Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium

L. Foresti[1], M. Reyniers[1], A. Seed[2] and L. Delobbe[1]

[1]{Royal Meteorological Institute of Belgium, Brussels, Belgium}

[2]{Bureau of Meteorology, Centre for Australian Weather and Climate Research, Melbourne, Australia}

Correspondence to: L. Foresti (loris.foresti@gmail.com)

Abstract

The Short-Term Ensemble Prediction System (STEPS) is implemented in real-time at the Royal Meteorological Institute (RMI) of Belgium. The main idea behind STEPS is to quantify the forecast uncertainty by adding stochastic perturbations to the deterministic Lagrangian extrapolation of radar images. The stochastic perturbations are designed to account for the unpredictable precipitation growth and decay processes and to reproduce the dynamic scaling of precipitation fields, i.e. the observation that large scale rainfall structures are more persistent and predictable than small scale convective cells. This paper presents the development, adaptation and verification of the system STEPS for Belgium (STEPS-BE). STEPS-BE provides in real-time 20 member ensemble precipitation nowcasts at 1 km and 5 min resolution up to 2 hours lead time using a 4 C-band radar composite as input. In the context of the PLURISK project, STEPS forecasts were generated to be used as input in sewer system hydraulic models for nowcasting urban inundations in the cities of Ghent and Leuven. Comprehensive forecast verification was performed in order to detect systematic biases over the given urban areas and to analyze the reliability of probabilistic forecasts for a set of case studies in 2013 and 2014. The forecast biases over the cities of Leuven and Ghent were found to be small, which is encouraging for future integration of STEPS nowcasts into the hydraulic models. Probabilistic forecasts of exceeding 0.5 mm hr^{-1} are reliable up to 60-90 min lead time, while the ones of exceeding 5.0 mm hr^{-1} are only reliable up to 30 min. The STEPS ensembles are slightly under-dispersive and represent only 75-90% of the forecast errors.

1 Introduction

The use of radar measurements for urban hydrological applications has substantially increased during the last years (e.g. Berne et al., 2004; Einfalt et al., 2004; Bruni et al., 2015). Given the fast response time of urban catchments and sewer systems, radar-based very short-term precipitation forecasting (nowcasting) has potential to extend the lead time of hydrological and hydraulic flow predictions.

Nowcasting concerns the accurate description of the current weather situation together with very-short term forecasts obtained by extrapolating the real-time observations. Quantitative precipitation nowcasting (QPN) is traditionally done by estimating the apparent movement of radar precipitation fields using optical flow or variational echo tracking techniques and extrapolating the last observed precipitation field into the future (e.g. Germann and Zawadzki, 2002; Bowler et al., 2004a). During recent years there has been significant progress in NWP modelling with radar data assimilation techniques (see a review in Sun et al., 2014), which reduces the useful lead time of extrapolation-based nowcasts compared with NWP forecasts. The development of seamless forecasting systems that optimally blend the extrapolation nowcast with the output of NWP models makes the definition of the nowcasting time range even fuzzier (see e.g. Pierce et al., 2010).

Due to the lack of predictability of rainfall growth and decay processes at small spatial scales (Radhakrishna et al., 2012), it is very important to provide together with a forecast an estimation of its uncertainty. The established method to represent the forecast uncertainty in Numerical Weather Prediction (NWP) is to generate an ensemble of forecasts by perturbing the initial conditions of the model in the directions exhibiting the largest error growth, which amplify more the spread of the obtained ensemble. However, in the nowcasting range the computation of large NWP ensembles (50-100 members) that resolve features at the scales of 1 km and are updated every 5 min is still impossible to achieve. Consequently, the efforts in nowcasting research have recently focused on developing heuristic techniques for probabilistic precipitation nowcasting, which was the topic of the *Heuristic Probabilistic Forecasting Workshop* that was organized in Munich, Germany (Foresti et al., 2014).

Probabilistic QPN methods can be categorized into three main classes: analogue, local Lagrangian and stochastic approaches. The analogue-based approach derives the forecast probability density function (p.d.f.) by retrieving a set of similar situations from an archive of precipitation events (Panziera et al., 2011; Foresti et al., 2015), the local Lagrangian approach

derives the p.d.f. by collecting the precipitation values in a neighborhood of a given grid point in Lagrangian frame of reference (Hohti et al., 2000; Germann and Zawadzki, 2004) and the stochastic approach exploits a random number generator to compute an ensemble of equally likely precipitation fields, for example by adding stochastic perturbations to a deterministic extrapolation nowcast (Pegram and Clothier, 2001a, 2001b; Bowler et al., 2006; Metta et al., 2009; Berenguer et al., 2011; Seed et al., 2013; Atencia and Zawadzki, 2014; Dai et al., 2015). The stochastic approach is also extensively used to produce ensembles of precipitation fields that characterize the radar measurement uncertainty (e.g. Jordan et al., 2003; Germann et al., 2009) and for design storm studies (e.g. Willems, 2001a; Paschalis et al., 2013).

Uncertainty quantification is nowadays an integral part of both weather and hydrological forecasting (Pappenberger and Beven, 2006). Not surprisingly, an important part of hydro-meteorological research aims at understanding how to propagate the uncertainty of precipitation observations and forecasts into the hydrological models (e.g. Willems, 2001b; Cloke and Pappenberger, 2009; Collier, 2009, Zappa et al., 2010).

Several studies already analyzed the value of deterministic nowcasting systems for catchment hydrology (e.g. Berenguer et al., 2005) and for better control of urban drainage systems (e.g. Achleitner et al., 2009; Verworn et al., 2009; Thorndahl and Rasmussen, 2013). Since an important fraction of the uncertainty of hydrological predictions is due to the uncertainty of the input rainfall observations and forecasts, radar-based ensemble nowcasting systems are increasingly used as inputs for flood and sewer system modeling (e.g. Ehret et al., 2008; Silvestro and Rebora, 2012; Silvestro et al., 2013; Xuan et al., 2009; Xuan et al., 2014). At longer forecast ranges, the NWP ensembles are also exploited for uncertainty propagation into hydrological models (see Roulin and Vannitsem, 2005; Thielen et al., 2009; Schellekens et al., 2011).

The Short-Term Ensemble Prediction System (STEPS) is a probabilistic nowcasting system developed at the Australian Bureau of Meteorology and the UK MetOffice (see the series of papers Seed, 2003; Bowler et al., 2006; Seed et al., 2013). STEPS is operationally used at both weather services and provides short-term ensemble precipitation forecasts using both the extrapolation of radar images and the downscaled precipitation output of NWP models. The main idea behind STEPS is to represent the uncertainty due to the unpredictable precipitation growth and decay processes by adding stochastic perturbations to the deterministic extrapolation of radar images. The stochastic perturbations are designed to represent the scale-

dependence of the predictability of precipitation and to reproduce the correct spatio-temporal correlation and growth of the forecast errors.

One of the first applications of STEPS in hydrology is presented in Pierce et al. (2005), who used the STEPS ensemble nowcasts to quantify the accuracy of flow predictions in a medium-sized catchment in UK. The value of STEPS nowcasts for urban hydrology was extensively analyzed by Liguori and Rico-Ramirez, 2012; Liguori et al., 2012; Liguori and Rico-Ramirez, 2013; Xuan et al., 2014). Liguori and Rico-Ramirez (2012) concluded that the performance of the radar-based extrapolation nowcast can be improved after 1 hour lead time if blended with the output of a NWP model. They also found that, according to the Receiver Operating Characteristic (ROC) curve, the probabilistic nowcasts have more discrimination power than the deterministic ones. Liguori et al. (2012) integrated STEPS nowcasts as inputs into sewer system hydraulic models in an urban catchment in Yorkshire (UK). They concluded that the blending of radar and NWP forecasts has potential to increase the lead time of flow predictions, but is strongly limited by the low accuracy of the NWP model in forecasting small scale features. Liguori and Rico-Ramirez (2013) performed a detailed verification of the accuracy of flow predictions and concluded that the STEPS ensembles provide similar performance than using a deterministic STEPS control forecast, but the ensembles lead to a slight underestimation of the flow predictions. Xuan et al. (2014) used ensemble STEPS nowcasts as inputs in a lumped hydrological model for a medium-sized catchment in the South-West of UK. The hydrological model calibrated with rain gauges had lower RMSE than the one using radar data, but the ability of STEPS in accounting for the forecast uncertainty was useful to capture some of the high flow peaks and extending the forecast lead time. However, the conclusions of the previous studies are strongly affected by the limited number of flood events analyzed. An extensive review of the usage of precipitation forecast systems for operational hydrological predictions in UK from very-short to long range (including STEPS) is provided in Lewis et al. (2015).

The goal of this paper is to present the development and verification of the STEPS system at the Royal Meteorological Institute of Belgium (RMI), referred to as STEPS-BE. STEPS-BE provides in real-time 20 member ensemble precipitation nowcasts at 1 km and 5 min resolutions up to 2 hours lead time on a 512x512 kilometer domain using the Belgian 4 C-band radar composite as input. It was developed in the framework of the Belspo project PLURISK for better management of rainfall-induced risks in the urban environment. With

respect to the original implementation of STEPS (Bowler et al., 2006), STEPS-BE includes two main improvements, which are designed to generate better STEPS nowcasts without NWP blending. The first one is related to the optical flow algorithm, which is extended with a kernel-based interpolation method to obtain smoother velocity fields. The second one concerns the generation of stochastic noise only within the advected radar composite. While the verification of STEPS nowcasts with NWP blending has already been extensive (Bowler et al., 2006; Seed et al., 2013), this paper analyzes the accuracy of STEPS ensemble nowcasts without NWP blending in the 0-2 hours forecasting range.

Ensemble STEPS nowcasts are computed for a set of sewer overflow cases that affected the cities of Leuven and Ghent in 2013 and 2014. The accuracy of the ensemble mean forecast is verified using both continuous verification scores (multiplicative bias, RMSE) and categorical scores derived from the contingency table (probability of detection, false alarm ratio and Gilbert skill score). However, the most interesting part of this paper is the probabilistic and ensemble verification of STEPS nowcasts using both stratiform and convective rainfall events. Probabilistic nowcasts are verified using reliability diagrams and ROC curves. On the other hand, the dispersion of the nowcast ensembles is verified using rank histograms and by comparing the ensemble spread to the error of the ensemble mean.

The paper is structured as follows. Section 2 presents the radar data processing and case studies that are used to generate and verify the STEPS forecasts. Section 3 describes the nowcasting system STEPS, its extension and local implementation for Belgium (STEPS-BE). Section 4 illustrates the forecast verification results. Section 5 concludes the paper and discusses future perspectives.

2 Radar data and precipitation case studies

STEPS-BE integrates as input a composite image produced from the C-band radars of Wideumont (RMI, single-pol), Zaventem (Belgocontrol, single-pol), Jabbeke (RMI, dual-pol) and Avesnois (Meteo-France, dual-pol). The composite is produced on a 500 m resolution grid by combining single-radar pseudo Constant Altitude Plan Position Indicators (CAPPI) at a height of 1500 m.a.s.l.. The compositing algorithm takes the maximum reflectivity value from each radar at each grid point.

The radars have different hardware, scanning strategies and are operated by different agencies (RMI, Belgocontrol and Meteo-France), which inevitably leads to differences in the data processing. The Wideumont and Zaventem radars eliminate the non-meteorological echoes using standard Doppler filtering. The Jabbeke radar includes an additional clutter filtering which uses a fuzzy logic algorithm based on the dual-polarization moments (essentially the co-polar correlation coefficient, the texture of the differential reflectivity and the texture of the specific differential phase shift). A static ground clutter map and a statistical filter are used by Meteo-France to remove the non-meteorological echoes of the Avensois radar. The French radar data processing chain is described in Tabary (2007) and in Figueras i Ventura and Tabary (2013).

Since the Zaventem radar is mainly used for aviation applications, its scanning strategy is optimized for the measurement of winds. Except for the lowest elevation scan, a dual PRF mode (1200/800 Hz) is used. The azimuths that are scanned with a high PRF (1200 Hz) only have a maximum range of 125 km and are more affected by the second trip echoes caused by convective cells located beyond the 125 km range.

All radars use the standard Marshall-Palmer relationship $Z=200R^{1.6}$ to convert the measured reflectivity to rainfall rate. A composite image with more advanced radar-based quantitative precipitation estimation (QPE), that includes better ground clutter removal algorithms and also a correction for the bright band, was recently developed and the verification of the new product is ongoing.

STEPS forecasts were generated and verified for a set of sewer system overflow cases that affected the cities of Ghent and Leuven (see Table 1). The Ghent cases have a more stratiform character and occurred in late autumn and winter. On the other hand, the Leuven cases are more convective and occurred in summer months. A detailed climatology of convective storms in Belgium can be found in Goudenhoofdt and Delobbe (2009).

3 Short-Term Ensemble Prediction System (STEPS)

3.1 STEPS description

The Short-Term Ensemble Prediction System (STEPS) was jointly developed by the Australian Bureau of Meteorology (BOM) and the UK MetOffice (Bowler et al., 2006).

STEPS forecasts are produced operationally at both weather services and are distributed to weather forecasters and a number of external users, in particular the hydrological services.

The key idea behind STEPS is to account for the unpredictable rainfall growth and decay processes by adding stochastic perturbations to the deterministic extrapolation of radar images (Seed, 2003). In order to be effective, the stochastic perturbations need to reproduce important statistical properties of both the precipitation fields and the forecast errors:

1. Spatial scaling of precipitation fields,
2. Dynamic scaling of precipitation fields,
3. Spatial correlation of the forecast errors,
4. Temporal correlation of the forecast errors.

The *spatial scaling* considers the precipitation field as arising from multiplicative cascade processes (Schertzer and Lovejoy, 1987; Seed, 2003). The presence of spatial scaling can be demonstrated by computing the 2D Fourier power spectrum (PS) of a precipitation field. A 1D PS can be obtained by radially averaging the 2D PS. The precipitation field is said to be *scaling* if the 1D PS draws a straight line on the log-log plot of the power against the spatial frequency (power law), which can be parametrized by a one or two spectral exponents (see e.g. Seed et al. 2013; Foresti and Seed, 2014). Within the multiplicative framework, a rainfall field is not represented as a collection of convective cells of a characteristic size but rather as a hierarchy of precipitation structures embedded in each other over a continuum of scales. STEPS considers the spatial scaling by decomposing the radar rainfall field into a multiplicative cascade using a fast Fourier transform (FFT) to isolate a set of 8 spatial frequencies (Seed, 2003; Bowler et al., 2006, Seed et al., 2013). The top cascade levels (0, 1 and 2) represent the low spatial frequencies (large precipitation structures), while the bottom cascade levels (5, 6, 7) represent the high spatial frequencies (small precipitation structures). Another important behavior of rainfall fields is known as *dynamic scaling*, which is the empirical observation that the rate of temporal development of rainfall structures is a power law function of their spatial scale (Venugopal et al., 1999; Foresti and Seed, 2014). This means that large precipitation features are more persistent and hence predictable compared with small precipitation cells, which is closely related to concept of scale-dependence of the predictability of precipitation (Germann and Zawadzki, 2002; Turner et al., 2004).

The stochastic perturbations should be able to reflect the properties of the forecast errors. Generating spatially and temporally correlated forecast errors is mandatory for hydrological applications, in particular when the correlation length of the errors is comparable or superior to the size and response time of the catchment. *Spatially correlated stochastic noise* can be constructed by applying a power law filter to a white noise field (Schertzer and Lovejoy, 1987). In practice it consists of three steps: computing the FFT of a white noise field, multiplying the obtained components in frequency domain by a given filter and applying the inverse FFT to return back to the spatial domain. The 1D or 2D power spectra of the rainfall field can be used as filter to obtain noise fields that have the same scaling and spatial correlation of the rainfall field. The 1D PS of the precipitation fields often appears to be a power law of the spatial frequency and explains why the procedure is also called power law filtering of white noise. In order to represent the anisotropies of the precipitation field the 2D PS can also be used as filter. In the absence of a target precipitation field from which to derive the PS, the filter can be parametrized by using a climatological power law (see Seed et al., 2013). Finally, the *temporal correlations* are imposed by auto-regressive (AR) filtering. A hierarchy of AR processes defines the temporal evolution of the cascade levels. With the exception of forecast lead times beyond 2-3 hours (Atencia and Zawadzki, 2014), an AR process of order 1 or 2 is already a good approximation to describe the temporal decorrelation of the forecast errors.

The practical implementation of STEPS to reproduce these important properties consists of the following steps (see Bowler et al., 2006; Foresti and Seed, 2014):

1. Estimation of the velocity field using optical flow on the last two radar rainfall images (Bowler et al., 2004a).
2. Decomposition of both rainfall fields into a multiplicative cascade using an FFT to isolate a set of 8 spatial frequencies.
3. Estimation of the rate of temporal evolution of rainfall features at each level of the cascade (Lagrangian auto-correlation).
4. Generation of a cascade of spatially correlated stochastic noise using as filter the 1D or 2D power spectra of the last observed radar rainfall field. A Gaussian filter is used to isolate a given spatial frequency (see Foresti and Seed, 2014).
5. Stochastic perturbation of the rainfall cascade using the noise cascade (level by level).

6. Extrapolation of the cascade levels using a semi-Lagrangian advection scheme.
7. Application of the AR(1) or AR(2) model for the temporal update of the cascade levels at each forecast lead time using the Lagrangian auto-correlations estimated in step (3).
8. Recomposition of the cascade into a rainfall field.
9. Probability matching of the forecast rainfall field with the original observed field (Ebert, 2001).
10. Computation of the forecast rainfall accumulations from the instant forecast rainfall rates. This procedure is known as advection correction and consists of advecting the instant rainfall rate forward over the 5 min period by discretizing the advection into smaller time steps.

3.2. STEPS implementation at RMI (STEPS-BE)

Bowler et al. (2006) introduced a general framework for blending a radar-based extrapolation nowcast with one or more outputs of downscaled NWP models (see also Pierce et al., 2010, and Seed et al., 2013). Because of being designed for urban applications, the maximum lead time of STEPS-BE is restricted to 2 hours. The operational NWP model of RMI (ALARO) runs only 4 times daily using a grid spacing of 4 km. Considering the model spin-up time and the absence of radar data assimilation, it is very unlikely that ALARO provides useful skill for blending its output with a radar-based extrapolation nowcast within the considered nowcasting range. It must also be reminded that the effective resolution of NWP models is much larger than the grid spacing. For instance, Grasso (2000) estimates the effective resolution to be at least 4 times the grid spacing, while Skamarock (2004) estimates it to be up to 7 times the grid spacing. ALARO would then only be able to resolve features that are greater than 20 km. For all these reasons, STEPS-BE only involves an extrapolation nowcast without NWP blending.

The STEPS-BE forecast domain is smaller than the extent of the 4 C-band radars composite (see Fig. 1). The radar field was upscaled from the original resolution of 500 m to 1 km and a sub-region of 512x512 grid points centered over Belgium was extracted. The forecast domain was extended by 32 pixels on each side to reduce the edge effects due to the FFT. This leads to an 8-levels multiplicative cascade representing the following spatial scales (rounded to the nearest integer): 576-256, 256-*114*-51, 114-*51*-23, 51-*23*-11, 23-*11*-4, 11-*4*-2, 4-*2*-1 and 2-*1*

km. Italic characters mark the scales on which the Gaussian filter is centered (see Foresti and Seed, 2014, for a more detailed explanation and visualization of the Gaussian FFT filter). One can notice that the spatial scales are not exact multiples of 2. In fact, a multiplication factor of 2,246 was employed to match the enlarged STEPS-BE domain size.

STEPS-BE includes a couple of improvements compared with the original implementation of the BOM:

1. Kernel interpolation of optical flow vectors,
2. Generation of stochastic noise only within the advected radar mask.

The optical flow algorithm of Bowler et al. (2004a) estimates the velocity field by dividing the radar domain into a series of blocks within which the optical flow equation is solved. The minimization of the field divergence is only performed at the level of the block, which leaves sharp discontinuities in the velocity field between the blocks. In order to overcome this issue, a Gaussian kernel regression was applied to interpolate the velocity vectors located at the center of the blocks onto the fine radar grid. The bandwidth of the Gaussian kernel was chosen to be σ = 24km = 0.4k, where k=60 grid points is the block size. This setting has the advantage of obtaining velocity fields that are less affected by the differential motion of small rainfall features and the presence of ground clutter. A too precise velocity field would provide increased predictability at very-short lead times but worse forecasts at longer lead times due to excessive convergence and divergence of precipitation features during the advection. Smooth velocity fields could also be obtained by using a smaller block size and by compensating with a larger bandwidth of the smoothing kernel.

In STEPS-BE the 1D power spectrum of the last observed rainfall field is used as filter to generate the spatially correlated stochastic perturbations. The PS is parameterized using two spectral slopes to account for a scaling break that is often observed at the wavelength of 40 km (see Seed et al., 2013; Foresti and Seed, 2014). To simplify the computations, an auto-regressive model of order 1 (AR(1)) was employed for imposing the temporal correlations and to model the growth of forecast errors.

The original STEPS implementation (Bowler et al., 2006) was designed to blend the radar extrapolation nowcasts with the output of NWP models. However, the domain covered by the radars is smaller than the rectangular domain of the NWP model and small amounts of stochastic noise are generated by default also outside of the radar composite. This setting was

not adapted for radar-based nowcasts without NWP blending and needed some adaptation. In fact, when advecting the radar mask over several time steps, large areas with small amounts of stochastic rain appear outside of the validity domain of the forecast and perturb the probability matching. In STEPS-BE the stochastic perturbations are only generated within the advected radar domain and set to zero elsewhere.

STEPS-BE can also account for the uncertainty in the estimation of the velocity field. The STEPS version that is implemented in UK (Bowler et al., 2006) includes a detailed procedure to generate velocity perturbations that reproduce various statistical properties of the differences between the forecast velocity and the actual future diagnosed velocity (see details in Bowler et al., 2004b). In the BOM and RMI implementations a simpler procedure is applied. The diagnosed velocity field is multiplied by a single factor C that is drawn from the following distribution:

$$C = 10^{1.5N/10} \quad , \qquad (1)$$

where N is a normally distributed random variable with zero mean and unit variance. In other words, the velocity field is accelerated or decelerated by a single random factor without affecting the direction of the vectors. In fact, the uncertainty on the diagnosed speed was observed to be higher than that of the direction of movement (Bowler et al., 2006).

The BOM and RMI versions of STEPS also include a stochastic model for the radar measurement error, a broken-line model to account for the unknown future evolution of the mean areal rainfall and the possibility to use time-lagged ensembles. However, a nowcasting model with too many components is harder to calibrate and complicates the interpretation of the forecast fields. Because of these reasons, STEPS-BE only exploits the basic stochastic model for the velocity field and for the evolution of rainfall fields (due to growth and decay processes).

The core of STEPS-BE is implemented in C/C++ and the production of figures in python. Bash scripts were written to call multiple STEPS instances and compute the ensemble members in parallel over several processors. Once all the ensemble members are computed, a separate script collects the corresponding netCDF files and calculates the forecast probabilities. Most of the computational cost of STEPS consists of filtering the white noise field with FFT, advecting and updating the radar cascade with the AR model. The re-

calculation of optical flow fields on each processor takes less than 10% of the total computational time.

The python matplotlib library is used to read the netCDF files, export the PNG figures and the time series of observed and forecast rainfall at the location of major cities and weather stations. A single STEPS nowcast generates more than 600 figures, which takes a significant fraction of the total computational time. In order to optimize the timing, a bash script monitors continuously the directory with incoming radar composites and triggers STEPS-BE once a field with a new time stamp is found. All these implementation details ensure that the user/forecaster can have access to an ensemble STEPS nowcast in less than 5 min after receiving the radar composite image.

The visualization system of STEPS-BE is very similar to the one of INCA-BE, the local Belgian implementation of the Integrated Nowcasting through Comprehensive Analysis system (INCA, Haiden et al., 2011) developed at the Austrian weather office (ZAMG). Figure 1 illustrates the web interface with an example of an ensemble mean nowcast. The user can highlight the major cities, weather stations and click to visualize the time series of observed and forecast precipitation/probability, which appears at the bottom of the web page. The navigation through the observations and forecast lead times is facilitated by the scroll wheel of the mouse. On the other hand, by clicking on the image it is possible to easily scroll through the various ensemble members or probability levels for a given lead time. Scrolling the ensemble members at different lead times is very instructive and can make the user aware of the forecast uncertainty. In fact, at a lead time of 5 min the ensemble members agree very well on the intensity and location of precipitation. This means that the ensemble spread is small and the probabilistic forecast is sharp, i.e. most of the forecast probabilities are close to 1 or 0 (see an explanation in Appendix A). On the other hand, at 1 or 2 hours lead time the ensemble members disagree on the location and intensity of rainfall, which enhances the ensemble spread and decreases the sharpness of the probabilistic forecast. The web page includes extensive documentation to guide the user and a set of case studies to help understanding the strengths and limitations of STEPS. The visualization system was implemented with great attention to take full advantage of the multi-dimensional information content of probabilistic and ensemble forecasts.

4 Forecast verification

4.1 Verification set-up

This section presents the verification of STEPS-BE forecasts using a set of case studies (see Sect. 2). The accumulated radar observations were employed as reference for the verification. The rainfall rates are accumulated over the last 5 min by reversing the field vectors based on the observations and then performing the advection correction. The 30 min ensemble mean forecast was verified against the observed 30 min radar accumulations using both continuous and categorical verification scores. The deterministic verification procedure follows the one presented in Foresti and Seed (2015), which was designed to analyze the spatial distribution of the forecast errors. More details about the forecast verification setup and scores are given in Appendix A.

The continuous scores include the multiplicative bias and the root mean squared error (RMSE), while the categorical scores include the probability of detection (POD), false alarm ratio (FAR) and Gilbert skill score (GSS) derived from the contingency table for rainfall thresholds of 0.5 mm hr^{-1} and 5.0 mm hr^{-1}. The rainfall thresholds are given in equivalent intensity independently of the forecast rainfall accumulation. Thus, a threshold of 5.0 mm hr^{-1} on a 30 min accumulation corresponds to 2.5 mm of rain. The multiplicative bias and the RMSE were evaluated only at the locations where the forecast or the verifying observations exceeded 0.1 mm hr^{-1}, which can be referred to as a *weakly conditional verification*. The probabilistic forecast of exceeding 0.1, 0.5, 5.0 mm hr^{-1} was verified using the reliability diagrams and ROC curves. Finally, the dispersion of the ensemble was analyzed by comparing the ensemble spread to the RMSE of the ensemble mean and by using rank histograms. The probabilistic and ensemble verification does not consider the spatial distribution of the errors and pools the data together in both space and time to derive the statistics.

4.2 Deterministic verification

Figures 2 and 3 show the average forecast and observed rainfall rates corresponding to the 0-30 min ensemble mean accumulation nowcast for the Ghent and Leuven cases respectively. In other words, they represent the average forecast and observed rainfall rates over the duration

of the precipitation event (for the 0-30 min lead time). The average was computed using the weak conditional principle explained above.

The average forecast and observed accumulations generally agree very well for the 0-30 min lead time forecast. The Ghent case on 10 November 2013 (Figs. 2a and 2b) is the only one with northwesterly flows and is characterized by the lowest average rainfall rates. The Avesnois radar demonstrates very well the range-dependence of the average rainfall rates, which gradually decrease with increasing distance from the radar. On the contrary, the smaller ring of high rainfall rates around the Zaventem radar is mostly due to the bright band (Fig. 2b).

The bright band effect influences the radar observations and hence the nowcasts based on their extrapolation. At longer lead times the larger rainfall estimates due to the bright band are extrapolated far from the location of the radar. The stochastic perturbations of STEPS can help to gradually dissolve the circular patterns introduced by the bright band effect. However, the bright band affects more the observations used for the verification, in particular when the rainfall is advected from upstream over the radar region. In such case, the local larger rainfall estimates lead to a verification bias and the forecasts are wrongly accused of rainfall under-estimation. In spite of these issues, bright band effects might not be so important for urban hydrological applications. In fact, except for one stratiform case presented in this paper, pluvial floods mainly happen in summer with convective precipitation events, during which the bright band is absent or negligible.

The Ghent case on 3 January 2014 has higher rainfall rates and the elongated structures of precipitation areas demonstrate well the southwesterly flow regime (Figs. 2c and 2d). For this case the measurements of the Zaventem radar are also affected by second trip echoes, which appear as a set of radially oriented rainfall structures North-West of the radar. These alternating patterns are explained by the dual PRF mode of Zaventem (see Sect. 2).

The Leuven cases on 9 June 2014 and 19-20 July 2014 have an important convective activity (Figs. 3a, 3b, 3c and 3d). The maximum average rainfall rates are located over the Ardennes mountain range and the city of Leuven respectively. Since urban flash floods can be triggered by a single convective cell, the average rainfall rate over the duration of the event may not be as high in the considered city (e.g. Fig. 3b).

Figure 4 illustrates the multiplicative bias of the 0-30 min nowcast averaged over each of the 4 events. A detailed interpretation of such forecast biases using Australian radar data and their

connection to orographic features is given in Foresti and Seed (2015), which point out that an important fraction of the forecast errors is caused by the biases of the verifying radar observations rather than systematic rainfall growth and decay processes due to orography. In Fig. 4a it is easy to notice the effect of bright band, which causes a series of systematic forecast biases around the Zaventem radar and perpendicularly oriented with respect to the prevailing flow direction (NW). Systematic rainfall underestimation occurs along the Belgian coast of the North Sea. One factor which contributes to this underestimation is the absence of visibility of the radar at longer ranges. The incoming precipitation is suddenly detected by the radar and therefore strongly underestimated by STEPS. The only situation where the range dependence of the rainfall estimation does not affect the forecast verification occurs when the velocity field is perfectly rotational and centered on the radar (assuming no beam blockage). All the other cases have to deal with the fact that the rainfall nowcast also extrapolates the biases of the radar observations! Contrary to expectation, on the upwind side of the Ardennes there is overestimation, which may depict a region of systematic rainfall decay. The bias over the city of Ghent is fortunately small and is included in the range from 0 to +0.5 dB (light overestimation, rainfall decay). Having small systematic biases over the cities of interest is very important for future integration of STEPS nowcasts as input in hydraulic models. In Fig. 4b the systematic underestimation is also located upstream with respect to the prevailing winds (SW). The strong overestimations in Germany and The Netherlands are mostly due to the underestimation of rainfall by the verifying radar observations rather than caused by systematic rainfall decay. This is particularly visible after a range of 125 km from the Zaventem radar, which demonstrates again that discontinuities and biases in the radar observations lead to biases in the extrapolation forecast. Also in this case the bias over the city of Ghent is small but in the range from -0.5 to 0 dB (light underestimation, rainfall growth). A similar radar bias is visible in Fig. 4c but this time located at a range of 240 km North of the Wideumont radar when entering the area covered by the Jabbeke radar. This forecast bias is mainly explained by the negative calibration bias of the Jabbeke radar, which is known to slightly underestimate the rainfall rates with respect to the Wideumont radar. Strong underestimation occurs over the Ardennes due to the systematic initiation and growth of convection that cannot be predicted by STEPS (Fig. 4c). Fortunately the city of Leuven is located in a region with small biases in the range from -0.5 to +0.5 dB. Figure 4d is quite interesting since strong underestimations are located in front of the rain band (from Charleroi to Leuven and beyond) and overestimations at the rear of the rain band (West of the Jabbeke

radar). The underestimations are due to systematic rainfall initiation in front of the rain band, while the overestimations are probably caused by a too slow extrapolation of rainfall, which tends to drag at the rear of the rain band. The two bands of underestimations South of Leuven are caused by two different thunderstorms. The first one passed over the city of Leuven and had a stronger westerly component with respect to the prevailing southerly flow. The second thunderstorm was weaker and had a stronger easterly component. When isolated convection does not follow the prevailing movement of the rainfall field, strong biases can appear in the nowcast during the first lead times.

Figure 5 shows the spatial distribution of the RMSE for the stratiform event on 3 January 2014 in Ghent and the convective event on 20 July 2014 in Leuven. If compared with Figs. 2d and 3d it is clear that the RMSE is strongly correlated with the regions having the highest mean rainfall accumulations (proportional effect). Thus, it is not surprising that the RMSE of the convective case (Fig. 5b) displays values exceeding 10 mm hr^{-1} over the city of Leuven. The winter case only shows RMSE values below 2 mm hr^{-1} over the city of Ghent.

Figure 6 illustrates an example of categorical verification of the 30-60 min ensemble mean forecast for the Leuven case on 19-20 July 2014 relative to the rainfall threshold of 0.5 mm hr^{-1}. The probability of detection is high everywhere (mean of 0.75) except in the neighborhood of Antwerp and South of Leuven, where the initiation of thunderstorms could not be predicted by STEPS (Fig. 6a). The false alarm ratio is quite low (mean of 0.36) and the regions with high values are mainly located at the rear of the front where the rainfall is advected too slowly compared with the actual movement of the front (Fig. 6b). A high Gilbert skill score generally coincides with the regions with the highest rainfall accumulations and becomes lower at the edges of the rain areas (Fig. 6c). This finding can be explained conceptually if one thinks about the verification of the future path of a single convective cell. The regions with the highest uncertainty are located along the edges of the predicted thunderstorm path and the highest skill is obtained in the center of the predicted path.

4.3 Probabilistic verification

Figure 7 shows the reliability diagrams relative to the probabilistic forecast of exceeding the 0.5 and 5.0 mm hr^{-1} rainfall thresholds for the Ghent case on 03 January 2014 (Figs. 7a and 7b) and the Leuven case on 19-20 July 2014 (Figs. 7c and 7d). The reference probabilistic forecast is taken as the climatological frequency of exceeding a given rainfall threshold during

that precipitation event (horizontal dashed line). Unexpectedly, the forecasts of the stratiform case in Ghent are less reliable than the ones of the convective case in Leuven for both rainfall thresholds. Probabilistic forecasts of exceeding 0.5 mm hr^{-1} for the Ghent case have a good reliability and positive Brier skill score (BSS) up to 60 min lead time (Fig. 7a). The higher rainfall threshold of 5.0 mm hr^{-1} is harder to predict and there is skill only up to 30 min lead time (Fig. 7b). The convective case in Leuven is more predictable and the probabilistic forecast of exceeding 0.5 mm hr^{-1} exhibits skill up to 90 min lead time (Fig. 7c). It is interesting to note that forecast probabilities that are close to the climatological frequency (intersection of lines around the probability 0.15) often fall outside of the skillful region. In fact, a small systematic forecast bias is likely to be worse than the event climatology at those frequencies. The rainfall threshold of 5.0 mm hr^{-1} shows again a limit of predictability of 30 min (Fig. 7d). Despite having a negative BSS, the following lead times (Fig. 7d) have higher resolution than the stratiform case in Ghent (Fig. 7b).

Figure 8 illustrates the ROC curves relative to the probabilistic forecast of exceeding 0.1 mm hr^{-1} for the Ghent case on 03 January 2014 (Figs. 8a) and the Leuven case on 19-20 July 2014 (Figs. 8b). All the ROC curves are very far from the diagonal line of no skill. The probability level that is marked with a cross is the one which maximizes the difference between the hit rate *HR* and the false alarm rate *F* (not to be confused with the false alarm ratio which is conditioned on the forecasts). This point is located within the probabilities 0.1 and 0.2, which means that an optimal forecast of the probability of rain is achieved when only 10-20% of the ensemble members exceed the 0.1 mm hr^{-1} threshold. A forecaster who is not scared of making false alarms would choose a lower probability level to increase the number of hits. On the contrary, an unconfident forecaster who would like to minimize the false alarms would choose a higher probability level, which has however the consequence of reducing the number of hits. As expected, the area under the ROC curves (AUC) decreases for increasing lead times. The discrimination skill for the convective event in Leuven is slightly higher than the one of the stratiform event in Ghent, which confirms the findings on the reliability diagrams (Fig. 7). This does not mean that small scale features are easier to forecast than larger scales features, which is known to be false (see Foresti and Seed, 2014). It means that the predictability of well defined and organized convective systems is higher than the one of more moderate convection with shorter lifetime, at least for the cases analyzed in this paper.

4.4 Ensemble verification

Figure 9 compares the error of the ensemble mean (RMSE) and the ensemble spread for the Ghent case on 03 January 2014 and the Leuven case on 19-20 July 2014 (see interpretation of ensemble spread in Appendix A). In both cases the RMSE increases up to a lead time of 50-60 min and then starts a slow decrease, which can be counter-intuitive. However, it must be reminded that the ensemble mean forecast becomes smoother for increasing lead times, which reduces the double penalty error due to forecasting a thunderstorm at the wrong location. The ensemble spread also increases up to 50-60 min lead time and then slowly stabilizes. For both the Ghent and Leuven cases the ensemble spread is lower than the error of the ensemble mean at all lead times, which suggests that the ensemble forecasts are under-dispersive. The degree of under-dispersion is highest at a lead time of 5 min, with the spread values being equal to 60% of the forecast error for the winter event in Ghent (Fig. 9a) and 70% for the summer event in Leuven (Fig. 9b). Except for the 5 min lead time, the ensemble spread represents 75-90% of the forecast error for the Ghent case (Fig. 9a) and 75-80% for the Leuven case (Fig. 9b), which is a good result. It is not yet clear why the RMSE at a lead time of 5 min is higher than the one at 10 min for the winter case in Ghent (Fig. 9a).

The under-estimation of the ensemble dispersion at the first lead time can be attributed to both the under-estimation of the ensemble spread and the over-estimation of the ensemble mean RMSE, but with different degrees according to the different causes. High RMSEs at the start of the nowcast can be due to using a very smooth velocity field for the advection (see Section 3.2), which does not exploit sufficiently the very short term predictability of small scale precipitation features, but is optimized for predictions at longer lead times. Another explanation for this underestimation of ensemble dispersion could be due to the space-time variability of the Z-R relationship. Spatial and temporal changes in the drop size distribution (DSD) can lead to changes in the estimated rainfall rate that is used for the verification. Therefore, there could be a mismatch between the "fixed" DSD of the forecasts and the variable DSD underlying the verifying observations. Another possible source of mismatch could be due to the advection correction with optical flow when computing the rainfall accumulations. The forecast accumulations are computed by advecting forward the previous rainfall field. On the other hand, the observed accumulations are computed by reversing the optical flow vectors and advecting the rainfall field backwards (see Section 4.1). This choice increases the differences when comparing the +0-5 min forecast accumulations (advection of

the 0 min image forward) with the +0-5 min observed accumulations a posteriori (advection of the +5 min image backwards). The ideal approach would be to derive the accumulation by advecting both the previous image forward and the last image backwards. An optimal accumulation could be computed by a weighted average of the two advected images by discretizing the 5 min interval. However, such approach is not very pragmatic and would require additional computational time in order to obtain a marginal improvement on the forecasts.

Figure 10 illustrates the rank histograms for the Leuven case on 19-20 July 2014 for lead times of 5 and 60 min. The U-shape of the rank histograms demonstrates again a certain degree of ensemble under-dispersion. In particular, all the ensemble members for the 5 min lead time are inferior to the observations in ~16% of the cases (Fig. 10a), while for the 60 min lead time it happens in more than ~30% of the cases (Fig. 10b). On the other hand, the fraction of observations falling below the value of the lowest ensemble member is only 8% for both lead times. Despite the fact that STEPS is designed to reproduce the space-time variability of rainfall, it underestimates a certain fraction of the observed rainfall extremes. This underestimation grows with increasing lead time and depicts an increasing smoothness of the STEPS ensembles, which is probably due to the advection of the radar rainfall cascade (see Sect. 3, step 6). In fact, the small scale rainfall features represented by the bottom cascade levels suffer more from numerical diffusion during the Lagrangian extrapolation, which is observed as a gradual loss of variability in the forecast ensembles.

4.5 Verification summary of the events

Table 2 provides a comparison of the verification scores for each event. The average standard deviation of the multiplicative biases of the 30 min lead time forecast is in the range 0.3-0.8. Except for the event on 19-20 July 2014 the biases remain well below 1 dB for all lead times, which is a positive result. Of course, these are average values and locally they can even exceed 3 dB (see Fig. 4).

On the other hand, the RMSE values mark more the distinction between the two winter cases in Ghent and the two summer cases in Leuven. For the winter cases the RMSE values increase from 0.38-0.95 at a lead time of 30 min to 0.78-1.48 at 120 min, while for the summer cases from 1.84-2.45 to 2.52-3.38 mm hr^{-1}. Thus, the RMSE of a 30 min lead time nowcast of the two convective cases is higher than the RMSE of a 120 min nowcast of the

two stratiform cases, as might be expected. It is interesting to mention that linear verification scores such as the RMSE strongly depend on the variance of the data. Consequently, it would be difficult to compare the error of the STEPS ensemble mean nowcast with the one of a deterministic nowcast, for example computed by INCA-BE. In fact, the ensemble averaging process filters out the unpredictable precipitation features and is rewarded in terms of RMSE. Similar results were already observed in Foresti et al. (2015), who also pointed out the difficulty of comparing ensemble prediction systems having a different number of ensemble members.

The probability of detection relative to the 0.5 mm hr^{-1} threshold decreases from 78-86% to 33-58%, while the false alarm ratio increases from 10-17% to 46-65%. The Gilbert skill score starts with values of 0.58-0.64 and 0.29-0.40 at the 30 and 60 min lead times respectively and decays to values of 0.08-0.20 at 120 min. Wang et al. (2009) reported a Critical success index value of 0.45 for STEPS nowcasts of 0-60 accumulations relative to the 1 mm hr^{-1} threshold. Considering that the GSS is the CSI corrected by random chance, this value is comparable with the ones of the 30-60 min accumulations obtained in this paper. The GSS values relative to the threshold of 5.0 mm hr^{-1} are much lower. They oscillate between 0.15 and 0.44 for the first lead time and become very low and close to 0 afterwards. Thus, the predictability of rainfall structures exceeding 5.0 mm hr^{-1} rarely exceeds 30 min according to the GSS.

The area under the ROC curve values characterizing the potential discrimination power of the probabilistic forecast of exceeding 0.5 mm hr^{-1} start at 0.92-0.95 at 30 min lead time and decrease to 0.69-0.79 at 120 min. For the probabilistic forecast of exceeding 5.0 mm hr^{-1} they start at 0.88-0.90 and decrease to 0.62 for the convective cases and to 0.50 for the stratiform cases (no discrimination).

From all these results we can conclude that there is not much predictability beyond 2 hours lead time by extrapolating the 4 C-band composite radar image in Belgium. Therefore, a maximum lead time of 2 hours in STEPS-BE is a good choice. Extending this lead time requires blending the radar-extrapolation nowcast with the output of NWP models to increase the predictability of precipitation.

5 Conclusions

The Short-Term Ensemble Prediction System (STEPS) is a probabilistic nowcasting system based on the extrapolation of radar images developed at the Australian Bureau of Meteorology in collaboration with the UK MetOffice. The principle behind STEPS is to produce an ensemble forecast by perturbing a deterministic extrapolation nowcast with stochastic noise. The perturbations are designed to reproduce the spatial and temporal correlations of the forecast errors and the scale-dependence of the predictability of precipitation.

This paper presented the local implementation, adaptation and verification of STEPS at the Royal Meteorological Institute of Belgium, referred to as STEPS-BE. STEPS-BE produces in real-time 20 member ensemble nowcasts at 1 km and 5 min resolutions up to 2 hours lead time using the 4 C-band radar composite of Belgium. Compared with the original implementation, STEPS-BE includes a kernel-based interpolation of optical flow vectors to obtain smoother velocity fields and an improvement to generate stochastic noise only within the advected radar composite to respect the validity domain of the nowcasts.

The performance of STEPS-BE was verified using the radar observations as reference on four case studies that caused sewer system floods in the cities of Ghent and Leuven during the years 2013 and 2014. The ensemble mean forecast of the next four 30 min accumulations was verified using the multiplicative bias, the RMSE as well as some categorical scores derived from the contingency table: the probability of detection, false alarm ratio and Gilbert skill score (Equitable skill score). The spatial distribution of multiplicative biases revealed regions of systematic over- and under-estimation by STEPS. The underestimations are often associated with the locations of convective initiation and thunderstorm growth, which cannot be predicted by STEPS. On the other hand, the regions of overestimation are mostly due to the underestimation of rainfall by the verifying observations rather than systematic rainfall decay (see Foresti and Seed, 2015, for a more detailed discussion). In order to disentangle the forecast and observation biases, detailed knowledge about the spatial distribution of the radar measurement errors for a given weather situation is needed. The multiplicative biases over the cities of Leuven and Ghent are very low (from -0.5 dB to + 0.5 dB), which is a good starting point to integrate STEPS nowcasts as inputs into sewer system hydraulic models. The categorical forecast verification helped discovering the places with low probability of detection due to convective initiation at the front of the rain band and high false alarm ratio at

the rear of the rain band, likely due to a too slow rainfall extrapolation by STEPS. Reliability diagrams demonstrated that probabilistic forecasts of exceeding 0.5 mm hr^{-1} have skill up to 60-90 min lead time. On the contrary, convective features exceeding 5.0 mm hr^{-1} are only predictable up to 30 min. In terms of reliability and discrimination ability, it was also observed that the forecasts of convective events have more skill than the ones on stratiform events. The STEPS ensembles are characterized by a certain degree of under-estimation of the forecast uncertainty, with values of the ensemble spread close to 75-90% of the forecast error.

The current contribution focused on the verification of STEPS-BE nowcasts using only four precipitation cases of different character. The deterministic and categorical verification require many more cases to analyze the climatological distribution of the forecast errors, e.g. as done in Foresti and Seed (2015). On the other hand, the probabilistic and ensemble verification pool the data in both space and time and converges much faster to stable statistics.

From a research perspective, STEPS-BE could also be extended by including a stochastic model to account for the residual radar measurement errors, in particular to obtain more accurate estimations of the forecast uncertainty at short range. The STEPS framework also allows blending the extrapolation nowcast with the output of NWP models, which is a necessary step to increase the predictability of precipitation for lead times beyond 2 hours.

Appendix A: Forecast verification scores

Forecast verification is an important aspect of a forecasting system. A forecast without an estimation of its accuracy is not very informative. For an in-depth description of forecast verification science and corresponding scores we refer to Jolliffe and Stephenson (2011) and the verification website maintained at the Bureau of Meteorology (http://www.cawcr.gov.au/projects/verification/).

The STEPS *ensemble mean forecast* was verified using the following scores:

- Multiplicative bias:

$$bias = \frac{1}{N}\sum_{i=1}^{N} 10\log_{10}\left(\frac{F_i + b}{O_i + b}\right), \quad (1)$$

where F_i is the forecast rainfall at a given grid point, O_i is the observed rainfall at a given grid point, b=2 mm hr^{-1} is an offset to eliminate the division by zero and to reduce the contribution of the forecast errors at low rainfall intensities, and N is the number of samples. For the specific case of the verification of the spatial distribution of forecast biases, the summation is performed over time. Thus, N corresponds to the number of forecasts where either the forecast or the observed rainfall are greater than 0.1 mm hr^{-1} at a given grid point (denoted as weak conditional verification). The bias is given in decibels [dB] in order to obtain a more symmetric distribution of the multiplicative errors centered at 0, which is not possible with the simple power ratio *F/O*. The following table summarizes the correspondence between the decibel scale and the power ratio:

dB	-6	-3	-1	-0.5	0	+0.5	+1	+3	+6
Power ratio (F/O)	0.251	0.501	0.794	0.891	1	1.122	1.259	1.995	3 .981

For example, a bias of +3 dB occurs when the forecast rainfall F is twice as much as the observed rainfall O.

- Root mean square error: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$. (2)

- Contingency table of a dichotomous (yes/no) forecast:

		Observed		
		Yes	No	Total
Forecast	Yes	hits	false alarms	forecast yes
	No	misses	correct negatives	forecast no
	Total	observed yes	observed no	total

where the *hits* is the number of times that both the observation and the forecast exceed a given rainfall threshold (at a given grid point), the *false alarms* is the number of times that the forecast exceeds the threshold but the observation does not, the *misses* is the number of times that the forecast does not exceed the threshold but the observation does and the *correct negatives* is the number of times that both the observation and the forecast do not exceed the threshold.

- Different scores can be derived from the contingency table to characterize a particular feature or skill of the forecasting system:

  - Probability of detection (hit rate): $POD = \frac{hits}{hits + misses} = \frac{hits}{observed\ yes}$, (3)

    The *POD* characterizes the fraction of observed events that were correctly forecast and is also known as hit rate (*HR*).

  - False alarm ratio:

    $$FAR = \frac{false\ alarms}{hits + false\ alarms} = \frac{false\ alarms}{forecast\ yes}, \quad (4)$$

    The *FAR* characterizes the fraction of forecast events that were wrongly forecast.

  - False alarm rate: $F = \frac{false\ alarms}{false\ alarms + correct\ negatives} = \frac{false\ alarms}{observed\ no}$. (5)

    The false alarm rate *F* is conditioned on the observations, while the false alarm ratio *FAR* on the forecasts.

- Gilbert skill score (Equitable threat score):

$$GSS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}, \quad (6)$$

where $$hits_{random} = \frac{(hits + misses)(hits + false\ alarms)}{total} = \frac{(observed\ yes)(forecast\ yes)}{total} \quad (7)$$

is the number of hits obtained by random chance, which is calculated by multiplying the marginal sums of the observed and forecast events (such as computing the theoretical frequencies for the Chi-squared test). The GSS characterizes the detection skill of the forecasting system w.r.t. random chance. In practice it corresponds to the Critical success index (CSI) adjusted for the hits obtained by random chance.

The accuracy of probabilistic forecasts can be verified in various ways. In this paper we employ the reliability diagram and the Receiver Operating Characteristic curve (ROC). The reliability diagram compares the forecast probability with the observed frequency. Reliability characterizes the agreement between the forecast probability and observed frequency. For a reliable forecasting system the two values should be the same, which happens for example when we observe rain 80% of the time when it is forecast with 80% probability (in average, diagonal line of Fig. 8). Unreliable forecasts exhibit departures from this optimum (bias). Resolution characterizes the ability of the forecasts to categorize the observed frequencies into distinct classes. The complete lack of resolution occurs when the forecast probabilities are completely unable to distinguish the observed frequencies, which generally corresponds to the climatological frequency of exceeding a given precipitation threshold (horizontal dashed line in Fig. 8). The Brier skill score (BSS) characterizes the relative accuracy of the probabilistic forecast compared to a reference system (see Jolliffe and Stephenson, 2011). Although the climatology or sample climatology of the event is often used as a reference, the BSS can also be computed against other reference forecasts, e.g. another probability forecasting method or even a deterministic forecasting method treated as a probabilistic binary forecast. However, in such cases it is not possible to draw a unique horizontal line of no skill in Figure 8. The region where the probabilistic forecast has a positive BSS, i.e. it is better than the climatological frequency, is grayed out. In fact, the points located below the no skill line are closer to the climatological frequency and produce a negative BSS. Reliability diagrams usually contain the histogram of the forecast probabilities to analyze the sharpness of the forecasts (small inset in Fig. 8). Sharpness characterizes the ability to forecast

probabilities that are different from the reference forecast. Sharp forecasting systems are "confident" about their predictions and give many probabilities around one and zero.

The ROC curve is used to analyze the discrimination power of a probabilistic forecast of exceeding a given rainfall threshold. It is constructed by plotting the hit rates and false alarm rates evaluated at increasing probability thresholds to make the binary decision whether it will rain or not. The ROC curve of a random probabilistic forecast system lies on the diagonal where the hit rate equals the false alarm rate (no skill): the forecast probabilities do not have discrimination power. When the false alarm rate is higher than the hit rate the forecast is worse than that obtained by random chance (below the diagonal). A skilled forecasting system is observed when the hit rates are higher than the false alarm rates, which draws a characteristic curve. The area under the ROC curve (AUC) measures the discrimination power of the probabilistic forecasts, with a maximum value of 1 (100% of hits and 0% of false alarms) and a minimum value of 0.5 for a random forecasting system. Values below 0.5 denote a forecasting system that performs worse than random chance. The AUC is computed by integrating over all the trapezoids that can be drawn below the ROC curve. The AUC is not sensitive to the forecast bias and the reliability of the forecast could be still improved through calibration. For this reason the AUC is only a measure of potential skill.

The ensemble forecasts are verified to detect whether there is over- or under-dispersion. It is common practice to compare the "skill" (error) of the ensemble mean with the ensemble spread (Whitaker and Loughe, 1998; Foresti et al., 2015):

$$spread = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(F_{im}-\overline{F_i}\right)^2} \quad (9)$$

where M is the number of ensemble members (ensemble size), F_{im} is the forecast of a given ensemble member and $\overline{F_i}$ is the ensemble mean forecast (at a given grid point). Since we are not analyzing the spatial or temporal distribution of the ensemble spread, N corresponds to the total number of samples in space and time, which is the number of forecasts within a rainfall event multiplied by the number of grid points within a radar field. The weak conditional verification is also applied to the computation of the spread. The ensemble spread characterizes the variability of the ensemble members about the ensemble mean (standard deviation). For a good ensemble prediction system, the ensemble spread should be equal to the average variability of the observations about the ensemble mean, as measured by the

RMSE of the ensemble mean (Eq. (2)). If the spread is larger than the RMSE, the ensemble is overestimating the forecast uncertainty (over-dispersion), otherwise it is underestimating it (under-dispersion). It is interesting to mentions that the ensemble mean RMSE and ensemble spread could also be computed starting from the logarithm of rainfall rates to account for the skewed distribution of precipitation (not used in this paper).

Another way to analyze the spread of ensemble forecasts is based on rank histograms (also known as Talagrand diagram). First, the precipitation values of the ensemble members are ranked in increasing order. Then, the rank of the observation is evaluated by checking in which of the $M+1$ bins it falls. By repeating the operation for a large number of cases and forecasts it is possible to construct a histogram. A good ensemble prediction system displays a flat histogram, i.e. the observations are indistinguishable from the forecasts and each ensemble member is an equi-probable realization of the future state of the atmosphere. A bell-shaped histogram with a peak in the middle is observed in case of ensemble over-dispersion. On the contrary, a U-shape histogram with peaks at the edges is observed in case of ensemble under-dispersion, which is more common (in particular for NWP ensembles). In this case the values of the observations often fall below or above the lowest or highest value of the ranked ensemble, which is not dispersive enough to capture the extremes.

Acknowledgments

This research was funded by the Belgian Science Policy Office (BelSPO) project PLURISK: “Forecasting and management of rainfall-induced risks in the urban environment” (SD/RI/01A). We thank Clive Pierce for the detailed discussions about the STEPS implementation and the guidance for various code improvements. We also acknowledge Meteo-France for providing the Avesnois data.

References

Achleitner, S., Fach, S., Einfalt, T., and Rauch, W.: Nowcasting of rainfall and of combined sewage flow in urban drainage systems. Water Sci. Technol., 59(6), 1145-51, 2009.

Atencia, A., and Zawadzki, I.: A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. Mon. Weather Rev., 142, 4036-4052, 2014.

Berenguer, M., Corral, C., Sánchez-Diezma, R., and Sempere-Torres, D.: Hydrological validation of a radar-based nowcasting technique. J. Hydrometeor, 6, 532-549, 2005.

Berenguer, M., Sempere-Torres, D., and Pegram, G. G. S.: SBMcast - An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. J. Hydrology, 404 (3), 226-240, 2011.

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

Bowler, N. E. H., Pierce, C. E., and Seed, A. W.: Development of a precipitation nowcasting algorithm based upon optical flow techniques. J. Hydrology, 288, 74-91, 2004a.

Bowler, N. E. H., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. Forecast Research Technical Report No. 433, MetOffice, 2004b.

Bowler, N. E. H., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. Q. J. R. Meteorol. Soc., 132, 2127-2155, 2006.

Bruni, G., Reinoso, R. van de Giesen, N. C., Clemens, F. H. L. R., and ten Veldhuis, J. A. E.: On the sensitivity of urban hydrodynamic modelling to rainfall spatial and temporal resolution. Hydrol. Earth Syst. Sci., 19, 691-709, 2015.

Cloke, H. L., and Pappenberger, F.: Ensemble flood forecasting: a review. J. Hydrology, 375(3), 613-626, 2009.

Collier, C. G.: On the propagation of uncertainty in weather radar estimates of rainfall through hydrological models. Meteorol. Appl., 16(1), 35-40, 2009.

Dai, Q., Rico-Ramirez, M. A., Han, D., Islam, T., and Liguori, S.: Probabilistic radar rainfall nowcasts using empirical and theoretical uncertainty models. Hydrol. Processes, 29, 66-79, 2015.

Ebert, E. E.: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Mon. Wea. Rev., 129, 2461-2480, 2001.

Ehret, U., Götzinger, J., Bárdossy, A., and Pegram, G. G. S.: Radar-based flood forecasting in small catchments, exemplified by the Goldersbach catchment, Germany. Int. J. River Basin Manag., 6 (4), 323-329, 2008.

Einfalt, T., Arnbjerg-Nielsen, K., Golz, C., Jensen, N. E., Quirmbachd, M., Vaes, G., and Vieux, B.: Towards a roadmap for use of radar rainfall data in urban drainage. J. Hydrology, 299, 186-202, 2004.

Figueras i Ventura, J. and Tabary, P.: The new French operational polarimetric radar rainfall rate product. J. Appl. Meteor. Climatol., 52, 1817-1835, 2013.

Foresti, L., and Seed, A.: The effect of flow and orography on the spatial distribution of the very short-term predictability of rainfall. Hydrol. Earth Syst. Sci., 18, 4671-4686, 2014.

Foresti, L., Seed, A., and Zawadzki, I.: Report of the Heuristic Probabilistic Forecasting workshop, Munich, Germany, 13 p., 30-31 August, 2014, https://sites.google.com/site/lorisforesti/projects/nowcasting/ScientificReport_HeuristicProbForecastingWorkshop_Munich_2014_121214.pdf.

Foresti, L., and Seed, A.: On the spatial distribution of rainfall nowcasting errors due to orographic forcing. Meteorol. Appl., 22(1), 60-74, 2015.

Foresti, L., Panziera, L., Mandapaka, P. V., Germann, U., and Seed, A.: Retrieval of analogue radar images for ensemble nowcasting of orographic rainfall. Meteorol. Appl., 22(2), 141-155, 2015.

Germann, U., and Zawadzki, I.: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Methodology. Mon. Wea. Rev., 130, 2859-2873, 2002.

Germann, U., and Zawadzki, I.: Scale-dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. J. Appl. Meteorol., 43, 74-89, 2004.

Germann, U., Zawadzki, I., and Turner, B.: Scale-dependence of the predictability of precipitation from continental radar images. Part IV: Limits to Prediction. J. Atmos. Sci., 63, 2092-2108, 2006.

Germann, U., Berenguer, M., Sempere-Torres, D., and Zappa, M.: REAL - Ensemble radar precipitation estimation for hydrology in a mountainous region. Q. J. R. Meteorol. Soc., 135, 445-456, 2009.

Goudenhoofdt, E., and Delobbe, L.: Evaluation of radar-gauge merging methods for quantitative precipitation estimates. Hydrol. Earth Syst. Sci., 13, 195-203, 2009.

Goudenhoofdt, E., and Delobbe, L.: Statistical characteristics of convective storms in Belgium derived from volumetric weather radar observations. J. Appl. Meteor. Climatol., 52, 918-934, 2013.

Grasso, L. D.: The differentiation between grid spacing and resolution and their application to numerical modeling. Bull. Am. Meteorol. Soc., 81, 579-580, 2000.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the eastern Alpine region. Wea. Forecasting, 26, 166-183, 2011.

Hohti, H., Koistinen, J., Nurmi, P., Saltikoff, E., and Holmlund, K.: Precipitation nowcasting using radar-derived atmospheric motion vectors. Proc. of the 1st European Conf. on Radar in Meteorology and Hydrology (ERAD), 2000.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science (2nd Edition). John Wiley and Sons, Chichester, 2011.

Jordan, P., Seed, A. W., and Weinnman, P. E.: A stochastic model of radar measurement errors in rainfall accumulations at catchment scale. J. Hydrometeorol., 4(5), 841-855, 2003.

Lewis, H., et al.: From months to minutes – exploring the value of high-resolution rainfall observation and prediction during the UK winter storms of 2013/2014. Meteorol. Appl., 22, 90-104, 2015.

Liguori, S., and Rico-Ramirez, M. A.: Quantitative assessment of short-term rainfall forecasts from radar nowcasts and MM5 forecasts. Hydrol. Processes, 26, 3842-3857, 2012.

Liguori, S., and Rico-Ramirez, M. A., Schellart, A., and Saul, A.: Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. Atmos. Res., 103, 80-95, 2012.

Liguori, S., and Rico-Ramirez, M. A.: A practical approach to the assessment of probabilistic flow predictions. Hydrol. Processes, 27, 18-32, 2013.

Metta, S., Rebora, N., Ferraris, L., von Hardernberg, J., and Provenzale, A.: PHAST: a phase-diffusion model for stochastic nowcasting. J. Hydrometeorol., 10, 1285-1297, 2009.

Panziera, L., Germann, U., Gabella, M., and Mandapaka, P. V.: NORA - Nowcasting of orographic rainfall by means of analogues. Q. J. R. Meteorol. Soc., 137(661), 2106-2123, 2011.

Pappenberger, F., and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resour. Res., 42(5), W05302, 2006.

Paschalis, A., Molnar, P., Fatichi, S., and Burlando, P.: A stochastic model for high-resolution space-time precipitation simulation. Water Resour. Res., 49(12), 8400-8417, 2013.

Pegram, G. G. S., and Clothier, A. N.: High resolution space-time modelling of rainfall: the "String of Beads" model. J. Hydrology, 241, 26-41, 2001a.

Pegram, G. G. S., and Clothier, A. N.: Downscaling rainfields in space and time, using the String of Beads model in time series mode. Hydrol. Earth Sys. Sci., 5(2), 175-186, 2001b.

Pierce, C., Bowler, N., Seed, A., Jones, A., Jones, D., and Moore, R.: Use of a stochastic precipitation nowcast scheme for fluvial flood forecasting and warning. Atmosph. Sci. Lett., 6, 78-83, 2005.

Pierce, C., Hirsch, T., and Bennett, A.C.: Formulation and evaluation of a post-processing algorithm for generating seamless, high resolution ensemble precipitation forecasts, Forecasting R&D Technical Report 550, MetOffice, Exeter, UK, 2010.

Radhakrishna, B., Zawadzki, I., and Fabry, F.: Predictability of precipitation from continental radar images. Part V: growth and decay. J. Atmos. Sci., 69, 3336-3349, 2012.

Roulin, E. and Vannitsem, S.: Skill of medium-range hydrological ensemble predictions. J. Hydrometeor, 6, 729-744, 2005.

Schellekens, J., Weerts, A. H., Moore, R. J., Pierce, C. E., and Hildon, S.: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales. Adv. Geosci., 29, 77-84, 2011.

Schertzer, D., and Lovejoy, S.: Physical modelling and analysis of rain and clouds by anisotropic scaling multiplicative processes. J. Geophys. Res., 92, 9696-9714, 1987.

Seed, A.: A dynamic and spatial scaling approach to advection forecasting. J. Applied Meteorol., 42, 381-388, 2003.

Seed, A. W., Pierce, C. E., and Norman, K.: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. Water Resour. Res., 49(10), 6624-6641, 2013.

Silvestro, F., Rebora, N.: Operational verification of a framework for the probabilistic nowcasting of river discharge in small and medium size basins. Nat. Hazards Earth Syst. Sci., 12, 763-776, 2012.

Silvestro, F., Rebora, N., and Cummings, G.: An attempt to deal with flash floods using a probabilistic hydrological nowcasting chain: a case study. Nat. Hazards Earth Syst. Sci. Discuss., 1, 7497-7515, 2013.

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.: Use of NWP for nowcasting convective precipitation: recent progress and challenges. *Bull. Amer. Meteor. Soc.*, **95**, 409–426, 2014.

Tabary, P.: The new French operational radar rainfall product. Part I: methodology. Wea. Forecasting, 22, 393–408, 2007.

Thielen, J., Bartholmes, J. Ramos, M. H., and de Roo, A.: The European Flood Alert System - Part 1: Concept and development. Hydrol. Earth Syst. Sci., 13, 125-140, 2009.

Thorndahl, S., and Rasmussen, M. R.: Short-term forecasting of urban storm water runoff in real-time using extrapolated radar rainfall data. J. Hydroinform., 15(3), 897-912, 2013.

Turner, B. J., Zawadzki, I., and Germann, U.: Predictability of precipitation from continental radar images. Part III: operational nowcasting implementation (MAPLE). J. Appl. Meteorol., 43, 231-248, 2004.

Venugopal, V., Foufoula-Georgiou, E., and Sapozhnikov, V.: Evidence of dynamic scaling in space-time rainfall. J. Geophys. Res., 104(D24), 31599-31610, 1999.

Verworn, H. R., Rico-Ramirez, M. A., Krämer, S., Cluckie, I., and Reichel, F.: Radar-based flood forecasting for river catchments. Water Management, 162(2), 159-168, 2009.

Wang, J. Keenan, T., Joe, P., Wilson, J., Lai, E. S. T., Liang, F., Wang, Y., Ebert, E. E., Ye, Q., Bally, J., Seed, A., Chen, M., Xue, J., Conway, B.: Overview of the Beijing 2008 Olympics project. Part I: Forecast Demonstration Project, A report to the WMO World Weather Research Programme, 2009.

Whitaker, J. S., Loughe, A. F.: The relationship between ensemble spread and ensemble mean skill. Mon. Weather Rev., 26, 3292-3302, 1998.

Willems, P.: A spatial rainfall generator for small spatial scales. J. Hydrology, 252(1-4), 126-144, 2001a.

Willems, P.: Stochastic description of the rainfall input errors in lumped hydrological models. Stoch. Env. Res. Risk Assess., 15(2), 132-152, 2001b.

Skamarock, W. C.: Evaluating mesoscale NWP models using kinetic energy spectra. Mon. Wea. Rev., 132, 3019-3032, 2004.

Xuan, Y., Cluckie, I. D., and Wang, Y.: Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction, Hydrol. Earth Syst. Sci., 13, 293-303, 2009.

Xuan, Y., Zhu, D., Triballi, P., and Cluckie, I.: Forecast uncertainty of a lumped hydrological model coupled with the STEPS radar rainfall nowcasts. Int. Symp. Weather Radar and Hydrol., Washington DC, US, April 2014.

Zappa, M., Beven, K., Bruen, M., Cofino, A., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Seed, A., Schroter, K., Szturc, J., Vehvilainen, B., Germann, U., and Rossa, A.: Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. Atmos. Sci. Lett., 11, 83-91, 2010.

1 Table 1. List of precipitation events that caused sewer system floods in Ghent and Leuven.

City	Date	Event start [UTC]	Event end [UTC]	Duration	Predominant precipitation	Main wind direction
Ghent	10 Nov. 2013	13:50	22:00	8h 10min	Stratiform	WNW → NNW
Ghent	3 Jan. 2014	03:00	14:00	11h	Stratiform	SW → WSW
Leuven	9-10 June 2014	06:30, 9th	15:30, 10th	33h	Convective	SW
Leuven	19-20 July 2014	22:00, 19th	06:30, 20th	8h 30min	Convective	SSW

Table 2. Summary of the forecast verification scores of the next four 30 min accumulation forecasts for the precipitation events in Ghent and Leuven. The lead time shown is the end of the 30 min accumulation period (e.g. 60 min is relative to the 30-60 min accumulation). The bias values correspond to the standard deviation of the multiplicative bias, which is more interesting than its mean (often close to 0).

Event	Bias 30min	Bias 60min	Bias 90min	Bias 120min	RMSE 30min	RMSE 60min	RMSE 90min	RMSE 120min
	[dB]				[mm hr^{-1}]			
10.11.2013	0.30	0.49	0.61	0.70	0.38	0.59	0.71	0.78
03.01.2014	0.54	0.74	0.82	0.89	0.95	1.39	1.53	1.48
9-10.06.2014	0.52	0.63	0.66	0.69	2.45	3.26	3.40	3.38
19-20.07.2014	0.84	1.18	1.30	1.35	1.84	2.36	2.49	2.52
	POD 30min	POD 60min	POD 90min	POD 120min	FAR 30min	FAR 60min	FAR 90min	FAR 120min
Event	Forecast >= 0.5 mm hr^{-1}				Forecast >= 0.5 mm hr^{-1}			
10.11.2013	0.83	0.71	0.62	0.54	0.17	0.30	0.38	0.46
03.01.2014	0.80	0.63	0.49	0.33	0.10	0.25	0.45	0.65
9-10.06.2014	0.78	0.65	0.55	0.46	0.15	0.32	0.44	0.54
19-20.07.2014	0.86	0.75	0.66	0.58	0.17	0.36	0.50	0.61
	GSS 30min	GSS 60min	GSS 90min	GSS 120min	GSS 30min	GSS 60min	GSS 90min	GSS 120min
Event	Forecast >= 0.5 mm hr^{-1}				Forecast >= 5.0 mm hr^{-1}			
10.11.2013	0.58	0.38	0.27	0.20	0.15	0.02	0.0	0.0
03.01.2014	0.64	0.40	0.20	0.08	0.28	0.06	0.0	0.0
9-10.06.2014	0.59	0.38	0.26	0.17	0.44	0.20	0.09	0.04
19-20.07.2014	0.58	0.29	0.14	0.07	0.27	0.09	0.04	0.02
	AUC 30min	AUC 60min	AUC 90min	AUC 120min	AUC 30min	AUC 60min	AUC 90min	AUC 120min
Event	Forecast >= 0.5 mm hr^{-1}				Forecast >= 5.0 mm hr^{-1}			
10.11.2013	0.95	0.89	0.84	0.79	0.88	0.67	0.56	0.50
03.01.2014	0.92	0.85	0.78	0.69	0.90	0.72	0.57	0.50
9-10.06.2014	0.93	0.86	0.81	0.76	0.89	0.77	0.68	0.62
19-20.07.2014	0.94	0.87	0.82	0.77	0.88	0.75	0.68	0.62

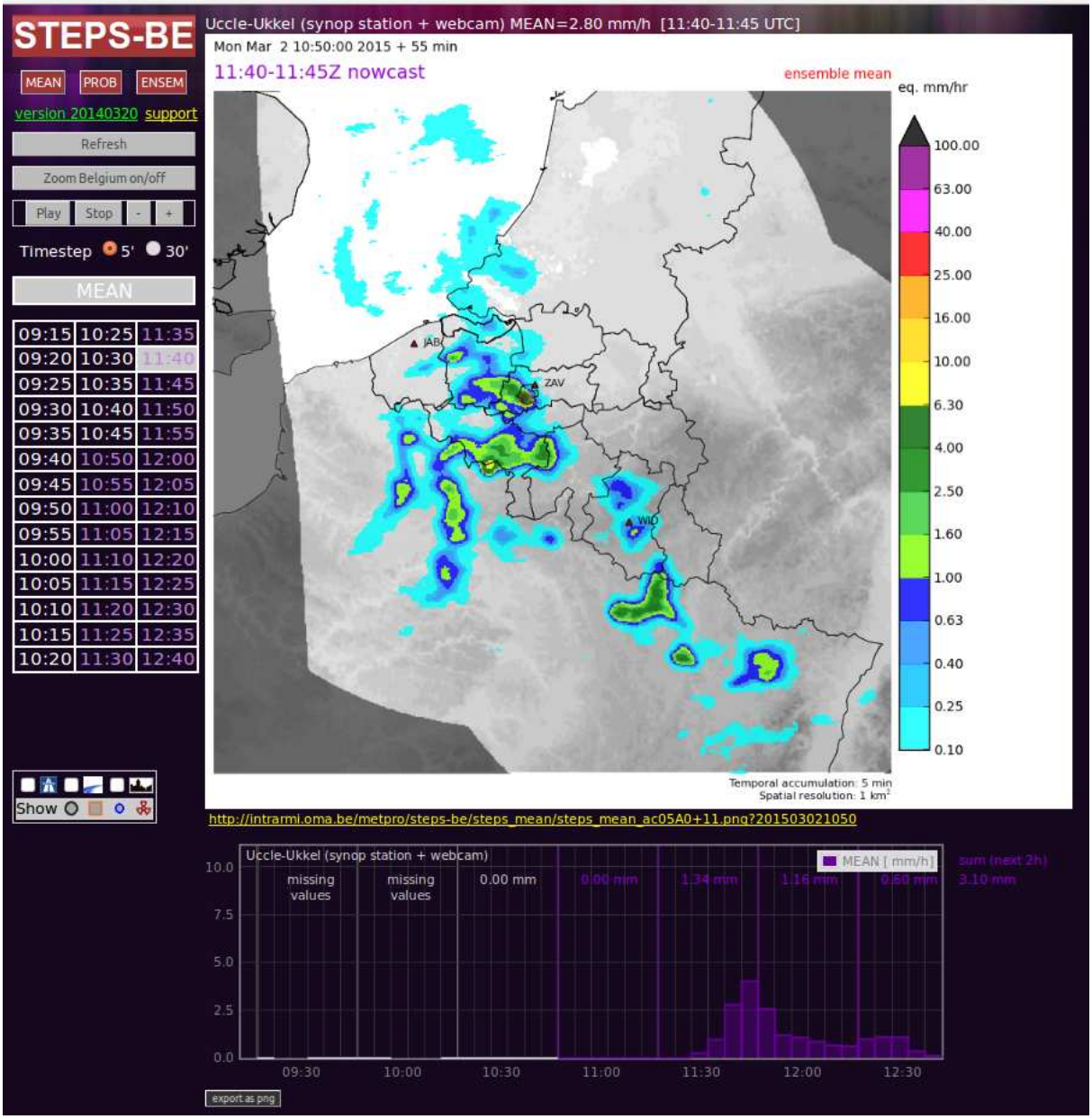
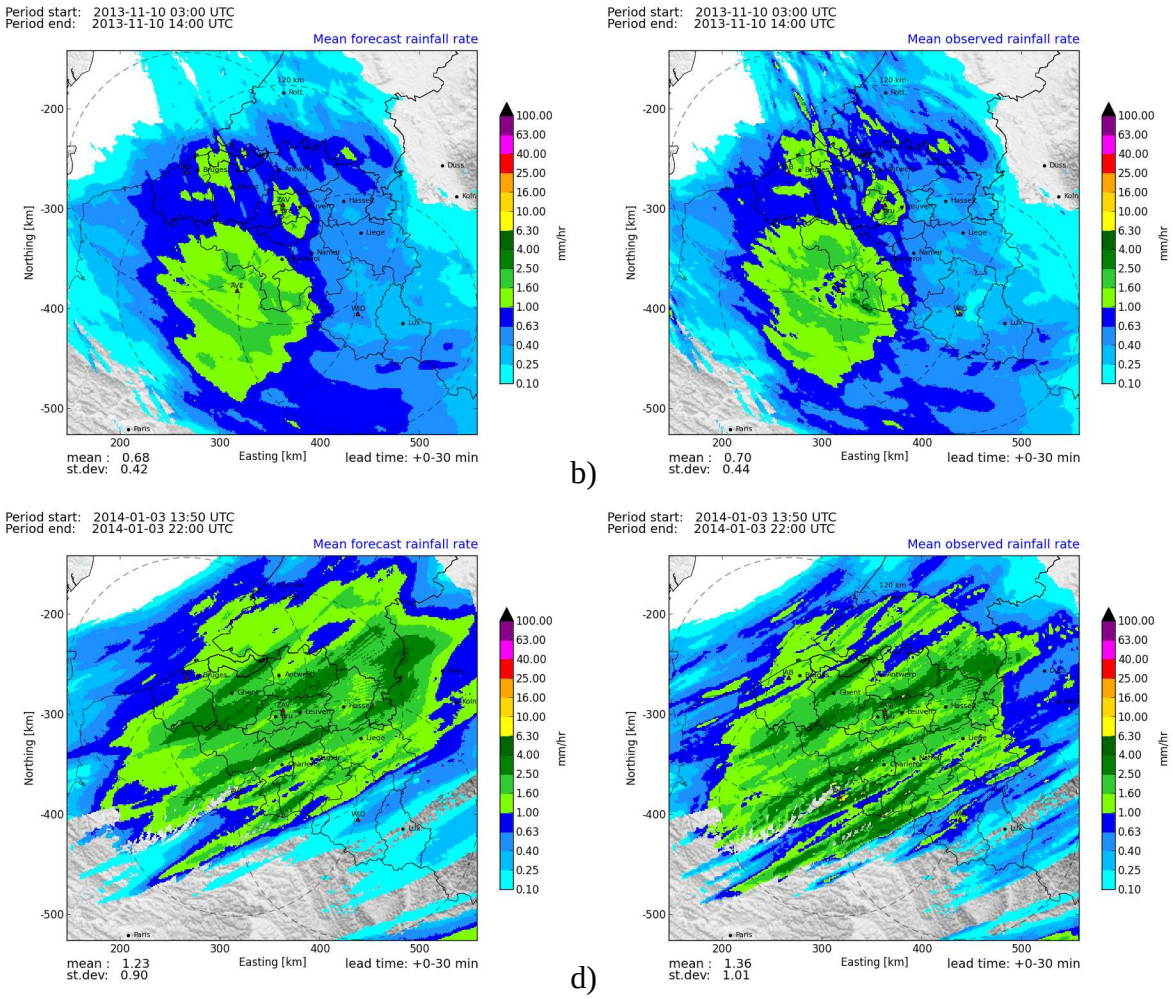


Figure 1. Web platform of STEPS-BE showing the ensemble mean forecast.



a) b)

c) d)

Figure 2. Average forecast and observed rainfall accumulations for the Ghent cases. a) Forecast and b) observed 0-30 min rainfall accumulations on 10 November 2013. c) Forecast and d) observed 0-30 min rainfall accumulations on 03 January 2014. The mean and standard deviation of the field within the 120 km range of the radars are shown on the bottom left corner. Field values are shown only if there are at least 10 samples for the computation of the mean. The red triangles denote the location of the Wideumont (WID, coordinates: 438 km East / -405 km North), Zaventem (ZAV, 363/-296), Jabbeke (JAB, 266/-263) and Avesnois (AVE, 317/-382) radars. The 120 km range from the radar is displayed as a dashed circle. The mountain range of the Ardennes covers the three most southern districts of Belgium and Luxembourg (LUX).

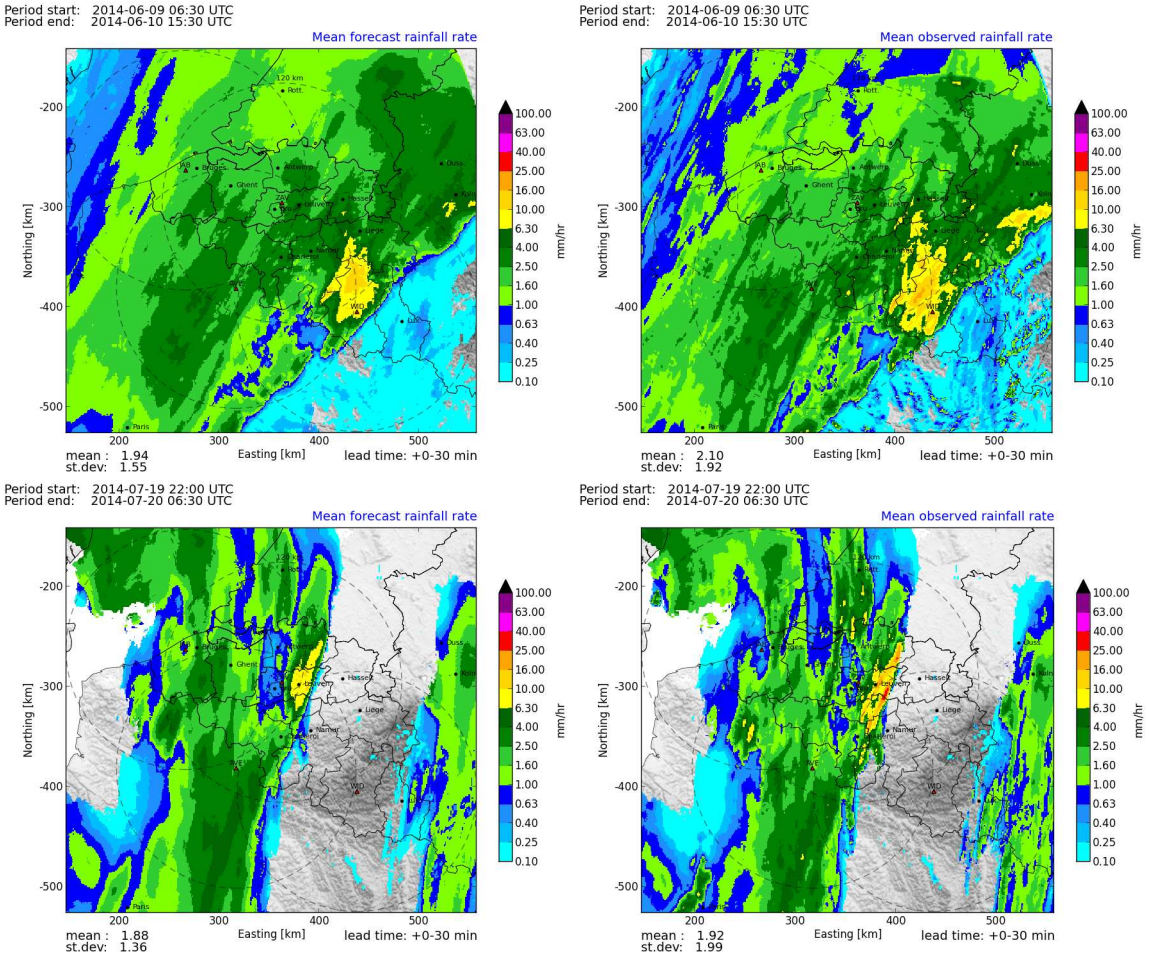
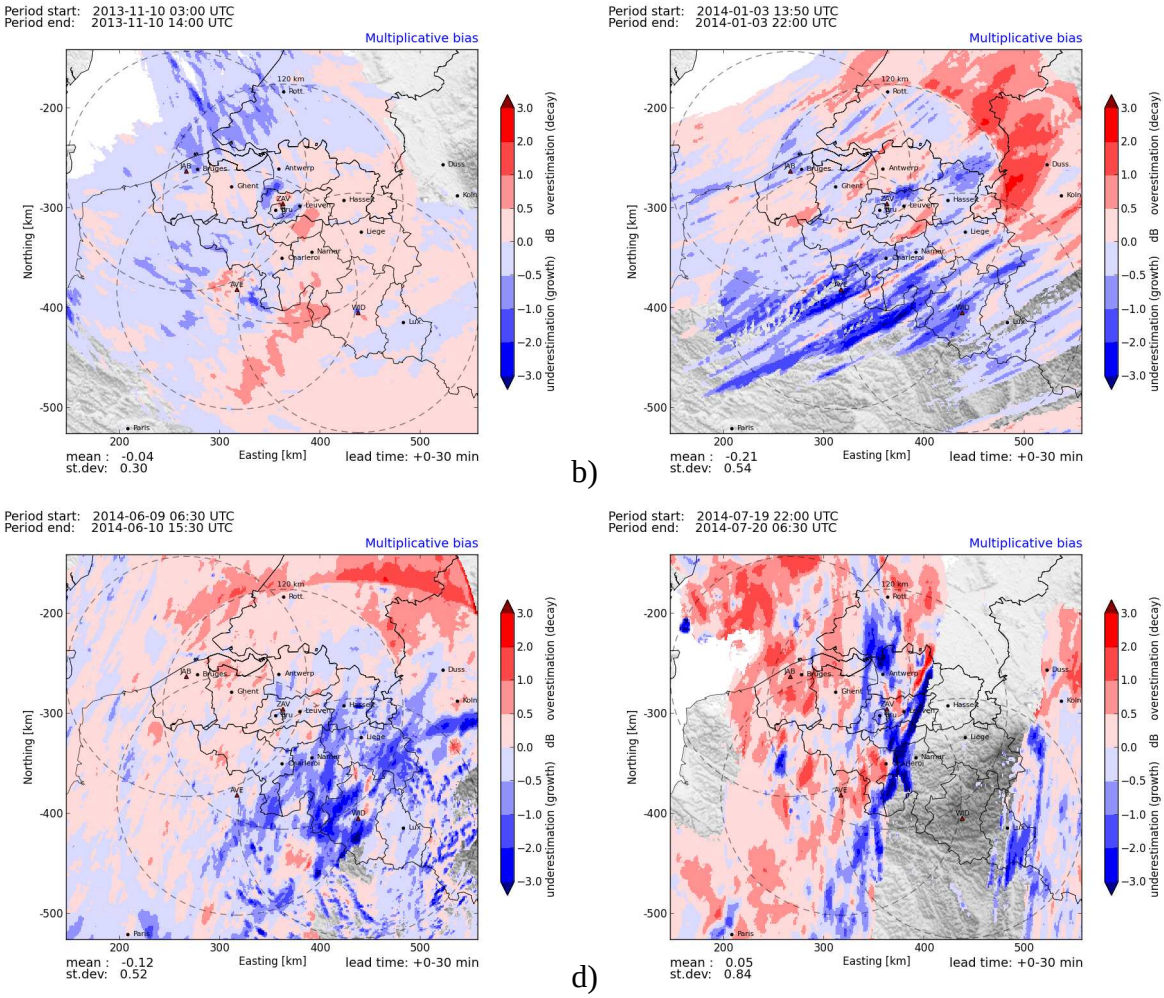


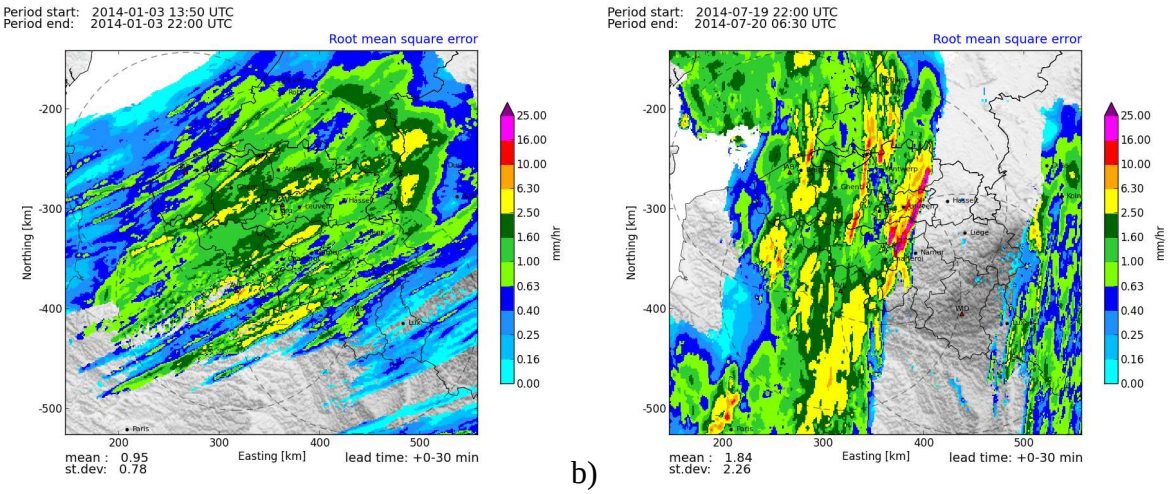
Figure 3. Average observed and forecast rainfall accumulations for the Leuven cases. a) Forecast and b) observed 0-30 min rainfall accumulations on 9-10 June 2014. c) Forecast and d) observed 0-30 min rainfall accumulations on 19-20 July 2014.



a) b)

c) d)

Figure 4. Average 0-30 min multiplicative forecast biases for the Ghent cases on a) 10 November 2013 and on b) 03 January 2014 and the Leuven cases on c) 9-10 June 2014 and on d) 19-20 July 2014. The interpretation of under- and over-estimations by STEPS as systematic rainfall growth and decay or simply as radar measurement biases is subject to interpretation as explained in text.



a) b)

Figure 5. Average 0-30 min forecast RMSE for a) the Ghent winter case on 03 January 2014 and b) the Leuven summer case on 19-20 July 2014.

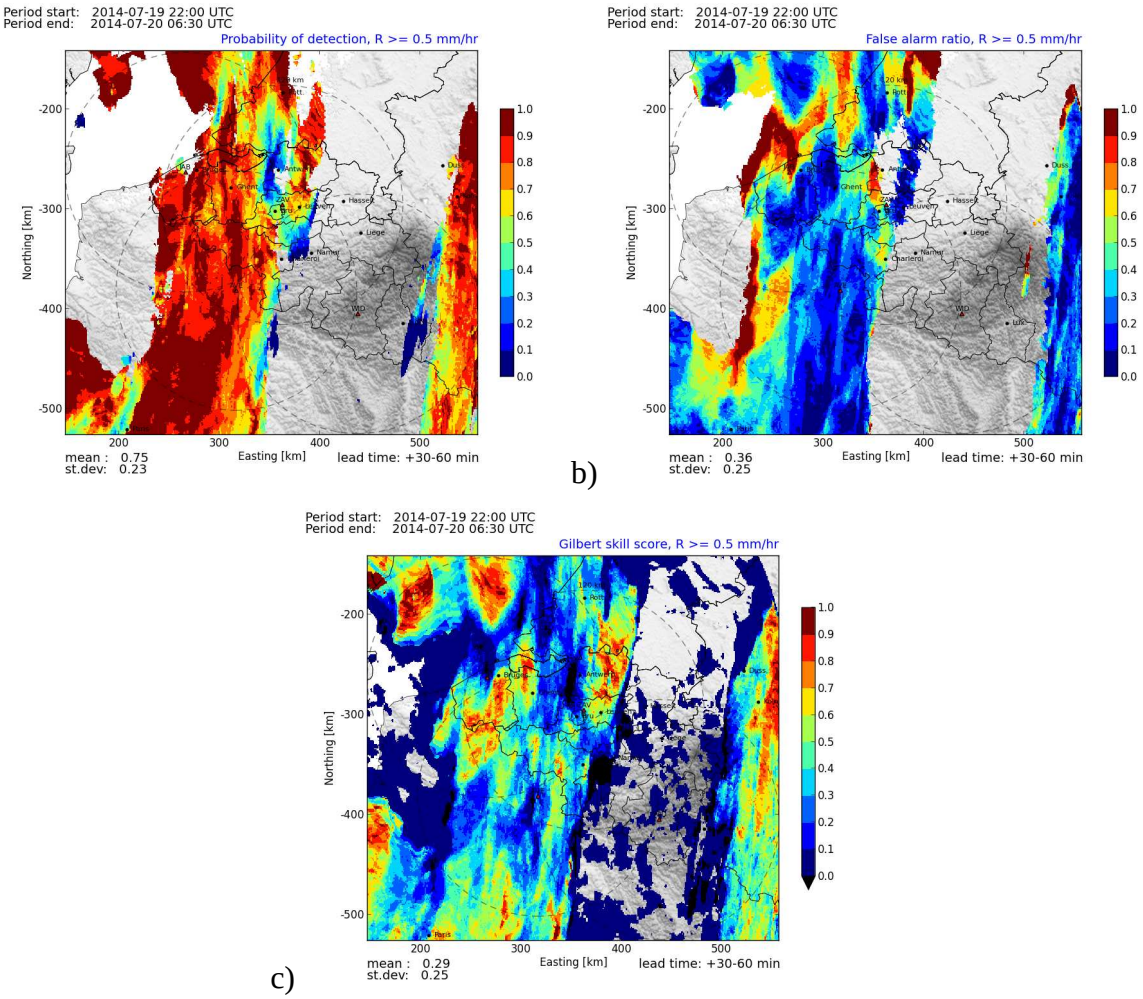


Figure 6. a) POD, b) FAR and c) GSS of the 30-60 min ensemble mean forecast of exceeding 0.5 mm hr^{-1} for the Leuven case on 19-20 July 2014.

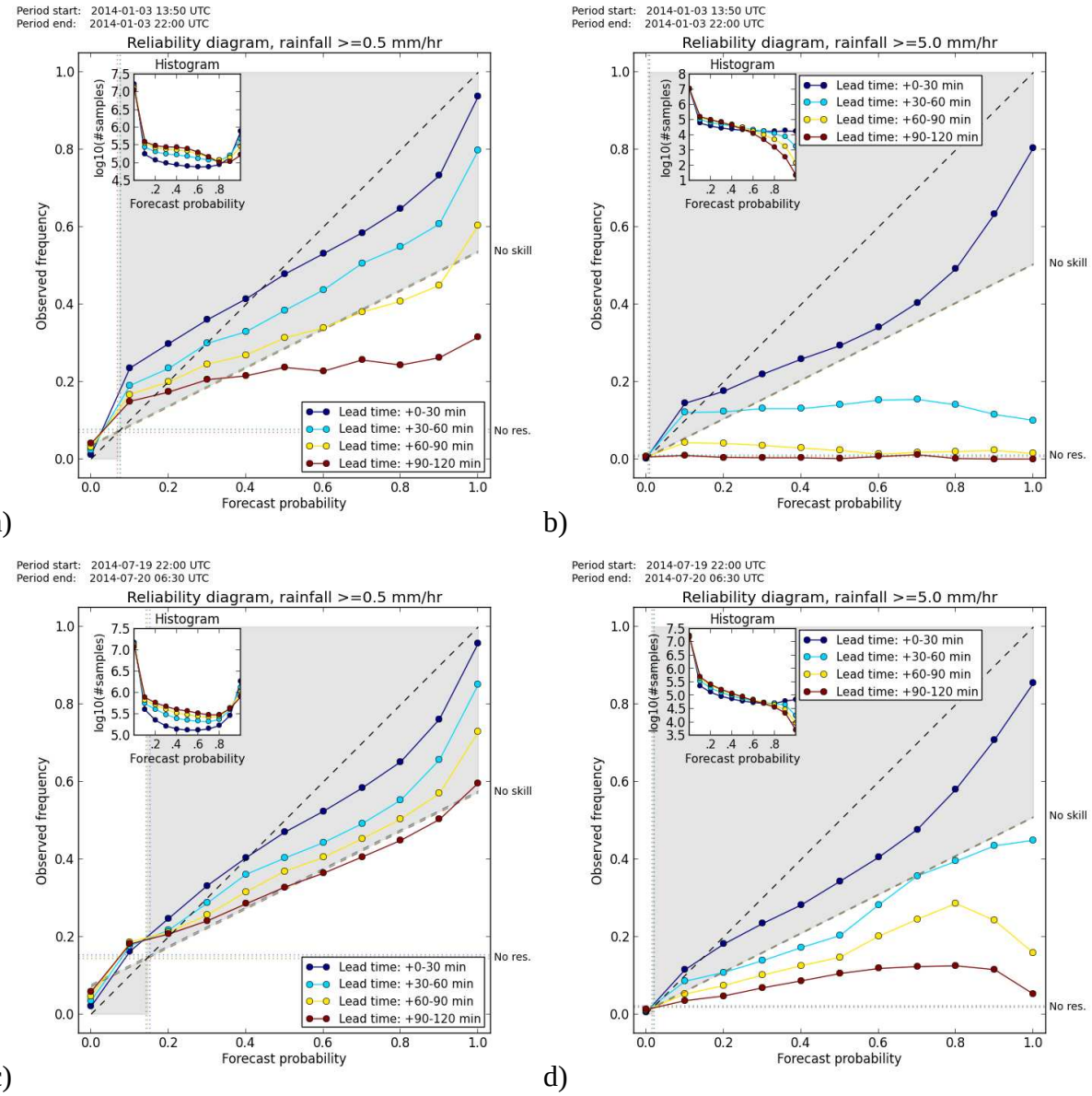
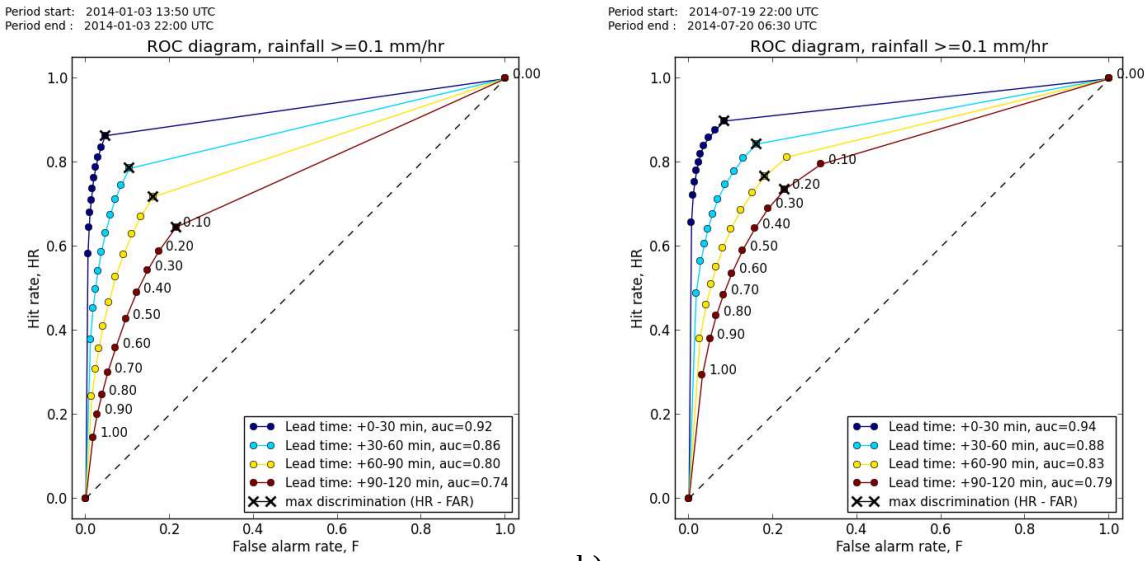
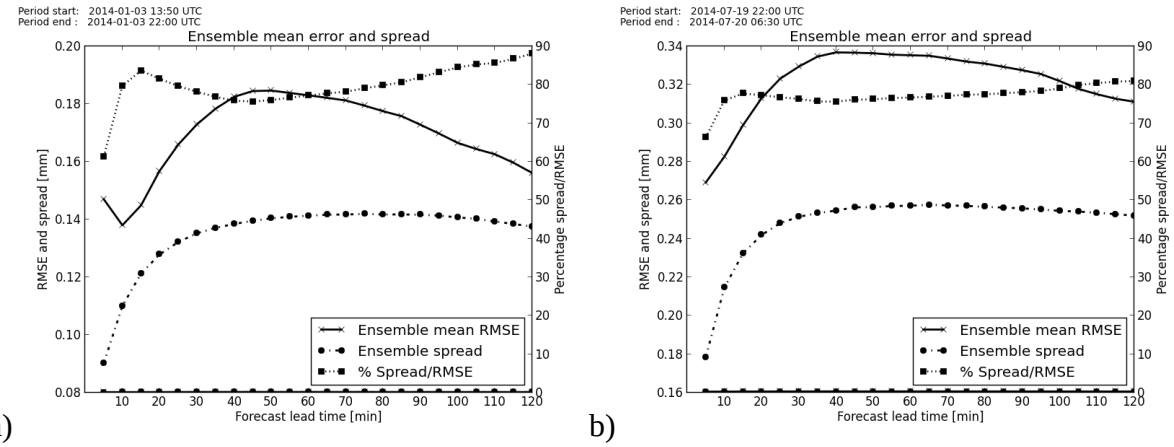


Figure 7. Reliability diagrams for the Ghent case on 03 January 2014 relative to the probabilistic forecast of exceeding a) 0.5 mm hr^{-1} and b) 5.0 mm hr^{-1}. c, d) Same as a,b) but for the Leuven case on 19-20 July 2014.



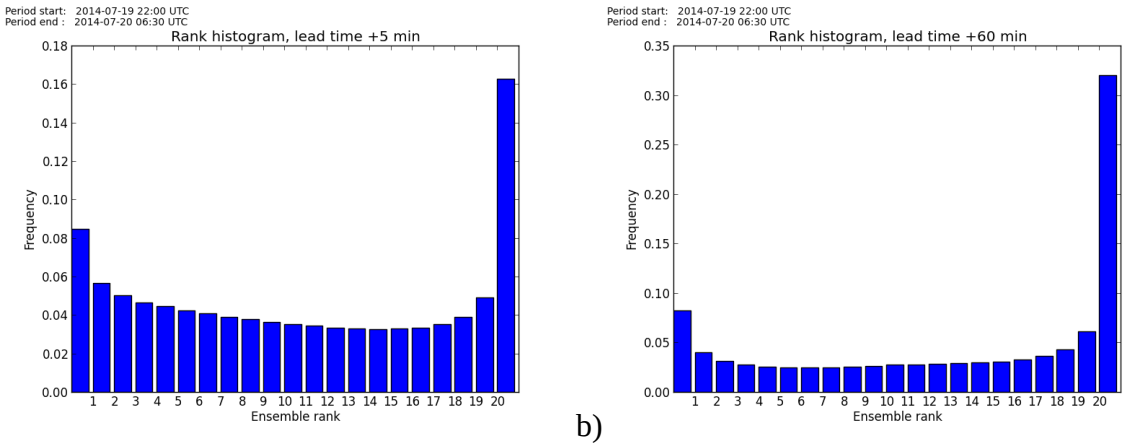
a) b)

Figure 8. ROC curves relative to the probabilistic forecast of exceeding 0.1 mm hr^{-1} for a) the Ghent case on 03 January 2014 and b) the Leuven case on 19-20 July 2014.



a) b)

Figure 9. Comparison of ensemble spread and RMSE of the ensemble mean forecast at 5 min resolution for a) the Ghent case on 03 January 2014 and b) the Leuven case on 19-20 July 2014.



a) b)

Figure 10. Rank histograms for the Leuven case on 19-20 July 2014 for a lead time of a) 5 minutes and b) 60 minutes.