WPT4.3.1:

A Beta Version of a Flash Flood
Early Warning System for
Several Meuse Tributaries

WPT4.3.1: A Beta Version of a Flash Flood Early Warning System for Several Meuse Tributaries

Jerom Aerts,   (TU Delft)
Dr. Ruben Imhoff,   (Deltares)
Athanasios Tsiokanos,   (Deltares)
Prof.dr. Remko Uijlenhoet   (TU Delft)

30-10-2023

Summary

This report investigates the potential of various rainfall forecasting products in accurately predicting flash flood discharge events across several catchments in the Meuse basin. The rainfall forecasting products include Nowcasting, Numerical Weather Prediction, and Blended Forecast. Conceptually, the Blended Forecast product optimally combines the short-term prediction accuracy of the Nowcasting product with the medium-term accuracy of the Numerical Weather Prediction product. The wflow_sbm hydrological model is employed for each rainfall forecasting product to calculate river discharges. This report evaluates each model's effectiveness in offering early warnings for flood events.

Findings reveal that, in general, a substantial benefit of using probabilistic forecasts as opposed to a determinist forecast. The Blended Forecast product provides the earliest warnings during stratiform and convective rainfall events. The combination of stratiform and convective rainfall during the July 2021 flood event, results in a more nuanced picture. With Blended Forecast performing providing early warnings for the Geul and Rur catchments, but to a lesser extent than the Numerical Weather Prediction model for the Vesdre catchment. This emphasizes that the results are case specific, depending on the selected rainfall event and catchment.

The results underscore the potential of integrating Blended Forecasting into operational flood early warning systems, leveraging the combined strengths of the Nowcasting and Numerical Weather Predictions to enhance prediction accuracy, rapid forecast updates, and skillful lead times. To this end, this work package has successfully developed a beta version of a flash flood early warning system for several tributaries of the Meuse and provides recommendations for the operational implementation of such a system for the tributaries of the Meuse.

Table of Contents

1 Introduction

In July 2021, flash floods ravaged the Netherlands, Belgium, and Germany, resulting in over 240 fatalities and an estimated 245 billion USD in damages (AON, 2021; (Koks et al., 2022; Kreienkamp et al., 2021). These floods stemmed from a persistent mesoscale low-pressure system over Northwestern and Central Europe. The affected catchments are characterized by minimal lag times between intense rainfall events (Berne et al., 2004; Gaume et al., 2009). The subsequent surge in water discharge and levels intensified the calamity's scope and destructive impact, see Figure 1.

Figure 1: Flood damages in Schuld, Germany (Source: Reuters).

The July 2021 floods highlight the challenges water managers face, even in countries theoretically equipped to implement robust flood risk adaptation strategies (Jongman, 2018). For improved crisis decision-making, accurate flash flood early warning systems (FFEWS) are essential. These systems predict floods and extend lead times for decision making, bolstering the ability to anticipate and address imminent flash floods. The resulting investments in flood early warning systems (FEWS) can result in very favourable cost-benefit ratios of approximately 1:400 per Euro invested depending on the country and frequency of flood occurrence (Pappenberger et al., 2015). To reduce monetary damages and mortality rates to hazardous events, the United Nations started in 2023 the Early Warnings for All agenda to ensure climate justice and to stress the importance of early warnings (*UNDRR, 2022).*

The Work Package (WP) titled "A beta version of a flash flood early warning system for several Meuse tributaries" is a pilot initiative focused on enhancing current FFEWS. By leveraging recent advancements in short-term rainfall forecasting and hydrologic/hydraulic modelling, the WP constructed a beta FFEWS tailored for multiple Meuse tributaries. This system underwent testing and evaluation to assess its viability for operational deployment. The WP's integration of cutting-edge meteorological and hydrological forecasting techniques holds promise to achieve longer lead times

for decision-makers to act during crises. The WP aims to achieve this by analysing the advantages and limitations of current rainfall forecast products and by leveraging their combined strengths.

Project Deliverable
The deliverable evaluates three rainfall forecasting products and their impact on discharge forecast accuracy during three extreme rainfall events, including July 2021, across five catchments in the Netherlands, Belgium, and Germany. The rainfall forecasting products are used to drive the wflow_sbm hydrological model (van Verseveld et al., 2022) to produce river discharge forecasts. The models, the combination of each rainfall forecasting product and the wflow_sbm hydrological model, are extensively tested to underscore the potential benefits of the proposed FFEWS beta system and its suitability for real-world applications.

The report details the project's methodology, testing, and evaluation of the FFEWS beta system. All data and scripts, compliant with the European Union's FAIR data management principles, are accessible at: https://doi.org/10.4121/a91158d6-887f-4d6c-bc17-fc79d73809e7.v1.

This Work Package (WP) is a segment of the European Interreg EMFloodResilience Project. Key collaborators include Delft University of Technology (NL), Waterboard Limburg (NL), and Waterboard Eifel-Rur (DE).

Activities
To achieve the project deliverables, the following key activities were undertaken:
1. **Data Collection**: Comprehensive data gathering encompassed meteorological, hydrological, and geographical aspects.
2. **Modelling Setup**: The modelling chains were established using three distinct rainfall forecasting products and the wflow_sbm hydrological model.
3. **Testing and Evaluation**: Three historical extreme rainfall events were used to test and evaluate the modelling chains.
4. **Collaboration with Partners**: Active collaboration was fundamental to the project's success:
   o Researchers attended multiple strategic meetings in the Netherlands, Belgium, and Germany, facilitating the exchange of expert knowledge, findings, and data.
   o Waterboard Limburg contributed data vital for the testing and evaluation of the modelling chain.
   o Waterboard Eifel-Rur provided crucial insights to refine the modelling chain specifically for the Rur catchment.
   o The Royal Meteorological Institute of Belgium contributed data vital for creating the rainfall forecast products.

2 Data and Methodology

2.1 Case Study Area

The following five catchments, tributaries of the Meuse River spanning the Netherlands, Belgium, and parts of Germany, have been selected as case studies for the evaluation of the FFEWS beta system (see Figure 2):

Geul (344 km²):

- **Location & Topography**: The Geul catchment stretches across the Netherlands, Belgium, and small parts of Germany, particularly near the three-border region. The catchment is characterized by a pronounced gradient, descending about 250 meters across a span of roughly 60 km.
- **Average Annual Specific Discharge**: The Geul has an average annual specific discharge measuring around 391 mm y^{-1}.
- **Characteristics**: Due to its reliance on rainfall, the discharge from the Geul displays significant variability. During drought conditions, discharge can be as low as 1 m^3 s^{-1}, whereas in flooding scenarios, it might surge to over 40 m^3 s^{-1}.

Rur (2354 km²):

- **Location & Topography**: The primary expanse of the Rur catchment is situated in western Germany, while smaller segments of it stretch into Belgium, accounting for 6.7 % of its overall area, and the Netherlands, making up 4.6 % of the area (Bogena et al., 2018). The elevation within the Rur catchment varies widely. In the south, it reaches 680 meters above mean sea level, while in the north, it is 30 meters above sea level.
- **Average Annual Specific Discharge:** 309 mm y^{-1}
- **Characteristics**:
    - **Upper Catchment (South - Eifel Region)**: This part of the Rur catchment is distinct due to its steep, forest-covered slopes, bedrock foundations, and the presence of large reservoirs.
    - **Lower Catchment (North)**: This northern part is defined by its urban, agricultural, and industrial land use. The region has permeable aeolian deposits and mining pits that affect groundwater levels (Bogena et al., 2005; Pyka et al., 2016).

Demer (2268 km²):

- **Location**: The Demer catchment acts as a tributary to the river Dijle, which subsequently merges with the river Scheldt.
- **Average Annual Specific Discharge**: Demer records an average annual specific discharge of approximately 193 mm y^{-1}.
- **Characteristics**: A noteworthy feature within the catchment is the Hasselt channel. Its presence adds complexity to hydrological modelling related to the broader catchment. As a result, most modelling activities, including this study, often utilize the sub-catchment leading up to Hasselt, which covers an area of around 136 km².

Vesdre (685 km²):

- **Location & Topography**: Within the Belgian Ardennes, the Vesdre catchment is notable for its steep terrain that results in swift hydrological responses.
- **Contribution**: The Vesdre feeds into the Ourthe basin, which subsequently merges with the river Meuse in Liège, Belgium.
- **Average Annual Specific Discharge**: The Vesdre has an average annual specific discharge estimated at 525 mm y^{-1}.
- **Characteristics**: Two key reservoirs, Eupen and La Gillepe, play a significant role in shaping its hydrological behaviour.

Dommel (1691 km²):

- **Location & Topography**: The Dommel catchment has its origins in northern Belgium. It meanders through the Netherlands, ultimately merging with the river Meuse in proximity to 's-Hertogenbosch. The terrain of the Dommel catchment is mainly flat.
- **Average Annual Discharge**: The Dommel catchment displays a mean annual discharge of approximately 261 mm y^{-1}.
- **Characteristics:** The Dommel catchment is characterized by sandy soils and slow hydrological responses.

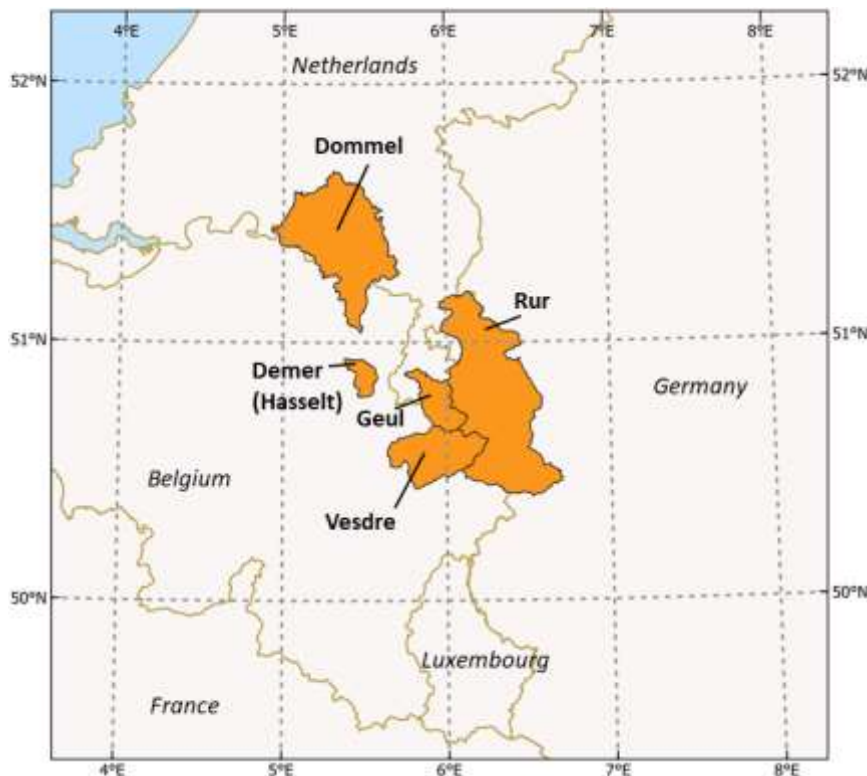The locations and boundaries of these catchments are illustrated in Figure 1.

Figure 2: Locations of the five selected catchments, highlighted in orange. Note: Only the section of the Demer catchment up to Hasselt is displayed, as this specific area is the focus of this report.

2.2 Extreme Rainfall Events

Three extreme rainfall events from 2021 are considered for analysis (namely, events in January, June, and July), each exhibiting a distinct rainfall pattern (stratiform, convective, or a combination of both) that contributed to flooding in varying degrees across the five catchments. Stratiform rainfall occurs when large air masses rise as larger-scale winds and atmospheric dynamics force them to move over each other, resulting in a more uniform cloud cover. These systems are in general less intense and of shorter duration than convective rainfall events, which occur when air rises vertically. The unique attributes of the selected events are outlined in Table 1.

The specific extreme rainfall events of 2021 (Imhoff, 2022):
- **The January event** is recognized as a typical winter stratiform episode. This event caused moderate to high rainfall sums*.*
- **The June event** is marked by convective dynamics, featuring small-scale, concentrated rainfall cells that led to localized (flash) flooding in the western Geul and eastern Demer areas*.*
- **The July event** stands out as an unprecedented and extreme rainfall event, resulting in catastrophic flooding in the larger area. This event is defined by a sustained mesoscale system encompassing both stratiform and convective rainfall patterns (Journée et al., 2023; Tradowsky et al., 2023)*.*

Table 1: Overview of the three extreme rainfall events investigated in this report, table based on Imhoff et al. (2023).

Time event (UTC)		Rainfall Pattern	Catchment-average rainfall sum (mm)				
Start	End		Vesdre	Demer	Geul	Dommel	Rur
27/01/2021 23:00	29/01/2021 09:00	Stratiform (S)	30.6	20.7	27.7	30.6	32.7
29/06/2021 11:30	30/06/2021 11:30	Convective (C)	30.9	54.5	28.8	9.6	34.5
12/07/2021 22:00	15/07/2021 21:00	S/C	131.5	92.0	109.2	101.9	102.0

2.3 Rainfall Forecast Products

This section describes the rainfall forecast products that are evaluated based on the potential for accurately forecasting discharges during extreme rainfall events. First the individual products are briefly described, followed by a summary of the potential advantages and limitations of these products.

2.3.1 Numerical Weather Prediction Models

Introduction to Numerical Weather Prediction models

Numerical Weather Prediction models (NWPs) are computational tools developed to describe complex atmospheric dynamics. These models enable meteorologists to evaluate current weather conditions and predict forthcoming atmospheric changes. Typically functioning on broader spatial and temporal scales, organizations such as the European Centre for Medium-range Weather Forecasts (ECMWF) supervise the development, maintenance and operation of these models. They produce global weather forecasts spanning up to ten days in advance.

Application in Local Meteorological Offices

These broad-scale NWPs are instrumental for local meteorological offices, offering boundary inputs for region-specific predictions. By leveraging smaller-scale NWPs, these offices produce finely-tuned

forecasts targeting particular geographic regions. Such refined predictions offer increased spatial and temporal resolutions, covering periods ranging from +12 to +72 hours. It is therefore important to make note that in this study NWP refers to local meteorological office implementations.

Case Application: The Royal Meteorological Institute of Belgium (RMI)
In the context of this study, the rainfall forecasts were sourced from a NWP system affiliated with the Royal Meteorological Institute of Belgium (RMI). The selection of the RMI-NWP was based on its comprehensive coverage of the study area, relative to other national meteorological institutions. The ALARO variant of the NWP system, an advancement from the ACCORD-developed ALADIN model, is employed by the RMI. Key attributes of this system include a spatial resolution of 1.3 km over a 548x548 grid. The system's forecasts are updated four times daily and forecast up to 48 hours in advance. An operational quality of the ALARO NWP model lies in its temporal precision, capable of 5-minute intervals (Bubnová et al., 1995; Termonia et al., 2018).

Limitations of NWPs in Rainfall Predictions
The NWP models have seen impressive refinements in their predictions. This includes extending lead times for accurate extreme event predictions in the medium-range (Bauer et al., 2015). An improvement in predictive skill is also present for rainfall events with shorter lead times (a window of up to six hours), altough their predictions remain suboptimal for flash flood forecasting. This limitation is particularly evident in smaller catchments or urban contexts (Bowler et al., 2006). Challenges like coarse temporal granularity, infrequent updates, and the ensuing inability to pinpoint the precise timing and location of rainfall pose significant challenges (Berenguer et al., 2012; Lin et al., 2005; Roberts & Lean, 2008). The infrequent update intervals, typically every 3 to 6 hours, coupled with output dissemination delays of 2 to 4 hours, further accentuate the limitations. In dynamic weather scenarios, especially during convective events, this latency can introduce significant errors.

2.2.2 Radar Rainfall Nowcasting
Introduction to radar rainfall nowcasting
Radar rainfall nowcasts (Nowcasting) relies on remotely-sensed quantitative rainfall estimates, using an ensemble of statistical methods to project future rainfall patterns (Pierce et al., 2012). This technique has evolved into commercial ventures, including mobile apps that offer real-time rainfall estimates and short-term predictions. The main strength of nowcasting lies in its ability to utilize high-resolution remotely sensed data, both spatially (around 1 km) and temporally (every 5 minutes) from weather radars (Overeem et al., 2009; Serafin & Wilson, 2000). In contrast to slower NWP models, nowcasting models are fast, producing results rapidly. Instead of relying on a comprehensive numerical modelling of the land-atmosphere system, nowcasting utilizes statistical projections, which (through stochastic perturbation) also allows for ensemble prediction. For convective rainfall cells, nowcasting is most effective for lead times less than 30 minutes, whereas it can predict up to 6 hours ahead using a network of weather radars for prolonged stratiform events on a continental scale.

Methodology
Modern nowcasting employs ensemble techniques, where multiple model runs (ensemble members) with slightly different initial conditions or physics are used. Weather, particularly rainfall, is inherently uncertain due to the chaotic nature of the atmosphere. Small inaccuracies in observations or initial conditions can result in significant differences in predictions, especially over short timescales. A probabilistic approach accounts for this uncertainty by providing a range of possible

outcomes with associated likelihoods. In this study, we employ 48 members that constitute the ensemble.

The Nowcasting methodology can be roughly divided into four stages (Imhoff, 2022):
1. **Advection Field Determination:** This involves gauging the direction of the observed rainfall fields.
2. **Rainfall Development:** A statistical evaluation is performed to track the evolution of rainfall over time.
3. **Projection:** Rainfall fields are extrapolated into the future.
4. **Post-processing:** After forecasting, an interpolation returns the output to the observation grid.

Application in the Work Package (WP)
The WP employs data from the Royal Meteorological Institute of Belgium (RMI) to create its nowcasting model. Specifically, it harnesses radar-based rainfall estimates. These estimates combine data from five distinct C-band weather radars (Imhoff et al., 2023). Before application, radar rainfall estimates undergo bias corrections in order to get closer to the true rainfall amount and in that way become useful for hydrological purposes. The methods for these corrections are detailed in Goudenhoofdt & Delobbe (2016). The Nowcasting forecast product is created using the "STEPS" nowcasting method of the pySTEPS platform (Pulkkinen et al., 2019).

2.2.3 Blended Rainfall Forecast
Introduction to blended rainfall forecasts
This WP conducts a thorough assessment of the advantages and limitations associated with the Nowcasting and NWP models. The aim is to combine the strengths of both products optimally, resulting in the creation of the Blended Forecast product. This product seeks to offer skilful short-term rainfall forecasts crucial for flash flood early warning systems while maintaining accuracy for longer lead times.

Methodology
The technique of merging multiple forecasts is commonly referred to as "blending," as documented in several studies (Atencia et al., 2011; Bailey et al., 2014; Bowler et al., 2006; Golding, 1998; Kober et al., 2012, 2014; Nerini et al., 2019; Radhakrishnan & Chandrasekar, 2020; Yoon, 2019). "STEPS" (Short-Term Ensemble Prediction System) blending is a more specific methodology used in blending radar rainfall nowcasts with longer-term forecasts (Bowler et al., 2006; Seed et al., 2013):

1. **Stochastic Generation**: The STEPS approach involves generating a number of short-term forecasts (ensemble) using radar data (Nowcasting). Each of these ensembles is based on a slight variation of the observed data, introducing a range of possibilities and capturing the inherent uncertainty in the short-term future.
2. **Weighting of Ensembles**: The forecasts created from radar observations have varying probabilities. Those closer to the recent radar observations are given more weight, whereas those that deviate significantly are weighted less.
3. **Combining with NWP**: As the forecast period progresses beyond the immediate short term, the weighted radar nowcast ensembles are gradually blended with outputs from numerical weather prediction models (NWP). This is done to smoothly transition from the high-

resolution, short-term predictions of the radar to the coarser but longer-term predictions of the NWP models.

4. **Dynamic Adjustment**: The STEPS system adapts based on evolving conditions. As new radar observations are made, the ensemble forecasts are updated including their weights, ensuring they remain as accurate and relevant as possible.

5. **Bias Correction**: The latest real-time observations are used to adjust the blended forecasts to align them better with observed rainfall patterns.

The essence of STEPS blending lies in its ability to represent the uncertainty in the immediate future and to combine that with the predictive power of numerical models to provide a seamless and adaptive rainfall forecast product.

Application in the Work Package (WP)
The WP employs the STEPS blending method to capitalize on the skill of the Nowcasting for short-term predictions and statistically blends them, depending on the type of rainfall system, with the better predictive capabilities of the NWP for medium-range forecasts.

2.2.4 Advantages and Limitations of the Rainfall Forecast Products

The forecast innovation central to the FFEWS beta system is rooted in the STEPS blending approach, a method integrating both radar-based rainfall nowcasting models and Numerical Weather Prediction (NWP) models. This approach, termed the "Blended Forecast," leverages the strengths of both aforementioned models to offer improved flash flood early warnings. Similar to the Nowcasting product the Blended Forecast product consists of 48 ensemble members. The main advantages and limitations are:

- **Radar-based Rainfall Nowcasting Models (Nowcasting)**
  - o **Advantages**: Provides real-time or near-real-time monitoring and short-term predictions, making it useful for immediate flood risk assessment and response.
  - o **Limitations**: Typically limited to short lead times, often up to 6 hours, and may not capture the full complexity of larger weather systems.
- **Numerical Weather Prediction Models (NWP)**
  - o **Advantages**: Can offer forecasts for extended periods, thereby giving ample preparation time for potential flood events.
  - o **Limitations**: May not be as precise for immediate short-term predictions as compared to Nowcasting.
- **Blended Forecast Models**
  - o **Advantages**: This methodology seamlessly integrates the Nowcasting model, which is precise for short lead times, with the NWP model that offers forecasts for more extended periods. The combination aims to provide a continuum of reliable predictions, thereby enhancing flood early warning systems' efficiency.
  - o **Limitations**: Sensitive to the weighting of "blending" the Nowcasting and NWP models.

A concise overview of the conceptual advantages and limitations of the three rainfall forecast products can be found in Table 2 and Figure 3.

Table 2: Overview of the conceptual advantages and limitations of the rainfall forecast products.

Rainfall Product	Spatial-Temporal Resolution	Delivery Time	Short Term Forecast	Long Term Forecast
Radar	++	++	--	--
Nowcasting	++	++	++	--
NWP	--	--	-	++
Blended	+	+	++	++
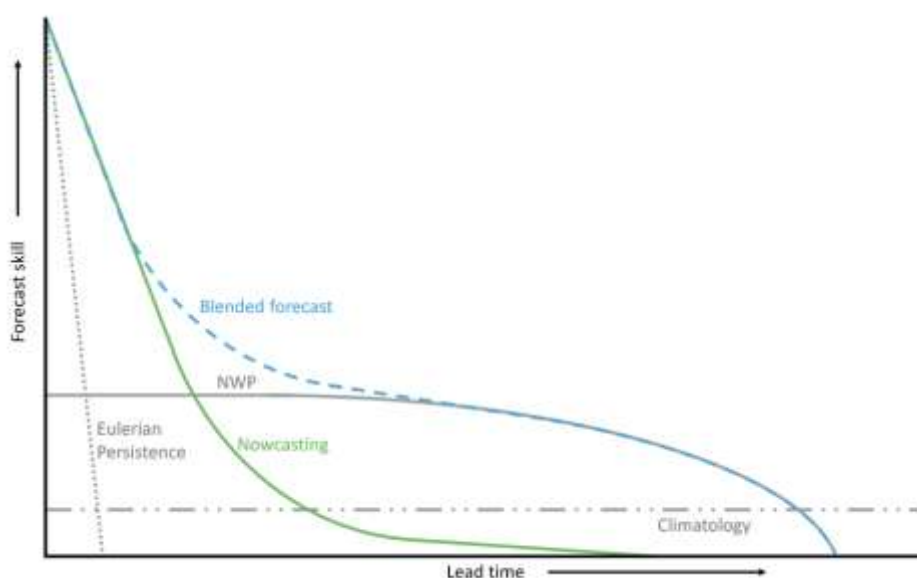
Figure 3: Schematic overview of the rainfall forecast skill as a function of lead time for the rainfall forecast products. In grey the Numerical Weather Prediction (NWP), in green the Nowcasting, and in blue the Blended Forecast. The forecast skill and lead time are case-specific. Note that the Blended Forecast optimally combines the high forecast skill of Nowcasting for short lead time with high forecast skill of NWP with long lead time. Figure is taken from Imhoff (2022).

2.3 Discharge Forecasting

The potential improvement of the Blended Forecast for discharge forecasting is assessed using the distributed wflow_sbm hydrological model (van Verseveld et al., 2022). The term "distributed" refers to the modelling grid, which subdivides the catchment into 1x1 km grid cells. A common alternative approach is to apply a conceptual model that divides the catchment in areas based on the availability of discharge observation stations. The benefit of employing a distributed hydrological model is that the spatial distribution of the rainfall fields is better captured in the numerical calculation of the states and fluxes that ultimately lead to river discharge. The wflow_sbm model is selected due to its integration in existing operational flood early warning systems (FEWS).

Between the start and the end of the rainfall event, every 15 minutes an "issue time", the moment at which the model run starts based on new rainfall predictions, is initiated by the hydrological model

for each rainfall product (Nowcasting, NWP, Blended Forecast). The modelling chain then calculates the discharge forecast 12 hours ahead using a 5-minute temporal resolution (model timestep), see Figure 4.
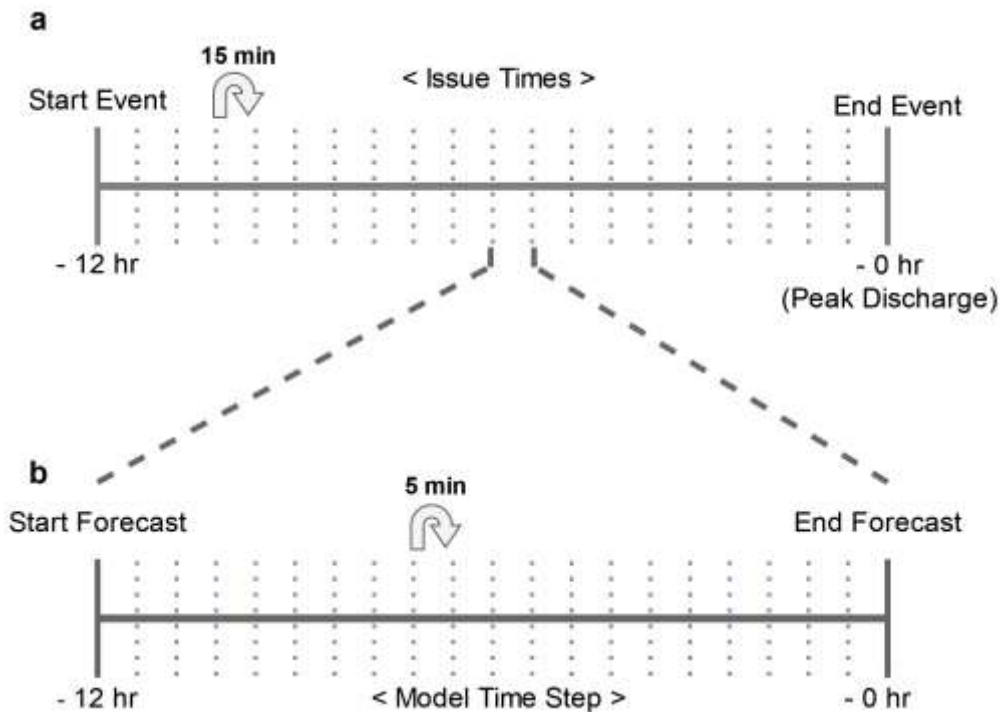

Figure 4: Schematic overview of (a) the temporal discretization of the event in issue times of 15 minutes between the start and the end of the extreme rainfall events. The events are based on the 12 hour period before the peak discharge occurrence. (b) The temporal discretization of the issue time in model time steps showing the 5 minute temporal resolution between the start and the end of the forecast.

To discern the impact of the individual rainfall product's forecast skill on forecasted discharge simulations, a reference discharge simulation is produced. The wflow_sbm hydrological model is driven by the observed bias-corrected radar rainfall data to produce the best estimate of discharge. This reference product serves as the baseline for evaluating the discharge forecasts instead of relying on actually observed discharges. This approach is taken due to the severity of the flood events which resulted in discharge measurement disruptions at observation stations. In addition, it reduces the influence of various sources of uncertainty, such as initial states, model configurations, parameterization, and calibration (Clark et al., 2017). A schematic overview of the modelling chain, that includes the forecasted discharge and reference discharge simulations, is provided in Figure 5.
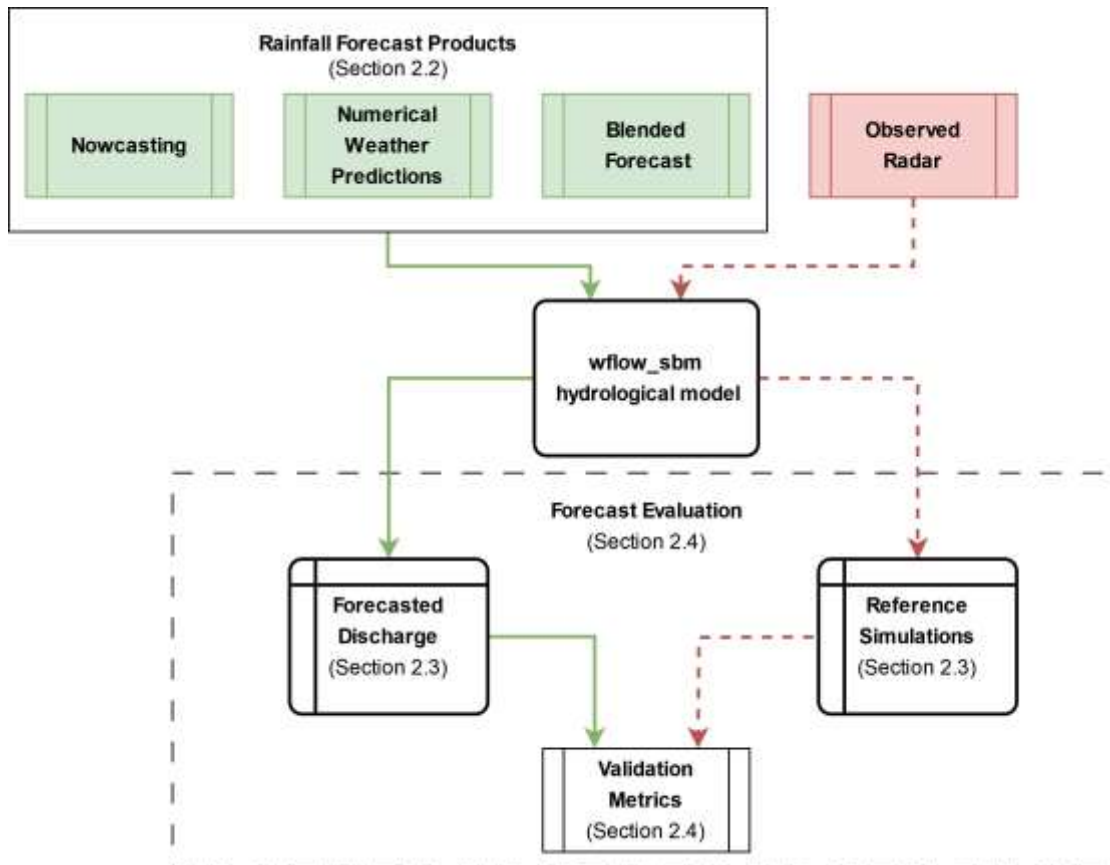
Figure 5: Schematic overview of the modelling chains employed in this study. In green the modelling chain of the rainfall forecast products that produce discharge forecasts is shown. In red the reference discharge simulations modelling chain based on observed bias-corrected radar rainfall is shown. Both the rainfall products and the observed radar rainfall are used to drive the wflow_sbm hydrological model. Subsequently, sets of forecasted discharge and reference simulations are produced that are used to evaluate the modelling chain's forecast skill.

First, reference discharge observations are created using the observed radar data that is bias-corrected using rain gauges, see Figure 5. The forecasted discharges of the wflow_sbm hydrological model based on the three rainfall products (Nowcasting, Numerical Weather Predictions, Blended Forecast) is evaluated against these reference simulations. The evaluation is based on the most downstream discharge observation station for each of the five catchments. The Demer catchment is an exception as the evaluation is based up to the Hasselt station. Table 2 shows the station used for each catchment and their respective upstream areas.

Table 2: Discharge stations used for validation per catchment and their respective upstream areas.

Catchment	Station used	Upstream area (km^2)
Demer	Hasselt	136
Geul	Meerssen	323
Vesdre	Chaudfontaine	685
Dommel	Vught	1691
Rur	Stah	2346

The nature of the three discharge forecast models differ. The NWP forecasts are deterministic, meaning that they only contain one outcome per forecast. The Nowcasting and Blended Forecast are ensemble forecasts (probabilistic), with 48 different realizations (forecasts) made to take the uncertainty of the forecast into account. The assessment of discharge forecast quality is based on hydrographs, timeseries of discharge observations and simulations, of each issue time between the start and end of the rainfall event. To illustrate, Figure 6 depicts the hydrograph of the June 2021 flood event within the Geul catchment at the Meerssen observation station.
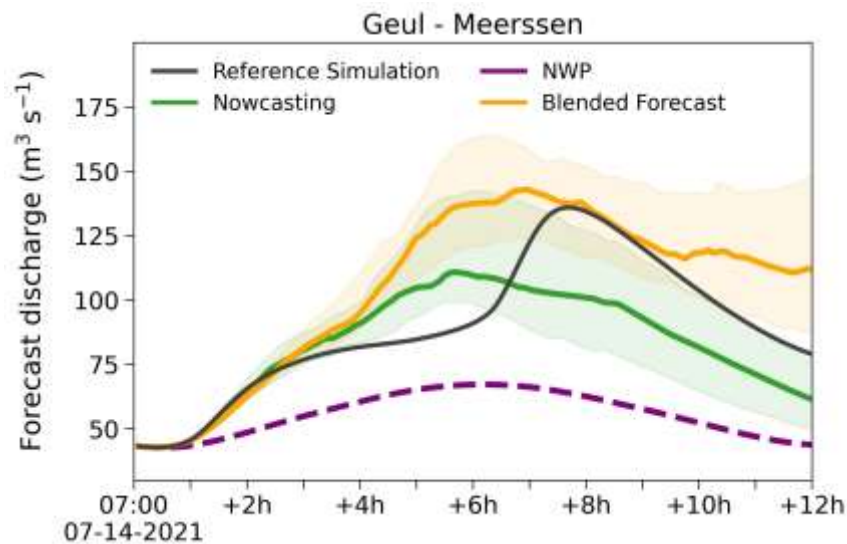
Figure 6: Example hydrograph of the July 2021 flood event in the Geul catchment at the Meerssen discharge measurement station. The reference discharge simulation is indicated in black, with the discharge peak at approximately 14:00 UTC on 07-14-2021. The deterministic NWP, the dashed purple line, shows an underestimation compared to the reference simulation. The probabilistic forecast products show the ensemble median in bold colors and the shading 25th and 75th percentile of the ensemble distribution. Orange indicates the Blended Forecast and green the Nowcasting.

Figure 6 portrays the deterministic essence of NWP forecast (dotted purple line) and reference observations (black line), presented as single lines. The Nowcasting and Blended Forecast adopt a probabilistic ensemble approach that exhibit a spectrum of potential discharge values. This results in a plume of predictions due to considering 48 distinct model realizations. The median values, indicative of a 50 % consensus among realizations, are represented by the bold lines, while the extent of variation between the 25th and 75th percentiles is depicted by the lighter shading.

2.4 Validation Metrics

Multiple validation metrics are used to determine the quality of the FFEWS discharge forecasts. These metrics can be categorized into three groups: (1) deterministic skill-based metrics, (2) probabilistic skill-based metrics, and (3) contingency skill-based metrics. By conducting the evaluation based on these three groups, the WP determines the match between simulations and observations as well as the skill in predicting the peak discharges.

2.4.1 Deterministic Forecast Skill Metric

The mean absolute error (MAE) is a metric utilized to measure the accuracy of predictive models by computing the average magnitude of errors between predicted and reference or observed values. It offers a linear score, where all individual differences have equal weight, making it intuitive and easy to interpret.

Mathematically, the MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad\qquad \text{(Eq. 1)}$$

Where:
   n is the total number of observations.
   $\hat{y}_i$ represents the predicted values.
   y_i represents the reference observed values.

The summation iterates over all observations, calculating the absolute difference between the predicted and actual value for each observation. The MAE then averages these absolute differences over all the observations. A smaller MAE indicates a higher accuracy of the prediction, while a larger MAE suggests that the model's predictions may be off-target.

2.4.2 Probabilistic Forecast Skill Metric

The brier threshold score (BTS) measures the accuracy of probabilistic forecasts for binary events, specifically in relation to a predefined threshold. This threshold is defined here as 90 % of the peak discharge reference simulation value per catchment and extreme rainfall event. It assesses how well the forecasted probabilities align with the actual outcomes given an exceedance threshold.

Mathematically, the BTS is defined as:

$$BTS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \qquad\qquad \text{(Eq. 2)}$$

Where:
   n is the total number of forecasts.
   f_i is the forecasted probability, the fraction of the 48 ensemble members for probabilistic forecasts, for the ith event, exceeding the predefined threshold.
   o_i is the reference observed outcome for the ith event, coded as 1 if the event exceeds the threshold and 0 otherwise.

The BTS is calculated for each simulation time step (5 minutes) the exceedance threshold of 90 % peak discharge is surpassed before averaging over the issue time. This exceedance can consist of multiple time steps within an issue time. The BTS ranges from 0 to 1, where a lower score indicates better forecast accuracy. A perfect forecast would have a BTS of 0, while a BTS of 0.25 suggests the forecast is not better than a random guess. The Brier Threshold Score is particularly useful when forecasters or decision-makers are concerned with the occurrence of an event surpassing a specific value or level, such as rainfall exceeding a specific threshold.

2.4.3 Peak Anticipation Time Metric

The ability to accurately forecast peak or threshold exceedances in water systems significantly ahead of their occurrence is crucial for water management. This anticipation time facilitates preparatory measures, enhancing flood risk management and safeguarding resources. Within this study, the peak anticipation time (PAT) metric serves as a primary tool for gauging the accuracy of the forecasting techniques.

For each issue time a 12-hour forecast is produced and evaluated using the PAT metric, see colored lines in Figure 7. To determine the match between forecast simulations and reference discharge a window is created around the discharge peak that encompasses 90 % to 110 % of this peak discharge. This 10 % margin is assumed as it is generous enough to not overly penalize the forecast skill but strict enough to remain relevant for decision-making.

When we utilize probabilistic forecasting methods, individual ensemble members are evaluated separately. This results in each issue time containing a variety of anticipation times, stemming from the different ensemble predictions. To express the accuracy of these ensemble forecasts, we calculate the percentage of ensemble members that correctly anticipate the peak discharge for a given issue time, terming this the "percentage of agreement" or consensus.
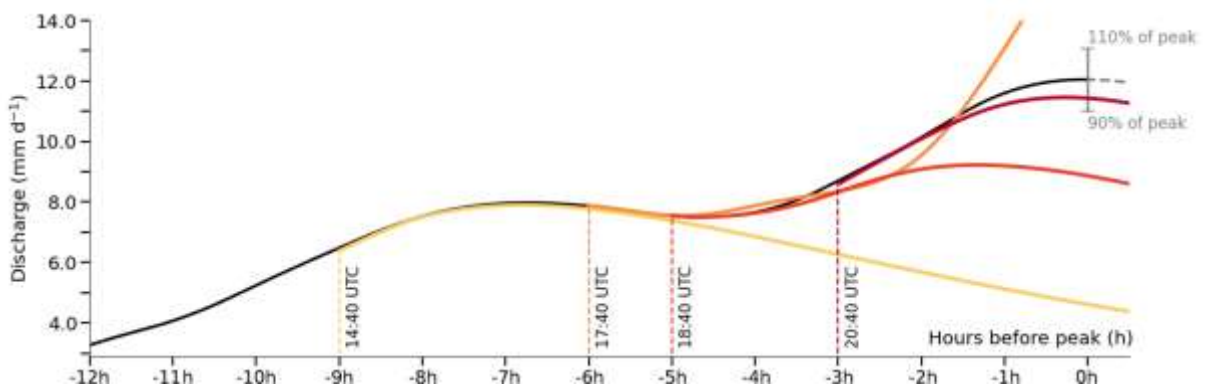

Figure 7: Example illustrating the peak anticipation time (PAT) method. The black line represents the reference discharge simulation. The colored lines indicate forecast for four different issue times: (a) yellow: 14:40 UTC-9 hours before the peak; (b) orange: 17:40 UTC-6 hours before the peak; (c) red: 18:40 UTC-5 hours before the peak and (d) burgundy: 20:40 UTC-3 hours before the peak. The 10 % error margin for characterizing a "correct" peak discharge forecast is shown in the upper right corner (adapted from Imhoff et al., 2022).

3 Results

This section begins by presenting the discharge reference simulations based on the three selected extreme rainfall events across the five selected tributaries of the Meuse. This is followed by an assessment of the deterministic discharge forecast skill of the three rainfall forecast products (Numerical Weather Prediction (NWP), Nowcasting, Blended Forecast) using the median value of the ensemble prediction. Subsequently, the probabilistic discharge forecast skill based on the whole ensemble prediction is assessed. Lastly, the section concludes by examining the peak anticipation time (PAT), which quantifies the lead time required for an accurate prediction of the peak discharge. Appendix A contains more analyses based on contingency statistics.

3.1 Discharge Reference Observations

The wflow_sbm hydrological model, driven by gauge-adjusted radar rainfall data, is employed to derive discharge reference simulations (detailed in Section 2.3). In Figure 8, see next page, the discharge reference simulations for the three distinct rainfall events within the five catchments are shown. These catchments showcase diverse responses to heavy rainfall as is exemplified by the fast responding Geul and Vesdre catchments in Figures 8a,b and the more gradual reacting Dommel catchment in Figure 8e. Similar response patterns are found for the Geul, Rur, and Vesdre catchments during the stratiform January 2021 event in Figures 8a,b,d, albeit with varying discharge volumes. The same catchments contain a characteristic double peak in the discharge time series of the July 2021 event (Figures 8k,l,n). Note that the upcoming analysis excludes the June 2021 event of the Dommel due to its unresponsive nature during this event, as shown in Figure 8j, due to the absence of a significant rainfall event during this period.
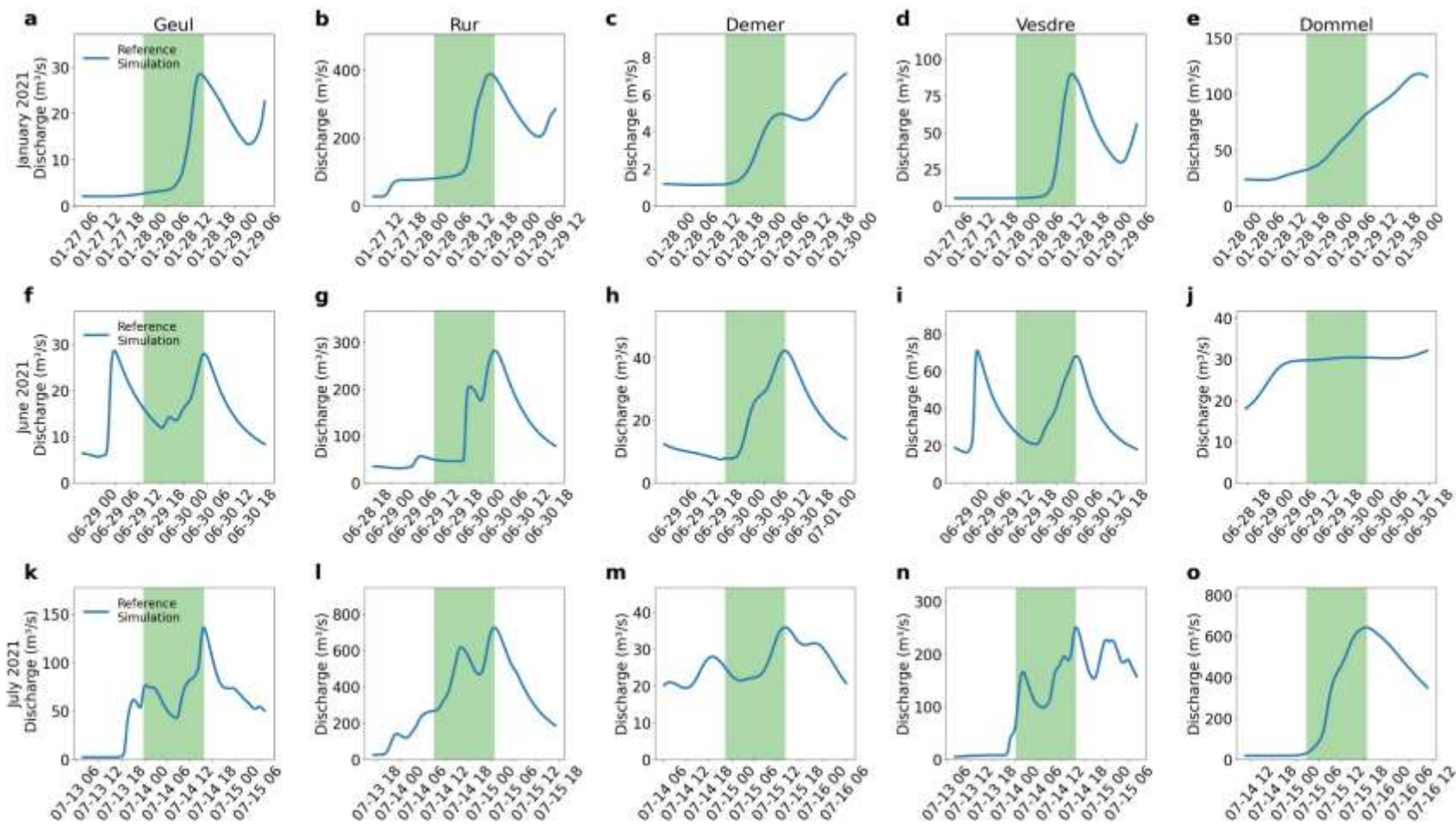
Figure 8: The discharge reference simulations obtained by employing the gauge-adjusted radar rainfall data as input to the wflow_sbm hydrological model are shown. For the January stratiform rainfall event, panels (a-e) present the estimated discharge reference simulations, with the estimates highlighted in blue and green shading indicating the chosen events. The June convective rainfall event's discharge reference simulations are displayed in panels (f-j). Panels (k-o) portray the discharge reference simulations concerning the July 2021 flood event, encompassing both stratiform and convective rainfall.

20

3.2 Deterministic Forecast Skill

The advantages of integrating Nowcasting with Numerical Weather Predictions (NWP) to form the Blended Forecast are assessed using the mean absolute error (MAE) metric (see Section 2.3.1). The calculation of this metric involves employing the probabilistic ensemble medians of both the Nowcasting and the Blended Forecast, while the NWP already functions as a deterministic model. The outcomes of this analysis provide insights into the extent of forecast errors in relation to the reference simulations. This assessment addresses the level of skill present, the magnitude of volume disparities, and the comparative performance of the discharge forecasting models.

The MAE results are shown in Figure 9. When the Nowcasting and Blended Forecast models transition from one issue time to the next, there is no temporal memory retained due to the model's initialization relying on the most recent observations. In contrast, the NWP forecasts possess temporal memory as the updates to the model's initialization are less frequent. This update frequency is shown by the discontinuities between successive issue times, exemplified by the leap at the 8-hour mark in Figures 8a,g,l,m.

The MAE indicates the difference in discharge volume in m^3s^{-1} between the predictions and reference simulations. In general, the discharge forecasts of the Demer and Dommel catchments remain within a 10 % MAE margin compared to the peak discharge volume of the reference simulations (see Figure 8). The estimates for the Rur catchment remain within 20 % MAE difference, while the Geul and Vesdre catchments exhibit more substantial differences.

The results of the forecast models for the stratiform January 2021 rainfall event are shown in Figures 9a-e. The Geul catchment results, Figure 9a, show a lower MAE at shorter lead times, between +2hr and +5hr, for the Nowcasting and Blended Forecast than the NWP model. The better forecast skill of NWP model for longer lead times (> +5hr) is not fully utilized by the Blended Forecast, although a small improvement compared to the Nowcasting model is present. The results of the Rur catchment in Figure 9b show that the Blended Forecast model performs best for the longest lead times (> +8hr). The Nowcasting and Blended Forecast models result in low MAE values until the +6hr mark, showing that these models are capable of predicting discharges in slower responding catchments under stratiform rainfall conditions. However, all models show large errors between consecutive lead times. The low MAE values for the Demer and Dommel catchments in Figure 9c,e indicate that all models predict accurately compared to the reference simulations. The NWP model has the lowest MAE values overall for the fast responding Vesdre catchment in Figure 9d. The Nowcasting and Blended Forecast models contain much higher MAE values beyond +6hr lead times for this catchment.

The June 2021 convective rainfall event results in Figures 9f-i, show higher MAE values than for the stratiform event. The NWP model performs best overall for the Geul catchment (Figure 9f). The Nowcasting and Blended Forecast models contain very large errors beyond the +11hr lead time. The benefits of employing Nowcasting or Blended Forecast models is evident for the Rur catchment between +4hr and +8hr lead times (Figure 9g). Here, the NWP model contains very high MAE values after the NWP is updated at +7.5hr lead time. The Demer catchment results in Figure 9h show that the Blended Forecast performs has the lowest MAE values until the +7hr lead time. Similarly, the Blended Forecast has the lowest MAE values for the Vesdre catchment until the +7hr lead time (Figure 9i).

The July 2021 event results, the combination of stratiform and convective rainfall, in Figures 9j-n show similar MAE values as for the convective June 2021 rainfall event. The relative error is smaller for the July 2021 event as the predicted discharge volumes are larger (see Figure 8). Therefore, the overall performance is better for the July than for the June 2021 rainfall event. Considering the individual catchments, the Geul catchment results in Figure 9j indicate that the forecast models perform similarly until +6hr lead time, with high MAE values present for the Nowcasting and Blended Forecast models at +6.5hr to +7hr lead times. The NWP model contains the lowest MAE values beyond the +6hr lead time. For the Rur catchment in Figure 9k, the Blended Forecast shows the lowest MAE, the exception is the NWP model between +6hr and +9hr lead times. The Demer and Dommel catchment results in Figure 9l,n show lower errors for the Nowcasting and Blended Forecast models compared to the NWP model until the +11hr lead time. The results of the Vesdre catchments in Figure 9m show lowest MAE values for the NWP model beyond the +3hr mark. This indicates that, based on the deterministic analysis, the NWP is expected to perform better than the Blended Forecast in this catchment.

Summary
The Nowcasting and Blended Forecast models exhibit improved forecast skill for short lead times in comparison to the NWP, while the latter shows more proficiency for predicting longer lead times (> +7hr). Specific catchments, such as the Rur catchment and the Geul and Demer within specific time windows, do derive benefits from the blending approach at longer lead times. The challenge of accurate prediction during convective or combined rainfall events is underscored by the larger MAE values, particularly for longer issue times. The MAE analysis gives a good first estimate of the error between forecasts and reference simulations but does not consider the benefits of probabilistic forecasting.
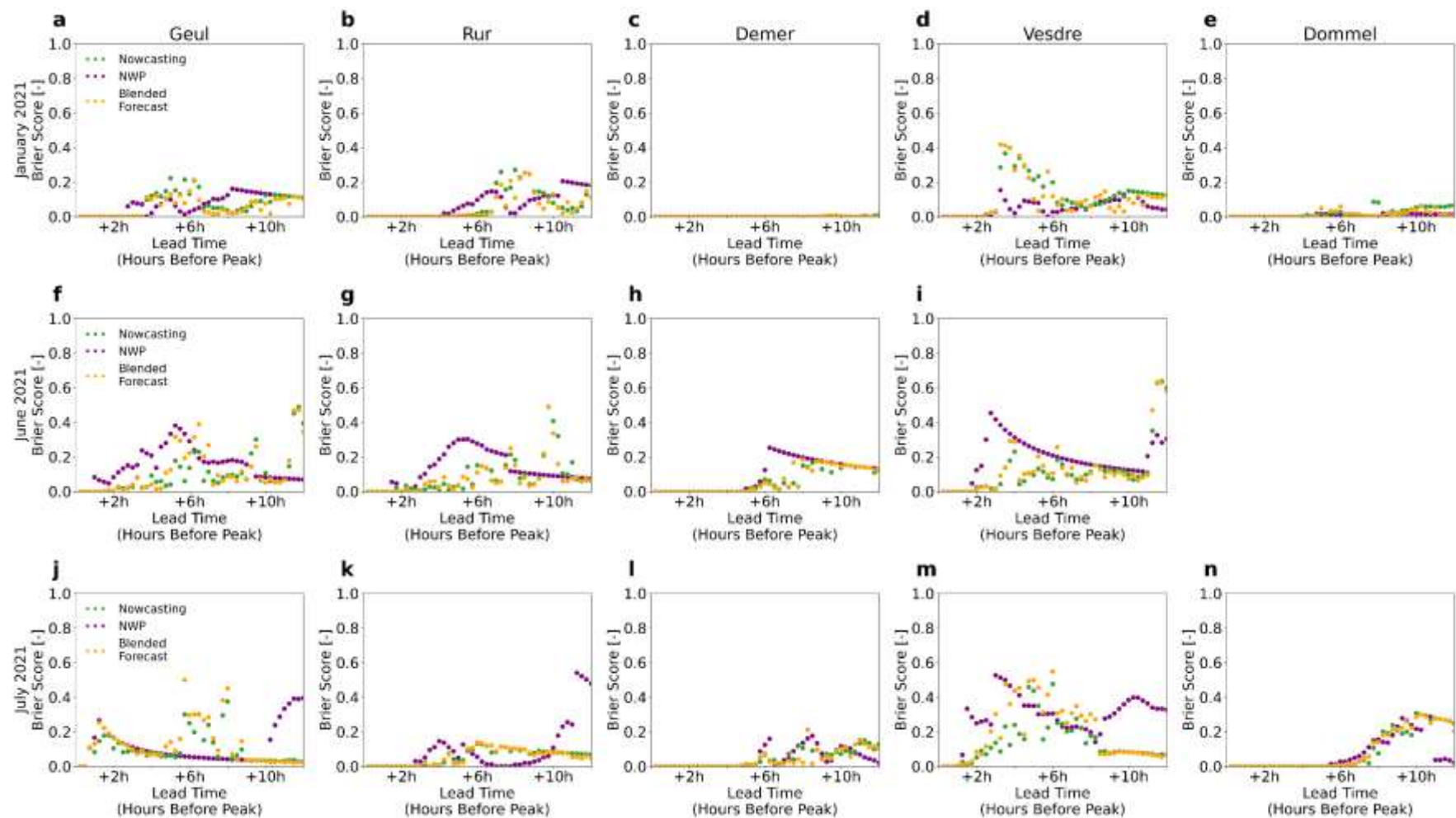
Figure 9: The mean absolute error (MAE) in cubic meters per second ($m^3 s^{-1}$), with on the horizontal axis the lead time in hours before peak discharge. The Nowcasting model is indicated in green, the NWP model in purple, and the Blended Forecast model in orange. For the January stratiform rainfall event, panels (a-e) present the MAE results. The June convective rainfall event's MAE results are displayed in panels (f-i). Panels (j-n) portray the MAE results concerning the July 2021 flood event, encompassing both stratiform and convective rainfall.

3.3 Probabilistic Forecast Skill

The Brier threshold score (see Section 2.3.2) is used to assess whether the discharge forecasts achieve an exceedance of 90 % of the reference simulation peak discharge. This evaluation addresses the forecast model's capacity to predict the discharge peak and to ascertain if ensemble members agree on this prediction. The Brier threshold score spans values from the optimal scenario of 0, indicating unanimous correct agreement among ensemble members, to the least favourable scenario of 1, denoting unanimous disagreement. The Brier score encompasses the entire ensemble forecast by averaging over all ensemble members and by averaging over the time steps during which 90 % of the reference simulation peak discharge is exceeded. In doing so, the consistency of correctly exceeding the threshold is addressed.

The outcomes shown in Figure 10 underscore across all rainfall events, catchments and discharge forecast models that there are no instances where all ensemble members fail to exceed 90 % of the reference simulation peak discharge. In a broader context, the results reveal better Brier scores (lower values) for the stratiform rainfall event (Figures 10a-e) compared to convective events (Figures 10f-i) or the combined impact of both rainfall types (Figures 10j-n). This is in line with previous findings based on the deterministic MAE analysis in Figure 9.

The results from the stratiform event in January 2021 for the Geul catchment, as presented in Figure 10a, reveal that both the Nowcasting and Blended Forecast model ensembles surpass the NWP in capturing the exceedance threshold. This is especially evident during the shortest (up to +3hr) and the longest (beyond +6.5hr) lead times, aligning again with the MAE results in Figure 9a. Similar to these results, there is a notable dip in performance for the Nowcasting and Blended Forecast models between +4hr and +6.5hr lead times. This is also present between +7hr and +9hr for the Rur catchment (Figure 10b) and between +3hr and +5hr for the Vesdre catchment (Figure 10c). These shifts in Brier scores are worrying, implying diminished accuracy hours before the peak discharge. One plausible reason might be that accurate forecasts with shorter issue times fall within the catchment response times, making them essentially an extrapolation of the current catchment response to rainfall. As such, the advantages of precise forecasts emerge more prominently after the catchment response times have elapsed.

Besides the Dommel catchment results shown in Figure 10e, both the Nowcasting and Blended Forecast models consistently show lower Brier scores than the NWP model for longer lead times. The opposite was determined by the deterministic MAE findings. This underscores the value of probabilistic ensemble forecasting in capturing the reference simulation discharge peak, when compared against the deterministic approach of the NWP forecast.

Examining the results of the convective rainfall event (Figures 10f-i), the Nowcasting and Blended Forecast models contain lower Brier scores than the NWP model, except for a few prolonged lead times. This distinction in Brier scores is less pronounced for the combined rainfall event (Figures 10j-n), where the NWP model has lower Brier scores for the Geul, Rur, and Dommel catchments (Figures 10j,k,n) than the Nowcasting and Blended Forecast models, except for the Vesdre catchment (Figure 10m) and longer lead times of the Geul and Rur Catchments (Figures 10j,k). The latter are especially noticeable improvements compared to the NWP model.

Summary

The probabilistic forecast skill analysis, Figure 10, highlights that across all rainfall events, catchments, and discharge forecast models, none of the ensemble members fall short of exceeding 90% of the reference simulation peak discharge. Overall, better Brier scores (lower) are found for the stratiform rainfall event (Figures 10a-e) compared to convective events (Figures 10f-i) or the combination of both (Figures 20j-n). The results in Figures 10a,b emphasize the Nowcasting and Blended Forecast model's better performance in capturing the exceedance threshold compared to the NWP model, particularly for longer lead times (> +7hr). This demonstrates the advantage of probabilistic ensemble forecasts over deterministic ones.

For convective rainfall (Figures 10f-i), the Nowcasting and Blended Forecast models generally outperform the NWP, except for a few longer lead times. This distinction is somewhat subdued for the combined event (Figures 10j-n), where the NWP model excels over the Nowcasting and Blended Forecast models, except for the Vesdre catchment (Figure 10m) and longer lead times of the Geul and Rur Catchments (Figures 10j,k). This establishes that the findings are rainfall event and catchment specific.
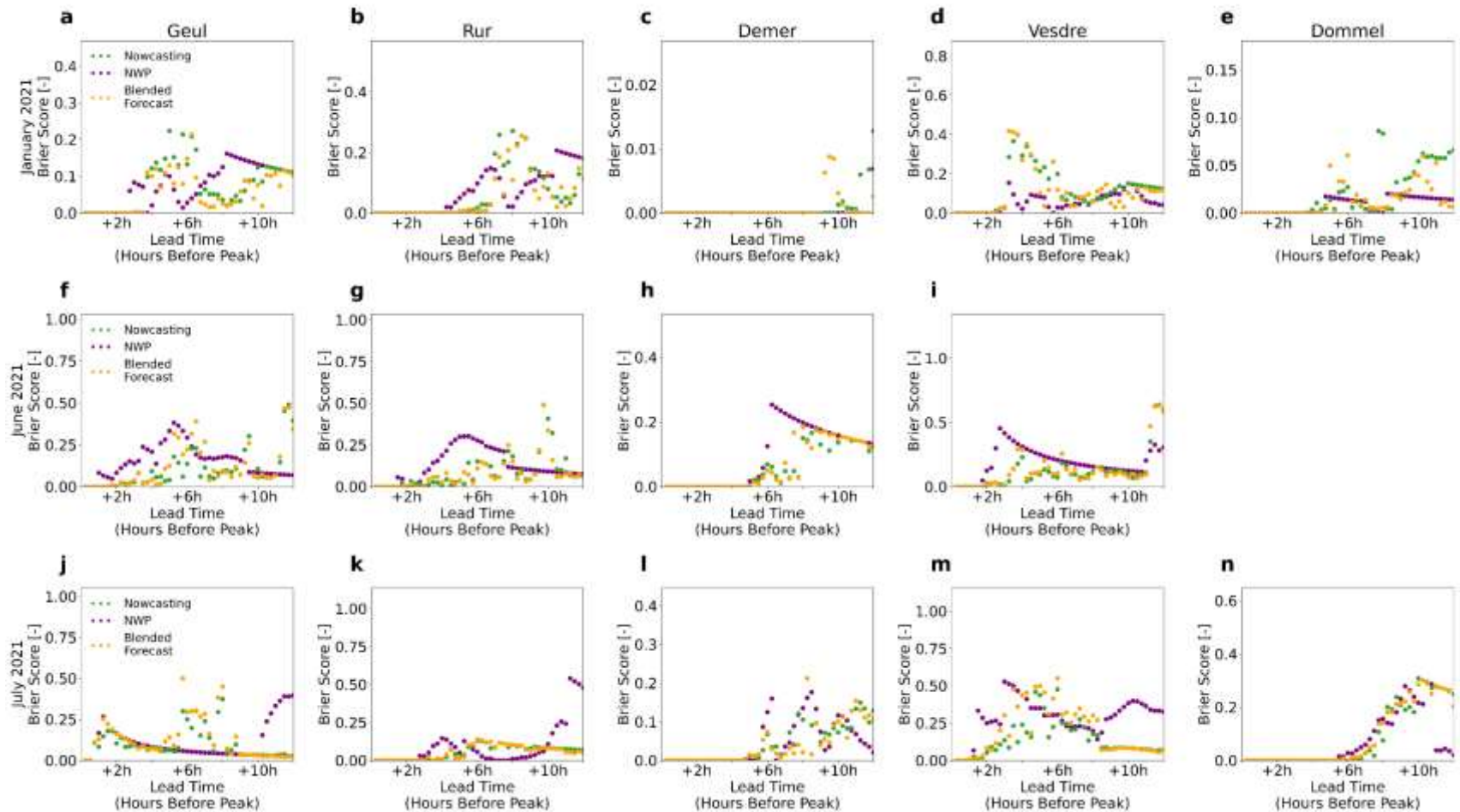
Figure 10: The Brier score results are shown for the 5 catchments, 3 rainfall events, and 3 discharge forecast models. The Brier score uses a threshold set at 90 % of the reference simulation discharge peak. A Brier score of 0 signifies the optimal prediction scenario, while a score of 1 represents the least favourable outcome. The Nowcasting model is indicated in green, NWP in purple, and the Blended Forecast in orange. Panels (a-e) display the stratiform January 2021 event results. Panels (f-i), show the June convective rainfall event results. Panels (j-n) focus on the July 2021 flood event, encapsulating both stratiform and convective rainfall.

3.4 Peak Anticipation Time

The Peak Anticipation Time (PAT; see Section 2.3.3) plays a pivotal role in assessing the benefits of adopting a Blended Forecast model for FFEWS. Other contingency based statistics include the hit rate, false alarm ratio, and critical success index, which are presented in Appendix A. The assessment in this section leverages the entire ensemble (comprising 48 realizations) to ascertain the lead time for predicting peak discharge of the reference simulations. Strong consensus within the prediction of peak discharge indicate likelihood of a flood event. Figure 11 shows the PAT in hours on the vertical axis and the three discharge forecast models on the horizontal axis for three rainfall events in the Geul catchment. The PAT, expressed in percentages, reflects the consensus among ensemble forecasts regarding peak discharge occurrence. Note that the NWP functions as a deterministic model, resulting in either 0 % or 100 % agreement.

The interpretation of probabilistic forecasts is difficult as it is can be subjective at which confidence level (percentage of ensembles in agreement) decision-makers act. For the following analysis we base our interpretation on the actions and personal communications with the Rijkswaterstaat fluvial flood forecasting service team. They faced multiple challenges during the July 2021 flood event, such as confirmation bias when visual confirmation of a flood was not yet possible (e.g. Demeritt et al., 2010). This resulted in attempts to discredit the forecast. A second bias surrounded the decision whether or not to initiate certain costs/disadvantages based on an uncertain forecast. Actions were taken by the forecast team based on 5 out of 51 ensemble members indicating a severe flood event, a 10 % agreement.

In Figure 11a the PAT results of the stratiform January 2021 rainfall event of the Geul catchment are shown. The NWP model has a PAT of +8hr before the discharge peak. This anticipation time is surpassed by the Nowcasting model, with an ensemble agreement of 29 % at +10.15hr and the Blended Forecast model with 31 % agreement at +11.15hr. Although the Nowcasting and Blended Forecast ensembles are less consistent over time in PAT agreement percentages than the deterministic NWP, a third of the ensemble agreeing on peak discharge is a clear indicator of a likely occurrence of a high discharge event. While longer PATs exhibit lower agreement for the Nowcasting and Blended Forecast models, they do provide water managers with an early indication of a potential significant event. The following interpretation assumes that a probability of 10 % is sufficient for authorities based on the actions of Rijkswaterstaat during the July 2021 flood event.

The convective June 2021 rainfall event results in Figure 11b show that the Nowcasting and Blended Forecast models exhibit at approximately +10hr PAT, an agreement percentage above 10 %. This is 5.5hr before the +4.5hr PAT of the NWP model. In addition, a gap of 45 minutes of no agreement on peak discharge occurs for the NWP model between +3.15hr and +4hr PAT. This gap is not due to the update frequency of the NWP model but due to an underestimation of the peak discharge.

A similar but larger gap in agreement between consecutive PAT is found for the NWP model during the combined July 2021 rainfall event results in Figure 11c. Starting with an early warning between +10.5hr and +12hr PAT, the following warning occurs only +1hr PAT before the peak discharge. Given the focus of this analysis on the discharge peak, an early warning might have occurred as can be established by the relatively good MAE results in Figure 9j. Although the Nowcasting and Blended Forecast models fall short of a clear early warning, the Nowcasting model has a 10 % agreement on PAT at +8.5hr and the Blended Forecast model at +11.5hr. Nonetheless, both models show inconsistent and low probabilities leading up to +1hr PAT.
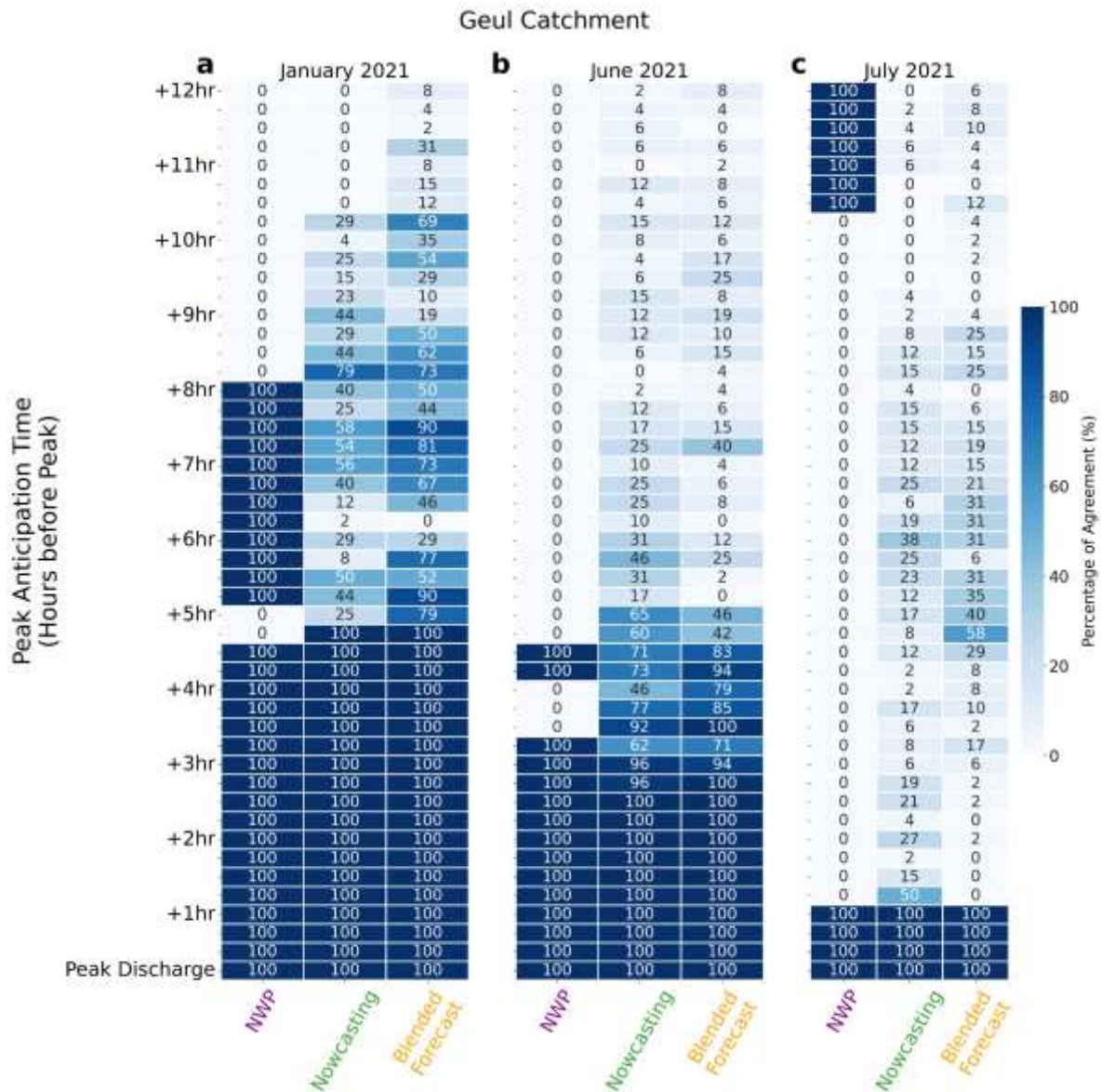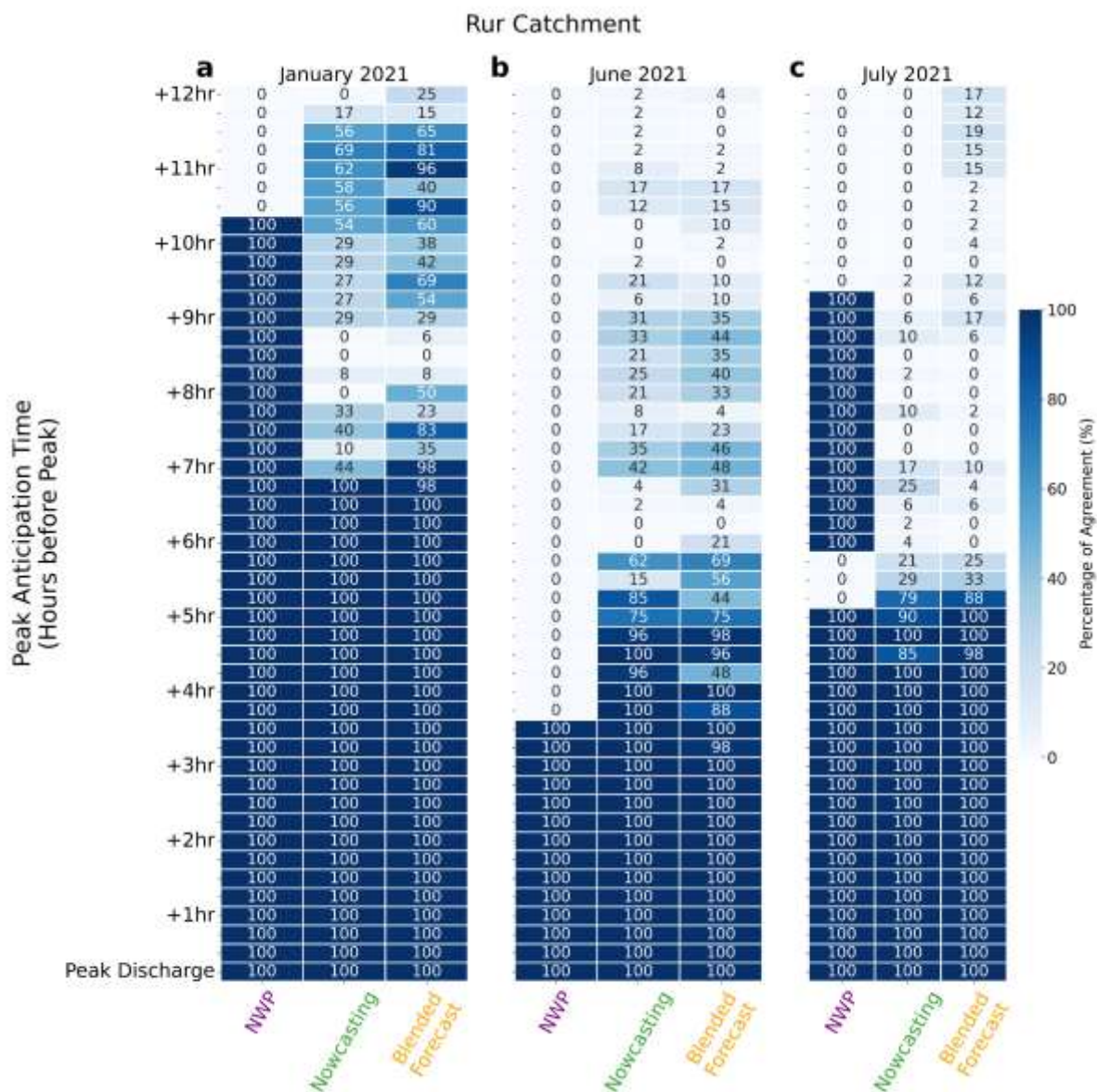
Figure 11: Heatmap displaying the peak anticipation time (PAT) in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. The number of ensemble members agreeing on the PAT per issue time ranges from 0 % in white to 100 % in dark blue. The panels show the results for the Geul catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Next, we consider the Rur catchment results in Figure 12. The Nowcasting and Blended Forecast show strong consensus on a PAT of +11.5hr in Figure 12a. One hour before the NWP model at +10.25hr PAT. Both the Nowcasting and Blended Forecast models experience a decline in agreement between +8.45hr and +8hr PAT. These results show that all models are capable of producing PATs with long lead times in larger catchments under stratiform rainfall conditions.

The opposite is the case for the convective rainfall conditions in Figure 12b. Here, the NWP model exhibits a shorter +3.5hr PAT. For the Nowcasting and Blended Forecast models there is an initial

early warning at +10.45 PAT with 17 % ensemble agreement. Strong consensus on the occurrence of an event with approximately 30 % agreement is present at +9hr PAT. This demonstrates the benefit of employing a probabilistic approach for forecasting under convective rainfall conditions.

The combined July 2021 rainfall event in Figure 12c shows a more balanced result with the NWP model indicating a PAT of +9.15hr. No early warning is given by the Nowcasting model until +9hr PAT, followed by inconsistent agreement percentages up to +5.45hr PAT. The Blended Forecast provides an early warning at +12hr PAT with percentages of agreement ranging between 12 % and 19 %. Stronger consensus between ensemble members occurs at +5.45hr PAT. This demonstrates the benefit of employing a Blended Forecast compared to a Nowcasting approach, as it also takes advantage of the information in the NWP product.
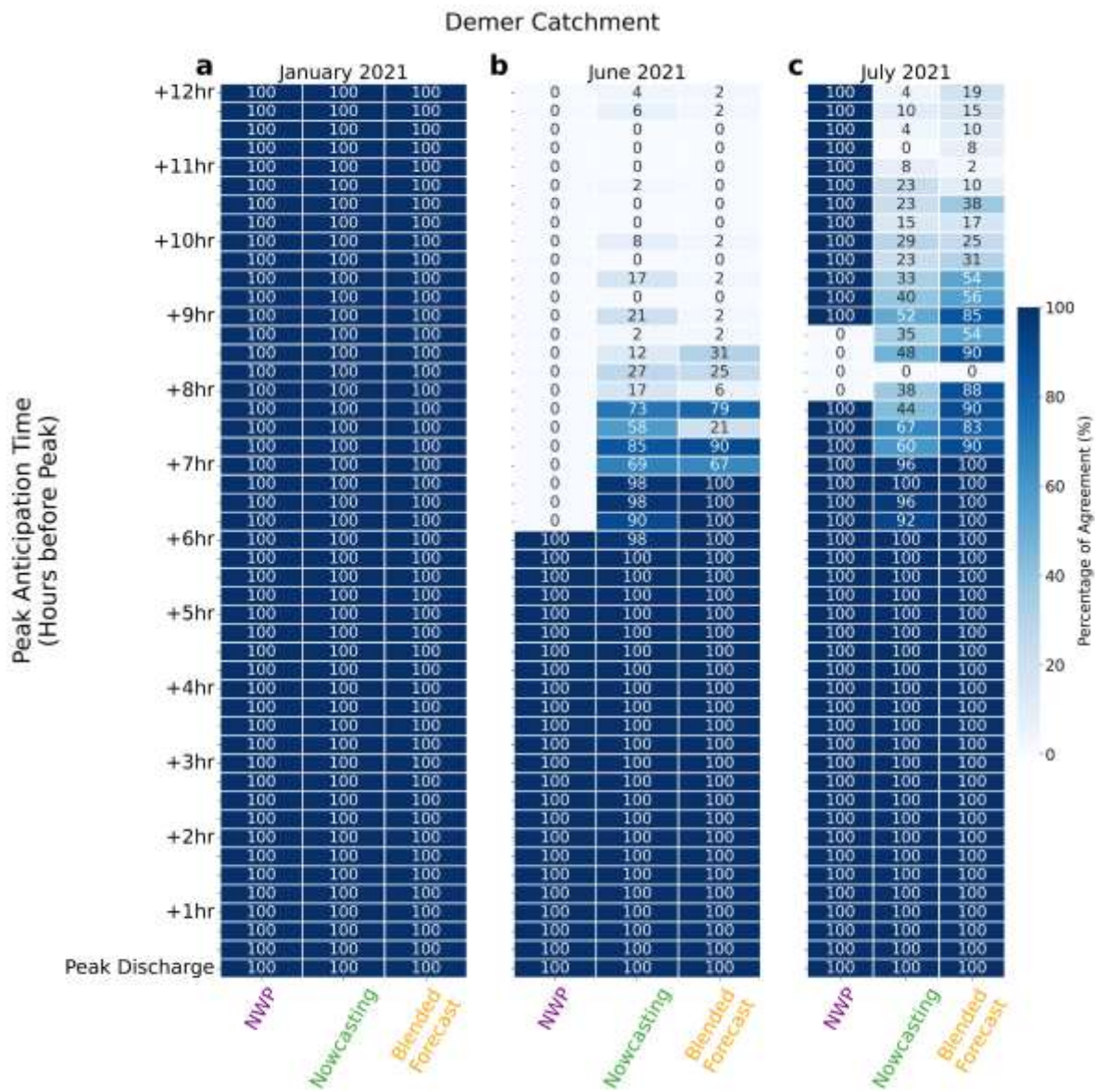
Figure 12: Heatmap displaying the peak anticipation time (PAT) in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. The number of ensemble members agreeing on the PAT per issue time ranges from 0 % in white to 100 % in dark blue. The panels show the results for the Rur catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Shifting focus to the June 2021 rainfall event results of the Demer catchment in Figure 13a, it is evident that all models are capable of predicting a consistent +12hr PAT. For the convective June 2021 result in Figure 13b, the NWP model has a +6hr PAT. The earliest warning is provided by the Nowcasting model at +9.5hr PAT followed by the Blended Forecast at +8.5hr PAT. Both show a clear improvement over the NWP model, with an average improvement of +3hr PAT. The July 2021 results in Figure 13c, demonstrate a +12hr PAT for the NWP model with a 1hr 0 % agreement gap between +7.45hr and +8.45hr PATs. The Nowcasting and Blended Forecast models provide an early warning around +12hr PAT but show stronger consensus on agreement around +10.45hr. Of interest is that all models show 0 % agreement at +8.15hr PAT due to an underestimation of the peak discharge. This

could imply an issue with the observations used to initialize the models, showing that consecutive agreement between subsequent model lead times should be considered.
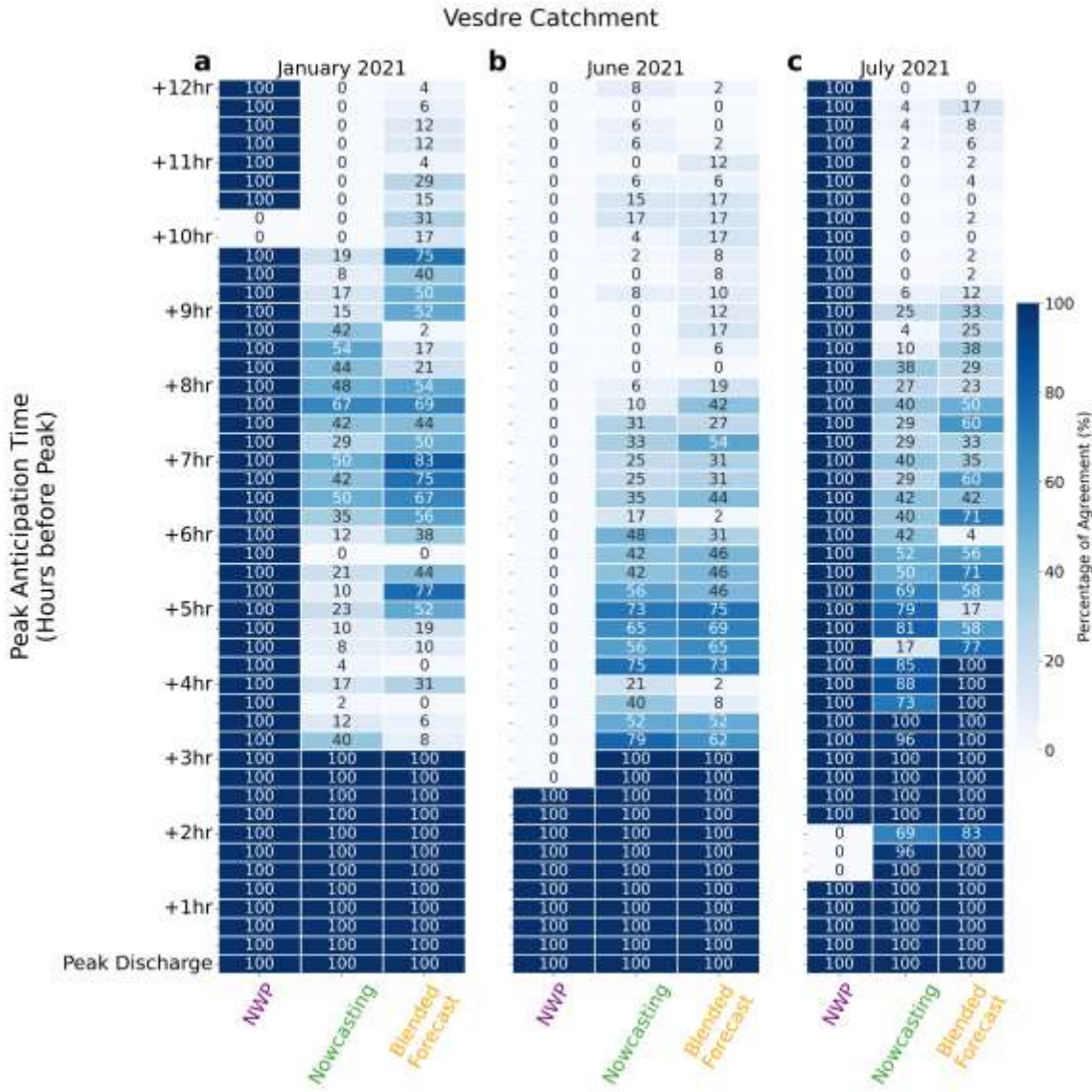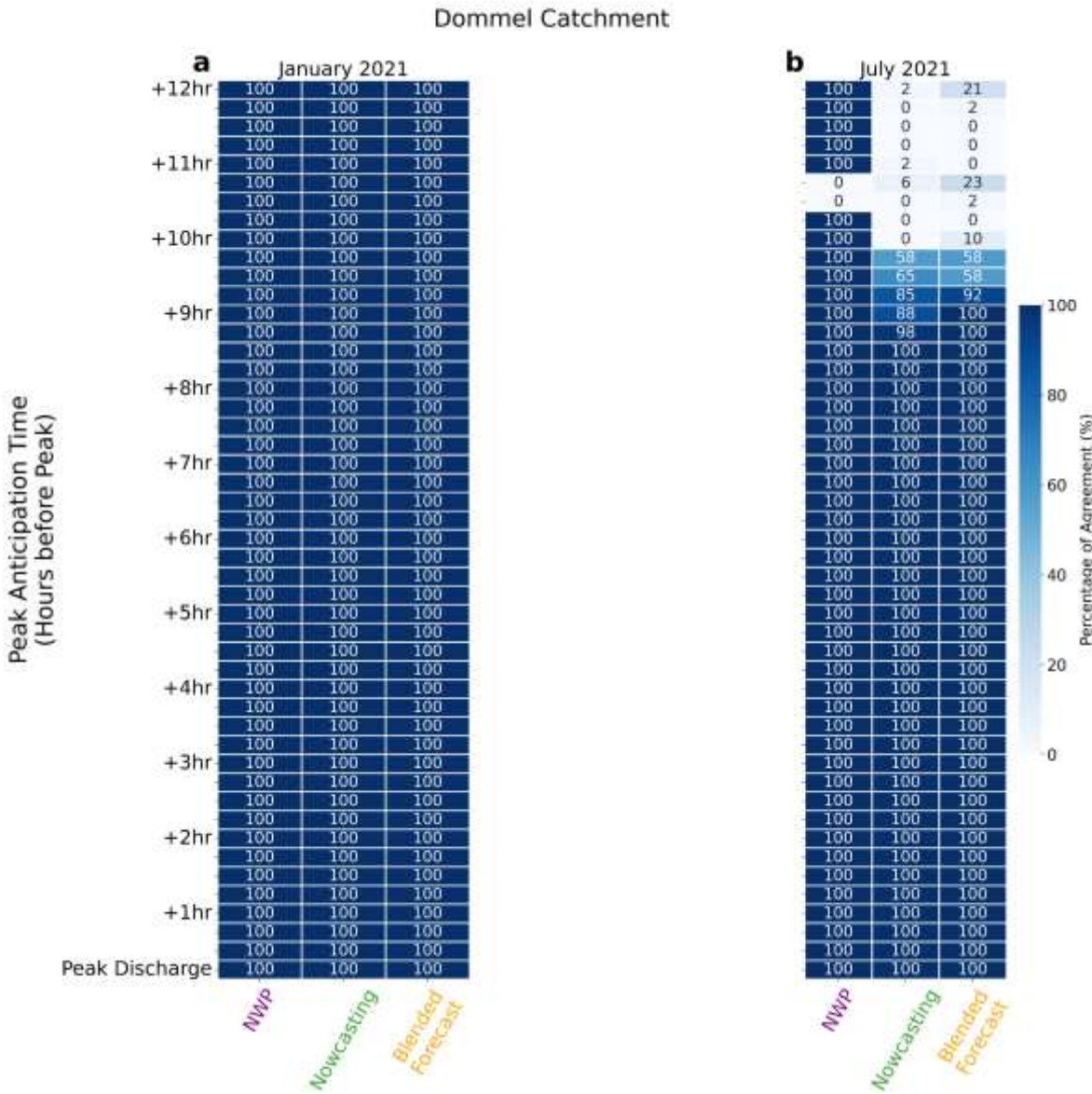

Figure 13: Heatmap displaying the peak anticipation time (PAT) in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. The number of ensemble members agreeing on the PAT per issue time ranges from 0 % in white to 100 % in dark blue. The panels show the results for the Demer catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

The PAT results of the Vesdre catchment deviate from previous findings. Here, the NWP model results in Figure 14a indicate a +12hr PAT. The Blended Forecast model shows an improvement over the Nowcasting model at +11.5hr compared to +9.45hr PAT. These results show superior skill by the NWP model in forecasting in a fast responding catchment under stratiform rainfall conditions.

The results shown in Figure 14b again underline the benefit of employing a probabilistic modelling approach under convective rainfall conditions. The Nowcasting model provides a +10.15 hr PAT with 15 % ensemble agreement and the Blended Forecast a +10.45hr PAT with 12 % ensemble agreement. The Blended Forecast model shows in the following lead times stronger consensus among ensemble members than the Nowcasting model. The NWP model provides a PAT of only +2.5hr before peak discharge.

The combined July 2021 rainfall event deviates the most from earlier findings. The NWP model has a PAT of +12hr, while the Blended Forecast model provides the first early warning at +11.45hr PAT. The first occurrence of strong consensus among the ensemble occurs at +9hr PAT. This is in contrast with the NWP model results for the Geul catchment (Figure 11c), where the NWP model did provide an early warning at +12hr PAT but lacked consistency for subsequent lead times.

Figure 14: Heatmap displaying the peak anticipation time (PAT) in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. The number of ensemble members agreeing on the PAT per issue

Concluding with the Dommel catchment's PAT results, shown in Figure 13a, the findings reveal a complete 100 % agreement among all models on the prediction of peak discharge at +12hr PAT. The July 2021 event results in Figure 13b show a PAT of +12hr, 2 hours before the Nowcasting model. The Blended Forecast also has a PAT of +12hr but is less consistent until +10hr PAT. Here, we find that the NWP model performs best followed by an improvement of the Blended Forecast model over the Nowcasting model.

Figure 15 Heatmap displaying the peak anticipation time (PAT) in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. The number of ensemble members agreeing on the PAT per issue time ranges from 0 % in white to 100 % in dark blue. The panels show the results for the Dommel catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Summary

Throughout the various catchments and rainfall events studied, the Peak Anticipation Time (PAT) results have highlighted several patterns. First the stratiform January 2021 rainfall event results show perfect agreement among all models on +12hr PAT for the Demer and Dommel catchments.

Based on at least 10 % agreement of the ensemble members on PAT for the Geul catchment, the Blended Forecast model best predicts the peak discharge, followed by the Nowcasting model, and the NWP model 3 hours later. This large improvement over using the NWP model might affect decision making. For the Rur catchment this improvement is +1.75hr, highlighting that the Nowcasting and Blended Forecasting models are capable of forecasting for larger catchments. The Vesdre catchments results show that the NWP model has a +1.25hr earlier warning than the Blended Forecast model. This demonstrates that the results are very case-specific, since the Geul and Vesdre are both smaller and fast responding catchments.

The convective June 2021 rainfall event results clearly demonstrate the added value of employing a probabilistic forecast (Nowcasting and Blended Forecast) as opposed to a deterministic forecast (NWP) due to uncertainty in forecasting convective events. Major improvements are found with approximately a +6hr earlier warning for the Nowcasting and Blended Forecast models than for the NWP model in the Geul catchment, +7hr earlier waring in the Rur catchment, +3hr in the Demer catchment, and +7hr in the Vesdre catchment.

The benefits of probabilistic forecasting for the convective June 2021 rainfall event are less present for the combined July 2021 rainfall event. The Geul catchments contained a large inconsistency in subsequent NWP model PATs. Therefore, the Nowcasting and Blended Forecast models performed considerably better in this catchment. The opposite was the case for the Rur catchment, where the Blended Forecast provided an early warning with subsequent inconsistencies in PAT agreement percentages. For the Demer, Vesdre and Dommel catchments the NWP model contained the earliest and most consistent PATs, indicating that the event was quite well captured by the NWP.

In summary, the NWP model, while robust in certain circumstances, displayed instances of limitations in close proximity to some peak discharge events. The Nowcasting and Blended Forecast models, on the other hand, frequently showcased an advantage in offering earlier warnings for convective rainfall events, thereby taking full advantage of the high update frequency (e.g. every 5 minutes) that these methods can give in contrast to most NWP models. The Blended Forecast model, combining attributes of both NWP and Nowcasting, often emerged as a consistent performer that balances early warnings with sustained consensus. However, in catchments with a slower response, such as the Dommel, the distinctiveness of the Nowcasting and Blended Forecast models became less pronounced for prolonged lead times. While each forecasting model possesses unique advantages and limitations, the Blended Forecast model consistently demonstrated efficacy by combining early PATs with strong consensus among ensemble members, emphasizing its potential utility in real-world applications.

For more contingency based statistics, the reader is referred to Appendix A.

4 Discussion

This section first discusses the main findings in relation to employing the Blended Forecast rainfall product to drive the wflow_sbm hydrological model for Flash Flood Early Warning Systems (FFEWS). Next, we reflect on the shortcomings of the current model setup and analyses. This is followed by a discussion on how we envision an operational system based on this work package. Next, we reflect on the strengths and challenges of decision making under uncertainty based on ensemble forecasts. Lastly, an outlook is provided of the most recent developments in the field of rainfall forecasting.

4.1 Advantages and Challenges of the Blended Forecast Model

In the context of the selected catchments and rainfall events, the Blended Forecast model has repeatedly demonstrated skill in predicting flash floods. For instance, during the stratiform January 2021 rainfall event, the model accurately predicted the peak discharge for the Geul catchment 11 hours in advance, 3 hours before the NWP model's prediction. Together with the Nowcasting model, the probabilistic ensemble models, based on at least 10 % ensemble agreement on PAT, consistently outperform the deterministic NWP model in providing early peak discharge warnings during the convective June 2021 event. For example, the Blended Forecast and Nowcasting models provided roughly 6-7 hours earlier warnings for the Geul and Rur catchments. By integrating the strengths of both the NWP and Nowcasting models, the Blended Forecast model issues timely warnings while ensuring a consistent agreement among ensemble members for diverse events. In addition, the the information on the chance of an event occurring by probabilistic ensembles is valuable on its own right compared to deterministic forecasts.

The challenges of the Blended Forecast model are present at longer lead times. A challenge is to resolve the volume error related to the ensemble median, as highlighted in the MAE analysis in Figure 9. In addition, the model should produce more similar results to the NWP model for longer lead times of combined rainfall events. For instance, this might improve the skill of providing early warnings during the July 2021 event for the Vesdre catchment. The Blended Forecast model might therefore be improved by giving more weight to the NWP model at longer lead times when applying the blending method.

4.2 Shortcomings of the Beta Version of the FFEWS

Several shortcomings emerge from the beta FFEWS setup, offering potential for improvement. Firstly, the reliance on a single hydrological model is a noticeable constraint. The setup overlooks structural model uncertainties, which may bias the outcomes (e.g. Butts et al., 2004). Incorporating multiple models not only addresses these uncertainties but also provides a better representation of potential variability in discharge forecasts. Secondly, the results may be further biased by depending on reference simulations of a single hydrological model for evaluation. The ideal would be to use actual observations for assessments that are independent of the model simulations. This, however, complicates the distinction between the impacts of various rainfall forecast products on discharge forecasts. Thirdly, the NWP considered in this study is deterministic. It should be noted that recent advances in NWP development shift towards probabilistic ensemble forecasts. Lastly, while the analyses mainly focusses on peak discharge, this might not give a holistic understanding of the system's performance. Therefore, Appendix A contains extended analyses based on contingency statistics.

4.3 Interpretation of Ensemble Forecasts

Ensemble forecasts are an improvement in meteorological and hydrological forecasting methodologies. The strength lies in their ability to provide a range of possible future scenarios instead of a singular deterministic prediction. By incorporating multiple model initializations, the ensemble includes model uncertainties, therefore offering a probabilistic perspective. This facilitates better risk assessment as users can gauge the likelihood of different outcomes, allowing for more informed and robust decision-making.

Despite the strength of ensemble forecasts, the interpretation of likelihoods can be complex and subjective. Many studies have investigated the concept of decision-making under uncertainty (e.g., (Kwakkel et al., 2016; Ramos et al., 2013; Zappa et al., 2013). For one, the sheer volume of data presented can be overwhelming, especially considering that the 48 model realizations in this study can be considered small. Another complexity is understanding the spread of the ensemble. While a wider spread indicates higher uncertainty or a smaller spread lower uncertainty, it does not necessarily pinpoint which particular outcome within that spread is the most probable. One of the main challenges is communicating the nuances of ensemble forecasts to non-experts, such as decision-makers or the general public, as misinterpretations can lead to inappropriate actions or responses.

A possible solution is to employ a structured approach to streamline the decision making process. This can be achieved by not communicating the whole ensemble forecast, but to use techniques like clustering to group together similar outcomes, essentially simplifying the decision-making process. If agreed upon beforehand, this clustering could even further reduce the ensemble to color codes that indicate the uncertainty and the severity of the forecasts. In addition, the results of this study show the strength of communicating the additional anticipation time that decision-makers are expected to have when implementing flood forecasting improvement. Most importantly, regular training sessions for end-users are required to ensure that they remain updated on best practices in ensemble forecast interpretation.

4.4 The Operational System: A Glimpse into the Near Future

To move forward from the "beta" stage of the WP towards an operational system, we envision an ideal system that is characterized by several advanced features. In essence, this envisioned operational system would be a culmination of modern technology, advanced modelling techniques, and user-centric design, aiming to address the multifaceted challenges of water management. This system would consist of the following components:

1. **Time step**
   The system will operate on a highly refined time step, ensuring precision in capturing the intricate rainfall dynamics of stratiform and convective systems. In addition, the time step needs to be adequate for modelling the hydrological and hydrodynamic response. Leading is the availability of high temporal resolution rainfall inputs. Based on the results of this WP, we recommend employing the highest possible time step, 5 minutes.
2. **Spatial resolution**
   Similar to the time step, the spatial resolution of the modelling grid should encompass the spatial variability of rainfall and hydrological processes. The spatial variability of rainfall in the Netherlands has a decorrelation-distance of 10 km in Summer (Leth et al., 2021), a 1 km spatial resolution should be sufficient. The hydrological processes descriptions of the

wflow_sbm are better resolved in steeper catchment at finer spatial scales due to better representation of the topography (Aerts et al., 2022). Therefore, we recommend using a high spatial resolution, e.g. 100 meters. However, the speed of the modelling chain and computational demands are leading. A trade-off between the number of model realizations and computational demand should be carefully considered.

3. **Modelling chain**
   The system would employ multiple conventional hydrological models, but also hydrodynamic models to model surface runoff and river flow. In addition, surrogate machine learning models (e.g. Sun et al., 2023), statistical approximation of conventional models, can be employed to reduce model runtime and computational demand. This has the potential to greatly boost the number of model realizations in the ensemble.

4. **Remote sensing integration**
   Through data assimilation multiple remote sensing products, such as soil moisture, can be integrated in the model ensemble (e.g. Houser et al., 1998; Lievens et al., 2017). This corrects the model states in addition to using the most recent rainfall observations during model initialization to reduce uncertainty.

5. **Cross-boundary system**
   The system needs to surpass political boundaries and encompass the whole cross-boundary Meuse basin. For instance, this increases anticipation time for the Rur catchment and fosters shared decision-making among different jurisdictions, promoting a unified approach to water management.

6. **Platform**
   The operational platform would be digital, intuitive, and user-friendly. It would be designed for seamless integration with other systems and would have the capability for future scalability, ensuring it remains relevant with evolving technological advancements. A widely used example of such a system is Delft-FEWS (Werner et al., 2013), for which an FFEWS system is integrated based on the findings of this work package.

7. **Users**
   Targeted are a wide array of users, from policy-makers and decision-makers to hydrologists and environmental engineers. The platform would cater to varied needs depending on the level of expertise and offer customization options based on user preferences and requirements.

4.5 Ongoing Scientific Developments in Quantitative Precipitation Forecasting

Here, we provide a brief overview of the ongoing scientific developments in the field of quantitative precipitation forecasting (QPF). In the past few years significant advancements have been made, driven by both technology and scientific research. Several groundbreaking research papers on QPF have been published in renowned journals. Recently a (physics-conditional) deep generative model for nowcasting has been proposed by Ravuri et al., 2021 and Zhang et al. (2023). This model uses physical principles to contain the statistical nowcasting model, resulting in more plausible forecasts. Other studies hold promise in incorporating or replacing numerical models using machine learning. Examples are 3D neural networks replacing NWP models (Bi et al., 2023) and machine learning models replacing hydrological models under physical constraints (Sun et al., 2023).

In addition, there are technical advances made in earth observations. The Meteosat satellite, operated by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), plays a pivotal role in improving QPF. This satellite has enhanced spatial and temporal resolution,

allowing for better tracking of cloud formations and movement. The infrared and water vapor sensors are improved that aid in understanding rainfall potential. The produced data is now more integrated with ground-based radar systems and weather stations, enhancing the overall prediction capability. Other technical advances are based on opportunistic sensing techniques such as the use of commercial microwave links necessary for telecommunication to quantify precipitation rates (*e.g.,* (de Vos et al., 2018; Heuvelink et al., 2020; Imhoff et al., 2020; Overeem et al., 2021; Uijlenhoet et al., 2018). These methods make it possible to complement current observation systems used for e.g. nowcasting or to enable these techniques in data-scarce regions.

5 Reflection on Project Deliverables

The core objective of this work package was to evaluate the efficacy of various forecast models (specifically Numerical Weather Prediction (NWP), Nowcasting, and Blended Forecasting in predicting peak discharge events, using the Peak Anticipation Time (PAT) as a pivotal metric. The ambition is to enhance early warning capacities and streamline responses to potential flash flood events.

Activities Undertaken

To realize this goal, a series of activities were executed:

- Detailed analysis and comparison of the models, especially with reference to their PAT estimates, were conducted across multiple catchments and rainfall events.
- Extensive ensemble simulations were performed, incorporating 48 realizations, to better understand and compare the lead time and consensus on predicting peak discharges.
- Figures are created to provide a visual summary of the outcomes for each model, rainfall event, and catchment.

Results in Context

The findings provide a comprehensive insight into the relative strengths and limitations of each model:

- Generally, the Blended Forecast model frequently demonstrated advantages over the Nowcasting and NWP models, particularly for stratiform rainfall events.
- The Nowcasting and Blended Forecast models often delivered early warnings, notably in scenarios where the NWP model struggled, especially during convective rainfall events.
- In certain catchments, like the Demer and Dommel, the results highlighted a near-unanimous consensus on peak discharge predictions, emphasizing the accuracy and reliability of the models used. In addition, this emphasizes the usefulness of Nowcasting techniques in smaller catchments.

Next Steps

To operationalize the beta system:

- Continued calibration and fine-tuning of the models should be undertaken, based on the feedback loops from the results.
- There should be an increased emphasis on integrating real-time data streams, to further improve prediction accuracy.
- Training sessions and workshops for water managers and other relevant stakeholders should be organized, ensuring that they are well-equipped to interpret and act upon the forecasts.

General Reflection

This project stands as a testament to the power of collaborative research and innovation. Choosing Interreg as the platform for this endeavour proved invaluable. This framework fostered cross-border collaboration, facilitating the sharing of knowledge, expertise, and best practices. The collective effort not only underscored the importance of flash flood early warning systems but also provided a new course for the future, pointing towards methodologies that prioritize both accuracy and timeliness.

References

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., & Hazenberg, P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, *26*(16), 4407–4430. https://doi.org/10.5194/hess-26-4407-2022

Atencia, A., Mediero, L., Llasat, M. C., & Garrote, L. (2011). Effect of radar rainfall time resolution on the predictive capability of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *15*(12), 3809–3827. https://doi.org/10.5194/hess-15-3809-2011

Bailey, M. E., Isaac, G. A., Gultepe, I., Heckman, I., & 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*(1), 257–276. https://doi.org/10.1007/s00024-012-0553-x

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), Article 7567. https://doi.org/10.1038/nature14956

Berenguer, M., Surcel, M., Zawadzki, I., Xue, M., & Kong, F. (2012). The Diurnal Cycle of Precipitation from Continental Radar Mosaics and Numerical Weather Prediction Models. Part II: Intercomparison among Numerical Models and with Nowcasting. *Monthly Weather Review*, *140*(8), 2689–2705. https://doi.org/10.1175/MWR-D-11-00181.1

Berne, A., Delrieu, G., Creutin, J.-D., & Obled, C. (2004). Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, *299*(3), 166–179. https://doi.org/10.1016/j.jhydrol.2004.08.002

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), Article 7970. https://doi.org/10.1038/s41586-023-06185-3

Bogena, H., Kunkel, R., Schöbel, T., Schrey, H. P., & Wendland, F. (2005). Distributed modeling of groundwater recharge at the macroscale. *Ecological Modelling*, *187*(1), 15–26. https://doi.org/10.1016/j.ecolmodel.2005.01.023

Bogena, H. r., Montzka, C., Huisman, J. a., Graf, A., Schmidt, M., Stockinger, M., von Hebel, C., Hendricks-Franssen, H. j., van der Kruk, J., Tappe, W., Lücke, A., Baatz, R., Bol, R., Groh, J., Pütz, T., Jakobi, J., Kunkel, R., Sorg, J., & Vereecken, H. (2018). The TERENO-Rur Hydrological Observatory: A Multiscale Multi-Compartment Research Platform for the Advancement of Hydrological Science. *Vadose Zone Journal*, *17*(1), 180055. https://doi.org/10.2136/vzj2018.03.0055

Bowler, N. E., Pierce, C. E., & Seed, A. W. (2006). STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society*, *132*(620), 2127–2155. https://doi.org/10.1256/qj.04.100

Bubnová, R., Hello, G., Bénard, P., & 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*(2), 515–535. https://doi.org/10.1175/1520-0493(1995)123<0515:IOTFEE>2.0.CO;2

Butts, M. B., Payne, J. T., Kristensen, M., & Madsen, H. (2004). An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology*, *298*(1), 242–266. https://doi.org/10.1016/j.jhydrol.2004.03.042

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., & Peters-Lidard, C. D. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

de Vos, L. W., Raupach, T. H., Leijnse, H., Overeem, A., Berne, A., & Uijlenhoet, R. (2018). High-Resolution Simulation Study Exploring the Potential of Radars, Crowdsourced Personal Weather Stations, and Commercial Microwave Links to Monitor Small-Scale Urban Rainfall. *Water Resources Research*, *54*(12), 10,293-10,312. https://doi.org/10.1029/2018WR023393

Demeritt, D., Nobert, S., Cloke, H., & Pappenberger, F. (2010). Challenges in communicating and using ensembles in operational flood forecasting. *Meteorological Applications*, *17*(2), 209–222. https://doi.org/10.1002/met.194

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., … Viglione, A. (2009). A compilation of data on European flash floods. *Journal of Hydrology*, *367*(1), 70–78. https://doi.org/10.1016/j.jhydrol.2008.12.028

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

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

Houser, P. R., Shuttleworth, W. J., Famiglietti, J. S., Gupta, H. V., Syed, K. H., & Goodrich, D. C. (1998). Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resources Research*, *34*(12), 3405–3420. https://doi.org/10.1029/1998WR900001

Imhoff, R. O. (2022). *Rainfall nowcasting for flood early warning*. https://doi.org/10.18174/573867

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

Imhoff, R. O., De Cruz, L., Dewettinck, W., Brauer, C. C., Uijlenhoet, R., van Heeringen, K.-J., Velasco-Forero, C., Nerini, D., Van Ginderachter, M., & Weerts, A. H. (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*, *149*(753), 1335–1364. https://doi.org/10.1002/qj.4461

Imhoff, R. O., Overeem, A., Brauer, C. C., Leijnse, H., Weerts, A. H., & Uijlenhoet, R. (2020). Rainfall Nowcasting Using Commercial Microwave Links. *Geophysical Research Letters*, *47*(19), e2020GL089365. https://doi.org/10.1029/2020GL089365

Jongman, B. (2018). Effective adaptation to rising flood risk. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-04396-1

Journée, M., Goudenhoofdt, E., Vannitsem, S., & Delobbe, L. (2023). Quantitative rainfall analysis of the 2021 mid-July flood event in Belgium. *Hydrology and Earth System Sciences*, *27*(17), 3169–3189. https://doi.org/10.5194/hess-27-3169-2023

Kober, K., Craig, G. C., & Keil, C. (2014). Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, *140*(681), 1179–1188. https://doi.org/10.1002/qj.2220

Kober, K., Craig, G. C., Keil, C., & 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*(664), 755–768. https://doi.org/10.1002/qj.939

Koks, E. E., van Ginkel, K. C. H., van Marle, M. J. E., & Lemnitzer, A. (2022). Brief communication: Critical infrastructure impacts of the 2021 mid-July western European flood event. *Natural Hazards and Earth System Sciences*, *22*(12), 3831–3838. https://doi.org/10.5194/nhess-22-3831-2022

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, rich M., Fowler, H. J., Goergen, K., Heinrich, D., … Wanders, N. (2021). Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021. *World Weather Atribution*. http://hdl.handle.net/1854/LU-8732135

Kwakkel, J. H., Haasnoot, M., & Walker, W. E. (2016). Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep

uncertainty. *Environmental Modelling & Software*, *86*, 168–183.

https://doi.org/10.1016/j.envsoft.2016.09.017

Leth, T. C. van, Leijnse, H., Overeem, A., & Uijlenhoet, R. (2021). Rainfall Spatiotemporal Correlation

and Intermittency Structure from Micro-γ to Meso-β Scale in the Netherlands. *Journal of*

Hydrometeorology, *22*(8), 2227–2240. https://doi.org/10.1175/JHM-D-20-0311.1

Lievens, H., Reichle, R. H., Liu, Q., De Lannoy, G. J. M., Dunbar, R. S., Kim, S. B., Das, N. N., Cosh, M.,

Walker, J. P., & Wagner, W. (2017). Joint Sentinel-1 and SMAP data assimilation to improve

soil moisture estimates. *Geophysical Research Letters*, *44*(12), 6145–6153.

https://doi.org/10.1002/2017GL073904

Lin, C., Vasić, S., Kilambi, A., Turner, B., & Zawadzki, I. (2005). Precipitation forecast skill of numerical

weather prediction models and radar nowcasts. *Geophysical Research Letters*, *32*(14).

https://doi.org/10.1029/2005GL023451

Nerini, D., Foresti, L., Leuenberger, D., Robert, S., & 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*(3), 987–1006.

https://doi.org/10.1175/MWR-D-18-0258.1

Overeem, A., Holleman, I., & Buishand, A. (2009). Derivation of a 10-Year Radar-Based Climatology of

Rainfall. *Journal of Applied Meteorology and Climatology*, *48*(7), 1448–1463.

https://doi.org/10.1175/2009JAMC1954.1

Overeem, A., Leijnse, H., Leth, T. C. van, Bogerd, L., Priebe, J., Tricarico, D., Droste, A., & Uijlenhoet,

R. (2021). Tropical rainfall monitoring with commercial microwave links in Sri Lanka.

Environmental Research Letters, *16*(7), 074058. https://doi.org/10.1088/1748-9326/ac0fa6

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., & Thielen, J. (2015). The

monetary benefit of early flood warnings in Europe. *Environmental Science & Policy*, *51*, 278–

291. https://doi.org/10.1016/j.envsci.2015.04.016

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L.

(2019). Pysteps: An open-source Python library for probabilistic precipitation nowcasting

(v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Pyka, C., Jacobs, C., Breuer, R., Elbers, J., Nacken, H., Sewilam, H., & Timmerman, J. (2016). Effects of water diversion and climate change on the Rur and Meuse in low-flow situations. *Environmental Earth Sciences*, *75*(16), 1206. https://doi.org/10.1007/s12665-016-5989-3

Radhakrishnan, C., & 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*(2), 211–228. https://doi.org/10.1175/JTECH-D-18-0192.1

Ramos, M. H., van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219–2232. https://doi.org/10.5194/hess-17-2219-2013

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., & Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), Article 7878. https://doi.org/10.1038/s41586-021-03854-z

Roberts, N. M., & Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Seed, A. W., Pierce, C. E., & Norman, K. (2013). Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resources Research*, *49*(10), 6624–6641. https://doi.org/10.1002/wrcr.20536

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

Sun, R., Pan, B., & Duan, Q. (2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, *624*, 129944. https://doi.org/10.1016/j.jhydrol.2023.129944

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., … Joly, A. (2018). The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geoscientific Model Development*, *11*(1), 257–281. https://doi.org/10.5194/gmd-11-257-2018

Tradowsky, J. S., Philip, S. Y., Kreienkamp, F., Kew, S. F., Lorenz, P., Arrighi, J., Bettmann, T., Caluwaerts, S., Chan, S. C., De Cruz, L., de Vries, H., Demuth, N., Ferrone, A., Fischer, E. M., Fowler, H. J., Goergen, K., Heinrich, D., Henrichs, Y., Kaspar, F., … Wanders, N. (2023). Attribution of the heavy rainfall events leading to severe flooding in Western Europe during July 2021. *Climatic Change*, *176*(7), 90. https://doi.org/10.1007/s10584-023-03502-7

Uijlenhoet, R., Overeem, A., & Leijnse, H. (2018). Opportunistic remote sensing of rainfall using microwave links from cellular communication networks. *WIREs Water*, *5*(4), e1289. https://doi.org/10.1002/wat2.1289

van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Bouaziz, L., Eilander, D., Hegnauer, M., ten Velden, C., & Russell, B. (2022). *Wflow_sbm v0.6.1, a spatially distributed hydrologic model: From global data to local applications* [Preprint]. Hydrology. https://doi.org/10.5194/gmd-2022-182

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

Yoon, S.-S. (2019). Adaptive Blending Method of Radar-Based and Numerical Weather Prediction QPFs for Urban Flood Forecasting. *Remote Sensing*, *11*(6), Article 6. https://doi.org/10.3390/rs11060642

Zappa, M., Fundel, F., & Jaun, S. (2013). A 'Peak-Box' approach for supporting interpretation and

verification of operational ensemble peak-flow forecasts. *Hydrological Processes*, *27*(1), 117–

131. https://doi.org/10.1002/hyp.9521

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023). Skilful nowcasting of

extreme precipitation with NowcastNet. *Nature*, *619*(7970), Article 7970.

https://doi.org/10.1038/s41586-023-06184-4

Appendix A

A.1 Contingency Statistics

The contingency based statistics are based on penalizing and rewarding the forecast models based on correct early warnings. The statistics use the ensemble median of the probabilistic Nowcasting and Blended Forecast Models.

For each time step it is determined if the reference simulation is or is not within the 90 % to 110 % peak discharge threshold, considering the following categories:
- **Hits** are registered when both the reference simulation and forecast simulation are within the peak discharge threshold.
- **Misses** are registered when the reference simulation is within the peak discharge threshold but the forecast simulation is not.
- **Correct Negatives** are registered when both the reference simulation and the forecast simulation are not within the peak discharge threshold.
- **False Alarms** are registered when the reference simulation is not within the peak discharge threshold but the forecast simulation is.

An overview of the resulting contingency table is provided in Figure A.1.

Figure A.1: Contingency table, showing hits occurring with both a "yes" for the reference and forecast simulations based on the peak discharge threshold (90 % - 110 %). Misses: "yes" for the reference and "no" for the forecast. False alarms: "no" for the reference and "yes" for the forecast. Correct negatives: "no" for the reference and "no" for the forecast.

A.2 Hit Rate

The hit rate or probability of detection ranges from 0 to 1 (perfect score) and answers the question: *What fraction of the observed "yes" events were correctly forecast?*

Mathematically, the hit rate is defined as:

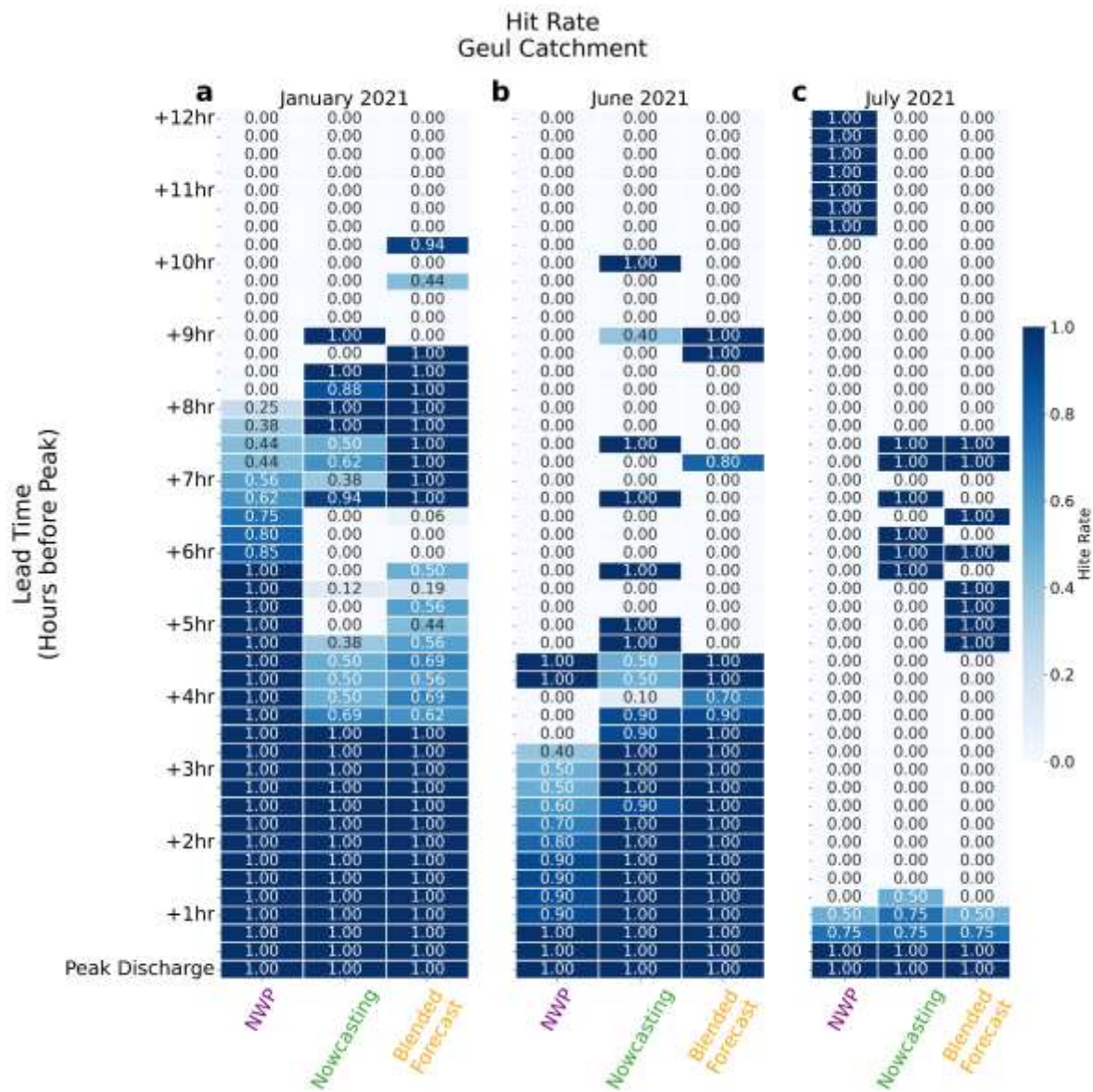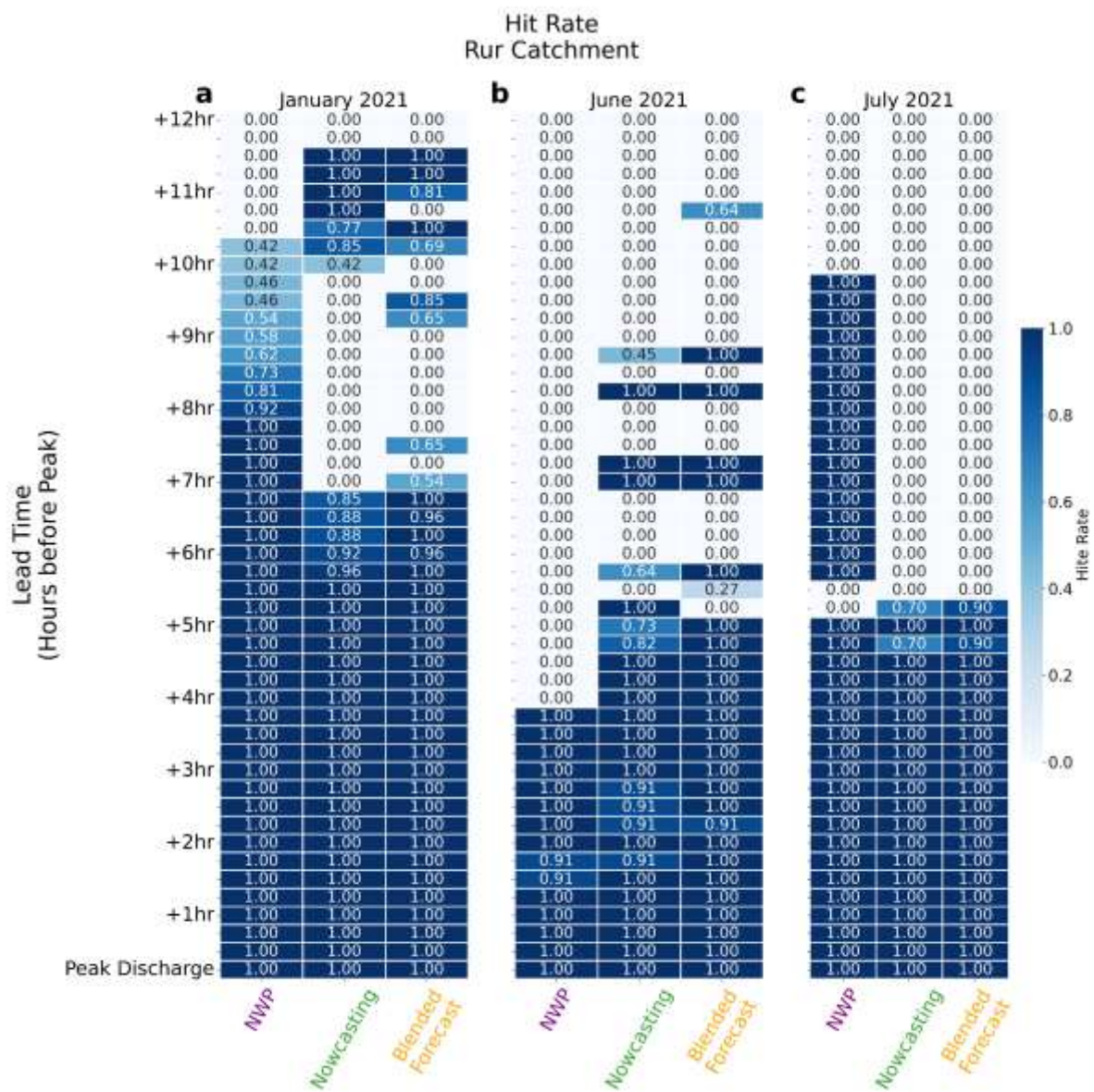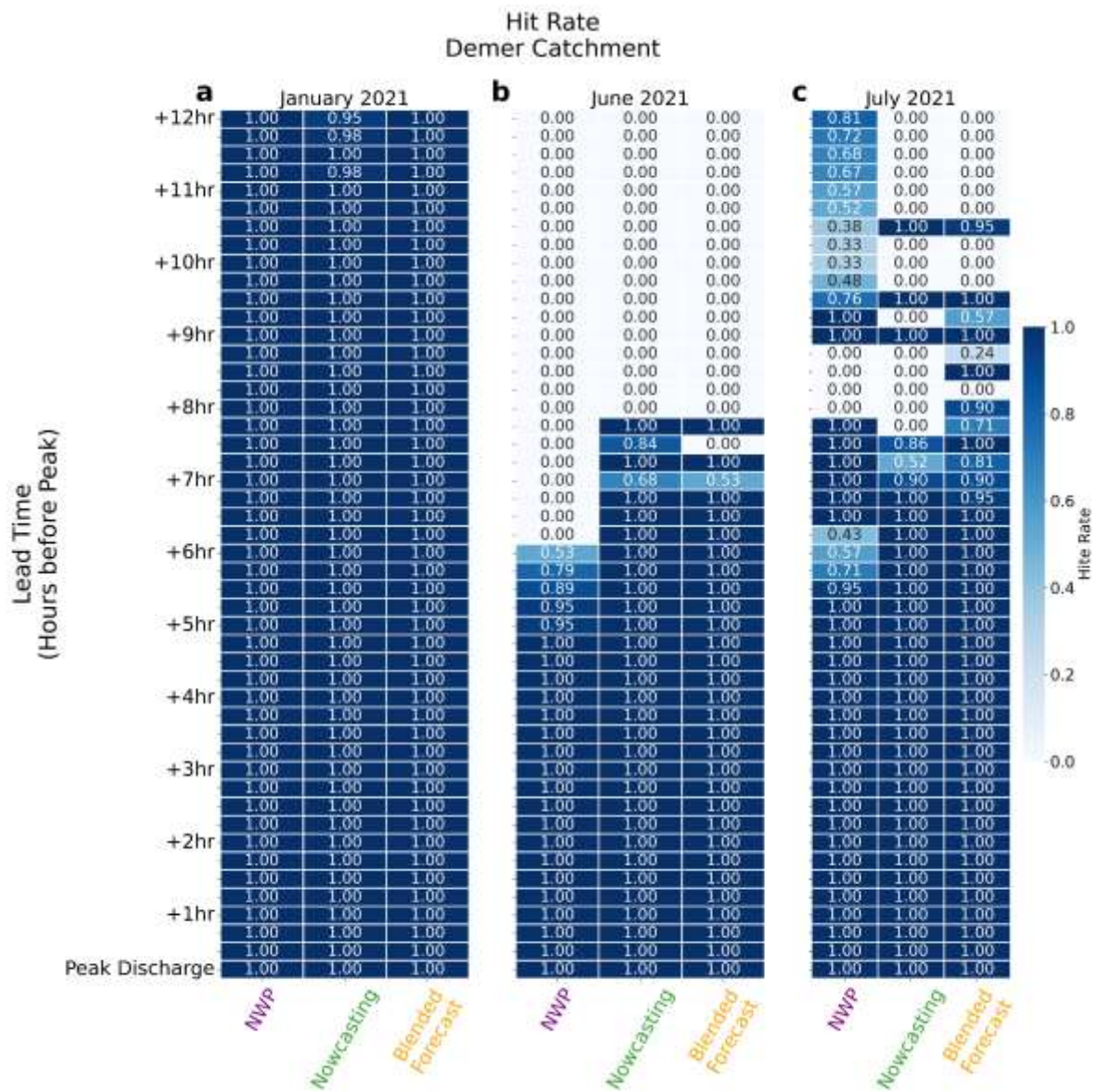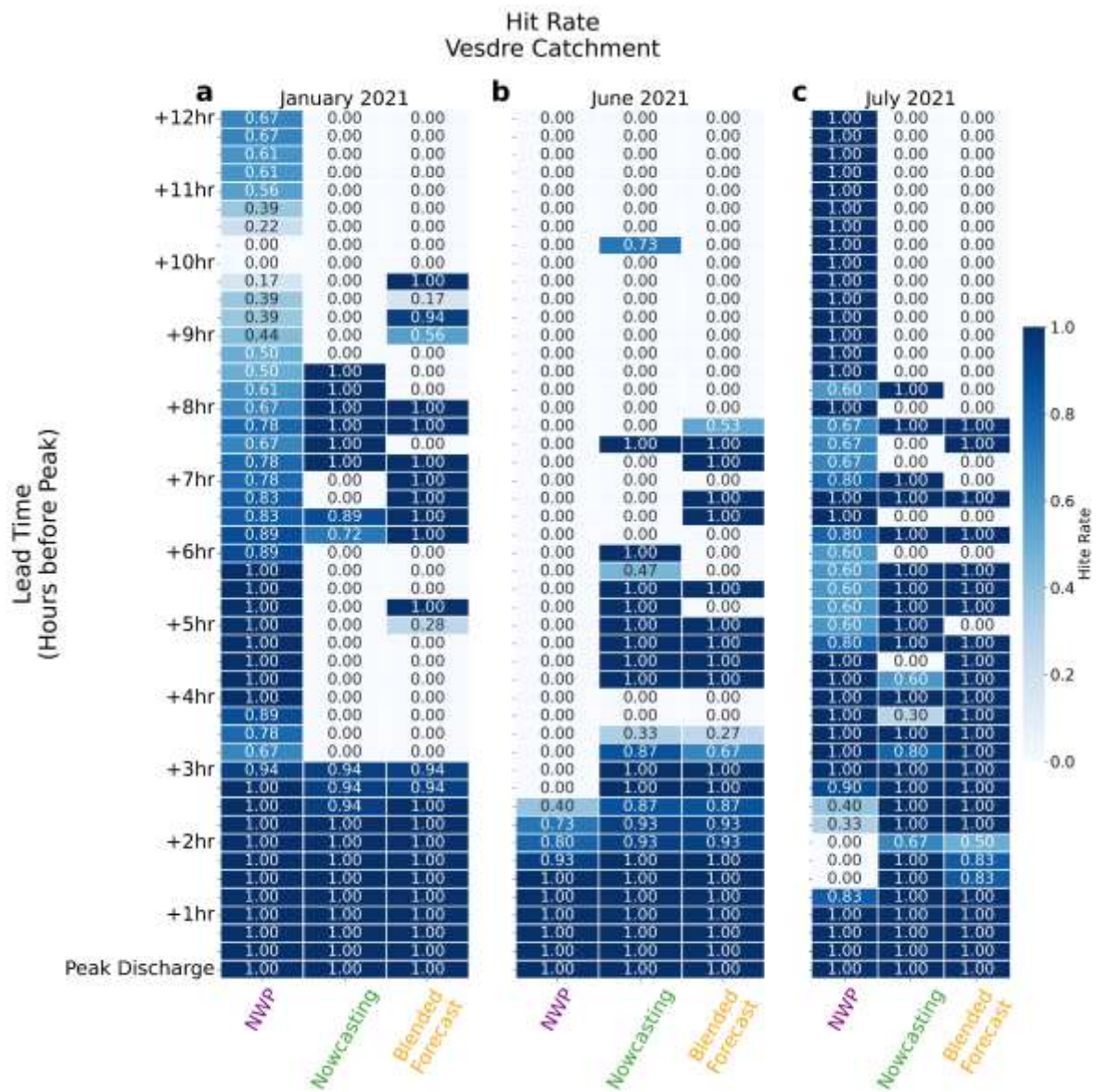$$Hit\ Rate = \frac{hits}{hits+misses}$$
(Eq. 3)

Figure A.2: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Hit rate ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Geul catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Figure A.3: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Hit rate ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Rur catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
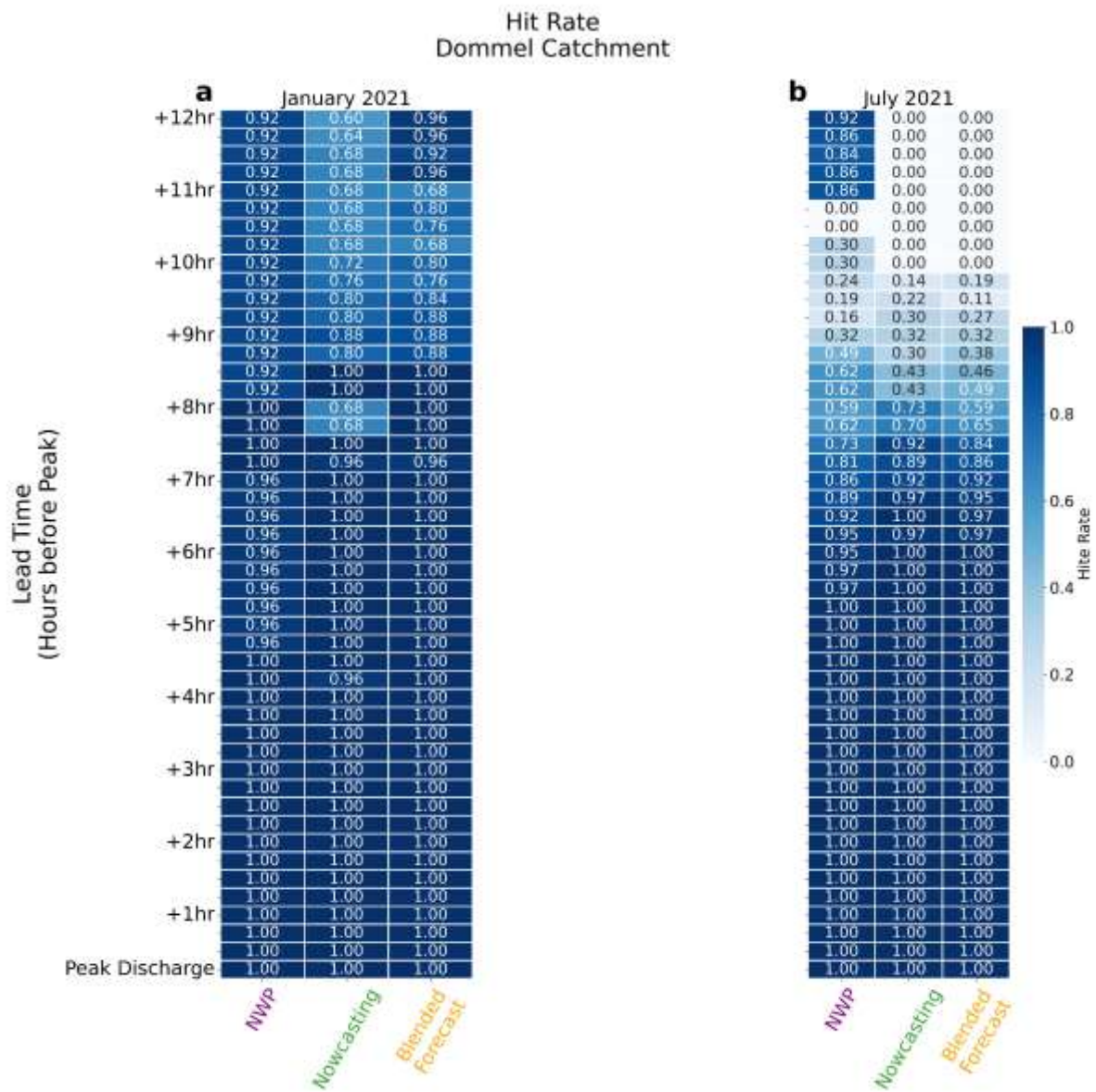
Figure A.4: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Hit rate ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Demer catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Figure A.5: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Hit rate ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Vesdre catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
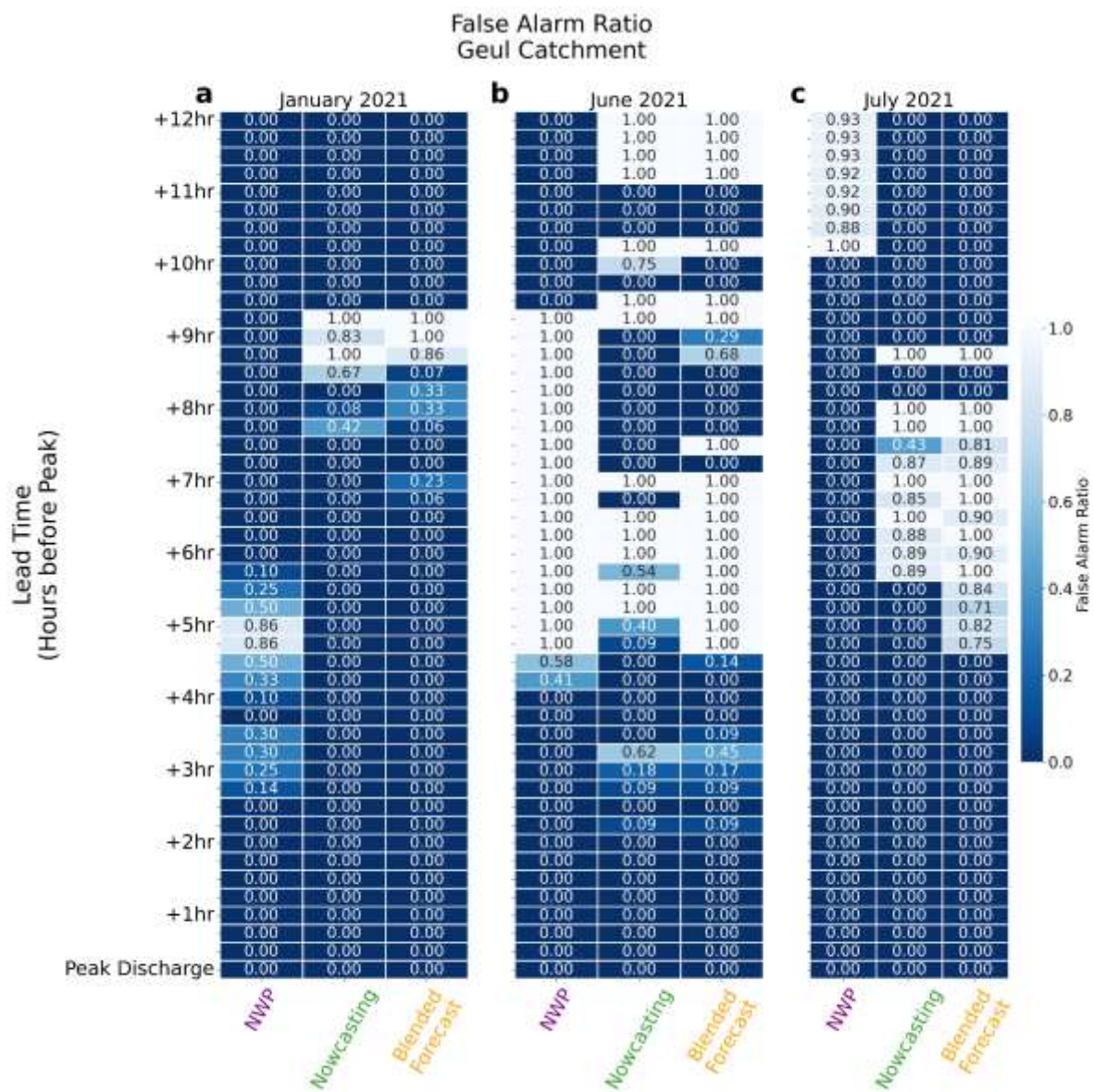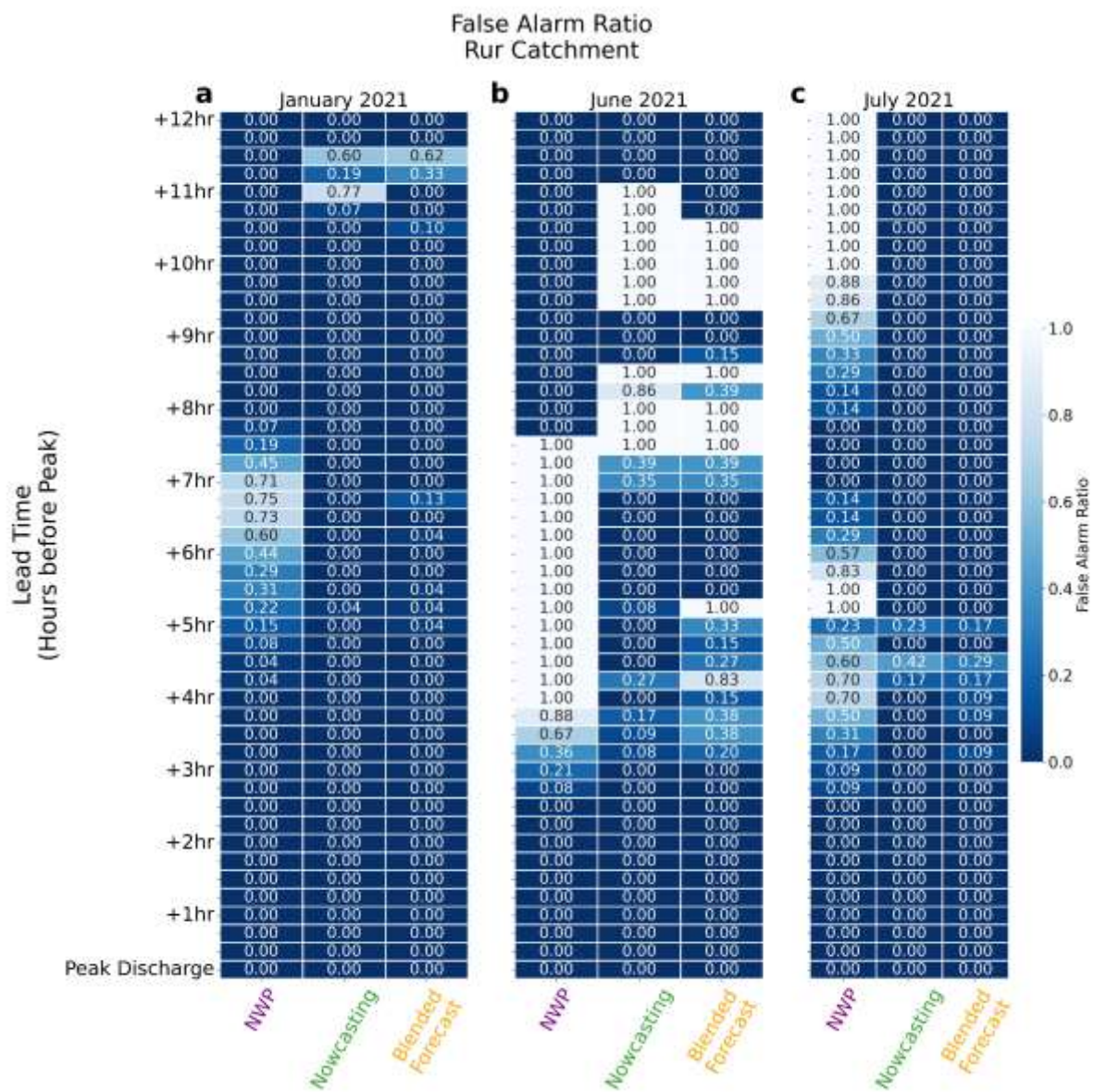
Figure A.6: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Hit rate ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Dommel catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

A.3 False Alarm Ratio
The false alarm ratio ranges from 0 (perfect score) to 1 and answers the question: *What fraction of the predicted "yes" events actually did not occur?*

Mathematically, the false alarm ratio is defined as:

$$False\ Alarm\ Ratio = \frac{false\ alarms}{hits + false\ alarms} \qquad \text{(Eq. 4)}$$
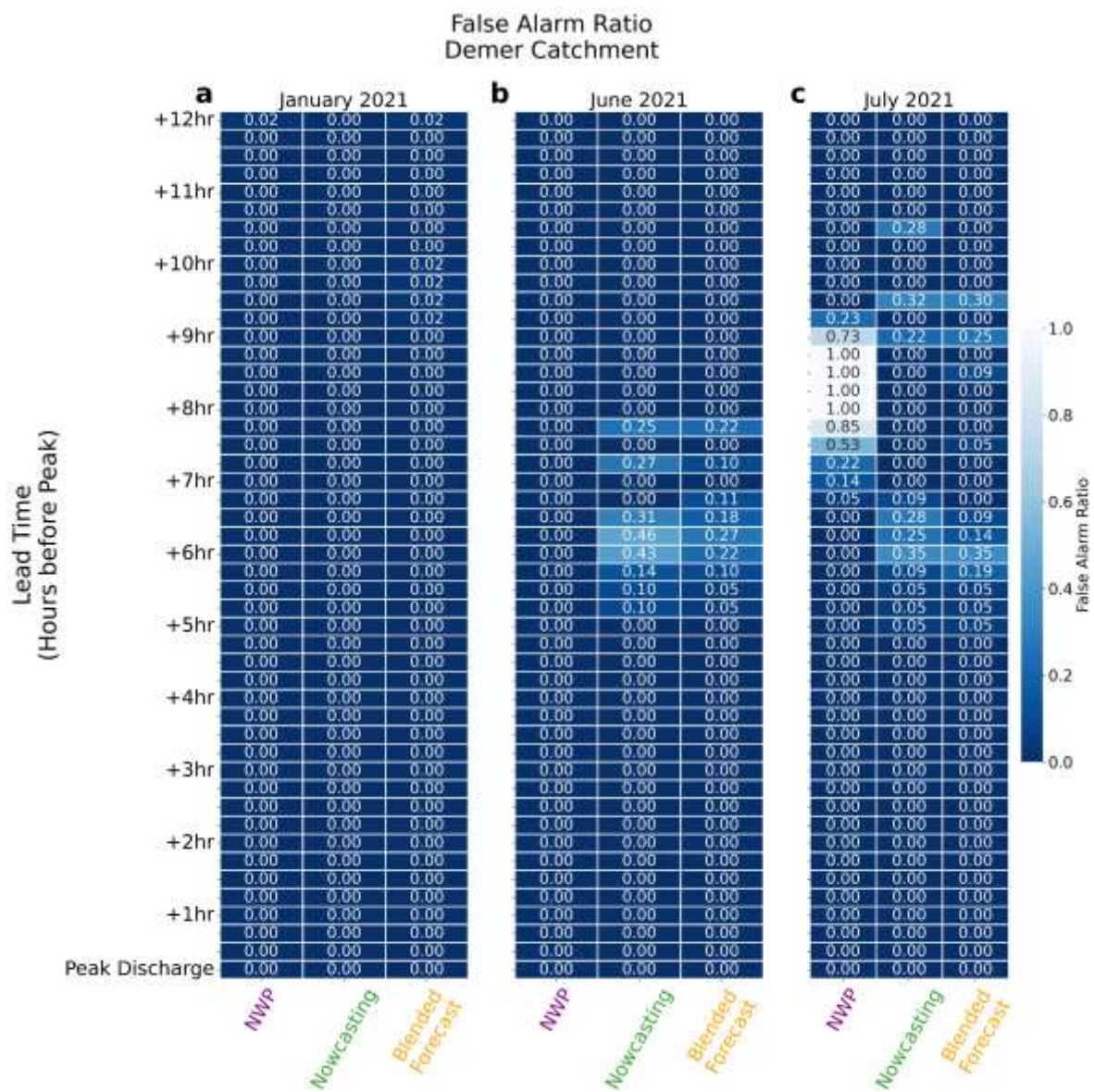
Figure A.7: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. False alarm ratio ranges from 0 (perfect score) in dark blue to 1. The panels show the results for the Geul catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Figure A.8: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. False alarm ratio ranges from 0 (perfect score) in dark blue to 1. The panels show the results for the Rur catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
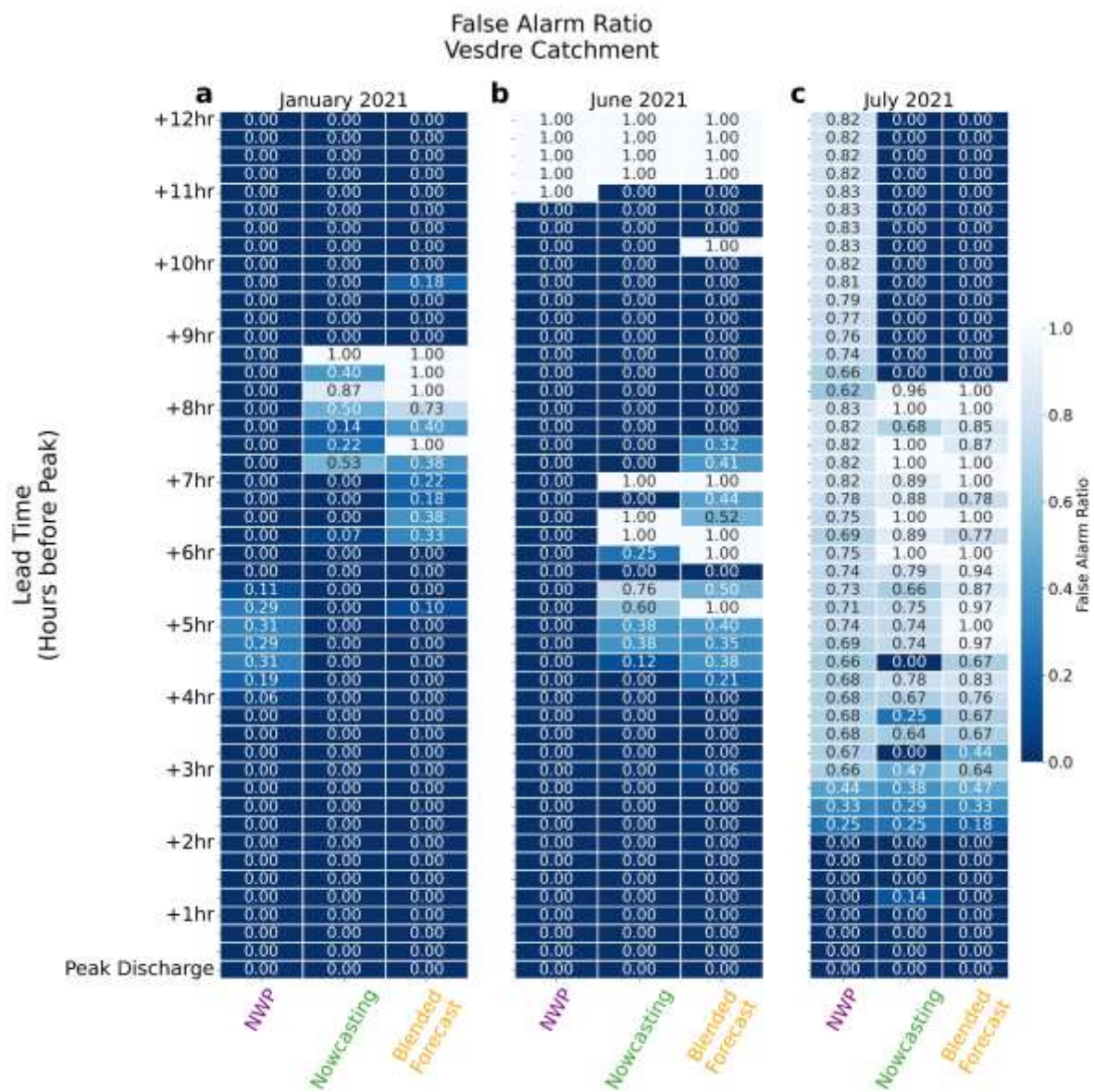
Figure A.9: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. False alarm ratio ranges from 0 (perfect score) in dark blue to 1. The panels show the results for the Demer catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
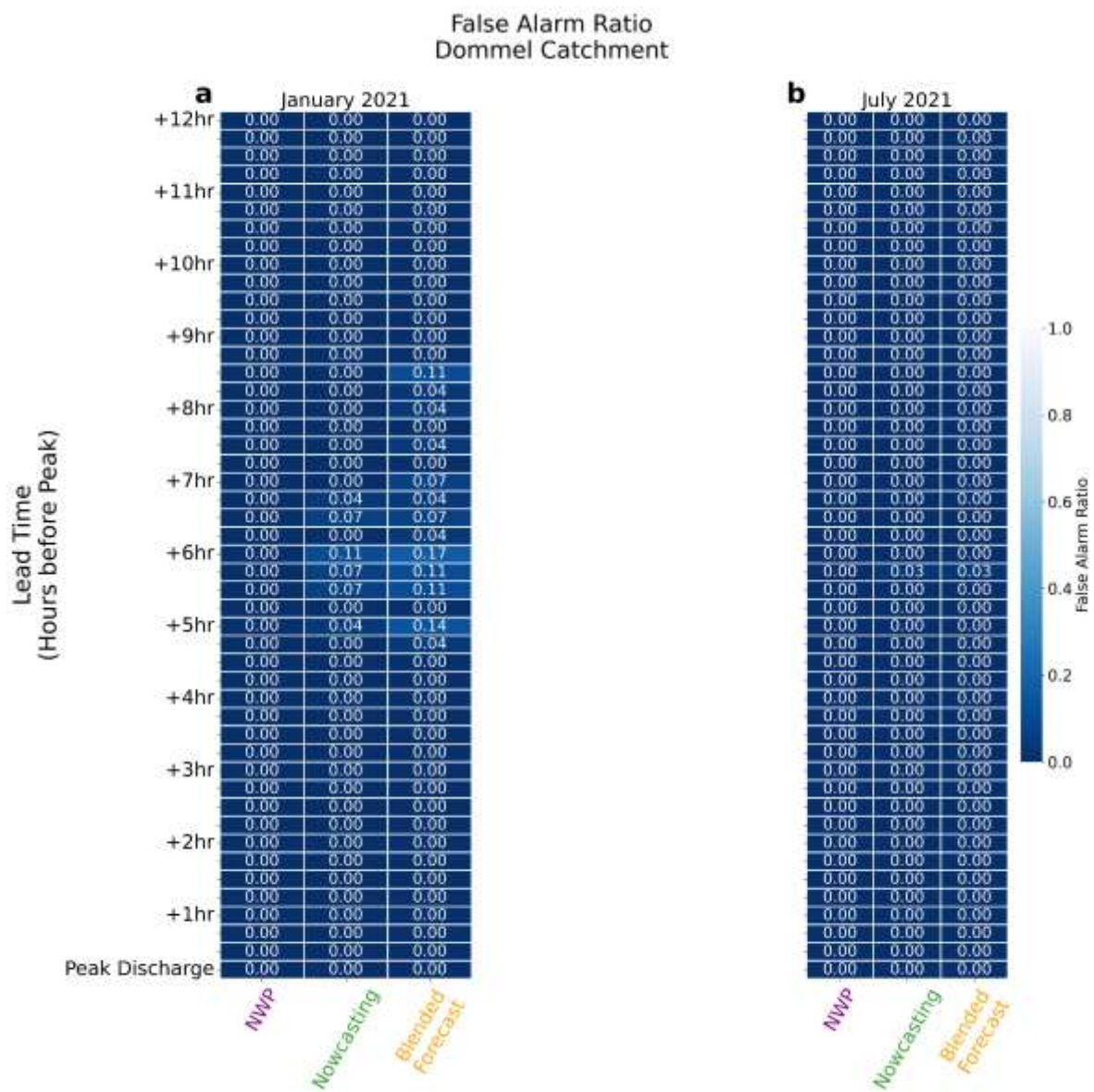
Figure A.10: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. False alarm ratio ranges from 0 (perfect score) in dark blue to 1. The panels show the results for the Vesdre catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
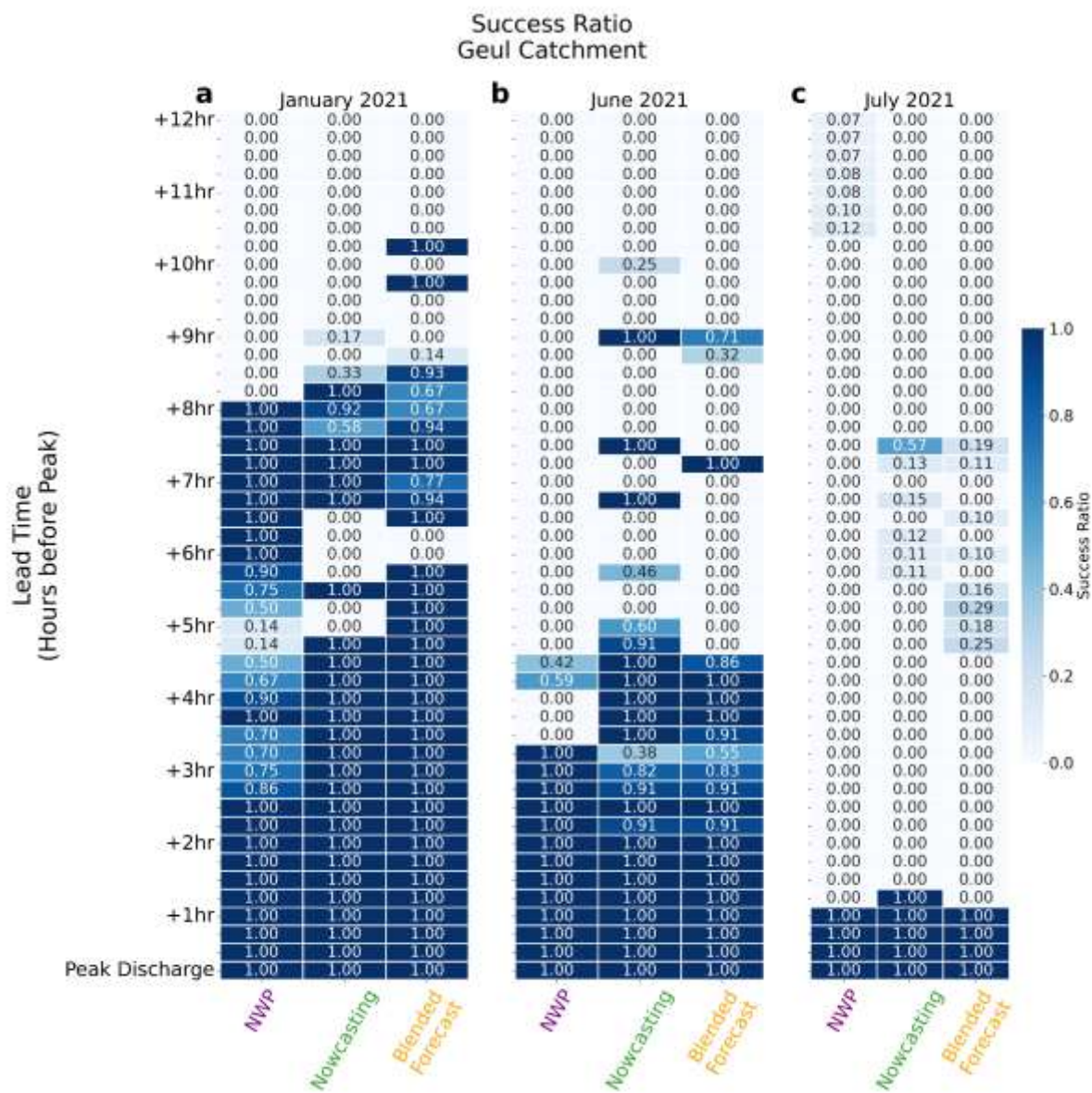
Figure A.11: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. False alarm ratio ranges from 0 (perfect score) in dark blue to 1. The panels show the results for the Dommel catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

A.4 Success Ratio

The success ratio ranges from 0 to 1 (perfect score) and answers the question: *What fraction of the forecast "yes" events were correctly observed?*

Mathematically, the success ratio is defined as:

$$Success\ Ratio = \frac{hits}{hits + false\ alarms} \qquad \text{(Eq. 5)}$$

Figure A.12: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Success ratio ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Geul catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
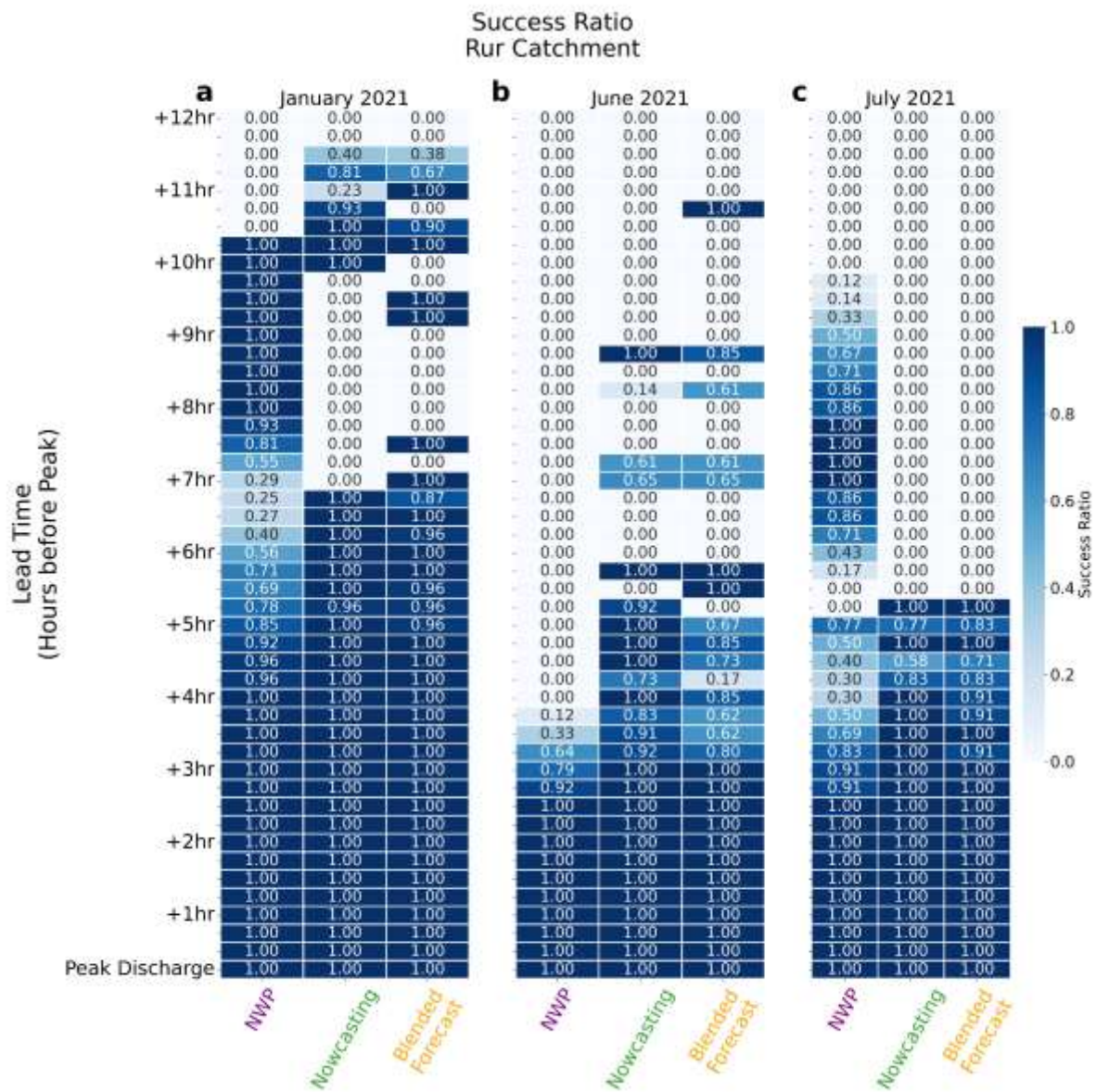
Figure A.13: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Success ratio ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Rur catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
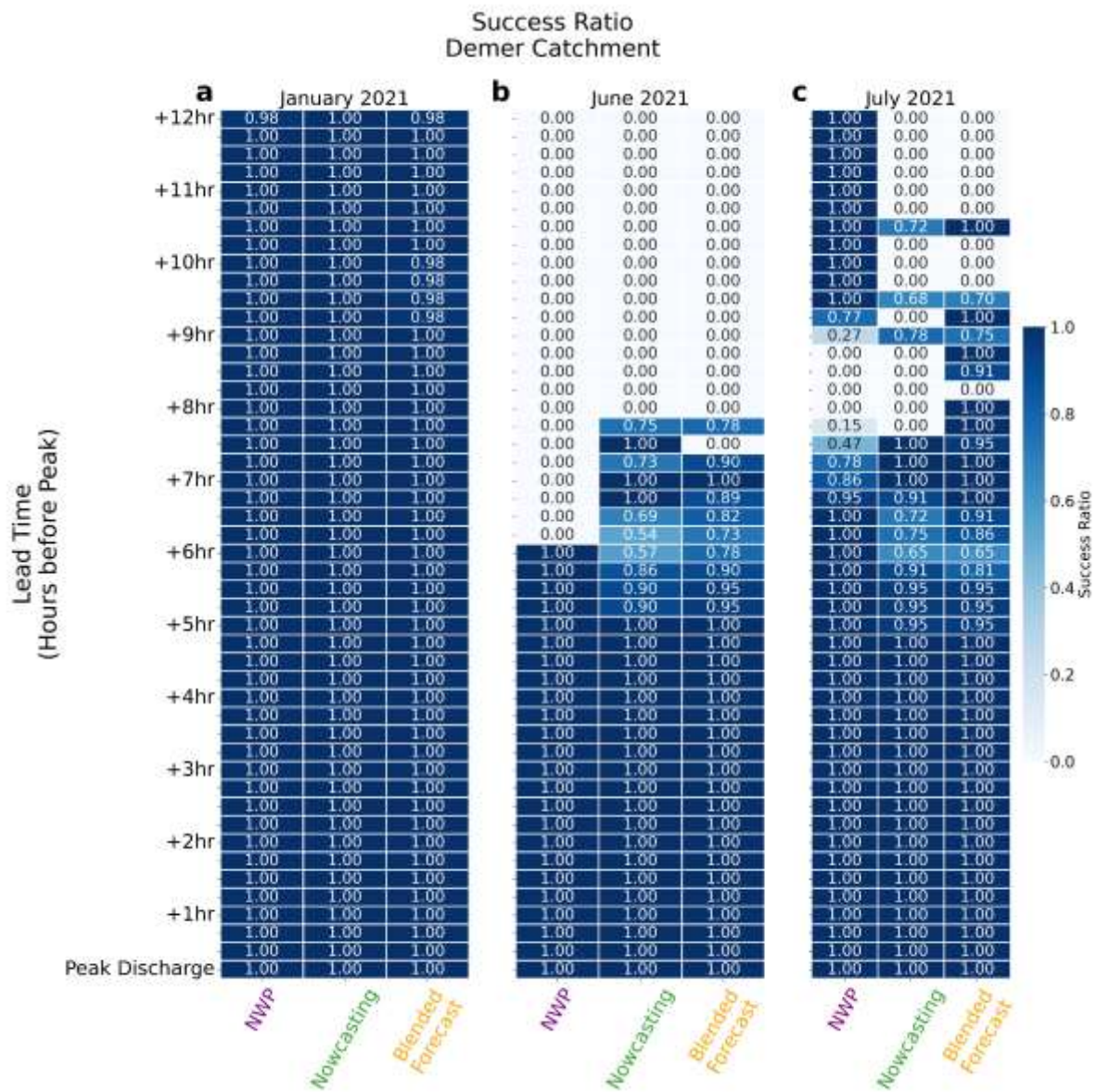
Figure A.14: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Success ratio ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Demer catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
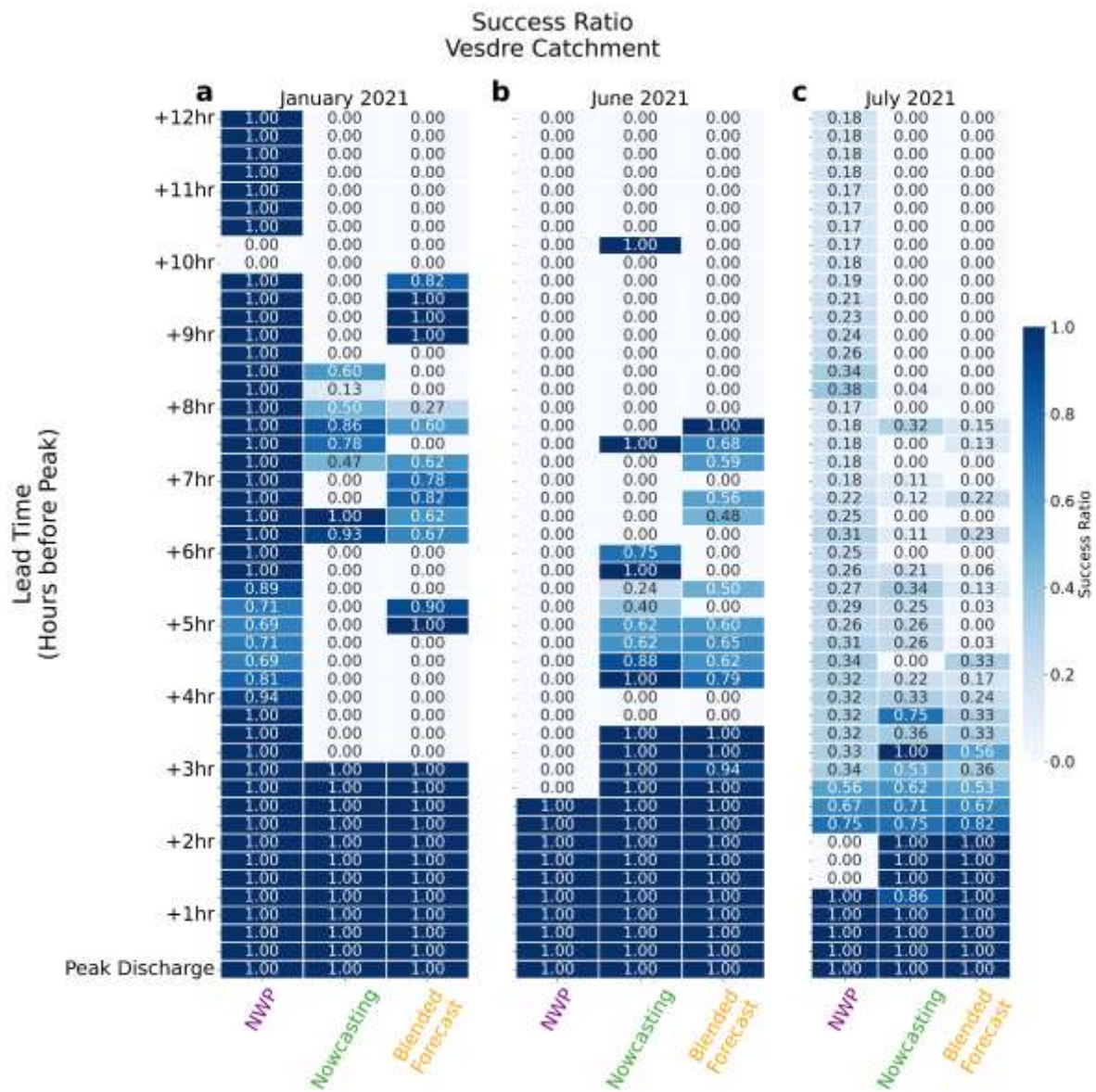
Figure A.15: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Success ratio ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Vesdre catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.
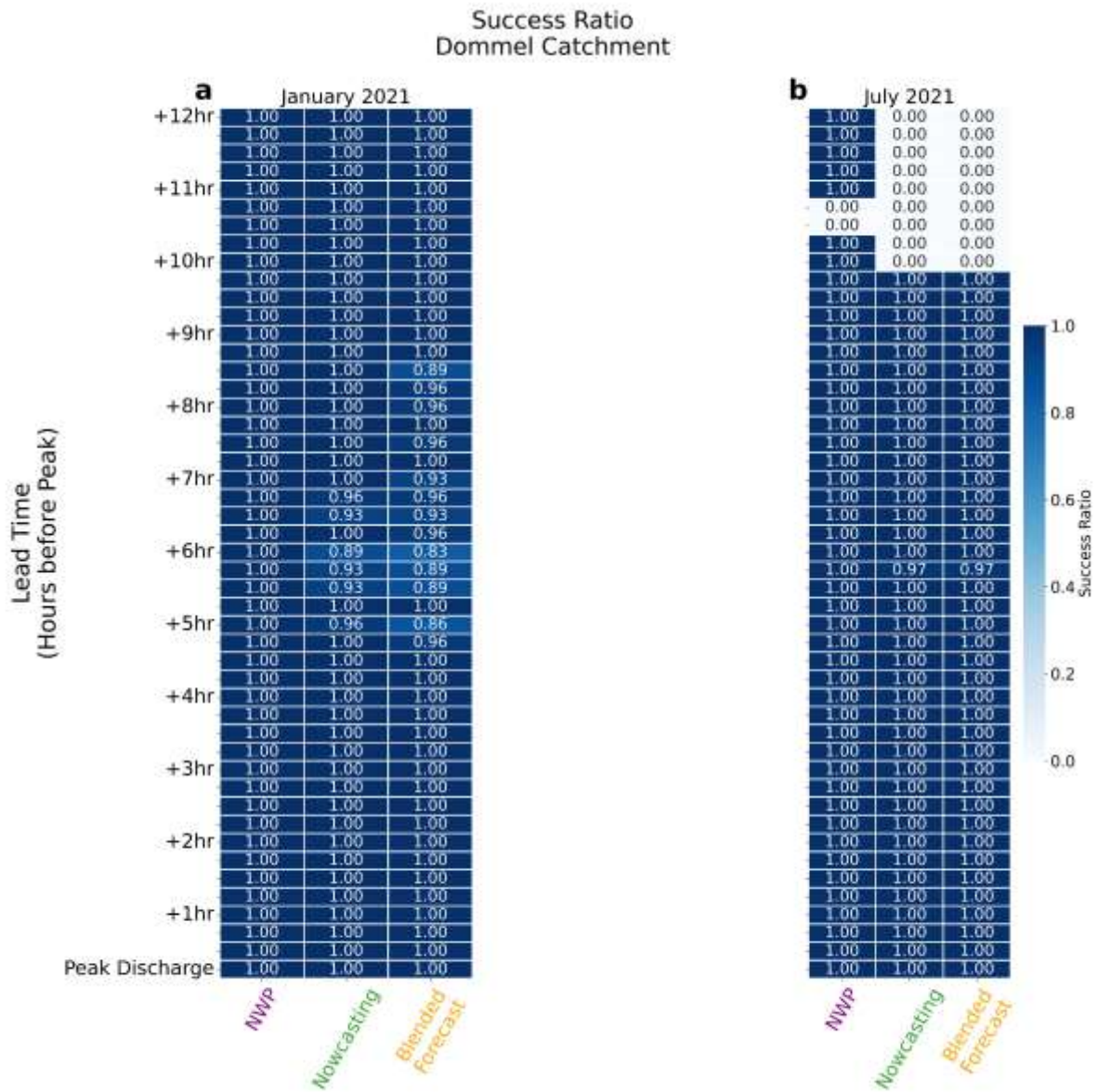
Figure A.16: Heatmap displaying the lead time in hours before peak discharge on the vertical axis. The horizontal axis shows the discharge forecast models, with in purple the NWP, in green the Nowcasting, and in orange the Blended Forecast models. Success ratio ranges from 0 to 1 (perfect score) in dark blue. The panels show the results for the Dommel catchment for (a) the stratiform January 2021 event, (b) the convective June 2021 event, and (c) the combined July 2021 event.

Appendix B
B.1 Official Declaration of Use by Waterschap Limburg
Niet openbaar