} the velocity fields for the extrapolation and NWP cascades, and $w_2^{\text{ext}*}(t + t_l)$ and $w_2^{\text{nwp}*}(t + t_l)$ the weights for the extrapolation and NWP cascades (note that more than two model cascades can be added to this equation), normalized by the sum of the two to ensure a total weight of 1.0. This approach is slightly different from the method in Bowler et al. (2006), where $w_2^{\text{nwp}*}(t + t_l) = 1 - w_2^{\text{ext}*}(t + t_l)$. To take into account the uncertainty in the motion field development, $\mathbf{v}^{\text{ext}}(t + t_l)$ can be stochastically perturbed prior to the blending step in Equation 6.17 (see Bowler et al., 2006; Pulkkinen et al., 2019).

6.2.5 | Blending the cascades and recomposing the forecast

After the temporal evolution and advection of the extrapolation and noise cascades (Sections 6.2.2 and 6.2.4), the cascades can be blended with the NWP cascade(s):

$$Y_{k,i,j}^{\text{blended}}(t+t_l) = w_k^{\text{ext}}(t+t_l) Y_{k,i,j}^{\text{ext}}(t+t_l) + w_k^{\text{nwp}}(t+t_l) Y_{k,i,j}^{\text{nwp}}(t+t_l) + w_k^{\epsilon}(t+t_l) Y_{k,i,j}^{\epsilon}(t+t_l), \quad (6.18)$$

where $Y_{k,i,j}^{\text{ext}}(t+t_l)$ and $Y_{k,i,j}^{\epsilon}(t+t_l)$ are extrapolated to time $t+t_l$ (Section 6.2.4). Note that outside the radar domain, only the NWP and noise cascade(s) are blended. In the current implementation, the aforementioned blending procedure can take place in three ways:

1. There is only one deterministic NWP model: Equation 6.18 is repeated for n requested (extrapolation) ensemble members, which differ only due to the different realizations of the noise cascade.
2. There are multiple NWP models or an ensemble NWP forecast and all members need to be blended per realization of the noise cascade: the procedure remains the same as above, but instead of one model cascade and weight, multiple model cascades and weights are introduced (see also Section 6.2.3.2). As the different model realizations can be correlated, it is recommended to use the Seed et al. (2013) weights method.
3. There are multiple NWP models or an ensemble NWP forecast, but these members are not blended together, rather they are individually blended per realization: Equation 6.18 is applied for NWP model realization 1 and noise cascade realization 1, followed by model and realization 2, 3, and so on. If the requested number of ensemble members is larger than the number of NWP model realizations, this process is simply repeated for the next set of noise cascade realizations in a round-robin fashion. In the latter case, it is recommended to use a number of pysteps ensemble members that is a multiple of the number of NWP ensemble members. The advantages of this method are that it can bypass the correlation problem between individual model realizations, it leads to less degradation of extremes and it increases the spread of the ensemble.

Once the cascades are blended, the result can be recomposed to one forecast field (Bowler et al., 2004, 2006):

$$\text{dBR}_{i,j}^{\text{blended}}(t+t_l) = \sum_{k=1}^K \sigma_k^{\text{blended}}(t+t_l) Y_{k,i,j}^{\text{blended}}(t+t_l) + \mu_k^{\text{blended}}(t+t_l), \quad (6.19)$$

where $\sigma_k^{\text{blended}}(t+t_l)$ and $\mu_k^{\text{blended}}(t+t_l)$ are the weighted sums of the means and standard deviations of the extrapolation and NWP model cascades:

$$\mu_k^{\text{blended}}(t+t_l) = \frac{w_k^{\text{ext}}(t+t_l)}{w_k^{\text{ext}}(t+t_l) + w_k^{\text{nwp}}(t+t_l)} \mu_k^{\text{ext}}(t) + \frac{w_k^{\text{nwp}}(t+t_l)}{w_k^{\text{ext}}(t+t_l) + w_k^{\text{nwp}}(t+t_l)} \mu_k^{\text{nwp}}(t), \quad (6.20)$$

$$\sigma_k^{\text{blended}}(t+t_l) = \frac{w_k^{\text{ext}}(t+t_l)}{w_k^{\text{ext}}(t+t_l) + w_k^{\text{nwp}}(t+t_l)} \sigma_k^{\text{ext}}(t) + \frac{w_k^{\text{nwp}}(t+t_l)}{w_k^{\text{ext}}(t+t_l) + w_k^{\text{nwp}}(t+t_l)} \sigma_k^{\text{nwp}}(t). \quad (6.21)$$

Table 6.1 | Overview of the three events in this study.

Time event (UTC)		Type*	Catchment-average rainfall sum (mm)			
Start	End		Vesdre	Demer	Geul	Dommel
2021-01-27 23:00	2021-01-29 09:00	S	30.6	20.7	27.7	30.6
2021-06-29 11:30	2021-06-30 11:30	C	30.9	54.5	28.8	9.6
2021-07-12 22:00	2021-07-15 21:00	S/C	131.5	92.0	109.2	101.9

*S: stratiform, C: convective.

6.2.6 | Post-processing steps

As a last step, pysteps ensures that the forecast precipitation fields have the same statistical properties as the most recent observation. Two post-processing methods are used for this: (1) masking, which avoids the generation of rainfall too far from the existing precipitation fields, and (2) probability matching, which matches the statistics (the total precipitation volumes) with the most recent observations within the mask. We have implemented two of the pysteps masking methods in the blending framework: one method which constrains the mask to the observed grid cells that exceed a given threshold, and an incremental masking method, which relaxes the mask to a wider area around the precipitation fields. Moreover, we have implemented both probability matching methods from pysteps: the first one, originally developed for S-PROG (Seed, 2003) matches the mean precipitation amount of the masked forecast field to the observed one, while the second method by Foresti et al. (2016) matches the cumulative distribution function (CDF) of the masked forecast field with the observed field (for more information, see Pulkkinen et al., 2019).

Different from the original pysteps implementation where the post-processing steps take place in Lagrangian coordinates (thus, prior to the extrapolation step), the post-processing steps have to take place after extrapolation and blending of the cascades (and incorporate the NWP model fields as well). Therefore, we have developed a Lagrangian blended probability matching (LPBM) scheme and incremental masking strategy, in which the mask consists of the blended radar observation and NWP forecast fields and which advects along with the forecast. In this procedure, the most recent radar observation is extrapolated to time $t + t_i$ with the velocity field from Equation 6.17. Subsequently, the extrapolated radar field is blended with the NWP forecast field(s) using the blending weights from the second cascade level, which is also used to blend the velocity fields. Finally, this blended field is used as a reference for the rainfall intensity distribution used in the post-processing steps.

6.3 | Evaluation of the blended rainfall forecasts

6.3.1 | Study area and rainfall events

The study areas to test and evaluate the blended forecast of Section 6.2 are Belgium and the south of the Netherlands (Figure 2.3). Besides a focus on domain-wide rainfall forecasting performance, we focus on four catchments: (1) Vesdre (685 km²), (2) Demer (2268 km²), (3) Geul (323 km²) and (4) Dommel (1691 km²), which are described in Section 2.3.2.

Table 6.2 | The pysteps configuration used in this study.

Configuration option	Value	Reference
Methods		
Optical flow method	Lucas-Kanade	Lucas et al. (1981)
Advection method	Semi-Lagrangian	Germann & Zawadzki (2002)
Nowcasting method	STEPS	Seed (2003); Seed et al. (2013); Bowler et al. (2006)
Perturbations	Non-parametric	Pulkkinen et al. (2019)
Mask method	Incremental	Seed et al. (2013)
Probability matching	cdf	Pulkkinen et al. (2019) Foresti et al. (2016)
Blending module settings		
Climatological skill window length	3 d	
Weights method	BPS	Bowler et al. (2006)
Forecast settings		
Number of lead times	144 (12 h)	
Number of ensemble members	48	
Precipitation threshold	0.1 mm h ⁻¹	
Order of the AR(p) model	2	
Number of cascade levels	8	
Transformation	R to dBR	Equation 6.1
Velocity perturbations	Turned off	

We focus on three heavy rainfall events in 2021 with different rainfall characteristics (see Table 6.1) that led to flood peaks in some or all of the four catchments. The event in January is a stratiform winter event, typical for the temperate maritime climate in the study area, resulting in moderate to high rainfall sums for winter (20–30 mm, on average) in the four catchments. The event in June is a convective event, more typical for the summers in the study area, with small and locally-occurring intense rainfall cells that locally led to more than 100 mm of rainfall and (flash) flooding in the western part of the Geul (catchment average of 28.8 mm) and eastern part of Demer (catchment average of 54.5 mm). Finally, the July event, as already mentioned in the introduction, was a persistent mesoscale system that contained both stratiform and convective rainfall. Over a large region, this system led to extreme rainfall amounts and devastating floods, with the Vesdre as one of the hotspots, and significant flooding in the Geul and Demer (Koks et al., 2021; Kreienkamp et al., 2021).

6.3.2 | Radar and NWP data

This chapter uses the radar QPE and NWP QPF data described in Sections 2.2.2.1 and 2.2.2.2. Both products have 5-min temporal resolution. The spatial resolution of the radar rainfall product is 1 km, whereas this is 1.3 km for the NWP product.

6.3.3 | Experimental setup

6.3.3.1 | Evaluation of rainfall forecasts

To test the blending setup as described in Section 6.2, we constructed blended forecasts (from here onwards referred to as STEPS blending) with 48 ensemble members and a 12-h forecast horizon for every 5-minute issue time during the three events (Section 6.3.1). For this blending approach, the radar QPE and NWP rainfall forecasts from Sections 6.3.2 were used. The used pysteps configuration is given in Table 6.2. We benchmarked the results against: (1) the NWP forecasts (Section 2.2.2.2), (2) ensemble nowcasts with 48 members, constructed with pysteps (v1.6.2) using the methods and forecast settings from Table 6.2 (except for the blending settings), and (3) a linear blending method (48 members) that was also added to pysteps as an additional blending functionality. The linear blending method linearly reduces the blending weight for the (48-member ensemble) extrapolation component from 1 to 0 between a given start and end time, while it linearly increases the blending weight for the (deterministic) NWP component from 0 to 1. For this study, the start and end times of the linear blending were fixed at 1 h and 3 h after the issue time, which is around the average skillful lead time of 2 h for nowcasting for catchments in this region (Imhoff et al., 2020a). The radar QPE, that was also used to construct the (blended) nowcasts, was considered the observation in this study.

In the evaluation of the STEPS blending approach and the comparison with the three aforementioned models, we focused on two spatial scales: the radar domain and the catchment scale. At the radar domain scale, which provides a measure of the forecast skill on a country level, the forecasts of the four models were validated using the CRPS and CSI (Equations 2.10 and 2.18). Per issue time, both metrics were calculated per grid cell, and, subsequently, averaged over all grid cells in the radar domain. The used thresholds to calculate the CSI were 1.0 and 5.0 mm h⁻¹. At the catchment scale, the point-wise precipitation intensities and precise grid point localisation becomes less relevant, but the accumulation over catchments and spatio-temporal consistency becomes more relevant. On this scale, we validated the forecasts of the four models using the CRPS on both the catchment-averaged rainfall (per lead time) and cumulative rainfall sums (from issue time until lead time t_l) for the catchments Vesdre, Demer, Geul and Dommel.

6.3.3.2 | Evaluation of climatological moving window size

The skill of the extrapolation and NWP components per lead time determines the blending weights for that lead time. The NWP skill regresses from the initial skill at the issue time of the forecast to its climatological skill, which is based on the past skill for a given moving window size (Section 6.2.3.1). The choice for this moving window size will depend on, for instance, the variety in weather patterns and seasons, and its influence on the skill of the NWP rainfall forecasts. A short moving window of only several days may better represent the current NWP skill for some regions and climatic zones, but may at the same time be too short and contain an insufficient number of rainy samples for others. Throughout this study, we focused on a 3-day moving window size, but we also tested other moving window sizes (1, 7, 14 and 21 days). In Section 6.4.2, we visualize the effects of these window sizes on the resulting climatological skill on all 8 spatial cascade levels for the months January, June and July 2021 (the months containing the three events). In addition, we tested these window sizes for one issue time (13:30 UTC) during the June event (Table 6.1), with a focus on the effects concerning both the domain-wide rainfall forecast skill and the catchment-averaged forecast skill. This particular test case for June is also used to illustrate the differences in rainfall forecasts between the methods (see Figures 6.3

and 6.4 in Section 6.4.1.1). This should give a first impression of the sensitivity of the blending approach to the moving window size choice.

6.3.3.3 | Evaluation of weights method

Throughout this study, the weights method by Bowler et al. (2006) was used. We also compared the effect of both the Bowler et al. (2006) and Seed et al. (2013) weights derivation method on the resulting forecast skill for the same issue time (13:30 UTC) during the June event. The resulting weights for this issue time on all 8 cascade levels are shown in Figure 6.2. In Section 6.4.3, we will discuss the effects of this approach on the rainfall forecast skill, both at the catchment and radar domain scale, for this forecast.

6.4 | Results

6.4.1 | Evaluation of rainfall forecasts

6.4.1.1 | Example case of 2021-06-29

Before we discuss the statistics per event based on all forecasts, Figure 6.3 illustrates the rainfall forecasts of all tested methods for just one issue time, the test case of 2021-06-29 13:30 UTC. This day consisted of convective rainfall, with locally high-intensity rainfall, especially near the Vesdre, Demer and Geul catchments, which is generally challenging to forecast well with both nowcasting and NWP (e.g. Roberts & Lean, 2008; Berenguer et al., 2012; Ayzel et al., 2019b). Up to at least the first hour ahead, the nowcast captures the location and intensity of the rainfall better than the NWP forecast. After 3 h, this reverses, which is also in line with the maximum skillful lead time of nowcasting (Lin et al., 2005; Ayzel et al., 2019b; Imhoff et al., 2020a). The linear blending approach resembles the nowcast during the first hour, after which the NWP forecast slowly gets more weight until three hours ahead, when the linear blending forecast is the same as the deterministic NWP forecast. As a consequence, there are no differences between the 48 ensemble members in the linear blending approach beyond the 3-h lead time.

As the nowcast fails to capture the localisation of rainfall and the rainfall tends to dissipate for the indicated lead times of 6 and 12 h (and to a lesser extent 3 h), the linear blending approach seems beneficial here. The same holds for the STEPS blending approach, which also adds perturbations to the NWP forecast. This increases the ensemble spread throughout the forecast (for more information about the ensemble spread, see Figures D.1–D.3 in the supplement). From visual inspection, it is hard to say which of the blending approaches performs better for this event, although the linear blending approach seems to benefit from the higher rainfall intensities in the NWP forecast during the 6 and 12-h lead time. However, as we are only focusing on one ensemble member, this is not an entirely fair comparison.

Figure 6.4a, instead, takes the entire ensemble into account by showing the CRPS for the forecasts of Figure 6.3. Averaged over the entire radar domain, the ensemble nowcast results in a lower error than the NWP forecast for the entire forecast horizon. This can be partly attributed to the frequent zero-rainfall forecasts at the grid cell level for the ensemble nowcast forecasts, which can benefit statistics such as the CRPS when a larger fraction of the radar domain has zero rainfall with a few scattered high-intensity rainfall cells. At the catchment scale (Figure 6.4b), this effect becomes clear with cumulative rainfall sums that stagnate after a lead time of 6 h and an overall underestimation of the rainfall amount by the ensemble nowcast for lead times of

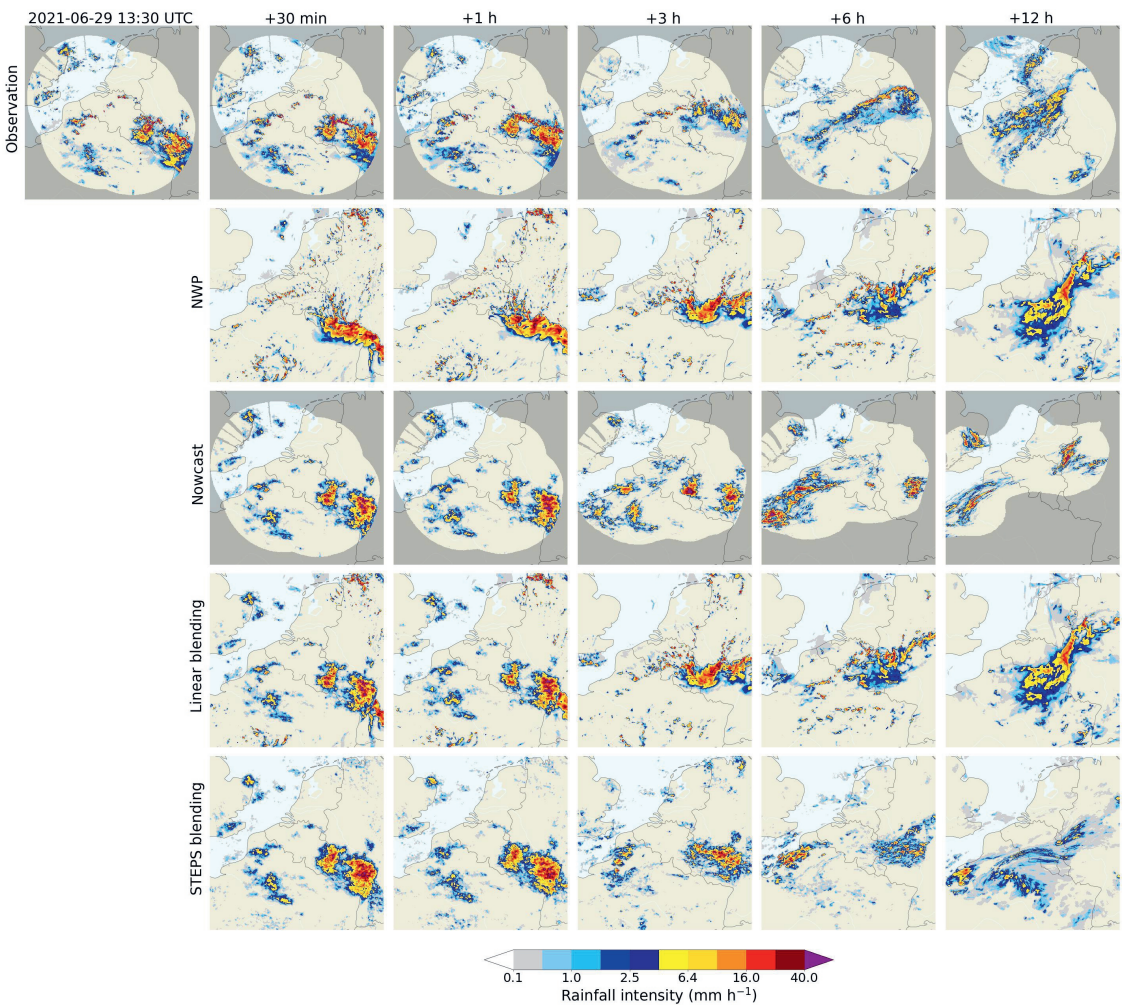


Figure 6.3 | Domain-wide forecast rainfall fields for the test case of 2021-06-29 13:30 UTC. The top row illustrates the observed radar rainfall fields and the rows below illustrate the forecast rainfall fields with the four tested methods (ensemble nowcasting, deterministic NWP, linearly blended forecasts and STEPS-blended forecasts) for five lead times. From the ensemble forecasts, only the first member is shown.

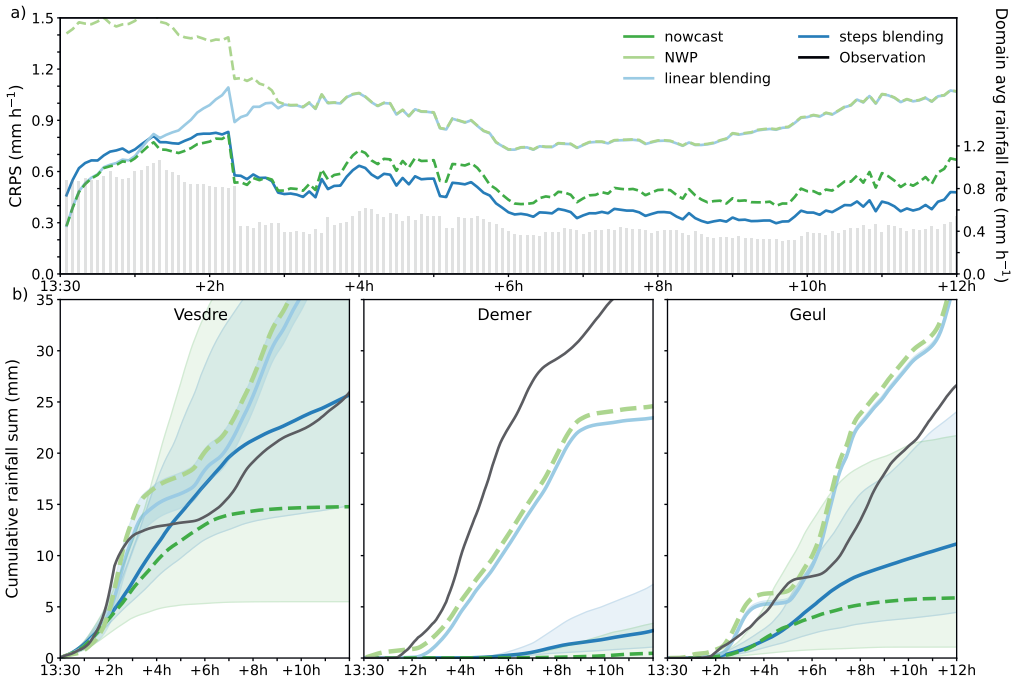


Figure 6.4 | Evaluation of the four forecasting methods for the test case of 2021-06-29 13:30 UTC. (a) The CRPS per lead time, averaged over all grid cells in the radar domain. The grey bars indicate the domain-averaged rainfall rates (mm h^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sums per catchment (Vesdre, Demer and Geul) as compared to the observations (the observed radar rainfall) in black. The thick coloured lines indicate the ensemble median, or the deterministic forecast (for NWP, light green). The shaded areas around the ensemble medians indicate the interquartile range (IQR).

more than 2 h (Vesdre and Geul) or for the entire forecast horizon (Demer). The NWP forecast, however, tends to overestimate the cumulative rainfall sums for the Vesdre and Geul, especially for lead times beyond 6 h. At the same time, it underestimates the rainfall sum for the Demer, though less than the nowcast.

In the linear blending forecast, the skill at the domain scale is the same as the nowcast skill for the first hour and the same as the NWP skill for three hours or more ahead (Figure 6.4a). In between, there is a transition from the skill of the nowcast to the skill of the NWP forecast, which is in line with the fixed blending weights of the linear blending approach (Section 6.3.3.1). At the catchment scale (Figure 6.4b), the results are similar to the NWP forecast.

The domain-averaged CRPS of the STEPS blending forecast is lowest of all tested methods for lead times of 3 h or more (Figure 6.4a). During the first 2–3 hours of the forecast, the STEPS blending forecast has a somewhat higher CRPS than the ensemble nowcast and linear blending method, which may be caused by an excessive (initial) weight for the NWP component during these lead times. At the catchment scale (Figure 6.4b), the STEPS blending approach outperforms the ensemble nowcast for all three catchments. Whether the linear blending or STEPS blending approach was a better choice, differs between the three catchments in this test case

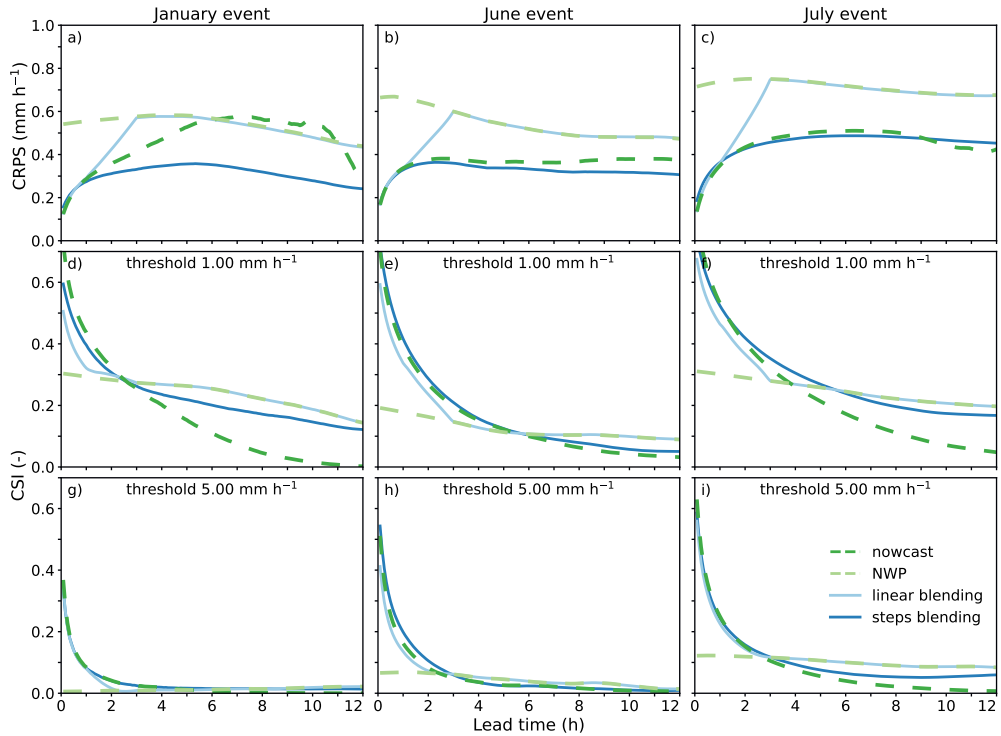


Figure 6.5 | Event-averaged CRPS and CSI per lead time for the four methods over the full radar domain. (a–c) The domain-averaged CSI per event. (d–f) The CSI per event for a threshold of 1 mm h^{-1} , and (g–i) for a threshold of 5 mm h^{-1} .

(and also varies per issue time, see e.g. Figure D.4). For the Vesdre, STEPS blending clearly outperforms all other methods, whereas for the Demer the linear blending approach (and the NWP forecast) are much closer to the observations. This also holds, to a lesser extent, for the Geul, although the observations fall at least within the spread of the STEPS blending approach. In the subsequent sections (6.4.1.2 and 6.4.1.3), we focus on the event-averaged statistics for the four tested methods, based on all forecasts.

6.4.1.2 | Evaluation for the three events on the domain scale

Averaged over the entire radar domain and event duration, STEPS blending attains the lowest CRPS values over the forecast horizon of 12 h (the CRPS values are similar to the ensemble nowcast for the June and July event; Figure 6.5a–c). Only during the January event, the average CRPS of the ensemble nowcast exceeds the CRPS of the NWP forecast at a lead time of approximately 6 h. The linear blending forecasts have similar CRPS values as the ensemble nowcasts and STEPS blending for the first hour of the forecast, but they increase to the CRPS of the NWP for longer lead times.

When focusing on rainfall intensity thresholds of 1.0 and 5.0 mm h^{-1} , the CSI of the ensemble nowcast, linear blending and STEPS blending forecasts are closer (Figure 6.5d–i) than for the

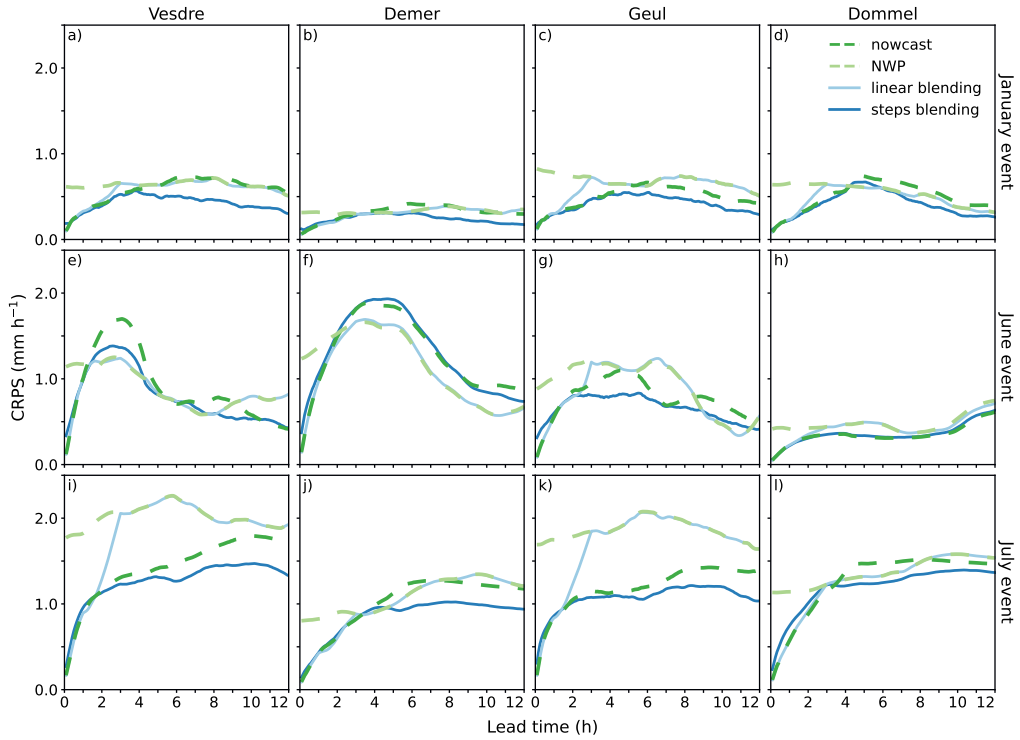


Figure 6.6 | Event-averaged CRPS of the catchment-averaged rainfall forecast for the four catchments and three events (January, a–d; June, e–h; and July, i–l). The ensemble nowcasting method is indicated in dark green, NWP in light green, the linear blending method in light blue and the STEPS blending method in dark blue.

CRPS (a–c). Overall, the CSI of STEPS blending remains somewhat higher for longer lead times than the ensemble nowcasts. It is expected that the CSI of the nowcasts reduces to values lower than those of the NWP forecast for longer lead times (Lin et al., 2005; Germann et al., 2006), which happens for the 1.0 mm h^{-1} threshold between a lead time of 2.5 and 6 h and 2–2.5 h for the 5.0 mm h^{-1} threshold. This is also the point where the linear blending approach starts to outperform the ensemble nowcasts. The CSI values for the STEPS blending approach remain closer to those of the NWP forecast than the CSI of the nowcasts after this transition point, which indicates that both blending approaches (particularly STEPS blending before this transition point) manage to get the best out of both products. Hence, overall at the radar domain scale, STEPS blending outperforms the other methods or has at least a similar performance.

6.4.1.3 | Evaluation for the three events on the catchment scale

When focusing on the four catchments instead of the entire domain, the event-average CRPS values of the stratiform January event are approximately half of those for the (more) convective June and July events, which can be attributed to both lower rainfall rates over the domain and a higher predictability of the event (Figure 6.6). STEPS blending generally attains lower CRPS values than the other methods for the January and July events. During the June event, the NWP and linear blending forecasts outperform the nowcasts and STEPS blending for most lead times

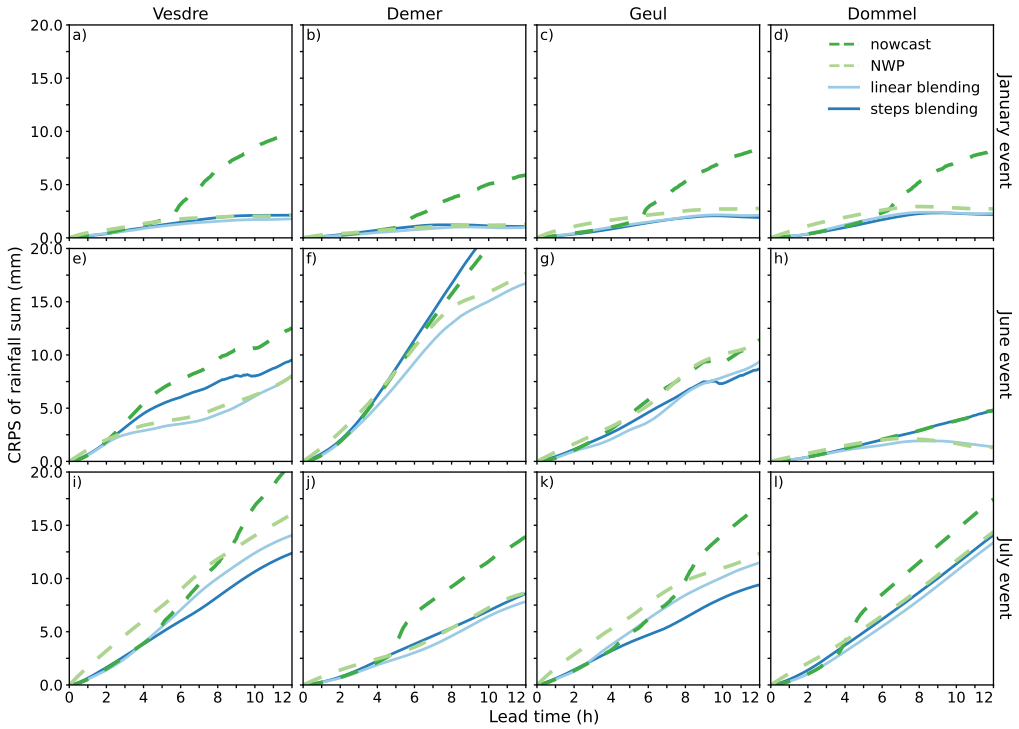


Figure 6.7 | Event-averaged CRPS of the cumulative catchment-averaged rainfall sum from the issue time until the indicated lead time. Shown are the event-averaged CRPS values for the four catchments and three events (January, a–d; June, e–h; and July, i–l). The ensemble nowcasting method is indicated in dark green, NWP in light green, the linear blending method in light blue and the STEPS blending method in dark blue.

of more than 1 – 2 h in catchments Vesdre and Demer.

For the January event (Figure 6.6a–d), the ensemble nowcasts outperform the NWP forecast for at least 3 – 5 h ahead. The linear blending approach blends the NWP too early for the Vesdre, Geul and Dommel in this event. That the optimal blending time for the linear blending approach should be longer for stratiform (winter) events is not surprising due to the higher predictability of these events and resulting better skill of the nowcast (Berenguer et al., 2012; Ayzel et al., 2019b; Imhoff et al., 2020a). The results for the July event are quite similar, though with significantly higher CRPS values for the NWP forecast in the Vesdre and Geul and therefore more skillful nowcasts than NWP for the entire forecast horizon (Figure 6.6i–l).

The June event was convective and therefore more challenging to forecast. Especially for the Vesdre and Demer, which had high-intensity convective rainfall locally, this is directly visible in the higher CRPS values for all methods (Figure 6.6e–h). The ensemble nowcasts already become less skillful after 1 – 2 h for these two catchments.

The aforementioned view changes somewhat for the forecasts of the catchment-average cumulative rainfall volumes, which are relevant for, for instance, (flash) flood forecasts. The nowcasts

often predict zero rainfall after lead times of approximately six hours or more, which can be partially attributed to rainfall leaving the domain, which increases the underestimation by the nowcasts from that lead time onward (see also the bias in supplement Figure D.5). The result is that the CRPS for the cumulative rainfall sum significantly increases after a lead time of 4 – 7 h (Figure 6.7). The overall performance of STEPS blending compared to the other methods remains similar to that in Figure 6.6, but the differences between STEPS blending and linear blending (and to a lesser extent the NWP for longer lead times) becomes smaller and is nearly absent for the January event (Figure 6.7a–d). For the July event, this is also visible, although STEPS blending still outperforms all other methods for the Vesdre and Geul.

To conclude, STEPS blending and linear blending match or even exceed the ensemble nowcasts' performance for the four catchments. Overall, STEPS blending outperforms the other methods for the months January and July, although the difference reduces when we focus on the cumulative rainfall sums for the catchments instead of instantaneous rainfall rates, particularly with respect to the linear blending approach.

6.4.2 | Evaluation of climatological moving window size

The variability in the climatological skill depends strongly on the size of the temporal moving window, and decreases for increasing window sizes (Figure 6.8). This holds for all spatial cascade levels, although from cascade level 2 onwards, the skill becomes close to zero and varies less than on levels 0 and 1. The 1-day window, and to a lesser extent the 3-day window, follows the current skill of the NWP forecast more closely, whereas larger window sizes give a more average skill over a longer period.

Compared to the fixed skill values per cascade level in Bowler et al. (2006), the climatological skill values for the Belgian NWP forecasts are generally lower, probably caused by the higher spatial resolution and time step at which evaluation took place, which is 5-min accumulations in this study and 15-min accumulations in Bowler et al. (2006). This illustrates that the fixed climatological skill values from Bowler et al. (2006) are not representative for the spatial and temporal resolution of the NWP product used in this study.

Another reason for using a moving window approach to estimate the climatological skill is the variability in the NWP skill from day to day (or even per 5-min step; the grey lines in Figure 6.8) and between seasons. For instance, at the largest spatial scale (level 0), the 21-day window mean skill is 0.53 in January (winter period with predominantly stratiform precipitation) and 0.34 in June-July (summer period with more convective precipitation). For the smaller moving window sizes, the difference is particularly observable in the variance of the climatological skill value over time (higher variance for June-July than for January).

Although the difference in climatological skill values is considerable for the tested moving window sizes (Figure 6.8), the effect on the rainfall forecasts for the test case of 2021-06-29 13:30 UTC is less pronounced (Figure 6.9). On this day, the climatological skill values were 0.50 (1-day window), 0.51 (3-day), 0.55 (7-day), 0.26 (14-day) and 0.27 (21-day), hence with a clear difference between the 14 and 21-day windows and the other three windows. At the radar domain scale (Figure 6.9a), the difference in CRPS between the tested moving window sizes gradually increases with lead time, which is expected as the climatological skill value impacts the longer lead times most. Differences in resulting domain-average CRPS are almost absent for the first

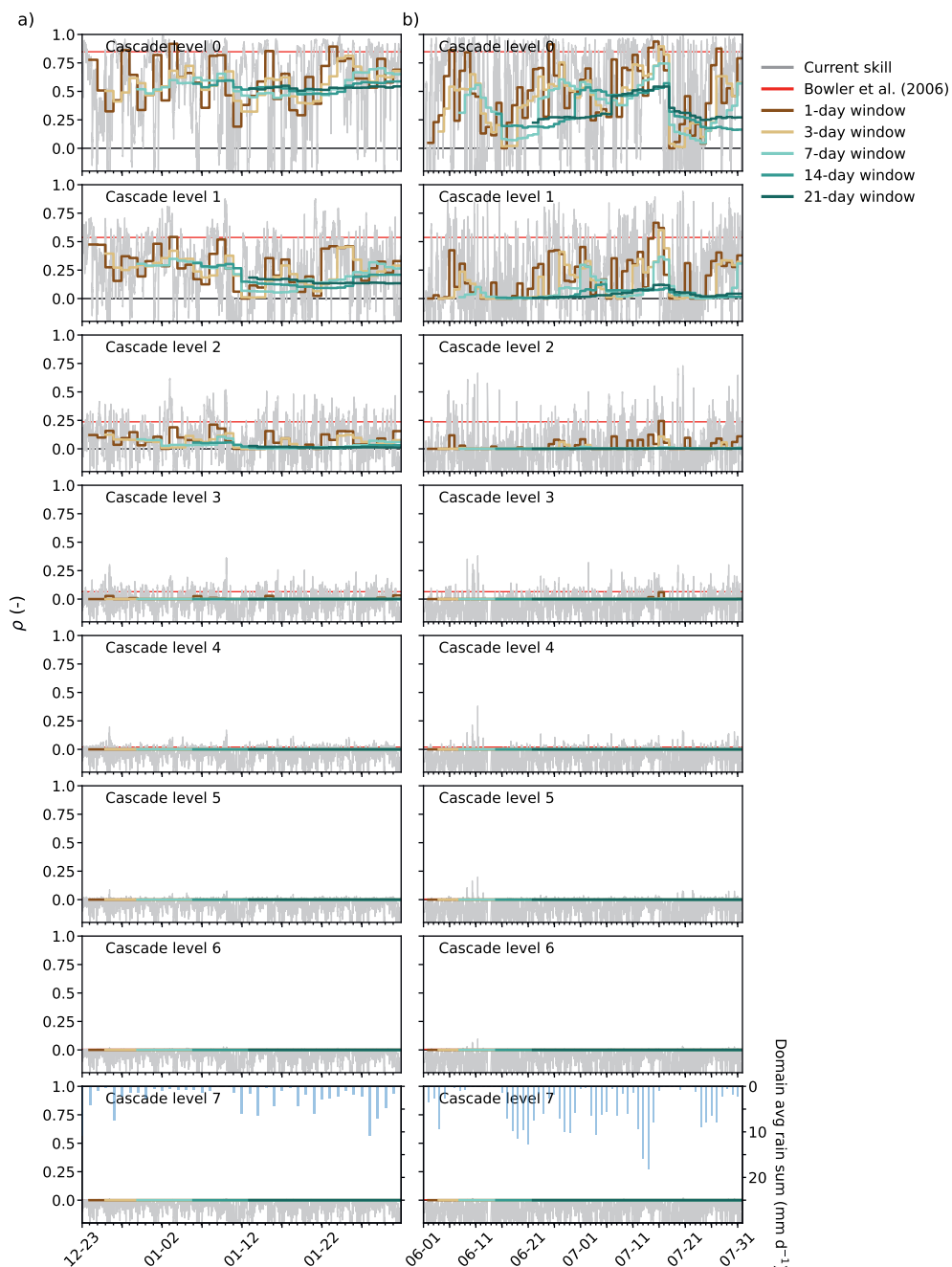


Figure 6.8 | Climatological NWP skill (Pearson correlation) per cascade level as a function of moving window size for (a) January and (b) June-July. The grey lines indicate the NWP skill of the most recently available NWP forecast for that time step as compared to the observed radar rainfall amount, the red lines indicate the climatological skill as provided by Bowler et al. (2006), the brown to blue coloured lines indicate the day-average climatological skill for a given moving window size (the longer window sizes start at a later date as they need t previous days to calculate an average skill). The blue bars indicated in the bottom right panels of (a) and (b) indicate the domain-average rainfall intensity per 5-min time step.

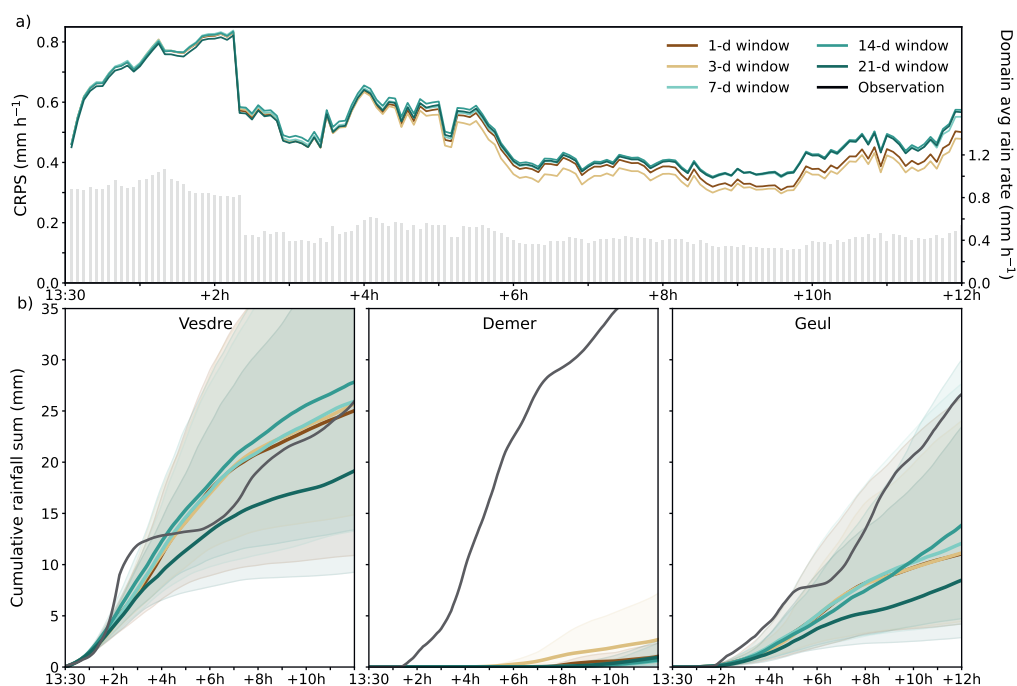


Figure 6.9 | Effect of the climatological skill window sizes on the resulting blended rainfall forecasts for the test case of 2021-06-29 13:30 UTC. (a) The CRPS per lead time, averaged over all grid cells in the radar domain. The grey bars indicate the domain-averaged rainfall rates (mm h^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sum per moving window size as compared to the observation in black. The thick coloured lines indicate the ensemble median and the shaded areas around it indicate the IQR.

four hours of the forecast, but eventually become at most 0.09 mm h^{-1} between the rainfall forecast with a 3-day and 14-d window. At the catchment scale, differences are generally also minor, although the 21-day window underestimates the rainfall more than the other tested window sizes (for this case) for longer lead times, particularly for the Vesdre and Geul (Figure 6.9b).

Concluding, the moving window approach for the climatological skill captures the temporal variability of the NWP skill better than fixed values, although the choice for the moving window size can have a considerable effect on the resulting climatological skill. The effect on the rainfall forecast on both the radar domain and the catchment scale is, however, limited. For the test case, the smaller (1 and 3-day) window sizes result in somewhat lower forecast errors, which is in favour of the used 3-day window in this study.

6.4.3 | Evaluation of weights method

The two methods to determine the blending weights (Bowler et al., 2006; Seed et al., 2013) result in fairly similar weights for the test case of 2021-06-29 13:30 UTC (Figure 6.10), except for the negative weights that occur during the first hours with the Seed et al. (2013) method at cascade level 1 and, to a lesser extent, at level 2. Smaller differences in the resulting weights between the two methods are: (1) the extrapolation component weight that exceeds 1.0 during the first and

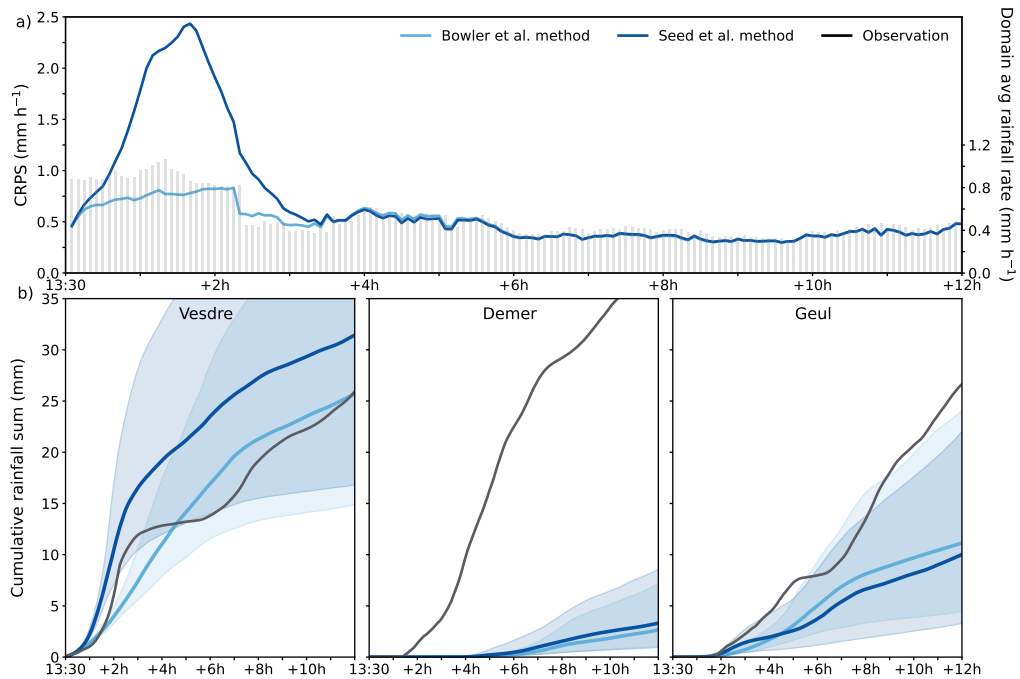


Figure 6.10 | Effect of the two weights methods on the resulting blended rainfall forecasts for the test case of 2021-06-29 13:30 UTC. (a) The CRPS per lead time, averaged over all grid cells in the radar domain. The method by Bowler et al. (2006) is illustrated in light blue and the method by Seed et al. (2013) in dark blue. The grey bars indicate the domain-averaged rainfall rates (mm h^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sum for both methods as compared to the observation in black. The thick coloured lines indicate the ensemble median and the shaded areas around it indicate the IQR.

second hour of the forecast at cascade level 1 with the Seed et al. (2013) method (Figure 6.2b), while this weight is continuously decreasing for the Bowler et al. (2006) method (Figure 6.2a), and (2) the NWP component weights that exceed zero at cascade levels 2 and 3 with the Seed et al. (2013) method, while these weights remain zero for the Bowler et al. (2006) method.

Due to particularly the negative NWP weights at cascade level 1 for the Seed et al. (2013) method, the CRPS of the forecast with the Seed et al. (2013) weights is significantly higher than the CRPS for the Bowler et al. (2006) method at the radar domain scale for the first three hours of the forecast (Figure 6.10a). The maximum difference, which occurs at a lead time of 95 minutes, is 1.7 mm h^{-1} . After more than three hours, the differences reduces and becomes almost absent, which corresponds to the weights that have become relatively similar from that point onward (Figure 6.2). At the catchment scale, this difference is particularly pronounced for the Vesdre, because the rainfall starts there at the issue time of the forecast. For the other two catchments, rainfall commences later (at lead times of 2 – 4 h) and therefore the difference in weights during the first three hours has a limited impact on the resulting forecast rainfall sums. Hence, the choice for the weights method can have impact on the forecast skill. For this example, the results favour the Bowler et al. (2006) weights method, provided that only one deterministic NWP model is blended with the extrapolation and noise components.

6.5 | Discussion

The STEPS blending approach has shown to provide rainfall forecasts that are at least as good as, but regularly outperform, the ensemble nowcasts and NWP forecasts at the radar domain scale. These results are in line with the results of other blending methods (Golding, 1998; Atencia et al., 2010; Kober et al., 2012; Nerini et al., 2019; Yoon, 2019; Radhakrishnan & Chandrasekar, 2020). This analysis differs from most of the aforementioned studies in that it also focuses on the rainfall forecasts at the catchment scale, where from a flood forecasting perspective both the rainfall location and volume over time are most relevant. At this scale, the STEPS blending approach yields convincing results, although when we focus on the cumulative rainfall sums for the entire forecast horizon, the differences with the other models reduce. This is particularly in favour of the linear blending and NWP forecasts, as the STEPS blending approach tends to give more weight to the nowcast component for longer lead times, which tends to underestimate more (see also Imhoff et al., 2020a).

6.5.1 | Bias towards the radar-based products

The presented approach bases the skill, and therefore the blending weights, on the highest resolution of the radar and NWP data, which is a 5-min temporal and 1-km spatial resolution (after spatial downscaling and temporal aggregation of the NWP forecasts). To better match the catchment perspective, it may be of interest to base the skill on hourly or multi-hour sums and on coarser spatial resolutions. NWP (rainfall) forecasts are known to perform better on a coarser resolution in space and time, which is generally done by upscaling in space and aggregating the forecasts in time (e.g. Gangopadhyay et al., 2004; Mittermaier, 2006). An advantage of such an approach, i.e. determining the ‘current skill’ on a coarser spatial and temporal resolution, is that (minor) displacement errors are less penalised and that rainfall sums over a longer aggregation period become more relevant. As the focus on cumulative sums in this study has been advantageous for the NWP forecasts, compared to a focus on instantaneous rainfall rates, this might lead to higher weights for the NWP component(s). Besides, current computation times (on 4 CPU cores) were between 120 and 165 min, which will strongly decrease at a coarser spatio-temporal resolution.

In addition, we regarded the radar QPE as the ‘true’ rainfall in this study, even though radar QPE products come with considerable (systematic) biases and other sources of error (Austin, 1987; Joss & Lee, 1995; Creutin et al., 1997; Gabella et al., 2000; Sharif et al., 2002; Uijlenhoet & Berne, 2008). The use of this product as observation favours the radar-based nowcasting component in the blending approach. Since the precise radar QPE quality is (generally) unknown at the issue time of the blended forecast, a Bayesian weight determination method could be considered (see Section 6.5.2). Besides, it is recommended to use a bias-adjusted radar QPE product to prevent that the blending method is steered towards the systematic bias in the unadjusted radar QPE (for an overview of adjustment methods, see Ochoa-Rodriguez et al., 2019).

6.5.2 | Future implementations and outlook

In the implementation considered here, we have incorporated two blending methods, namely the one by Bowler et al. (2006) and the one by Seed et al. (2013). In essence, the optimal weights are based on multiple (forecasting) products and a ‘true’ value that is unknown (for instance,

due to the aforementioned biases in the radar QPE product), which is a dilemma that is very similar to typical data assimilation problems. A way to tackle this was proposed by Nerini et al. (2019), who introduced an ensemble Kalman filter-based Bayesian blended forecasting system, which would be a logical next implementation in this open-source blending approach.

Another advantage of the approach by Nerini et al. (2019) is that it enables a resampling of the NWP and nowcasted rainfall amounts per grid cell, which could be beneficial for the Lagrangian blended probability matching (LPBM) scheme that was introduced in this study (Section 6.2.6). The current implementation advects the latest radar rainfall observations to lead time $t + t_i$, without any further perturbations and auto-regression steps, blends this with the NWP forecast and uses this as the ‘observation’, at $t + t_i$, to determine the statistics for the probability matching steps. A disadvantage of this implementation is that peak values can be dampened, especially when the weights for the NWP and extrapolation components are (nearly) equal. This means that the target distribution for probability matching can become smoother than the original radar and NWP fields, something that can be prevented with the resampling scheme in Nerini et al. (2019), which preserves the target distribution. This approach will be implemented in the pysteps blending scheme in the near future.

Moreover, both the blending weights and the (blended) normalized multiplicative cascades are variance-based. Therefore, the estimation of the variance in the individual blending components is an important aspect of the STEPS blending procedure. To better cope with the non-normal distribution of rainfall, a log-transform was used. This transform cannot deal with zeroes and therefore, these zeroes are masked. This leads to an unnatural (sharp) transition between rain and no rain, which can influence the estimation of the current and future variance, especially if the variance is estimated and blended in space (not done in this implementation, but a possible future implementation). This could be solved by implementing a transformation that can deal with zeroes, for example the log-sine transformation. This is a recommendation for future developments of the blending code.

Ultimately, this open-source blending implementation (of both STEPS and linear blending) in pysteps should pave the way to implement other and new blending methods, and could be used as a benchmark for future algorithm development. Besides the aforementioned future implementations, current plans in the pysteps blending module are, among others, to include deep-learning methods, as well as the blending method by Atencia et al. (2010) in which the NWP forecast is first phase-corrected with the latest (radar) observations for displacement errors, before blending the individual components. This should increase the NWP forecast skill during the blending procedure and prevents blending of misplaced rainfall fields.

6.6 | Conclusions

Although the first few hours ahead (in the order of 6 h) in rainfall forecasting are crucial for, for example, (flash) flood warnings, this time scale is generally not sufficiently well captured by the rainfall forecasts of numerical weather prediction (NWP) models. Radar rainfall nowcasting, an observation-based rainfall forecasting technique that statistically extrapolates current observations into the future, provides opportunities at this time scale, but has as disadvantage that it quickly loses skill after approximately the first 2 h of the forecast for individual radars. To extend the skillful lead time of short-term rainfall forecasts and improve flash flood early warning,

we have to bridge the gap between nowcasting and short-range NWP model forecasts. One way to do so, is by combining both products, so-called blending. In this study, we have implemented an adaptive scale-dependent ensemble blending method in the open-source Python library *pysteps*, based on earlier work on the STEPS scheme. In this implementation, the extrapolation (ensemble) nowcast, (ensemble) NWP and noise components are blended with weights that vary per spatial cascade level. We described the implementation details and new functionalities, and evaluated the method on three events in 2021 that led to high discharge peaks in the Belgian and Dutch catchments Vesdre, Demer, Geul and Dommel (including the dramatic July 2021 case that caused more than 200 casualties and enormous economic damage). To benchmark the results of the tested 48-member blended forecasts, we compared the results against the original deterministic NWP forecast, a 48-member ensemble nowcast with *pysteps* and a simple (48-member) ensemble linear blending approach.

At the radar domain scale, the implemented STEPS blending approach performs on par with or better than the other three tested methods. This also holds for higher intensity rainfall cells, although the difference between nowcasting, linear blending and STEPS blending is less pronounced for higher intensity rainfall than for the average of all forecasts. At the catchment level, the linear and STEPS blending approaches result in lower forecast errors than only nowcasting, particularly for lead times of approximately 4 h or longer (depending on the rainfall type). Both methods outperform the NWP forecasts for the first few hours of the forecasts, followed by a similar skill for longer lead times. Overall, STEPS blending outperforms the other methods for the two events in January (stratiform) and July (stratiform-convective), although the difference, particularly with the linear blending method, reduces when we focus on the cumulative rainfall sums for the catchments instead of instantaneous rainfall rates.

The scale-dependent blending weights in the STEPS blending implementation are computed from the recent skill (Pearson's correlation) of the forecast components, and converge to a climatological value. In contrast to the original STEPS blending approach, this implementation bases the climatological skill value for the NWP component(s) on the recent NWP skill with a multi-day moving (averaging) window, instead of fixed values which do not take into account the temporal variability in the NWP forecast skill. Although a 3-day moving window was used for the aforementioned evaluation, we also tested moving window sizes of 1, 7, 14 and 21 days. For the test case considered, the tested moving window sizes give minimal differences in the results, even though the skill values can vary considerably between the moving window sizes.

In addition, we have implemented two methods (the ones by Bowler et al., 2006; Seed et al., 2013) to determine the blending weights from the estimated skill of the components. As the Seed et al. (2013) weights can result in negative weights or weights that exceed 1.0 for the individual blending components, both the resulting weights and forecasts can differ significantly from the Bowler et al. (2006) approach. The results from the test case in this study favours the Bowler et al. (2006) weights method, but that is provided that only one deterministic NWP model is blended with the extrapolation and noise components. For multi-model ensembles, the Seed et al. (2013) method is recommended as it takes into account the cross-correlation between the models.

Concluding, we consider this open-source blending approach in *pysteps* as a starting point for further implementations of other blending methods and future collaborations. In this way, we

envision an acceleration of developments in the realm of short-term rainfall forecasting. The pysteps initiative has already demonstrated that this is feasible in the nowcasting domain, a development which we strongly support.

A stylized illustration of a storm. A large, dark grey, billowing cloud dominates the upper half of the frame. Below the cloud, several vertical, textured grey bars represent rain falling against a teal background. In the lower portion, a light blue, swirling shape suggests a storm or a path. A large, white, serif number '7' is superimposed over the center of the image, partially overlapping the cloud and the rain.

7

Evaluation of rainfall nowcasting for flood early warning

This chapter was originally published as:

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. doi: 10.1029/2021WR031591

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-h 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.

“If, then, there is any error whatever in observing the
present state—and in any real system such errors seem inevitable—
an acceptable prediction of an instantaneous state
in the distant future may well be impossible.”

—Edward N. Lorenz, *Deterministic Nonperiodic Flow* (1963)

7.1 | Introduction

The livability and economy of many areas worldwide are endangered by floods (European Environment Agency, 2004; Merz et al., 2010; Jongman et al., 2012; Ward et al., 2013; Ceola et al., 2014). 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 (Hirabayashi et al., 2013; Klein Tank et al., 2014; Arnell & Gosling, 2016). 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 (UNISDR, 2002; Pappenberger et al., 2015). In less extreme situations water managers and indirectly citizens can also benefit from improved hydrological predictions and early warnings on the short term for e.g. 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, e.g. as a result of model structure, initial conditions, setup or calibration procedures (Beven, 1993; Melsen et al., 2016; Clark et al., 2017), 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 h) numerical weather prediction (NWP) model output as quantitative precipitation forecast (QPFs). On the short term (up to approximately 6 h 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: (1) a too coarse temporal resolution, (2) a too low update frequency or (3) the mislocation of rainfall events (Roberts & Lean, 2008; Lin et al., 2005; Berenguer et al., 2012; Pierce et al., 2012). 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 h and regularly arrives at the end users 4 h after the issue time of the forecast. Within those 4 h, let alone the 6-h 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: (1) increasingly rapid update cycle NWP models, (2) nowcasting, possibly incorporating machine learning techniques, and (3) a blended system using the former two (e.g. Golding, 1998; Germann & Zawadzki, 2002; Turner et al., 2004; Bowler et al., 2006; Sun et al., 2014). In this chapter, 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 et al., 2009b). 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 h for larger-scale and more persistent rainfall events, up to a maximum of 6 h for persistent stratiform events on a continental scale (Germann & Zawadzki, 2002; Lin et al., 2005; Germann et al., 2006; Berenguer et al., 2011; Liguori & Rico-Ramirez, 2012; Berenguer et al., 2012; Foresti et al., 2016; Mejsnar et al., 2018; Ayzel et al., 2019b, and Chapter 4 of this thesis).

Nowcasted rainfall has already been successfully used as input for various hydrological models and forecasting systems (Berenguer et al., 2005; Pierce et al., 2005; Sharif et al., 2006; Vivoni et al., 2006, 2007; Germann et al., 2009; Liguori et al., 2012; Liguori & Rico-Ramirez, 2013; Moreno et al., 2013; Poletti et al., 2019; Heuvelink et al., 2020). Berenguer et al. (2005) and Heuvelink et al. (2020) have found significant improvements in discharge forecasts, with a gain in anticipation time of 10 to 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, 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 chapter 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 Chapter 4, we analysed nowcasts for 1,500+ events spread over 12 catchments in the Netherlands to evaluate rainfall predictability. Four open-source nowcasting algorithms were used: two benchmarking advection algorithms from the rainymotion library (Ayzel et al., 2019b) and two from the pysteps library (Pulkkinen et al., 2019). This chapter 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 km²). 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 7.2 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 7.3, the discussion in section 7.4 and the conclusions in section 7.5.

7.2 | Materials and methods

7.2.1 | Study area

The study area is the same as in Chapter 4 and comprises 12 lowland catchments in the Netherlands (Figure 2.2 and Table 2.3). 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 chapter. More information about these catchments, their characteristics and the employed hydrological models can be found in Section 2.3.1 of Chapter 2.

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 4.2.4.3 and 4.3.3.2 of Chapter 4). 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 chapter, 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.

7.2.2 | Nowcasts

The nowcasts used in this chapter were the same as the nowcasts constructed for the large sample of events in Chapter 4. In this section, we briefly introduce the underlying radar rainfall product (Section 7.2.2.1), specifics about the available nowcasts (Section 7.2.2.2) and the set of algorithms used to construct these nowcasts (Section 7.2.2.3). For more information, see Chapter 4.

7.2.2.1 | Bias-adjusted radar rainfall product

In this chapter, rainfall estimates from the unadjusted radar datasets with a 1-km² spatial and 5-min temporal resolution were used (described in Section 2.2.1 of chapter 2). As this product is not bias-corrected, underestimations of the true rainfall amount of 50% or more can be expected in the Netherlands (Hazenbergh et al., 2014, and Chapter 3 in this thesis). 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 (Chapter 3). 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 chapter. The factors are based on a 10-year historical dataset 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 in Chapter 3 to be outperformed by the CARROTS correction for hydrological simulation for most of the twelve catchments in this chapter. 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 Chapter 4, 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.

7.2.2.2 | Available nowcasts

The nowcasts from Chapter 4 were selected for the period 2008–2018 in a systematic manner (see also Figure 4.1). 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-h event can fall within the time span of a longer (e.g. 24-h) 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 h). For more information on the event procedure, statistics and model runs, see Sections 4.2.2.2 and 4.2.4, and Table 4.1.

For this large sample of events, nowcasts were produced for each 5-min time step in the event duration and the 6 h 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 h and a temporal resolution of 5 min (information about the event-averaged rainfall intensities and durations can be found in Appendix Figure E.1). Summarizing, over 940,000 separate 5-min nowcasts were constructed for each nowcasting algorithm (see Section 7.2.2.3). Because the events of different durations can have an overlap, i.e. an event with a 1-h duration for catchment x in season y can fall within the time window of an event with a 24-h 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 chapter.

7.2.2.3 | Nowcasting algorithms

The nowcasts for the events described in Section 7.2.2.2 were constructed with four nowcasting algorithms: two from the rainymotion library (Sparse, referred to as RM-S, and DenseRotation, referred to as RM-DR; Ayzel et al., 2019b) and two from the pysteps library (a deterministic and probabilistic setup with 20 ensemble members, referred to as PS-D and PS-P, respectively; Pulkkinen et al., 2019), as described in Section 2.5 of Chapter 2. 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 chapter). Hence, this limits this analysis to only a focus on the benefit of field-based nowcasting for hydrological forecasting.

7.2.3 | Experimental and forecast verification setup

7.2.3.1 | Hydrological model setup

The employed hydrological models for the twelve catchments and their calibration are described in Section 2.4 of Chapter 2. Table 2.3 list the catchments, information about their water balance

and the used models for the study area. As mentioned in Section 2.4, the calibration of the hydrological models for the twelve catchments resulted by no means in the optimal parameter set, due to e.g. the equifinality issue (Beven, 1993). However, the effects of this are excluded from this chapter 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 7.2.3.2).

7.2.3.2 | 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 h (the 6 h forecast horizon of nowcast rainfall as input, followed by zero precipitation for 6 – 12 h in advance), which is more than the average response time of the catchments (Table 2.3, 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 h^{-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 Chapter 4, the rainfall intensity in the latest radar QPE ($t = 0$) was used as the forecast for the coming 12 h. Hence, if it rained with 1.0 mm h^{-1} at $t = 0$, it was assumed to keep on raining with 1.0 mm h^{-1} for every time step in the subsequent 12 h.

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; Melsen et al., 2016; Clark et al., 2017). 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 chapter was the hydrological model simulation with the ‘observed’ radar rainfall after bias correction with the CARROTS factor (see Section 7.2.2.1). 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 7.4.4). On one occasion, in Figure 7.2 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 7.2.3.1).

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 analyse the effect of different rainfall inputs on the model runs in isolation. Besides, van Osnabrugge et al. (2019) showed that including ET_{pot} forecast has little impact on the discharge forecasts, so we expect minor impacts from this decision.

7.2.3.3 | 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 h⁻¹), the NSE score (Equation 2.19 in Section 2.6) 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-h discharge accumulations (rolling sum) of forecast and reference run. This was chosen, instead of the instantaneous discharges, to smooth the ‘pump on - pump off’ behaviour of the polders. Besides, a 1-h accumulation is still less than the response time of all catchments (Table 2.3). The NSE was calculated following the method of Berenguer et al. (2005), who made a discharge time series per lead time, e.g. all 1-h 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 (Equation 2.9) 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-h QPF sum (hence, the first three hours of the nowcast) was compared to the 3-h sum of the reference rainfall. The first 3 h instead of 6 h of the rainfall nowcast were used, because the last 3 h of the nowcast is seldom skillful (Chapter 4). A similar approach was taken for the discharge forecast, though here the 12-h forecast (the full forecast horizon) sum was used.

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 (l_σ) was calculated following the method of Lobligeois et al. (2014):

$$l_\sigma = \frac{\sum_{t=1}^{N_t} \sigma_t \cdot P_t}{\sum_{t=1}^{N_t} P_t}, \quad (7.1)$$

with

$$\sigma_t = \sqrt{\frac{\sum_{i=1}^{N_i} [P_i(t)]^2}{N} - \frac{[\sum_{i=1}^{N_i} P_i(t)]^2}{N^2}}. \quad (7.2)$$

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 l_σ , only the time steps where it rained on at least one grid cell were taken into account.

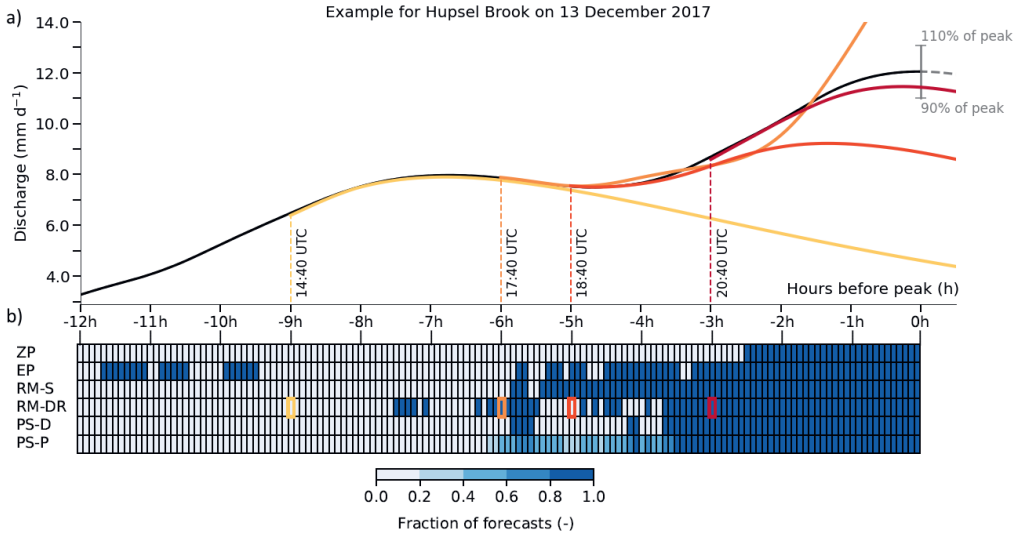


Figure 7.1 | Example of discharge peak verification procedure using the discharge peak that occurred on 13 December 2017 in the Hupsel Brook catchment. (a) The forecast (using RM-DR; colors) and reference (black) discharge. Shown are four issue times of the forecasts: 14:40 UTC (yellow; 9 h before peak), 17:40 UTC (orange; 6 h before peak), 18:40 UTC (red; 5 h before peak) and 20:40 UTC (burgundy; 3 h 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 false alarm ratio, the forecast consistency only uses the 90% threshold exceedance at the time of the peak (see the forecast consistency paragraph in Section 7.2.3.4). (b) Visualization of the forecast consistency for the forecasts issued from 12 h 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).

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 modelled with a distributed hydrological model, which is outside the scope of this paper.

7.2.3.4 | 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 advantage that it can provide robust statistics about the model simulation quality, it has as dis-

advantage that it becomes nearly impossible to verify peak discharge or threshold exceedance forecast quality in an automated way. The reason for that is twofold: (1) for the larger catchments, the peak discharge may arrive later than the 12 h 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 (2) the use of a standard threshold as set by e.g. 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-h 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 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 $\pm 10\%$ of the maximum discharge magnitude (see Figure 7.1a for an example of this magnitude range). As the forecast horizon was 12 h in this chapter, it was possible to verify this from t_{-12} until t_0 h, 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.

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 h) 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 7.1 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 e.g. PS-D occurs around 6 h prior to the maximum observed discharge in the event (indicated with the first dark blue cell in Figure 7.1b) 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 h 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.

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 (Equation 2.14) 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 7.1 for RM-DR. At 6 h prior to the observed maximum discharge, the forecast (orange line in Figure 7.1a) 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 7.1b), 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 h 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.

Forecast consistency

Finally, a forecast can also be inconsistent. It is damaging for water managers to issue, recall and re-issue warnings or flood prevention measures based on a previous erroneous forecast. Figure 7.1 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 7.1b for RM-DR, the forecast hydrographs are visualized in Figure 7.1a, illustrating how a “correct” forecast can be succeeded by a miss an hour later. Such inconsistencies lower the trust in the forecast. In this chapter, 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 7.1b, 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 7.1, 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.

7.3 | Results

In the following results, the quality of the hydrological models is assessed in Section 7.3.1, followed by an example forecast, focusing on one event, in Section 7.3.2. In the remaining part of the results (Sections 7.3.3 and 7.3.4), all events are jointly taken into account.

7.3.1 | Hydrological model validation

The difference between the response of the two polders Gouwepolder and Beemster, and the freely-draining catchments is directly visible in Figure 7.2b and e. 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

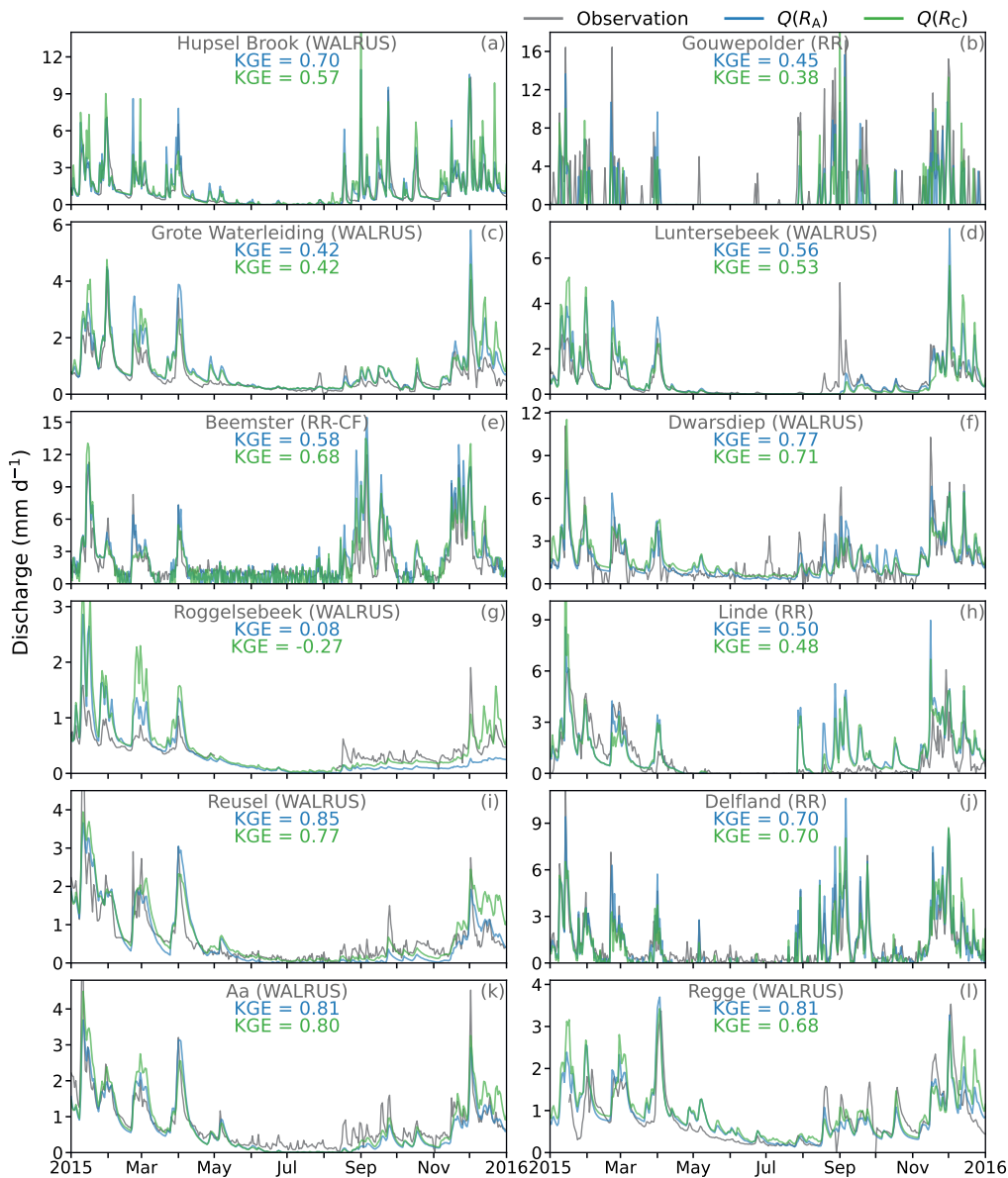


Figure 7.2 | Simulated discharges for the twelve catchments as part of the hydrological model validation. The observed discharge is shown in grey, the model simulations with the gauge-adjusted radar rainfall product ($Q(R_A)$); the reference) in blue and the model simulations with the bias-corrected real-time available product ($Q(R_C)$) in green. Only the results for 2015 are shown, but the KGE-values are based on the full period (2008 – 2018).

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 (Figure 7.2f, h and j).

The hydrological models for most catchments perform well (Figure 7.2). This specifically holds for the catchments Dwarsdiep, Reusel, Delfland, Aa and Regge where the KGE (for a description, see Section 2.6.7.2) 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 7.2h). The opposite happens in the simulations for the Linde, where the discharge is overestimated during fall (Figure 7.2d). 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 7.2.2.1 and 7.2.3.1). 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 .

7.3.2 | Example forecast

Prior to analysing all events, we zoom in on one event to highlight some typical differences between the tested models. Figure 7.3 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 three hours later (and the discharge peak approximately 8 h 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 (Figure 7.3a and b). The nowcasts issued an hour later do forecast rainfall, but substantially less than the approximately 11 mm that would eventually fall between two and three hours later (Figure 7.3c and d). Besides, the timing of the rainfall in the RM-DR forecast is more than 3 h off. For PS-P, more ensemble members forecast a substantial amount of rainfall

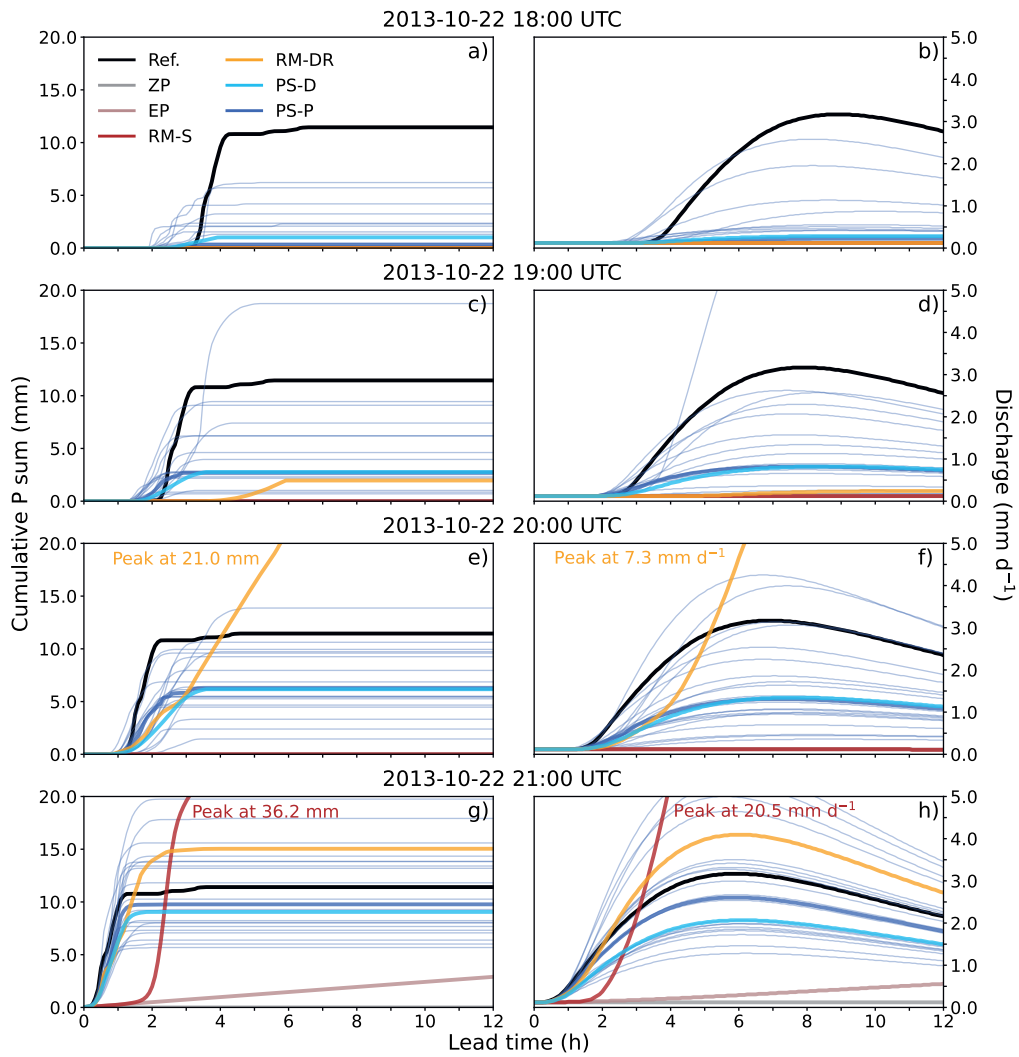


Figure 7.3 | Rainfall and discharge forecasts for four issue times during the discharge peak that occurred in the Hupsel Brook catchment on 22 October 2013. Shown are the forecasts using rainfall inputs from all nowcasting algorithms and methods (colors), compared to the reference (black). The thin dark blue lines indicate the separate ensemble members of PS-P, whereas the thick dark blue line indicates the ensemble median.

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 (Figure 7.3e and f) 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, Figure 7.3g and h), 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 h. 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, e.g. 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 e.g. also Chapter 4. Hence, although the forecast timing and spatial location of the rainfall fields is generally good, the rainfall amount is somewhat underestimated, leading to larger underestimations in discharge (Figure 7.3e–h). 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 7.3.4.

7.3.3 | Dependency on the rainfall characteristics

7.3.3.1 | Rainfall intensity

The skillfulness of the discharge forecast decreases with increasing rainfall intensity (Figures 7.4a–c), which is similar to the effect on the rainfall forecast (Foresti et al., 2016; Ayzel et al., 2019b; Pulkkinen et al., 2019, e.g.). 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 h⁻¹, to 110 min (between 2.0 and 5.0 mm h⁻¹) and to 75 min for rainfall intensities of 5.0 mm h⁻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 4.3. The relative gain, the 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 h⁻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 h⁻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.

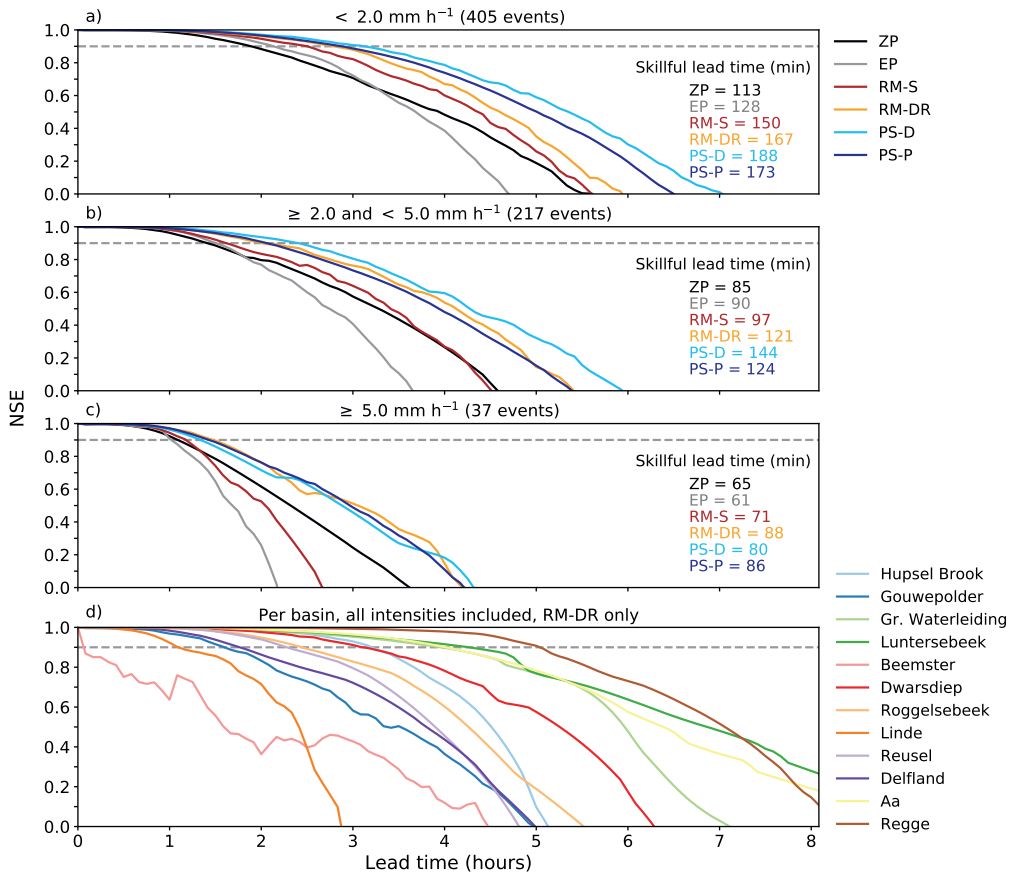


Figure 7.4 | Event-average NSE for all studied methods as function of forecast lead time based on the 1-h accumulated forecast discharge (rolling sum). (a–c) All catchments together for different rainfall thresholds: (a) $< 2.0 \text{ mm h}^{-1}$, (b) ≥ 2.0 and $< 5.0 \text{ mm h}^{-1}$ and (c) $\geq 5.0 \text{ mm h}^{-1}$. 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 RM-DR is shown.

Table 7.1 | Statistics of the linear regression lines (forced through zero) that were adjusted to the data points of Figure 7.5b, which corresponds to the mean event rainfall intensity analysis. 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 7.5b) to the deep groundwater depths (class 5; the red lines in Figure 7.5b).

Catchment	Class 1		Class 2		Class 3		Class 4		Class 5	
	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

^aNo (summer) discharge simulated in this class.

In addition, Figure 7.4d 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 five 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 4.6 in Chapter 4), the size of the catchment (e.g. Figure 4.5 in Chapter 4) and catchment response time.

The decreasing skill with increasing rainfall intensity is also present in Figure 7.5 (and corresponding Table 7.1), 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 three hours of the forecast (Figure 7.5a) or the MAE of the discharge forecast for the entire forecast horizon (12-h sum, Figure 7.5b). For the rainfall nowcasts (Figure 7.5a), the relationship between the increase in rainfall intensity and MAE in the rainfall forecast is linear, with Pearson correlation 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-h) 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 h (Brauer et al., 2011).

Although the error in the discharge forecast increases with increasing rainfall intensity as well (see also Figure 7.4a–c), 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 7.5b). However, one should realize that the hydrological response and predictability depend also on the initial conditions. In Figure 7.5b, the data points 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

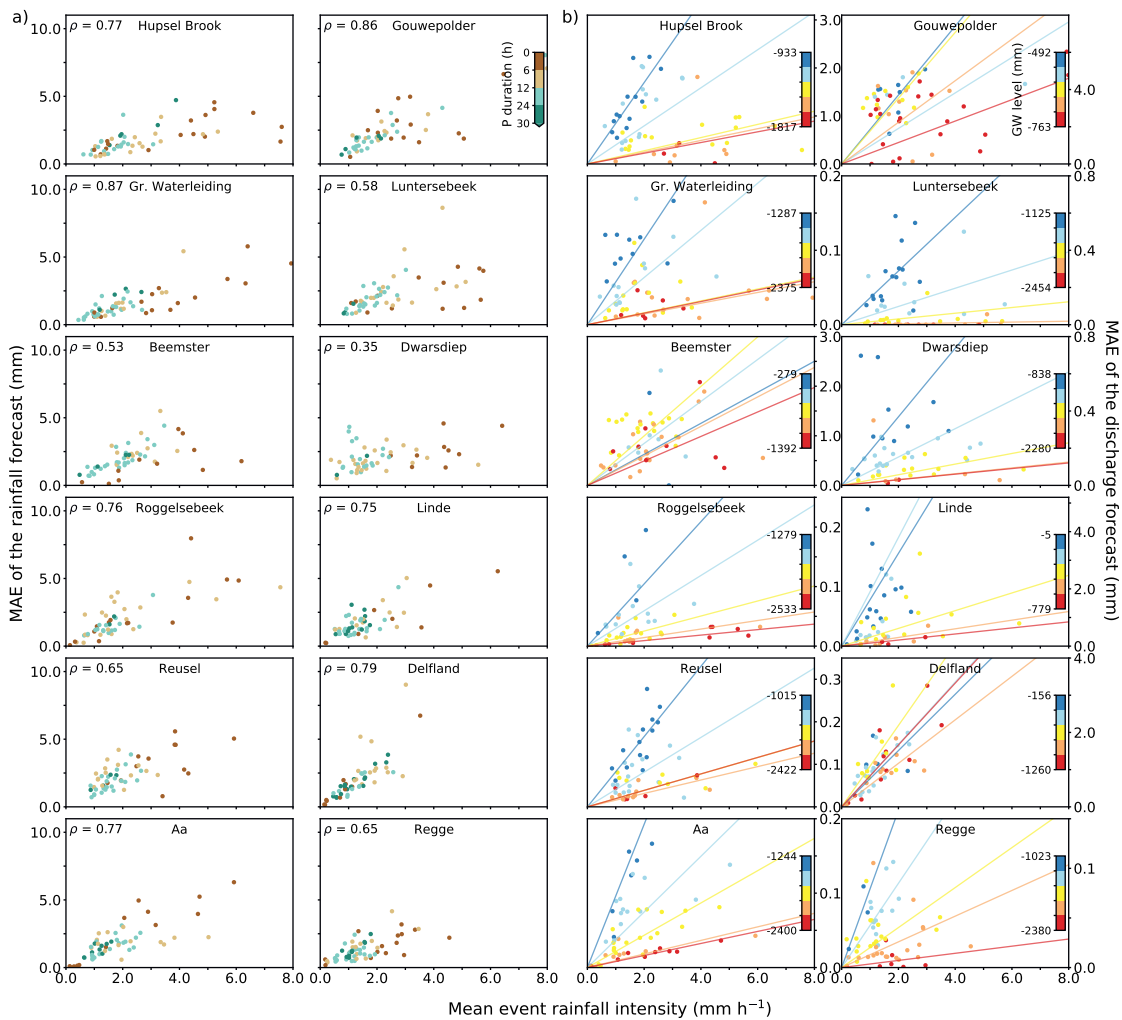


Figure 7.5 | Relationship between the mean event rainfall intensity (based on all rainy 5-min instances) and the MAE of the: (a) rainfall forecast and (b) the discharge forecast per catchment. The rainfall MAE was calculated for the 3-h sum (the first three hours of the nowcast) of the forecast and reference rainfall. The 12-h 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 7.1.

Table 7.2 | Statistics of the linear regression lines (forced through zero) that were adjusted to the data points of Figure 7.6, which corresponds to the mean event rainfall variability dependence analysis. Indicated are the slope of the line and the Pearson correlation coefficient for the regression line Figure 7.6a and per class in Figure 7.6b, going from shallow groundwater table depths (class 1; the blue lines in Figure 7.6b) to the deep groundwater depths (class 5; the red lines in Figure 7.6b).

Catchment	P forecast		Class 1		Class 2		Class 3		Class 4		Class 5	
	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

^aNo (summer) discharge simulated in this class.

most catchments steeper for shallower groundwater depths (Table 7.1). 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 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.

7.3.3.2 | 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 (l_r , Figure 7.6). Overall, the results in this figure are similar to those presented in Figure 7.5. Nonetheless, the catchment size appears to play a role here. In Figure 7.6a, 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 7.2). 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

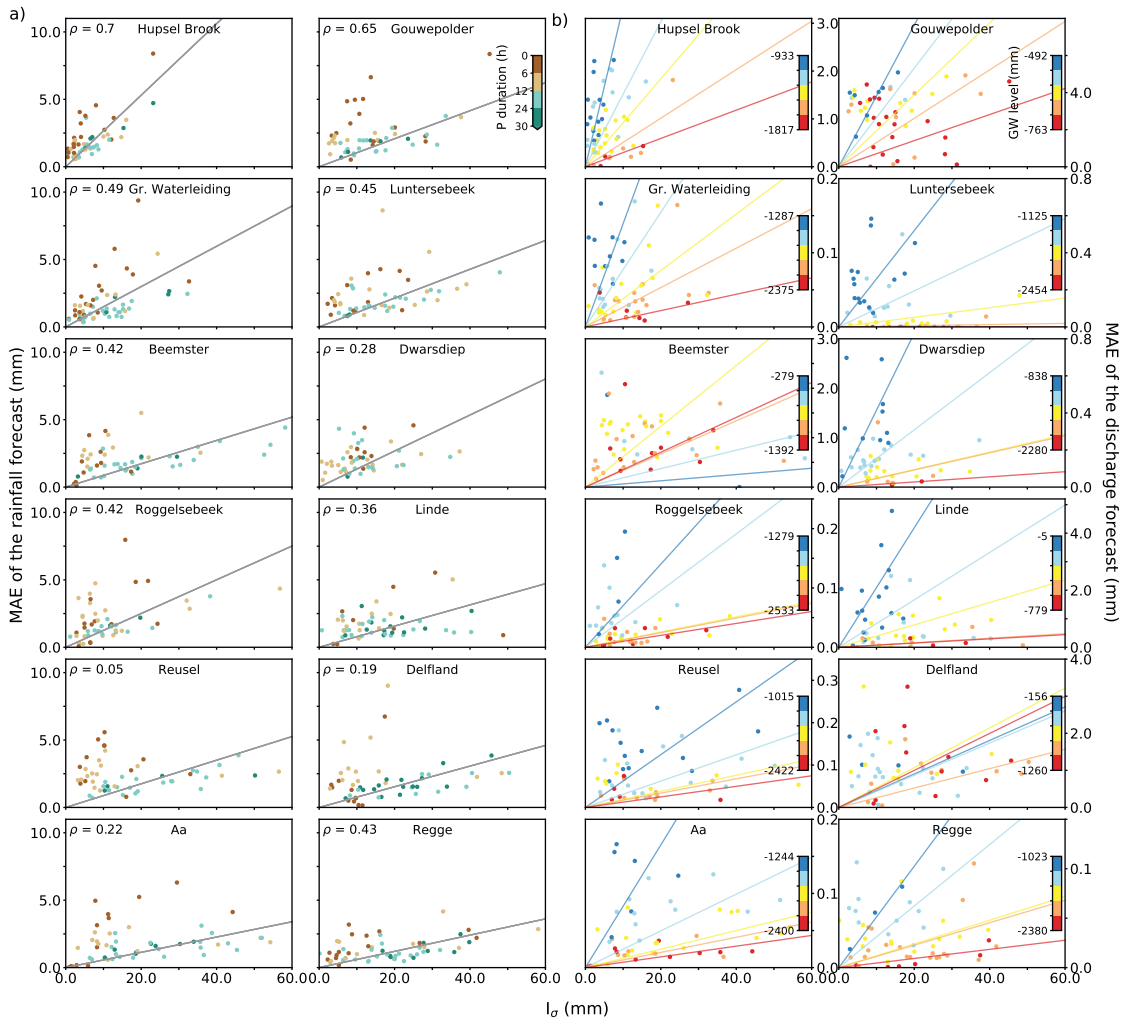


Figure 7.6 | Relationship between the mean event rainfall variability (based on all rainy 5-min instances) and the MAE of the: (a) rainfall forecast and (b) the discharge forecast per catchment. The rainfall MAE was calculated for the 3-h sum (the first three hours of the nowcast) of the forecast and reference rainfall. The 12-h sum (the entire forecast horizon) was used for the discharge MAE. The grey 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 7.2.

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 catchment size dependency is present in the MAE of the discharge forecast (Figure 7.6b), 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 chapter. 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 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.

7.3.4 | Discharge peak forecast verification

7.3.4.1 | Peak anticipation time

Figure 7.7a shows per catchment and method how many hours before the highest discharge during the events, a “correct” forecast, i.e. a hit (having a maximum peak magnitude error of $\pm 10\%$ 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 h 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 h 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, i.e. 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 Chapter 4). 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

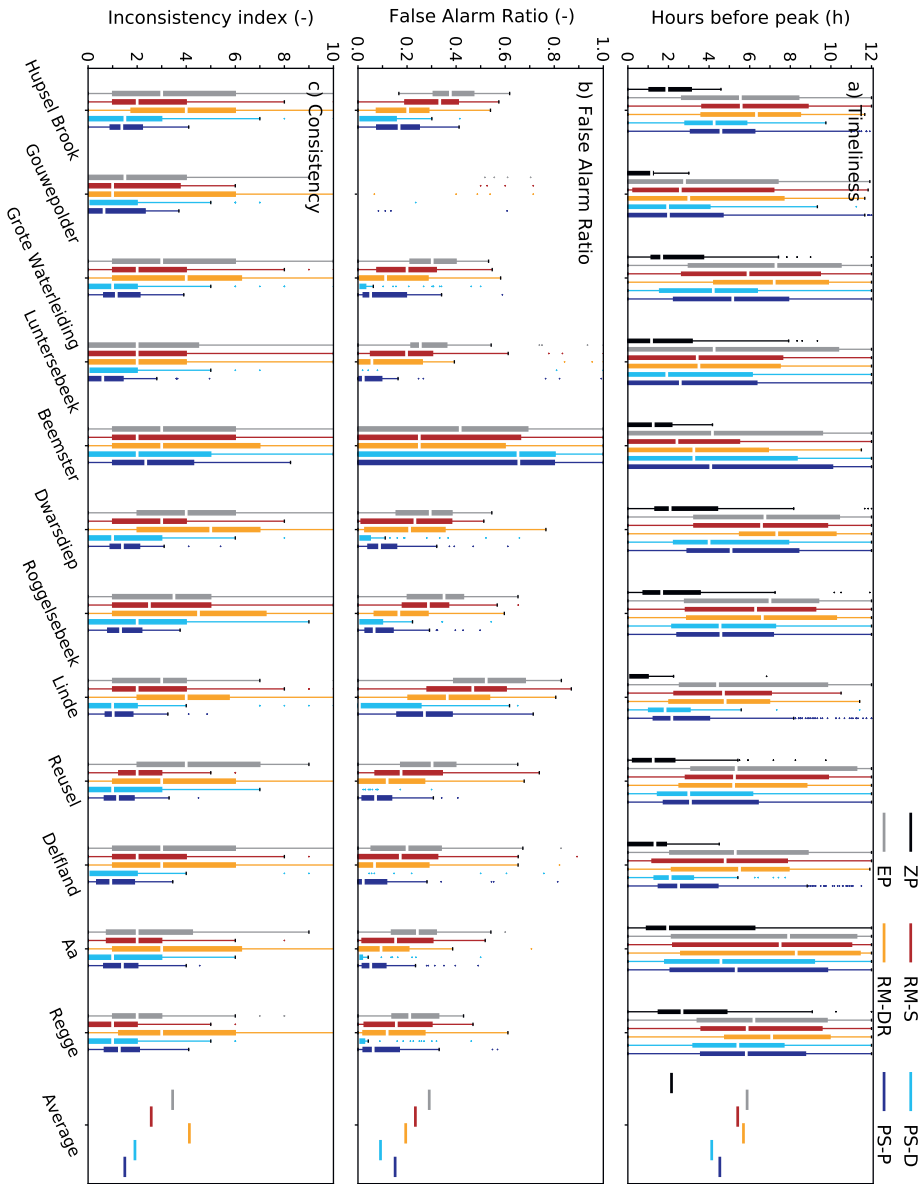


Figure 7.7 | Discharge peak forecast verification per catchment for all methods in this chapter. All events were taken into account, but only the longest duration was selected when there is an overlap between events of different durations (1, 3, 6 or 24 h). Panel (a) shows the timeliness of the peak 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 IQR in colored boxes, the whiskers indicating $1.5 \times$ 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.

forecast rainfall volume is generally crucial in the magnitude of the peak discharge.

7.3.4.2 | 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 chapter was defined as a forecast discharge above the 10% magnitude error of the highest discharge (Section 7.2.3.4). On average for all 12 catchments and the considered events, the false alarm ratio (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 7.7b). 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).

7.3.4.3 | 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 7.7c). Similarly 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 7.1b) 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 7.1).

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.

7.4 | Discussion

7.4.1 | Relation to other studies

In line with Chapter 4, 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 to 170 min compared to a zero precipitation (ZP) forecast. The results of this chapter 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 h^{-1} or more. Heuvelink et al. (2020) studied three catchments that were also present in this chapter, 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 chapter, 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 analyse 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, e.g. 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 chapter 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 7.3.4 and Figure 7.7).

7.4.2 | 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 7.2.3.4, 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 h, which was generally within the range of response times of the 12 catchments considered (Table 2.3). 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 ap-

proach, 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 of an allowed peak magnitude error of $\pm 10\%$ of the difference between the initial discharge at the start of the event and the highest discharge during the event, was subjective and 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, i.e. 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 Appendix Figures E.6 and E.7. The relative differences, i.e. the timeliness compared to benchmark ZP, are small, with e.g. for RM-DR relative gains of 213 min (method in this chapter), 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 chapter), 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 E.6 and E.6, which indicates that the results from this analysis should be interpreted in a relative sense, i.e. compared to a benchmark or to the other nowcasting methods.

7.4.3 | Transferability of results to other regions

This chapter focused on the Netherlands, with its typical lowland catchments and polder systems. Although we expect that the results in this chapter 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 (Gabella et al., 2000; Borga, 2002; Anagnostou et al., 2010). In addition, growth and decay processes, the pitfall of the tested nowcasting methods in this chapter, dominate over advection in these regions (Foresti & Seed, 2015; Foresti et al., 2019), 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 chapter 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. Sharif et al., 2006; Liguori et al., 2012).

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, i.e. 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 chapter. However, to test this, a fully distributed or semi-distributed (containing a sufficient number of sub catchments to capture the catchment

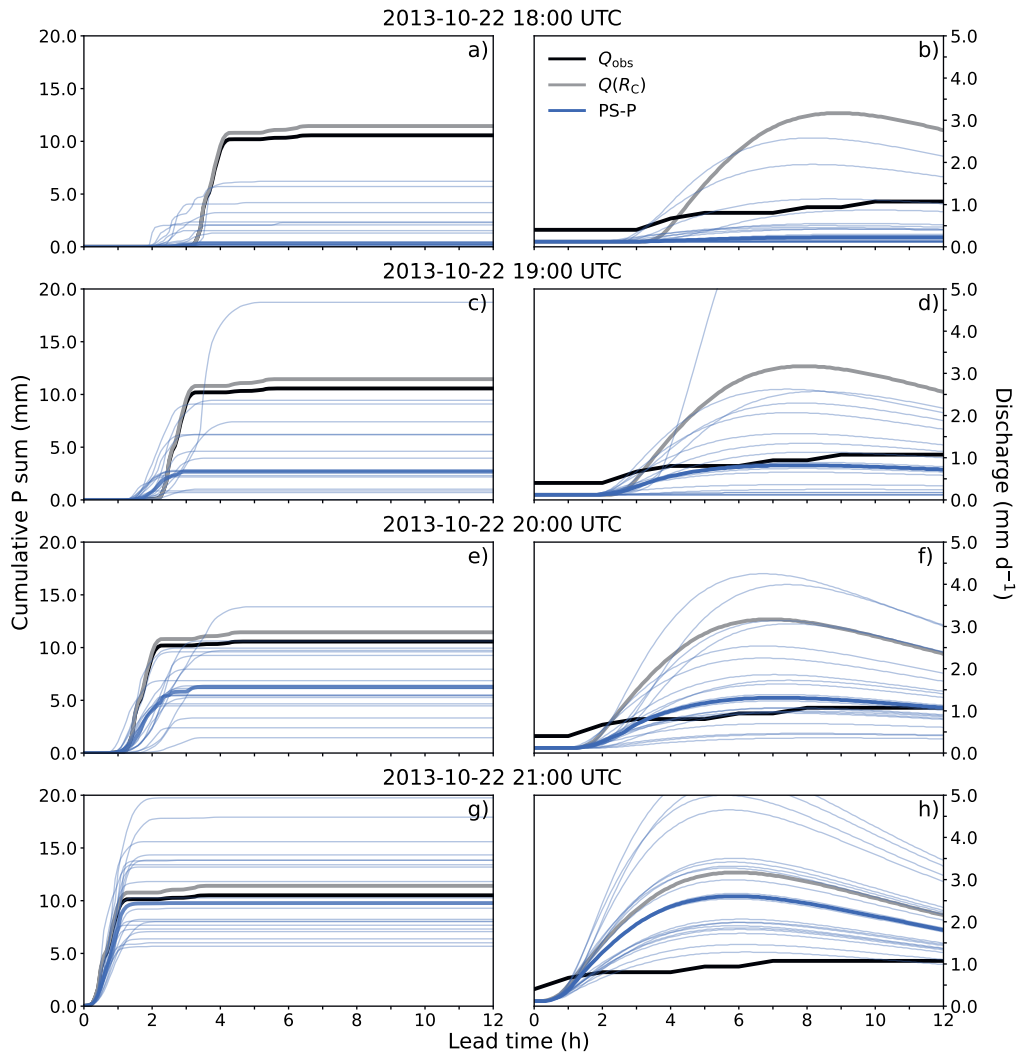


Figure 7.8 | Rainfall and discharge forecasts for four issue times during the discharge peak that occurred in the Hupsel Brook catchment on 22 October 2013. Shown are the rainfall and discharge observations (Q_{obs} ; black), the CARROTS-corrected QPE ($Q(R_C)$; used as reference in this chapter) and subsequent hydrological simulation (grey), 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 7.3 provides the results for this event for the other nowcasting methods considered in this chapter.

heterogeneity) hydrological model should be used, similar to the setup in Lobligeois et al. (2014), which is a recommendation for future analyses.

7.4.4 | Actual error with observations

Finally, the reference in this chapter was a model run with the observed CARROTS-corrected radar data (Chapter 3). This approach discarded any hydrological model and radar QPE errors, as actual discharge 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 7.3 (only PS-P shown this time for clarity) is compared to the observed discharge in Figure 7.8. Although the difference between the CARROTS-corrected QPE and observed rainfall were minimal (Figure 7.8a, c, e and g), 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.

7.5 | 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 chapter, nowcasts for 659 individual events from Chapter 4 were used to construct discharge forecasts. The events were systematically selected for all four seasons, four durations (1, 3, 6 and 24 h) and 12 catchments in the Netherlands, with sizes varying from 6.5 to 957 km². Four open-source nowcasting algorithms were tested: *rainymotion Sparse* (RM-S), *rainymotion DenseRotation* (RM-DR), *pysteps deterministic* (PS-D; similar to S-PROG) and *pysteps probabilistic* (PS-P) with 20 ensemble members. In addition, two benchmark forecasting setups were considered: a hydrological simulation using *Eulerian Persistence* (EP) and a forecast without any precipitation input (ZP: *Zero Precipitation*). For every 5-min time step in the considered events, a discharge forecast with a 12-h 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 twelve catchments. The reference for verification in this chapter 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 chapter. 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 chapter.

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 chapter. 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 h 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 behaviour. For trust in the forecast, it is essential that the false alarm ratio (FAR) is low and that the forecast is consistent, i.e. 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) modelled. 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 chapter, but could provide a step towards 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.



8

Synthesis

“Oh, take me back to the start
I was just guessing at numbers and figures
Pulling the puzzles apart
Questions of science, science and progress”

—Coldplay, *The Scientist* (2002)

IN this thesis, the aim is to *identify if and how operational flood forecasting can be improved with (radar) rainfall nowcasting-based techniques*. In the previous chapters, an extensive analysis and several pathways of nowcasting-based short-term rainfall forecasting for flood early warning were presented. In this final chapter, I will discuss the main scientific findings of this thesis (Section 8.1), followed by a look ahead at existing and emerging opportunities, along with my view on these pathways, in Section 8.2.

The research underlying this thesis was originally motivated by a need from practitioners, the Dutch water managers, to receive more accurate rainfall forecasts on the short term (ideally up to six hours ahead, was the idea), with the hope to further improve discharge and water level forecasts for their management areas. The need for accurate short-term forecasts for (flash) flood early warning was (re-)emphasized during recent floods, in particular the July 2021 floods that were described in Chapter 1 and that were part of the analysis in Chapter 6. In Section 8.3, I will come back to this original request from the water management field and translate the main findings of this thesis into a recommendation for practitioners. This is followed by a final concluding remark with respect to the value of a short-term forecasting chain for operational flood forecasting.

8.1 | Main findings of this thesis

8.1.1 | Radar rainfall bias reduction

The starting point of radar rainfall nowcasting, the quantitative precipitation estimates (QPE) that are used as input, generally come with considerable (systematic) biases. In the Netherlands, these errors result in systematic underestimations of on average 55% in real time. Therefore, the first sub-question of this thesis was “*How can real-time radar rainfall bias adjustments be improved?*”. In Chapter 3, we introduced CAROTS (Climatology-based Adjustments for Radar Rainfall in an OperaTional Setting): a set of gridded climatological adjustment factors for every day of the year. The factors were based on a historical set of 10 years of 5-min radar rainfall data and a gridded reference dataset for the Netherlands, and give bias-adjustment factors that vary both in space and in time. A contrast with already existing bias-adjustment methods is that this method is based on historical data and therefore does not rely on real-time rain gauge data availability. This has as advantage that this bias-adjustment method can be operationally used even when insufficient rain gauge data are available in real time. The main disadvantage of using historical data for bias-adjustment factors is that it does not consider real-time biases caused by event-specific phenomena (for instance, weather patterns that are not part of the climatology for that day of the year).

When we focus on country-average rainfall amounts for the Netherlands, the CAROTS-corrected QPE is outperformed by the operationally-used mean field bias (MFB) correction

method. However, as the MFB adjustment factors are spatially uniform, the perspective changes when we focus on the catchment scale for different locations (tested in this chapter in a hydrometeorological testbed with twelve catchments). Close to the weather radars, the MFB adjustment still performs equally well to the CARROTS adjustment method. For catchments that are located further away from the weather radars, however, where systematic biases generally increase, the MFB-adjusted QPE starts to heavily underestimate rainfall volumes, while this bias is almost absent with the CARROTS-adjusted QPE. This indicates that the spatial corrections of CARROTS outperform the uniform MFB method for application areas (for instance catchments) further away from the radars.

As the focus of this thesis is on improving operational flood forecasting, the subsequent step was to apply both tested bias-adjusted QPE products to simulate discharges with the operational rainfall-runoff models of the twelve Dutch catchments (and later on used to test the subsequent QPF for discharge forecasting in Chapter 7). The effects of any biases in the QPE are generally amplified when used to force rainfall-runoff models, which is especially affecting the discharge simulations with the MFB-adjusted product. For all catchments, the discharge simulations with the CARROTS QPE outperform those using the MFB-adjusted QPE.

Hence, for hydrological applications in the Netherlands, the current operational use of a country-wide MFB adjustment may be reconsidered as it often performs worse than the proposed climatological adjustment factor, which can be seen as the minimum benchmark to beat. CARROTS has shown to be an effective alternative and can bridge the gap between the spatial uniform MFB adjustment method and more advanced geostatistical and Bayesian methods when insufficient rain gauges are available in real time to apply these methods.

8.1.2 | Short-term rainfall forecasting

8.1.2.1 | Rainfall forecast skill

To assess the rainfall forecast skill of nowcasting at the catchment level, its dependencies on specific (environmental) characteristics should be known, as indicated by the sub-question in Chapter 1: *“To what extent does the rainfall nowcasting skill at the catchment scale depend on the factors event type, event duration, seasonality, catchment location and catchment size?”*. Chapter 4 reports an assessment of the skill of radar rainfall nowcasting for rainfall and discharge forecasting at the catchment scale (for 12 Dutch lowland catchments) with a large sample of events. This analysis has shown that the rainfall forecasting skill heavily depends on the event duration, with maximum skillful lead times ranging from on average 25 min for 1-h events (generally convective events) to 116 min for 24-h events (generally more persistent stratiform events). Nowcasting is also more skillful than NWP during at least the aforementioned maximum skillful lead times, but often for even longer lead times (Chapter 6). The dependency on the event type and duration also translates into the seasonality of the forecasting skill, with up to three times lower rainfall forecasting errors during winter (predominantly stratiform precipitation) than during summer (more convective rainfall). This reduction in radar rainfall nowcasting skill for higher-intensity, convective summer rainfall is also shown in Chapter 5, where the forecast skill reduces with approximately 50% for a 5.0 mm h^{-1} threshold compared to a 1.0 mm h^{-1} threshold.

In addition, the rainfall forecasting skill increases with increasing catchment size and increases in the downwind direction, with almost a factor two higher forecast skill for catchments downwind

than upwind in the Netherlands (with respect to the prevailing southwesterly wind direction). The more downwind a catchment is located, the more upwind information is available, which is beneficial for nowcasting.

Another sub-question that was posed in Chapter 1, was “*What is the added value of rainfall nowcasting techniques that statistically take into account rainfall field development, compared to persistence-based techniques?*”. In this thesis, pysteps was the most advanced tested method, and was tested both in a deterministic and probabilistic sense. Pysteps takes growth and dissipation statistically into account by assigning different lifetimes to distinct spatial scales, and by adding stochastic noise, which correlates in space and times with the observed rainfall fields. Chapter 4 highlights that more advanced nowcasting methods (here, pysteps) often outperform the other tested methods (in this thesis: the advection-based rainymotion and Eulerian persistence), with skillful lead times that are generally 35–60% longer. An important remark in this chapter is that the pysteps approaches tend to underestimate the true rainfall amounts (especially) for longer lead times, which can impact the subsequent discharge forecasts (see Section 8.1.3 and Chapter 7).

8.1.2.2 | Alternative rainfall data for nowcasting

The underlying data for rainfall nowcasting are commonly QPE products from operational weather radars, but weather radar data are not globally available (Saltikoff et al., 2019; WMO, 2020) and come with considerable biases (Chapter 3). Therefore, at the start of this thesis (Chapter 1), we stated the sub-question “*How can alternative rainfall information sources be used for rainfall nowcasting?*”. Chapter 5 demonstrates that ensemble nowcasts can also be constructed with country-wide rainfall maps as estimated with received signal level data from commercial microwave links (CMLs) in the Netherlands.

In a test case for twelve summer days in the Netherlands in 2011, Chapter 5 illustrates that for low rainfall intensities, radar rainfall nowcasts still outperform the CML-based nowcasts, due to a more coherent advection field and more detailed rainfall structures in the QPE product. As radar QPE often tends to underestimate rainfall volumes (see Section 8.1.1 and Chapter 3), however, this difference reverses for forecasting higher-intensity rainfall events, a limitation that was also highlighted in Chapter 4. The CML QPE and resulting nowcasts give more accurate rainfall volumes, making this data source suitable for real-time nowcasting of high-intensity rainfall events (provided that a sufficient calibration of the CML rainfall estimation algorithm is possible).

Two main limitations are present when using CML QPE for nowcasting: (1) the data generally has to be upscaled to coarser grid cell sizes than state-of-the-art radar data, due to regionally low CML densities, and (2) the absence or low density of CMLs in some regions and over (inland) water bodies hampers the rainfall advection derivation, which is a vital part of most nowcasting systems. Despite these remarks, Chapter 5 clearly indicates a potential for CML-based nowcasting, which suggests that the potential added value of rainfall nowcasting is not only limited to locations with weather radars. This, in theory, can facilitate nowcasting (or rainfall forecasting in general) and possible short-term (flash) flood forecasting (see also Chapter 7) in urban or relatively densely-populated regions worldwide.

8.1.2.3 | Extending the maximum skillful lead time

The first few hours ahead (in the order of 6 h) in rainfall forecasting are crucial for flash flood forecasting. Although Chapters 4 and 5 indicate that nowcasting advances our ability to make accurate forecasts for this time scale, both chapters also indicate that the methods still lose skill too quickly (generally already after 2 h). Thus, additional steps are necessary to provide skillful rainfall forecasts within the flash flood early warning time scale. Hence, *“To what extent can blending between nowcasts and NWP extend the skillful lead time of rainfall forecast?”* (Chapter 1). One way to bridge the gap between nowcasting and short-range NWP model forecasts is by combining both products, so-called blending. In Chapter 6, the STEPS blending approach (Bowler et al., 2006; Seed et al., 2013) was implemented in the open-source pysteps framework (Pulkkinen et al., 2019). The approach and resulting rainfall forecasts were tested on both the radar-domain scale and the catchment scale (for the Belgian and Dutch catchments Vesdre, Demer, Geul and Dommel).

At the radar domain scale, the tested blending approach performs on par with or better than ensemble nowcasting with pysteps and a deterministic NWP rainfall forecasting model, and also generally outperforms a linear blending approach (the benchmark in Chapter 6). This also holds for higher intensity rainfall cells, although the difference between nowcasting, linear blending and STEPS blending is smaller than for lower intensities.

At the catchment level, both the linear and STEPS blending approaches result in lower forecast errors than nowcasting for lead times of approximately 4 h or longer (depending on the rainfall type). Both methods outperform the NWP forecasts for the first few hours of the forecasts, followed by a similar skill for longer lead times. Overall, the STEPS blending approach outperforms the other tested methods for most of the tested events, but the difference, in particular, with the linear blending benchmark, reduces when we focus on forecast cumulative rainfall sums (up to 6–12 h ahead) instead of on instantaneous rainfall rates. These cumulative rainfall sums (volumes) matter when we move from rainfall forecasting to discharge forecasting, as described in Chapter 7 and in the next Section 8.1.3.

In conclusion, rainfall nowcasting can advance short-term rainfall forecasting up to at most a few hours ahead at the catchment scale, which depends on the factors event type, event duration, seasonality, catchment location and catchment size, but also on the employed nowcasting algorithm (Chapters 4 and 5). This forecast horizon can be extended with an adequate blending method between nowcasts and NWP (Chapter 6). In the absence of radar rainfall estimates, QPE from CMLs can be used for rainfall nowcasting as well and this has shown to be a well-performing source for higher intensity rainfall estimation and forecasting (Chapter 5).

8.1.3 | Short-term flood forecasting

The potential application of rainfall nowcasting for (flash) flood forecasting raised the sub-question *“What are the strengths and limitations of rainfall nowcasting for discharge forecasting and how does this relate to the rainfall forecasting skill of the previous theme?”* in Chapter 1. From Chapter 7, we learn that the dependencies of the rainfall nowcasting skill on the aforementioned environmental characteristics (Chapter 4 and Section 8.1.2) are also present in the subsequent discharge forecasts. Particularly the rainfall event characteristics and hydrological initial conditions determine most of the forecast quality. With increasing rainfall intensity, the skill of both rainfall

and discharge forecasts decreases. This error in the rainfall nowcasts increases nearly linearly with rainfall intensity. Besides the rainfall intensity, however, the discharge forecast error also depends on the initial groundwater table depths. For shallower initial groundwater depths, the discharge forecast error is generally more pronounced and shows a steeper increase with rainfall intensity, than for drier initial conditions.

In addition, the spatial rainfall variability (which is generally higher for convective events) plays a role in both the rainfall and discharge forecasts. Similar to the rainfall intensity, the errors in both the rainfall and discharge forecasts increase with increasing spatial variability of the rainfall events. This effect is stronger for smaller catchments, where a high spatial variability of the rainfall fields can make it more challenging to forecast whether the rainfall will fall inside or outside the catchment, eventually leading to a discharge peak, or not. For larger catchments, this is less of a concern, as the rainfall likely falls somewhere in the catchment 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 are used in Chapter 7.

Similar to the rainfall forecast skill, one sub-question in Chapter 1 focused on the determination of the effect of the nowcasting algorithm on the discharge forecast skill: *“What is the added value of rainfall nowcasting techniques that statistically take into account rainfall field development for (peak) discharge forecasting, compared to persistence-based techniques?”*. Corresponding to the results in Chapter 4, the more advanced methods generally outperform the other tested methods for flood forecasting as well, although the difference, especially between the advection-based rainymotion DenseRotation approach and the tested pysteps approaches, is less than in Chapter 4. The finding that the pysteps approaches tend to underestimate the true rainfall amounts, however, becomes more apparent in the discharge forecasting step with these nowcasts in Chapter 7. Especially with a focus on discharge peak or threshold forecasting, the discharge volumes start to play an important role. The aforementioned tendency of pysteps to underestimate the true rainfall volumes leads to a peak anticipation time (the issue time before a peak when the peak is captured by the forecast within a given margin) that is generally smaller than the exclusively advection-based rainymotion models and even lower than Eulerian persistence. Hence, this suggests that from a (flash) flood forecasting perspective, a focus on volume-preservation in nowcasting methods is warranted. Nevertheless, all tested nowcasting methods manage to forecast peak discharge at least two to three hours earlier than without a forecasting method.

In addition, the resulting gain in anticipation time is only part of the desired model behaviour. A forecaster needs trust in the forecast, and that is only achieved when the false alarm ratio (FAR) is low and the forecast is consistent, i.e. when subsequent forecasts do not often disagree with each other concerning the (non) exceedance of a threshold. When focusing on these aspects, we can conclude that the more advanced nowcasting methods generally result in both a lower FAR and inconsistency index, 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 (with pysteps), as indicating the uncertainty associated with the forecast leads to a lower inconsistency than the binary view of the tested deterministic forecasting methods.

Hence, all tested nowcasting methods advance short-term flood forecasting compared to issuing no rainfall forecasts at all, but the tested methods in Chapter 7 all have their shortcomings.

For reliable and valuable discharge forecasts on the short term, a focus on volume preservation (a main limitation in rainfall nowcasting for discharge forecasting), for instance through better capturing growth and dissipation processes, or blending with NWP (Chapter 6), in nowcasting is warranted and the use of ensemble nowcasts is recommended.

8.2 | Synthesis and outlook on rainfall nowcasting for flood early warning

8.2.1 | Stronger focus on the input data

It is generally acknowledged that poor data quality results in poor modelling and forecasting results, and this holds, in particular, for rainfall-runoff modelling (Borga et al., 2006; Sampson et al., 2014). As data, or the observations, are the starting point of the forecasting chain, the initial focus should be on that. Radar rainfall estimates are a typical example of observations with a large potential, because of their large spatial coverage and fine spatio-temporal resolution. At the same time, radar provides remotely sensed observations of the atmosphere, which is by no means a direct measurement of the rainfall that actually reached the ground. This means that a conversion, an estimation, has to be made to move from measured radar reflectivities aloft to rainfall estimates at the ground. This estimation process produces a number of issues and errors, which were already mentioned in Chapter 1 (Section 1.4): sources of error related to the radar reflectivity measurements, errors in the conversion from radar reflectivity to QPE and spatio-temporal sampling errors (Austin, 1987; Joss & Lee, 1995; Gabella et al., 2000; Sharif et al., 2002; Uijlenhoet & Berne, 2008; Zawadzki, 2018; Ochoa-Rodriguez et al., 2019). These errors can be amplified when used for discharge forecasts in hydrological models, as shown in Chapter 3, and by Borga (2002), Borga et al. (2006) and Brauer et al. (2016). The result is that many radar-based studies, for instance nowcasting studies, use the biased radar QPE as reference instead of the true rainfall amounts, as was also the case in Chapters 4–6 of this thesis. This is not necessarily a problem in a scientific testing environment, but it is exemplary for the fact that these methods remain in the research and development phase, rather than in a full operational phase.

Fortunately, correction methods to improve radar QPE are available, as was also mentioned in Chapter 3. These methods are a combination of (1) techniques to correct the radar observations prior to the reflectivity to rainfall rate conversion, for example corrections for physical phenomena such as ground clutter, attenuation, the vertical profile of reflectivity and variations in raindrop size distributions, but also fully exploiting dual polarisation techniques (Joss & Pittini, 1991; Germann & Joss, 2002; Berenguer et al., 2006; Cho et al., 2006; Uijlenhoet & Berne, 2008; Kirstetter et al., 2010; Qi et al., 2013; Hazenberg et al., 2013, 2014), and (2) post-processing steps to adjust the radar QPE by making use of for instance rain gauge measurements (see Ochoa-Rodriguez et al., 2019, for an overview of these methods). As many of the post-processing methods are not sufficient or not feasible in an operational setting, we introduced a new and easy-to-apply radar rainfall bias adjustment method (Chapter 3).

All these methods are not comprehensive enough to fully reduce the errors in the radar QPE, but combined, they can provide a good step in this direction. Hence, this should be the first step in data processing prior to considering for example nowcasting techniques.

8.2.2 | Pathways for improved rainfall forecasts

Apart from an improved radar QPE quality, the continuous model developments in both the nowcasting and NWP realm help to further improve the rainfall forecasting quality. Chapters 4 and 7 identified the strengths and weaknesses of state-of-the-art nowcasting methods and point out that convective events are particularly challenging to forecast well for more than 30 minutes into the future. Except for Chapter 6, where a well-known blending method was implemented and new functionalities were added, most chapters have mainly identified points of improvement for short-term rainfall forecasting. The main point of improvement remains forecasting convective rainfall cells. In the following sections, a number of initiatives are highlighted that give directions to producing improved short-term rainfall forecasts, particularly with a focus on convective events. In addition to these developments, the movement to more open-source models, such as rainymotion (Ayzel et al., 2019b) and pysteps (Pulkkinen et al., 2019) in this thesis, can accelerate model improvements and collaborations and is therefore a recommended way forward.

8.2.2.1 | Numerical weather prediction improvements

One initiative is the movement to convection-permitting rapid-update-cycle NWP models, which generally have a shorter range of for instance only 12–24 h ahead, which makes it computationally feasible to run these models at a higher spatio-temporal resolution and update frequency. An example is the High-Resolution Rapid Refresh (HRRR) model of NCEP, which is updated every hour and runs at 3-km spatial resolution (Benjamin et al., 2016). HRRR models have shown to be beneficial in the United States (Turner et al., 2022). In addition, time-lagged ensembles can improve the short-term rainfall forecasting results from NWP models (Lu et al., 2007), for example when a large ensemble run is not feasible due to computational and time constraints.

As already mentioned in Chapter 1, some of the limitations of NWP models originate from computational capacity issues. To partly overcome this issue, it helps when (hydro)meteorological offices across smaller countries work together, combine their knowledge and share computational power to facilitate computationally demanding and rapid-update-cycle NWP models. A recent step to share a supercomputer and facilities by the meteorological offices of Denmark, Iceland, Ireland and the Netherlands is a good example of a move in this direction (KNMI, 2022).

8.2.2.2 | Improved nowcasting techniques

The nowcasting field has made some progress in improving convective rainfall forecasts, too. Object-oriented nowcasting methods, sometimes referred to as centroid-tracking algorithms, can follow individual storm cells and model growth, dissipation, splits and mergers of these cells based on the past radar reflectivity observations in the full 3D scan of the weather radar. TITAN (Dixon & Wiener, 1993) is an example of this and has shown to be advantageous in convective conditions (Dixon & Wiener, 1993; Ebert et al., 2004; Han et al., 2009). Another pathway is to consider more predictors than only (2D) rainfall fields in a nowcasting system. Examples are the ANVIL system by Pulkkinen et al. (2020), which uses the Vertically Integrated Liquid (VIL) content as an additional predictor and uses an auto-regressive process (similar to S-PROG and STEPS; Seed, 2003; Bowler et al., 2006) on the VIL estimations instead of on the radar QPE to model growth and dissipation. The NCAR Auto-Nowcast System (ANC; Mueller et al., 2003)

goes even a step further, and uses boundary layer information, storm and cloud characteristics (from various remote sensing sources) to identify boundary layer convergence lines. This information, combined with a fuzzy logic routine, is then used as additional information about storm initiation, growth and dissipation on top of the extrapolation nowcast.

A reason to keep working with just the 2D rainfall fields, is that it is less computationally demanding, it does not limit us to radar data, and we can make use of post-processed or even bias-adjusted rainfall fields, which should have fewer artefacts. The STEPS algorithm, which was used in Chapters 4–7 of this thesis, can for instance be further improved to better simulate small-scale (convective) structures. A first attempt to do so was made by Nerini et al. (2017b), who introduced the Short-Space Fourier Transform (SSFT) as a replacement of the Fast Fourier Transform in the original STEPS and pysteps systems. This approach takes the spatial heterogeneity of the rainfall statistics better into account, which shows to be advantageous for rainfall forecasting for, among others, convective rainfall systems (Nerini et al., 2017b). In addition, Pulkkinen et al. (2021) used SSFT with a spatially variable marginal distribution, along with a convolution kernel, in pysteps (the LINDA algorithm) and show that this manages to better preserve the anisotropic and small-scale structures than the original STEPS implementation.

8.2.2.3 | Blending approaches

On top of the previous developments, a blending approach can get most out of both the NWP and nowcasting developments. See for instance Chapter 6, which has shown that a blending method, which takes into account the initial skill of both NWP and nowcasting components (the STEPS blending method in this chapter; Bowler et al., 2006; Seed et al., 2013), generally outperforms a simple linear blending approach. Finding the optimal weights based on the initial skill and unknown future skill can be further improved with for instance Bayesian (Nerini et al., 2019) or deep learning methods. In addition, a blending method that preserves, and more heavily weighs, high-intensity rainfall amounts of both products (nowcasts and NWP), might be of interest for forecasting more extreme situations in a (flash) flood forecasting system. The saliency-based blending method by Hwang et al. (2015) is an interesting step in this direction.

Chapter 6 also indicated that a grid cell-based verification score does not necessarily favour rainfall volumes at the catchment scale, which is not beneficial for the blending results at this scale. Therefore, it is recommended to try out a blending approach that either determines the skill on a coarser resolution in space and time, which would favour the NWP forecast more and penalises (minor) displacement errors less (Gangopadhyay et al., 2004; Mittermaier, 2006), or that focuses on rainfall volumes over a longer aggregation period (for instance 1, 3 or 6 h) for specific areas, such as catchments and urban areas.

Hence, there are various pathways to improve both the NWP rainfall forecasts and nowcasting. Deep learning methods are another way to further improve forecasting, particularly for nowcasting which is already statistics-based anyway. Section 8.2.3 will deal with the opportunities, and limitations, of machine learning for nowcasting.

8.2.3 | The pleasure and burden of artificial intelligence for nowcasting

Artificial intelligence methods for nowcasting are gaining popularity, in particular machine learning (e.g. French et al., 1992; Han et al., 2017; Sprenger et al., 2017; Ukkonen et al., 2017;

Shehu & Haberlandt, 2022) and deep learning methods (e.g. Shi et al., 2017; Ayzel et al., 2019a; Foresti et al., 2019; Chen et al., 2020; Franch et al., 2020; Ravuri et al., 2021; Cuomo & Chandrasekar, 2022). These methods have quite some potential for short-term rainfall forecasting. As nowcasting is an observation-based statistical method, it is a logical step to replace a ‘conventional’ nowcasting system with a machine learning method, which we can regard as a statistics-based black box model. Compared to ‘conventional’ nowcasting, machine learning methods have as advantage that they can offer reduced run times. This makes it for instance feasible to frequently run large ensembles. The algorithms can also be trained to perform well for specifically high-intensity rainfall events (Franch et al., 2020), and it is possible to include other information sources along with radar QPE, even when the initial relationship between these information sources and the current and future QPE is not (well) known. This latter option shows similarities with for instance the NCAR ANC system (Mueller et al., 2003).

Unfortunately, there are also disadvantages to machine learning methods for short-term rainfall forecasting. To train a machine learning model well, a large, ideally endless, historical dataset of radar rainfall estimates should be present, with a minimal amount of errors in the dataset. Although radar archives are increasing and nowadays at least 40 countries have a radar archive of ten years or more (Saltikoff et al., 2019), this limits the applicability of machine learning for nowcasting to a few countries. However, data augmentation techniques and the upcoming field of transfer learning may help here (Zhuang et al., 2021).

Nevertheless, the larger the archive that is used to train the deep learning algorithms and the higher the resolution of the data, the greater the computational requirements become. Ravuri et al. (2021), for instance, needed 2,500 computer hours on the Google Cloud Tensor Processing Unit (TPU) to train their model on less than three years of 5-min radar data for the UK. Besides, many machine learning methods for nowcasting have a smoothing character due to the skewed distribution of rainfall, which leads to underestimations during high-intensity rainfall (Franch et al., 2020), even when a large dataset is used for training. This, however, to a lesser extent also holds for ‘conventional’ nowcasting methods, as was shown in Chapters 4 and 7 of this thesis.

The way forward in machine learning for rainfall nowcasting is, in my opinion, to use these techniques for processes that we cannot capture well with ‘conventional’ nowcasting techniques. A typical process that is simulated well by nowcasting algorithm nowadays, is the advection of the rainfall fields, as derived with optical flow methods (Pulkkinen et al., 2019). Thus, it is more advantageous to keep this process from existing nowcasting methods and add machine learning to the processes that are harder to forecast or simply not well understood on a small scale, such as growth and dissipation of rainfall storms (Chapter 4 and 5). Foresti et al. (2019), for instance, have shown that a deep learning algorithm is capable of learning where growth and decay typically take place under different circumstances in the Swiss Alps. Adding different predictors (for example atmospheric or model variables), as described in the first paragraph of this section, is another way to get the most out of a machine learning algorithm. This especially holds when focusing on growth and dissipation processes in convective cells, which have a short lifetime. Recent rainfall observations do not provide much information about future rainfall cells under these conditions, but other predictors might and a machine learning algorithm can be capable of picking up this information.

8.2.4 | The future of nowcasting

With the development of improved and higher-resolution NWP models, more computational power and the rapid evolution of machine learning models, one could wonder if ‘conventional’ nowcasting, as predominantly used in this thesis, will have a future one or two decades from now. Nowcasting is placed in between physics-based forecasting and the black box of deep learning methods, which makes the system statistical and fast, but still understandable and also applicable whenever a smaller archive of spatial observations is present (in contrast to deep learning methods).

Some pathways to further improve nowcasting methods were already discussed in Section 8.2.2, but ultimately the two systems (nowcasting and NWP) become intertwined. This basically could be the next step in blending or data assimilation, where NWP rainfall forecasts are initialised by the rainfall nowcasts, as, for instance, already done by Golding (1998) and Honda et al. (2022), and the nowcast is later on updated and blended with the NWP forecast (that will likely have a lower update frequency than the nowcasting system). Besides, not only the NWP forecasts continuously move to finer resolutions, so do the spatial observations with for example radar (e.g. Heinzelman & Torres, 2011; Honda et al., 2022). The issues of NWP rainfall forecasts for flood early warning, already mentioned in Chapter 1, will never entirely disappear, but rather reduce over time, due to the always present chaos problem (Lorenz, 1993). Therefore, in my opinion, nowcasting can always add additional skill and information to operational flood forecasting systems, whether that is through ‘conventional’ or machine learning methods and stand-alone or intertwined with an NWP model.

8.2.5 | Taking advantage of present-day spatially distributed hydrological models

In this thesis, all employed hydrological models were lumped (WALRUS) and semi-distributed (SOBEK-RR; Section 2.4), which makes the model results hardly sensitive, or not sensitive at all, to the location of (the forecast) rainfall in the catchment. Where the rainfall falls in the catchment, i.e. upstream or near the outlet, influences the catchment response, especially in larger catchments with more heterogeneous terrain or in mountainous catchments (Lobligeois et al., 2014), where this effect becomes more pronounced than in the 12 Dutch lowland catchment of this thesis. To take these effects into account, a fully distributed or semi-distributed (containing a sufficient number of sub-catchments to capture the catchment heterogeneity) hydrological model should be used, for example similar to the setup in Lobligeois et al. (2014).

A spatially distributed hydrological model can, in theory, take full advantage of the high spatial resolution that current radar QPE and QPF provide (Sharif et al., 2006; Thorndahl et al., 2017). Simulating the aforementioned effect of the location of rainfall on the catchment response is one advantage (Moreno et al., 2013; Lobligeois et al., 2014). Another advantage is that the higher resolution, and modelled hydrological processes at this resolution, allow us to forecast local flooding and couple results of hydrological models to, for instance, inundation models, which could be a good first step towards more impact-based forecasting (see also Section 8.2.6). Some studies have already successfully tested the use of nowcasting as forcing in spatially distributed hydrological models (Vivoni et al., 2006, 2007; Berenguer et al., 2011; Moreno et al., 2013; Sharif et al., 2006; Poletti et al., 2019), indicating that it is logical step to test the use of distributed hydrological models in the presented forecasting chain in this thesis for the Netherlands and Belgium.

The move towards distributed hydrological models for discharge simulations comes, however, at a prize. This move requires us to consider the link between temporal and spatial scales, which is similar to the scaling issue mentioned in Section 1.2 (Moulin et al., 2009; Melsen et al., 2016; Clark et al., 2017). It is potentially feasible to make this step with the current high spatio-temporal resolution of, for instance, radar QPE (Serafin & Wilson, 2000; Overeem et al., 2009b). In addition, model structure, parameter estimation and process representation at this high-resolution grid-cell representation become important and are challenging (Beven, 2006; Samaniego et al., 2010; Archfield et al., 2015; Bierkens, 2015; Paniconi & Putti, 2015; Clark et al., 2016, 2017; Mizukami et al., 2017; Feigl et al., 2020; Imhoff et al., 2020c; Schweppe et al., 2022). Hence, despite the mentioned (potential) advantages of moving towards fully spatially distributed hydrological models in our (short-term) forecasting systems, it brings an additional level of complexity to the forecasting chain, which should be considered prior to making this step.

8.2.6 | Flood forecast dissemination and decision making under uncertainty

This thesis has only focused on three aspects of the flood forecasting chain: the input (radar) QPE, the subsequent rainfall forecasts that force the hydrological models (mainly based on now-casting) and the way this translates into discharge forecasts with lumped and semi-distributed hydrological models. Although the rainfall forecasts are a significant part of and possible error source in this forecasting chain (Moulin et al., 2009; Zappa et al., 2011; Sampson et al., 2014), this is not where the forecasting chain ends and therefore also not the only source of uncertainty in the flood forecasting chain. Other sources of errors originate from the other inputs to the hydrological (and sometimes hydrodynamic) model, such as uncertainty in the initial conditions at the start of the forecast, evapotranspiration forecasts (although this is shown to have a minimal effect on the forecast quality; van Osnabrugge et al., 2019) and other input fluxes (for instance upward seepage and surface water supply). In addition, there are uncertainties in the hydrological model setup, such as parameterization, model structure and model quality under extreme (weather) conditions (Section 8.2.5 and Beven, 1993; Melsen et al., 2016; Clark et al., 2017). These forecasting errors have not been taken into account in this thesis, although Figure 7.8 in Chapter 7 gives an impression of the effect of these errors on the final discharge forecast.

Observational, meteorological and hydrological errors can introduce a high level of uncertainty to a flood forecast (Cloke & Pappenberger, 2009), but the flood forecasting chain, and inherent errors, actually does not end there. A forecast has to be disseminated to an end user (a local or regional authority, crisis manager or inhabitant) and based on this information a decision has to be made whether or not action measures have to be taken, for example an evacuation of a town. This last step, the decision, can on its own already be quite a challenge due to everything that is at stake: people's safety, economic damages due to evacuating or not evacuating and trust in the crisis management in case evacuations happen too often without it being necessary in hindsight (Demeritt et al., 2007). This decision in reality has to be made under uncertain conditions, due to the way the information is disseminated (if the information is present at all), the uncertainty in the discharge forecast, which is in turn a result of the aforementioned sources of observational, meteorological and hydrological errors. The uncertainty in the forecast can be visualized by means of ensemble forecasts (Demeritt et al., 2007), which, in principle, can depict the probability that for example a discharge threshold will be exceeded. This is a useful way to visualize and take into account uncertainty in the forecast (Verkade & Werner, 2011),

as was also done in Chapters 4–7 of this thesis. However, it remains a challenge to interpret and act on these probabilities (Tversky & Kahneman, 1974; Demeritt et al., 2007; Jain et al., 2018), because it is sometimes unclear how a crisis manager is supposed to respond when the probability of threshold exceedance is, for instance, ‘just’ 20%. Hence, this asks in turn for an improved way of disseminating this probabilistic information, communication between different information providers and end users, and training our water (crisis) managers in working with this information (e.g. Werner et al., 2009).

In addition, from the perspective of modellers and forecasters the current information provision may not be sufficient or is simply not understood by the end user. The outcome of our flood forecasting systems are generally discharges or water levels, as was also the case in this thesis. Despite that this is useful information, it leaves out the actual impact a forecast discharge or water level can have on the environment. For a crisis manager or any other end user, it may be more valuable to get a so-called impact-based forecast, which gives an indication of the effect of such a discharge or water level on possible dike breaches, flood extents and damages (Basher, 2006; Merz et al., 2020). Impact-based flood forecasting will shift the focus to the end user and their needs to make well-informed decisions. An impact-based forecasting approach will not reduce the level of uncertainty that occurs within the forecasting chain. In fact, it may even further increase the level of uncertainty. Nevertheless, an impact-based forecast may at least give a better perception of the effect a forecast discharge or water level (and the range in the ensemble forecast) has (Merz et al., 2020). Impact-based forecasting can therefore be seen as the next step in flood forecasting systems.

8.3 | Advice for practitioners in the flood forecasting domain

Implementing a flood early warning system is the way to go, as these systems have shown to be beneficial and a good means to respond (accordingly and timely) to flooding (UNISDR, 2002; Pappenberger et al., 2015). The way these systems are set up, depends on the needs of the user. For short-term discharge forecasts in flood early warning systems, which are generally needed in quickly responding small, urban, mountainous or polder catchments, rainfall nowcasting can provide forecasting skill compared to NWP. The first step in the nowcasting chain will generally be the use of weather radar data in the forecasting system. Radar rainfall estimates can come with substantial biases, something which is known by most practitioners and radar rainfall estimates are, therefore, still not widely used in flood forecasting systems. Chapter 3 of this thesis has highlighted this issue, but has also shown that with a simple real-time applicable climatological correction factor a large part of these biases can be reduced and with that, hopefully, the reluctance to use radar rainfall data.

In principle, the use of radar rainfall data for operational flood forecasting is beneficial due to its high spatial and temporal resolution and possibilities for, for instance, nowcasting. Yet, the use of radar rainfall data requires (bias) corrections, which should be taken into account prior to using this data. Such corrections can consist of simple corrections, as introduced in Chapter 3, but also geostatistical or Bayesian correction methods. The latter post-processing methods require a high density of (automatic) rain gauges, which would advocate for either more rain gauges or alternative rainfall sensors, such as the CMLs in Chapter 5. As mentioned in Chapter 3, corrections prior to the rainfall estimations (for instance, correction for ground clutter, attenuation, variations in the raindrop size distribution and the vertical profile of reflectivity) are possible as

well, and generally require a different product from the (meteorological) organizations supplying them. Hence, the use of radar rainfall data for operational flood forecasting systems requires a consideration of bias reduction steps prior to the usage of the data.

With a good radar rainfall product in place, flood early warning systems that should operate within the space and time domain of flash floods (small, quickly responding catchments that require forecast horizons in the order of six hours), can benefit from rainfall nowcasting methods. In the absence of weather radar information, rainfall estimates from commercial microwave links in cellular telecommunication networks can be used for nowcasting, too (Chapter 5). Chapter 4 has indicated that radar rainfall nowcasts in the Netherlands are, on average, skillful up to 2 h ahead, but this decreases to approximately 30 min for convective rainfall. Especially for highly variable convective situations, an ensemble forecast can be advantageous. For these short forecast horizons, nowcasts are more skillful than NWP forecasts (Chapter 6) and this translates into discharge forecasts that capture threshold exceedances or discharge peaks at least 2–3 h earlier than without a (nowcasting-based) flood early warning system (Chapter 7).

Although this is a considerable time profit and a move in the right direction, nowcasting is not sufficient to provide skillful rainfall and discharge forecasts on the entire flash flood time scale when the response time of the catchment is in the order of or more than those 2–3 h. This is where blending, the combination of for example nowcasting and NWP, comes into play. Optimal blending techniques allow to take advantage of the nowcasting skill on the very short term and to make use of the NWP information for lead times beyond that. Chapter 6 has shown that this combination of both products tends to get the best out of both worlds and extends the skillful lead times of nowcasting beyond the aforementioned two hours. It is, however, important to realize that a blended forecasting system has higher computational and data requirements and ideally contains NWP forecasts with a temporal and, if possible, spatial resolution similar to the (ensemble) nowcasts. This is feasible, but often not provided in real time.

In conclusion, nowcasting-based methods can advance operational (flash) flood forecasting. This thesis has also shown that the entire forecasting chain should be considered when implementing nowcasting in a flood forecasting system. To take full advantage of nowcasting or blended forecasting systems, we should improve our rainfall estimates in real time prior to using nowcasting-based rainfall and discharge forecasting techniques. Or, in simpler terms: one does not go without the other, the forecasting chain is as good as its weakest component. In my opinion, this requires a closer collaboration between water managers, meteorological offices and researchers. A streamlined plan, taking into account all forecasting components, can really advance (flash) flood forecasting. We should not be afraid to use test environments and shadow systems (in which the system runs operationally in a test environment next to the current system), as this allows testing new methods in an operational setting, generally giving different outcomes than an idealised research setting. This can provide a feedback loop between practitioners, developers and researchers, which eventually could tailor our developments more to what is really needed in the flood forecasting domain.

As a final remark of this section, this thesis is a result of a close collaboration between Deltares, Wageningen University & Research, and, to a lesser extent, the Royal Netherlands Meteorological Institute (KNMI) and a steering group consisting of people from the water authorities, Rijkswaterstaat and engineering firms. Therefore, these chapters and the recommendations in

this section have a strong focus on the Netherlands (and to a lesser extent Belgium). I believe, however, that the results from these chapters and the recommendations in this section can give an idea in broader terms about the applicability of rainfall nowcasting for operational flood forecasting elsewhere, provided that we are dealing with similar catchment sizes and response times as in this thesis. Ultimately, this applicability stands or falls with local water management practices and (forecast) communication.

8.4 | Concluding remark

This thesis has shown the potential of (radar) rainfall nowcasting for operational (flash) flood forecasting. A nowcasting-based forecasting system can advance (flash) flood forecasting, especially when optimally combined with NWP forecasts. A main recommendation that follows from the previous chapters (and Section 8.3), is to take the entire forecasting chain into account when moving towards a nowcasting-based operational flood forecasting system. This means that we are not finished after implementing a nowcasting algorithm, but that we should consider the quality and need for bias corrections of the input rainfall estimations, the NWP output (e.g. spatial and temporal resolution) to allow for an optimal blending, and eventually data assimilation, dissemination of forecast results and the uncertainty in these results. I believe that nowcasting can truly advance a (flash) flood forecasting system and once all of the aforementioned components are in place, we can keep improving the short-term (impact-based) forecasting chain.



9

- **Appendices**
- **Bibliography**
- **Summary**
- **Samenvatting**
- **Authorship contribution**
- **Acknowledgments**
- **List of publications**
- **Graduate school certificate**

Appendices

A | Additional results of Chapter 3

Table A.1 shows the country-average FSE between R_A and the three QPE products for every year and the winter and summer seasons. The method to calculate the FSE score is described in Section 3.2.3.

Table A.1 | Country-average Fractional Standard Error (FSE) between the hourly reference rainfall (R_A) and the three QPE products (R_U , R_{MFB} and R_C) per year for the winter (DJF) and summer (JJA) seasons. The FSE was only calculated for hours where the country-average rainfall rate was larger than 0.0 mm h^{-1} .

Season	Year	Avg rain rate (mm h^{-1})	FSE		
			R_U	R_{MFB}	R_C
DJF	2009	0.32	1.10	0.49	0.74
	2010	0.26	1.23	0.61	0.82
	2011	0.38	1.12	0.50	0.73
	2012	0.36	1.09	0.51	0.65
	2013	0.30	1.04	0.56	0.90
	2014	0.33	1.06	0.51	0.72
	2015	0.34	1.04	0.51	0.84
	2016	0.34	1.15	0.61	0.84
	2017	0.37	0.56	0.32	0.44
	2018	0.37	1.22	0.65	0.76
JJA	2009	0.33	1.18	0.80	1.08
	2010	0.43	1.34	0.71	1.02
	2011	0.37	1.31	0.78	1.03
	2012	0.36	1.19	0.72	0.99
	2013	0.36	1.34	0.86	1.20
	2014	0.33	1.37	0.91	1.28
	2015	0.44	1.24	0.69	1.08
	2016	0.30	1.46	1.00	1.46
	2017	0.37	1.29	0.76	1.09
	2018	0.34	1.26	0.78	1.20

B | Additional results of Chapter 4

This supplement contains six figures that complement the sections on ‘Event type dependency’, ‘Seasonal dependency’, ‘Dependency on catchment size and location’, and ‘Ensemble forecast verification’ in the main text of Chapter 4. An extended description of the results, complementing their brief mentioning in the main text, is present for Figures B.3 and B.4 in Section B.1, and for Figure B.6 in Section B.3. Section B.2 gives the a brief introduction to the methodology behind the bias calculations.

The presented results are based on the same model runs and number of events (1536) as in the main text. Both in the text and the figures, the algorithm names are abbreviated conform the abbreviation used in the main text: EP (*Eulerian Persistence*), RM-S (*rainymotion Sparse*), RM-DR (*rainymotion DenseRotation*), PS-D (*pysteps deterministic*) and PS-P (*pysteps probabilistic*).

In Figures B.1 and B.2, the results of an extra analysis with pysteps, as run with only advection implemented, are shown and compared to the other methods (similar to Figures 4.3 and 4.4 in the main text of Chapter 4). This run is from here-on referred to as *pysteps advection* (PS-A). The run was setup with the following configuration: a Lucas-Kanade optical flow method (using the QPE from time $t - 3$ to t) (Lucas et al., 1981) and a backward semi-Lagrangian advection method (Germann & Zawadzki, 2002). As such, this run is somewhat similar to RM-DR.

B.1 | Critical Success Index

As for Figure 4.3 in the main text of Chapter 4, a similar analysis of the maximum skillful lead time is possible with the CSI (described in Equation 2.18 of Section 2.6). Average maximum skillful lead times, seen as the mean of the intersections between the $1/e$ line and the mean CSI of an event, are different in Figure B.3 than in Figure 4.3 of the main text of Chapter 4. Still, PS-D and PS-P outperform the other methods, but the absolute differences in skill between the methods are smaller than in Figure 4.3 of the main text. Especially RM-DR has a performance that is closer to the pysteps algorithms. The average maximum skillful lead times in Figure B.4 for PS-D are: 48 min for a duration of 1-h, 55 min (3-h), 56 min (6-h) and 51 min (24-h). Based on the correlation as metric, this is: 25 min (1-h), 39 min (3-h), 56 min (6-h) and 116 min (24-h). Note that in Figure B.3, a threshold of 1.0 mm h^{-1} is used.

A limitation of the CSI metric is that this score depends on the used metric, see also Figure B.4 for the effect of four different thresholds on the CSI. As the average rainfall intensity clearly varies per duration (see also Table 1 in the main text), it becomes challenging to compare the four durations based on one threshold.

B.2 | Bias determination

The bias in Figure B.5 is a deterministic categorical bias, formulated as:

$$B = \frac{TP + FP}{TP + FN} , \quad (\text{B.1})$$

where TP is the number of true positives, the ‘hits’: both the forecast and observation exceed the threshold. FP is the number of false positives, the ‘false alarms’: the forecast exceeds the threshold, but the observation does not. FN is the number of false negatives, the ‘misses’: the observation exceeds the threshold, but the forecast does not.

The bias is calculated with thresholds of 0.1 and 1.0 mm h⁻¹ per grid cell. The mean biases, as illustrated in Figure B.5, are the mean of all cells in that catchment averaged over all events and catchments.

B.3 | Ensemble spread versus error

The ensemble spread, defined as the standard deviation per grid cell of all rainfall amounts forecast by the (20) ensemble members, is an indication of the wideness of the forecast ensemble. An ensemble with a small spread has a high sharpness. However, if there is high uncertainty, the spread will be larger in order to realistically represent the uncertainties. Ideally, the spread is as small as possible for a sharp forecast, but at the same time, the observed rainfall amount should fall within the ensemble spread. Hence, the ensemble should just be large enough to capture the uncertainties in the forecast. This is reached when the ensemble spread equals the ensemble error, defined as the root mean squared error (RMSE; described in Equation 2.8 of Section 2.6) between the ensemble mean and the observation.

For PS-P, the ensemble spread is on the average already after 5 min in between 20 and 40% of the ensemble error (Figure B.6). This means that in many forecasts, the observed rainfall amount falls outside the ensemble, indicating that the ensemble is underdispersive. This problem can be overcome by increasing the ensemble spread, i.e. by taking more uncertainty into account, or by decreasing the ensemble error. The latter is generally challenging, but a good start could be a bias correction (see Figure B.5).

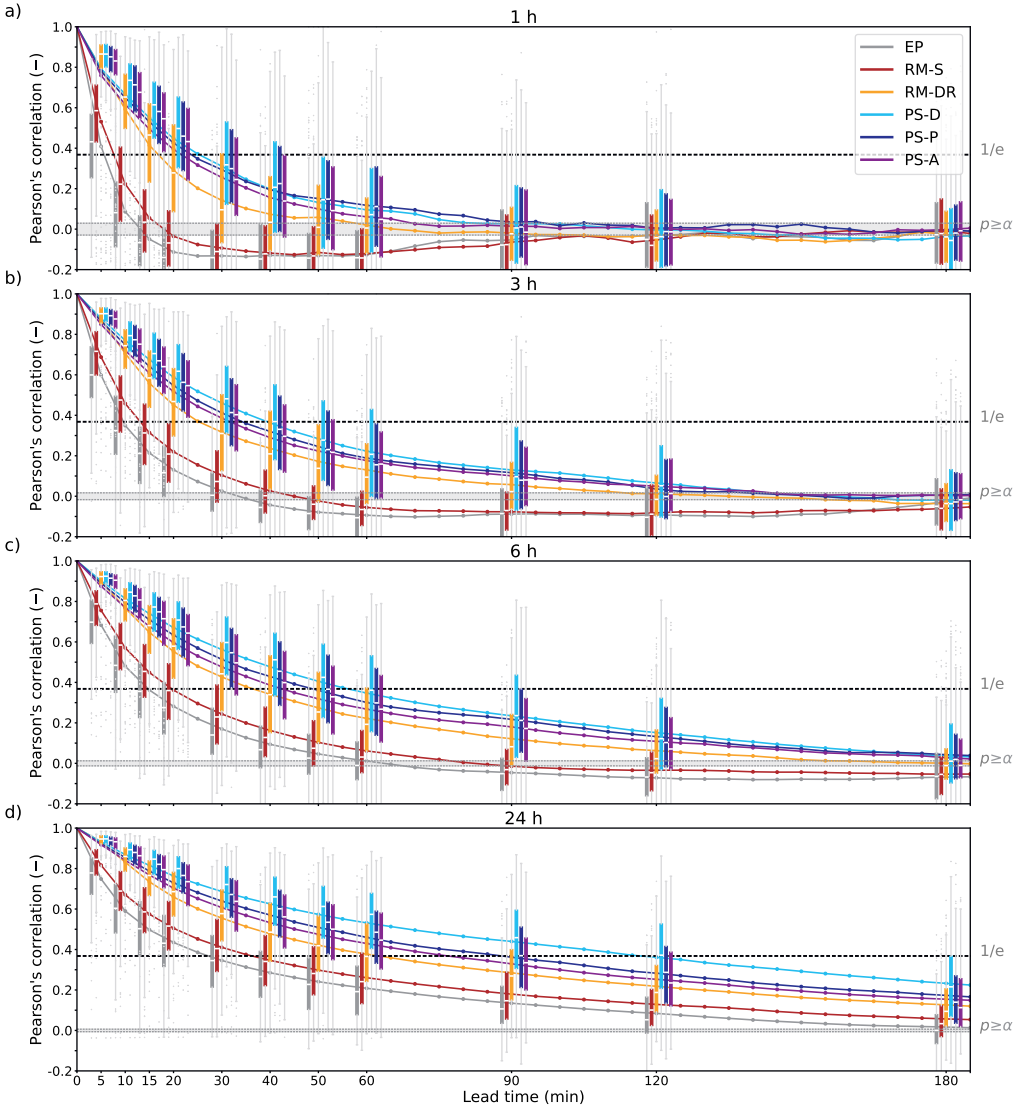


Figure B.1 | Pearson's correlation as a function of lead time (5-min steps), averaged over all cells within the catchment and events (in that order), for event durations of 1-h (a), 3-h (b), 6-h (c) and 24-h (d). This figure is similar to Figure 4 in the main text, but with PS-A added. The dotted line indicates a correlation of $\frac{1}{e}$, the minimum correlation for a skillful nowcast. The boxes indicate the variability in results per event, with: the median in white, the interquartile (25th–75th percentile) range (IQR) in colored boxes, $1.5 \times \text{IQR}$ starting outside the boxes in grey bars and the outliers in grey dots. The horizontal grey band around a correlation of 0.0 indicates insignificant correlations, based on a two-tailed T-test with $\alpha = 5\%$.

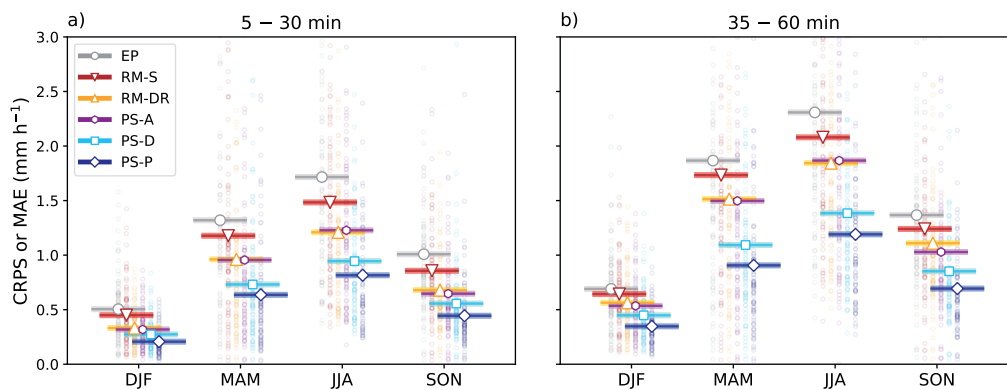


Figure B.2 | CRPS and MAE per season for all events and catchments for the 6-h event duration, averaged over lead times of 5–30 min (a) and 35–60 min (b). This figure is similar to Figure 5 in the main text, but with PS-A added. For all deterministic runs, the MAE is shown and for PS-P, CRPS is shown. The thick lines with a marker indicate the mean CRPS or MAE for all runs and catchments in that season. The scattered points are the mean CRPS or MAE per event.

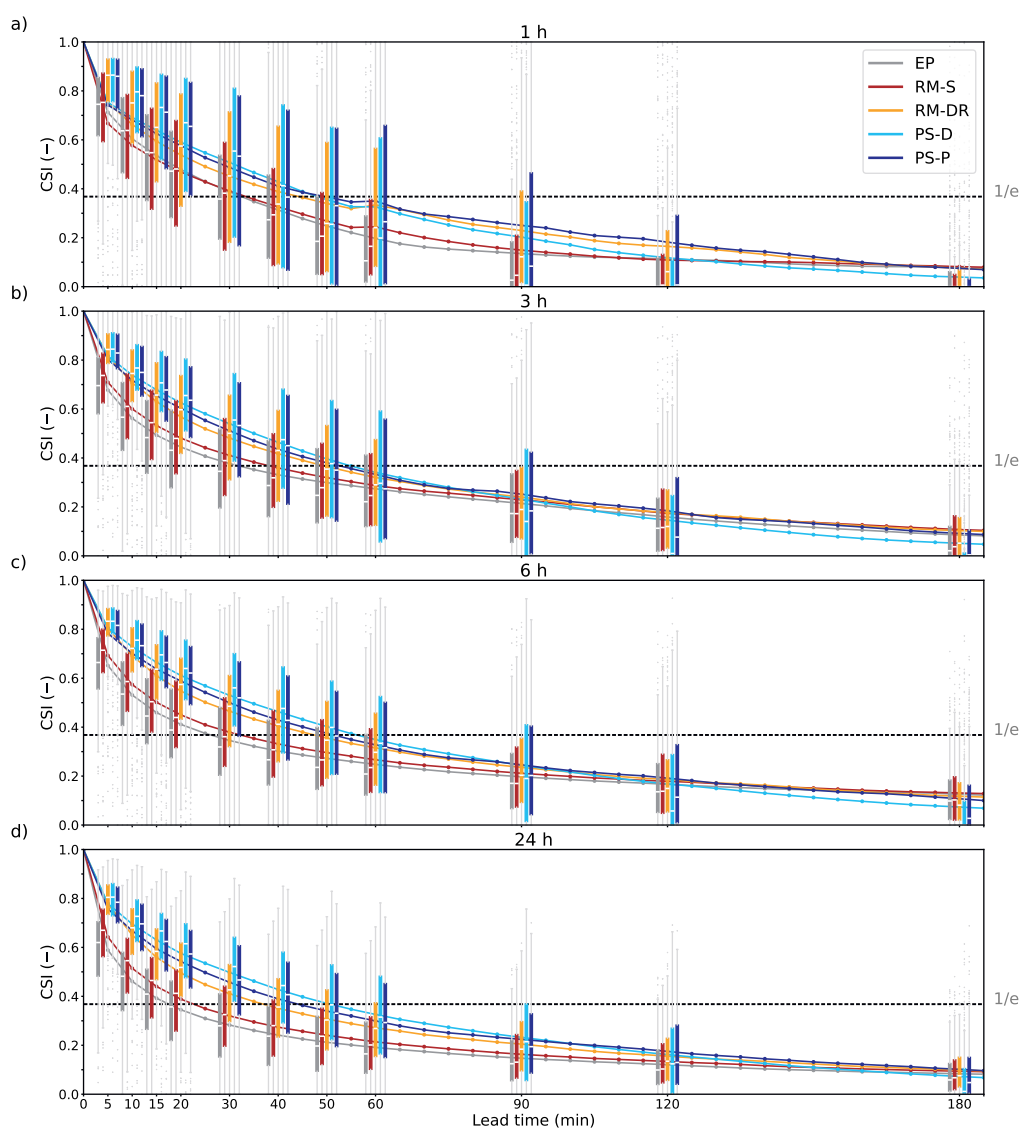


Figure B.3 | CSI score as a function of lead time (5-min steps) for a threshold of 1.0 mm h^{-1} , averaged over all cells within the catchment and events. The CSI is calculated for the events in four durations: 1-h (a), 3-h (b), 6-h (c) and 24-h (d). The dotted line indicates a CSI of $\frac{1}{e}$, the minimum CSI for a skillful nowcast. The boxes indicate the variability in results per event, with: the median in white, the interquartile (25th–75th percentile) range (IQR) in colored boxes, $1.5 \times \text{IQR}$ starting outside the boxes in grey bars and the outliers in grey dots.

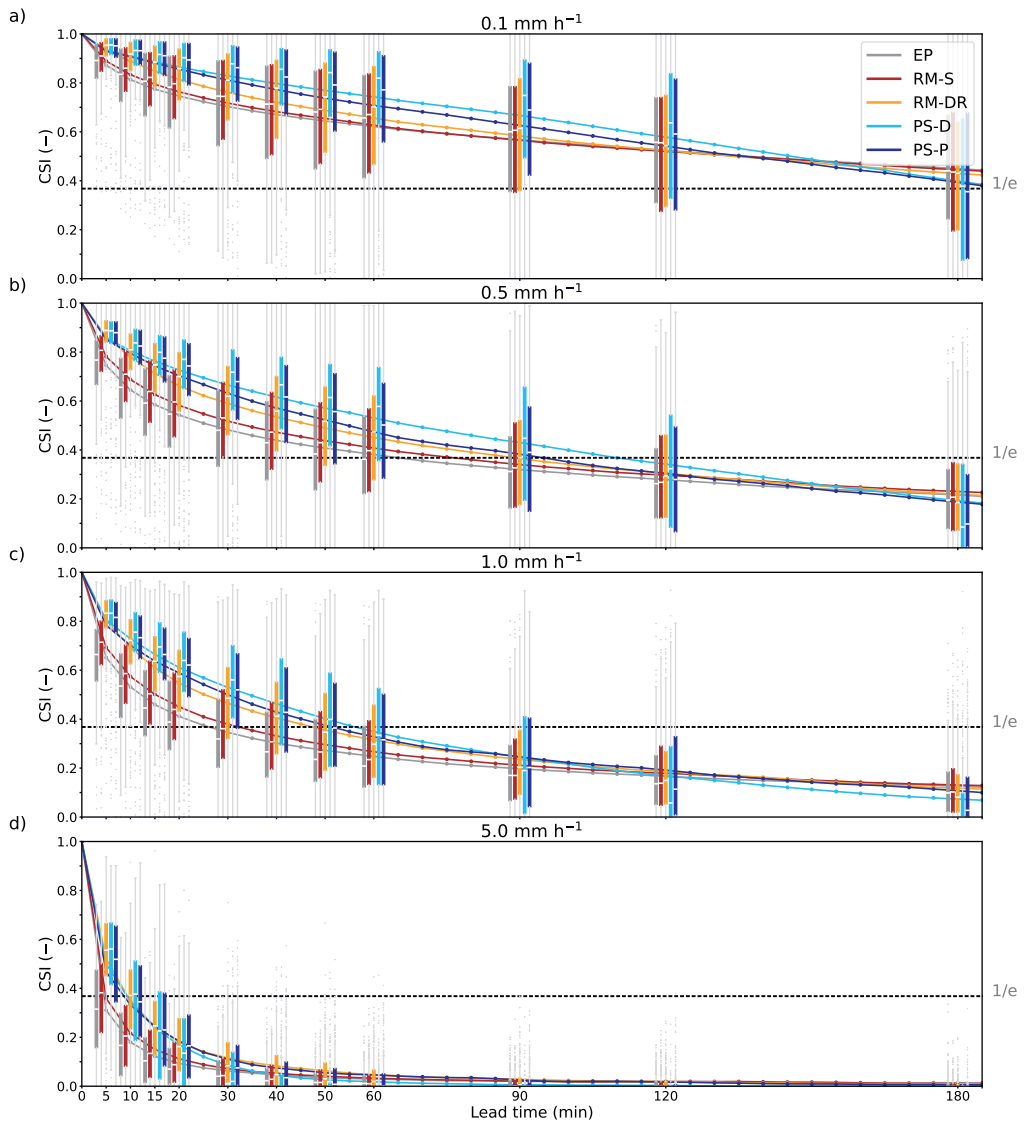


Figure B.4 | CSI score as a function of lead time (5-min steps), averaged over all cells within the catchment and events. The CSI is only calculated for the 6-h duration and for four thresholds: 0.1 mm h⁻¹ (a), 0.5 mm h⁻¹ (b), 1.0 mm h⁻¹ (c) and 5.0 mm h⁻¹ (d). The dotted line indicates a CSI of $\frac{1}{e}$, the minimum CSI for a skillful nowcast. The boxes indicate the variability in results per event, with: the median in white, the interquartile (25th–75th percentile) range (IQR) in colored boxes, 1.5 × IQR starting outside the boxes in grey bars and the outliers in grey dots.

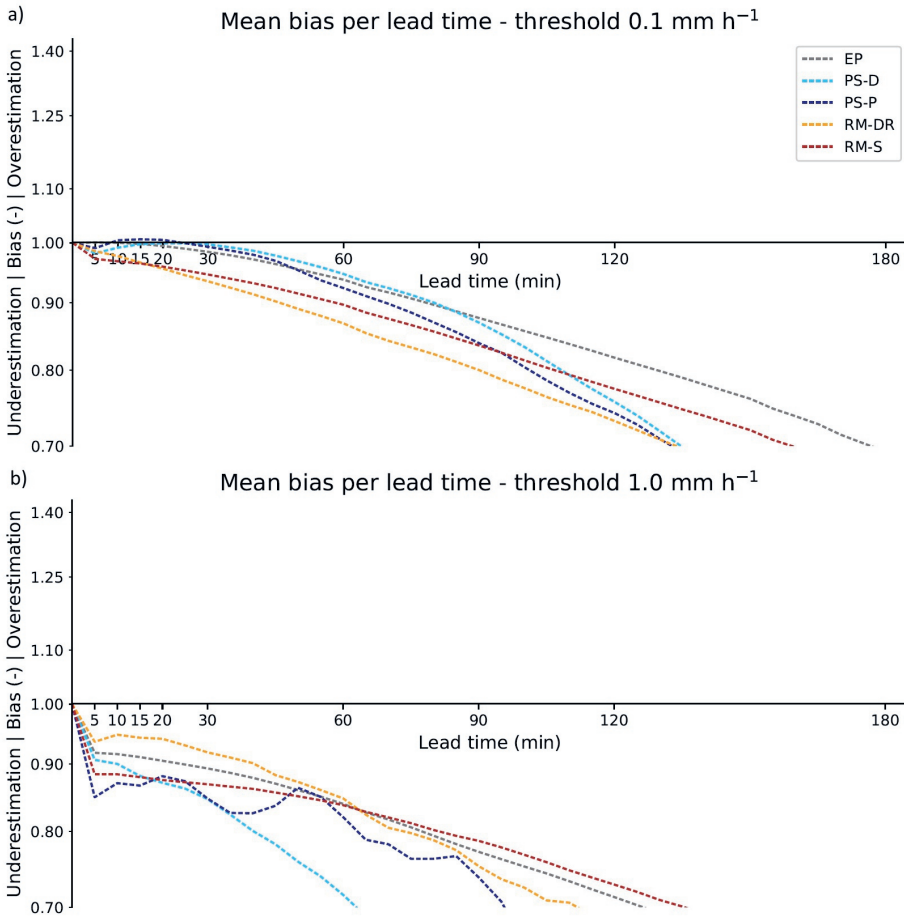


Figure B.5 | Mean bias per algorithm for pixels with a rainfall intensity of (a) $\geq 0.1 \text{ mm h}^{-1}$ and (b) $\geq 1.0 \text{ mm h}^{-1}$ for the 6-h event duration. The bias is plotted on a logarithmic scale, with biases of > 1.0 indicating higher rainfall volumes than observed and the opposite for biases < 1.0 .

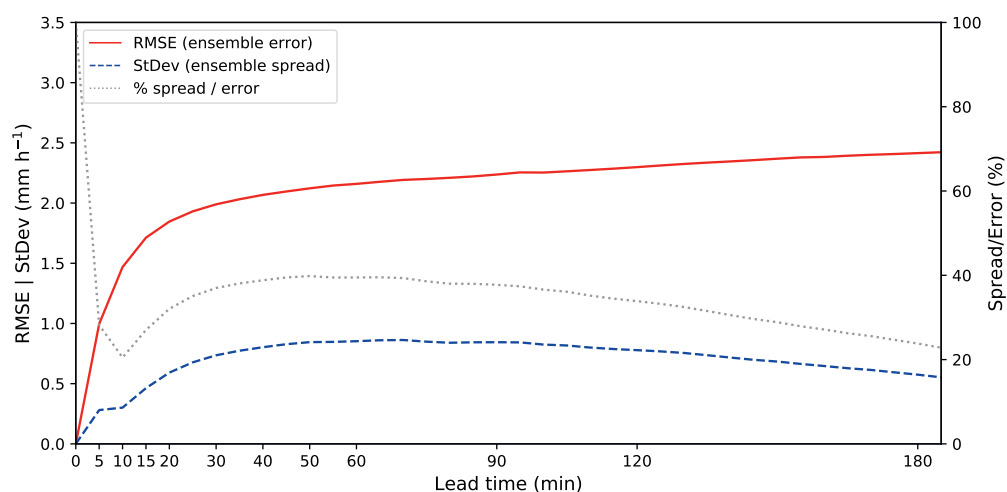


Figure B.6 | Mean ensemble spread (the standard deviation of the ensemble) versus the error between observation and ensemble mean for the 6-h event duration. The dotted grey line indicates which percentage of the forecast errors falls within the ensemble spread. The plot is made for the events in the 6-h event duration as ran with PS-P. A threshold of 0.1 mm h^{-1} is used to only calculate the metrics for cases with rainfall.

C | Additional results of Chapter 5

In this appendix to Chapter 5, we provide thirteen extra figures that complement the main text. Figure C.1 shows the cumulative distribution function of the rainfall intensities from the reference data set for every time interval in the twelve summer days. Figures C.2 – C.12 provide twelve additional figures based on the spatial analysis with the Critical Success Index (CSI; see Equation 2.18), which are similar to Figure 5.4 in Chapter 5, but for the other eleven summer days. Figure (C.13) shows an analysis of all events with the Continuous Ranked Probability Score (CRPS). This approach is introduced in Section C.1.

C.1 | Continuous Ranked Probability Score

To assess the full forecast distribution of the probabilistic nowcasts for every 15-min interval in the nowcasts, the CRPS was used. This metric is a probabilistic application of the mean absolute error and uses the area between the cumulative distribution functions of the forecasts and observations. Hence, the CRPS is 0 for a perfect forecast. The numerical application of this approach for a finite number of ensemble members is described in Hersbach (2000). The CRPS was normalized with the standard deviation of the reference data to be able to compare the different events.

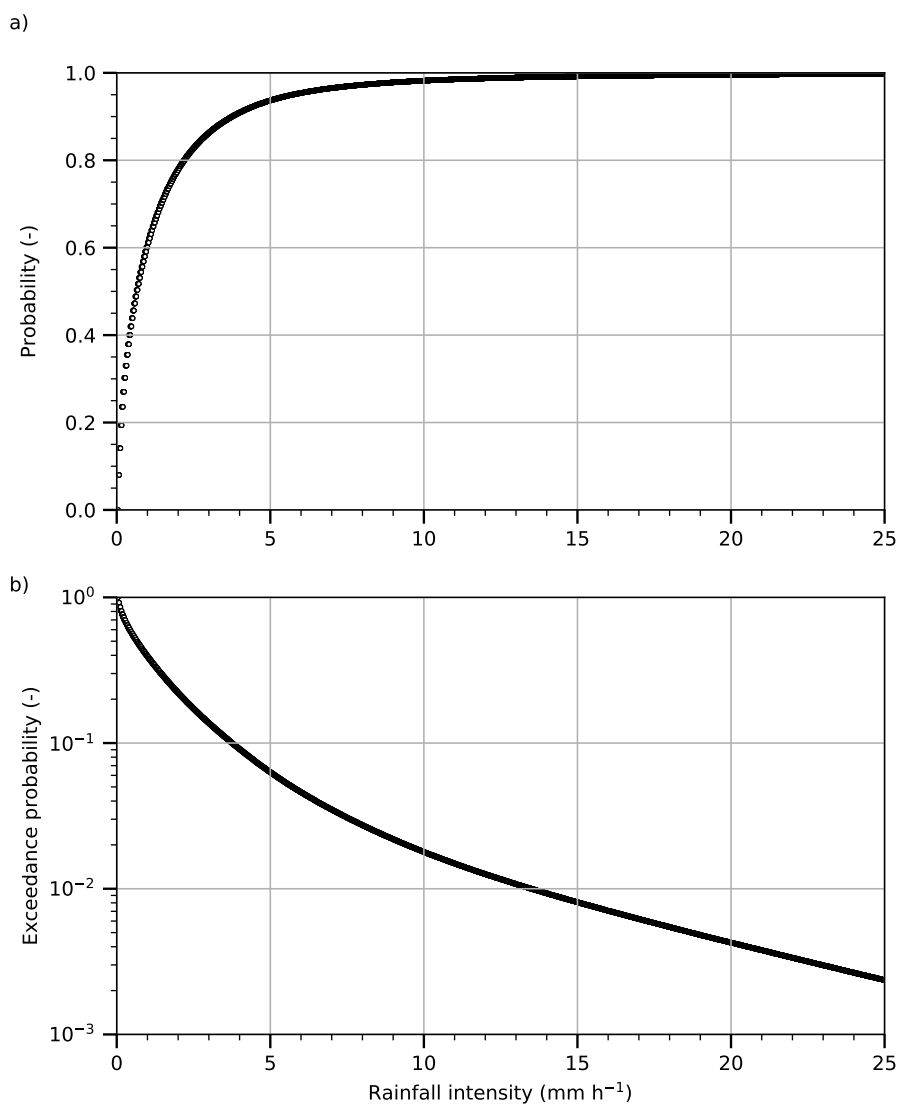


Figure C.1 | Cumulative distribution function (a) and exceedance probabilities (b) of rainfall intensities (mm h⁻¹) in the reference data (1 km² spatial and 5-min temporal resolution) set for the land surface area of the Netherlands and all twelve summer days. Only grid cells with rainfall intensities of 0.1 mm h⁻¹ or higher are taken into account.

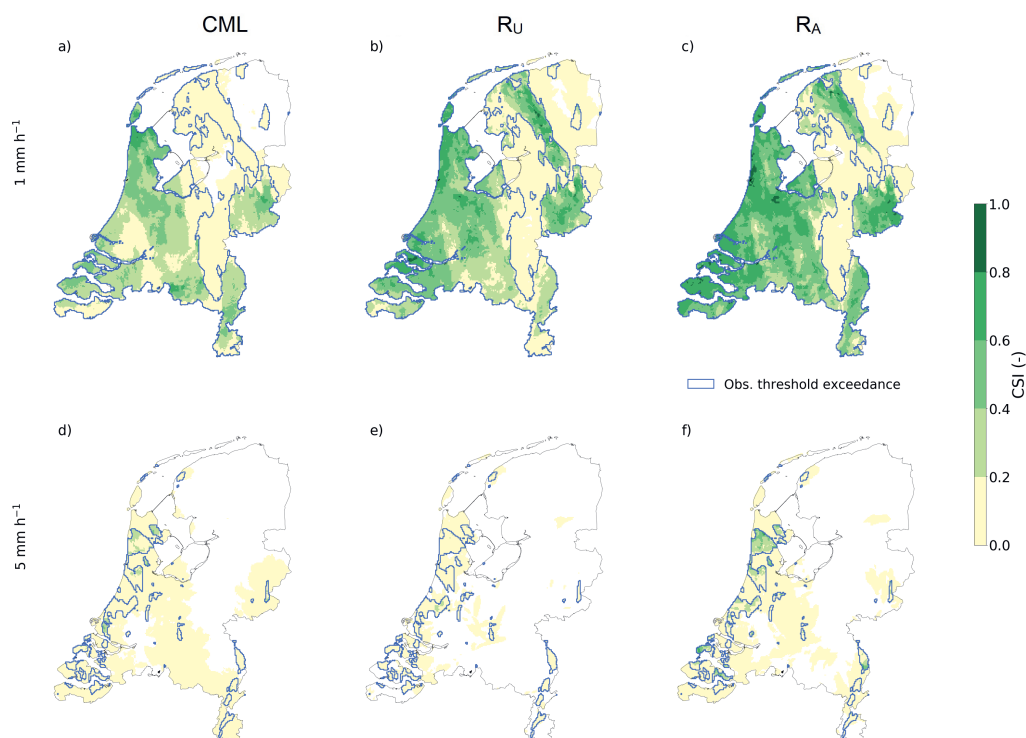


Figure C.2 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 10 – 11 June 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

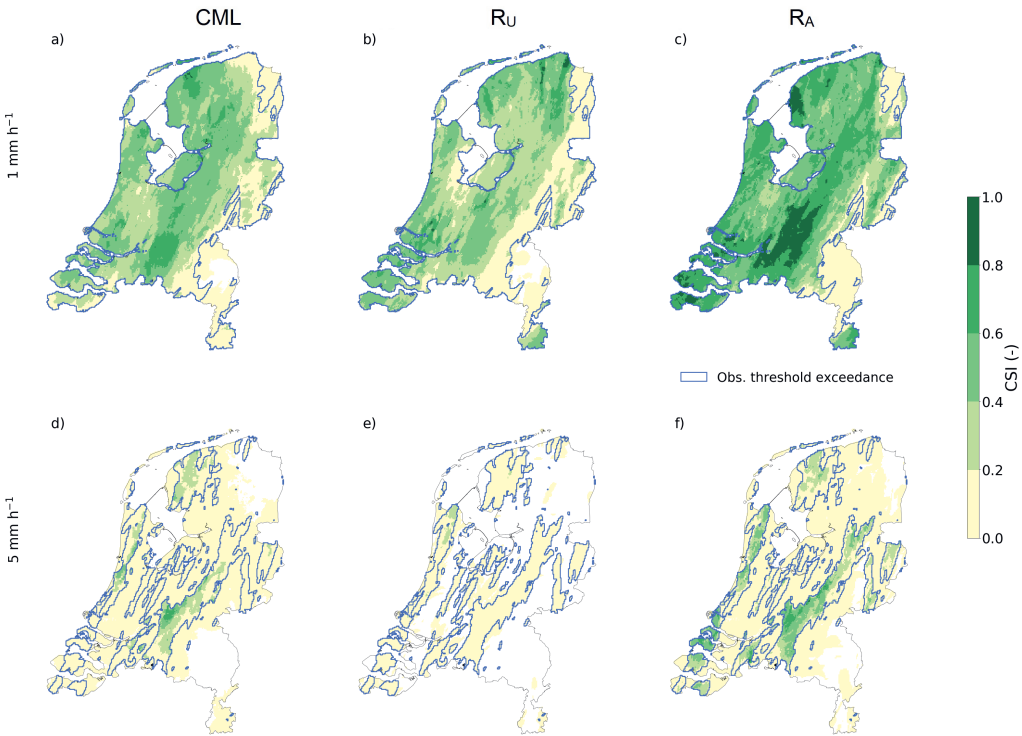


Figure C.3 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 06 – 07 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

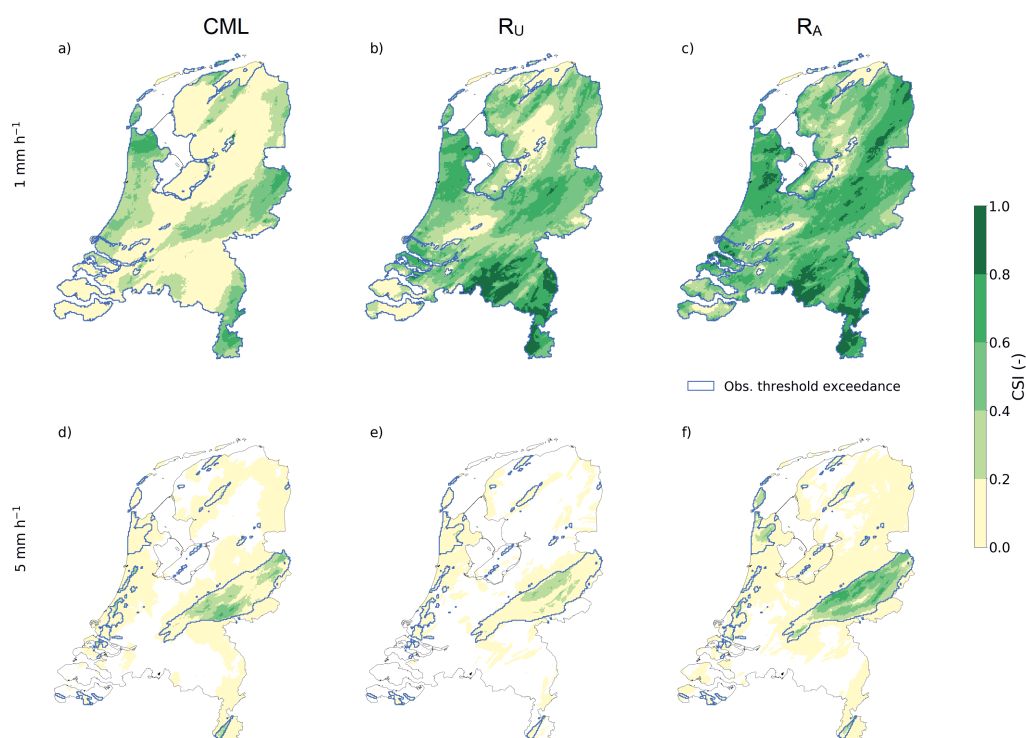


Figure C.4 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 07 – 08 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

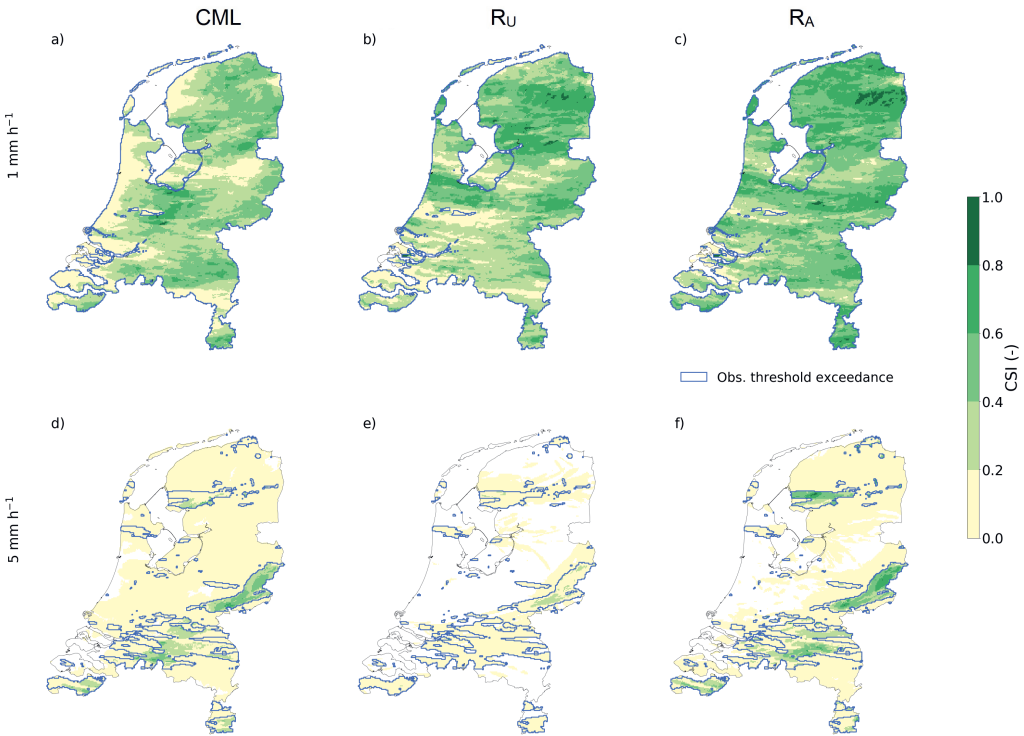


Figure C.5 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 08 – 09 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

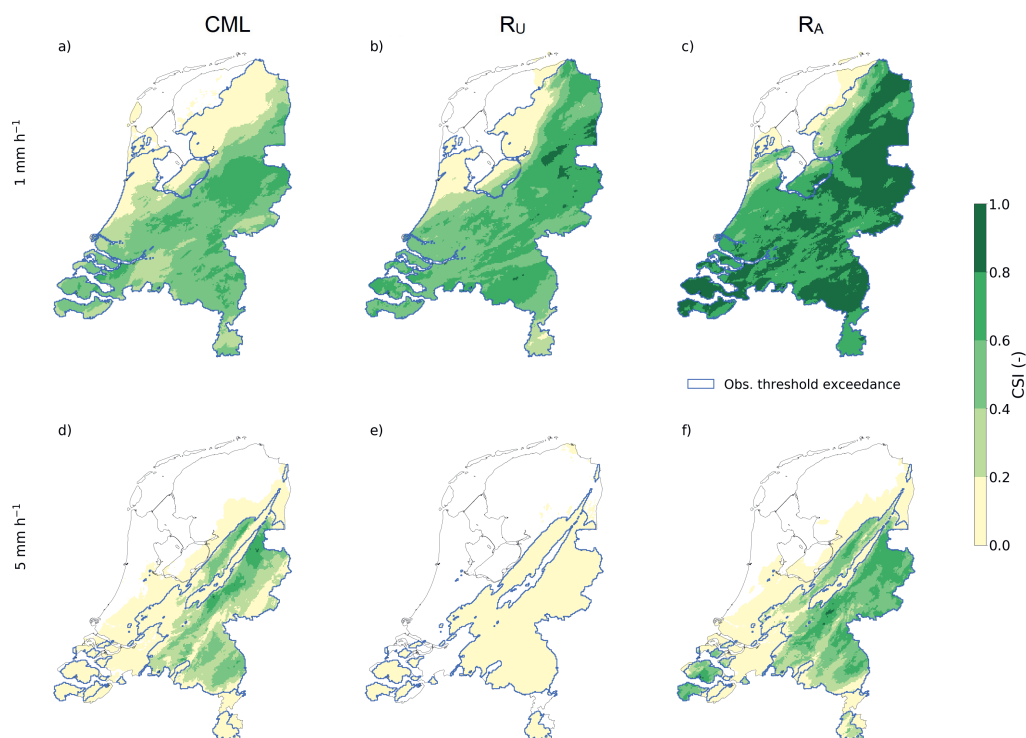


Figure C.6 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 13 – 14 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

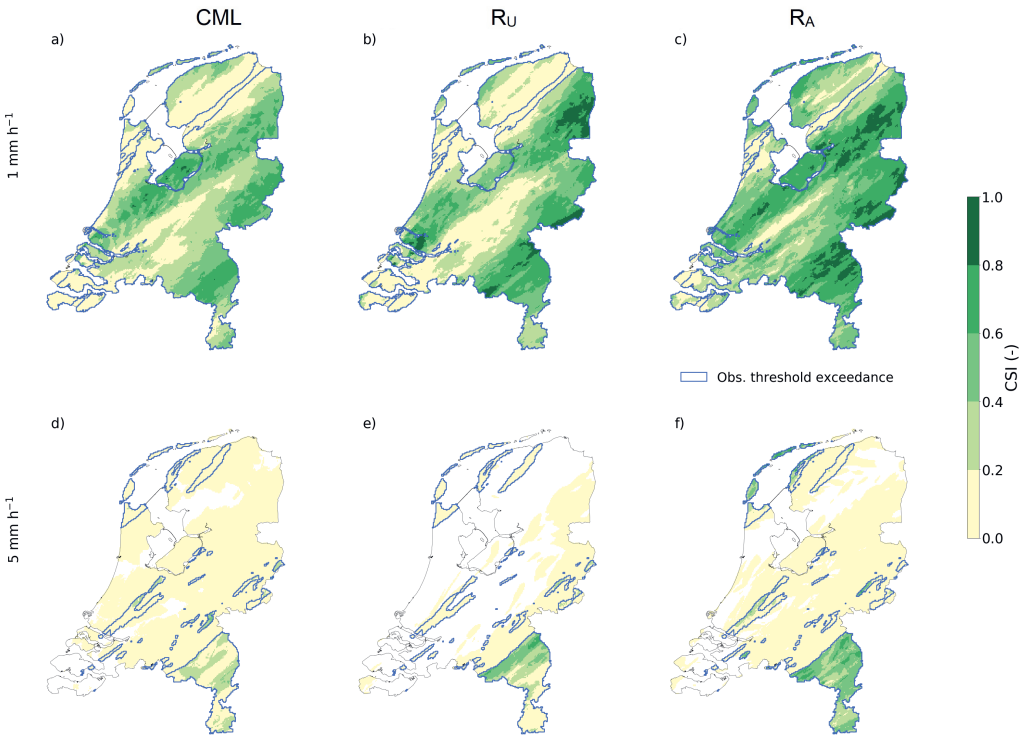


Figure C.7 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 18 – 19 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

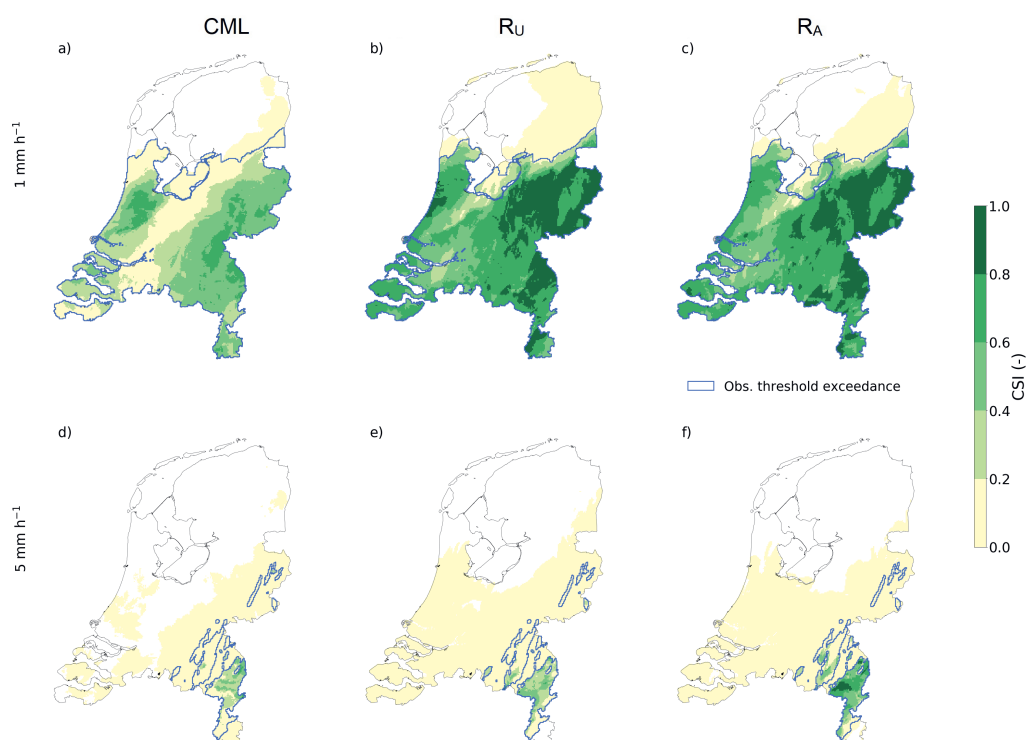


Figure C.8 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 22 – 23 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

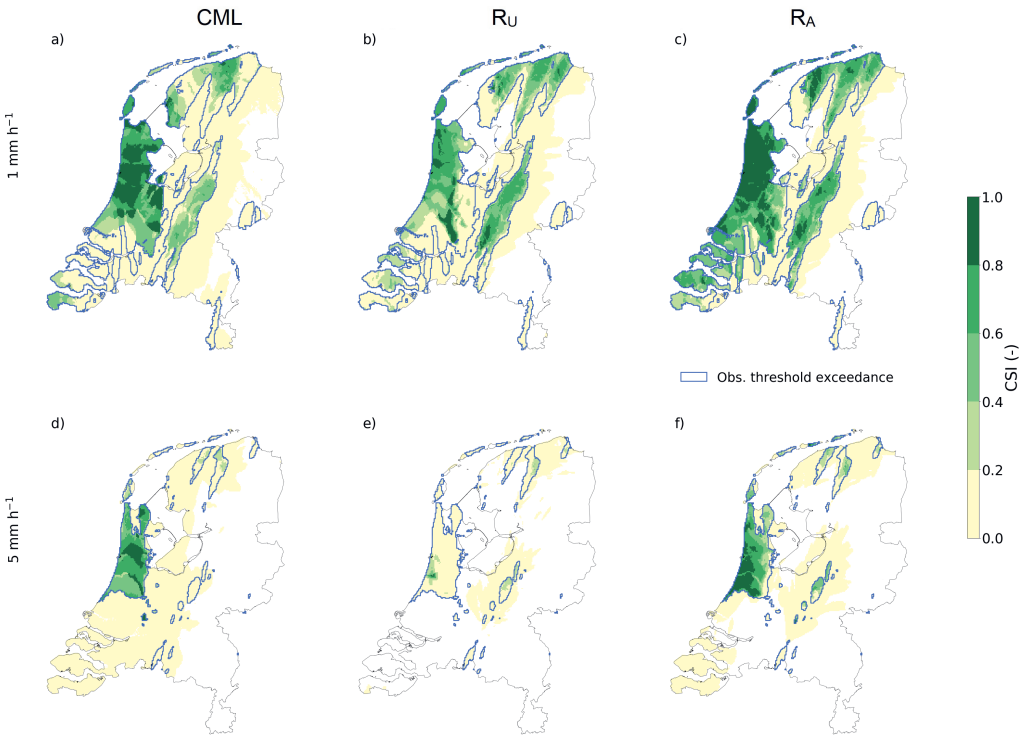


Figure C.9 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 25 – 26 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h⁻¹ for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h⁻¹. The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

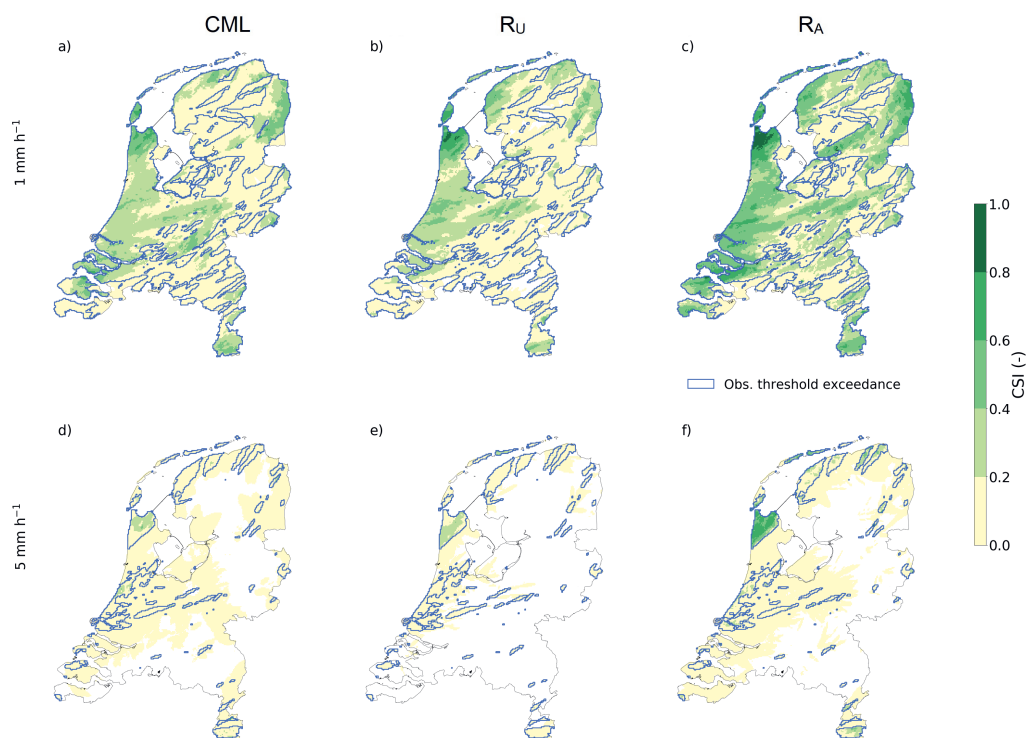


Figure C.10 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 27 – 28 August 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_u) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

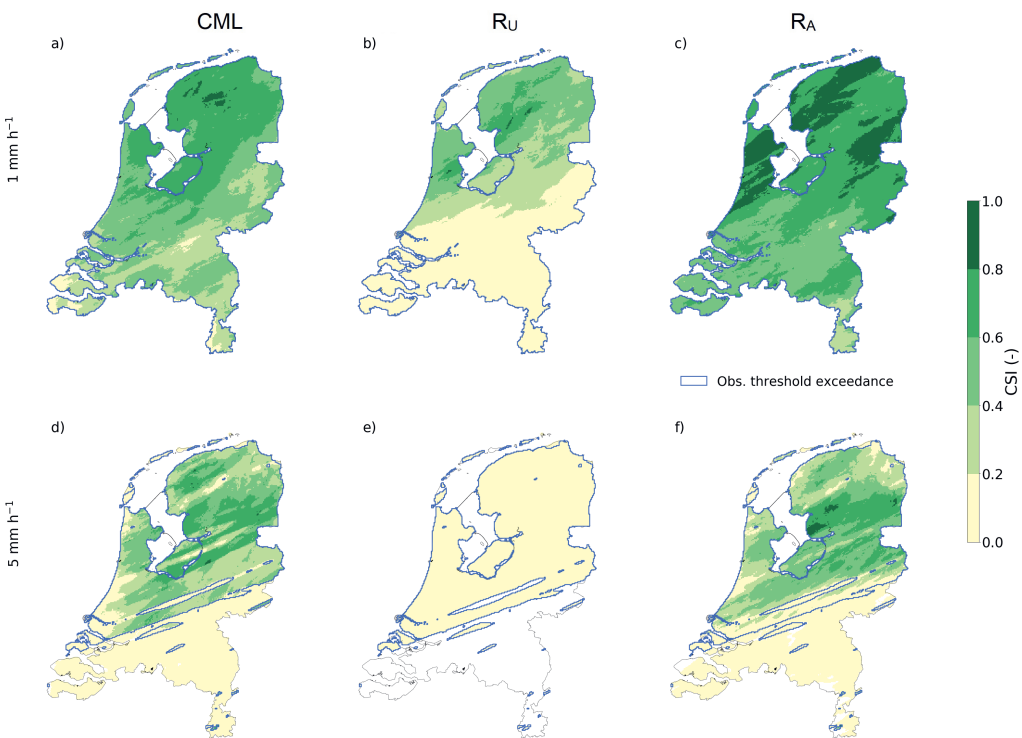


Figure C.11 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 6 – 7 September 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

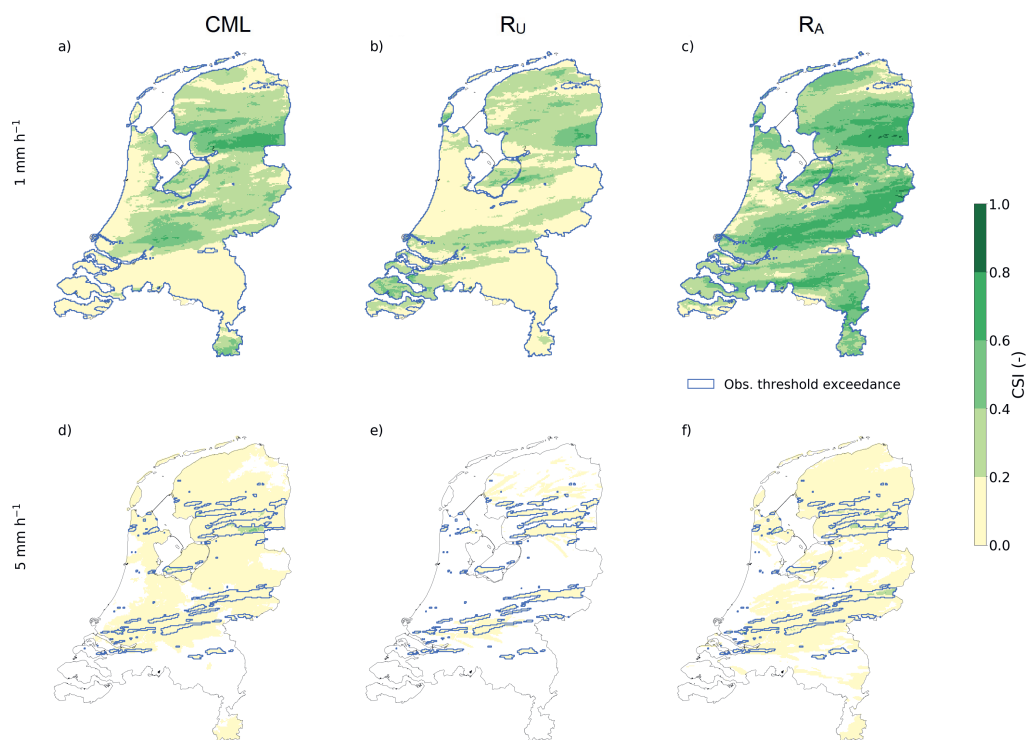


Figure C.12 | Critical success index (CSI) of forecast hourly rainfall sums above two thresholds for the land surface area of the Netherlands during the event of 7 – 8 September 2011. All available nowcasts on a 15-min interval are taken into account. (a) – (c), attained CSI for a threshold of 1.0 mm h^{-1} for the CML nowcast (a), the unadjusted radar (R_U) nowcast (b) and for a nowcast with the gauge-adjusted radar data set (c). (d) – (f) Same as (a) – (c), but for a threshold of 5.0 mm h^{-1} . The blue contours indicate the threshold exceedances in the reference data set for all studied time steps.

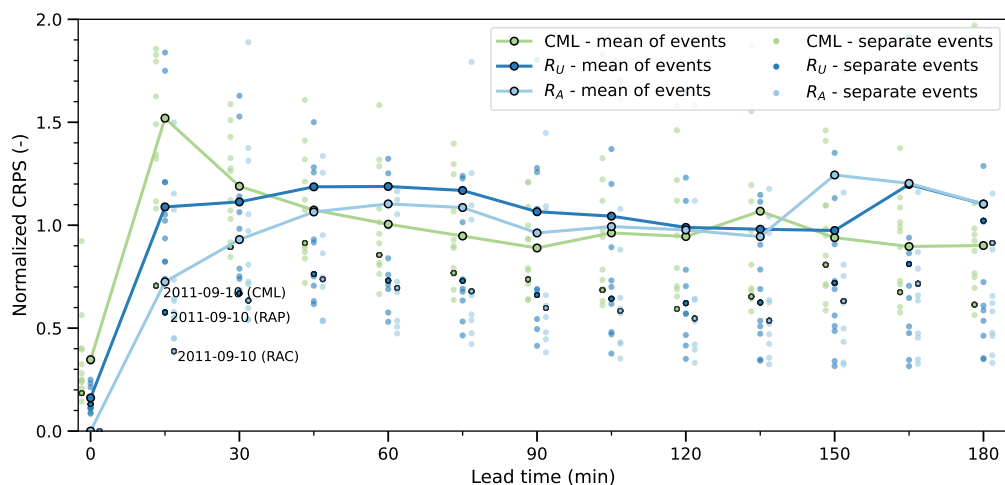


Figure C.13 | Mean normalized CRPS per lead time based on all forecast 15 min rainfall accumulations of the twelve analyzed events for the nowcasts constructed with the CML QPE (green), the unadjusted radar (R_U) QPE (dark blue) and the reference data (light blue). Individual circles indicate the mean normalized CRPS per event.

D | Additional results of Chapter 6

In this appendix to Chapter 6, we provide five additional figures that complement the main text. Figures D.1–D.3 show the mean ensemble spread versus the mean forecast error of the ensemble nowcasts, linear blending method and STEPS blending method. The figures indicate, per lead time, what fraction of the observations fell within the ensemble spread. Figure D.4 is similar to Figure 6.4, but is made for an issue time of one hour later. Figure D.5 illustrates the bias in the forecasts per catchment and event.

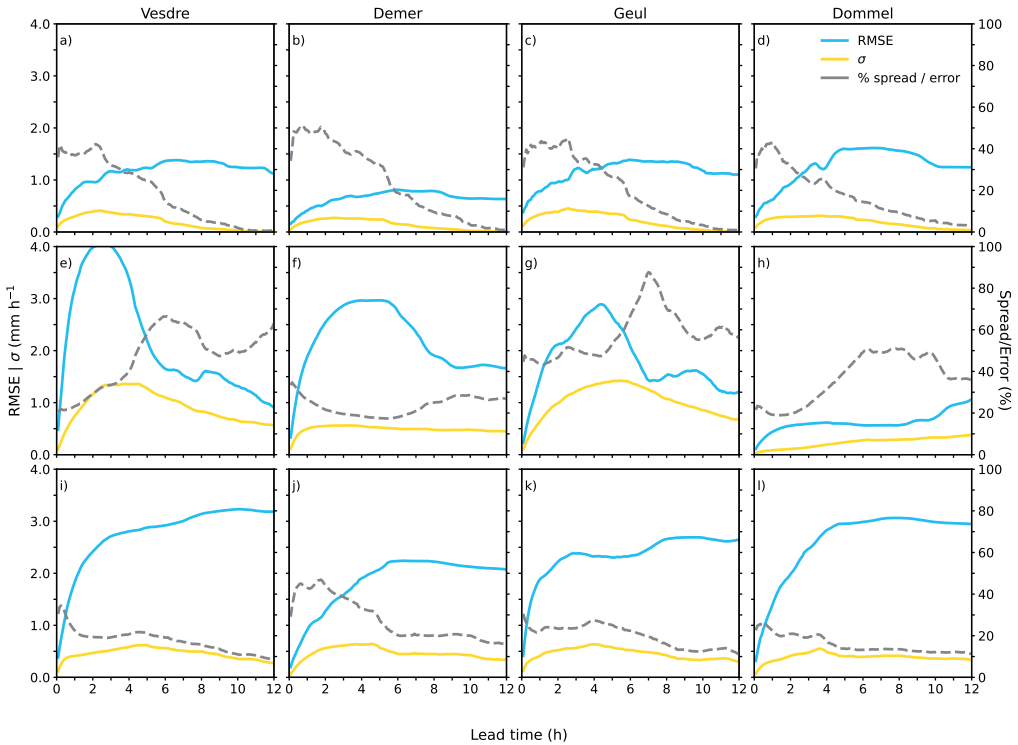


Figure D.1 | Mean ensemble spread (the standard deviation of the ensemble; yellow) versus the error between observation and ensemble mean (blue) for the catchment-averaged rainfall forecasts of the 48-member ensemble nowcasts, for the four catchments and three events (January, a–d; June, e–h; and July, i–l). Shown are event-averaged values for the standard deviation and RMSE. The dotted grey lines indicate which percentage of the forecast errors falls within the ensemble spread.

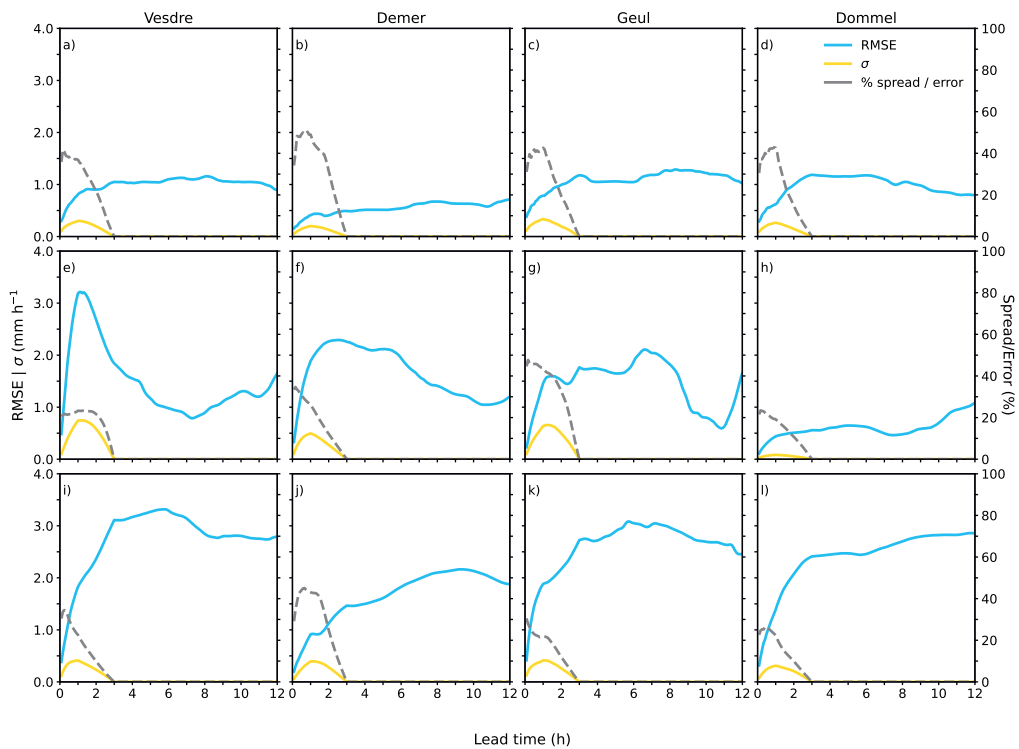


Figure D.2 | Mean ensemble spread (the standard deviation of the ensemble; yellow) versus the error between observation and ensemble mean (blue) for the catchment-averaged rainfall forecasts of the 48-member ensemble linear blending method, for the four catchments and three events (January, a–d; June, e–h; and July, i–l). Shown are event-averaged values for the standard deviation and RMSE. The dotted grey lines indicate which percentage of the forecast errors falls within the ensemble spread.

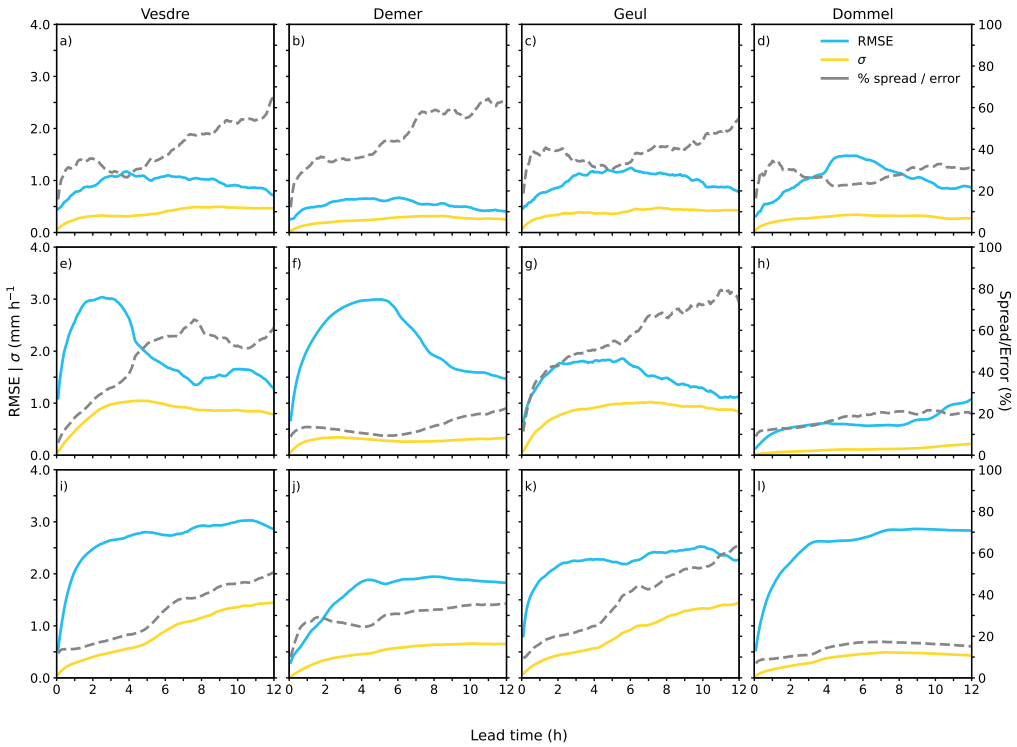


Figure D.3 | Mean ensemble spread (the standard deviation of the ensemble; yellow) versus the error between observation and ensemble mean (blue) for the catchment-averaged rainfall forecasts of the 48-member ensemble STEPS blending method, for the four catchments and three events (January, a–d; June, e–h; and July, i–l). Shown are event-averaged values for the standard deviation and RMSE. The dotted grey lines indicate which percentage of the forecast errors falls within the ensemble spread.

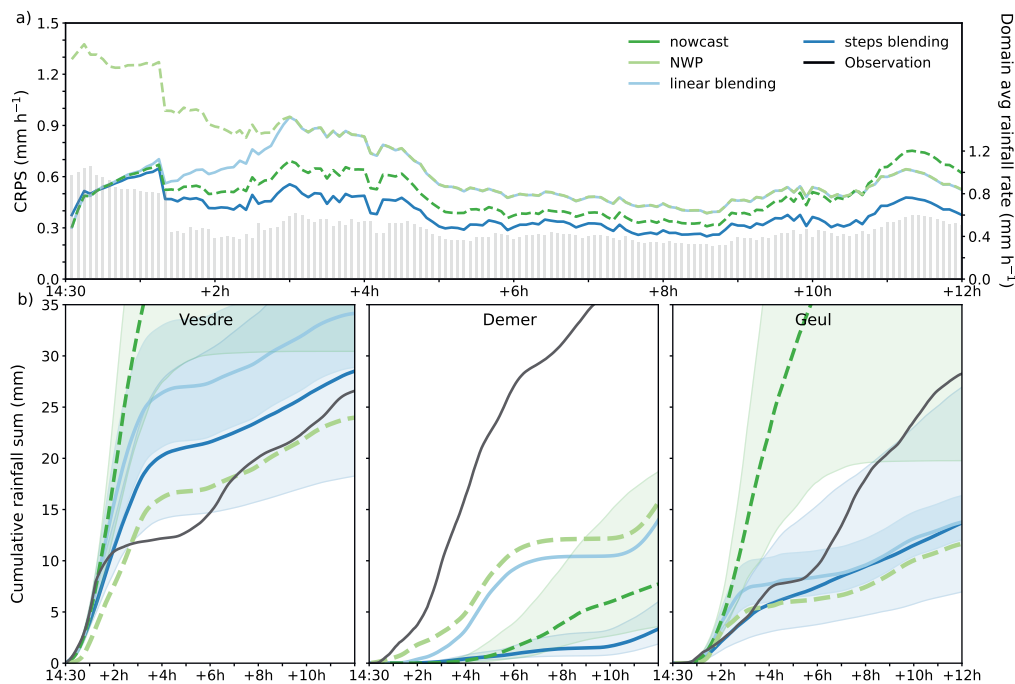


Figure D.4 | Evaluation of the four forecasting methods for the test case of 2021-06-29 14:30 UTC (which is one hour later than the test case in Figure 6.4 of the main text). (a) The CRPS per lead time, averaged over all grid cells in the radar domain. The grey bars indicate the domain-averaged rainfall rates (mm h^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sums per catchment (Vesdre, Demer and Geul) as compared to the observation in black. The thick coloured lines indicate the ensemble median, or the deterministic forecast (for NWP, light green). The shaded areas around the ensemble medians indicate the IQR.

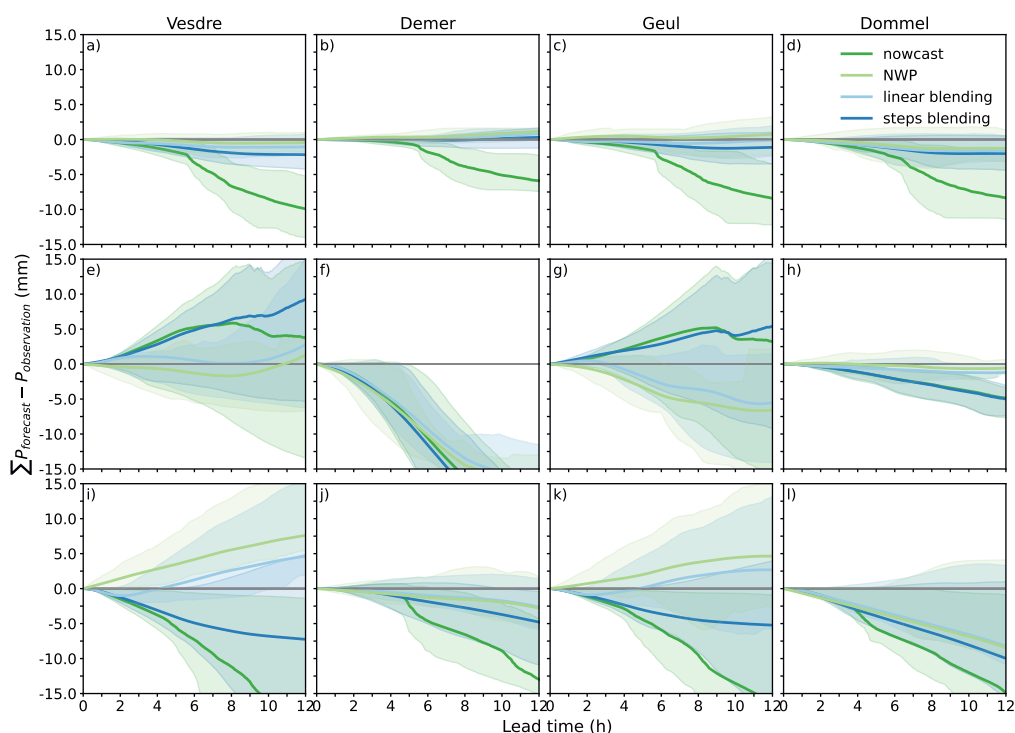


Figure D.5 | Event-averaged cumulative bias (sum of $P_{\text{forecast}} - P_{\text{observation}}$) of the catchment-averaged rainfall sum from the issue time until the indicated lead time. Shown are the event-averaged cumulative bias values for the four catchments and three events (January, a–d; June, e–h; and July, i–l). The ensemble nowcasting method is indicated in dark green, NWP in light green, the linear blending method in light blue and the STEPS blending method in dark blue. The thick coloured lines indicate the ensemble median, or the deterministic forecast (for NWP, light green). The shaded areas around the ensemble medians indicate the IQR of the ensemble.

E | Additional results of Chapter 7

This appendix contains seven additional figures and one table that complement the Sections 7.2.3.1 (Hydrological model setup), 7.3.3 (Dependency on the rainfall characteristics), 7.3.4 (Discharge peak forecast verification) and 7.4.2 (Discharge peak verification) in Chapter 7.

Figure E.1 provides event information regarding the event-averaged rainfall intensity and duration for the 1,533 events in Chapter 4.

Figures E.2–E.4 give the results of the calibration period for catchments Dwarsdiep, Roggelsebeek and Luntersebeek. The WALRUS model (Brauer et al., 2014a) was used for these models and the calibration procedure is further described in Section 7.2.3.1 of Chapter 7. The validation period (2015 – 2016) is shown in Figure 7.2 in Chapter 7, along with model validation for the other catchments. In addition, Table E.1 provides the parameters used to set up WALRUS for 8 out of 12 catchments.

Figure E.5 illustrates the relationship between the spatial variability of the rainfall fields (l_r) and the mean rainfall intensity for the rainy periods during an event (as described in Chapter 7). This gives an explanation for the quite similar response of the rainfall and discharge forecast error to increasing l_r and the mean event rainfall intensity, as illustrated in Figures 7.5 and 7.6 in the main text.

Finally, Figures E.6 and E.7 are similar to Figure 7.7 in Chapter 7. Both figures give the results of the analysis in Section 7.4.2 of Chapter 7 for a either a different allowed timing error or a different allowed peak magnitude. The results are further elaborated on in discussion Section 7.4.2 of Chapter 7.

Table E1 | WALRUS parameters used for the hydrological model setup and simulations of the (partly) freely-draining catchments (8 out of 12 catchments). c_W is the wetness index parameter (mm), c_V the vadose zone relaxation time (h), c_G the groundwater reservoir constant (mm h^{-1}), c_Q the quickflow reservoir constant (h), c_S the surface water parameter, indicating the bankfull discharge (mm h^{-1}), c_D the channel depth (mm) and a_S the surface water area fraction (-).

Catchment name	Size (km^2)	c_W (mm)	c_V (h)	c_G (10^6 mm h)	c_Q (h)	c_S (mm h^{-1})	c_D (mm)	a_S (-)	soil type	reference
Hupsel Brook	6.5	356	0.2	5	3.0	4.0	1500	0.01	sand on clay	[1]
Grote Waterleiding	40	340	10	20	35.0	3.0	2200	0.01	loamy sand	[2]
Luntersebeek	63	250	10	80	25.0	4.0	2500	0.01	sand	-
Dwarsdiep	83	255	1	30	24.7	4.5	2292	0.015	loamy sand	-
Roggelsebeek	88	268	10	87	28.0	4.0	3000	0.01	sand	-
Reusel	176	255	100	75	175.0	7.0	1500	0.01	sand	[3]
Aa	836	350	10	30	85.0	6.0	2250	0.01	loamy sand	[4]
Regge	957	396	45	16	7.5	0.2	2450	0.01	loamy sand	[5]

¹Brauer et al. (2014b)

²Heuvelink et al. (2020)

³Loos (2015a)

³Gerritsen (2019)

³Loos (2015b)

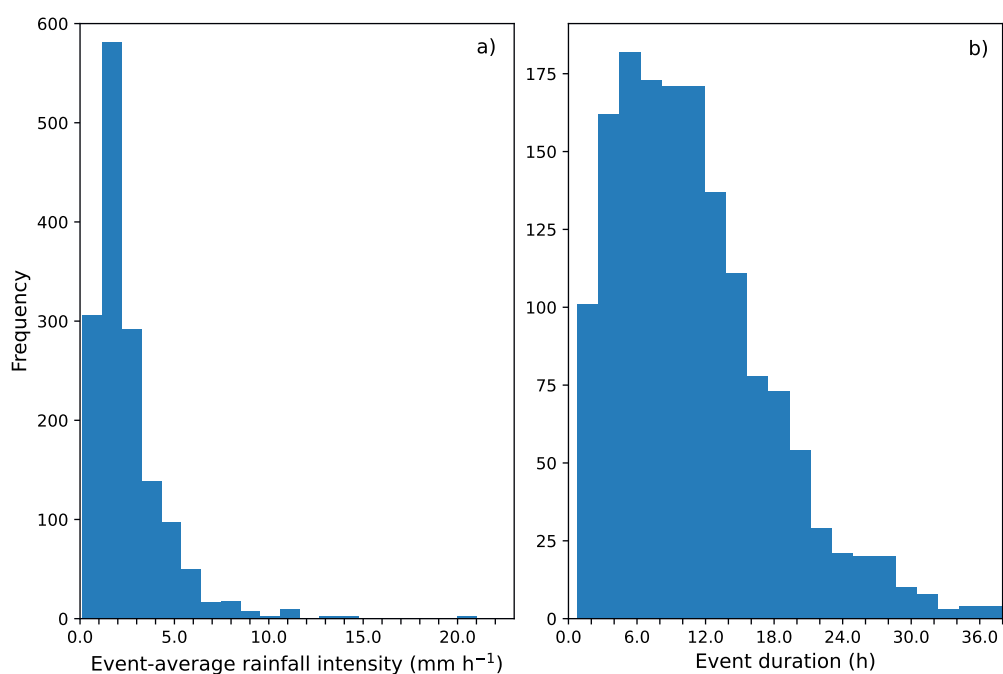


Figure E.1 | Event information about the 1,533 events in Chapter 4. Panel (a) shows the distribution of the event-average rainfall intensity for the events and panel (b) shows the distribution of the rainfall duration for the events, based on the number of 5-min time steps with rainfall in the events. Note that the maximum duration can exceed the longest “event duration” of 24 h, because the nowcasts were run for a period from 6 h prior to to 6 h after the defined event.

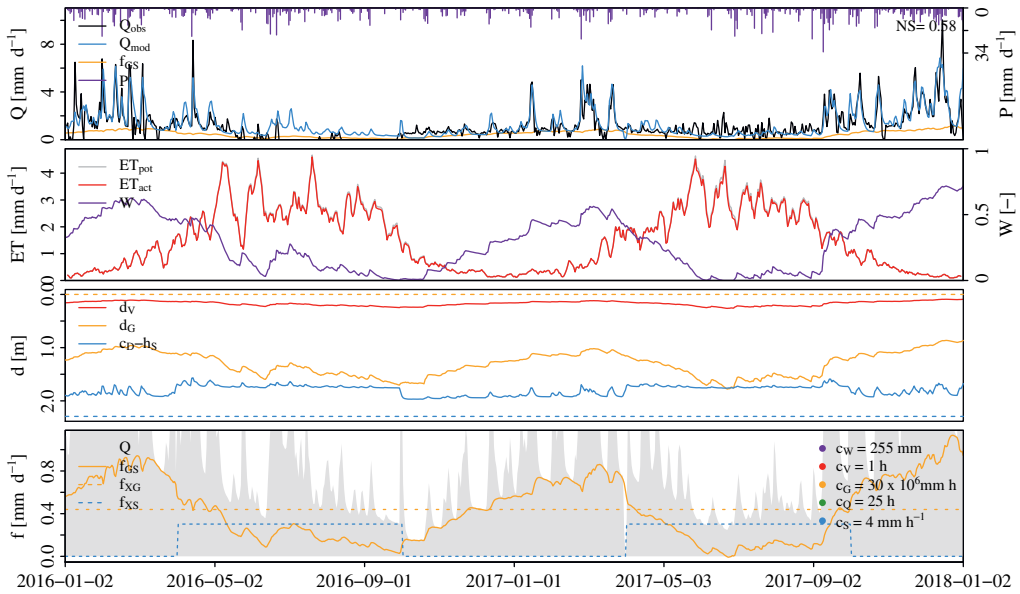


Figure E.2 | Modeled states and fluxes with the WALRUS model for the Dwarsdiep catchment during the calibration period (2016 – 2017). The top sub panel shows the simulated (blue) and observed (black) discharge. The Nash-Sutcliffe efficiency (NSE) is indicated in the top right (Nash & Sutcliffe, 1970). The purple bars indicate the rainfall input, the yellow line the groundwater drainage. Sub panel two shows the actual (red) and potential (grey) evapotranspiration, and the wetness index in purple. Sub panel three shows the storage deficit (red), the groundwater level below the surface (yellow) and the surface water level below the surface (blue). Sub panel four shows the managed groundwater influx (dotted yellow line) and surface water influx (dotted blue line) over the year. All fluxes are in mm d^{-1} .

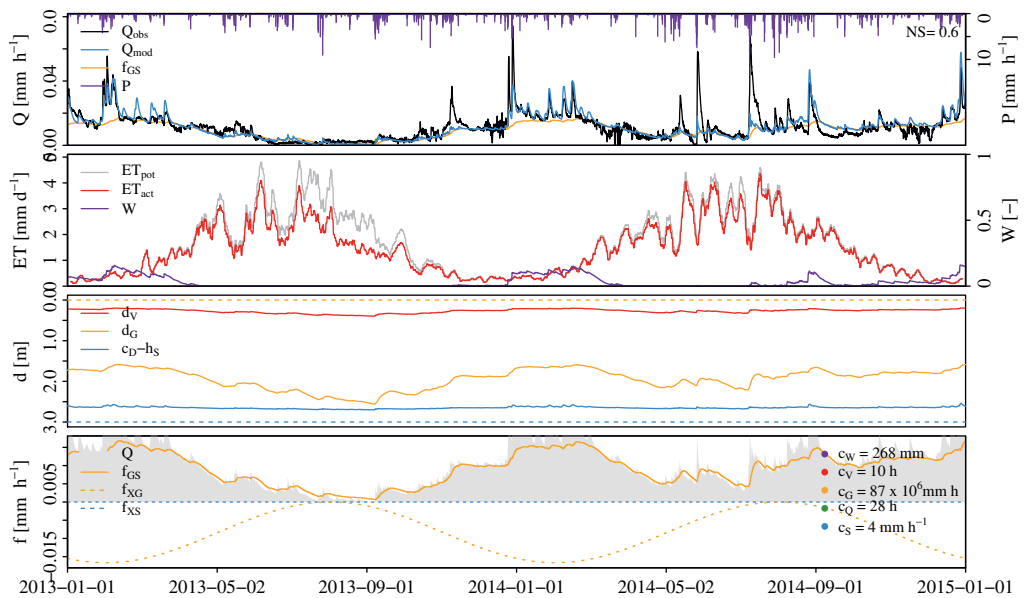


Figure E.3 | Modeled states and fluxes with the WALRUS model for the Roggelsebeek catchment during the calibration period (2013 – 2015). The top sub panel shows the simulated (blue) and observed (black) discharge. The Nash-Sutcliffe efficiency (NSE) is indicated in the top right (Nash & Sutcliffe, 1970). The purple bars indicate the rainfall input, the yellow line the groundwater drainage. Sub panel two shows the actual (red) and potential (grey) evapotranspiration, and the wetness index in purple. Sub panel three shows the storage deficit (red), the groundwater level below the surface (yellow) and the surface water level below the surface (blue). Sub panel four shows the managed groundwater outflow (downward seepage; dotted yellow line) over the year. All fluxes are in mm d^{-1} .

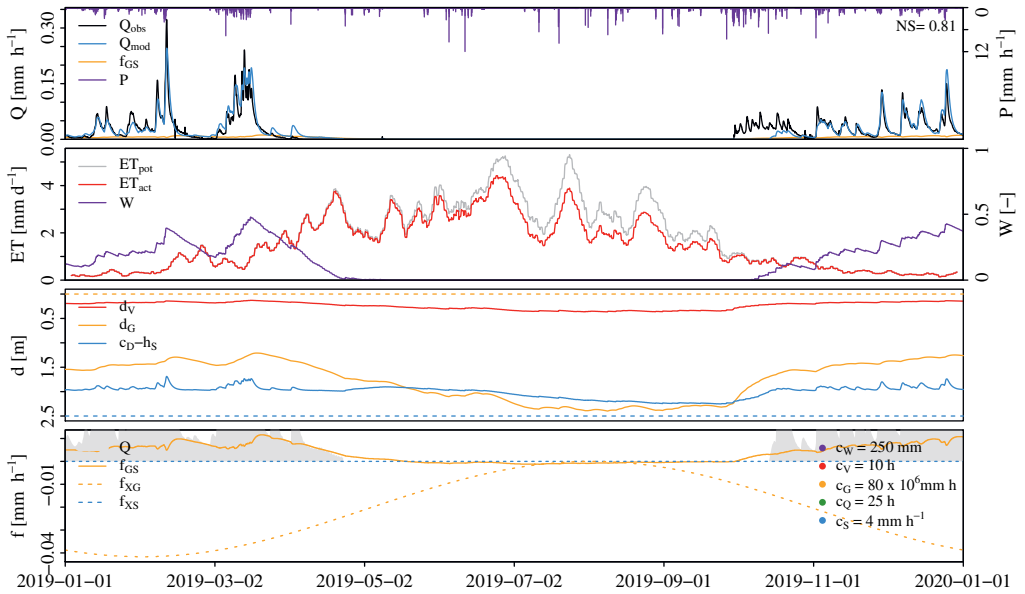


Figure E.4 | Modeled states and fluxes with the WALRUS model for the Luntersebeek catchment during the calibration period (2019). The top sub panel shows the simulated (blue) and observed (black) discharge. The Nash-Sutcliffe efficiency (NSE) is indicated in the top right (Nash & Sutcliffe, 1970). The purple bars indicate the rainfall input, the yellow line the groundwater drainage. Sub panel two shows the actual (red) and potential (grey) evapotranspiration, and the wetness index in purple. Sub panel three shows the storage deficit (red), the groundwater level below the surface (yellow) and the surface water level below the surface (blue). Sub panel four shows the managed groundwater outflow (downward seepage; dotted yellow line) over the year. All fluxes are in mm d^{-1} .

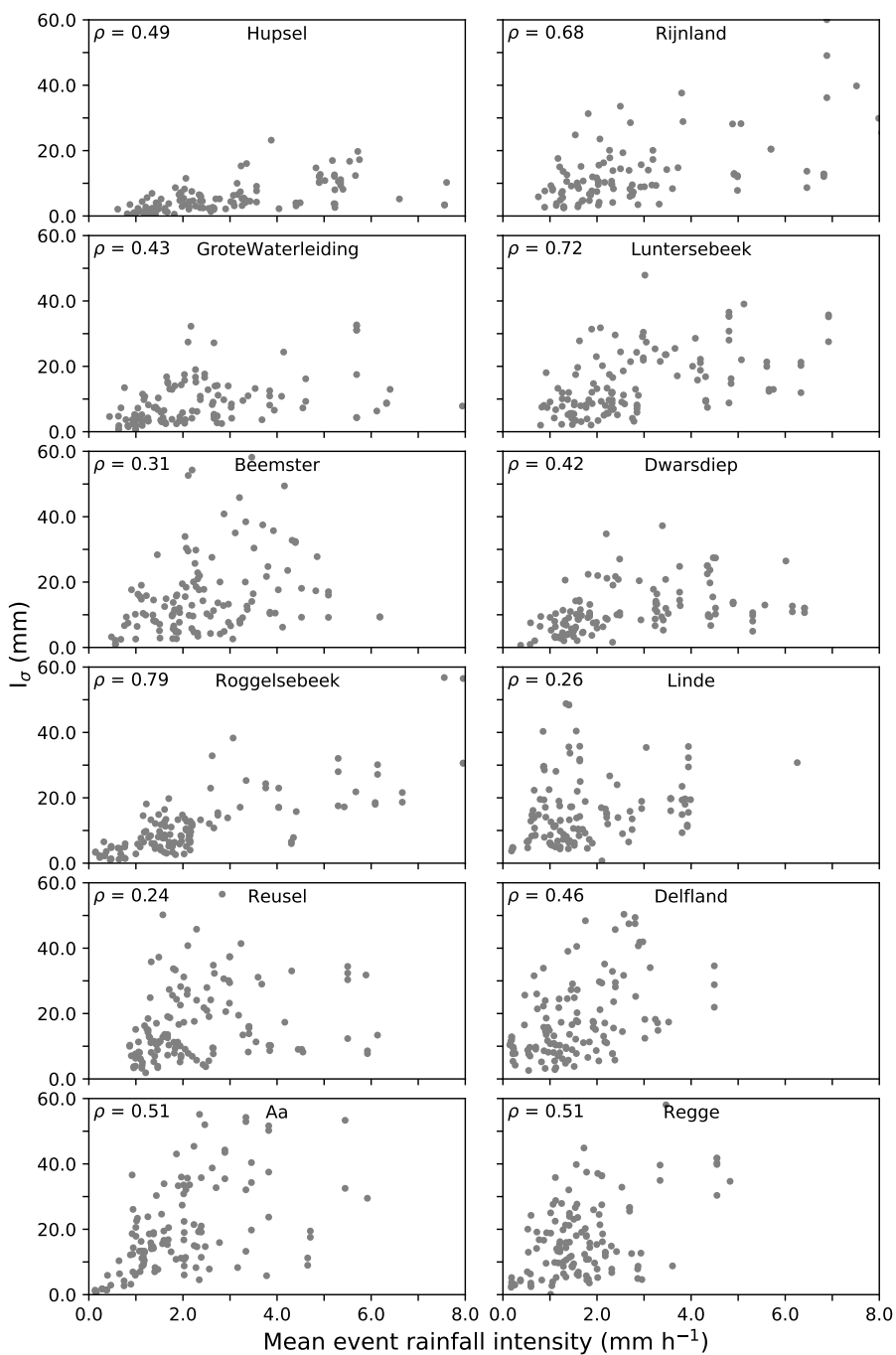


Figure E.5 | Relationship between the mean event rainfall intensity and the mean event rainfall variability (I_{σ}). Both variables are based on all rainy 5-min instances in an event. The mean event values of the rainfall intensity and I_{σ} are the same as in figures 6 and 7.

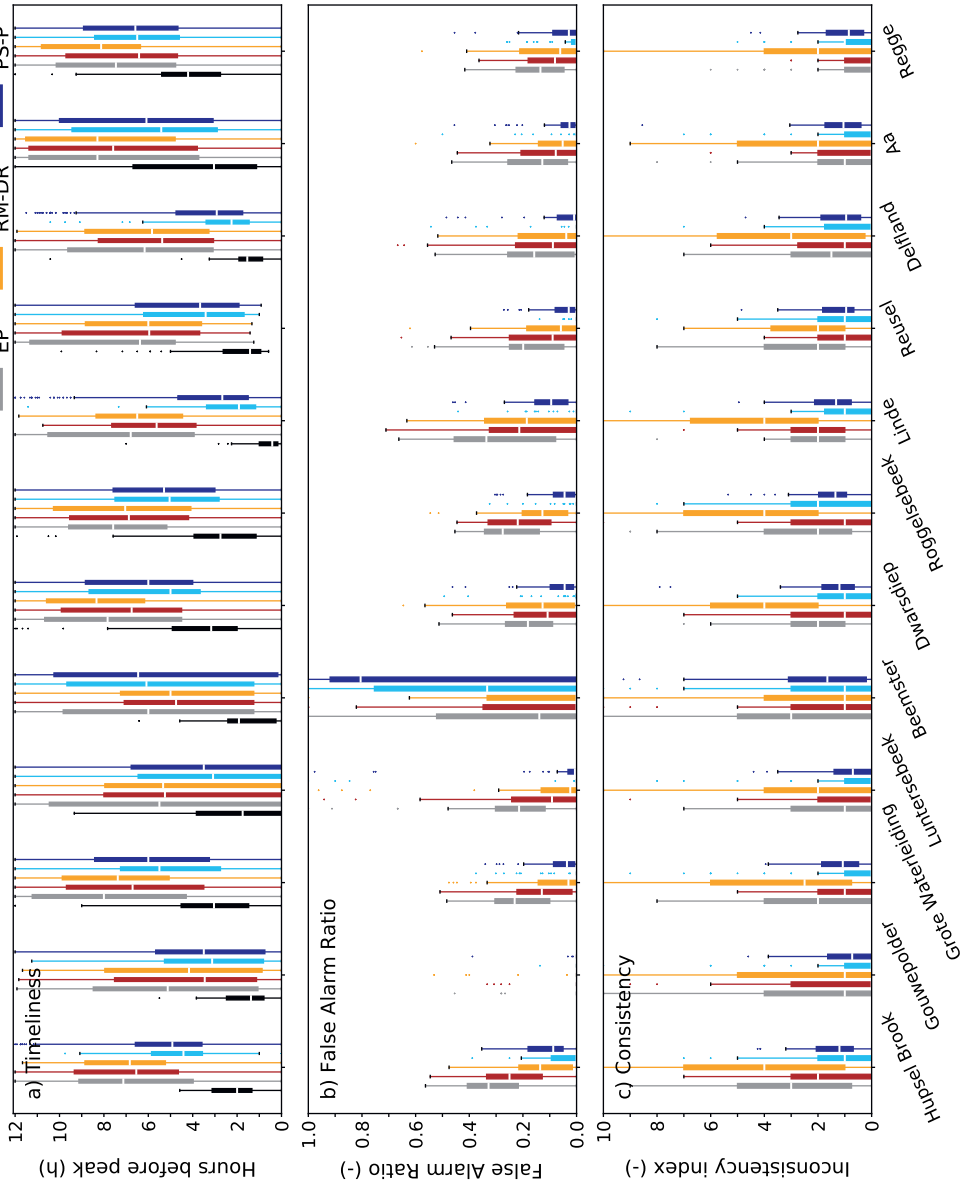


Figure E.6 | Discharge peak forecast verification per catchment for all methods in Chapter 7 based on a $\pm 10\%$ allowed magnitude error and a ± 30 -min allowed timing error. All events were taken into account, but only the longest duration was selected when there is an overlap between events of different 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 IQR indicated by colored boxes, the whiskers indicating $1.5 \times \text{IQR}$, and the outliers indicated by dots.

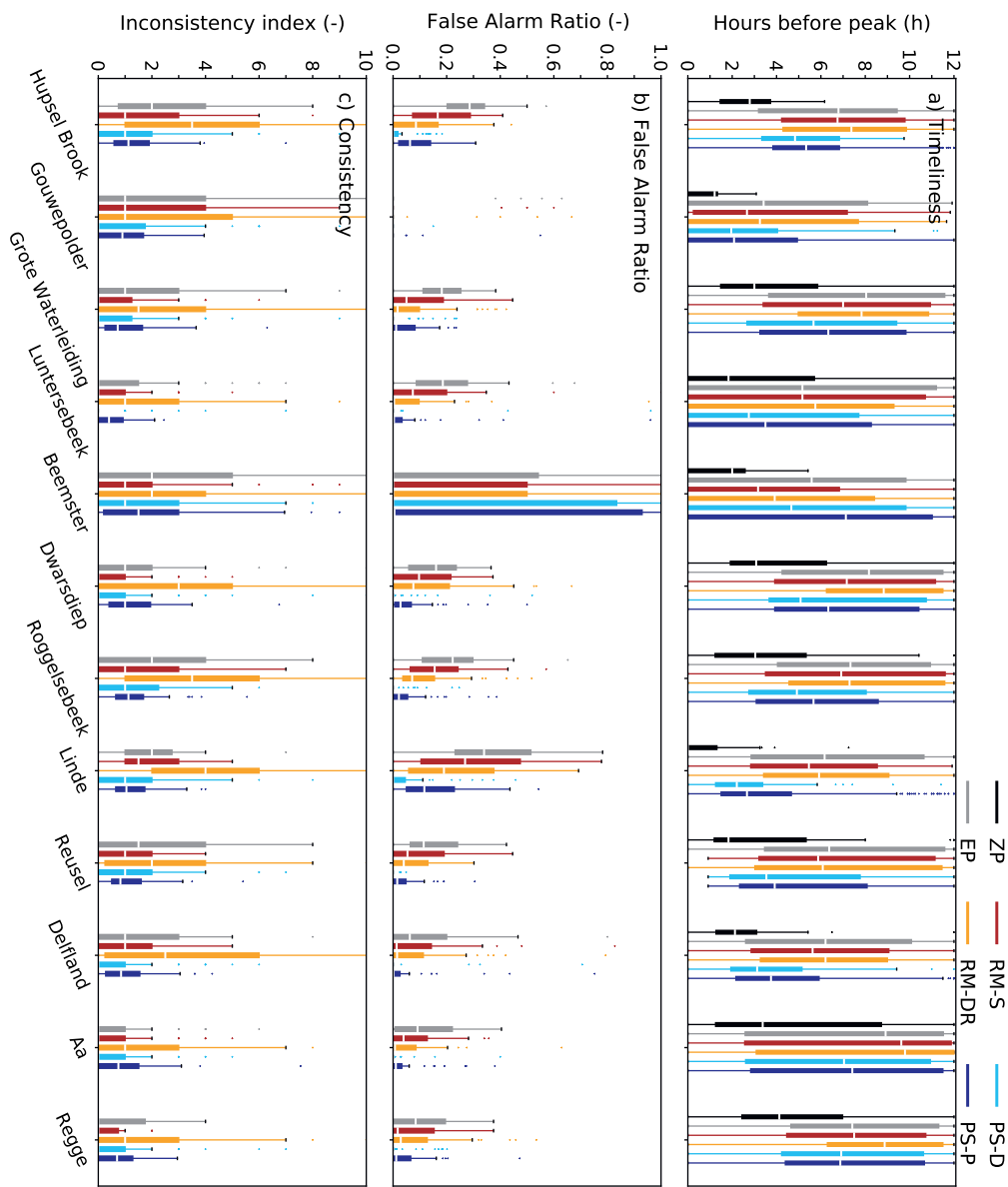


Figure E7 | Discharge peak forecast verification per catchment for all methods in Chapter 7 based on a $\pm 25\%$ allowed magnitude error. All events were taken into account, but only the longest duration was selected, when there is an overlap between events of different durations (1, 3, 6 or 24 hr). 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 IQR indicated by colored boxes, the whiskers indicating $1.5 \times \text{IQR}$, and the outliers indicated by dots.

Bibliography

- ABBE, C. (1901). The physical basis of long-range weather forecasts. *Monthly Weather Review*, 29(12), 551–561. DOI: 10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2.
- ANAGNOSTOU, E. N. & W. F. KRAJEWSKI (1999). Real-time radar rainfall estimation. Part I: Algorithm formulation. *Journal of Atmospheric and Oceanic Technology*, 16(2), 189–197. DOI: 10.1175/1520-0426(1999)016<0189:RTRREP>2.0.CO;2.
- ANAGNOSTOU, M. N., J. KALOGIROS, E. N. ANAGNOSTOU, M. TAROLLI, A. PAPADOPOULOS & M. BORGA (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. DOI: 10.1016/j.jhydrol.2010.06.026.
- ANDRIEU, H., M. N. FRENCH, W. F. KRAJEWSKI & K. P. GEORGAKAKOS (2003). Stochastic–dynamical rainfall simulation based on weather radar volume scan data. *Advances in Water Resources*, 26(5), 581–593. DOI: 10.1016/S0309-1708(02)00168-9.
- AON (2021). *Global catastrophe recap: July 2021*. Technical report, AON, London, United Kingdom. http://thoughtleadership.aon.com/Documents/20211008_analytics-if-july-global-recap.pdf.
- ARCHFIELD, S., M. CLARK, B. ARHEIMER, L. HAY, H. McMILLAN, J. KIANG & ET AL. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, 12(12), 10078–10091. DOI: 10.1002/2015WR017498.
- ARNBJERG-NIELSEN, K. & H. S. FLEISCHER (2009). Feasible adaptation strategies for increased risk of flooding in cities due to climate change. *Water Science and Technology*, 60(2), 273–281. DOI: 10.2166/wst.2009.298.
- ARNELL, N. W. & S. N. GOSLING (2016). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, 134(3), 387–401. DOI: 10.1007/s10584-014-1084-5.
- ATENCIA, A., T. RIGO, A. SAIROUNI, J. MORÉ, J. BECH, E. VILA CLARA, J. CUNILLERA, M. C. LLASAT & L. GARROTE (2010). Improving QPF by blending techniques at the Meteorological Service of Catalonia. *Natural Hazards and Earth System Sciences*, 10(7), 1443–1455. DOI: 10.5194/nhess-10-1443-2010.
- ATENCIA, A. & I. ZAWADZKI (2014). A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Monthly Weather Review*, 142(11), 4036–4052. DOI: 10.1175/MWR-D-13-00117.1.
- ATENCIA, A. & I. ZAWADZKI (2015). A comparison of two techniques for generating nowcasting ensembles. Part II: Analogs selection and comparison of techniques. *Monthly Weather Review*, 143(7), 2890–2908. DOI: 10.1175/MWR-D-14-00342.1.
- ATLAS, D. & C. W. ULBRICH (1977). Path- and area-integrated rainfall measurement by microwave attenuation in the 1–3 cm band. *Journal of Applied Meteorology*, 16(12), 1322–1331. DOI: 10.1175/1520-0450(1977)016<1322:PAAIRM>2.0.CO;2.
- AUSTIN, P. M. (1987). Relation between measured radar reflectivity and surface rainfall. *Monthly Weather Review*, 115(5), 1053–1070. DOI: 10.1175/1520-0493(1987)115<1053:RBMRR>2.0.CO;2.
- AYZEL, G., M. HEISTERMANN, A. SOROKIN, O. NIKITIN & O. LUKYANOVA (2019a). All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Computer Science*, 150, 186–192. DOI: 10.1016/j.procs.2019.02.036.
- AYZEL, G., M. HEISTERMANN & T. WINTERRATH (2019b). Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0.1). *Geoscientific Model Development*, 12(4), 1387–1402. DOI: 10.5194/gmd-12-1387-2019.
- BAILEY, M. E., G. A. ISAAC, I. GULTEPE, I. HECKMAN & J. REID (2014). Adaptive blending of model and observations for automated short-range forecasting: Examples from the Vancouver 2010 Olympic and Paralympic winter games. *Pure and Applied Geophysics*, 171(1), 257–276. DOI: 10.1007/s00024-012-0553-x.
- BASHER, R. (2006). Global early warning systems for natural hazards: systematic and people-centred. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 364(1845), 2167–2182. DOI: 10.1098/rsta.2006.1819.
- VAN DE BEEK, C. Z., H. LEIJNSE, P. HAZENBERG & R. UIJLENHOET (2016). Close-range radar rainfall estimation and error analysis. *Atmospheric Measurement Techniques*, 9(8), 3837–3850. DOI: 10.5194/amt-9-3837-2016.
- VAN DE BEEK, C. Z., H. LEIJNSE, P. J. J. F. TORFS & R. UIJLENHOET (2012). Seasonal semi-variance of Dutch rainfall at honotey to daily scales. *Advances in Water Resources*, 45, 76–85. DOI: 10.1016/j.advwatres.2012.03.023.
- BEEKHUIS, H. & I. HOLLEMAN (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. https://cdn.knmi.nl/system/data_center_publications/files/000/068/061/original/erad2008drup_0120.pdf?1495621011.
- BEEKHUIS, H. & T. MATHIJSSSEN (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 (editors), *10th European Conference on Radar in Meteorology*

- and Hydrology (ERAD 2018) : 1-6 July 2018, Ede-Wageningen, The Netherlands, pages 960–965. Wageningen University & Research, Wageningen, the Netherlands. DOI: 10.18174/454537.
- BELLON, A., G. W. LEE & I. ZAWADZKI (2005). Error statistics of VPR corrections in stratiform precipitation. *Journal of Applied Meteorology and Climatology*, 44(7), 998–1015. DOI: 10.1175/JAM2253.1.
- BENGTSOON, L., U. ANDRAE, T. ASPELIEN, Y. BATRAK, J. CALVO, W. DE ROOY, E. GLEESON, B. HANSEN-SASS, M. HOMLEID, M. HORTAL, K. I. IVARSSON, G. LENDERINK, S. NIEMELÄ, K. P. NIELSEN, J. ONVLEE, L. RONTU, P. SAMUELSSON, D. S. MUÑOZ, A. SUBIAS, S. TIJM, V. TOLL, X. YANG & M. Ø. KØLTZOW (2017). The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Monthly Weather Review*, 145(5), 1919–1935. DOI: 10.1175/MWR-D-16-0417.1.
- BENJAMIN, S. G., S. S. WEYGANDT, J. M. BROWN, M. HU, C. R. ALEXANDER, T. G. SMIRNOVA, J. B. OLSON, E. P. JAMES, D. C. DOWELL, G. A. GRELL, H. LIN, S. E. PECKHAM, T. L. SMITH, W. R. MONINGER, J. S. KENYON & G. S. MANIKIN (2016). A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, 144(4), 1669–1694. DOI: 10.1175/MWR-D-15-0242.1.
- BERENGUER, M., C. CORRAL, R. SÁNCHEZ-DIEZMA & D. SEMPERE-TORRES (2005). Hydrological validation of a radar-based nowcasting technique. *Journal of Hydrometeorology*, 6(4), 532–549. DOI: 10.1175/JHM433.1.
- BERENGUER, M. & D. SEMPERE TORRES (2013). Radar-based rainfall nowcasting at european scale: long-term evaluation and performance assessment. In *36th Conference on Radar Meteorology*, pages 15B.3–1–15B.3–7.
- BERENGUER, M., D. SEMPERE-TORRES, C. CORRAL & R. SÁNCHEZ-DIEZMA (2006). A fuzzy logic technique for identifying nonprecipitating echoes in radar scans. *Journal of Atmospheric and Oceanic Technology*, 23(9), 1157–1180. DOI: 10.1175/JTECH1914.1.
- BERENGUER, M., D. SEMPERE-TORRES & G. G. PEGRAM (2011). SBMcst – An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *Journal of Hydrology*, 404(3-4), 226–240. DOI: 10.1016/j.jhydrol.2011.04.033.
- BERENGUER, M., M. SURCEL, I. ZAWADZKI, M. XUE & F. KONG (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. DOI: 10.1175/MWR-D-11-00181.1.
- BERNDTSSON, R. & J. NIEMCZYNOWICZ (1988). Spatial and temporal scales in rainfall analysis — Some aspects and future perspectives. *Journal of Hydrology*, 100(1), 293–313. DOI: 10.1016/0022-1694(88)90189-8.
- BEVEN, K. (1989). Changing ideas in hydrology—the case of physically-based models. *Journal of Hydrology*, 105(1-2), 157–172. DOI: 10.1016/0022-1694(89)90101-7.
- BEVEN, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, 16, 41–51. DOI: 10.1016/0309-1708(93)90028-E.
- BEVEN, K. (2006). Searching for the holy grail of scientific hydrology: $Q_t = (s, r, \delta t)a$ as closure. *Hydrology and Earth System Sciences*, 10(5), 609–618. DOI: 10.5194/hess-10-609-2006.
- BIERKENS, M. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, 51(7), 4923–4947. DOI: 10.1002/2015WR017173.
- BLÖSCHL, G. & M. SIVAPALAN (1995). Scale issues in hydrological modelling: a review. *Hydrological processes*, 9(3-4), 251–290. DOI: 10.1002/hyp.3360090305.
- BORGA, M. (2002). Accuracy of radar rainfall estimates for streamflow simulation. *Journal of Hydrology*, 267(1), 26–39. DOI: 10.1016/S0022-1694(02)00137-3.
- BORGA, M., E. N. ANAGNOSTOU & E. FRANK (2000). On the use of real-time radar rainfall estimates for flood prediction in mountainous basins. *Journal of Geophysical Research: Atmospheres*, 105(D2), 2269–2280. DOI: 10.1029/1999JD900270.
- BORGA, M., S. D. ESPOSTI & D. NORBIATO (2006). Influence of errors in radar rainfall estimates on hydrological modeling prediction uncertainty. *Water Resources Research*, 42(8). DOI: 10.1029/2005WR004559.
- BOTZEN, W. J. W. & J. C. J. M. VAN DEN BERGH (2008). Insurance against climate change and flooding in the Netherlands: Present, future, and comparison with other countries. *Risk Analysis*, 28(2), 413–426. DOI: 10.1111/j.1539-6924.2008.01035.x.
- BOUWER, L. M., P. BUBECK & J. C. J. H. AERTS (2010). Changes in future flood risk due to climate and development in a Dutch polder area. *Global Environmental Change*, 20(3), 463–471. DOI: 10.1016/j.gloenvcha.2010.04.002.
- BOWLER, N., C. E. PIERCE & A. W. SEED (2004). *STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP*. Forecasting Research Technical Report 433, Met Office, Wallingford, United Kingdom.

- BOWLER, N. E., C. E. PIERCE & A. W. SEED (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. DOI: 10.1256/qj.04.100.
- BRAUER, C. C., A. OVEREEM, H. LEIJNSE & R. UIJLENHOET (2016). The effect of differences between rainfall measurement techniques on groundwater and discharge simulations in a lowland catchment. *Hydrological Processes*, 30(21), 3885–3900. DOI: 10.1002/hyp.10898.
- BRAUER, C. C., A. J. TEULING, A. OVEREEM, Y. VAN DER VELDE, P. HAZENBERG, P. M. M. WARMERDAM & R. UIJLENHOET (2011). Anatomy of extraordinary rainfall and flash flood in a Dutch lowland catchment. *Hydrology and Earth System Sciences*, 15(6), 1991–2005. DOI: 10.5194/hess-15-1991-2011.
- BRAUER, C. C., A. J. TEULING, P. J. J. F. TORFS & R. UIJLENHOET (2014a). The Wageningen Lowland Runoff Simulator (WAL-RUS): a lumped rainfall-runoff model for catchments with shallow groundwater. *Geoscientific Model Development*, 7(5), 2313–2332. DOI: 10.5194/gmd-7-2313-2014.
- BRAUER, C. C., P. J. J. F. TORFS, A. J. TEULING & R. UIJLENHOET (2014b). The Wageningen Lowland Runoff Simulator (WAL-RUS): application to the Hupsel Brook catchment and the Cabauw polder. *Hydrology and Earth System Sciences*, 18(10), 4007–4028. DOI: 10.5194/hess-18-4007-2014.
- BROWNING, K. A. (1980). Review Lecture: Local weather forecasting. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 371(1745), 179–211. DOI: 10.1098/rspa.1980.0076.
- BRUTSAERT, W. (2013). *Evaporation into the atmosphere: Theory, history and applications*. Springer Science & Business Media, Berlin, Germany.
- BRUWIER, M., S. ERPICUM, M. PIROTTON, P. ARCHAMBEAU & B. J. DEWALS (2015). Assessing the operation rules of a reservoir system based on a detailed modelling chain. *Natural Hazards and Earth System Sciences*, 15(3), 365–379. DOI: 10.5194/nhess-15-365-2015.
- BUBNOVÁ, R., G. HELLO, P. BÉNARD & J. F. GELEYN (1995). Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP System. *Monthly Weather Review*, 123(2), 515–535. DOI: 10.1175/1520-0493(1995)123<0515:IOTFEE>2.0.CO;2.
- BUISHAND, T. & C. VELDS (1980). *Klimaat van Nederland 1: Neerslag en verdamping*. Royal Netherlands Meteorological Institute (KNMI).
- CEOLA, S., F. LAIO & A. MONTANARI (2014). Satellite nighttime lights reveal increasing human exposure to floods worldwide. *Geophysical Research Letters*, 41(20), 7184–7190. DOI: 10.1002/2014GL061859.
- CHEN, L., Y. CAO, L. MA & J. ZHANG (2020). A deep learning-based methodology for precipitation nowcasting with radar. *Earth and Space Science*, 7(2), e2019EA000812. DOI: 10.1029/2019EA000812.
- CHO, Y. H., G. LEE, K. E. KIM & I. ZAWADZKI (2006). Identification and removal of ground echoes and anomalous propagation using the characteristics of radar echoes. *Journal of Atmospheric and Oceanic Technology*, 23(9), 1206–1222. DOI: 10.1175/JTECH1913.1.
- CHWALA, C., A. GMEINER, W. QIU, S. HIPPE, D. NIENABER, U. SIART, T. EIBERT, M. POHL, J. SELTMANN, J. FRITZ & H. KUNSTMANN (2012). Precipitation observation using microwave backhaul links in the alpine and pre-alpine region of Southern Germany. *Hydrology and Earth System Sciences*, 16(8), 2647–2661. DOI: 10.5194/hess-16-2647-2012.
- CHWALA, C., F. KEIS & H. KUNSTMANN (2016). Real-time data acquisition of commercial microwave link networks for hydrometeorological applications. *Atmospheric Measurement Techniques*, 9(3), 991–999. DOI: 10.5194/amt-9-991-2016.
- CHWALA, C. & H. KUNSTMANN (2019). Commercial microwave link networks for rainfall observation: Assessment of the current status and future challenges. *WIREs Water*, 6(2), e1337. DOI: 10.1002/wat2.1337.
- CLARK, M., B. SCHAEFLI, S. SCHYMANSKI, L. SAMANIEGO, C. LUCE, B. JACKSON & ET AL. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, 52(3), 2350–2365. DOI: 10.1002/2015WR017910.
- CLARK, M. P., M. F. P. BIERKENS, L. SAMANIEGO, R. A. WOODS, R. UIJLENHOET, K. E. BENNETT, V. R. N. PAUWELS, X. CAI, A. W. WOOD & C. D. PETERS-LIDARD (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. DOI: 10.5194/hess-21-3427-2017.
- CLOKE, H. & F. PAPPENBERGER (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375(3–4), 613–626. DOI: 10.1016/j.jhydrol.2009.06.005.
- COIFFIER, J. (2011). *Fundamentals of Numerical Weather Prediction*. Cambridge University Press, Cambridge, United Kingdom.

- COX, D., J. HUNT, P. MASON, H. WHEATER, P. WOLF, H. GUPTA, S. SOROOSHIAN, X. GAO, B. IMAM, K. L. HUS, L. BASTIDAS, J. LI & S. MAHANI (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. DOI: 10.1098/rsta.2002.1015.
- CREUTIN, J. D., H. ANDRIEU & D. FAURE (1997). Use of a weather radar for the hydrology of a mountainous area. Part II: radar measurement validation. *Journal of Hydrology*, 193(1), 26–44. DOI: 10.1016/S0022-1694(96)03203-9.
- CREUTIN, J. D., G. DELRIEU & T. LEBEL (1988). Rain measurement by raingage-radar combination: A geostatistical approach. *Journal of Atmospheric and Oceanic Technology*, 5(1), 102–115. DOI: 10.1175/1520-0426(1988)005<0102:RMBRRC>2.0.CO;2.
- CUOMO, J. & V. CHANDRASEKAR (2022). Developing deep learning models for storm nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13. DOI: 10.1109/TGRS.2021.3110180.
- DAVENPORT, F. V., M. BURKE & N. S. DIFFENBAUGH (2021). Contribution of historical precipitation change to US flood damages. *Proceedings of the National Academy of Sciences*, 118(4), e2017524118. DOI: 10.1073/pnas.2017524118.
- DEMERRIT, D., H. CLOKE, F. PAPPENBERGER, J. THIELEN, J. BARTHOLMES & M. RAMOS (2007). Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*, 7(2), 115–127. DOI: 10.1016/j.envhaz.2007.05.001.
- DIXON, M. & G. WIENER (1993). TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting — A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, 10(6), 785–797. DOI: 10.1175/1520-0426(1993)010<0785:TITIAA>2.0.CO;2.
- DOUMOUNIA, A., M. GOSSET, F. CAZENAVE, M. KACOU & F. ZOUGMORE (2014). Rainfall monitoring based on microwave links from cellular telecommunication networks: First results from a West African test bed. *Geophysical Research Letters*, 41(16), 6016–6022. DOI: 10.1002/2014GL060724.
- EBERT, E. E., L. J. WILSON, B. G. BROWN, P. NURMI, H. E. BROOKS, J. BALLY & M. JAENEKE (2004). Verification of nowcasts from the WWRP Sydney 2000 forecast demonstration project. *Weather and Forecasting*, 19(1), 73–96. DOI: 10.1175/1520-0434(2004)019<0073:VONFTW>2.0.CO;2.
- ERICSSON (2016). *Ericsson microwave outlook: Trends and needs in the microwave industry*. Technical report, Ericsson, Stockholm, Sweden. <https://www.ericsson.com/4adebb/assets/local/reports-papers/microwave-outlook/2016/ericsson-microwave-outlook-report-2016.pdf>.
- ERICSSON (2019). *Ericsson microwave outlook: Enhancing 5G with microwave*. Technical report, Ericsson, Stockholm, Sweden. <https://www.ericsson.com/en/reports-and-papers/microwave-outlook/reports/2019>.
- EUROPEAN ENVIRONMENT AGENCY (2004). *Mapping the impacts of recent natural disasters and technological accidents in Europe*. 35. Office for Official Publications of the European Communities, Luxembourg.
- FABRY, F., G. L. AUSTIN & D. TEES (1992). The accuracy of rainfall estimates by radar as a function of range. *Quarterly Journal of the Royal Meteorological Society*, 118(505), 435–453. DOI: 10.1002/qj.49711850503.
- FEIGENBAUM, M. J. (1983). Universal behavior in nonlinear systems. *Physica D: Nonlinear Phenomena*, 7(1), 16–39. DOI: 10.1016/0167-2789(83)90112-4.
- FEIGL, M., M. HERRNEGGER, D. KLOTZ & K. SCHULZ (2020). Function space optimization: A symbolic regression method for estimating parameter transfer functions for hydrological models. *Water Resources Research*, 56(10), e2020WR027385. DOI: 10.1029/2020WR027385.
- FENCL, M., M. DOHNAL, P. VALTR, M. GRABNER & V. BAREŠ (2020). Atmospheric observations with E-band microwave links - challenges and opportunities. *Atmospheric Measurement Techniques Discussions*, pages 1–29. DOI: 10.5194/amt-2020-28.
- FERRARIS, L., R. RUDARI & F. SICCARDI (2002). The uncertainty in the prediction of flash floods in the northern Mediterranean environment. *Journal of Hydrometeorology*, 3(6), 714–727. DOI: 10.1175/1525-7541(2002)003<0714:TUITP0>2.0.CO;2.
- FORESTI, L., M. REYNIERS, A. SEED & L. DELOBBE (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. DOI: 10.5194/hess-20-505-2016.
- FORESTI, L. & A. SEED (2015). On the spatial distribution of rainfall nowcasting errors due to orographic forcing. *Meteorological Applications*, 22(1), 60–74. DOI: 10.1002/met.1440.
- FORESTI, L., I. V. SIDERIS, D. NERINI, L. BEUSCH & U. GERMANN (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. DOI: 10.1175/WAF-D-18-0206.1.

- FORESTI, L., I.V. SIDERIS, L. PANZIERA, D. NERINI & U. GERMANN (2018). A 10-year radar-based analysis of orographic precipitation growth and decay patterns over the Swiss Alpine region. *Quarterly Journal of the Royal Meteorological Society*, 144(716), 2277–2301. DOI: 10.1002/qj.3364.
- FOUFOULA-GEORGIOU, E. (1998). On scaling theories of space-time rainfall: some recent results and open problems. In *Stochastic methods in hydrology*, volume 7 of *Advanced series on statistical science & applied probability*, pages 25–72. World Scientific, Singapore, Singapore. DOI: 10.1142/9789812839725\$\$_{0002}.
- FRAME, D.J., S.M. ROSIER, I. NOY, L.J. HARRINGTON, T. CAREY-SMITH, S.N. SPARROW, D.A. STONE & S.M. DEAN (2020). Climate change attribution and the economic costs of extreme weather events: a study on damages from extreme rainfall and drought. *Climatic Change*, 162(2), 781–797. DOI: 10.1007/s10584-020-02729-y.
- FRANCH, G., D. NERINI, M. PENDESINI, L. COVIELLO, G. JURMAN & C. FNOTEANELLO (2020). Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events. *Atmosphere*, 11(3), 267. DOI: 10.3390/atmos11030267.
- FRENCH, M.N., W.F. KRAJEWSKI & R.R. CUYKENDALL (1992). Rainfall forecasting in space and time using a neural network. *Journal of Hydrology*, 137(1), 1–31. DOI: 10.1016/0022-1694(92)90046-X.
- FRIEDMAN, R.M. (1993). *Appropriating the weather: Vilhelm Bjerknes and the construction of a modern meteorology*. Cornell University Press, Ithaca, NY, United States of America.
- FRISINGER, H.H. (2018). *History of Meteorology to 1800*. Springer, Berlin, Germany.
- GABELLA, M., J. JOSS & G. PERONA (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. DOI: 10.1029/1999JD900420.
- GANGOPADHYAY, S., M. CLARK, K. WERNER, D. BRANDON & B. RAJAGOPALAN (2004). Effects of spatial and temporal aggregation on the accuracy of statistically downscaled precipitation estimates in the Upper Colorado River basin. *Journal of Hydrometeorology*, 5(6), 1192–1206. DOI: 10.1175/JHM-391.1.
- GEORGAKAKOS, K.P. & R.L. BRAS (1984a). A hydrologically useful station precipitation model: 1. Formulation. *Water Resources Research*, 20(11), 1585–1596. DOI: 10.1029/WR020i011p01585.
- GEORGAKAKOS, K.P. & R.L. BRAS (1984b). A hydrologically useful station precipitation model: 2. Case studies. *Water Resources Research*, 20(11), 1597–1610. DOI: 10.1029/WR020i011p01597.
- GERARD, L., J.M. PIRIOU, R. BROŽKOVÁ, J.F. GELEYN & D. BANCUI (2009). Cloud and precipitation parameterization in a Meso-Gamma-scale operational weather prediction model. *Monthly Weather Review*, 137(11), 3960–3977. DOI: 10.1175/2009MWR2750.1.
- GERMANN, U., M. BERENGUER, D. SEMPERE-TORRES & M. ZAPPA (2009). REAL-Ensemble radar precipitation estimation for hydrology in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 135(639), 445–456. DOI: 10.1002/qj.375.
- GERMANN, U. & J. JOSS (2002). Mesobeta profiles to extrapolate radar precipitation measurements above the Alps to the ground level. *Journal of Applied Meteorology and Climatology*, 41(5), 542–557. DOI: 10.1175/1520-0450(2002)041<0542:MPTEPF>2.0.CO;2.
- GERMANN, U. & I. ZAWADZKI (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. DOI: 10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2.
- GERMANN, U. & I. ZAWADZKI (2004). Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *Journal of Applied Meteorology*, 43(1), 74–89. DOI: 10.1175/1520-0450(2004)043<0074:SDOTPO>2.0.CO;2.
- GERMANN, U., I. ZAWADZKI & B. TURNER (2006). Predictability of precipitation from continental radar images. Part IV: Limits to prediction. *Journal of the Atmospheric Sciences*, 63(8), 2092–2108. DOI: 10.1175/JAS3735.1.
- GERRITSEN, T. (2019). *Hydrological intercomparison of rain gauge, weather radar and satellite observations*. Master's thesis, Wageningen University & Research, Wageningen, The Netherlands.
- GOLDING, B.W. (1998). Nimrod: A system for generating automated very short range forecasts. *Meteorological Applications*, 5(1), 1–16. DOI: 10.1017/S1350482798000577.
- GOSSET, M., H. KUNSTMANN, F. ZOUGMORE, F. CAZENAVE, H. LEIJNSE, R. UIJLENHOET, C. CHWALA, F. KEIS, A. DOUMOUNIA, B. BOUBACAR, M. KACOU, P. ALPERT, H. MESSER, J. RIECKERMANN & J. HOEDJES (2015). Improving rainfall measurement in gauge poor regions thanks to mobile telecommunication networks. *Bulletin of the American Meteorological Society*, 97(3),

- ES49–ES51. DOI: 10.1175/BAMS-D-15-00164.1.
- GOUDENHOOFDT, E. & L. DELOBBE (2009). Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrology and Earth System Sciences*, 13(2), 195–203. DOI: 10.5194/hess-13-195-2009.
- GOUDENHOOFDT, E. & L. DELOBBE (2016). Generation and verification of rainfall estimates from 10-Yr volumetric weather radar measurements. *Journal of Hydrometeorology*, 17(4), 1223–1242. DOI: 10.1175/JHM-D-15-0166.1.
- GRAF, M., C. CHWALA, J. POLZ & H. KUNSTMANN (2020). Rainfall estimation from a German-wide commercial microwave link network: optimized processing and validation for 1 year of data. *Hydrology and Earth System Sciences*, 24(6), 2931–2950. DOI: 10.5194/hess-24-2931-2020.
- GSMA (2019). *Mobile technology for rural climate resilience: The role of mobile operators in bridging the data gap*. Technical report, GSMA, London, UK. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/10/GSMA_AgriTech_Climate_Report.pdf.
- GUPTA, H. V., H. KLING, K. K. YILMAZ & G. F. MARTINEZ (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. DOI: 10.1016/j.jhydrol.2009.08.003.
- GUPTA, V. K. & E. C. WAYMIRE (1979). A stochastic kinematic study of subsynoptic space-time rainfall. *Water Resources Research*, 15(3), 637–644. DOI: 10.1029/WR015i003p00637.
- HAASE, G., S. CREWELL, C. SIMMER & W. WERGEN (2000). Assimilation of radar data in mesoscale models: Physical initialization and latent heat nudging. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 25(10), 1237–1242. DOI: 10.1016/S1464-1909(00)00186-6.
- HALLEGATTE, S. (2012). *A cost effective solution to reduce disaster losses in developing countries: Hydro-meteorological services, early warning, and evacuation*. Policy Research Working Papers. The World Bank, Washington D.C., United States of America. DOI: 10.1596/1813-9450-6058.
- HAMILTON, J. D. (1994). *Time series analysis*. Princeton University Press, NJ, United States of America.
- HAN, L., S. FU, L. ZHAO, Y. ZHENG, H. WANG & Y. LIN (2009). 3D convective storm identification, tracking, and forecasting — An enhanced TITAN algorithm. *Journal of Atmospheric and Oceanic Technology*, 26(4), 719–732. DOI: 10.1175/2008JTECHA1084.1.
- HAN, L., J. SUN, W. ZHANG, Y. XIU, H. FENG & Y. LIN (2017). A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, 122(7), 4038–4051. DOI: 10.1002/2016JD025783.
- HARRIS, D., M. MENABDE, A. SEED & G. AUSTIN (1996). Multifractal characterization of rain fields with a strong orographic influence. *Journal of Geophysical Research: Atmospheres*, 101(D21), 26405–26414. DOI: 10.1029/96JD01656.
- HARRISON, D. L., R. W. SCOVELL & M. KITCHEN (2009). High-resolution precipitation estimates for hydrological uses. *Proceedings of the Institution of Civil Engineers - Water Management*, 162(2), 125–135. DOI: 10.1680/wama.2009.162.2.125.
- HAZENBERG, P., H. LEIJNSE & R. UIJLENHOET (2011). Radar rainfall estimation of stratiform winter precipitation in the Belgian Ardennes. *Water Resources Research*, 47(2). DOI: 10.1029/2010WR009068.
- HAZENBERG, P., H. LEIJNSE & R. UIJLENHOET (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. DOI: 10.1016/j.jhydrol.2014.09.057.
- HAZENBERG, P., P. J. J. F. TORFES, H. LEIJNSE, G. DELRIEU & R. UIJLENHOET (2013). Identification and uncertainty estimation of vertical reflectivity profiles using a Lagrangian approach to support quantitative precipitation measurements by weather radar: VPR estimation and uncertainty. *Journal of Geophysical Research: Atmospheres*, 118(18), 10,243–10,261. DOI: 10.1002/jgrd.50726.
- HEINSELMAN, P. L. & S. M. TORRES (2011). High-temporal-resolution capabilities of the national weather radar testbed phased-array radar. *Journal of Applied Meteorology and Climatology*, 50(3), 579–593. DOI: 10.1175/2010JAMC2588.1.
- HERING, A. M., C. MOREL, G. GALLI, S. SENESI, P. AMBROSETTI & M. BOSCACCI (2006). Nowcasting thunderstorms in the Alpine region using a radar based adaptive thresholding scheme. In *3th European Conference on Radar in Meteorology and Hydrology (ERAD 2004)*, pages 206–211. Copernicus, Visby, Sweden, Visby, Sweden.
- HERSBACH, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5), 559–570. DOI: 10.1175/1520-0434(2000)015<0559:D0TCRP>2.0.CO;2.
- HEUVELINK, D., M. BERENGUER, C. C. BRAUER & R. UIJLENHOET (2020). Hydrological application of radar rainfall nowcasting in the Netherlands. *Environment International*, 136, 105431. DOI: 10.1016/j.envint.2019.105431.

- HIEMSTRA, P. & R. SLUITER (2011). *Interpolation of Makkink evaporation in the Netherlands*. Technical Report TR-327, Royal Netherlands Meteorological Institute, De Bilt, The Netherlands. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.705.9459rep=rep1type=pdf>.
- HILL, P.G., T.H. STEIN, A.J. ROBERTS, J.K. FLETCHER, J.H. MARSHAM & J. GROVES (2020). How skilful are nowcasting satellite applications facility products for tropical Africa? *Meteorological Applications*, 27(6), e1966. DOI: 10.1002/met.1966.
- HIRABAYASHI, Y., R. MAHENDRAN, S. KOIRALA, L. KONOSHIMA, D. YAMAZAKI, S. WATANABE, H. KIM & S. KANAE (2013). Global flood risk under climate change. *Nature Climate Change*, 3(9), 816–821. DOI: 10.1038/nclimate1911.
- HOGG, D.C. (1968). Millimeter-wave communication through the atmosphere. *Science*, 159(3810), 39–46. DOI: 10.1126/science.159.3810.39.
- HOLLEMAN, I. (2007). Bias adjustment and long-term verification of radar-based precipitation estimates. *Meteorological Applications*, 14(2), 195–203. DOI: 10.1002/met.22.
- HONDA, T., A. AMEMIYA, S. OTSUKA, J. TAYLOR, Y. MAEJIMA, S. NISHIZAWA, T. YAMAURA, K. SUEKI, H. TOMITA & T. MIYOSHI (2022). Advantage of 30-s-updating numerical weather prediction with a phased-array weather radar over operational nowcast for a convective precipitation system. *Geophysical Research Letters*, 49(11), e2021GL096927. DOI: 10.1029/2021GL096927.
- HWANG, Y., A.J. CLARK, V. LAKSHMANAN & S.E. KOCH (2015). Improved nowcasts by blending extrapolation and model forecasts. *Weather and Forecasting*, 30(5), 1201–1217. DOI: 10.1175/WAF-D-15-0057.1.
- IMHOFF, R.O., C. BRAUER, K.J. VAN HEERINGEN, H. LEIJNSE, A. OVEREEM, A. WEERTS & R. UIJLENHOET (2021). A climatological benchmark for operational radar rainfall bias reduction. *Hydrology and Earth System Sciences*, 25(7), 4061–4080. DOI: 10.5194/hess-25-4061-2021.
- IMHOFF, R.O., C.C. BRAUER, K.J. VAN HEERINGEN, R. UIJLENHOET & A.H. WEERTS (2022). Large-sample evaluation of radar rainfall nowcasting for flood early warning. *Water Resources Research*, 58(3), e2021WR031591. DOI: 10.1029/2021WR031591.
- IMHOFF, R.O., C.C. BRAUER, A. OVEREEM, A.H. WEERTS & R. UIJLENHOET (2020a). Spatial and temporal evaluation of radar rainfall nowcasting techniques on 1,533 events. *Water Resources Research*, 56(8), e2019WR026723. DOI: 10.1029/2019WR026723.
- IMHOFF, R.O., A. OVEREEM, C.C. BRAUER, H. LEIJNSE, A.H. WEERTS & R. UIJLENHOET (2020b). Rainfall nowcasting using commercial microwave links. *Geophysical Research Letters*, 47(19), e2020GL089365. DOI: 10.1029/2020GL089365.
- IMHOFF, R.O., W.J. VAN VERSEVELD, B. VAN OSNABRUGGE & A.H. WEERTS (2020c). Scaling point-scale (pedo)transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the Rhine River. *Water Resources Research*, 56(4), e2019WR026807. DOI: 10.1029/2019WR026807.
- INGRAM, K.T., M.C. RONCOLI & P.H. KIRSHEN (2002). Opportunities and constraints for farmers of West Africa to use seasonal precipitation forecasts with Burkina Faso as a case study. *Agricultural systems*, 74(3), 331–349. DOI: 10.1016/S0308-521X(02)00044-6.
- INNESS, P.M. & S. DORLING (2012). *Operational weather forecasting*. John Wiley & Sons, Hoboken, NJ, United States of America.
- IPCC (2011). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY.
- IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY.
- IPCC (2014). *Part A: Global and Sectoral Aspects, Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Climate Change 2014: Impacts, Adaptation and Vulnerability. Cambridge University Press, Cambridge, United Kingdom and New York, NY.
- IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY.
- JAIN, S.K., P. MANI, S.K. JAIN, P. PRAKASH, V.P. SINGH, D. TULLOS, S. KUMAR, S.P. AGARWAL & A.P. DIMRI (2018). A brief review of flood forecasting techniques and their applications. *International Journal of River Basin Management*, 16(3), 329–344. DOI: 10.1080/15715124.2017.1411920.
- JOLLIFFE, I.T. & D.B. STEPHENSON (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons,

- Chichester, West Sussex, United Kingdom, 2 edition.
- JONGMAN, B. (2018). Effective adaptation to rising flood risk. *Nature Communications*, 9(1), 1986. DOI: 10.1038/s41467-018-04396-1.
- JONGMAN, B., P.J. WARD & J.C.J.H. AERTS (2012). Global exposure to river and coastal flooding: Long term trends and changes. *Global Environmental Change*, 22(4), 823–835. DOI: 10.1016/j.gloenvcha.2012.07.004.
- JOSS, J. & R. LEE (1995). The application of radar–gauge comparisons to operational precipitation profile corrections. *Journal of Applied Meteorology*, 34(12), 2612–2630. DOI: 10.1175/1520-0450(1995)034<2612:TAORCT>2.0.CO;2.
- JOSS, J. & A. PITTINI (1991). Real-time estimation of the vertical profile of radar reflectivity to improve the measurement of precipitation in an Alpine region. *Meteorology and Atmospheric Physics*, 47(1), 61–72. DOI: 10.1007/BF01025828.
- KIDD, C., A. BECKER, G.J. HUFFMAN, C.L. MULLER, P. JOE, G. SKOFRONICK-JACKSON & D.B. KIRSCHBAUM (2017). So, how much of the Earth's surface is covered by rain gauges? *Bulletin of the American Meteorological Society*, 98(1), 69–78. DOI: 10.1175/BAMS-D-14-00283.1.
- KIRSTETTER, P.E., H. ANDRIEU, G. DELRIEU & B. BOUDEVILLAIN (2010). Identification of vertical profiles of reflectivity for correction of volumetric radar data using rainfall classification. *Journal of Applied Meteorology and Climatology*, 49(10), 2167–2180. DOI: 10.1175/2010JAMC2369.1.
- KITCHEN, M. & P.M. JACKSON (1993). Weather radar performance at long range - simulated and observed. *Journal of Applied Meteorology and Climatology*, 32(5), 975–985. DOI: 10.1175/1520-0450(1993)032<0975:WRPALR>2.0.CO;2.
- KLEIN TANK, A., J. BEERSMA, J. BESSEMBINDER, B. VAN DEN HURK & G. LENDERINK (2014). *KNMI '14 : climate scenarios for the Netherlands : a guide for professionals inclimate adaption*. KNMI, De Bilt, The Netherlands. <https://edepot.wur.nl/328690>.
- KNMI (2009). *KNMI - Jaar 2008: Twaalfde warme jaar op rij*. KNMI, De Bilt, The Netherlands. <https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2008/jaar>.
- KNMI (2011). *Klimaatatlas: Langjarige gemiddelden 1981-2010*. KNMI, De Bilt, The Netherlands. <http://www.klimaatatlas.nl/>.
- KNMI (2015). *KNMI'14: climate scenarios for the Netherlands; A guide for professionals in climate adaption*. KNMI, De Bilt, The Netherlands.
- KNMI (2022). *KNMI - Vier landen bundelen krachten voor betere weersverwachting*. KNMI, De Bilt, The Netherlands. <https://www.knmi.nl/over-het-knmi/nieuws/vier-landen-bundelen-krachten-voor-betere-weersverwachting>.
- KOBER, K., G.C. CRAIG & C. KEIL (2014). Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 140(681), 1179–1188. DOI: 10.1002/qj.2220.
- KOBER, K., G.C. CRAIG, C. KEIL & A. DÖRNBROCK (2012). Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(664), 755–768. DOI: 10.1002/qj.939.
- KOISTINEN, J., R. KING, E. SALTIKOFF & A. HARJU (1999). Monitoring and assessment of systematic measurement errors in the NORDRAD network. In *29th International Conference on Radar Meteorology*, pages 765–768. Montreal, Canada.
- KOISTINEN, J. & T. PUHAKKA (1981). An improved spatial gauge-radar adjustment technique. In *20th Conference on Radar Meteorology*, pages 179–186. Boston, MA, USA.
- KOKS, E., K. VAN GINKEL, M. VAN MARLE & A. LEMNITZER (2021). Brief Communication: Critical Infrastructure impacts of the 2021 mid-July western European flood event. *Natural Hazards and Earth System Sciences Discussions*, pages 1–11. DOI: 10.5194/nhess-2021-394.
- KOOPMANS, S., R. VAN HAREN, G.J. STEENEVELD, N.E. THEEUWES, R.J. RONDA, R. UIJLENHOET & A.A.M. HOLTSLAG (2018). Data assimilation of urban weather observations in WRF to create a re-analysis of the urban climate of Amsterdam. AMS, New York, United States of America.
- KRAJEWSKI, W.F. (1987). Cokriging radar-rainfall and rain gage data. *Journal of Geophysical Research: Atmospheres*, 92(D8), 9571–9580. DOI: 10.1029/JD092iD08p09571.
- KREIENKAMP, F., S.Y. PHILIP, J.S. TRADOWSKY, S.F. KEW, P. LORENZ, J. ARRIGHI, A. BELLEFLAMME, T. BETTMANN, S. CALUWAERTS, S.C. CHAN, A. CIAVARELLA, L. DE CRUZ, H. DE VRIES, N. DEMUTH, A. FERRONE, R.M. FISCHER, H.J. FOWLER, K. GOERGEN, D. HEINRICH, Y. HENRICH, G. LENDERINK, F. KASPAR, E. NILSON, F.E.L. OTTO, F. RAGONE, S.I. SENEVIRATNE, R.K. SINGH, A. SKÅLEVÅG, P. TERMONIA, L. THALHEIMER, M. VAN AALST, J. VAN DEN BERGH, H. VAN DE VYVER, S. VANNITSEM, G.J. VAN OLDENBORGH, B. VAN SCHAEYBROECK, R. VAUTARD, D. VONK & N. WANDERS (2021). Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021. *World Weather*

Atribution.

- KROEGER, T., R. TIMOFTE, D. DAI & L. VAN GOOL (2016). Fast optical flow using dense inverse search. In B. Leibe, J. Matas, N. Sebe, & M. Welling (editors), *European Conference on Computer Vision*, volume 9908 of *Lecture Notes in Computer Science*, pages 471–488. Springer. DOI: 10.1007/978-3-319-46493-0_8_\$29.
- KULL, D., L.P. RIISHOJGAARD, J. EYRE & R.A. VARLEY (2021). *The Value of Surface-based Meteorological Observation Data*. Technical report, World Bank, Washington DC, DC, United States of America. <https://openknowledge.worldbank.org/bitstream/handle/10986/35178/The-Value-of-Surface-based-Meteorological-Observation-Data.pdf?sequence=1&isAllowed=y>.
- KUMAR, A., T. ISLAM, Y. SEKIMOTO, C. MATTMANN & B. WILSON (2020). Convcast: An embedded convolutional LSTM based architecture for precipitation nowcasting using satellite data. *PLOS ONE*, 15(3), e0230114. DOI: 10.1371/journal.pone.0230114.
- KUNDZEWICZ, Z. W., S. KANAE, S. I. SENEVIRATNE, J. HANDMER, N. NICHOLLS, P. PEDUZZI, R. MECHLER, L. M. BOUWER, N. ARNEILL, K. MACH, R. MUIR-WOOD, G. R. BRAKENRIDGE, W. KRON, G. BENITO, Y. HONDA, K. TAKAHASHI & B. SHERSTYUKOV (2014). Flood risk and climate change: global and regional perspectives. *Hydrological Sciences Journal*, 59(1), 1–28. DOI: 10.1080/02626667.2013.857411.
- LADSON, A. R., R. BROWN, B. NEAL & R. NATHAN (2013). A standard approach to baseflow separation using the Lyne and Hollick filter. *Australasian Journal of Water Resources*, 17(1), 25–34. DOI: 10.7158/13241583.2013.11465417.
- LEIJNSE, H., R. UIJLENHOET & J. N. M. STRICKER (2007). Rainfall measurement using radio links from cellular communication networks. *Water Resources Research*, 43(3). DOI: 10.1029/2006WR005631.
- LEIJNSE, H., R. UIJLENHOET & J. N. M. STRICKER (2008). Microwave link rainfall estimation: Effects of link length and frequency, temporal sampling, power resolution, and wet antenna attenuation. *Advances in Water Resources*, 31(11), 1481–1493. DOI: 10.1016/j.advwatres.2008.03.004.
- 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., M. RICO-RAMIREZ, A. SCHELLART & A. SAUL (2012). Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. *Atmospheric Research*, 103, 80–95. DOI: 10.1016/j.atmosres.2011.05.004.
- LIGUORI, S. & M. A. RICO-RAMIREZ (2012). Quantitative assessment of short-term rainfall forecasts from radar nowcasts and MM5 forecasts. *Hydrological Processes*, 26(25), 3842–3857. DOI: 10.1002/hyp.8415.
- LIGUORI, S. & M. A. RICO-RAMIREZ (2013). A practical approach to the assessment of probabilistic flow predictions. *Hydrological Processes*, 27(1), 18–32. DOI: 10.1002/hyp.9468.
- LIN, C., S. VASIĆ, A. KILAMBI, B. TURNER & I. ZAWADZKI (2005). Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, 32(14), L14801. DOI: 10.1029/2005GL023451.
- LOBLIGEIS, F., V. ANDRÉASSIAN, C. PERRIN, P. TABARY & C. LOUMAGNE (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. DOI: 10.5194/hess-18-575-2014.
- LOOS, R. (2015a). *Making WALRUS applicable for large catchments: a case study in the Reusel catchment*. Master’s thesis, Wageningen University & Research, Wageningen, The Netherlands.
- LOOS, R. (2015b). *MSc internship report at water authority Vechtstromen: Development of WALRUS models for FEWS Vecht*. Technical report, Wageningen University & Research, Wageningen, The Netherlands.
- LORENZ, C. & H. KUNSTMANN (2012). The hydrological cycle in three state-of-the-art reanalyses: Intercomparison and performance analysis. *Journal of Hydrometeorology*, 13(5), 1397–1420. DOI: 10.1175/JHM-D-11-088.1.
- LORENZ, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2), 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
- LORENZ, E. N. (1993). *The essence of chaos*. UCL Press, London, United Kingdom.
- LOVEJOY, S. (2019). *Weather, macroweather, and the climate: Our random yet predictable atmosphere*. Oxford University Press, Oxford, United Kingdom.
- LOVEJOY, S. & D. SCHERTZER (1995). Multifractals and rain. , ed. aw kundzewicz, 62–103. In *New Uncertainty Concepts in Hydrology and Hydrological Modelling*, page 61. Cambridge University Press, New York, NY.
- LU, C., H. YUAN, B. E. SCHWARTZ & S. G. BENJAMIN (2007). Short-range numerical weather prediction using time-lagged

- ensembles. *Weather and Forecasting*, 22(3), 580–595. DOI: 10.1175/WAF999.1.
- LUCAS, B. D., T. KANADE ET AL. (1981). An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, pages 674–679. Vancouver, British Columbia, Canada.
- LYNCH, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7), 3431–3444. DOI: 10.1016/j.jcp.2007.02.034.
- 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.
- MARSAN, D., D. SCHERTZER & S. LOVEJOY (1996). Causal space-time multifractal processes: Predictability and forecasting of rain fields. *Journal of Geophysical Research: Atmospheres*, 101(D21), 26333–26346. DOI: 10.1029/96JD01840.
- MARSHALL, J. S., W. HITSCHFELD & K. L. S. GUNN (1955). Advances in radar weather. In H. E. Landsberg (editor), *Advances in Geophysics*, volume 2, pages 1–56. Academic Press Inc., New York, NY.
- McKAY, M. D., R. J. BECKMAN & W. J. CONOVER (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. DOI: 10.2307/1268522.
- MEJSNAR, J., Z. SOKOL & J. MINÁŘOVÁ (2018). Limits of precipitation nowcasting by extrapolation of radar reflectivity for warm season in Central Europe. *Atmospheric Research*, 213, 288–301. DOI: 10.1016/j.atmosres.2018.06.005.
- MELSEN, L. A., A. J. TEULING, P. J. F. TORFS, R. UIJLENHOET, N. MIZUKAMI & M. P. CLARK (2016). HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models. *Hydrology and Earth System Sciences*, 20(3), 1069–1079. DOI: 10.5194/hess-20-1069-2016.
- MERZ, B., H. KREIBICH, R. SCHWARZE & A. THIEKEN (2010). Review article “Assessment of economic flood damage”. *Natural Hazards and Earth System Sciences*, 10(8), 1697–1724. DOI: 10.5194/nhess-10-1697-2010.
- MERZ, B., C. KUHLICKE, M. KUNZ, M. PITTORE, A. BABEYKO, D. N. BRESCH, D. I. V. DOMEISEN, F. FESER, I. KOSZALKA, H. KREIBICH, F. PANTILLON, S. PAROLAI, J. G. PINTO, H. J. PUNGE, E. RIVALLA, K. SCHRÖTER, K. STREHLOW, R. WEISSE & A. WURPTS (2020). Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics*, 58(4), e2020RG000704. DOI: 10.1029/2020RG000704.
- MESSER, H., A. ZINEVICH & P. ALPERT (2006). Environmental monitoring by wireless communication networks. *Science*, 312(5774), 713–713. DOI: 10.1126/science.1120034.
- MICHELSON, D. B. & J. KOISTINEN (2000). Gauge-Radar network adjustment for the baltic sea experiment. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 25(10), 915–920. DOI: 10.1016/S1464-1909(00)00125-8.
- MIRZA, M. M. Q. (2011). Climate change, flooding in South Asia and implications. *Regional Environmental Change*, 11(1), 95–107. DOI: 10.1007/s10113-010-0184-7.
- MITTERMAIER, M., N. ROBERTS & S. A. THOMPSON (2013). A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorological Applications*, 20(2), 176–186. DOI: 10.1002/met.296.
- MITTERMAIER, M. P. (2006). Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmospheric Science Letters*, 7(2), 36–42. DOI: 10.1002/asl.127.
- MITTERMAIER, M. P. & N. ROBERTS (2010). Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Weather and Forecasting*, 25(1), 343–354. DOI: 10.1175/2009WAF2222260.1.
- MIZUKAMI, N., M. CLARK, A. NEWMAN, A. WOOD, E. GUTMANN, B. NIJSSEN & ET AL. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, 53(9), 8020–8040. DOI: 10.1002/2017WR020401.
- MORENO, H. A., E. R. VIVONI & D. J. GOCHIS (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. DOI: 10.1175/JHM-D-12-0129.1.
- MOULIN, L., E. GAUME & C. OBLED (2009). Uncertainties on mean areal precipitation: assessment and impact on streamflow simulations. *Hydrology and Earth System Sciences*, 13(2), 99–114. DOI: 10.5194/hess-13-99-2009.
- MUELLER, C., T. SAXEN, R. ROBERTS, J. WILSON, T. BETANCOURT, S. DETTLING, N. OIEN & J. YEE (2003). NCAR auto-nowcast system. *Weather and Forecasting*, 18(4), 545–561. DOI: 10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2.
- NA, W. & C. YOO (2018). A bias correction method for rainfall forecasts using backward storm tracking. *Water*, 10(12), 1728. DOI: 10.3390/w10121728.
- NASH, J. E. & J. V. SUTCLIFFE (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. DOI: 10.1016/0022-1694(70)90255-6.

- NAVON, I. M. (2009). Data assimilation for Numerical Weather Prediction: A review. In S. K. Park & L. Xu (editors), *Data assimilation for atmospheric, oceanic and hydrologic applications*, pages 21–65. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-540-71056-1_5_2.
- NEBEKER, F. (1995). *Calculating the weather: Meteorology in the 20th century*. Elsevier, Amsterdam, The Netherlands.
- NERINI, D., N. BESIC, I. SIDERIS, U. GERMANN & L. FORESTI (2017a). 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. doi: 10.5194/hess-21-2777-2017.
- NERINI, D., N. BESIC, I. SIDERIS, U. GERMANN & L. FORESTI (2017b). 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. doi: 10.5194/hess-21-2777-2017.
- NERINI, D., L. FORESTI, D. LEUENBERGER, S. ROBERT & U. GERMANN (2019). A reduced-space ensemble Kalman filter approach for flow-dependent integration of radar extrapolation nowcasts and NWP precipitation ensembles. *Monthly Weather Review*, 147(3), 987–1006. doi: 10.1175/MWR-D-18-0258.1.
- OCHOA-RODRIGUEZ, S., M. RICO-RAMIREZ, S. A. JEWELL, A. N. A. SCHELLART, L. WANG, C. ONOF & V. MAK-SIMOVIĆ (2013). Improving rainfall nowcasting and urban runoff forecasting through dynamic radar-rain gauge rainfall adjustment. In *7th International Conference on Sewer Processes & Networks*. Sheffield, United Kingdom. <http://spiral.imperial.ac.uk/handle/10044/1/14662>.
- OCHOA-RODRIGUEZ, S., L. P. WANG, P. WILLEMS & C. ONOF (2019). A review of radar-rain gauge data merging methods and their potential for urban hydrological applications. *Water Resources Research*, 55(8), 6356–6391. doi: 10.1029/2018WR023332.
- OLSEN, R. L., D. V. ROGERS & D. B. HODGE (1978). The aR^b relation in the calculation of rain attenuation. *IEEE Transactions on Antennas and Propagation*, 26(2), 318–329. doi: 10.1109/TAP.1978.1141845.
- ORLANSKI, I. (1975). A rational subdivision of scales for atmospheric processes. *Bulletin of the American Meteorological Society*, 56(5), 527–530.
- VAN OSNABRUGGE, B., R. UIJLENHOET & A. WEERTS (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. doi: 10.5194/hess-23-1453-2019.
- OVEREEM, A., T. A. BUISSHAND & I. HOLLEMAN (2009a). Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar. *Water Resources Research*, 45(10), W10424. doi: 10.1029/2009WR007869.
- OVEREEM, A., I. HOLLEMAN & A. BUISSHAND (2009b). Derivation of a 10-year radar-based climatology of rainfall. *Journal of Applied Meteorology and Climatology*, 48(7), 1448–1463. doi: 10.1175/2009JAMC1954.1.
- OVEREEM, A., H. LEIJNSE & R. UIJLENHOET (2011). Measuring urban rainfall using microwave links from commercial cellular communication networks. *Water Resources Research*, 47(12), W12505. doi: 10.1029/2010WR010350.
- OVEREEM, A., H. LEIJNSE & R. UIJLENHOET (2013). Country-wide rainfall maps from cellular communication networks. *Proceedings of the National Academy of Sciences*, 110(8), 2741–2745. doi: 10.1073/pnas.1217961110.
- OVEREEM, A., H. LEIJNSE & R. UIJLENHOET (2016a). Retrieval algorithm for rainfall mapping from microwave links in a cellular communication network. *Atmospheric Measurement Techniques*, 9(5), 2425–2444. doi: 10.5194/amt-9-2425-2016.
- OVEREEM, A., H. LEIJNSE & R. UIJLENHOET (2016b). Two and a half years of country-wide rainfall maps using radio links from commercial cellular telecommunication networks. *Water Resources Research*, 52(10), 8039–8065. doi: 10.1002/2016WR019412.
- PANICONI, C. & M. PUTTI (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. *Water Resources Research*, 51(9), 7090–7129. doi: 10.1002/2015WR017780.
- PAPPENBERGER, F., H. L. CLOKE, D. J. PARKER, F. WETTERHALL, D. S. RICHARDSON & J. THIELEN (2015). The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, 51, 278–291. doi: 10.1016/j.envsci.2015.04.016.
- PARK, S., M. BERENGUER & D. SEMPERE-TORRES (2019). Long-term analysis of gauge-adjusted radar rainfall accumulations at European scale. *Journal of Hydrology*, 573, 768–777. doi: 10.1016/j.jhydrol.2019.03.093.
- PEGRAM, G. G. & A. N. CLOTHIER (2001). Downscaling rainfields in space and time, using the string of beads model in time series mode. *Hydrology and Earth System Sciences Discussions*, 5(2), 175–186. doi: 10.5194/hess-5-175-2001.
- PIELKE, R. A. & M. W. DOWNTON (2000). Precipitation and damaging floods: Trends in the United States, 1932–97. *Journal of Climate*, 13(20), 3625–3637. doi: 10.1175/1520-0442(2000)013<3625:PADFTI>2.0.CO;2.

- PIERCE, C., N. BOWLER, A. SEED, A. JONES, D. JONES & R. MOORE (2005). Use of a stochastic precipitation nowcast scheme for fluvial flood forecasting and warning. *Atmospheric Science Letters*, 6(1), 78–83. DOI: 10.1002/asl.102.
- PIERCE, C., A. SEED, S. BALLARD, D. SIMONIN & Z. LI (2012). Nowcasting. In J. Bech (editor), *Doppler radar observations - Weather radar, wind profiler, ionospheric radar, and other advanced applications*. InTech. DOI: 10.5772/39054.
- POINCARÉ, H. (1952). *Science and Method*. Dover Publications, Mineola, NY, United States of America.
- POLETTI, M. L., F. SILVESTRO, S. DAVOLIO, F. PIGNONE & N. REBORA (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. DOI: 10.5194/hess-23-3823-2019.
- PRAKASH, S., A. K. MITRA, I. M. MOMIN, E. N. RAJAGOPAL, S. F. MILTON & G. M. MARTIN (2016). Skill of short- to medium-range monsoon rainfall forecasts from two global models over India for hydro-meteorological applications. *Meteorological Applications*, 23(4), 574–586. DOI: 10.1002/met.1579.
- PRINSEN, G., H. HAKVOORT & R. DAHM (2010). Neerslag-afvoermodellering met SOBEK-RR. *Stromingen*, 15(4), 8–24.
- PULKINEN, S., V. CHANDRASEKAR & A. M. HARRI (2018). Nowcasting of precipitation in the high-resolution Dallas–Fort Worth (DFW) urban radar remote sensing network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8), 2773–2787. DOI: 10.1109/JSTARS.2018.2840491.
- PULKINEN, S., V. CHANDRASEKAR, A. VON LERBER & A. M. HARRI (2020). Nowcasting of convective rainfall using volumetric radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–15. DOI: 10.1109/TGRS.2020.2984594.
- PULKINEN, S., V. CHANDRASEKAR & T. NIEMI (2021). Lagrangian integro-difference equation model for precipitation nowcasting. *Journal of Atmospheric and Oceanic Technology*, 38(12), 2125–2145. DOI: 10.1175/JTECH-D-21-0013.1.
- PULKINEN, S., D. NERINI, A. A. PÉREZ HORTAL, C. VELASCO-FORERO, A. SEED, U. GERMANN & L. FORESTI (2019). Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12(10), 4185–4219. DOI: 10.5194/gmd-12-4185-2019.
- QI, Y., J. ZHANG, P. ZHANG & Q. CAO (2013). VPR correction of bright band effects in radar QPEs using polarimetric radar observations. *Journal of Geophysical Research: Atmospheres*, 118(9), 3627–3633. DOI: 10.1002/jgrd.50364.
- RADCHENKO, P., A. L. VASNEV & W. WANG (2021). Too similar to combine? On negative weights in forecast combination. *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2021.08.002.
- RADHAKRISHNAN, C. & V. CHANDRASEKAR (2020). CASA prediction system over Dallas–Fort Worth urban network: Blending of nowcasting and high-resolution numerical weather prediction model. *Journal of Atmospheric and Oceanic Technology*, 37(2), 211–228. DOI: 10.1175/JTECH-D-18-0192.1.
- RAVURI, S., K. LENC, M. WILLSON, D. KANGIN, R. LAM, P. MIROWSKI, M. FITZSIMONS, M. ATHANASSIADOU, S. KASHEM, S. MADGE, R. PRUDDEN, A. MANDHANE, A. CLARK, A. BROCK, K. SIMONYAN, R. HADSELL, N. ROBINSON, E. CLANCY, A. ARRIBAS & S. MOHAMED (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878), 672–677. DOI: 10.1038/s41586-021-03854-z.
- RAYITSFELD, A., R. SAMUELS, A. ZINEVICH, U. HADAR & P. ALPERT (2012). Comparison of two methodologies for long term rainfall monitoring using a commercial microwave communication system. *Atmospheric research*, 104, 119–127. DOI: 10.1016/j.atmosres.2011.08.011.
- RICHARDSON, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563), 649–667. DOI: 10.1002/qj.49712656313.
- RICHARDSON, L. F. (1922). *Weather prediction by numerical process*. Cambridge University Press, Cambridge, United Kingdom.
- RIOS GAONA, M. F., A. OVEEREEM, H. LEIJNSE & R. UIJLENHOET (2015). Measurement and interpolation uncertainties in rainfall maps from cellular communication networks. *Hydrology and Earth System Sciences*, 19(8), 3571–3584. DOI: 10.5194/hess-19-3571-2015.
- RIOS GAONA, M. F., A. OVEEREEM, T. H. RAUPACH, H. LEIJNSE & R. UIJLENHOET (2018). Rainfall retrieval with commercial microwave links in São Paulo, Brazil. *Atmospheric Measurement Techniques*, 11(7), 4465–4476. DOI: 10.5194/amt-11-4465-2018.
- ROBERTS, N. M. & H. W. LEAN (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1), 78–97. DOI: 10.1175/2007MWR2123.1.
- ROGERS, R. F., J. M. FRITSCH & W. C. LAMBERT (2000). A simple technique for using radar data in the dynamic initialization of a mesoscale model. *Monthly Weather Review*, 128(7), 2560–2574. DOI: 10.1175/1520-0493(2000)128<2560:ASTFUR>2.0.CO;2.

- DE ROOY, W. C., P. SIEBESMA, P. BAAS, G. LENDERINK, S. R. DE ROODE, H. DE VRIES, E. VAN MEIJGAARD, J. F. MEIRINK, S. TIJM & B. VAN 'T VEEN (2022). Model development in practice: a comprehensive update to the boundary layer schemes in HARMONIE-AROME cycle 40. *Geoscientific Model Development*, 15(4), 1513–1543. DOI: 10.5194/gmd-15-1513-2022.
- SALTIKOFF, E., K. FRIEDRICH, J. SODERHOLM, K. LENGFELD, B. NELSON, A. BECKER, R. HOLLMANN, B. URBAN, M. HEISTERMANN & C. TASSONE (2019). An overview of using weather radar for climatological studies: successes, challenges, and potential. *Bulletin of the American Meteorological Society*, 100(9), 1739–1752. DOI: 10.1175/BAMS-D-18-0166.1.
- SAMANIEGO, L., R. KUMAR & S. ATTINGER (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5). DOI: 10.1029/2008WR007327.
- SAMPSON, C. C., T. J. FEWTRELL, F. O'LOUGHLIN, F. PAPPENBERGER, P. B. BATES, J. E. FREER & H. L. CLOKE (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. DOI: 10.5194/hess-18-2305-2014.
- SCHAEFER, J. T. (1990). The Critical Success Index as an indicator of warning skill. *Weather and Forecasting*, 5(4), 570–575. DOI: 10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.
- SCHERTZER, D. & S. LOVEJOY (1987). Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *Journal of Geophysical Research: Atmospheres*, 92(D8), 9693–9714. DOI: 10.1029/JD092iD08p09693.
- SCHLEISS, M., J. OLSSON, P. BERG, T. NIEMI, T. KOKKONEN, S. THORND AHL, R. NIELSEN, J. ELLERBÆK NIELSEN, D. BOZHINOVA & S. PULKKINEN (2020). The accuracy of weather radar in heavy rain: a comparative study for Denmark, the Netherlands, Finland and Sweden. *Hydrology and Earth System Sciences*, 24(6), 3157–3188. DOI: 10.5194/hess-24-3157-2020.
- SCHUURMANS, J. M., M. F. P. BIERKENS, E. J. PEBESMA & R. UIJLENHOET (2007). Automatic prediction of high-resolution daily rainfall fields for multiple extents: The potential of operational radar. *Journal of Hydrometeorology*, 8(6), 1204–1224. DOI: 10.1175/2007JHM792.1.
- SCHWEPPE, R., S. THOBER, S. MÜLLER, M. KELBLING, R. KUMAR, S. ATTINGER & L. SAMANIEGO (2022). MPR 1.0: a stand-alone multiscale parameter regionalization tool for improved parameter estimation of land surface models. *Geoscientific Model Development*, 15(2), 859–882. DOI: 10.5194/gmd-15-859-2022.
- SEED, A. W. (2003). A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology*, 42(3), 381–388. DOI: 10.1175/1520-0450(2003)042<0381:ADASSA>2.0.CO;2.
- SEED, A. W., C. E. PIERCE & K. NORMAN (2013). Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resources Research*, 49(10), 6624–6641. DOI: 10.1002/wrcr.20536.
- SEED, A. W., R. SRIKANTHAN & M. MENABDE (1999). A space and time model for design storm rainfall. *Journal of Geophysical Research: Atmospheres*, 104(D24), 31623–31630. DOI: 10.1029/1999JD900767.
- SEO, D. J., J. BREIDENBACH, R. FULTON, D. MILLER & T. O'BANNON (2000). Real-time adjustment of range-dependent biases in WSR-88D rainfall estimates due to nonuniform vertical profile of reflectivity. *Journal of Hydrometeorology*, 1(3), 222–240. DOI: 10.1175/1525-7541(2000)001<0222:RTAORD>2.0.CO;2.
- SEO, D. J., J. P. BREIDENBACH & E. R. JOHNSON (1999). Real-time estimation of mean field bias in radar rainfall data. *Journal of Hydrology*, 223(3), 131–147. DOI: 10.1016/S0022-1694(99)00106-7.
- SERAFIN, R. J. & J. W. WILSON (2000). Operational weather radar in the United States: Progress and opportunity. *Bulletin of the American Meteorological Society*, 81(3), 501–518. DOI: 10.1175/1520-0477(2000)081<0501:OWRITU>2.3.CO;2.
- SHARIF, H. O., F. L. OGDEN, W. F. KRAJEWSKI & M. XUE (2002). Numerical simulations of radar rainfall error propagation. *Water Resources Research*, 38(8), 15–1–15–14. DOI: 10.1029/2001WR000525.
- SHARIF, H. O., D. YATES, R. ROBERTS & C. MUELLER (2006). The use of an automated nowcasting system to forecast flash floods in an urban watershed. *Journal of Hydrometeorology*, 7(1), 190–202. DOI: 10.1175/JHM482.1.
- SHEHU, B. & U. HABERLANDT (2022). Improving radar-based rainfall nowcasting by a nearest-neighbour approach – Part 1: Storm characteristics. *Hydrology and Earth System Sciences*, 26(6), 1631–1658. DOI: 10.5194/hess-26-1631-2022.
- SHI, J. & C. TOMASI (1994). Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600. IEEE, Seattle, WA, USA. DOI: 10.1109/CVPR.1994.323794.
- SHI, X., Z. GAO, L. LAUSEN, H. WANG, D. Y. YEUNG, W. K. WONG & W. C. WOO (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- SIDERIS, I. V., M. GABELLA, R. ERDIN & U. GERMANN (2014). Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society*,

- 140(680), 1097–1111. DOI: 10.1002/qj.2188.
- SIMMONS, A. J. & A. HOLLINGSWORTH (2002). Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 128(580), 647–677. DOI: 10.1256/003590002321042135.
- SIVAPALAN, M., G. BLÖSCHL, L. ZHANG & R. VERTESSY (2003). Downward approach to hydrological prediction. *Hydrological processes*, 17(11), 2101–2111. DOI: 10.1002/hyp.1425.
- SMITH, J. A. & W. F. KRAJEWSKI (1991). Estimation of the mean field bias of radar rainfall estimates. *Journal of Applied Meteorology and Climatology*, 30(4), 397–412. DOI: 10.1175/1520-0450(1991)030<0397:EOTMFB>2.0.CO;2.
- SOKOL, Z., J. MEJSNAR, L. POP & V. BLIŽŇÁK (2017). Probabilistic precipitation nowcasting based on an extrapolation of radar reflectivity and an ensemble approach. *Atmospheric research*, 194, 245–257. DOI: 10.1016/j.atmosres.2017.05.003.
- SPRENGER, M., S. SCHEMM, R. OECHSLIN & J. JENKNER (2017). Nowcasting Foehn wind events using the AdaBoost machine learning algorithm. *Weather and Forecasting*, 32(3), 1079–1099. DOI: 10.1175/WAF-D-16-0208.1.
- STELLING, G. S. & S. P. A. DUINMEIJER (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. DOI: 10.1002/flid.537.
- STELLING, G. S. & A. VERWEY (2006). Numerical flood simulation. In *Encyclopedia of Hydrological Sciences. Part 2: Hydroinformatics*. John Wiley & Sons, Ltd. DOI: 10.1002/0470848944.hsa025a.
- SUBBIAH, A. R., L. BILDAN & R. NARASIMHAN (2008). *Background paper on assessment of the economics of early warning systems for disaster risk reduction*. The World Bank, Washington D.C., United States of America. https://ral.ucar.edu/hopson/Verkade/Economics/Subbiah_EWS.pdf.
- SUN, J., M. XUE, J. W. WILSON, I. ZAWADZKI, S. P. BALLARD, J. ONVLEE-HOOIMEYER, P. JOE, D. M. BARKER, P. W. LI, B. GOLDING, M. XU & J. PINTO (2014). Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95(3), 409–426. DOI: 10.1175/BAMS-D-11-00263.1.
- SUN, Y., W. BAO, K. VALK, C. C. BRAUER, J. SUMIHAR & A. H. WEERTS (2020). Improving forecast skill of lowland hydrological models using ensemble kalman filter and unscented kalman filter. *Water Resources Research*, 56(8), e2020WR027468. DOI: 10.1029/2020WR027468.
- TANG, Y., H. W. LEAN & J. BORNEMANN (2013). The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteorological Applications*, 20(4), 417–426. DOI: 10.1002/met.1300.
- TAUB, L. (2003). *Ancient Meteorology*. Routledge, London, United Kingdom. DOI: 10.4324/9780203634288.
- TEL4RAIN (2020). *About TEL4RAIN*. Czech Technical University, Prague, Czech Republic. <http://www.tel4rain.cz/tel4rain-about-eng.php>.
- TERMONIA, P., C. FISCHER, E. BAZILE, F. BOUYSSSEL, R. BROŽKOVÁ, P. BÉNARD, B. BOCHENEK, D. DEGRAUWE, M. DERKOVÁ, R. EL KHATIB, R. HAMD, J. MAŠEK, P. POTTIER, N. PRISTOV, Y. SEITY, P. SMOLÍKOVÁ, O. ŠPANIEL, M. TUDOR, Y. WANG, C. WITTMANN & A. JOLY (2018). The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geoscientific Model Development*, 11(1), 257–281. DOI: 10.5194/gmd-11-257-2018.
- THORNDahl, S., T. EINFALT, P. WILLEMS, J. E. NIELSEN, M. C. TEN VELDHUIS, K. ARNBjERG-NIELSEN, M. R. RASMUSSEN & P. MOLNAR (2017). Weather radar rainfall data in urban hydrology. *Hydrology and Earth System Sciences*, 21(3), 1359–1380. DOI: 10.5194/hess-21-1359-2017.
- THORNDahl, S., J. E. NIELSEN & M. R. RASMUSSEN (2014). Bias adjustment and advection interpolation of long-term high resolution radar rainfall series. *Journal of Hydrology*, 508, 214–226. DOI: 10.1016/j.jhydrol.2013.10.056.
- THORNDahl, S., T. S. POULSEN, T. BØVITH, M. BORUP, M. AHM, J. E. NIELSEN, M. GRUM, M. R. RASMUSSEN, R. GILL & P. S. MIKKELSEN (2013). Comparison of short-term rainfall forecasts for model-based flow prediction in urban drainage systems. *Water Science and Technology*, 68(2), 472–478. DOI: 10.2166/wst.2013.274.
- TODINI, E. (2001). A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences*, 5(2), 187–199. DOI: 10.5194/hess-5-187-2001.
- TURNER, B. J., I. ZAWADZKI & U. GERMANN (2004). Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE). *Journal of Applied Meteorology*, 43(2), 231–248. DOI: 10.1175/1520-0450(2004)043<0231:POPFRC>2.0.CO;2.
- TURNER, D. D., H. CUTLER, M. SHIELDS, R. HILL, B. HARTMAN, Y. HU, T. LU & H. JEON (2022). Evaluating the economic impacts of improvements to the high-resolution rapid refresh (HRRR) numerical weather prediction model. *Bulletin of the American Meteorological Society*, 103(2), E198–E211. DOI: 10.1175/BAMS-D-20-0099.1.
- TVERSKY, A. & D. KAHNEMAN (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

DOI: 10.1126/science.185.4157.1124.

- UIJLENHOET, R. & A. BERNE (2008). Stochastic simulation experiment to assess radar rainfall retrieval uncertainties associated with attenuation and its correction. *Hydrology and Earth System Sciences*, 12(2), 587–601. DOI: 10.5194/hess-12-587-2008.
- UIJLENHOET, R., A. OVEREEM & H. LEIJNSE (2018). Opportunistic remote sensing of rainfall using microwave links from cellular communication networks. *WIREs Water*, 5(4), e1289. DOI: 10.1002/wat2.1289.
- UKKONEN, P., A. MANZATO & A. MÄKELÄ (2017). Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *Journal of Applied Meteorology and Climatology*, 56(8), 2335–2352. DOI: 10.1175/JAMC-D-16-0361.1.
- UNISDR (2002). *Guidelines for reducing flood losses*. Technical report, United Nations International Strategy for Disaster Reduction (UNISDR), New York, NY. <https://www.undrr.org/publication/guidelines-reducing-flood-losses>.
- VANNITSEM, S., J. B. BREMNES, J. DEMAEYER, G. R. EVANS, J. FLOWERDEW, S. HEMRI, S. LERCH, N. ROBERTS, S. THEIS, A. ATENCIA, Z. B. BOUALLÈGUE, J. BHEND, M. DABERNIG, L. DE CRUZ, L. HIETA, O. MESTRE, L. MORET, I. O. PLENKOVIĆ, M. SCHMEITS, M. TAILLARDAT, J. VAN DEN BERGH, B. VAN SCHAEYBROECK, K. WHAN & J. YLHAISI (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3), E681–E699. DOI: 10.1175/BAMS-D-19-0308.1.
- VENUGOPAL, V., E. FOUFOULA-GEORGIU & V. SAPOZHNIKOV (1999). Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysical Research: Atmospheres*, 104(D24), 31599–31610. DOI: 10.1029/1999JD900437.
- VERKADE, J. S. & M. G. F. WERNER (2011). Estimating the benefits of single value and probability forecasting for flood warning. *Hydrology and Earth System Sciences*, 15(12), 3751–3765. DOI: 10.5194/hess-15-3751-2011.
- VIVONI, E. R., D. ENTEKHABI, R. L. BRAS, V. Y. IVANOV, M. P. VAN HORNE, C. GRASSOTTI & R. N. HOFFMAN (2006). Extending the predictability of hydrometeorological flood events using radar rainfall nowcasting. *Journal of Hydrometeorology*, 7(4), 660–677. DOI: 10.1175/JHM514.1.
- VIVONI, E. R., D. ENTEKHABI & R. N. HOFFMAN (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. DOI: 10.1175/JAM2506.1.
- VOS, L. W. D., H. LEIJNSE, A. OVEREEM & R. UIJLENHOET (2019). Quality control for crowdsourced personal weather stations to enable operational rainfall monitoring. *Geophysical Research Letters*, 46(15), 8820–8829. DOI: 10.1029/2019GL083731.
- WACKERNAGEL, H. (2003). *Multivariate geostatistics: An introduction with applications*. Springer, Berlin Heidelberg, Germany, 3 edition. DOI: 10.1007/978-3-662-05294-5.
- WAGNER, A., J. SELTMANN & H. KUNSTMANN (2012). Joint statistical correction of clutters, spokes and beam height for a radar derived precipitation climatology in southern Germany. *Hydrology and Earth System Sciences*, 16(11), 4101–4117. DOI: 10.5194/hess-16-4101-2012.
- WARD, P. J., B. JONGMAN, F. S. WEILAND, A. BOUWMAN, R. VAN BEEK, M. F. P. BIERKENS, W. LIGTVOET & H. C. WINSEMIUS (2013). Assessing flood risk at the global scale: model setup, results, and sensitivity. *Environmental Research Letters*, 8(4), 044019. DOI: 10.1088/1748-9326/8/4/044019.
- WASKO, C. & D. GUO (2021). *Package hydroEvents*. <https://cran.rstudio.com/web/packages/hydroEvents/index.html>.
- WERNER, M., M. CRANSTON, T. HARRISON, D. WHITFIELD & J. SCHELLEKENS (2009). Recent developments in operational flood forecasting in England, Wales and Scotland. *Meteorological Applications*, 16(1), 13–22. DOI: 10.1002/met.124.
- WERNER, M., J. SCHELLEKENS, P. GIJSBERS, M. VAN DIJK, O. VAN DEN AKKER & K. HEYNERT (2013). The Delft-FEWS flow forecasting system. *Environmental Modelling & Software*, 40, 65–77. DOI: 10.1016/j.envsoft.2012.07.010.
- WILLEMS, P. & J. OLSSON (2012). *Impacts of climate change on rainfall extremes and urban drainage systems*. IWA Publishing, London, United Kingdom. <https://library.oapen.org/handle/20.500.12657/25812>.
- WILLEMS, P. & M. VRAC (2011). Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. *Journal of Hydrology*, 402(3), 193–205. DOI: 10.1016/j.jhydrol.2011.02.030.
- WILSON, J. W., Y. FENG, M. CHEN & R. D. ROBERTS (2010). Nowcasting challenges during the Beijing Olympics: Successes, failures, and implications for future nowcasting systems. *Weather and Forecasting*, 25(6), 1691–1714. DOI: 10.1175/2010WAF2222417.1.
- WMO (2020). *WMO Radar Database*. World Meteorological Organization, Geneva, Switzerland. <https://wrd.mgm.gov.tr/Home/Wrd>.
- WONG, W. K., V. T. L. CHENG & W. C. WOO (2016). Community SWIRLS Nowcasting System (Com-SWIRLS). In *WMO WWRP 4th International Symposium on Nowcasting and Very-short-range Forecast 2016 (WSN2016)*. Hong Kong.

- YOON, S.S. (2019). Adaptive blending method of radar-based and numerical weather prediction QPFs for urban flood forecasting. *Remote Sensing*, 11(6), 642. DOI: 10.3390/rs11060642.
- ZAPPA, M., S. JAUN, U. GERMANN, A. WALSER & F. FUNDEL (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, 100(2), 246–262. DOI: 10.1016/j.atmosres.2010.12.005.
- ZAWADZKI, I. (2018). Republication of Zawadzki, I. (1984): Factors affecting the precision of radar measurements of rain. In L. de Vos, H. Leijnse, & R. Uijlenhoet (editors), *10th European Conference on Radar in Meteorology and Hydrology (ERAD 2018) : 1-6 July 2018, Ede-Wageningen, The Netherlands*, pages 960–965. Wageningen University & Research, Wageningen, the Netherlands. DOI: 10.18174/454537.
- ZAWADZKI, I.I. (1973). Statistical properties of precipitation patterns. *Journal of Applied Meteorology*, 12(3), 459–472. DOI: 10.1175/1520-0450(1973)012<0459:SPOPP>2.0.CO;2.
- ZHUANG, F., Z. QI, K. DUAN, D. XI, Y. ZHU, H. ZHU, H. XIONG & Q. HE (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76. DOI: 10.1109/JPROC.2020.3004555.
- ZINEVICH, A., H. MESSER & P. ALPERT (2009). Frontal rainfall observation by a commercial microwave communication network. *Journal of Applied Meteorology and Climatology*, 48(7), 1317–1334. DOI: 10.1175/2008JAMC2014.1.
- ZOU, X., Q. DAI, K. WU, Q. YANG & S. ZHANG (2020). An empirical ensemble rainfall nowcasting model using multi-scaled analogues. *Natural Hazards*. DOI: 10.1007/s11069-020-03964-3.

Statement of authorship contribution

THE general research direction was proposed by my promotors and co-promotor, but was also influenced by a clear need from the practitioners (predominantly the water managers in the Netherlands). Based on these ideas, I formulated the research plan and specific research questions. I wrote all text in Chapters 1 and 8, with minor suggestions from my (co-)promotors. Chapter 2 combines the study area, meteorological data information, hydrological models and overarching methods from Chapters 3–7 and therefore has the same contributions as the below listed contributions to Chapters 3–7. An overview of the contribution from (co-)promotors and others co-authors for the other chapters is listed below. Names have been abbreviated in the following way:

RI = Ruben Imhoff (Deltares/WUR)

AO = Aart Overeem (KNMI)

CB = Claudia Brauer (WUR)

DN = Daniele Nerini (MeteoSwiss)

KH = Klaas-Jan van Heeringen (Deltares)

MG = Michiel Van Ginderachter (RMI)

WD = Wout Dewettinck (Ghent Univ.)

AW = Albrecht Weerts (Deltares/WUR)

CV = Carlos Velasco-Forero (BoM)

HL = Hidde Leijnse (KNMI)

LD = Lesley De Cruz (RMI)

RU = Remko Uijlenhoet (TUD)

Chapter 3

Conceptualization of the work:

Data collection:

Modelling, data analysis and interpretation:

Drafting of manuscript:

Revision of the article and approval for publishing:

AW, CB, KH, RI, RU

AO, CB, KH, RI

RI in consultation with
AW, AO, CB, HL, KH, RU

RI

AW, AO, CB, HL, KH, RU

Chapter 4

Conceptualization of the work:

Data collection:

Modelling, data analysis and interpretation:

Drafting of manuscript:

Revision of the article and approval for publishing:

AW, CB, RI, RU

AO, CB, RI

RI in consultation with
AW, AO, CB, RU

RI

AW, AO, CB, RU

Chapter 5

Conceptualization of the work:

Data collection:

Modelling, data analysis and interpretation:

Drafting of manuscript:

Revision of the article and approval for publishing:

AO, AW, CB, HL, RI, RU

AO, RI

RI in consultation with
AO, AW, CB, HL, RU

RI

AO, AW, CB, HL, RU


Chapter 6

Conceptualization of the work:	AW, CB, LC, KH, RI, RU
Data collection:	AW, LC, RI
Model development:	LC, RI, WD in consultation with AW, CB, CV, DN, KH, MG, RU
Modelling, data analysis and interpretation:	RI in consultation with AW, CB, KH, LC, RU
Drafting of manuscript:	RI
Revision of the article and approval for publishing: (in progress)	AW, CB, CV, DN, KH, LC, MG, RU, WD

Chapter 7

Conceptualization of the work:	AW, CB, KH, RI, RU
Data collection:	CB, KH, RI
Modelling, data analysis and interpretation:	RI in consultation with AW, CB, KH, RU
Drafting of manuscript:	RI
Revision of the article and approval for publishing:	AW, CB, KH, RU

Summary

HORT-term forecasting of rainfall is crucial for a plethora of applications, among which flood forecasting. Yet, rainfall forecasting is challenging, even for short forecast horizons, due to the high spatial and temporal variability of rainfall (**Chapter 1**). A fast, observation-based forecasting technique, such as nowcasting, has potential to facilitate better flood forecasts through improved short-term rainfall forecasts. This thesis aims to *identify if and how operational flood forecasting can be improved with (radar) rainfall nowcasting-based techniques*. To do so, we have focused on several scientific questions (mentioned in **Chapter 1**) throughout this thesis and we have analysed, tested and improved parts of the short-term flood forecasting chain in **Chapters 3–7**. **Chapter 2** introduced the study areas, data and methods that have been used throughout this thesis.

In **Chapter 3**, a simple, operationally applicable, climatological bias-adjustment method was introduced to cope with the significant biases in real-time radar quantitative precipitation estimations (QPE). The method, CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting), consists of a set of gridded climatological adjustment factors for every day of the year, as derived with historical radar data. The method was tested on the resulting rainfall estimates and discharge simulations for twelve Dutch catchments and polders. The operational mean field bias (MFB) adjusted rainfall estimates acted as benchmark. When focusing on country-average rainfall sums, the MFB-adjusted QPE outperforms the CARROTS-corrected QPE. This perspective changes on the catchment scale, where annual rainfall sums from CARROTS outperform the MFB-adjusted rainfall estimates with increasing distance from the radars, because the MFB-adjusted QPE significantly underestimates the rainfall amounts further from the radars. The effects of any biases in the QPE are amplified when used in rainfall-runoff models to simulate discharge. Discharge simulations using the CARROTS QPE outperform those with the MFB-adjusted product for catchments. The presented method can provide an easy-to-apply real-time correction factor and can be a benchmark for future bias-adjustment method development.

In **Chapter 4**, a large-sample analysis of 1,533 systematically selected events was performed to construct nowcasts and determine the predictive skill of nowcasting for short-term rainfall forecasting at the catchment scale in the Netherlands. Four algorithms were tested and compared with Eulerian Persistence: rainymotion Sparse, rainymotion DenseRotation, pysteps deterministic and pysteps probabilistic with 20 ensemble members. The maximum skillful lead times of the tested nowcasting methods increase for longer event durations, mainly due to the more persistent character of longer-duration events. For all four event durations, pysteps deterministic attains the longest average decorrelation times, with 25 min for 1-h durations, 40 min for 3 h, 56 min for 6 h and 116 min for 24 h. Mean absolute errors are, on average, three times lower during winter (predominantly persistent stratiform precipitation) than during summer (more convective events). Higher skill is found after spatially upscaling the forecast to coarser resolutions and up to two times higher skillful lead times are found downwind than upwind towards the edge of the radar domain. This indicates that both catchment size and location matter for the nowcasting skill. Most errors originate from growth and dissipation processes, which are not or only partially (stochastically) accounted for. Therefore, the largest improvements can still be attained in this direction.

In **Chapter 5**, an alternative QPE source was tested for nowcasting. Probabilistic nowcasts were constructed with country-wide rainfall maps estimated from signal level data of commercial microwave links (CMLs) for twelve summer days in the Netherlands. Radar rainfall nowcasts are found to outperform the CML nowcasts for low rainfall intensities, due to a more coherent advection field and more detailed rainfall structures in the input QPE. This reverses for forecasting higher-intensity rainfall cells, due to the underestimations in the operational radar QPE, which makes the CML QPE more suitable for real-time nowcasting of high-intensity rainfall events. Two limitations are identified for nowcasting with CML data. First, regionally low CML densities lead to the need to upscale the data and forecasts from the tested 1-km² resolution to at least 11×11 km². Second, the rainfall advection derivation in the nowcasting approach is sensitive to areas with sparse or no data availability, which advocates for incorporating other data sources, such as satellites and personal weather stations, in an operational application. Hence, there is potential for rainfall nowcasting with CML data, for example in regions where weather radars are absent.

In **Chapter 6**, an adaptive scale-dependent ensemble blending method, based on the STEPS scheme, was implemented in the open-source pysteps library, to extend the skillful lead times beyond the average of 2 h. In this implementation, the extrapolation (ensemble) nowcast, (ensemble) NWP and noise components were blended with skill-dependent weights that vary per spatial scale level. The method was evaluated on both the radar domain and catchment scale using three events in 2021, including the devastating rain and flood event striking Belgium, Germany and the Netherlands in mid-July. A 48-member nowcast, a deterministic NWP model run and a 48-member linear blending approach acted as benchmark. Both at the radar domain and at the catchment scale, the introduced blending approach generally performs similarly or better (in terms of event-averaged CRPS and CSI values) than the other three tested methods, although the difference, particularly with the linear blending method, reduces when focusing on forecast catchment-average cumulative rainfall sums instead of instantaneous rainfall rates. By properly combining observations and NWP forecasts, blending methods such as these can extend the skillfulness of rainfall forecasts beyond that of the individual components.

In **Chapter 7**, the skill of radar rainfall nowcasting for (flash) flood forecasting was assessed across the Netherlands. For this purpose, the nowcasts from the large-sample analysis of Chapter 4 were used to construct discharge forecasts with the operationally-used hydrological models for the same twelve Dutch catchments. The results indicate that rainfall and discharge forecast errors increase with both increasing rainfall intensity and spatial variability. Although this increase is linear for the rainfall forecast errors, the relationship depends on the initial conditions for the discharge forecast errors. Discharge forecast errors increase more quickly with rainfall intensity when groundwater tables are shallow than for deeper groundwater tables. Similar to the rainfall forecast skill in Chapter 4, discharge forecasts with the more advanced nowcasting methods outperform the other tested methods. However, this does not always hold for threshold exceedance forecasts. Compared to no forecast at all, an exceedance is, on average, forecast between 119 and 223 min in advance. Especially pysteps performs worse than the other methods in forecasting threshold exceedances, due to a tendency to underestimate rainfall volumes. At the same time, pysteps counterbalances this by both a low false alarm ratio and more consistent forecasts than the other tested methods. All methods can advance short-term rainfall and discharge forecasting, though all have shortcomings. As forecast rainfall volumes are a crucial factor in discharge forecasting, a future focus on improving this aspect in

nowcasting is recommended.

This research has tested and shown the potential of rainfall nowcasting for operational (flash) flood forecasting. A nowcasting-based forecasting system can advance (flash) flood forecasting, especially when optimally combined with NWP forecasts. It is recommended to take the entire forecasting chain into account when moving towards a nowcasting-based operational flood forecasting system. This entails taking into account bias-adjusted rainfall estimates, the nowcasting or blended forecasting approach, the subsequent discharge forecast with hydrological models, and ultimately hydrological data assimilation to update the modelled system initial states, and the dissemination and translation into impact of results and inherent uncertainty in those results.



Samenvatting

KORTE-termijn neerslagverwachtingen zijn cruciaal voor een verscheidenheid aan toepassingen, waaronder overstromingsverwachtingen. Ondanks dit belang, is het een enorme uitdaging om accurate neerslagverwachtingen te maken, zelfs op de korte termijn, als gevolg van de grote variabiliteit in neerslag in zowel tijd als ruimte (**Hoofdstuk 1**). Een techniek die snel is en recente observaties kan gebruiken is nowcasting. Nowcasting heeft de potentie om betere overstromingsverwachtingen te faciliteren door middel van betere neerslagverwachtingen op de korte termijn. Dit proefschrift heeft als doel om *te identificeren of en hoe operationele hoogwaterverwachtingen kunnen worden verbeterd met regenvalnowcastingstechnieken*. Om dit te bereiken, is dit onderzoek toegespitst op enkele wetenschappelijke vragen (vermeld in Hoofdstuk 1) en zijn onderdelen van de korte-termijn overstromingsverwachtingsketen geanalyseerd, intensief getest en waar mogelijk verbeterd, zie Hoofdstukken 3–7. **Hoofdstuk 2** introduceert het studiegebied, data en methoden die door de thesis heen gebruikt zijn.

In **Hoofdstuk 3** wordt een simpele, operationeel toepasbare, klimatologische correctiemethode voor systematische fouten (bias) geïntroduceerd, waarmee gecorrigeerd kan worden voor systematische afwijkingen in het operationele radar-regenvalproduct (bestaande uit kwantitatieve neerslagschattingen, ‘QPE’ in het Engels). De methode, CARROTS (Climatology-based Adjustments for Radar Rainfall in an Operational Setting), is een gegridde, klimatologische QPE-correctie, die is afgeleid uit historische radardata, en kan zowel in de tijd (variërend per jaardag) als in de ruimte corrigeren. De methode is getest op de resulterende regenvalschattingen en daaropvolgende afvoersimulaties voor twaalf Nederlandse stroomgebieden en polders. In de validatiefase zijn deze schattingen vergeleken met de operationele mean field bias (MFB) correctiemethode. De MFB-gecorrigeerde QPE geeft betere resultaten dan CARROTS op nationale schaal voor domeingemiddelde neerslag. Echter, dit beeld verandert op de stroomgebiedsschaal, waar de jaarlijkse neerslagsommen van CARROTS dichter bij de referentie liggen dan de MFB-gecorrigeerde QPE, voornamelijk op grote afstand tot de radars. Dit komt doordat de MFB-gecorrigeerde QPE flink onderschat met toenemende afstand van de weerradars. Deze afwijkingen nemen verder toe wanneer de QPE-producten worden gebruikt in neerslag-afvoermodellen. Afvoersimulaties met de CARROTS-gecorrigeerde QPE liggen dichter bij de referentie dan de MFB-methode voor alle twaalf stroomgebieden. De gepresenteerde methode kan voorzien in de noodzaak om een makkelijk toepasbare, operationeel beschikbare correctiefactor te hebben en kan als referentie dienen voor toekomstige correctiemethode-ontwikkelingen.

In **Hoofdstuk 4** wordt de waarde van nowcasting voor korte-termijn neerslagverwachtingen getest met een grootschalige analyse, bestaande uit 1533 systematisch geselecteerde neerslaggebeurtenissen, op de stroomgebiedsschaal in Nederland. In deze analyse zijn nowcasts geproduceerd, en vergeleken met Eulerian Persistence, met vier algoritmes: rainymotion Sparse, rainymotion DenseRotation, pysteps deterministisch en pysteps probabilistisch met 20 ensemble members. Voor alle methoden neemt de maximaal bruikbare looptijd van de verwachting toe voor langere neerslaggebeurtenissen, wat voornamelijk veroorzaakt wordt door de voorspelbaarheid van de meer grootschalige neerslaggebieden voor langere duraties. Pysteps deterministisch bereikt voor alle geteste duren de langste decorrelatie-afstanden, een maat voor de maximaal bruikbare looptijd, met 25 min voor 1-uurs buien, 40 min voor 3 uur, 56 min voor 6 uur en 116 min voor 24 uur. De gemiddelde fout in de verwachtingen is drie keer zo laag in de winter

(met voornamelijk stratiforme neerslag) als in de zomer (voornamelijk convectieve neerslag). Daarnaast is de verwachtingswaarde hoger na het ruimtelijk aggregeren van de verwachting en zijn de de maximaal bruikbare looptijden twee keer hoger benedenwinds dan bovenwinds, wat aangeeft dat er een locatieafhankelijkheid is bij nowcasting. De meeste fouten komen voort uit groei- en uitdovingsprocessen van neerslag, die niet of enkel gedeeltelijk (stochastisch) worden meegenomen in de nowcastingmethodes. Verbeteringen in deze richting kunnen dus nog winst opleveren.

In **Hoofdstuk 5** wordt een alternatief QPE-product getest voor nowcasting. Ensemble nowcasts zijn gemaakt voor twaalf zomerdagen in Nederland met neerslagkaarten die voortkomen uit gemeten signaaldempingen (als gevolg van neerslag) in de radiostraalverbindingen van telecommunicatienetwerken (CMLs). Radarnowcasting presteert beter dan de CML-nowcasting voor lage neerslagintensiteiten, voornamelijk veroorzaakt door coherentere advectionvelden en gedetailleerdere neerslagvelden in de radar-QPE. Dit beeld draait echter om voor de verwachtingen van hoge neerslagintensiteiten, voornamelijk veroorzaakt door het onderschattende karakter van de operationele radar-QPE, wat de CML data geschikter maakt voor nowcasting van hoge neerslagintensiteiten. Twee beperkingen van CML-nowcasting zijn dat (1) een regionaal lage CML-dichtheid ervoor kan zorgen dat de data en verwachtingen geaggregeerd moeten worden naar grovere resoluties: van 1 km^2 naar ten minste $11 \times 11 \text{ km}^2$. En (2), de afleiding van de neerslagbewegingsrichting is erg gevoelig voor gebieden met weinig tot geen databeschikbaarheid, wat vraagt om het combineren van deze databron met andere bronnen, zoals satellietmetingen en persoonlijke weerstations, in een operationele context. Er is zeker potentie voor regenvalnowcasting met CML-data, vooral in regio's waar weerradars niet beschikbaar zijn.

In **Hoofdstuk 6** wordt een adaptieve, schaalafhankelijke ensemble-combinatiemethode, gebaseerd op STEPS, geïmplementeerd in de open-source pysteps-bibliotheek, om de maximaal bruikbare looptijd van de verwachtingen toe te laten nemen naar meer dan het gemiddelde van 2 uur. De toepassing combineert een (ensemble) extrapolatie nowcast met (ensemble) NWP en een perturbatiecomponent door gebruik te maken van gewichten die variëren per ruimtelijk schaalniveau. De methode is getest op zowel de radardomein- als stroomgebiedsschaal voor drie gebeurtenissen in 2021, waaronder de neerslag die de desastreuze overstromingen in België, Duitsland en Nederland in juli heeft veroorzaakt. Als referentie fungeerde een nowcast bestaande uit 48 realisaties, een deterministisch NWP-model en een lineaire blendingmethode met 48 realisaties. Zowel op de radardomein- als stroomgebiedsschaal weet de geïntroduceerde methode over het algemeen vergelijkbare of zelfs betere resultaten (gemeten naar gemiddelde CRPS en CSI-waarden) te behalen dan de andere drie methoden. Het verschil met, voornamelijk, de lineaire blindingmethode neemt wel af wanneer cumulatieve neerslagsommen worden getest in plaats van instantane neerslagintensiteiten. Een optimale combinatie van observaties en NWP-resultaten kan de maximaal bruikbare looptijd van de verwachtingen verder laten toenemen dan de afzonderlijke componenten.

In **Hoofdstuk 7** wordt de waarde van radar-regenvalnowcasting voor de verwachtingen van snelle afvoergolven getest. De nowcasts van Hoofdstuk 4 zijn hier gebruikt om afvoerverwachtingen te maken met de operationele hydrologische modellen van dezelfde twaalf Nederlandse stroomgebieden. De resultaten laten zien dat regenval- en afvoerverwachtingsfouten toenemen met zowel toenemende neerslagintensiteit als neerslagvariabiliteit. Alhoewel deze toename lineair is voor de neerslagverwachtingsfouten, hangt deze toename af van de initiële condities voor

de afvoerverwachtingsfouten, waarbij de fouten sneller toenemen voor natte initiële condities met hoge grondwaterstanden dan voor droge initiële condities. Vergelijkbaar met Hoofdstuk 4 zijn ook voor de afvoerverwachtingen de meer geavanceerde nowcastingmethoden vaak kwalitatief het beste. Dit is echter niet altijd het geval wanneer alleen naar overschrijdingskansen van drempelwaarden wordt gekeken. Vergeleken met een situatie zonder verwachtingssysteem, kan een drempelwaardeoverschreiding tussen de 119 en 223 min eerder worden verwacht met de geteste methoden. Voornamelijk pysteps presteert hierbij minder dan de andere methoden, als gevolg van de significante neerslagvolume-onderschattingen in pysteps. Tegelijkertijd weet pysteps wel het laagste aantal false alarms en de meest consistente verwachtingen te behalen. Kortom, korte-termijn regenval- en afvoerverwachtingen kunnen verbeterd met nowcasting, maar alle methoden hebben ook gebreken. Vooral vanuit een afvoerverwachtingsperspectief zijn volumes cruciaal en daarom zou een toekomstige focus moeten liggen op het verbeteren van dit aspect in nowcastigmethoden.

In dit proefschrift is de (toegevoegde) waarde van regenvalnowcasting voor operationele hoogwatervoorspellingen voor de korte termijn geanalyseerd. Een verwachtingssysteem kan verbeteren wanneer nowcasting gebruikt wordt, al helemaal wanneer de nowcasts optimaal gecombineerd worden met NWP-verwachtingen. Het is aan te raden om de gehele verwachtingsketen mee te nemen bij het implementeren van nowcasting in een overstromingsvoorspellingssysteem. Dit houdt dat de volgende aspecten moeten worden meegenomen: correcties voor de systematische fouten in de neerslagschattingen, de (gecombineerde) nowcastingmethode, de daaropvolgende afvoersimulaties met hydrologische modellen, en (idealiter) hydrologische data-assimilatie om de initiële condities van het systeem te updaten, de verspreiding en communicatie van de verwachte impact en daarmee gepaard gaande onzekerheden.

Acknowledgements

“Well the sky could be blue
Could be grey
Without you, I’m just miles away
Well the sky could be blue
I don’t mind
Without you, it’s a waste of time”

—Coldplay, *Strawberry Swing* (2008)



WELL done! You have made it to the acknowledgments and this means that you have read through most of this thesis, or did you skip all other chapters to directly end up here (yes, I know you!)? In any case, a PhD is a journey, which is only possible with the help of others. Therefore, I would like to take the opportunity to thank everyone who has made this journey possible.

A PhD trajectory can be a scary path and it is something you can hardly explain to a normal person (the scientific world, the fact that your work can be lonely and at the same time not, the discrepancy between having ample time to investigate something and at the same time being stressed along the way, or even as simple as explaining that you are both an employee and still studying). The realisation that you are going to deliver a book containing your research outcomes and ideas in four years from the start date of your PhD, is almost unimaginable. Yet, here you are with a thesis laid out in front of you. Was doing a PhD a pleasure or a burden? Well, as there is no love without pain, so is there no pleasure in a PhD without its burdens. Though, I have mainly experienced these four years as a pleasure. The past four years have taught me a lot, both on a professional and personal level, and it still is a strange idea to have an expertise on a topic now.

This thesis would not have been possible without the great help, advice, ideas, fun talks and occasionally pragmatic views of my advisors Albrecht, Claudia and Remko. It was wonderful to work with you and I think that throughout the whole thesis we have had the same end goal, which has really helped to get things done easily. Albrecht, already since my MSc thesis we are working together and I think (and hope) we will for the coming years after this thesis. I admire how you manage to combine a scientific interest with also a very practical view on matters; it really helps to focus more on innovations tailored to its users. Thanks for giving me a platform inside and outside Deltares, which has really helped to grow and start making a network. Claudia, thanks a lot for the time you manage to make for small things, such as designing a figure together, talks on a personal level (also with students you supervise!) and the thorough reviews you can give on all the manuscript versions I have written over the past four years. During meetings, you often manage to look at things from a somewhat different angle, which has really helped the overall process. Remko, I am always impressed how you manage to make time, also on a personal level, in your already completely filled agenda. That is really valuable! During meetings you always manage to ask a few in-depth (and sometimes difficult statistical) questions, which in the end steer things in the right direction. Thanks for facilitating presentations during conferences, opportunities to work with MSc thesis students at TU Delft and possible collaborations after the PhD.

Many chapters in this thesis would not have been possible without the help of a few other co-authors. Klaas-Jan, many thanks for your interest in nowcasting and its application for flood forecasting. You have paved the way to implement this work operationally by introducing the work (and myself) to Dutch water managers. It has been, and still is, fun to discuss on a more practical level what this work means for possible applications. Aart and Hidde, apart from the much appreciated occasional funny jokes, it has been a pleasure to work with you and to make the link with KNMI more explicit. I am sure we will stay in touch to keep working on radar and nowcasting-related issues! Lesley, in the middle of the COVID pandemic we came to the conclusion that we were about to work on a similar topic (blending nowcasting with NWP), so we made the great decision to do this together. My time in Brussels was a lot of fun (and it felt super far away after more than a year of not travelling) and I think we (Wout, you and me) can be proud of the work we managed to do in those few weeks. Finally, the pysteps framework was used in most chapters of this thesis and this open-source initiative has turned out to be an excellent network to be part of. Thanks for your help in discussing issues and new opportunities, and in particular thanks to Daniele Nerini for taking the lead in this network.

Without my colleagues, both at Deltares and in Wageningen, this journey would have been completely different. Thanks for all the inspiration you have given me, the fun talks, drinks, conferences and the team feeling. Although this thank you is for all colleagues, a special thanks to Jelte for four years of sharing an office in Wageningen (it was fun!), to Linda for our frequent "Regenuurtjes", that generally were accompanied by some delicious beers, and to Janneke and Linda for helping to revive the Young Hydrologic Society NL together with colleagues from TU Delft, VU Amsterdam and ITC Twente.

And then there are two friend groups that have supported me throughout this process, particularly by *not* talking about my PhD. Instead, having fun together was a great and necessary distraction sometimes! Joost, Kees, Koen, Luuk, Rik and Ruud (yes, I am the only one with a two-syllable name), I hope we can have many BBQ moments at the Rhine again (I do like my beer cold next time, though..). And Luuk, you are the next one to defend your thesis! Kaj, Niels, Roeland, Servaes, Sjoerd and Wif, it is great that we have stayed friends ever since high school. You have been calling me the "Avatar" of the group with my Soil, Water, Atmosphere background. This might not be entirely accurate, but I am glad to be the Avatar of our group.

Mam, pap, Annefloor, ondanks dat jullie misschien van tijd tot tijd geen idee hebben gehad waar ik precies mee bezig was (en jullie zijn niet de enigen hoor), hebben jullie me enorm gesteund gedurende het proces. Zolang ik jullie mijn onderwerp in grote lijnen uit kon leggen, wist ik in ieder geval zeker dat ik nog met een 'pakbaar' onderwerp bezig was en dat heeft me ook een beetje uit de onderzoeksbubbel gehouden.

Last but not least, Mary Rose, love of my life, thanks for being there throughout the entire journey and everything that will still follow!

Ruben Imhoff
Wageningen, November 2022

List of publications

Peer-reviewed journal papers | In this thesis

Imhoff, R. O., Brauer, C. C., Overeem, A., Weerts, A. H., and Uijlenhoet, R. (2020). Spatial and temporal evaluation of radar rainfall nowcasting techniques on 1,533 events, *Water Resources Research*, 56(8), e2019WR026723, doi:10.1029/2019WR026723.

Imhoff, R. O., Overeem, A., Brauer, C. C., Leijnse, H., Weerts, A. H., and Uijlenhoet, R. (2020). Rainfall nowcasting using commercial microwave links, *Geophysical Research Letters*, 19, e2020GL089365, doi:10.1029/2020GL089365.

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

Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., and Weerts, A. H. (2022). Large-sample evaluation of radar rainfall nowcasting for flood early warning, *Water Resources Research*, 58(3), e2021WR031591, doi:10.1029/2021WR031591.

Imhoff, R. O., De Cruz, L., Dewettinck, W., Velasco-Forero, C., Nerini, D., Van Genderachter, M., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., and Weerts, A. H. (2022). Scale-dependent blending of ensemble rainfall nowcasts with NWP in the open-source pysteps library, *Submitted to Quarterly Journal of the Royal Meteorological Society*.

Peer-reviewed journal papers | Other

Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., and Weerts, A. H. (2020). Scaling point-scale (pedo)transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: An example for the Rhine River, *Water Resources Research*, 56(4), e2019WR026807, doi:10.1029/2019WR026807.

van Hateren, T. C., Jongen, H. J., Al-Zawaidaha, H., Beemster, J. G. W., Boeke, J., Bogerd, L., Gao, S., Kannen, C., van Meerveld, I., de Lange, S., Linke, F., Pinto, R. B., Remmers, J., Ruijsch, J., Rusli, S. R., van de Vijzel, R. C., Aerts, J., Agoungbome, S. M. D., Anys, M., van Emmerik, T., Gallitelli, L., Gesualdo, G., Hanus, S., Hea, Z., Hoffmesiter, S., Imhoff, R. O., Meshram, S., Meyer, J., Oliveira, A. M., Müller, A. C. T., Nijzink, R., Schellerberg, M., Schreyers, L., Schymanski, S., Sehgal, D., Tasser, P., Teuling, A. J., Trevisson, M., Waldschläger, K., Walraven, B., Wannasin, C., Wienhöfer, J., Zander, M., Zhanga, S., Zhoua, J., Zomer, J. Y., and Zwartendijk, B. W. (2022). Where should hydrology go? An early-career perspective on the next IAHS Scientific Decade, *Submitted to Hydrological Sciences Journal*.

van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Eilander, D., Bouaziz, L., Hegnauer, M., ten Velden, C., and Russell, B. (2022). Wflow_sbm v0.6.1, a spatially distributed hydrologic model: from global data to local applications, *Geoscientific Model Development Discussions*, doi:10.5194/gmd-2022-182.

Rombek, N., Imhoff, R. O., and Brauer, C. C. (2022). Rainfall nowcasting for urban applications, *To be submitted to Hydrology and Earth System Sciences*.

Publications in Dutch journals

Imhoff, R. O. (2017). De Radar to Catchment-applicatie, *Stromingen*, 30(4), 59–64.

Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., and Weerts, A. H. (2020). Ruimtelijk schaalbare hydrologische modelparameters uit open-source omgevingsdata: een voorbeeld voor de Rijn, *Stromingen*, 26(3), 19–36.

Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Leijnse, H., Weerts, A. H., and Uijlenhoet, R. (2022). CARROTS: een klimatologische correctie voor radarneerslag in een operationele setting, *Stromingen*, 28(2), 57–72.

Media interaction

Interviews, as a result of the publication of and news item about *Rainfall nowcasting using commercial microwave links* (GRL, 2020), on:

- Dutch national radio (Radio 1 and BNR)
- Televised news (Editie NL, RTL4)



*Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment*

D I P L O M A

for specialised PhD training

The Netherlands research school for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Ruben Olaf Imhoff

born on 13 August 1995 in 's-Hertogenbosch, the Netherlands

has successfully fulfilled all requirements of the
educational PhD programme of SENSE.

Wageningen, 1 November 2022

Chair of the SENSE board

Prof. dr. Martin Wassen

The SENSE Director

Prof. Philipp Pattberg

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)



K O N I N K L I J K E N E D E R L A N D S E
A K A D E M I E V A N W E T E N S C H A P P E N



The SENSE Research School declares that **Ruben Olaf Imhoff** has successfully fulfilled all requirements of the educational PhD programme of SENSE with a work load of 42.2 EC, including the following activities:

SENSE PhD Courses

- o Environmental research in context (2019)
- o Research in context activity: 'News, radio and television items about published paper' (2020)

Other PhD and Advanced MSc Courses

- o Delft-FEWS basiscursus, Deltares (2019)
- o Training course: Predictability and ensemble forecast systems. ECMWF (2019)
- o HWM writing week (2020 and 2022)
- o Machine Learning Workshop, eScience Centre (2021)

Selection of Management and Didactic Skills Training

- o Supervising five MSc and three BSc students with thesis (2020-2022)
- o Assisting practicals of the BSc course 'Water' and the MSc courses 'Urban Hydrometeorology', 'Catchment and Climate Hydrology' and 'Radar rainfall estimation and adjustment' (2018-2022)
- o 2 articles in the magazine of the Dutch Hydrology Society (2020 and 2022)
- o Workshop 'Nowcasting products for the Dutch water authorities', Deltares (2021)
- o pySTEPS nowcasting short course, European Conference on Radar in Meteorology and Hydrology (2022)
- o Break-out session about radar rainfall estimations and nowcasting, International Delft-FEWS User Days (2020)
- o Dutch-German workshop for water managers, LHW, Magdeburg, Germany (2021)
- o Chairman Young Hydrologic Society NL (2021-2022)

Selection of Oral Presentations

- o *Scaling point-scale pedotransfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrological modelling: An example for the Rhine river.* EGU General Assembly, 7-21 April 2019, Vienna, Austria
- o *Spatial and temporal evaluation of radar rainfall nowcasting techniques on 1481 events.* Boussinesq Lecture, 17 October 2019, Amsterdam, The Netherlands
- o *Opportunistic Rainfall Nowcasting with Commercial Microwave Link Data.* AGU Fall Meeting, 7-17 December 2020, San Francisco, United States of America / Online
- o *Radar rainfall nowcasting for flash flood forecasting.* 11th Assembly of the International Association of Hydrological Sciences. 29 April- 3 June 2022, Montpellier, France

SENSE coordinator PhD education

Dr. ir. Peter Vermeulen

The presented research was financially supported by funding from the DAISY2 project, supported by the European Regional Development Fund, and by funding from Deltares' Strategic Research Program. The research was carried out at the Operational Water Management & Early Warning Department of Deltares (Delft, the Netherlands) and at the Hydrology and Quantitative Water Management Group, Department of Environmental Sciences, Wageningen University & Research (Wageningen, the Netherlands).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

