Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library

Imhoff, Ruben O.; De Cruz, Lesley; Dewettinck, Wout; Brauer, Claudia C.; Uijlenhoet, Remko; van Heeringen, Klaas Jan; Velasco-Forero, Carlos; Nerini, Daniele; Van Ginderachter, Michiel; Weerts, Albrecht H.

DOI
10.1002/qj.4461

Publication date
2023

Document Version
Final published version

Published in
Quarterly Journal of the Royal Meteorological Society

Received: 12 August 2022 | Revised: 4 March 2023 | Accepted: 7 March 2023

DOI: 10.1002/qj.4461

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library

Ruben O. Imhoff[1,2] | **Lesley De Cruz**[3,4] | **Wout Dewettinck**[5] | **Claudia C. Brauer**[2] | **Remko Uijlenhoet**[6] | **Klaas-Jan van Heeringen**[1] | **Carlos Velasco-Forero**[7] | **Daniele Nerini**[8] | **Michiel Van Ginderachter**[3] | **Albrecht H. Weerts**[1,2]

[1]Operational Water Management & Early Warning, Deltares, Delft, Netherlands

[2]Hydrology and Quantitative Water Management, Wageningen University & Research, Wageningen, Netherlands

[3]Department of Meteorological and Climatological Research, Royal Meteorological Institute of Belgium, Brussels, Belgium

[4]Electronics and Informatics Department, Vrije Universiteit Brussel, Brussels, Belgium

[5]Department of Physics and Astronomy, Ghent University, Ghent, Belgium

[6]Department of Water Management, Delft University of Technology, Delft, Netherlands

[7]Radar Science and Nowcasting Team, Bureau of Meteorology, Melbourne, Victoria Australia

[8]Federal Office of Meteorology and Climatology, MeteoSwiss, Locarno-Monti, Switzerland

Correspondence
Albrecht H. Weerts, Operational Water Management & Early Warning, Deltares, Delft, Netherlands.
Email: Albrecht.Weerts@deltares.nl

Funding information
Belgian Federal Science Policy Office, Grant/Award Number: Prf-2020-017; Deltares' Strategic Research Program

Abstract

Flash flood early warning requires accurate rainfall forecasts with a high spatial and temporal resolution. As the first few hours ahead are already 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 2 hr of the forecast, it needs to be combined with NWP forecasts to extend the skillful lead time of short-term rainfall forecasts, which should increase decision-making times. We implemented an adaptive scale-dependent ensemble blending method in the open-source pysteps library, based on the Short-Term Ensemble Prediction System scheme. In this implementation, the extrapolation (ensemble) nowcast, (ensemble) NWP, and noise components are combined 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 than only nowcasting (in terms of event-averaged continuous ranked probability score and critical success index values) and adds value compared with NWP for the first hours of the forecast, 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.

KEYWORDS

blending, early warning, nowcasting, numerical weather prediction, open source, rainfall

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 were in Belgium and Germany, and led to more than 25 billion USD in economic and infrastructural damages (AON, 2021; Koks *et al.*, 2022; 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 hr) 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 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 6 hr in the Netherlands and Belgium, although this is not the case everywhere) for usage in flash flood early warning systems. In addition, most operational NWP systems have a latency of several hours between model initialization and delivery at the end user. Consequently, the timing and location of intense rainfall events are often missed (Lin *et al.*, 2005; Roberts and Lean, 2008; Berenguer *et al.*, 2012; Pierce *et al.*, 2012).

One way to tackle this problem, at least for the first hours into the future, is to use rainfall nowcasting techniques, which (statistically and heuristically) 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 km^2 and 5 min for the QPE of current weather radars, Serafin and Wilson, 2000; Overeem *et al.*, 2009). 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; Vivoni *et al.*, 2007; Germann *et al.*, 2009; Liguori and Rico-Ramirez, 2012; Liguori and Rico-Ramirez, 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 are available, which can be categorized in field-based nowcasting methods that derive an advection field from the gridded rainfall observations and can add stochastic processes to simulate rainfall field evolutions (e.g., Seed, 2003; Bowler *et al.*, 2006; Berenguer *et al.*, 2011; Seed *et al.*, 2013; Sokol *et al.*, 2017; Ayzel *et al.*, 2019), object-oriented methods that track individual rainfall (storm) cells (e.g., Dixon and Wiener, 1993; Han *et al.*, 2009), analogue-based methods that look for a similar state in a historical dataset (e.g., Atencia and Zawadzki, 2014; Atencia and Zawadzki, 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.* (2017), and Pulkkinen *et al.* (2020, 2021).

One of pysteps' main features is an efficient Python implementation of the probabilistic field-based nowcasting scheme Short-Term Ensemble Prediction System (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 and Schertzer, 1995; Harris *et al.*, 1996; Marsan *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, STEPS generates an ensemble of rainfall forecasts. As a result, STEPS 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

2–3 hr. This maximum skilful 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 hr for larger-scale, more persistent rainfall events, and up to 6 hr for continental-scale persistent stratiform events (Germann and Zawadzki, 2002; Lin *et al.*, 2005; Germann *et al.*, 2006; Berenguer *et al.*, 2011; Berenguer *et al.*, 2012; Liguori and Rico-Ramirez, 2012; Foresti *et al.*, 2016; Simonin *et al.*, 2017; Mejsnar *et al.*, 2018; Ayzel *et al.*, 2019; Imhoff *et al.*, 2020).

To extend the skilful 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 are a plethora of blending techniques present (e.g. Golding, 1998; Bowler *et al.*, 2006; Atencia *et al.*, 2010; Kober *et al.*, 2012; Bailey *et al.*, 2014; Kober *et al.*, 2014; Simonin *et al.*, 2017; Nerini *et al.*, 2019; Yoon, 2019; Radhakrishnan and 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 combination of the extrapolation nowcast, NWP, and stochastic noise components is performed at different spatial scales using varying blending weights per cascade level. The main objective of this article is to describe the implementation of the STEPS ensemble blending approach in the pysteps framework. In this description, we introduce some new functionalities that are in line with existing pysteps functionalities or allow for the operational usage of the system. 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 scales. We benchmark the results against the NWP rainfall forecasts as issued by the Royal Meteorological Institute of Belgium (RMI; which covers the entire study area and has a 5 min temporal resolution), radar-based ensemble nowcasts with pysteps, and a simple linear blending between the former two.

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 or adjusted in this study. The modular set-up of pysteps allows the user to choose whether to include these functionalities. When these functionalities are not included, the system uses the basic functionalities and set-up that are described in Bowler *et al.* (2006) and Pulkkinen *et al.* (2019). For more information regarding specific STEPS or pysteps functionalities, we refer the reader to Bowler *et al.* (2006), Seed *et al.* (2013), and Pulkkinen *et al.* (2019).

The largest uncertainty in nowcasting is due to the unknown growth and decay of rainfall fields. This uncertainty requires an ensemble approach, which is provided by STEPS. It uses the multifractal properties of rainfall, characterized by power-law scaling behaviour in the spatial structure and the predictability of different scales. To exploit this property, it decomposes rainfall fields into a multiplicative cascade with, generally, a fast Fourier transform (FFT), representing different spatial scales and lifetimes, with smaller scale features having shorter lifetimes than larger scale features do (Lovejoy and Schertzer, 1995; Harris *et al.*, 1996; Marsan *et al.*, 1996; Foufoula-Georgiou, 1998; Seed *et al.*, 1999). STEPS tries to take the uncertainty in the rainfall field evolution into account by applying different spatially and temporally correlated stochastic perturbations to each spatial scale in the cascades. This results in large-scale features to evolve more slowly than small-scale features, and it results in the construction of an ensemble of outcomes, which, together, should ensure a more appropriate representation of the uncertainty associated with the rainfall field evolution. This is applied to the radar-based nowcasting part, for instance in pysteps, but this principle is also used when combining nowcasts and NWP in STEPS blending. Hence, the blending takes place at every individual cascade level, which requires both the radar data and the NWP forecasts to be decomposed into a multiplicative cascade. A third component, containing the stochastic perturbations, is added to account for the uncertainty in the rainfall field evolution and to construct an ensemble of outcomes in the blending procedure (Seed, 2003; Bowler *et al.*, 2006).

Figure 1 gives an overview of the workflow to compute a blended precipitation forecast in pysteps. In this figure, we have indicated with stars which functionalities are new or adjusted from their original implementations in Bowler *et al.* (2006), Seed *et al.* (2013), and Pulkkinen *et al.* (2019),

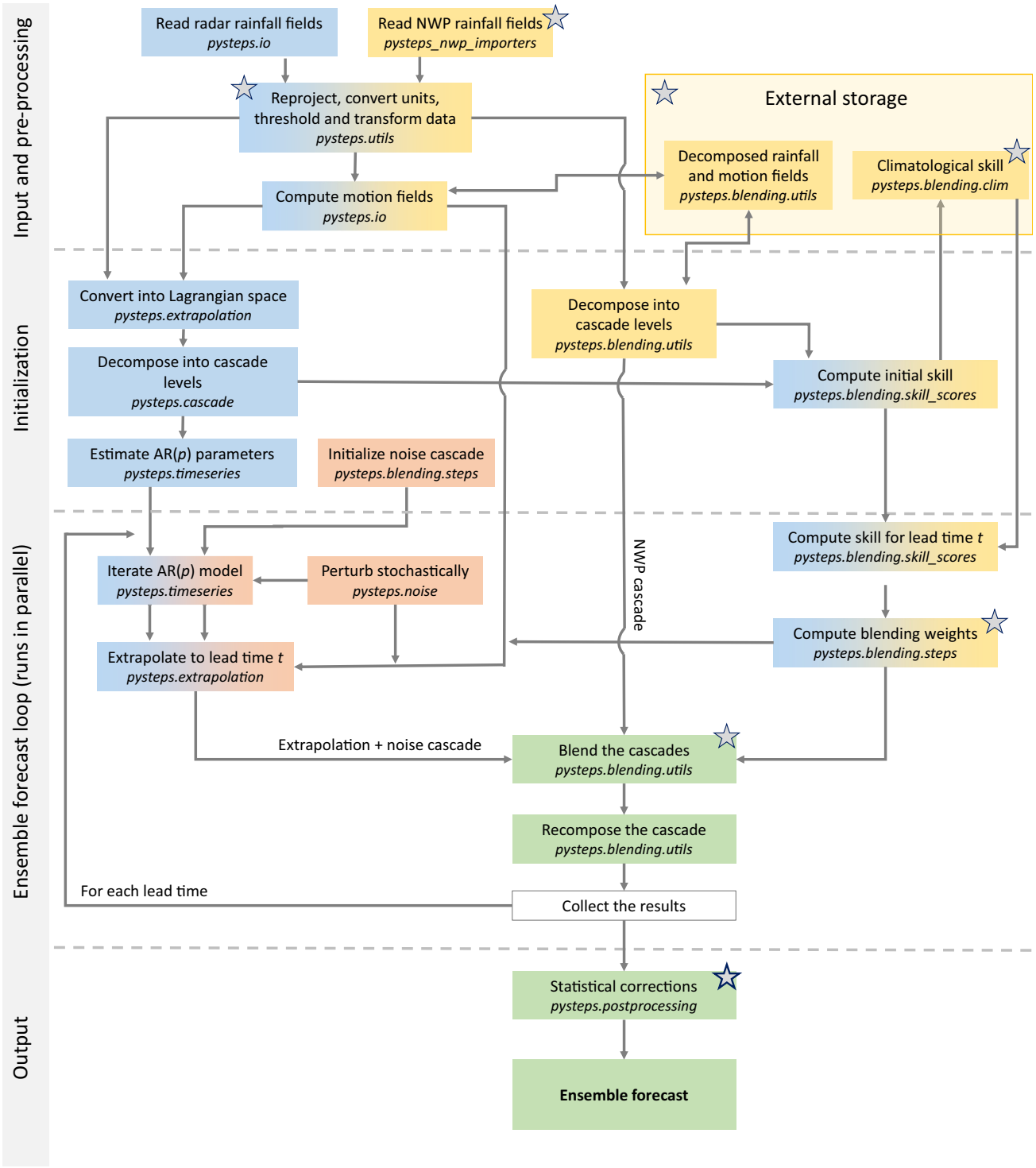

FIGURE 1 Schematic overview of the workflow for computing blended precipitation forecasts using pysteps. For each chart element, the bottom row in italics describes the pysteps module used to execute the task described. Blue colours represent the elements that are part of the extrapolation cascade, red colours those of the noise cascade, and yellow those of the NWP model cascade. An overlap in colours 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 colours indicate the merged cascades up to the final output. A star on top of the element indicates that this module is new or adjusted from the original Short-Term Ensemble Prediction System or pysteps implementation. AR(p): pth-order autoregressive process; NWP: numerical weather prediction.

and we also specifically mention this in the subsequent sections. In principle, this implementation combines, per ensemble member, an extrapolation nowcast 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 2.1). As the rainfall fields in these cascade levels evolve over time, a *p*th-order autoregressive process per cascade level is used (described in Section 2.2). The blending takes place level-by-level, meaning that the weights of the different components are specific to each cascade level (Section 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 2.3.1), meaning that the blending weights vary both per spatial scale level and in time (per issue time, but also over the forecast horizon). After the blending of the components per cascade level (Section 2.5), the cascade levels are recomposed to one blended forecast member and some post-processing steps take place (Section 2.6), resulting in a blended ensemble forecast for that issue time.

2.1 | Constructing the cascades for the three components

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

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

with $R_{i,j}$ (mm·hr^{-1}) the precipitation rate at grid cell i, j and time t. $\mathrm{dBR}_{i,j}(t)$ is −15 for precipitation intensities of less than 0.1 mm·hr^{-1} (this is adjustable) to ensure a clear rain–no rain transition for the nowcasting step. The transformation ensures that the precipitation data have 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 2.4.

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 the main motivation to decompose precipitation fields into a multiplicative cascade. The levels in this cascade represent different spatial scales, which are treated individually in the nowcasting scheme. An advantage of the log-transformation in Equation (1) is that this decomposition into a multiplicative cascade becomes an additive cascade in the log-transformed space (Seed, 2003). With an FFT, the precipitation field is decomposed in such an additive cascade by applying Gaussian filtering with weight functions and also by back-transforming this to the grid space (Pulkkinen *et al.*, 2018). This results in k cascade levels representing different spatial scales, which are normalized afterwards:

$$\mathrm{dBR}_{i,j}(t) = \sum_{k=1}^{K} \sigma_k(t) Y_{k,i,j}(t) + \mu_k(t). \tag{2}$$

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 2.2).

Once the radar data have 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 (a new functionality that has been added to pysteps in this study), which saves calculation time when the NWP data have an update frequency that is lower than the update frequency of the blended nowcasts (for instance, a 6 hr versus a 5 min update frequency).

After the aforementioned steps, three cascades are present that represent the extrapolation, NWP, and noise components and which will be combined with a set of blending weights per cascade level k. We will elaborate on this in Section 2.5.

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 pth-order (generally second-order) autoregressive (AR(p)) process per cascade level. This AR(p) process injects a stochastic perturbation term that represents the uncertainty in growth and decay of the precipitation field over time. In the original pysteps implementation, which focuses on nowcasting only, this stochastic term is added to the extrapolation component at every time step, avoiding the need for a separate noise component (Pulkkinen *et al.*, 2019). The addition of an NWP component in the blending approach (Bowler *et al.*, 2006) means that the weights of the noise terms are no longer determined by the AR(p) parameters alone, and we can no longer simply aggregate the noise into the extrapolation cascade step by step. Instead, we need to model the temporal evolution of the extrapolation and noise cascades as two separate processes, where the extrapolation cascade regresses without added noise, as

$$Y^{\text{ext}}_{k,i,j}(t+t_1) = \sum_{p=1}^{p_{\max}} \phi_{k,p} Y^{\text{ext}}_{k,i,j}(t+t_1-p\Delta t), \tag{3}$$

and the noise cascade regresses with added noise, according to

$$Y^{\epsilon}_{k,i,j}(t+t_1) = \sum_{p=1}^{p_{\max}} \phi_{k,p} Y^{\epsilon}_{k,i,j}(t+t_1-p\Delta t) + \phi_{k,0}\epsilon_{k,i,j}(t+t_1). \tag{4}$$

In these equations, Y^{ext} and Y^{ϵ} represent the extrapolation and noise cascades at cascade level k and lead time $t+t_1$, p is the AR order, Δt is 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, constructed using FFT filtering, that ensures the noise field has the desired correlation structure – (for more information and the available filtering methods, see Schertzer and Lovejoy, 1987; Pegram and Clothier, 2001; Bowler *et al.*, 2006; Pulkkinen *et al.*, 2019).

2.3 | Blending weights

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 with regard to the latest observation at the issue time of the blended nowcast. As the forecast lead time advances, the weights increase for the noise component and 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 parametrization of the blending process and weights determination and ensures that the blending weights represent the real-time state of the components that are combined (Bowler *et al.*, 2006).

The AR(p) model (Section 2.2) determines the skill decrease of the extrapolation component per cascade level k as follows (Bowler *et al.*, 2004):

$$\rho^{\text{ext}}_k(t+t_1) = \sum_{p=1}^{p} \phi_{k,p} \rho^{\text{ext}}_k(t+t_1-p\Delta t), \tag{5}$$

with t the issue time of the forecast, t_1 the lead time, and starting value

$$\rho^{\text{ext}}_k(t) = 1. \tag{6}$$

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

$$\rho^{\text{ext}}_k(t+\Delta t) = \frac{\phi_{k,1}}{1-\phi_{k,2}}. \tag{7}$$

The NWP skill per lead time is based on the initial skill, Pearson correlation at cascade level k, $\rho^{\text{nwp}}_k(t)$, and regresses toward a climatological value (Bowler *et al.*, 2004):

$$\rho^{\text{nwp}}_k(t+\Delta t) = q^{\text{nwp}}_k \rho^{\text{nwp}}_k(t) + (1-q^{\text{nwp}}_k)\overline{\rho^{\text{nwp}}_k}, \tag{8}$$

with

$$q^{\text{nwp}}_k = \mathrm{e}^{-t_1/L_{1,k}}(2-\mathrm{e}^{-t_1/L_{2,k}}). \tag{9}$$

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 Supporting Information Table S1, which is based in Bowler

et al. (2004). $\overline{\rho_k^{\text{nwp}}}$ is the climatological skill value toward which Equation (8) regresses.

Bowler and co-workers (Bowler *et al.*, 2004; Bowler *et al.*, 2006) used fixed values of $\overline{\rho_k^{\text{nwp}}}$ 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, as a new component in the blending procedure, 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 in a new (operationally usable) storage module in pysteps (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.

2.3.2 | Weights determination

The three components are combined 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. In the following, we describe the principle of both blending methods and show the difference in resulting weights for a test case on June 29, 2021, 1330 UTC in Figure 2. In Sections 3.3.4 and 4.3, we describe the evaluation and effect of both methods on the resulting rainfall forecast for this case.

The 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 5 and 8) and are calculated per lead time t_l as

$$w_k^{\text{ext}}(t+t_l) = \rho_k^{\text{ext}}(t+t_l)\sqrt{\frac{\lambda_k^{\text{ext}}(t+t_l)}{\lambda_k^{\text{ext}}(t+t_l)+\lambda_k^{\text{nwp}}(t+t_l)}}, \quad (10)$$

$$w_k^{\text{nwp}}(t+t_l) = \rho_k^{\text{nwp}}(t+t_l)\sqrt{\frac{\lambda_k^{\text{nwp}}(t+t_l)}{\lambda_k^{\text{ext}}(t+t_l)+\lambda_k^{\text{nwp}}(t+t_l)}}, \quad (11)$$

$$w_k^{\epsilon}(t+t_l) = \sqrt{1-[w_k^{\text{ext}}(t+t_l)]^2-[w_k^{\text{nwp}}(t+t_l)]^2}, \quad (12)$$

with $w_k^{\text{ext}}(t+t_l)$, $w_k^{\text{nwp}}(t+t_l)$, and $w_k^{\epsilon}(t+t_l)$ 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

$$\lambda_k^{\text{ext}}(t+t_l) = \frac{[\rho_k^{\text{ext}}(t+t_l)]^2}{1-[\rho_k^{\text{ext}}(t+t_l)]^2}, \quad (13)$$

$$\lambda_k^{\text{nwp}}(t+t_l) = \frac{[\rho_k^{\text{nwp}}(t+t_l)]^2}{1-[\rho_k^{\text{nwp}}(t+t_l)]^2}. \quad (14)$$

The Seed et al. method

A disadvantage of the method by Bowler *et al.* (2006) is that it assumes all cascades are uncorrelated. To enable the blending of more than two forecasts (e.g., 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 normalized covariance matrix, which consists of the cross-correlations between the extrapolation and model cascades (Seed, personal communication, December 2021). Considering these adjustments, the weights are determined as

$$\vec{w}_k(t+t_l) = \begin{bmatrix} \rho_k^{1,1}(t+t_l) & \rho_k^{1,2}(t+t_l) & \cdots \\ \rho_k^{1,2}(t+t_l) & \rho_k^{2,2}(t+t_l) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}^{-1} \cdot \begin{bmatrix} \rho_k^{1}(t+t_l) \\ \rho_k^{2}(t+t_l) \\ \cdots \end{bmatrix}, \quad (15)$$

with $\rho_k^{1,2}(t+t_l)$ the cross-correlation between models 1 and 2 (e.g., 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 combined, the matrix increases from 2×2 to $n_{\text{models}} \times n_{\text{models}}$, as indicated with the ellipses in Equation (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 $t+t_l$.

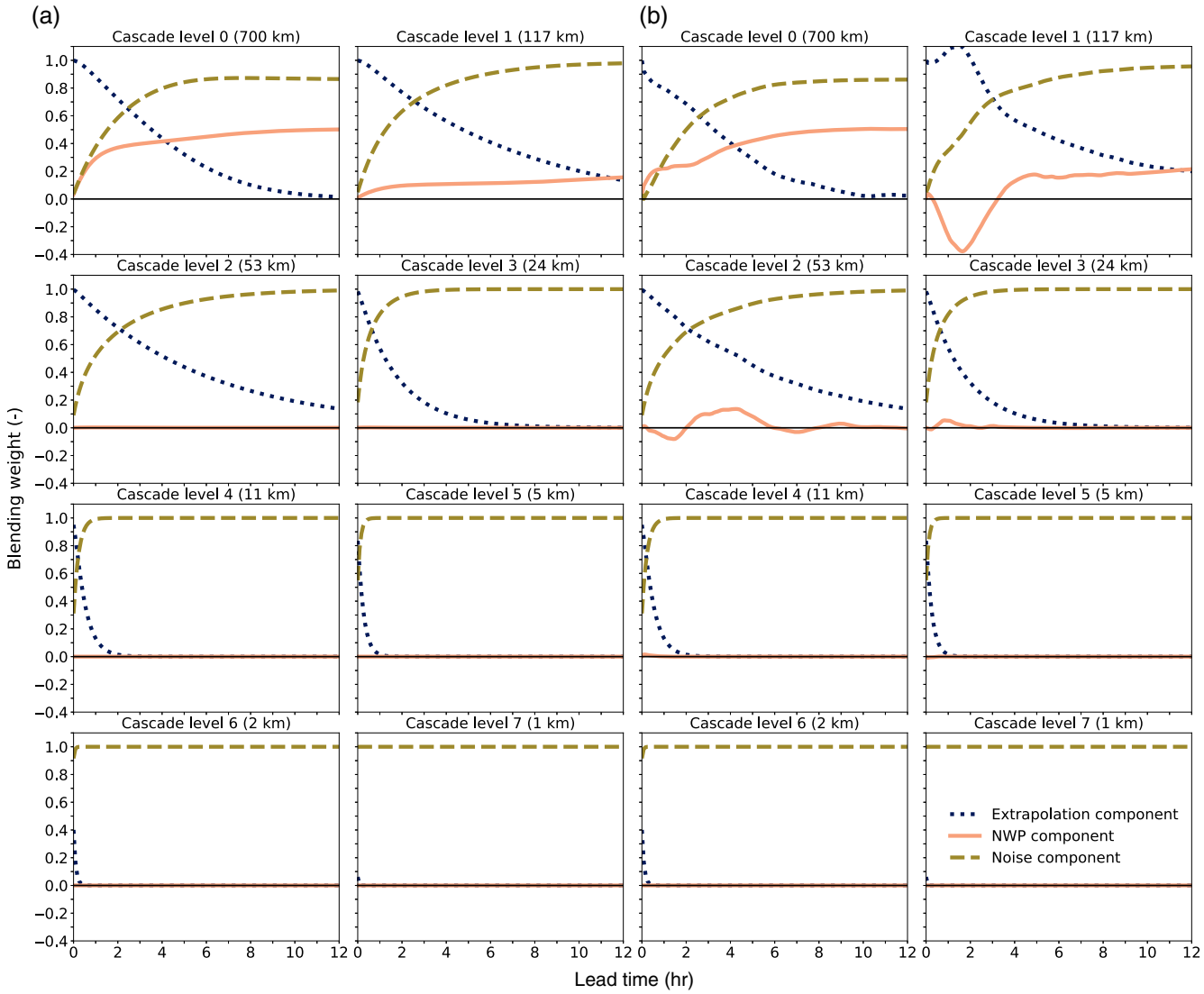

FIGURE 2 Resulting blending weights per cascade level for the test case of June 29, 2021, 1330 UTC, when using the approaches of (a) Bowler *et al.* (2006) and (b) Seed *et al.* (2013). The blue lines correspond to the blending weights of the extrapolation component, the pink lines to the numerical weather prediction (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. NWP: numerical weather prediction.

Subsequently, the noise weight is calculated as

$$w_k^{\epsilon}(t+t_1) = \sqrt{1 - \vec{w}_k(t+t_1) \cdot \begin{bmatrix} \rho_k^1(t+t_1) \\ \rho_k^2(t+t_1) \\ \cdots \end{bmatrix}}. \quad (16)$$

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

$$\vec{w}_k(t+t_1) \cdot \begin{bmatrix} \rho_k^1(t+t_1) \\ \rho_k^2(t+t_1) \\ \cdots \end{bmatrix} > 1.0. \quad (17)$$

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 zero. 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).

2.4 | Advection of the extrapolation and noise cascades

Before we can combine the different cascades, the extrapolation and noise cascades have to be extrapolated from the

issue time (the most recent observation) to time $t + t_1$. Both cascades are in Lagrangian space, because this allows for the temporal evolution of the cascades through an AR(p) process (Section 2.2). In pysteps, this extrapolation step is one of the last steps in the forecasting framework, but since the NWP forecast for time $t + t_1$ 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 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 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, the derived motion fields are also combined (Section 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 side or multiple sides is allowed to move into the domain on the opposite side(s) again (so-called mirroring). The extrapolation cascade can eventually also be advected out of the domain. This is a known limitation to nowcasting and constitutes another motivation to include information from other sources, such as NWP forecasts.

2.4.1 | Combining the velocity fields

In this implementation, we combine the motion fields for time $t + t_1$ using the second cascade level weights, as was done in Bowler *et al.* (2006), in the following way:

$$v(t + t_1) = w_2^{\text{ext}^*}(t + t_1)v^{\text{ext}}(t + t_1) + w_2^{\text{nwp}^*}(t + t_1)v^{\text{nwp}}(t + t_1), \tag{18}$$

with v^{ext} and v^{nwp} the velocity fields for the extrapolation and NWP cascades, and $w_2^{\text{ext}^*}(t + t_1)$ and $w_2^{\text{nwp}^*}(t + t_1)$ 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 $w_2^{\text{nwp}^*}(t + t_1) = 1 - w_2^{\text{ext}^*}(t + t_1)$, as that method does not ensure a total weight of 1.0, which can lead to either too small weights or negative $w_2^{\text{nwp}^*}$ weights when $w_2^{\text{ext}^*}$ is larger than 1.0. To take into account the uncertainty in the motion field development, $v^{\text{ext}}(t + t_1)$ can be stochastically perturbed prior to the blending step in Equation (18) (see Bowler *et al.*, 2006; Pulkkinen *et al.*, 2019). Although the use of the second cascade level to combine the velocity fields is the same as in the original implementation by Bowler *et al.* (2006), it is unclear if this leads to the best results. Further testing is recommended for this.

2.5 | Combining the cascades and recomposing the forecast

After the temporal evolution and advection of the extrapolation and noise cascades (Sections 2.2 and 2.4), the cascades can be blended with the NWP cascade(s):

$$\begin{aligned} Y_{k,i,j}^{\text{blended}}(t + t_1) = {} & w_k^{\text{ext}}(t + t_1)Y_{k,i,j}^{\text{ext}}(t + t_1) \\ & + w_k^{\text{nwp}}(t + t_1)Y_{k,i,j}^{\text{nwp}}(t + t_1) \\ & + w_k^{\epsilon}(t + t_1)Y_{k,i,j}^{\epsilon}(t + t_1), \end{aligned} \tag{19}$$

where $Y_{k,i,j}^{\text{ext}}(t + t_1)$ and $Y_{k,i,j}^{\epsilon}(t + t_1)$ are extrapolated to time $t + t_1$ (Section 2.4). Note that outside the radar domain only the NWP and noise cascade(s) are combined using weights that are determined based on the presence of only NWP and noise components. New in this implementation is that the aforementioned blending procedure can take place in three ways:

1 There is only one deterministic NWP model. Equation (19) 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 NWP members need to be combined with extrapolation and noise cascades. This means that each member of the blended nowcast contains information from all NWP members; the procedure remains the same as before, but, instead of one model cascade and weight, multiple model cascades and weights are introduced (see also Section 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 combined individually per realization, with the extrapolation and noise cascades. Equation (19) 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. An advantage of this method is that it can bypass the correlation problem between individual model realizations; and, with this approach, the ensemble spread can become larger than for method 1, which computes a (weighted) mean of the NWP members, thereby effectively reducing the ensemble spread.

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

$$\mathrm{dBR}_{i,j}^{\mathrm{blended}}(t+t_1) = \sum_{k=1}^{K} \sigma_k^{\mathrm{blended}}(t+t_1) Y_{k,i,j}^{\mathrm{blended}}(t+t_1) + \mu_k^{\mathrm{blended}}(t+t_1), \quad (20)$$

where $\sigma_k^{\mathrm{blended}}(t+t_1)$ and $\mu_k^{\mathrm{blended}}(t+t_1)$ are the weighted sums of the means and standard deviations of the extrapolation and NWP model cascades (Bowler *et al.*, 2004):

$$\mu_k^{\mathrm{blended}}(t+t_1) = \frac{w_k^{\mathrm{ext}}(t+t_1)}{w_k^{\mathrm{ext}}(t+t_1) + w_k^{\mathrm{nwp}}(t+t_1)} \mu_k^{\mathrm{ext}}(t) + \frac{w_k^{\mathrm{nwp}}(t+t_1)}{w_k^{\mathrm{ext}}(t+t_1) + w_k^{\mathrm{nwp}}(t+t_1)} \mu_k^{\mathrm{nwp}}(t), \quad (21)$$

$$\sigma_k^{\mathrm{blended}}(t+t_1) = \frac{w_k^{\mathrm{ext}}(t+t_1)}{w_k^{\mathrm{ext}}(t+t_1) + w_k^{\mathrm{nwp}}(t+t_1)} \sigma_k^{\mathrm{ext}}(t) + \frac{w_k^{\mathrm{nwp}}(t+t_1)}{w_k^{\mathrm{ext}}(t+t_1) + w_k^{\mathrm{nwp}}(t+t_1)} \sigma_k^{\mathrm{nwp}}(t). \quad (22)$$

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 that constrains the mask to the observed grid cells that exceed a given threshold, and an incremental masking method that 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, and 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). This ensures that the cdf of the rainfall intensities in the blended forecast is not enhanced or tuned down by a component with a low weight, or by rainfall from the extrapolation component that advects out of the domain. Therefore, we have developed a Lagrangian blended probability matching scheme and incremental masking strategy, in which the mask consists of the combined 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 $t+t_1$ with the velocity field from Equation (18). Subsequently, the extrapolated radar field is combined with the NWP forecast field(s) using the blending weights from the second cascade level, which is also used to combine the velocity fields. These blending weights do not include the noise weight, increasing both the nowcasting and NWP component weights. The use of the second cascade level weights is merely for consistency with the use of these weights for blending the velocity fields, but is not, by definition, the best choice. We will elaborate on the limitations and possible future improvements of the proposed Lagrangian blended probability matching scheme in the Section 5. Finally, after the previous step, the blended field is used as a reference for the rainfall intensity distribution used in the post-processing steps.

3 | EVALUATION OF THE BLENDED RAINFALL FORECASTS

3.1 | Study area and rainfall events

The study areas to test and evaluate the blended forecast of Section 2 are Belgium and the south of the Netherlands (Figure 3). Besides a focus on domain-wide rainfall forecasting performance, we focus on four catchments: (1) Vesdre (685 km^2), which is a steep and quickly responding catchment in the Belgian Ardennes; (2) Demer (2,268 km^2), which starts in the hilly parts of Belgian Limburg and flattens out downstream in Flemish Brabant; (3) Geul (323 km^2), which has its origins in the foothills of the Belgian Ardennes and flows into the hills of southern Limburg in the Netherlands; and (4) Dommel (1,691 km^2), a relatively flat catchment that has its origin in the north of Belgium and flows through the south of the Netherlands.

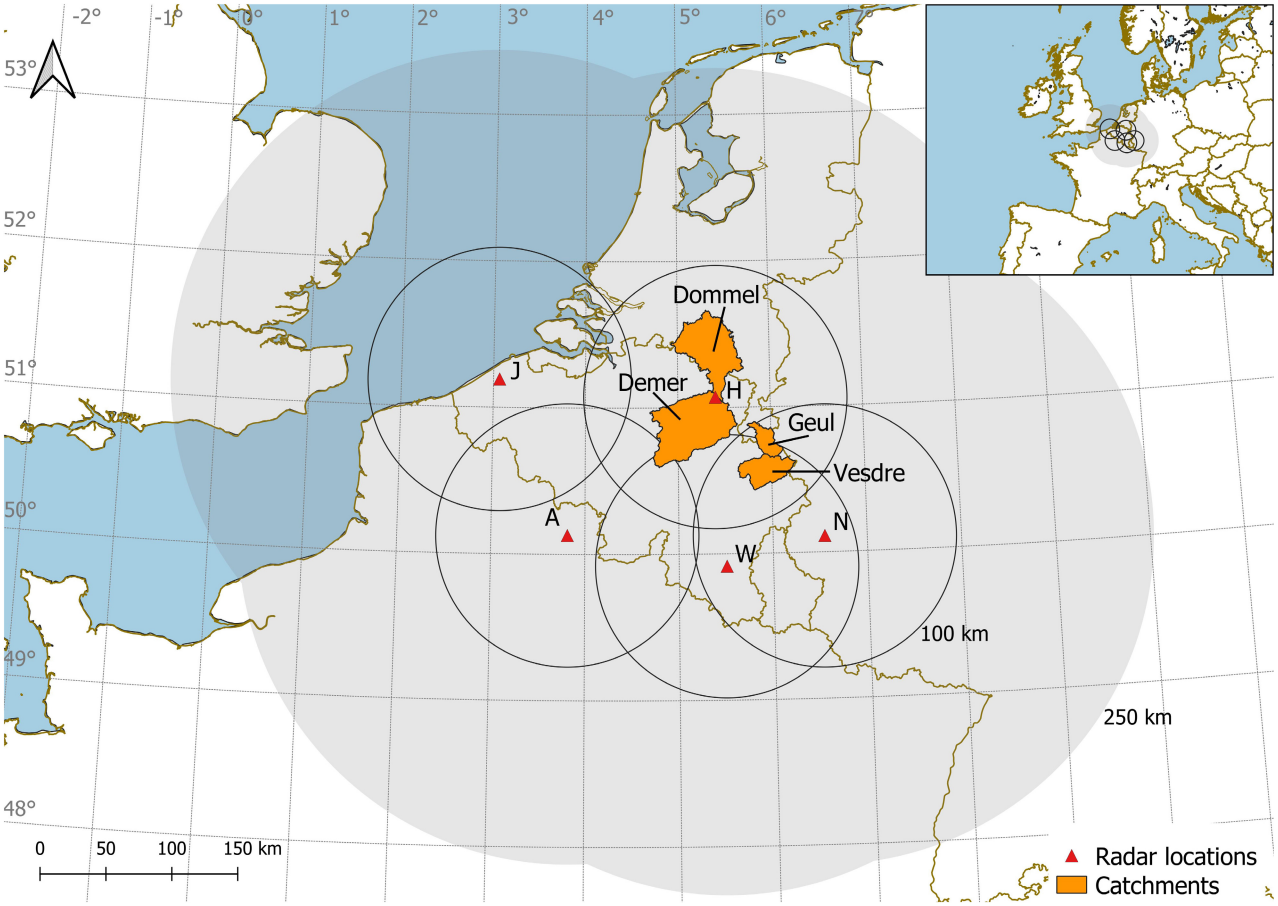

FIGURE 3 Map of the study area with the four catchments indicated in orange and the five radar locations as red triangles ('J' is Jabbeke, 'H' is Helchteren, 'A' is Avesnois, 'W' is Wideumont, and 'N' is Neuheilenbach). The circles indicate a distance of 100 km from the radar, and the grey shaded area indicates the maximum range of the radars based on a 250 km distance from the nearest radar.

We focus on three heavy rainfall events in 2021 with different rainfall characteristics (see Table 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 was caused by a relatively stationary low-pressure system west of the Netherlands that led to a continuation of frontal systems (predominantly occlusion fronts) passing over Belgium and the Netherlands, resulting in hours of rainfall with only short dry time spans in between. 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). This event was part of a low-pressure system above the north of France, which caused a convergence line between Belgium and the Netherlands that resulted in relatively stationary convective cells and high amounts of rainfall in a short time span (most of the rainfall fell between 1500 and 2000 UTC, and locally often within 1 hr). Finally, the July event, as already mentioned in Section 1, was a persistent mesoscale system that contained both stratiform and convective rainfall and resulted in almost continuous rainfall for days. 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.*, 2022; Kreienkamp *et al.*, 2021).

3.2 | Rainfall products

3.2.1 | Radar data

The RMI provides radar rainfall estimates from a composite of five C-band weather radars (Figure 3). The

TABLE 1 Overview of the three events in this study.

Start time event (UTC)	End time event (UTC)	Type[a]	Catchment-average rainfall sum (mm)				No. of runs verified[b]
			Vesdre	Demer	Geul	Dommel	
Jan 27, 2021, 2300	Jan 29, 2021, 0900	S	30.6	20.7	27.7	30.6	409
Jun 29, 2021, 1130	Jun 30, 2021, 1130	C	30.9	54.5	28.8	9.6	281
Jul 12, 2021, 2200	Jul 15, 2021, 2100	S/C	131.5	92.0	109.2	101.9	854

[a] C, convective; S, stratiform.
[b] The number of issue times (runs) that were part of the validation.

reflectivity measurements of the five radars are processed in four steps:

1 Non-meteorological echoes are removed with Doppler filtering, and clutter is identified and removed based on the vertical profile of reflectivity, image texture, a satellite cloud mask, and information from the dual polarization of two of the radars (Helchteren and Jabbeke).
2 Per radar, the reflectivity at the ground is estimated with an averaged vertical profile of reflectivity to extrapolate non-convective rainfall to the ground level, and by interpolating missing data and identifying the convective precipitation. This is followed by a conversion from reflectivity into rain rate using the Marshall–Palmer relationship (Marshall *et al.*, 1955), where low intensities (stratiform precipitation) and higher intensities (convective precipitation) are treated differently – (as done in Wagner *et al.*, 2012):

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

A maximum reflectivity of 55 dBZ (approximately 88 mm·hr^{-1}) is used to deal with hail and a minimum reflectivity of 7 dBZ is used as a rain–no rain threshold to filter out artefacts.
3 The rainfall rates of every radar are adjusted with a mean-field-bias factor that is based on the sums of the radar rainfall and rain-gauge measurements (from 152 gauges) over the past hour.
4 The resulting rainfall rates of all radars are combined with weights based on the distance of the grid cell to the radars. This range-weighted combination takes place with only the three closest radars during the months with predominantly convective precipitation (May through August). Finally, the composited rain rates are mean-field-bias adjusted (the whole domain at once), in a similar way as in step 3.

For more information about the product, we refer the reader to Goudenhoofdt and Delobbe (2016).

3.2.2 | NWP rainfall forecasts

The NWP rainfall forecasts are produced by 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 parametrizations of the ALARO model include the multiscale precipitation and cloud scheme "Modular Multiscale Microphysics and Transport" (3MT; Gerard *et al.*, 2009). The ALARO model is run operationally in a deterministic setup at 1.3 km resolution four times a day on a 548 × 548 domain covering Belgium, the Netherlands, and Luxembourg (Benelux), with a lead time of up to 48 hr. The forecast is available approximately 4 hr after the analysis time, and this is at present the best (also in terms of spatial and temporal resolution and update frequency) available NWP product provided by RMI. 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 an exceptionally high operational temporal resolution of 5 min (internal calculation time step is 45 s). This high-frequency precipitation output is produced over a smaller subdomain that covers the Belgian radar composite and the Benelux.

3.3 | Experimental set-up

3.3.1 | Verification metrics

This section describes the two verification metrics that were used to evaluate the performance of the blended forecasting product. The continuous ranked probability score (CRPS) takes the entire distribution of an ensemble forecast into account for comparison with the observations. It

uses the cdf of the probabilistic forecast, a curve, and the cdf of the observation, which is a single step function. The area between these two cdfs is a measure of the error of the probabilistic forecast, which is formulated in the CRPS as (given a lead time in the forecast)

$$\mathrm{CRPS} = \frac{1}{N_\mathrm{f}} \sum_{n=1}^{N_f} \int_{-\infty}^{+\infty} (P_{F_n}(x) - P_{O_n}(x))^2 \, dx. \tag{24}$$

In this equation, $P_{F_n}(x)$ and $P_{O_n}(x)$ are the forecast and observed non-exceedance probability, for the nth forecast (or the nth grid cell for spatial averaging) from a total of N_f forecasts. x represents the forecast and observed rainfall amount, which can be approximated numerically as an interval with variable step dx, depending on the rainfall sum per ensemble member. The decomposition into this stepwise function is explained in Hersbach (2000), which is the approach we follow here. The CRPS reduces to the mean absolute error for a deterministic forecast, which makes it applicable to and comparable for both deterministic and probabilistic forecasts. Nevertheless, one should acknowledge that issues such as the double penalty for high-resolution deterministic forecasts (e.g. Mittermaier and Csima, 2017) will cause CRPS to favour probabilistic forecasts.

The critical success index (CSI; Schaefer, 1990) is a threshold-based categorical score that is formulated as (given a lead time in the forecast)

$$\mathrm{CSI} = \frac{H}{H + M + \mathrm{FP}}, \tag{25}$$

where H is the number of hits where both forecast and observation exceed the threshold, M is the number of misses where only the observation exceeds the threshold, and FP is the number of false positives, where the forecast exceeds the threshold but the observation does not. In this study, the used thresholds to calculate the CSI were 1.0 and 5.0 $\mathrm{mm \cdot hr^{-1}}$. For ensemble forecasts, the individual contingency table members (H, M, and FP) are calculated per ensemble member and, finally, the CSI is based on the combined contingency table results of all members together.

3.3.2 | Evaluation of rainfall forecasts

To test the blending set-up as described in Section 2, we constructed blended forecasts (from here onwards referred to as STEPS blending) with 48 ensemble members and a 12 hr forecast horizon for every 5 min issue time during the three events (Section 3.1). The choice for 48 ensemble members originates from the operational ambition of the RMI. We used a forecast horizon of 12 hr to be able to distinguish the difference in predictability among the methods tested (which will be described hereafter), where we expect the radar-based nowcasting to have added value for the first hours, NWP for the longer lead times, and blending to have added value in the transition zone between those two (Lin *et al.*, 2005; Germann *et al.*, 2006). Note that the radar-based nowcasting will likely have no information to advect into the domain after the first hours of the forecast. The added stochastic perturbations to form the nowcast ensemble will somewhat expand this window, but eventually no rain will be forecast after a few hours of lead time (depending on the location and advection velocity of the rainfall fields). The choice to also use a 12 hr forecast horizon for the radar-based nowcasting is merely to identify the point of no predictability compared with the blended and NWP forecasts. In addition, as a logical follow-up would be to apply this work to flood forecasting, we have considered a sufficiently long forecast horizon to capture the response times of the selected catchments.

For the proposed blending approach, the radar QPE and NWP rainfall forecasts from Section 3.2 were used. The pysteps configuration used is given in Table 2. We benchmarked the results against (1) the deterministic NWP forecasts (Section 3.2.2), (2) radar-based ensemble nowcasts with 48 members, constructed with pysteps (v1.6.2) using the methods and forecast settings from Table 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 one to zero between a given start and end time, whereas it linearly increases the blending weight for the (deterministic) NWP component from zero to one. For this study, the start and end times of the linear blending were fixed at 1 hr and 3 hr respectively after the issue time, which is around the average skilful lead time of 2 hr for nowcasting for catchments in this region (Imhoff *et al.*, 2020). The radar QPE, which was also used to construct the (blended) nowcasts, was considered the observation in this study, and only data within the observed radar domain were used for comparison. Only the QPE inside the radar domain was used for comparison in the forecast validation. Note that a comparison between an ensemble forecast (the radar-based nowcasts and blended forecasts) and a deterministic forecast (the NWP forecast) is, strictly speaking, not fair. In this small test case to evaluate the proposed system, using the operational products of RMI, we are limited to this test set-up, but we recommend a more thorough future analysis using ensemble NWP forecasts and a larger sample of events.

In the evaluation of the STEPS blending approach and the comparison with the three aforementioned models, we

TABLE 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 and Zawadzki (2002)
Nowcasting method	STEPS	(Seed, 2003; Seed *et al.*, 2013; Bowler *et al.*, 2006), Pulkkinen *et al.* (2019)
Perturbations	Non-parametric	Seed *et al.* (2013)
Mask method	Incremental	Pulkkinen *et al.* (2019)
Probability matching	cdf	Foresti *et al.* (2016)
Blending module settings		
Climatological skill window length	3 days	
Weights method	BPS	Bowler *et al.* (2006)
Forecast settings		
Number of lead times	144 (12 hr)	
Number of ensemble members	48	
Precipitation threshold	0.1 mm·hr^{-1}	
Order of the AR(p) model	2	
Number of cascade levels	8	
Transformation	R to dBR	Equation (1)
Velocity perturbations	Turned off	

Abbreviations: BPS, Bowler–Pierce–Seed; cdf, cumulative distribution function; STEPS, short-term ensemble prediction system.

focused on two spatial scales: the radar-domain and the catchment scales. On 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. Per issue time, both metrics were calculated per grid cell, and, subsequently, averaged over all grid cells in the radar domain. On the catchment scale, the pointwise precipitation intensities and precise grid point localization becomes less relevant, but the accumulation over basins and spatio-temporal consistency becomes more relevant. On this scale, we validated the forecasts of the four models using the CRPS and CSI (for the results of the latter analysis, see the Supporting Information) 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.

3.3.3 | 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 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 4.2, we visualize the effects of these window sizes on the resulting climatological skill on all eight 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 (1330 UTC) during the June event (Table 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 4 and 5. This should give a first impression of the sensitivity of the blending approach to the moving window size choice.

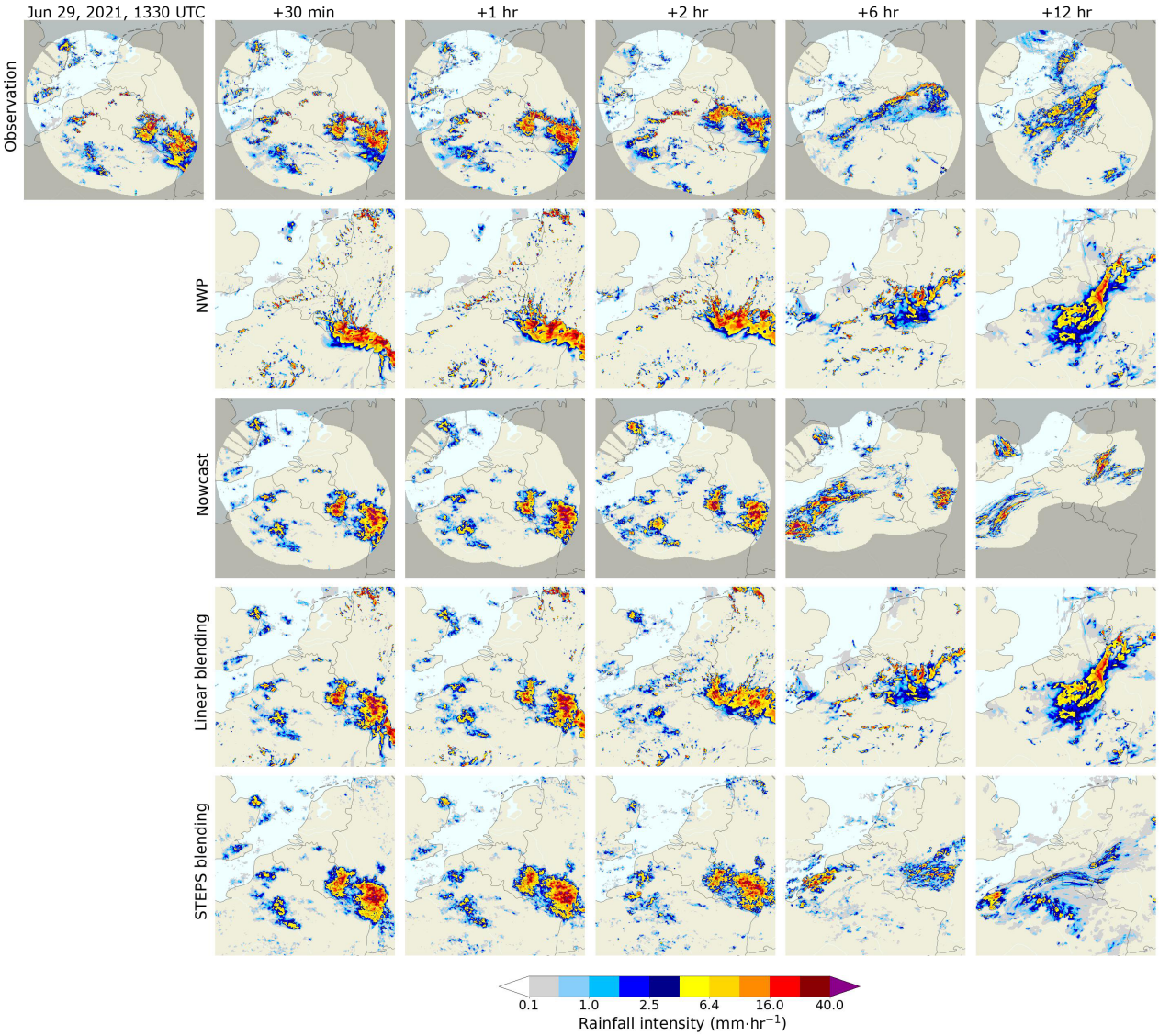

FIGURE 4 Domain-wide forecast rainfall fields for the test case of June 29, 2021, 1330 UTC (an example for the other events is illustrated in Supporting Information Figures S1 and S2). The top row illustrates the observed radar rainfall fields, and the rows below illustrate the forecast rainfall fields with the four methods tested (radar-based ensemble nowcasting, deterministic numerical weather prediction [NWP], linearly blended forecasts and Short-Term Ensemble Prediction System [STEPS]-blended forecasts) for five lead times. From the ensemble forecasts, only the first member is shown (the probabilities of exceeding a given threshold from these ensembles are illustrated in Supporting Information Figures S3–S5).

3.3.4 | 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 (1330 UTC) during the June event. The resulting weights for this issue time on all eight cascade levels are shown in Figure 2. In Section 4.3, we will discuss the effects of this approach on the rainfall forecast skill, both at the catchment and radar domain scales, for this forecast.

4 | RESULTS

4.1 | Evaluation of rainfall forecasts

4.1.1 | Example case of June 29, 2021

Before we discuss the statistics per event based on all forecasts, Figure 4 illustrates the rainfall forecasts of all methods tested for just one issue time, the test case of June 29, 2021, 1330 UTC. This day consisted of convective rainfall, with locally high-intensity rainfall, especially near the Vesdre, Demer, and Geul catchments, which is generally

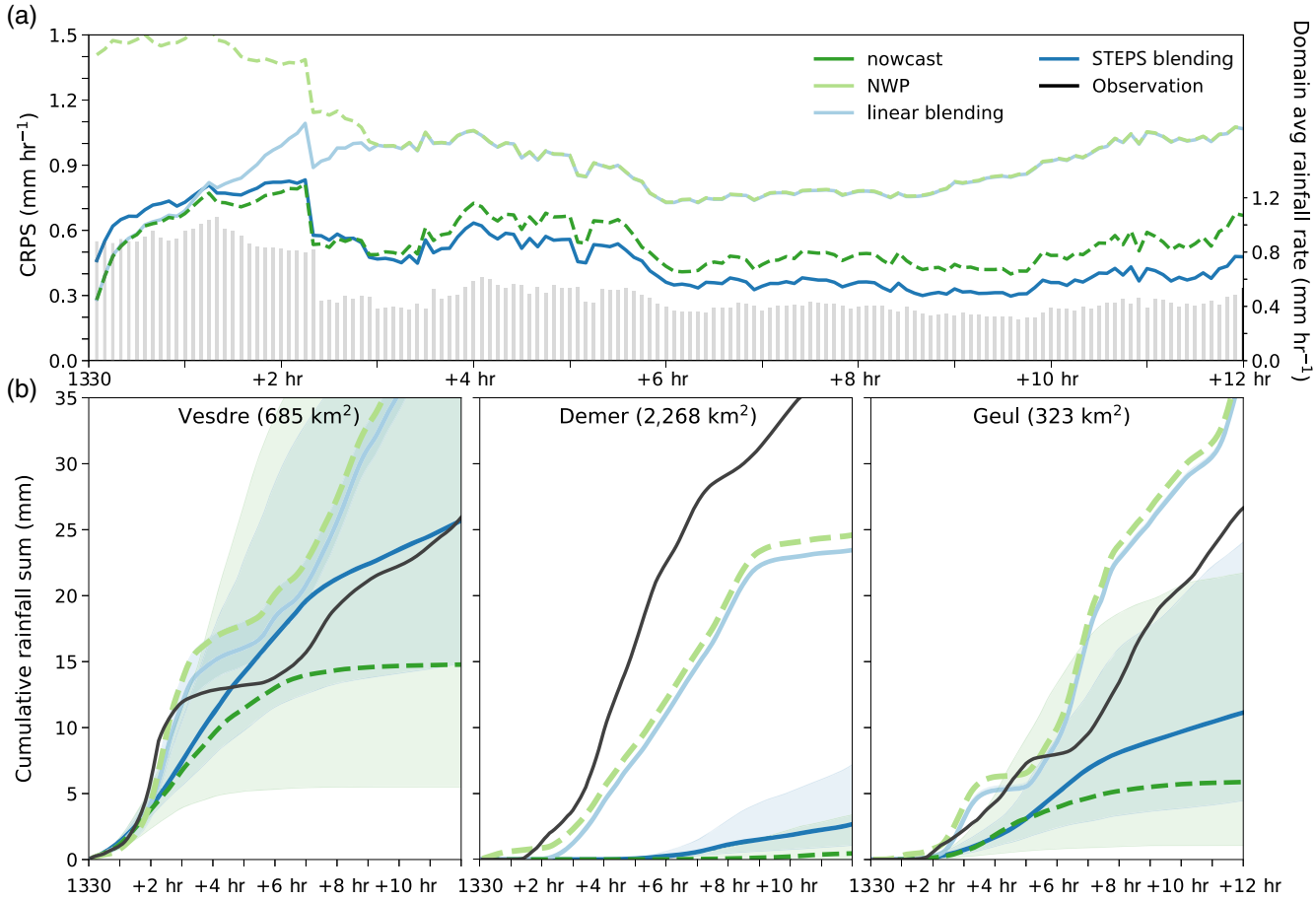

FIGURE 5 Evaluation of the four forecasting methods for the test case of June 29, 2021, 1330 UTC. (a) The continuous ranked probability score (CRPS) per lead time, averaged over all grid cells in the radar domain. The grey bars indicate the domain-averaged rainfall rates (mm·hr^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sums per catchment (Vesdre, Demer, and Geul) compared with the observations (the observed radar rainfall) in black. The thick coloured lines indicate the ensemble median, or the deterministic forecast (for numerical weather prediction [NWP], light green). The shaded areas around the ensemble medians indicate the interquartile range. STEPS: Short-Term Ensemble Prediction System.

challenging to forecast well with both nowcasting and NWP (e.g. Roberts and Lean, 2008; Berenguer *et al.*, 2012; Ayzel *et al.*, 2019). With the selection of this challenging event, we hope to more clearly visualize the differences between the introduced methods (for an example of the other two events and for the resulting ensemble probabilities, see Supporting Information Figures S1–S4). Up to at least the first hour ahead, the (radar-based ensemble) nowcast captures the location and intensity of the rainfall better than the NWP forecast does. For the longer lead times, for instance 6 hr or more, the radar-based nowcast loses skill, which is also in line with the maximum skilful lead time of nowcasting (Lin *et al.*, 2005; Ayzel *et al.*, 2019; Imhoff *et al.*, 2020). The linear blending approach resembles the radar-based nowcast during the first hour, after which the NWP forecast slowly gets more weight until 3 hr 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 hr lead time.

As the radar-based nowcast fails to capture the localization of rainfall and the rainfall tends to dissipate for the indicated lead times of 6 and 12 hr, the linear blending approach seems beneficial here. The same holds for the STEPS blending approach, which also adds perturbations to the NWP forecast. This can increase the ensemble spread throughout the forecast, especially compared with the linear blending approach (for more information about the ensemble spread, see Supporting Information Figures S6–S8). 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 hr lead times, due to the larger weight that the NWP is given for this forecast compared with the STEPS blending approach. However, as we

are only focusing on one ensemble member, this is not an entirely fair comparison. Therefore, in the subsequent sections, we will take the full ensemble into account.

Figure 5a takes the entire ensemble into account by showing the CRPS for the forecasts of Figure 4. Averaged over the entire radar domain, the radar-based 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 radar-based 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 5b), this effect becomes clear with cumulative rainfall sums that stagnate after a lead time of 6 hr and an overall underestimation of the rainfall amount by the radar-based ensemble nowcast for lead times of more than 2 hr (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 hr. At the same time, it underestimates the rainfall sum for the Demer, though less than the nowcast. The Dommel is not shown for this issue time because the observed rainfall was near zero and this was forecast well by all methods.

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 3 hr or more ahead (Figure 5a). 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 3.3.2). At the catchment scale (Figure 5b), the results are similar to the NWP forecast.

The domain-averaged CRPS of the STEPS blending forecast is lowest of all the methods tested for lead times of 3 hr or more (Figure 5a). During the first 2–3 hr of the forecast, the STEPS blending forecast has a somewhat higher CRPS than the radar-based 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 5b), the STEPS blending approach outperforms the radar-based 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 (and also varies per issue time; e.g., see Supporting Information Figure S9). For the Vesdre, STEPS blending clearly outperforms all other methods, whereas for the Demer the linear blending approach (and the NWP forecast) is 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 subsequent Sections 4.1.2 and 4.1.3 we focus on the event-averaged statistics for the four methods tested, based on all forecasts.

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 hr (the CRPS values are similar to the radar-based ensemble nowcast for the June and July events; Figure 6a–c). Only during the January event does the average CRPS of the radar-based ensemble nowcast exceed the CRPS of the NWP forecast at a lead time of approximately 6 hr. The linear blending forecasts have CRPS values similar to the radar-based 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 $mm \cdot hr^{-1}$, the CSI of the radar-based ensemble nowcast, linear blending, and STEPS blending forecasts are closer (Figure 6d–i) than for the CRPS (Figure 6a–c), which is also a result of the sensitivity of the CRPS score to the many zeroes in the radar domain and forecasts. Overall, the CSI of STEPS blending remains somewhat higher for longer lead times than of the radar-based 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 $mm \cdot hr^{-1}$ threshold between a lead time of 2.5 and 6 hr and 2–2.5 hr for the 5.0 $mm \cdot hr^{-1}$ threshold. This is also the point where the linear blending approach starts to outperform the radar-based ensemble nowcasts and sometimes also the STEPS blending approach (e.g., Figure 6d). 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, to a certain extent. Hence, overall at the radar-domain scale, STEPS blending outperforms the other methods or has at least a similar performance, especially when considering the ensemble verification metric CRPS.

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

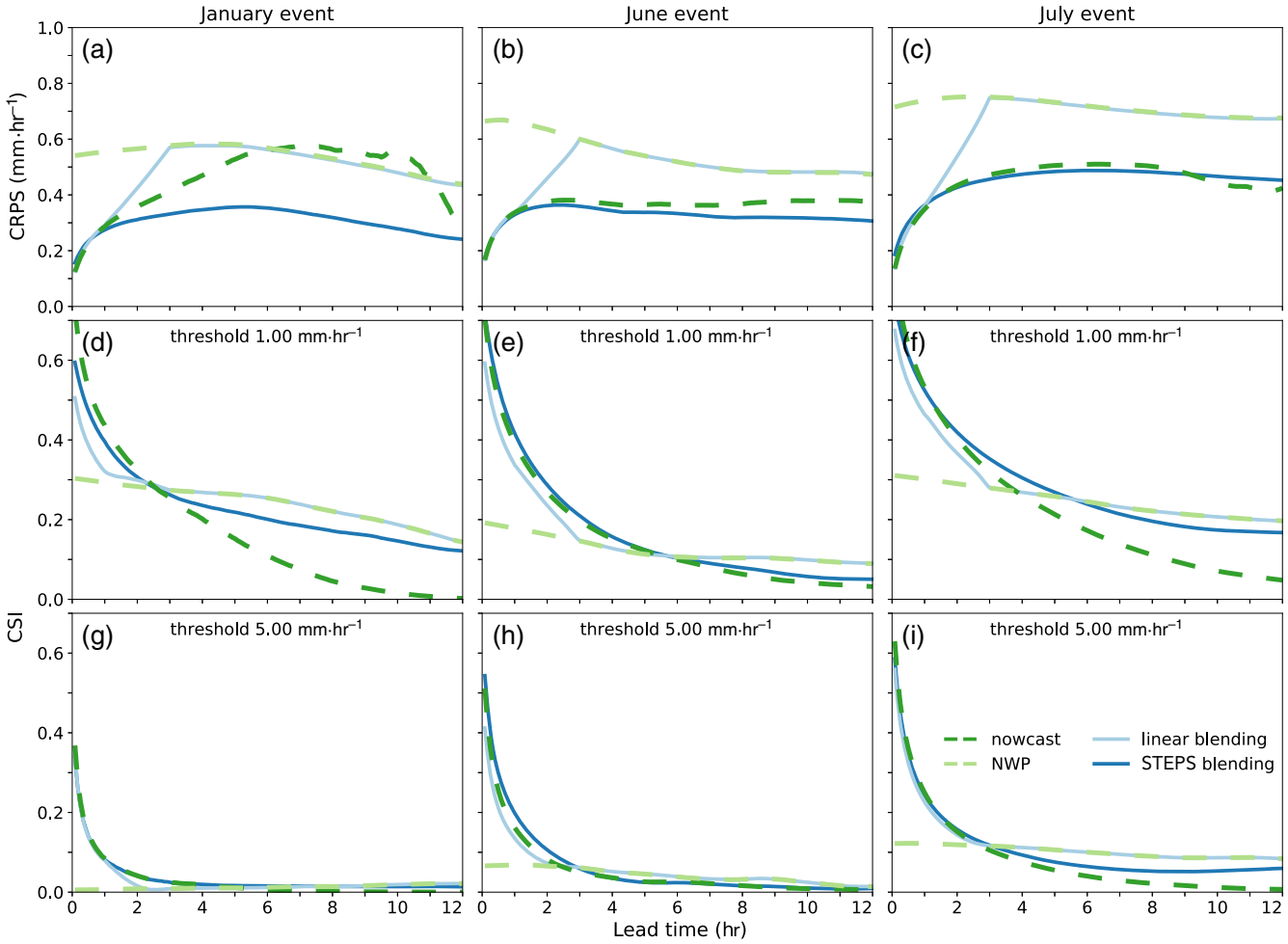

FIGURE 6 Event-averaged continuous ranked probability score (CRPS) and critical success index (CSI) per lead time for the four methods over the full radar domain. (a–c) The domain-averaged CSI per event. (d–f) The CSI per event for a threshold of 1 mm·hr^{-1}, and (g–i) for a threshold of 5 mm·hr^{-1}. NWP: numerical weather prediction; STEPS: Short-Term Ensemble Prediction System.

be attributed to both lower rainfall rates over the domain and a higher predictability of the event (Figure 7). STEPS blending generally attains lower CRPS values than the other methods for the January and July events. For the January event (Figure 7a–d), the radar-based ensemble nowcasts outperform the NWP forecast for at least 3–5 hr 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.*, 2019; Imhoff *et al.*, 2020). 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 skilful nowcasts than NWP for the entire forecast horizon (Figure 7i–l). Similar to Figure 6d–i, these differences become smaller when we look at the CSI (Supporting Information Figure S10), where both blended forecasts manage to follow the best-performing method for that lead time (nowcasting or NWP) and regularly have more skill (up to 1 hr more than nowcasting alone) in the transition region where NWP becomes more skilful than nowcasting (which always happens, in contrast to the CRPS-based analysis).

During the June event (Figure 7e–h), the NWP and linear blending forecasts outperform the nowcasts and STEPS blending for most lead times of more than 1–2 hr in catchments Vesdre and Demer. 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 7e–h). The radar-based ensemble nowcasts already become less skilful after 1–2 hr for these two catchments.

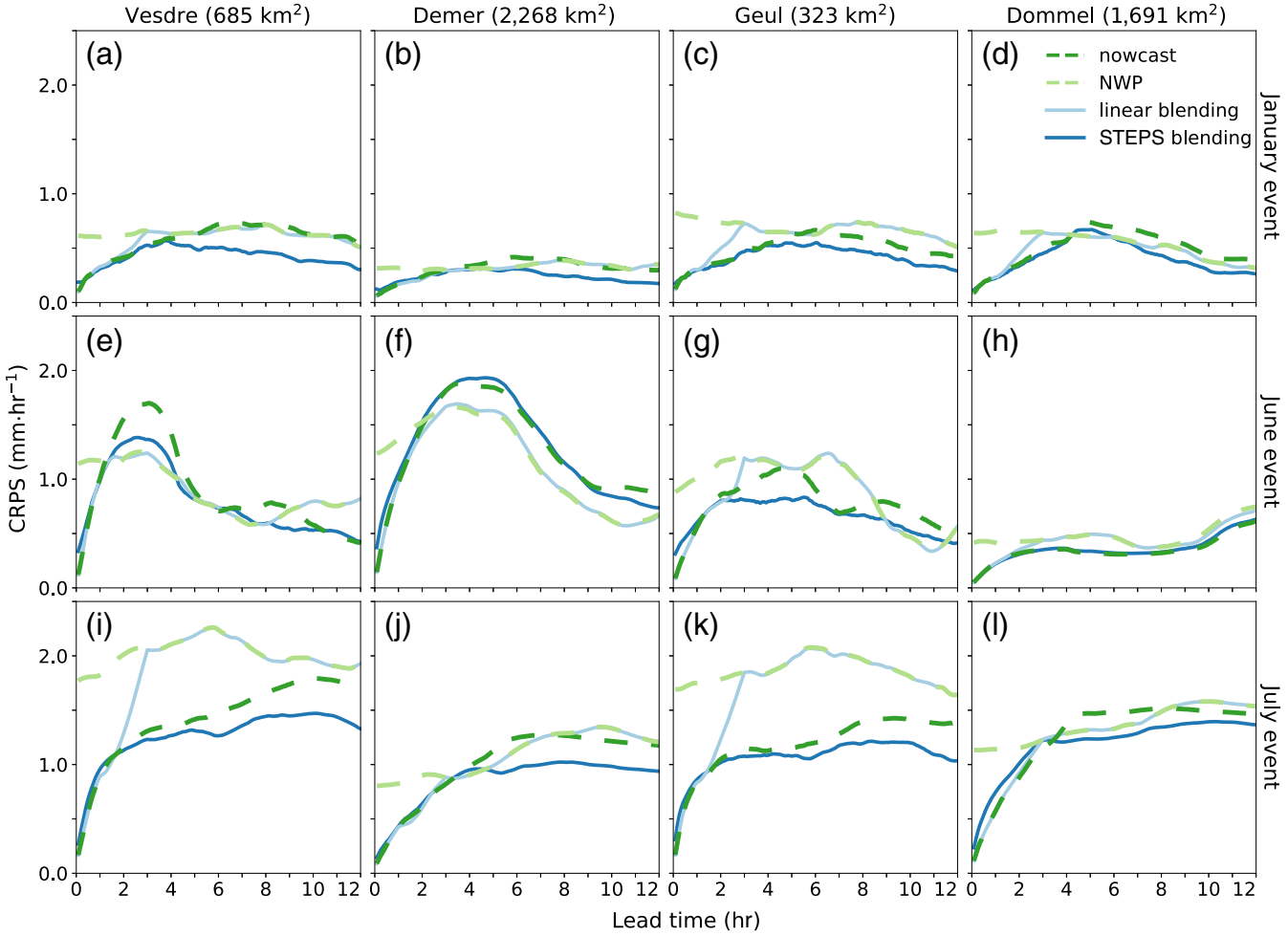

FIGURE 7 Event-averaged continuous ranked probability score (CRPS) of the catchment-averaged rainfall forecast for the four catchments and three events: (a)–(d) January; (e)–(h) June; (i)–(l) July. The radar-based ensemble nowcasting method is indicated in dark green, numerical weather prediction (NWP) in light green, the linear blending method in light blue, and the Short-Term Ensemble Prediction System (STEPS) blending method in dark blue.

The aforementioned view changes somewhat for the forecasts of the catchment-average cumulative rainfall volumes (Figure 8), which are relevant, for instance, for (flash) flood forecasts. The radar-based nowcasts often predict zero rainfall after lead times of approximately 6 h 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 Supporting Information Figure S13). The result is that the CRPS for the cumulative rainfall sum significantly increases after a lead time of 4–7 hr (Figure 8). The overall performance of STEPS blending compared with the other methods remains similar to that in Figure 7, 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 8a–d). For the July event, this is also visible, although STEPS blending still outperforms all other methods for the Vesdre and Geul. The CSIs for these events for higher rainfall thresholds (Supporting Information Figures S11 and S12) support the impression that, at the catchment scale, for cumulative rainfall sums, but also for higher rainfall intensities as indicated by the CSI, the nowcast quickly loses skill and STEPS blending tends to give too much weight to the nowcasts for too long, leading to a higher skill for NWP and linear blending for the longer lead times. However, note that, for high rainfall intensity thresholds, the NWP forecast has hardly any to no skill, whereas the nowcasts and blended forecasts still provide some skill for the first hours of the forecast (Figures 6 and S7).

To conclude, STEPS blending and linear blending match or even exceed the radar-based ensemble nowcasts' performance for the four catchments. Overall, STEPS blending outperforms the other methods for the months January and July, although the difference reduces

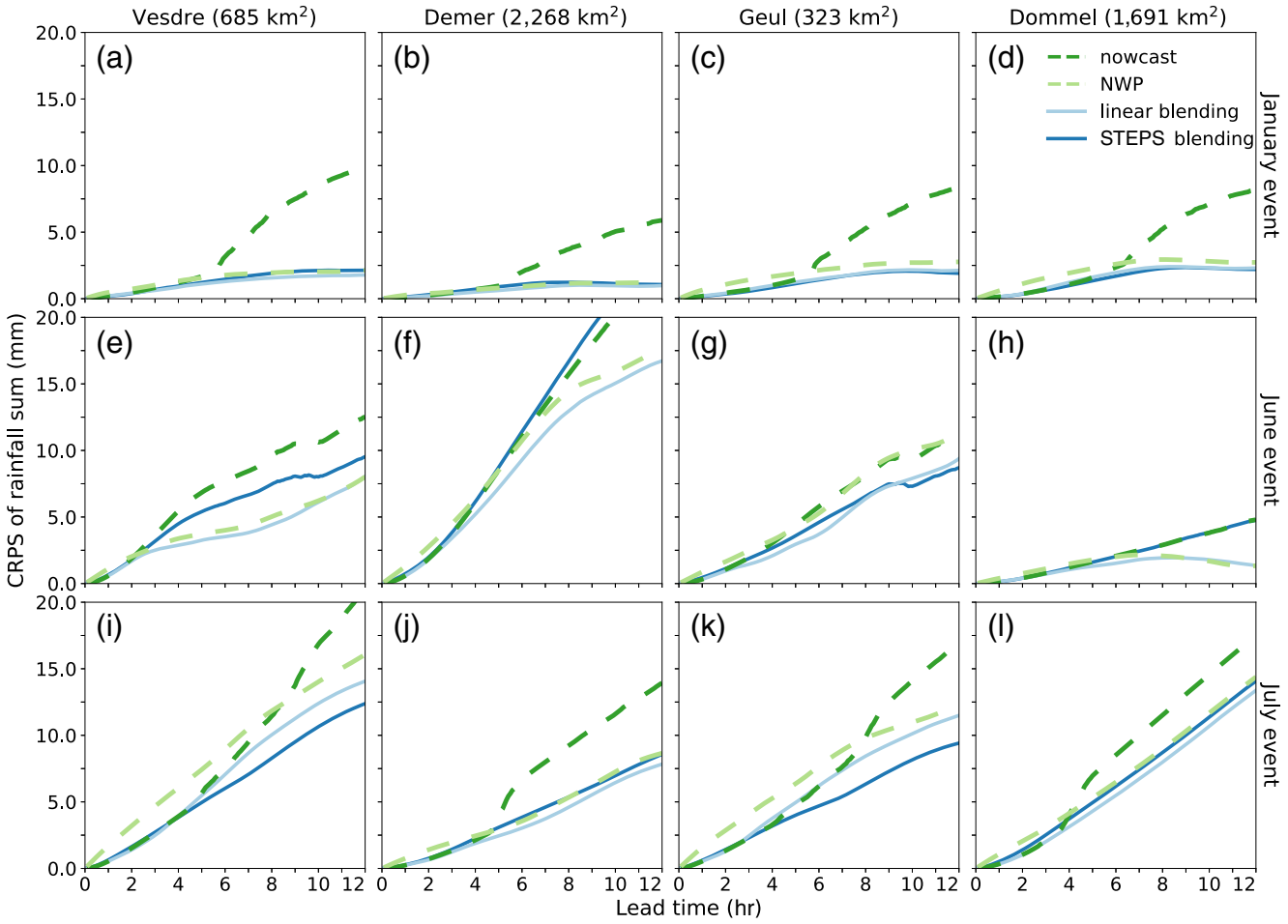

FIGURE 8 Event-averaged continuous ranked probability score (CRPS) of the cumulative catchment-averaged rainfall sum from the issue time until the indicated lead time. Shown are the event-averaged CRPS values for the four catchments and three events (a)–(d) January; (e)–(h) June; (i)–(l) July. The radar-based ensemble nowcasting method is indicated in dark green, numerical weather prediction (NWP) in light green, the linear blending method in light blue, and the Short-Term Ensemble Prediction System (STEPS) blending method in dark blue.

when we focus on the cumulative rainfall sums for the catchments instead of instantaneous rainfall rates, particularly with respect to the linear blending approach.

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 9). 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. A window size of a few days intuitively makes sense, since this is the typical persistence time of weather patterns (Neal *et al.*, 2016) that affect the atmospheric predictability and NWP model skill in the climate of the study area.

Compared with 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 9) and between seasons. For instance, at the largest spatial scale (level 0), the 21-day window mean skill

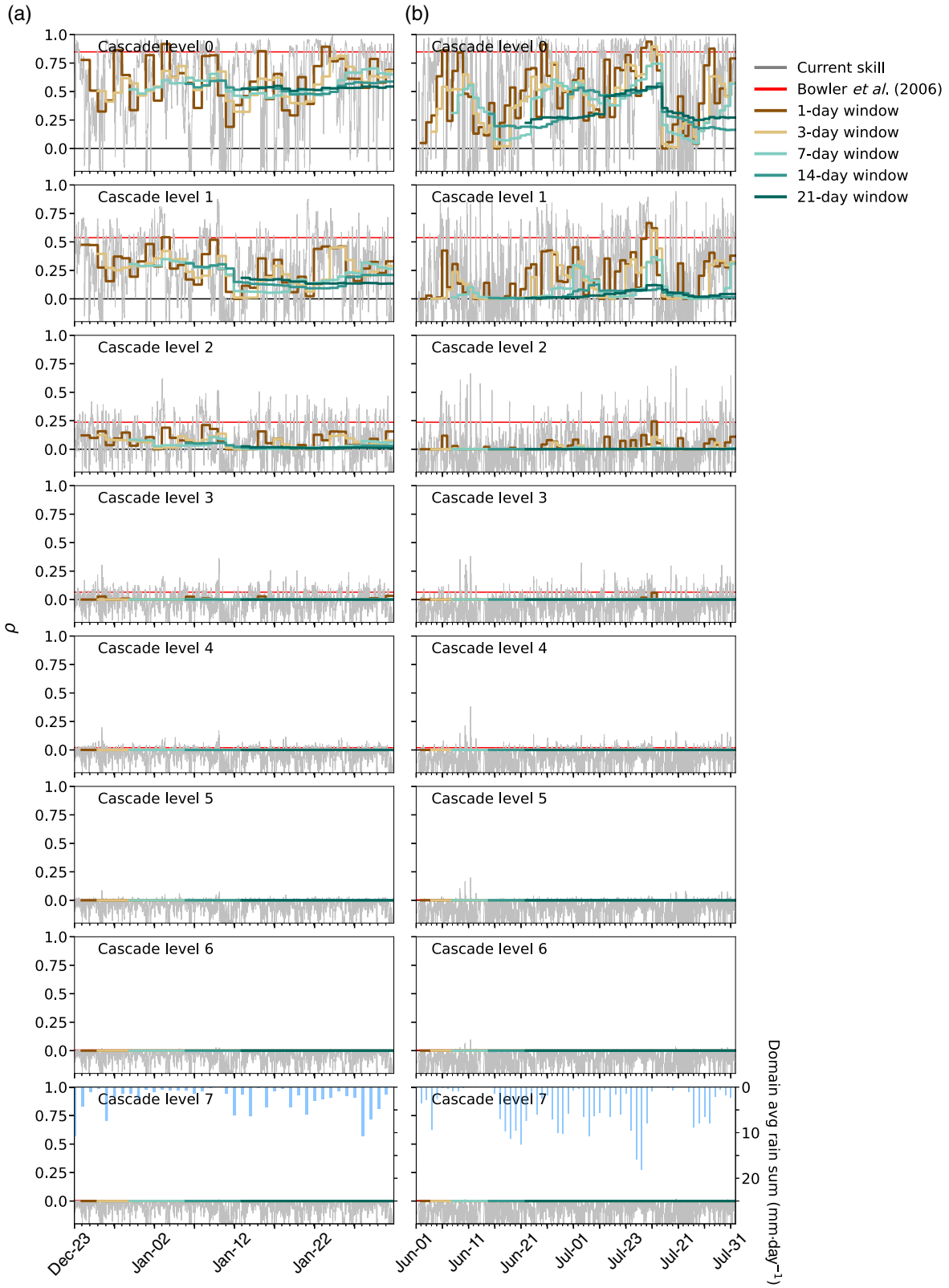

FIGURE 9 Climatological numerical weather prediction (NWP) skill (Pearson correlation) per cascade level as a function of moving window size for (a) January and (b) June–July. The grey lines indicate the NWP skill of the most recently available NWP forecast for that time step compared with the observed radar rainfall amount, the red lines indicate the climatological skill as provided by Bowler *et al.* (2006), and 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.

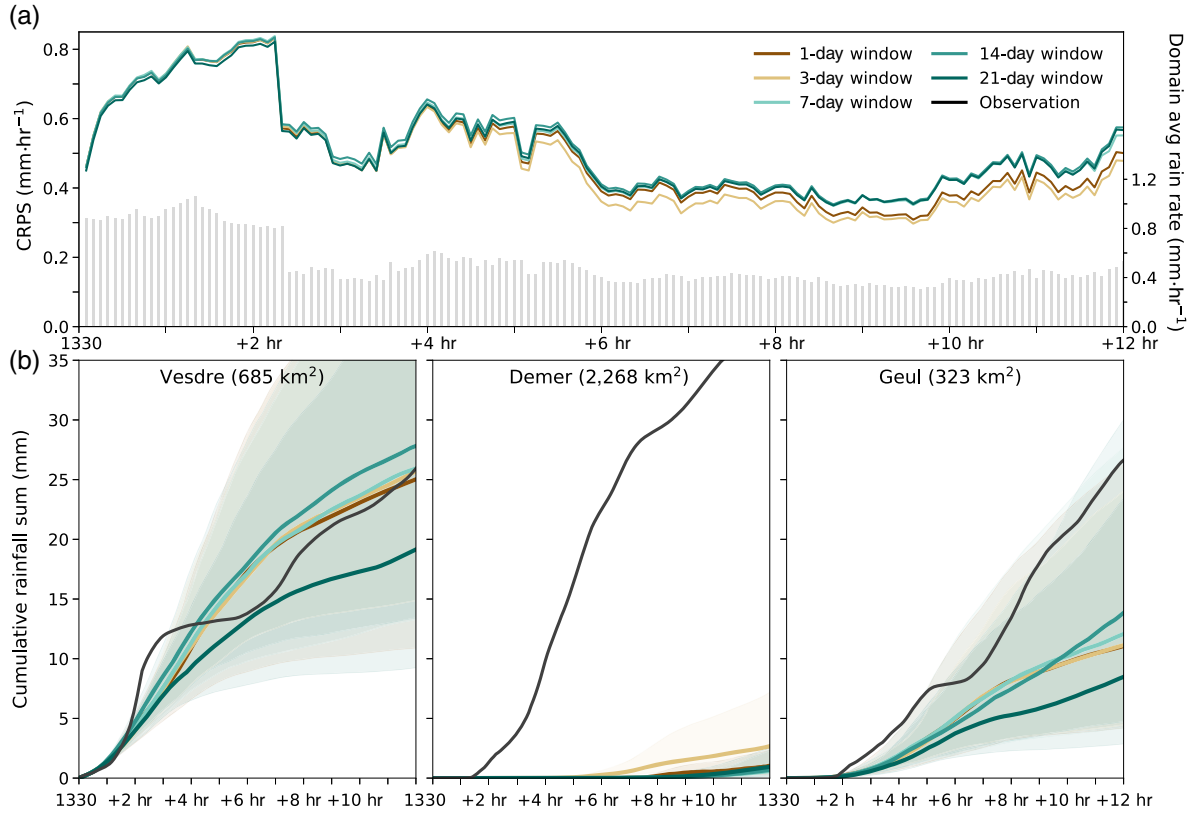

FIGURE 10 Effect of the climatological skill window sizes on the resulting blended rainfall forecasts for the test case of June 29, 2021, 1330 UTC. (a) The continuous ranked probability score (CRPS) per lead time, averaged over all grid cells in the radar domain. The grey bars indicate the domain-averaged rainfall rates (mm·hr^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sum per moving window size compared with the observation in black. The thick coloured lines indicate the ensemble median, and the shaded areas around it indicate the interquartile range.

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 moving window sizes tested (Figure 9), the effect on the rainfall forecasts for the test case of 2June 29, 2021, 1330 UTC is less pronounced (Figure 10). 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 10a), the difference in CRPS between the moving window sizes tested 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 4 hr of the forecast, but eventually become at most 0.09 mm·hr^{-1} between the rainfall forecast with a 3-day and 14-day window. At the catchment scale, differences are generally also minor, although the 21-day window underestimates the rainfall more than the other window sizes tested (for this case) for longer lead times, particularly for the Vesdre and Geul (Figure 10b).

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 3-day window used in this study. However, note that this is a result that is only based on one example forecast. We go into more detail on this topic in Section 5.

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

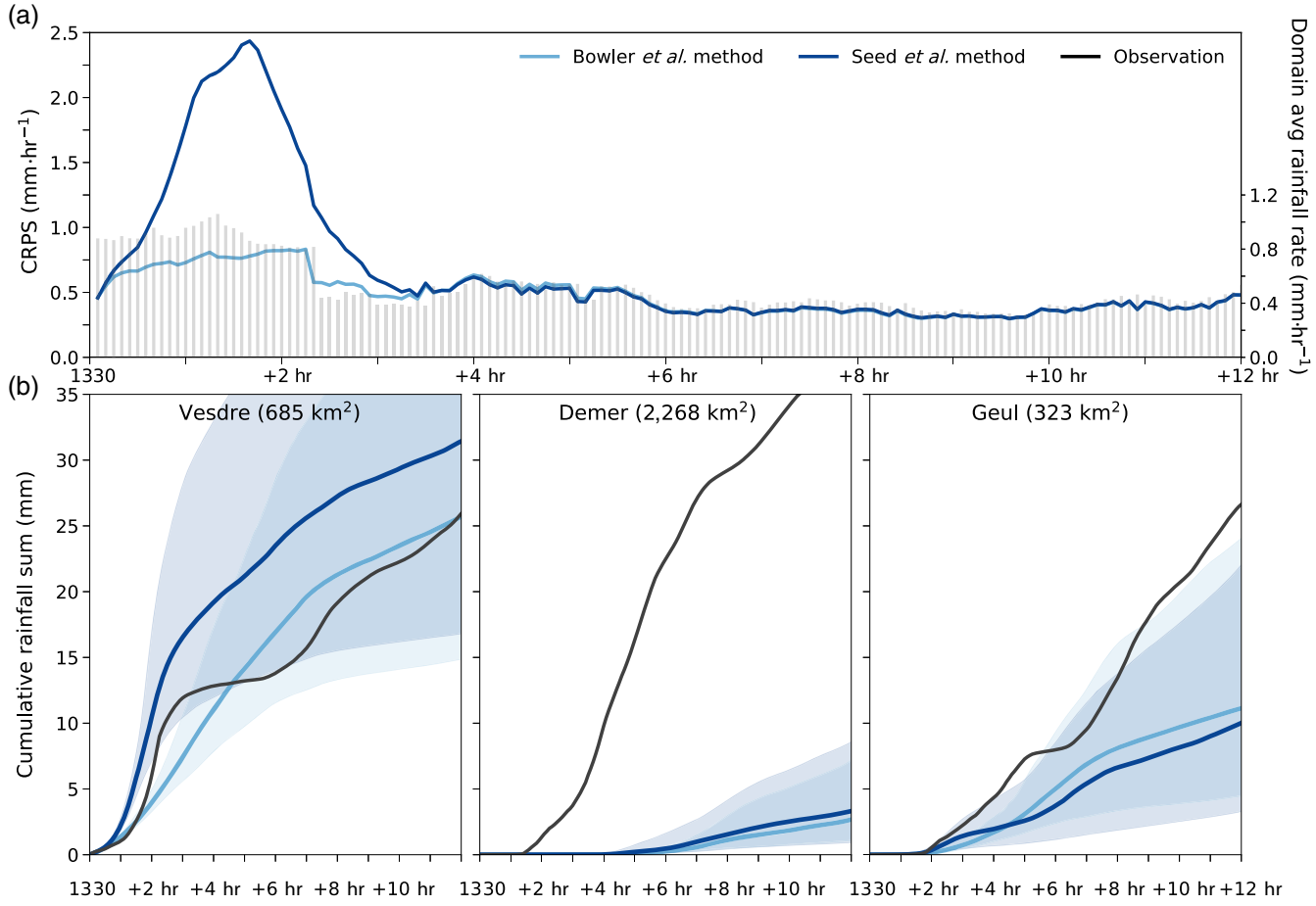

FIGURE 11 Effect of the two weights methods on the resulting blended rainfall forecasts for the test case of June 29, 2021, 1330 UTC. (a) The continuous ranked probability score (CRPS) per lead time, averaged over all grid cells in the radar domain. The method by Bowler *et al.* (2006) is illustrated in light blue, and the method by Seed *et al.* (2013) is in dark blue. The grey bars indicate the domain-averaged rainfall rates (mm·hr^{-1}) as observed during that lead time. (b) The forecast catchment-averaged cumulative rainfall sum for both methods compared with the observation in black. The thick coloured lines indicate the ensemble median, and the shaded areas around it indicate the interquartile range.

weights for the test case of June 29, 2021, 1330 UTC (Figure 11), 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. The reasons for the smaller differences in the resulting weights between the two methods are as follows: (1) the extrapolation component weight that exceeds 1.0 during the first and second hours of the forecast at cascade level 1 with the Seed *et al.* (2013) method (Figure 2b), though this weight is continuously decreasing for the Bowler *et al.* (2006) method (Figure 2a); and (2) the NWP component weights that exceed zero at cascade levels 2 and 3 with the Seed *et al.* (2013) method, though these weights remain zero for the Bowler *et al.* (2006) method.

Owing 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 3 hr of the forecast (Figure 11a). The maximum difference, which occurs at a lead time of 95 min, is 1.7 mm·hr^{-1}. After more than 3 hr, the differences reduces and becomes almost absent, which corresponds to the weights that have become relatively similar from that point onward (Figure 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 hr), and therefore the difference in weights during the first 3 hr 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 combined with the extrapolation and noise components. Note that this does not have to hold for other forecasts, and it is

therefore recommended to study the effect of the weights method more extensively in future research.

5 | DISCUSSION

In this study, we have described and evaluated the STEPS blending principle in the open-source Python library pysteps. We have added a few new functionalities that are indicated in Figure 1 and that are described in more detail in the subsections of Section 2. The STEPS blending approach was shown to provide rainfall forecasts that are generally as good as, and sometimes even outperform, the radar-based 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 and 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 or at high rainfall thresholds (Supporting Information Figures S11 and S12), 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.*, 2020). Sections 5.1 and 5.3 go into more detail about possible causes and solutions for this issue, but the climatological moving window size introduced (Sections 3.3.3 and 4.2) can play a role here, too, as this may not have been optimal for the individual events. This impression does not directly follow from the results in Figure 10, but this is only based on one issue time. A more extensive analysis of the role, strengths, and weaknesses of the climatological skill window is a recommendation for future work.

Furthermore, we should note that the simple benchmark linear blending could be optimized by making the blending start and end time variable and dependent on the skill of the different components. This would require incorporating some of the procedures introduced in STEPS blending, to allow for a blending that varies in time, while leaving out the spatial scale dependence. Doing so might be a good trade-off between the blending method introduced here and faster run times (by leaving out the cascade decomposition), especially considering the already satisfactory results for the linear blending method when we focus on the CSI metric (Sections 4.1.2 and 4.1.3).

5.1 | Bias towards the radar-based products

The approach presented 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 – that is, determining the "current skill" on a coarser spatial and temporal resolution – is that (minor) displacement errors are less penalized and that rainfall sums over a longer aggregation period become more relevant. As the focus on cumulative sums and some higher rainfall thresholds in this study has been advantageous for the NWP forecasts, compared with a focus on instantaneous rainfall rates, this might lead to higher weights for the NWP component(s). Moreover, current computation times (on four 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 and Lee, 1995; Creutin *et al.*, 1997; Gabella *et al.*, 2000; Sharif *et al.*, 2002; Uijlenhoet and 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 5.3). However, it is recommended to use a bias-adjusted radar QPE product to prevent the blending method being steered towards the systematic bias in the unadjusted radar QPE – (for an overview of adjustment methods, see Ochoa-Rodriguez *et al.*, 2019).

5.2 | Forecast timeliness for early warning

Core strengths of nowcasting are fast run times (in principle, a new forecast can be available within minutes after the last observation) and initial conditions that correspond to the latest observations. Compared with NWP, which had an update frequency of 6 hr in this study and is available approximately 4 hr after the analysis time, this can

provide crucial information for emergency managers and inhabitants during the first hours of an event. The narrow skilful forecast horizon of nowcasting limits its use to only the first hours of an event. We have introduced the blending procedure in this study to bridge the gap between nowcasting and NWP, which should extend skilful lead times and should still be available in a timely fashion. The run times of the blending approach introduced are, however, significantly longer than radar-based nowcasting only (Section 5.1). This process can, in the current implementation, already be significantly accelerated by including more CPU cores for parallelization, by decomposing the NWP forecast outside the blending procedure and by having the NWP forecast on the same grid as the radar data (or the other way around). In addition, the transition towards rapid-update-cycle NWP models (for instance, Benjamin *et al.*, 2016; Yussouf and Knopfmeier, 2019; Porson *et al.*, 2020; Turner *et al.*, 2022) increases the timeliness and skill of NWP model runs, and we expect that this could also enhance the blended forecast quality.

5.3 | 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, owing 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 scheme that was introduced in this study (Section 2.6). The current implementation advects the latest radar rainfall observations to lead time $t + t_1$, without any further perturbations and autoregression steps, combines this with the NWP forecast and uses this as the "observation", at $t + t_1$, 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, which occurred in the example test case for June 29, 2021, 1330 UTC with increasing lead time (see Supporting Information Figure S14). This is 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 prevent blending of misplaced rainfall fields. The scale-dependent blending method gives a first correction for errors such as misplaced rainfall fields, which are common in NWP forecasts. Nevertheless, the implementation of a phase correction could still help to process the NWP forecast prior to the blending procedure and, in that way, make better use of the information in the NWP forecast. Hence, we see this is a meaningful next implementation step in the pysteps blending module.

6 | CONCLUSIONS

Although the first few hours ahead (in the order of 6 hr) in rainfall forecasting are crucial – for example, for (flash) flood warnings – this time-scale is generally not sufficiently well captured by the rainfall forecasts of 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 a disadvantage that it quickly loses skill after approximately the first 2 hr of the forecast for individual radars. To extend the skilful 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 combined using weights that vary per spatial cascade level. We described the implementation details and some new functionalities, and we 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 48-member blended forecasts tested, we compared the results against the original deterministic NWP forecast, a 48-member radar-based ensemble nowcast with pysteps, and a simple (48-member) ensemble linear blending approach.

At the radar-domain scale, the STEPS blending approach implemented performs on par with or better than the other three methods tested, when focusing on the CRPS method, and generally manages to provide a smoother transition for the lead times (>2 hr) when the nowcast quickly loses skill. To a lesser extent, this also holds for higher intensity rainfall cells, where the difference between nowcasting, linear blending, and STEPS blending is less pronounced and STEPS blending occasionally tends to give too much weight to the nowcast. 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 hr 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 generally performs similarly or even better than 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 that 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 moving window sizes tested 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 and 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 favour the Bowler *et al.* (2006) weights method, but that is provided that only one deterministic NWP model is combined with the extrapolation and noise components. For multimodel ensembles, the Seed *et al.* (2013) method is recommended, as it takes into account the cross-correlation between the models, although negative weights or individual weights exceeding 1.0 can still occur. A further analysis of the strengths and weaknesses of both methods, and possible improvements such as Bayesian methods, is a recommendation for future research.

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 that we strongly support.

AUTHOR CONTRIBUTIONS

Ruben O. Imhoff: conceptualization; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft. **Lesley De Cruz:** conceptualization; dataCuration; funding acquisition; investigation; methodology; project administration; resources; software; writing – review and editing. **Wout Dewettinck:** formal analysis; investigation; software; writing – review and editing. **Claudia C. Brauer:** conceptualization; methodology; supervision; visualization; writing – review and editing. **Remko Uijlenhoet:** conceptualization; funding acquisition; methodology; resources; supervision; writing – review and editing. **Klaas-Jan van Heeringen:** conceptualization;

methodology; supervision; writing – review and editing. **Carlos Velasco-Forero:** methodology; software; writing – review and editing. **Daniele Nerini:** software; writing – review and editing. **Michiel Van Ginderachter:** software; writing – review and editing. **Albrecht H. Weerts:** conceptualization; data curation; funding acquisition; methodology; project administration; resources; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

We would like to thank Alan Seed for his time to discuss the STEPS implementation at the Bureau of Meteorology, the weights method, and his view on the new functionalities. In addition, we would like to thank Edouard Goudenhoofdt for answering all our questions with regard to the RMI radar QPE product.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Pysteps v1.6.2 is available via https://doi.org/10.5281/zenodo.6416615. A subset of the data is available as a data sample in the pysteps data github repository (https://github.com/pySTEPS/pysteps-data). The extended dataset is available upon request from RMI (info@meteo.be), if not used for commercial purposes. All property and intellectual rights are an exclusivity of the RMI and may not be transferred. The source of the data must be cited in each publication in which the data of the RMI have been used.

ORCID

Ruben O. Imhoff https://orcid.org/0000-0002-4096-3528
Lesley De Cruz https://orcid.org/0000-0003-4458-8953
Albrecht H. Weerts https://orcid.org/0000-0002-3249-8363

REFERENCES

AON. (2021) *Global catastrophe recap: July 2021. Tech. rep*. London, United Kingdom: AON. URL: http://thoughtleadership.aon.com/Documents/20211008_analytics%-if-july-global-recap.pdf.

Atencia, A., Rigo, T., Sairouni, A., Moré, J., Bech, J., Vilaclara, E., Cunillera, J., Llasat, M.C. and Garrote, L. (2010) Improving QPF by blending techniques at the meteorological service of Catalonia. *Natural Hazards and Earth System Sciences*, 10, 1443–1455.

Atencia, A. and Zawadzki, I. (2014) A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Monthly Weather Review*, 142, 4036–4052.

Atencia, A. and Zawadzki, I. (2015) A comparison of two techniques for generating nowcasting ensembles. Part II: Analogs selection and comparison of techniques. *Monthly Weather Review*, 143, 2890–2908.

Austin, P.M. (1987) Relation between measured radar reflectivity and surface rainfall. *Monthly Weather Review*, 115, 1053–1070.

Ayzel, G., Heistermann, M. and Winterrath, T. (2019) Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0.1). *Geoscientific Model Development*, 12, 1387–1402.

Bailey, M.E., Isaac, G.A., Gultepe, I., Heckman, I. and Reid, J. (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, 257–276.

Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Alexander, C.R., Smirnova, T.G., Olson, J.B., James, E.P., Dowell, D.C., Grell, G.A., Lin, H., Peckham, S.E., Smith, T.L., Moninger, W.R., Kenyon, J.S. and Manikin, G.S. (2016) A North American hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, 144, 1669–1694.

Berenguer, M., Corral, C., Sánchez-Diezma, R. and Sempere-Torres, D. (2005) Hydrological validation of a radar-based nowcasting technique. *Journal of Hydrometeorology*, 6, 532–549.

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

Berenguer, M., Surcel, M., Zawadzki, I., Xue, M. and Kong, F. (2012) The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. Part II: Intercomparison among numerical models and with nowcasting. *Monthly Weather Review*, 140, 2689–2705.

Bowler, N., Pierce, C.E. and Seed, A.W. (2004) STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. In: *Forecasting Research Technical Report 433*. Wallingford, United Kingdom: Met Office.

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

Bubnová, R., Hello, G., Bénard, P. and Geleyn, J.-F. (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, 515–535.

Cox, D., Hunt, J., Mason, P., Wheater, H., Wolf, P., Gupta, H., Sorooshian, S., Gao, X., Imam, B., Hus, K.-L., Bastidas, L., Li, J. and Mahani, S. (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, 1363–1371.

Creutin, J.D., Andrieu, H. and Faure, D. (1997) Use of a weather radar for the hydrology of a mountainous area. Part II: radar measurement validation. *Journal of Hydrology*, 193, 26–44.

Dixon, M. and Wiener, G. (1993) TITAN: thunderstorm identification, tracking, analysis, and nowcasting—A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, 10, 785–797.

Ferraris, L., Rudari, R. and Siccardi, F. (2002) The uncertainty in the prediction of flash floods in the northern Mediterranean environment. *Journal of Hydrometeorology*, 3, 714–727.

Foresti, L., Reyniers, M., Seed, A. and Delobbe, L. (2016) Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium. *Hydrology and Earth System Sciences*, 20, 505–527.

Foresti, L., Sideris, I.V., Nerini, D., Beusch, L. and Germann, U. (2019) Using a 10-year radar archive for nowcasting precipitation growth and decay: A probabilistic machine learning approach. *Weather and Forecasting*, 34, 1547–1569.

Foufoula-Georgiou, E. (1998) On scaling theories of space-time rainfall: some recent results and open problems. In: *Stochastic methods in hydrology, vol. 7 of Advanced series on statistical science & applied probability*. Singapore, Singapore: World Scientific, pp. 25–72.

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

Gangopadhyay, S., Clark, M., Werner, K., Brandon, D. and Rajagopalan, B. (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, 1192–1206.

Gerard, L., Piriou, J.-M., Brožková, R., Geleyn, J.-F. and Banciu, D. (2009) Cloud and precipitation parameterization in a Meso-Gamma-scale operational weather prediction model. *Monthly Weather Review*, 137, 3960–3977.

Germann, U., Berenguer, M., Sempere-Torres, D. and Zappa, M. (2009) REAL-Ensemble radar precipitation estimation for hydrology in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 135, 445–456.

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

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

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

Goudenhoofdt, E. and Delobbe, L. (2016) Generation and verification of rainfall estimates from 10-Yr volumetric weather radar measurements. *Journal of Hydrometeorology*, 17, 1223–1242.

Hamilton, J.D. (1994) *Time series analysis*. NJ, United States of America: Princeton University Press.

Han, L., Fu, S., Zhao, L., Zheng, Y., Wang, H. and Lin, Y. (2009) 3D convective storm identification, tracking, and forecasting — An enhanced TITAN algorithm. *Journal of Atmospheric and Oceanic Technology*, 26, 719–732.

Harris, D., Menabde, M., Seed, A. and Austin, G. (1996) Multifractal characterization of rain fields with a strong orographic influence. *Journal of Geophysical Research: Atmospheres*, 101, 26405–26414.

Hering, A.M., Morel, C., Galli, G., Senesi, S., Ambrosetti, P. and Boscacci, M. (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)*, 206–211. Visby, Sweden: Copernicus, Visby, Sweden.

Hersbach, H. (2000) Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15, 559–570.

Heuvelink, D., Berenguer, M., Brauer, C.C. and Uijlenhoet, R. (2020) Hydrological application of radar rainfall nowcasting in the Netherlands. *Environment International*, 136, 105431.

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, e2019WR026723.

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, e2021WR031591.

Joss, J. and Lee, R. (1995) The application of radar–gauge comparisons to operational precipitation profile corrections. *Journal of Applied Meteorology*, 34, 2612–2630.

Kober, K., Craig, G.C. and Keil, C. (2014) Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 140, 1179–1188.

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

Koks, E., Van Ginkel, K., Van Marle, M. and Lemnitzer, A. (2022) Brief Communication: Critical Infrastructure impacts of the 2021 mid-July western European flood event. *Natural Hazards and Earth System Sciences Discussions*, 22, 3831–3838.

Kreienkamp, F., Philip, S.Y., Tradowsky, J.S., Kew, S.F., Lorenz, P., Arrighi, J., Belleflamme, A., Bettmann, T., Caluwaerts, S., Chan, S.C., Ciavarella, A., De Cruz, L., de Vries, H., Demuth, N., Ferrone, A., Fischer, R.M., Fowler, H.J., Goergen, K., Heinrich, D., Henrichs, Y., Lenderink, G., Kaspar, F., Nilson, E., Otto, F.E.L., Ragone, F., Seneviratne, S.I., Singh, R.K., Skålevåg, A., Termonia, P., Thalheimer, L., van Aalst, M., Van den Bergh, J., Van de Vyver, H., Vannitsem, S., van Oldenborgh, G.J., Van Schaeybroeck, B., Vautard, R., Vonk, D. and Wanders, N. (2021) *Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021*. Oxford, United Kingdom: World Weather Atribution. URL: http://hdl.handle.net/1854/LU-8732135.

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

Liguori, S. and Rico-Ramirez, M.A. (2013) A practical approach to the assessment of probabilistic flow predictions. *Hydrological Processes*, 27, 18–32.

Lin, C., Vasic, S., Kilambi, A., Turner, B. and Zawadzki, I. (2005) Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, 32, L14801.

Lovejoy, S. and Schertzer, D. (1995) Multifractals and rain, ed. aw kundzewicz, 62-103. In: *New uncertainty concepts in hydrology and hydrological modelling*, Vol. 61. New York, NY: Cambridge University Press.

Lucas, B.D., Kanade, T., et al. (1981) An iterative image registration technique with an application to stereo vision. In: *DARPA image understanding workshop*. British Columbia, Canada: Vancouver, pp. 674–679.

Marsan, D., Schertzer, D. and Lovejoy, S. (1996) Causal space-time multifractal processes: Predictability and forecasting of rain fields. *Journal of Geophysical Research: Atmospheres*, 101, 26333–26346.

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

Mejsnar, J., Sokol, Z. and Minářová, J. (2018) Limits of precipitation nowcasting by extrapolation of radar reflectivity for warm season in Central Europe. *Atmospheric Research*, 213, 288–301.

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

Mittermaier, M.P. (2006) Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmospheric Science Letters*, 7, 36–42.

Mittermaier, M.P. and Csima, G. (2017) Ensemble versus deterministic performance at the kilometer scale. *Weather and Forecasting*, 32, 1697–1709.

Moreno, H.A., Vivoni, E.R. and Gochis, D.J. (2013) Limits to flood forecasting in the Colorado Front Range for two summer convection periods using radar nowcasting and a distributed hydrologic model. *Journal of Hydrometeorology*, 14, 1075–1097.

Neal, R., Fereday, D., Crocker, R. and Comer, R.E. (2016) A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, 23, 389–400.

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

Nerini, D., Foresti, L., Leuenberger, D., Robert, S. and Germann, U. (2019) A reduced-space ensemble Kalman filter approach for flow-dependent integration of radar extrapolation nowcasts and NWP precipitation ensembles. *Monthly Weather Review*, 147, 987–1006.

Ochoa-Rodriguez, S., Wang, L.-P., Willems, P. and Onof, C. (2019) A review of radar-rain gauge data merging methods and their potential for urban hydrological applications. *Water Resources Research*, 55, 6356–6391.

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

Pappenberger, F., Cloke, H.L., Parker, D.J., Wetterhall, F., Richardson, D.S. and Thielen, J. (2015) The monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, 51, 278–291.

Pegram, G.G. and Clothier, A.N. (2001) Downscaling rainfields in space and time, using the string of beads model in time series mode. *Hydrology and Earth System Sciences Discussions*, 5, 175–186.

Pierce, C., Bowler, N., Seed, A., Jones, A., Jones, D. and Moore, R. (2005) Use of a stochastic precipitation nowcast scheme for fluvial flood forecasting and warning. *Atmospheric Science Letters*, 6, 78–83.

Pierce, C., Seed, A., Ballard, S., Simonin, D. and Li, Z. (2012) Nowcasting. In: Bech, J. (Ed.) *Doppler radar observations–Weather radar, wind profiler, ionospheric radar, and other advanced applications*. London, United Kingdom: InTech. URL: http://www.intechopen.com/books/doppler-radar-observations-weather-radar-wind-profiler-ionospheric-radar-and-other-advanced-applications/nowcasting.

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

Porson, A.N., Carr, J.M., Hagelin, S., Darvell, R., North, R., Walters, D., Mylne, K.R., Mittermaier, M.P., Willington, S. and Macpherson, B. (2020) Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Quarterly Journal of the Royal Meteorological Society*, 146, 3245–3265.

Prakash, S., Mitra, A.K., Momin, I.M., Rajagopal, E.N., Milton, S.F. and Martin, G.M. (2016) Skill of short- to medium-range monsoon rainfall forecasts from two global models over India for hydro-meteorological applications. *Meteorological Applications*, 23, 574–586.

Pulkkinen, S., Chandrasekar, V. and Harri, A.-M. (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, 2773–2787.

Pulkkinen, S., Chandrasekar, V. and Niemi, T. (2021) Lagrangian integro-difference equation model for precipitation nowcasting. *Journal of Atmospheric and Oceanic Technology*, 38, 2125–2145.

Pulkkinen, S., Chandrasekar, V., von Lerber, A. and Harri, A.-M. (2020) Nowcasting of convective rainfall using volumetric radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11), 7845–7859.

Pulkkinen, S., Nerini, D., Pérez Hortal, A.A., Velasco-Forero, C., Seed, A., Germann, U. and Foresti, L. (2019) Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 12, 4185–4219.

Radchenko, P., Vasnev, A.L. and Wang, W. (2021) Too similar to combine? On negative weights in forecast combination. *International Journal of Forecasting*, 39(1), 18–38.

Radhakrishnan, C. and Chandrasekar, V. (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, 211–228.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A. and Mohamed, S. (2021) Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597, 672–677.

Roberts, N.M. and Lean, H.W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97.

Schaefer, J.T. (1990) The Critical Success Index as an indicator of warning skill. *Weather and Forecasting*, 5, 570–575.

Schertzer, D. and Lovejoy, S. (1987) Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *Journal of Geophysical Research: Atmospheres*, 92, 9693–9714.

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

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

Seed, A.W., Srikanthan, R. and Menabde, M. (1999) A space and time model for design storm rainfall. *Journal of Geophysical Research: Atmospheres*, 104, 31623–31630.

Serafin, R.J. and Wilson, J.W. (2000) Operational weather radar in the United States: Progress and opportunity. *Bulletin of the American Meteorological Society*, 81, 501–518.

Sharif, H.O., Ogden, F.L., Krajewski, W.F. and Xue, M. (2002) Numerical simulations of radar rainfall error propagation. *Water Resources Research*, 38, 15–1–15–14.

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

Simonin, D., Pierce, C., Roberts, N., Ballard, S.P. and Li, Z. (2017) Performance of Met Office hourly cycling NWP-based nowcasting for precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143, 2862–2873.

Sokol, Z., Mejsnar, J., Pop, L. and Bližňák, V. (2017) Probabilistic precipitation nowcasting based on an extrapolation of radar reflectivity and an ensemble approach. *Atmospheric Research*, 194, 245–257.

Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J., Joe, P., Barker, D.M., Li, P.-W., Golding, B., Xu, M. and Pinto, J. (2014) Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95, 409–426.

Termonia, P., Fischer, C., Bazile, E., Bouyssel, F., Brožková, R., Bénard, P., Bochenek, B., Degrauwe, D., Derková, M., El Khatib, R., Hamdi, R., Mašek, J., Pottier, P., Pristov, N., Seity, Y., Smolíková, P., Španiel, O., Tudor, M., Wang, Y., Wittmann, C. and Joly, A. (2018) The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geoscientific Model Development*, 11, 257–281.

Turner, D.D., Cutler, H., Shields, M., Hill, R., Hartman, B., Hu, Y., Lu, T. and Jeon, H. (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, E198–E211.

Uijlenhoet, R. and Berne, A. (2008) Stochastic simulation experiment to assess radar rainfall retrieval uncertainties associated with attenuation and its correction. *Hydrology and Earth System Sciences*, 12, 587–601.

UNISDR. (2002) Guidelines for reducing flood losses. *Tech. rep*., United Nations International Strategy for Disaster Reduction (UNISDR)., New York, NY. URL: https://www.undrr.org/publication/guidelines-reducing-flood-losses.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z.B., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Plenkovic, I.O., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K. and Ylhaisi, J. (2021) Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102, E681–E699.

Venugopal, V., Foufoula-Georgiou, E. and Sapozhnikov, V. (1999) Evidence of dynamic scaling in space-time rainfall. *Journal of Geophysical Research: Atmospheres*, 104, 31599–31610.

Vivoni, E.R., Entekhabi, D., Bras, R.L., Ivanov, V.Y., Van Horne, M.P., Grassotti, C. and Hoffman, R.N. (2006) Extending the predictability of hydrometeorological flood events using radar rainfall nowcasting. *Journal of Hydrometeorology*, 7, 660–677.

Vivoni, E.R., Entekhabi, D. and Hoffman, R.N. (2007) Error propagation of radar rainfall nowcasting fields through a fully distributed flood forecasting model. *Journal of Applied Meteorology and Climatology*, 46, 932–940.

Wagner, A., Seltmann, J. and Kunstmann, H. (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, 4101–4117.

Yoon, S.-S. (2019) Adaptive blending method of radar-based and numerical weather prediction QPFs for urban flood forecasting. *Remote Sensing*, 11, 642.

Yussouf, N. and Knopfmeier, K.H. (2019) Application of the Warn-on-Forecast system for flash-flood-producing heavy convective rainfall events. *Quarterly Journal of the Royal Meteorological Society*, 145, 2385–2403.

Zou, X., Dai, Q., Wu, K., Yang, Q. and Zhang, S. (2020) An empirical ensemble rainfall nowcasting model using multi-scaled analogues. *Natural Hazards*, 103, 165–188.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Imhoff, R.O., De Cruz, L., Dewettinck, W., Brauer, C.C., Uijlenhoet, R., van Heeringen, K.-J. *et al*. (2023) Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library. *Quarterly Journal of the Royal Meteorological Society*, 1–30. Available from: https://doi.org/10.1002/qj.4461