Meteorological Applications

RMetS Royal Meteorological Society

RESEARCH ARTICLE OPEN ACCESS

N-FOREWARNS: Regional Nowcasting of Surface Water Flooding With a 3-h Lead Time Predictability Limit

Joseph Smith[1] | Cathryn Birch[1] | Julia Perez[2] | Linda Speight[3] | John Mooney[2] | Ben Maybee[1] | John Marsham[1]

[1]School of Earth and Environment, University of Leeds, Leeds, UK | [2]Flood Forecasting Centre, Exeter, UK | [3]School of Geography and the Environment, University of Oxford, Oxford, UK

Correspondence: Joseph Smith (eejasm@leeds.ac.uk)

Received: 12 May 2025 | **Revised:** 28 November 2025 | **Accepted:** 28 January 2026

Keywords: expert engagement | nowcasting | surface water flooding

ABSTRACT

Surface water flooding is generally caused when intense rainfall overwhelms drainage systems and/or floods areas before reaching a major watercourse. A key driver of surface water flooding is rapidly developing, localised convection, which is poorly represented in numerical weather prediction models. Nowcasting uses current atmospheric observations to make more accurate predictions of future convective activity on hourly timescales, making it a well-suited approach for surface water flood (SWF) prediction. This study presents Nowcasting-FOREWARNS (N-FOREWARNS), a development of the daily Flood fOREcasts for Surface WAter at a RegioNal Scale (FOREWARNS) tool, which has been modified here to produce hourly SWF nowcasts for lead times of 1–6 h. User feedback from a nowcast comparison testbed indicated that N-FOREWARNS produced few false SWF predictions. When predictions were made by N-FOREWARNS, they were reported to have high spatial accuracy, with users noting an incorrect location only 7% of the time. However, users also reported that N-FOREWARNS missed a high proportion of smaller-scale SWF events with minor impacts. Quantitative verification shows that N-FOREWARNS's 1 h lead time nowcasts produce useful skill, comparable to FOREWARNS, but N-FOREWARNS reaches a limit of predictability at 3 h lead time, and that the skill is primarily limited by the accuracy of its rainfall inputs. Overall, these results show promise for nowcasting regional-scale SWFing and provide a benchmark for further development, which should focus on improving the accuracy of convective rainfall inputs.

1 | Introduction

Surface water flooding (also known as pluvial flooding) is defined as rainfall-generated water that pools on or flows over the ground surface, before it enters a major watercourse or drainage system (Falconer et al. 2009; Speight et al. 2021). This type of flooding is typically caused by intense convective weather events that produce high volumes of rainfall over localised regions, in short periods of time. Due to the rapid onset of these events, there is often little time to prepare for surface water flooding, which results in severe impacts to local communities (Carlyon et al. 2012). It is estimated that 4.6 million properties in England are at risk of surface water flooding (Environment Agency 2025). The Coverack, Cornwall flood on 18 July 2017 provides an example of the impacts that a surface water flood (SWF) can cause, during which an estimated 165–201 mm of rainfall fell in 3 h (Flack et al. 2019), destroying many houses in the village with flowing water up to 1 m in depth (Baker 2017). This required helicopter rescues to evacuate local residents from their homes (BBC News 2025). The event was caused by an extremely localised convective storm over Coverack—the nearby rain gauges recorded no significant rainfall. Improved early warning that the event might happen, even an hour ahead of the event, could have at least allowed the emergency services to be on standby for a quicker response.

Since the Flood and Water Management Act 2010 (UK Parliament 2010a), UK local government authorities have been responsible for the management of local SWF risk. However, varying levels of SWF exposure across the United Kingdom mean that the capability to undertake this management role differs between local authorities (Speight et al. 2025). To support local authorities in England and Wales, the Flood Forecasting Centre (FFC), a partnership between the Environment Agency and Met Office to improve flood risk guidance, issues a daily Flood Guidance Statement (Flood Forecasting Centre 2025). The Flood Guidance Statement provides a 5-day flood risk for surface water, river, groundwater and coastal flooding by using a risk matrix approach that details both likelihoods and impacts (Pilling 2016). The Flood Guidance Statement is primarily targeted at those involved in strategic and operational planning decisions for emergency response to flooding (Figure 5 provides an example of some of the information included in the Flood Guidance Statement).

Over the last 15 years, NWP modelling has vastly improved capability in forecasting convection development (Clark et al. 2016). To produce a skilful forecast, these models require atmospheric precursors or existing convective features to be well represented in the initial conditions. However, surface water flooding is often caused by rapidly developing, localised convection on scales smaller than even convection-permitting NWP grid resolutions (Clark et al. 2016; Flack et al. 2019), and skill at NWP lead times of a few hours is problematic due to the required spin-up period (Lean et al. 2008). Nowcasting provides a solution to this by making use of the latest atmospheric information to generate rapid, short-term forecasts (approximately ~0–6 h ahead) of convective weather (Roberts et al. 2022). Typically, methods such as optical flow (Bowler et al. 2006; Hill et al. 2020) or machine learning (Ayzel et al. 2020; Ravuri et al. 2021; Zhang et al. 2023) are employed to forward extrapolate rainfall in radar images, producing near-instantaneous predictions on the same grid scales as the rainfall observations. Nowcasting methods, therefore, provide an opportunity for SWF prediction for rapid emergency response.

Although nowcasting can improve skill at capturing rainfall on short lead times, the inherent chaotic nature of convection means that this is still a challenging problem. In particular, the stochastic behaviour of initiation results in low skill, even for high spatiotemporal nowcasts (Smith et al. 2024). Not only does SWF prediction require an accurate rainfall prediction, but it also relies on capturing non-meteorological factors such as land use, soil and orography. These additional factors add further complexity into the problem and emphasise the challenge of SWF prediction.

A wide range of methodologies have been developed for SWF forecasting, for urban to national scales (Speight et al. 2021). The least complex methods use empirically-based rainfall thresholds to derive SWF warnings (Acosta-Coll et al. 2018; Javelle et al. 2016). Warnings are triggered if rainfall observations (e.g., rain gauges or radar estimates), exceed pre-defined depth-duration thresholds. These approaches offer a good initial warning estimate but do not take into consideration the dynamically changing conditions that can affect SWF, for example, ground saturation levels. The most advanced approaches run real-time hydrodynamic simulations that are input into hydrological models to produce high resolution warnings, predominantly for urban areas (Green et al. 2017). These methods provide street level detail, while the use of high spatiotemporal resolution rainfall data allows capability for SWF nowcasting (Yu and Coulthard 2015). High computational costs and long run-times mean that these models are not currently suitable for operational SWF forecasting or producing national-scale warnings. However, as computational capacity increases in the future, it will be possible to operationally run these models on urban scales, improving capability for areas where flooding has the highest impact.

As a compromise between the approaches discussed so far, hydrological models are combined with realistic rainfall fields and predefined impact scenarios to produce warnings on local, regional and national scales (Speight et al. 2018; Maybee et al. 2024; Dottori et al. 2017). Scenario-dependent SWF impact warnings are produced by applying appropriate flood inundation maps to the output from a hydrological model. These approaches account for the spatial variability in rainfall, while also maintaining low computational expense.

Flood fOREcasts for Surface WAter at a RegioNal Scale (FOREWARNS) was originally developed to provide national-scale SWF forecasts for whole calendar days, at 1–4 days lead times (Maybee et al. 2024). It generates reasonable worst case rainfall scenarios (RWCRSs) (Böing et al. 2020) from convective-scale ensemble NWP forecasts (MOGREPS-UK, Porson et al. 2020) and compares them to predefined look-up flood thresholds from offline hydrological modelling. Maybee et al. (2024) showed that FOREWARNS is a suitable tool for operational forecasting of surface water flooding. However, a key outcome of this study was that, given the challenges of forecasting rapidly intensifying convection with NWP, producing SWF predictions at shorter lead times (0–6 h) using rainfall nowcasts, could fulfil the user-need for more reliable forecasts.

This paper presents Nowcasting-FOREWARNS (N-FOREWARNS), a modified version of FOREWARNS that is driven by probabilistic nowcast rainfall from the Short Term Ensemble Predictions System (STEPS, Bowler et al. 2006), and can produce maps of SWF risk predictions every hour with lead times of 1–6 h. The aim of this development is to test the ability of reducing the latency and increasing the temporal resolution of the flood forecasts, while maintaining good forecast skill. This paper presents the results of a SWF forecasting testbed, hosted by the Met Office and FFC during the 2024 summer season, in which expert participants provided feedback on the performance of N-FOREWARNS alongside current operational SWF forecast tools. Additionally, a quantitative verification of N-FOREWARNS is presented across lead times of 1–6 h for 70 recorded SWF events from 2013 to 2022, and for 3470 potential issuances at 1 h lead time across the 2024 summer period.

2 | Methods

2.1 | Data

2.1.1 | STEPS

STEPS is a rainfall nowcasting model that was jointly developed by the Met Office and the Bureau of Meteorology, Australia (Bowler et al. 2006). It produces an ensemble forecast that blends together nowcast radar extrapolations with NWP forecasts. For up to ~3 h lead time, it uses an optical flow method (Pulkkinen et al. 2019), which forward propagates storm cells that have been tracked through radar image series (producing a future radar prediction). After ~3 h lead time the model transitions to using downscaled output from the 12-member MOGREPS-UK model.

To address the uncertain nature of convection, STEPS adds varying intensities of stochastic noise into its prediction. It assumes that smaller-scale convection is associated with a greater uncertainty, and therefore injects it with a more intense noise field than relatively larger scale convection. Furthermore, as lead times increase, STEPS adds increasingly more intense noise into the extrapolation fields to account for the temporal uncertainty of convection. It uses 12 different initialisations of stochastic noise fields to produce a 12-member ensemble.

Overall, after the blending and noise injection stages, STEPS produces a 24-ensemble member nowcast (12 ensemble members from MOGREPS-UK plus 12 ensemble members from noise injections) that covers the United Kingdom with a resolution of 2.2 km, every 15 min, for lead times up to 6 h.

2.1.2 | Rainfall Radar Data

For ground truth in the quantitative verification, this study uses ground-based radar retrievals of rainfall as input into N-FOREWARNS. Rainfall data sets derived from 15 C-band radars across the UK are used to form the Met Office Nimrod system (Golding 1998; Harrison et al. 2012), which was used in this study. The output is produced on a 1 km grid every 5 min and has been verified against rain gauge measurements across the United Kingdom (Harrison et al. 2012).

2.1.3 | SWF Impact Reports

Identifying SWF events and their observed impacts presents a key challenge in SWF prediction. Currently, the most effective approach at national scale is to collate information using algorithms that scan a variety of sources such as social media, news publications and official reports, which mention surface water flooding (De Bruijn et al. 2019; Flood Tags 2025).

During the testbed part of this study, Hydrometeorologists at the FFC collated reports of SWF events across the United Kingdom, which provided a source of observational data for verification. Reported SWF events were identified from social media and professional partner sources. The location, dates and impact severity of each reported event are identified, and then the data is interpolated onto a county-scale map (e.g., Figure 1).

Maybee et al. (2024) built a record of observed minor and significant impact SWF events for Northern England using the Global Flood Monitor (GFM; https://www.globalfloodmonitor.org/, last accessed 6 March 2025) to identify observed events between 2013 and 2022 May–October. In order to make a comparison, this study chooses 70 SWF event days from that dataset (chosen based on data availability) as part of the quantitative verification of N-FOREWARNS. A verification over the full summer 2024 period is also presented to understand N-FOREWARNS's performance on a more climatologically representative dataset.

Applying methods that scan various sources for SWF events enables a range of useful impact information to be captured. However, SWF events that are not reported through these sources will be missed (De Bruijn et al. 2019), for example, events with localised impacts. Acknowledging this key caveat when interpreting any impact-based SWF verification is important (Maybee et al. 2024).

2.2 | SWF Forecast Tools

2.2.1 | N-FOREWARNS

N-FOREWARNS is a modified version of FOREWARNS (Maybee et al. 2024). It was co-designed with users (Birch et al. 2021) and is intended to be used by professional forecasters

FIGURE 1 | *Left*. An example of a county scale map of reported surface water flood impacts, with their associated severity level, for 15 July 2024 (impacts may have been reported at any time during the day). *Right* An image extracted from a post on https://x.com/ giving evidence of minor surface water flooding in Worthing, Sussex. This SWF event report is scaled up to the county level to produce the mapped impact level shown (*left*).

in the routine production of national guidance such as the daily Flood Guidance Statement by the FFC.

The FOREWARNS method has two key steps:

1. It generates RWCRSs for a range of rainfall accumulation periods (1, 3 and/or 6 h) across a 24-h forecast period. This is done by applying a 98th percentile-based neighbourhood processing method (Böing et al. 2020) to ensemble rainfall NWP forecasts of rainfall.
2. Rainfall accumulations from the RWCRSs are then translated into catchment-level SWF return periods by conducting look-up comparisons against pre-run rainfall modelling (Vesuviano 2022) underpinning the UK Risk of Flooding from Surface Water maps (a dataset of national street-level SWF risk available at https://check-long-term-flood-risk.service.gov.uk/map). The outputs from the look-up comparison of each scenario are then combined to create a single, reasonable worst case scenario SWF risk map of flood return periods (5, 10, 30, 100 and 1000 years).

N-FOREWARNS is run with three key modifications:

I. The NWP ensemble rainfall forecasts used in Stage 1 of the FOREWARNS setup are replaced with STEPS nowcast rainfall predictions.

II. A 3-h period of STEPS rainfall data is used as input instead of the 24-h period used in FOREWARNS. Therefore, N-FOREWARNS calculates RWCRSs using 1- and 3-h accumulation periods only (unlike 1, 3 and 6 h for FOREWARNS). A 3-h input period was chosen as a balance between capturing the majority of past rainfall that causes a SWF event and the need for short lead times. The 3-h period is selected so that, for a given SWF valid time, the period ends at the valid time. For example, for an N-FOREWARNS prediction of SWF flooding at valid time 1600 UTC, STEPS rainfall over the 3 h period 1300–1600 UTC would be used, from the STEPS run.

III. The catchment-level output is linearly interpolated onto a 20 km grid for direct comparison against other SWF nowcasting tools, whereas the output was presented at the catchment scale (30–40 km) in FOREWARNS.

As discussed in Section 2.1.3, the lack of comprehensive observed SWF datasets presents a key challenge in evaluating SWF nowcast output directly to SWF impacts. Therefore, to enable a robust quantitative verification, a radar-driven version of N-FOREWARNS—referred to as R-FOREWARNS—is produced to act as a proxy for SWF ground truth. To do this, the STEPS rainfall data is replaced by the Nimrod (radar) rainfall observations from the same 3-h period (as in Step II). Figure 2 provides an example of N-FOREWARNS and the corresponding R-FOREWARNS. Defining R-FOREWARNS as the ground truth allows a direct evaluation of N-FOREWARNS; however, this comparison is an assessment on the STEPS rainfall input (and its ability to capture SWF risk) only and does not provide insight into N-FOREWARNS's capability in capturing recorded SWF events.

The 3-h periods in N-FOREWARNS for step II above are only available from a single STEPS run for the 3, 4, 5 and 6 h lead times (i.e., 0–3, 1–4, 2–5 and 3–6 h from the STEPS initialisation). In order to extract a 3-h rainfall period for the 1 and 2 h lead time, data are extracted from previously initialised STEPS runs as shown in Figure 3.

2.2.2 | Surface Water Flooding Hazard Impact Model (SWFHIM)

SWFHIM is a SWF flood forecast tool that is currently in operational use by the FFC (Aldridge et al. 2020) for lead times of ~6-h to 3 days. It produces probabilistic output on a 1 km grid scale for minimal, minor, significant and severe SWF impact levels (Figure 4). The SWFHIM method is split into three parts: hazard model, impact model and risk mapping. The hazard model component comprises the FFC's Grid-to-Grid hydrological model (Cole and Moore 2009), which is driven by ensemble NWP rainfall forecasts from MOGREPS-UK to produce a 1 km gridded surface runoff field. The impact model is based on the Impact Library—a static dataset describing predefined impact scenarios in terms of their trigger criteria and their associated impact levels (upon transport, property, population and infrastructure). The Impact Library is mapped onto a 1 km grid, enabling the impact model to translate hazard model output to impact levels at this scale. The hazard and impact models both operate on the individual ensemble members of the NWP rainfall forecast. In the risk mapping component, the outputs from the impact model and the hazard model are combined and upscaled to produce ensemble-based likelihood maps for each SWF impact level (minimal, minor, significant and severe).

For this study, SWFHIM was also run in nowcast mode using 3-h periods. As with the modifications to N-FOREWARNS outlined in Section 2.2.1, 3-h periods of STEPS ensemble predictions were used. Outputs were upscaled to a 20 km grid (Figure 4). Table 1 provides the key comparison of N-FOREWARNS and SWFHIM.

2.2.3 | Rapid Flood Guidance (RFG)

The FFCs RFG operational trial ran from 14 May to 30 September 2024 (Flood Forecasting Centre 2024) to support emergency responders to prepare for rapidly developing SWF events that occur due to summertime convection over England and Wales. The key aims of the RFG service were

- To give short notice updates for England and Wales to supplement the daily Flood Guidance Statement.
- To be designed for responders who need to make decisions at a timescale of 0 to 6 h.

During the trial, operational forecasters at the Met Office were responsible for issuing the RFG. Traditional observation sources such as composite and single-site radar parameters, rain gauges, satellite imagery and radiosonde ascents were analysed in combination with operational nowcasting techniques (STEPS), high-resolution short-range NWP (MOGREPS-UK) and forecaster experience to generate the RFG output. Note N-FOREWARNS

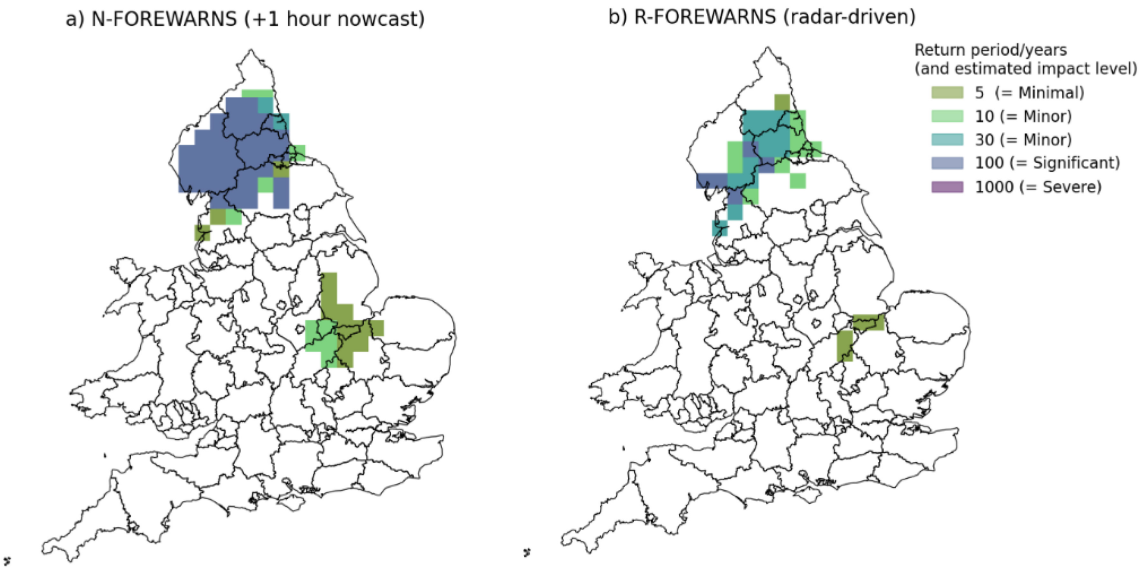


FIGURE 2 | An example from 10 September 2023 valid at 1500 UTC of a SWF warning from (a) N-FOREWARNS (nowcast-driven with a 1 h lead time), that is, the prediction and (b) R-FOREWARNS (radar-driven), that is, the verification.

and SWFHIM in nowcasting mode were not used in the RFG production process. Each RFG contained a forecaster-drawn polygon over the region of impact, along with written information describing the timing, likelihood and characteristics of the weather that is driving the rapid flooding (Figure 5).

2.3 | UK Summer 2024 SWF Testbed

2.3.1 | Testbed Summary and Aims

Aligned in time with the RFG operational trial (summer 2024), the FFC hosted a SWF nowcasting testbed in collaboration with researchers from the University of Leeds and the University of Oxford. During the testbed, the outputs from N-FOREWARNS and SWFHIM in nowcasting mode were evaluated by meteorological/hydrological expert participants and compared to the RFG and to observations of SWF impacts. The key aims of the testbed were to:

1. Compare and evaluate the performance of N-FOREWARNS and SWFHIM SWF nowcasts.
2. Assess whether the use of N-FOREWARNS and/or SWFHIM nowcasts by operational forecasters would have added value to the production of the RFGs.

2.3.2 | Testbed Structure

During the testbed, 20 events were analysed covering 11 different days (see Figure S1 in all events) with each event aligning with the issuance of an RFG warning. N-FOREWARNS and SWFHIM were then run at the same initialisation and lead time to enable a direct comparison between all three tools. A total of 50 individual participants took part across the duration of the testbed. For each event 5–20 participants (including the authors of this paper) were split into groups, with each group providing an independent subjective assessment of the N-FOREWARNS, SWFHIM and RFG predictions. Across all events a total of 92 groups provided subjective assessments (i.e., 92 individual data points). Participants came from a range of background expertise: hydrometeorologists, operational forecasters and university academics. The session focussed on a post-event assessment and was structured in the following way:

1. The flooding event (a recent event, if possible, from the previous week) was introduced to the participants with an overview of the meteorological conditions.
2. Participants groups were provided with the SWF nowcasts from N-FOREWARNS and SWFHIM and, if appropriate, asked to draw a flooding impact polygon over England and Wales based on each forecast (Figure 6). They were then required to identify the most appropriate impact severity for each polygon from Minor, Significant and Severe.
3. The rainfall radar observations, R-FOREWARNS and the observed impacts (Figure 7), plus the issued RFG were revealed to the groups to compare and subjectively verify. The groups were then asked to complete a survey (Figure S2), which focused on evaluating different aspects of the performance of N-FOREWARNS, SWFHIM and RFG.
4. Finally, each group was asked to feedback their key points during an all-participant discussion.

Drawing the impact polygons in Step 2 meant that participants were forced to interpret and consider strengths and weaknesses of the N-FOREWARNS and SWFHIM tools. Furthermore, it provided benchmarks to compare with the RFG polygon.

2.4 | Verification Metrics

During the testbed the participants were asked to qualitatively categorise the SWF nowcasts that were produced by N-FOREWARNS and SWFHIM into one of: hit (a; SWF nowcast event aligns with the SWF observed event), miss (b; the SWF

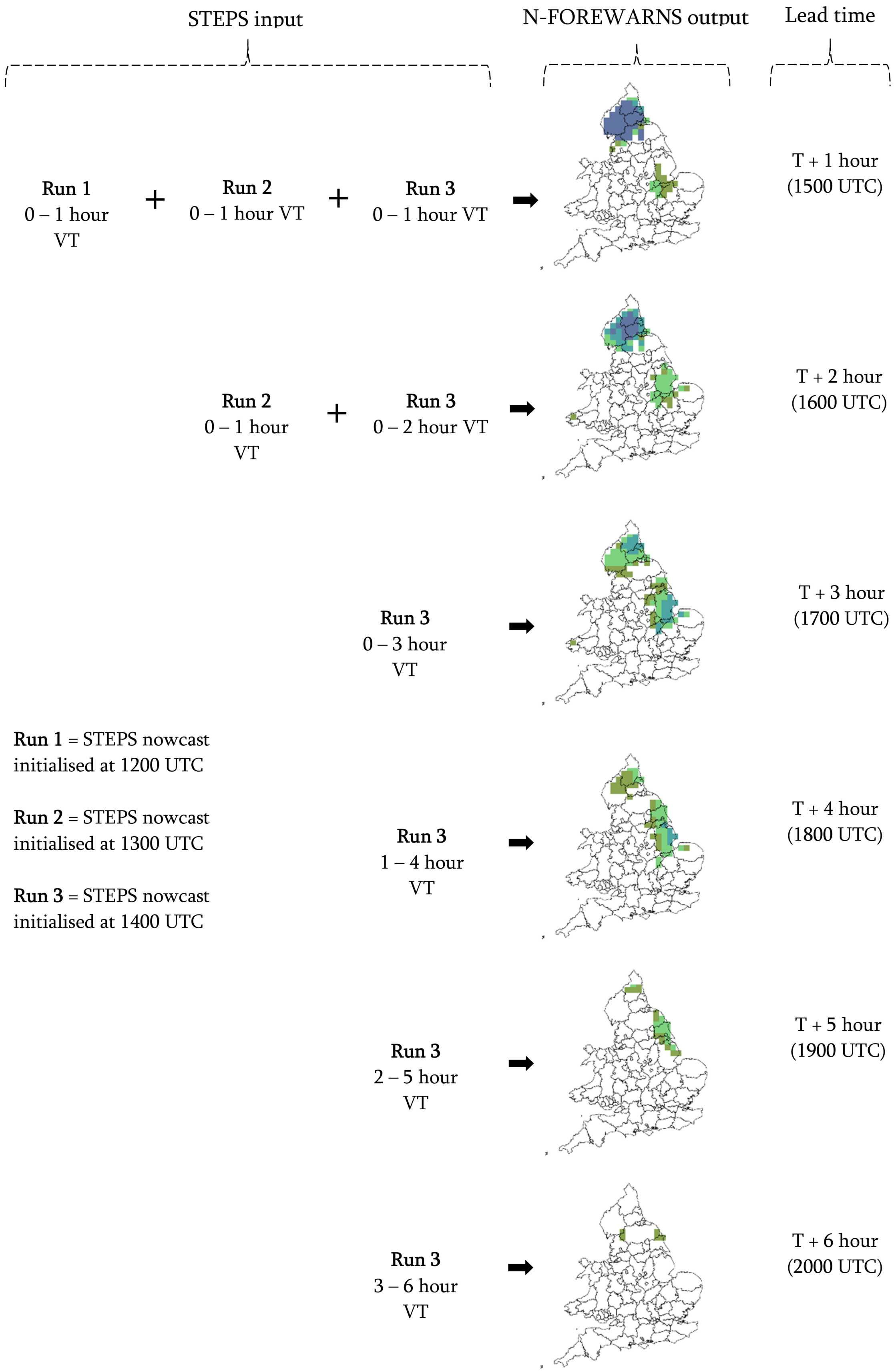


FIGURE 3 | A diagram showing how the 3-h period of STEPS rainfall data is inputted for each lead time (represented by each row). The STEPS run and period of valid time (VT) is shown for each N-FOREWARNS lead time output. An example from 10 September 2023 is shown, which has been initialised at 1400 UTC.

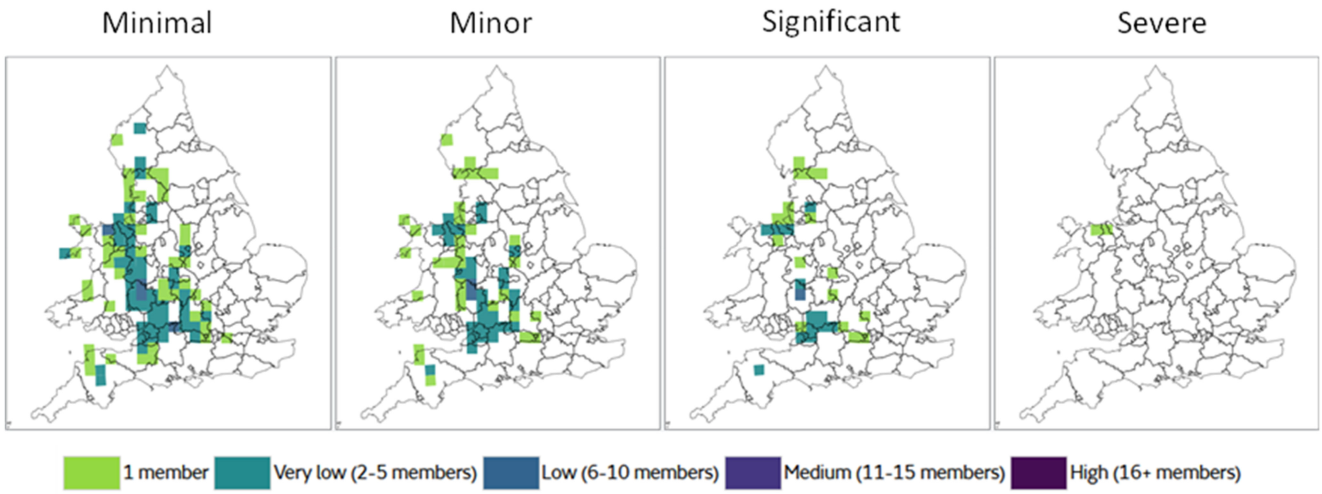


FIGURE 4 | An example of a SWFHIM SWF forecast valid at 1900 UTC on 12 May 2024.

TABLE 1 | Comparison table between N-FOREWARNS and SWFHIM.

	N-FOREWARNS	SWFHIM
Nowcast rainfall field	STEPS	STEPS
Testbed output grid scale	20 km	20 km
SWF risk measure	Return period	Ensemble member count
Output format	Single reasonable worst case scenario map	Impact map for each severity level

observed event is not captured by the SWF nowcast), correct rejection (c; no SWF event is produced in the nowcast or observation) and false alarm (d; a SWF event is nowcasted but no SWF event is observed). These results are presented in Section 3.1.

For the quantitative analysis presented in Section 3.2, skill scores are derived from the verification of N-FOREWARNS against R-FOREWARNS. The Hit Rate (H) is given by:

$$H = \frac{a}{a+b},$$

which captures the proportion of observed SWF events that were correctly nowcasted. An H value of 1 represents a perfect score and 0 represents the least skill (in terms of H).

The False Alarm Rate (F) is given by:

$$F = \frac{d}{c+d},$$

which captures the proportion of observed no-SWF events that were nowcasted as a SWF event. An F value of 0 represents a perfect score and 1 represents the least skill (in terms of F). To be consistent with H, $1-F$ is presented in the results (meaning a higher score is always better).

The success ratio (SR) is given by

$$SR = \frac{a}{a+d},$$

which provides the likelihood of an observed SWF event, given that a SWF nowcast is produced. This skill metric provides information on the balance between false alarms and hits. An SR value of 1 gives a perfect score and 0 represents the least skill (in terms of SR).

3 | Results

3.1 | Testbed Results

3.1.1 | Testbed Case Study Events

Table 2 contains five example events selected from the 20 covered in the testbed. These five events were subjectively chosen as they capture a range of different weather scenarios and SWF nowcast outputs, enabling the key advantages and disadvantages of each tool to be discussed (only minor impacts presented for SWHIM output).

During Event 1, rainfall accumulations of approximately 20 mm were observed in the north of England, but no observed flood impacts were reported. N-FOREWARNS produced a minimal prediction that aligned with the rainfall accumulation, whereas the SWFHIM minor impact prediction was more widespread with lower precision, covering the majority of northern England and reaching south to Norfolk. The RFG contained a polygon constrained to the northeast, covering the rainfall regions, but also extending into regions where no impactful rain was observed.

In Event 2, SWFHIM produced a more constrained and coherent minor impact prediction. Its low–medium likelihood regions aligned closely with the accumulated rainfall and coincided

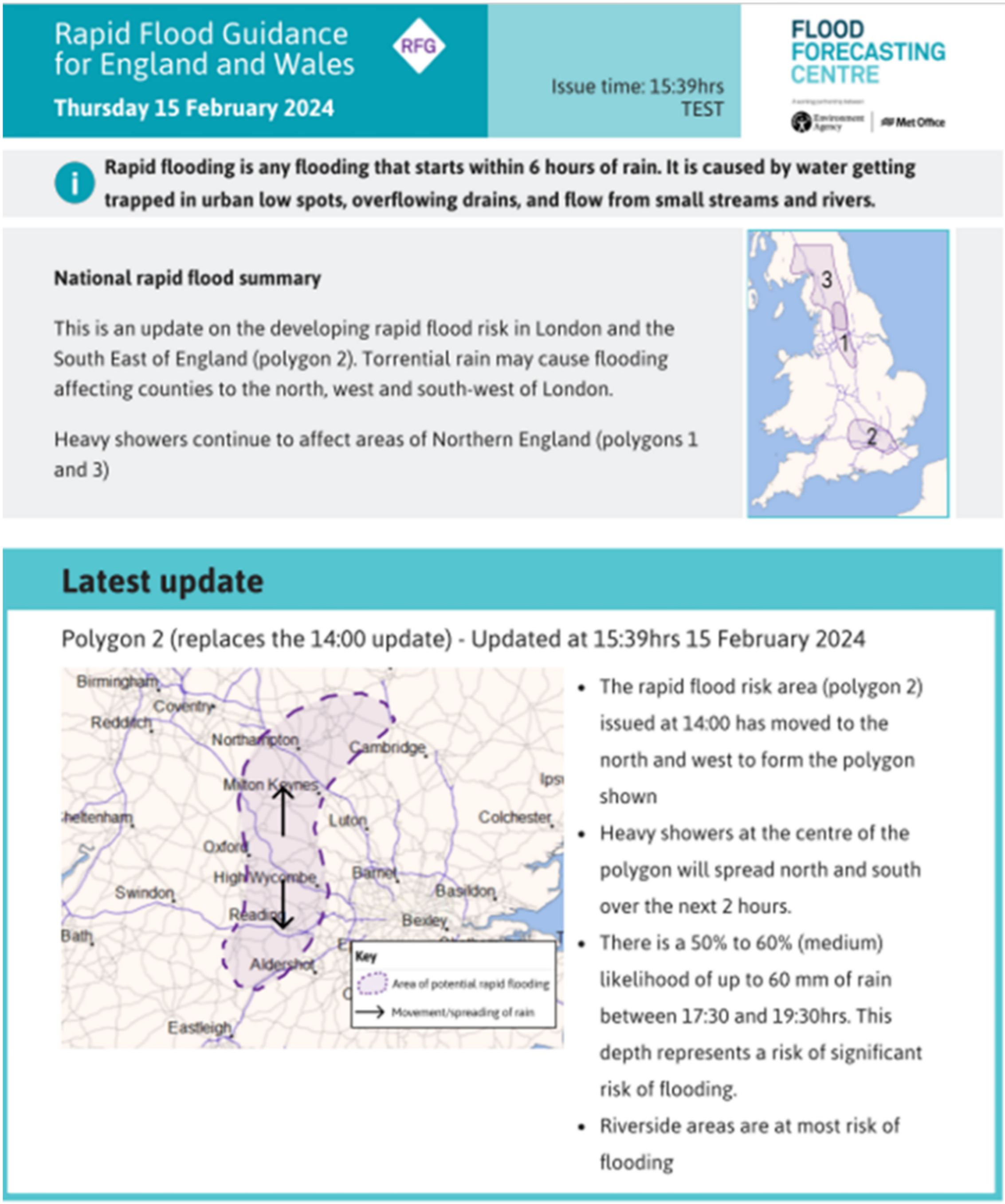


FIGURE 5 | An example of a Rapid Flood Guidance issued at 15:39 UTC on 15th.

with the regions where minor and significant observed impacts were reported. N-FOREWARNS, on the other hand, produced minor and minimal impact prediction over northern Wales, missing the majority of impacts further south. In contrast to both SWFHIM and N-FOREWARNS, the RFG warning was much larger and widespread across Wales, covering both affected and unaffected regions.

Event 3 is focused in the southeast of England (circled in red), with observed impacts reported over London and Sussex. All three SWF tools produced predictions over the southeast, with N-FOREWARNS and SWFHIM being more tightly constrained to London. Again, the RFG warning polygon was much larger and more widespread, containing regions where no rainfall or impacts are observed. Isolated, low likelihood predictions in Cumbria and Lincolnshire exemplify the low precision nature of SWFHIM, whereas N-FOREWARNS is larger but more focused over the southeast region.

One of the most impactful flooding events of the testbed was Event 4 on 1 August, which caused significant impacts to businesses, buildings and transport networks in Winchester, Hampshire. 50–60 mm of rainfall accumulated between 15Z and 18Z, with minor impacts also observed in Berkshire and Surrey. N-FOREWARNS produced a widespread significant impact prediction over the region, while SWFHIM produced minor

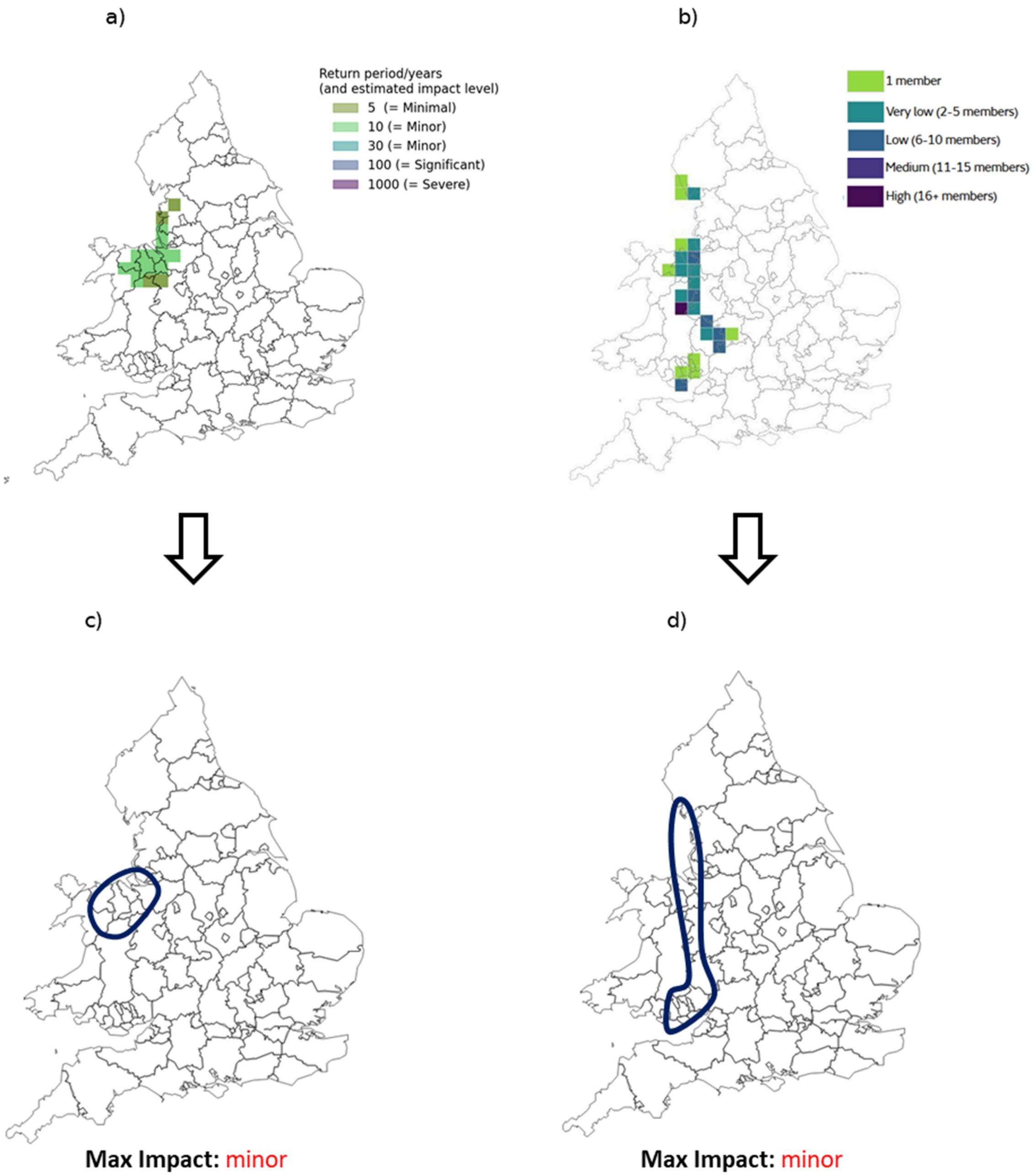


FIGURE 6 | Using the N-FOREWARNS and SWFHIM SWF nowcasts (a and b; only minor impact level shown here for SWFHIM example), participants drew SWF warning polygons (c and d) and associated them with a maximum impact. This example is taken from an event on 12 May 2024. The SWF nowcasts have a lead time of 3 h and are valid at 1800 UTC.

impact predictions that were more focussed over the surrounding regions of Hampshire. The RFG warning, on the other hand, produced a polygon over Sussex, which did capture some isolated rainfall there, but missed the bulk of the convective rainfall and impacts further west.

During Event 5, localised convective rainfall over northeastern Hampshire and the Isle of Wight produced SWF events that were categorised as minor (circled in red). This is a typical example of N-FOREWARNS under-representing the flood risk for a minor observed event when the rainfall is particularly isolated. SWFHIM approximately identifies the correct area, producing a minor prediction just to the east of Hampshire. However, it also produces a minor prediction over central Wales, where no rainfall is observed and no SWF events are reported. The RFG polygon is more constrained to Hampshire, capturing the rainfall and impacts well.

These examples capture the key behaviours of each SWF prediction tool. N-FOREWARNS generally produces fewer predictions of SWF than SWFHIM and RFG (e.g., Events 1, 2 and 5), which for low rainfall accumulation events (e.g., Event 1) enables it to correctly suppress the prediction severity.

However, for higher, localised rainfall amounts that cause minor events (e.g., Event 5), it can fail to capture potential SWF events. When N-FOREWARNS does identify high likelihood of a SWF event (during higher rainfall accumulation events, e.g.,

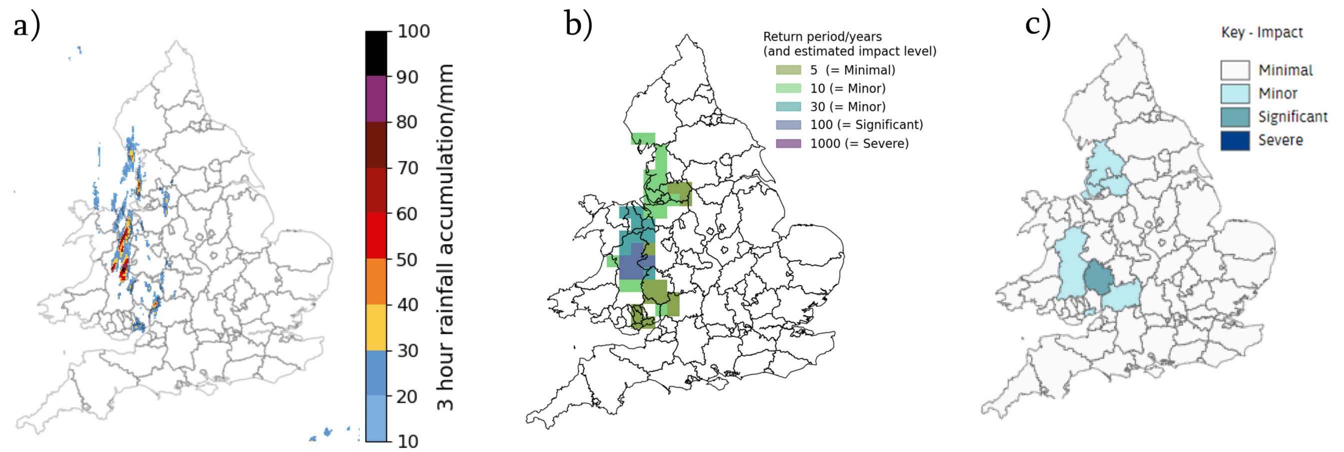


FIGURE 7 | An example (valid at 1800 UTC on 12 May 2025) of the observation sets that were revealed to the participant groups after they completed the polygon drawing exercise (Step 2). They provide the (a) 3-h radar rainfall accumulation taken from the 3 h before the valid time, (b) corresponding R-FOREWARNS output at the valid time and (b) the reported SWF impacts collected by the hyrdrometeorologists at the FFC.

Events 3 and 4), it is often in the correct location. SWFHIM on the other hand is more sensitive to the rainfall input and will produce a SWF prediction for the majority of events (e.g., Events 1–5), capturing minor impact events more often than N-FOREWARNS (e.g., Event 1). However, due to this sensitivity, SWFHIM's output has low spatial precision that can make it difficult to interpret (e.g., Events 1–4). Furthermore, this can often make SWFHIM predictions appear too spatially widespread (e.g., Event 1). In contrast, the RFG warning, which is forecaster-drawn, is always a coherent polygon and mostly captures the events (e.g., Events 2–4), but is often too widespread, meaning unaffected regions are included in the flood warning. It is important to note that RFG warnings were produced by the particular forecaster that was on shift for each event. Therefore, due to forecaster subjectivity, this results in a range of RFG polygon shapes/sizes, as highlighted in the RFG column of Figure 5 (also shown in Figure S1).

3.1.2 | Testbed Survey Responses

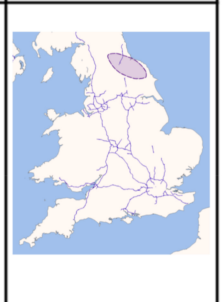
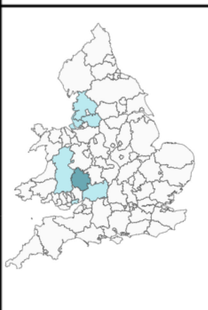
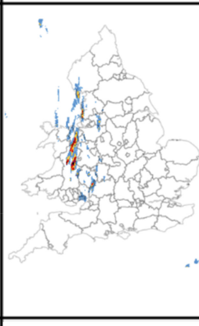
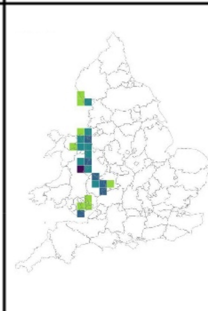
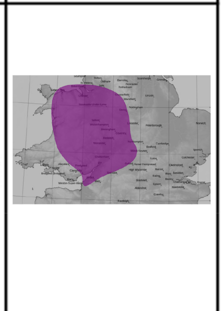
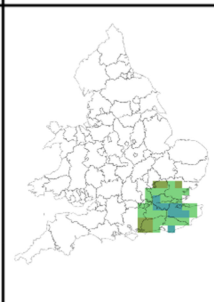
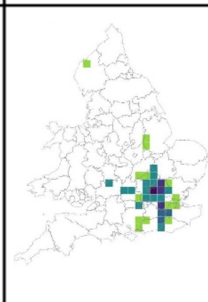
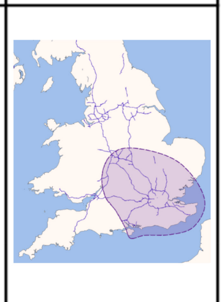
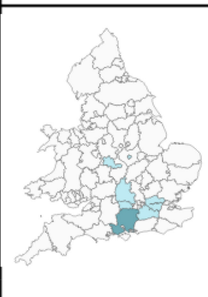
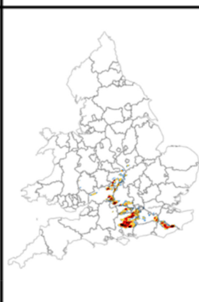
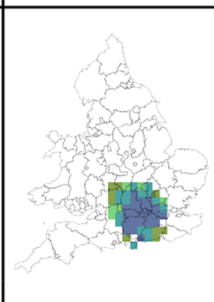
Figures 8–10 show the results from three survey questions that participant groups were asked during the testbed for all 20 events (92 subjective assessments in total). These questions aimed to cover participant opinions on three key areas of the SWF nowcast tools: location, severity and overall performance assessment.

In terms of impact severity, the most common participant response was that the participant-drawn SWF polygons (see Step 2 of Section 2.3.2) of N-FOREWARNS and SWFHIM were 'about right' (47% and 40% respectively, Figure 8a,b). However, for both SWFHIM and RFG, a large proportion (50% and 65%, respectively) believed that the polygons had impact severities that were 'Slightly too high' or 'Much too high' (Figure 8b,c). Across the testbed SWFHIM produced low likelihoods (1–5 members) of significant impacts for the majority of events, when only 5 observed significant events were reported (Figure S1). For N-FOREWARNS the second highest impact severity category was 'Miss', with only 23% of participants thinking that it was too high or too low. This suggests that, although N-FOREWARNS often does not produce enough predictions, when it does, they are likely to have the correct severity. Both N-FOREWARNS and SWFHIM show a wide distribution of severity scores, reflecting a range in participant opinions on their performance (Figure 8d). The RFG, however, produces a much narrower distribution with a mean score of 2, that is, its tendency to over-predict on severity levels is more consistent than N-FOREWARNS and SWFHIM.

When considering participants' opinions on polygon location, N-FOREWARNS produced the highest proportion (34%) of 'Miss' category (Figure 9a). This is exemplified in Events 2 and 5 in Table 2, when the nowcast does not capture the spatial distribution of the impacts. However, only 7% of participants reported that the N-FOREWARNS polygon was in the 'Incorrect location', suggesting that when N-FOREWARNS does produce a prediction, it is often considered to be in the correct location (as also seen from 33% of participants voting for the 'About right category'). The reported participant opinions of SWFHIM (Figure 9b), on the other hand, oppose that of N-FOREWARNS, with far fewer participants voting SWFHIM's location as a 'Miss'. The largest proportion of SWFHIM location votes was 'About right', however, 26% agreed that its polygon was in an 'Incorrect location'. As demonstrated in Table 2, SWFHIM will produce predictions in approximately the right location, resulting in a hit; however, predictions often have low spatial precision, meaning flood predictions are given to regions that are unaffected, resulting in false alarms. Similar to SWFHIM, a large proportion of participants (29%) reported that the RFG polygon locations were 'About right' (Figure 9c). However, the strongest participant response was that the RFG polygons were too large ('Slightly too large' and 'Much too large' categories combined to 47%). This can clearly be seen throughout Table 2 where the RFG often produces large, widespread polygons (Events 2, 3 and 5 in Table 2). The distribution of scores for location (Figure 9c) show a large spread of participant opinions for polygon location for N-FOREWARNS, SWFHIM and RFG. Again, this reflects the ranging opinions of participants and the tools' inconsistent ability to predict the location across a range of event types.

Figure 10 shows the participants' responses when asked to give an overall assessment of N-FOREWARNS and SWFHIM in

TABLE 2 | Information on the five selected events from the testbed. For each event (row), the reported SWF impacts, 3-h radar accumulation, N-FOREWARNS warning, SWFHIM minor impact warning and RFG output are shown. For each event, the SWF prediction tools were initialised with a lead time to match the RFG (ranging from 1 to 3 h lead time). For days when multiple observed impacts were reported, the red circles identify those that occurred during the valid time.

	Observed Impact Minimal Minor Significant Severe	3 hour radar accumulations (>10mm) 10 20 30 40 50 60 70 80 90 100 3 hour rainfall accumulation/mm	N-FOREWARNS Return period/years 5 (= Minimal) 10 (= Minor) 30 (= Minor) 100 (= Significant) 1000 (= Severe)	SWFHIM (minor) 1 member Very low (2-5 members) Low (6-10 members) Medium (11-15 members) High (16+ members)	RFG (warning area enclosed by polygon)
Event 1 15th Jun 2025 Initialised: 1300Z Valid until: 1600Z (+3 hour lead time)					
Event 2 12th May 2025 Initialised: 1500Z Valid until: 1800Z (+3 hour lead time)					
Event 3 15th Jul 2025 Initialised: 1800Z Valid until: 2000Z (+2 hour lead time)					
Event 4 1st Aug 2025 Initialised: 1600Z Valid until: 1800Z (+2 hour lead time)					
Event 5 8th Sep 2025 Initialised: 0900Z Valid until: 1100Z (+2 hour lead time)					

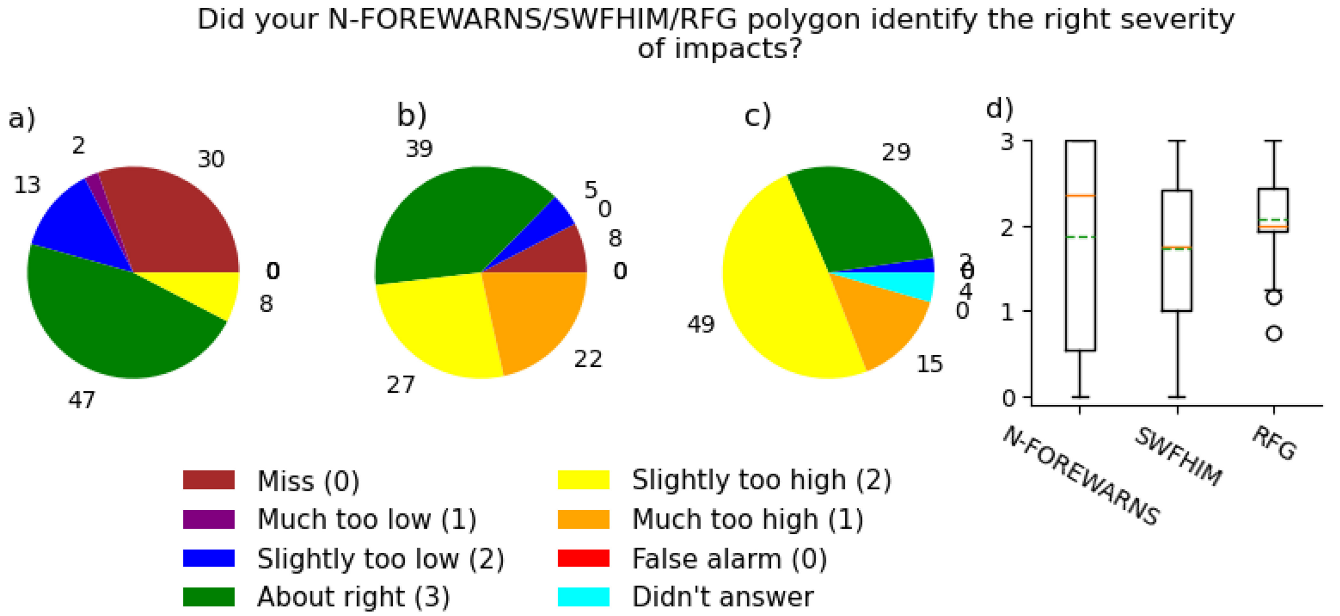


FIGURE 8 | Across all events, the percentage of participants who voted for each severity level evaluation category for (a) N-FOREWARNS (b) SWFHIM and (c) RFG. Each category has an associated score (shown in brackets), with a higher score representing a better performance. (d) Shows the distribution of these scores, with median (orange) and mean (green) scores, for each SWF prediction tool across all events.

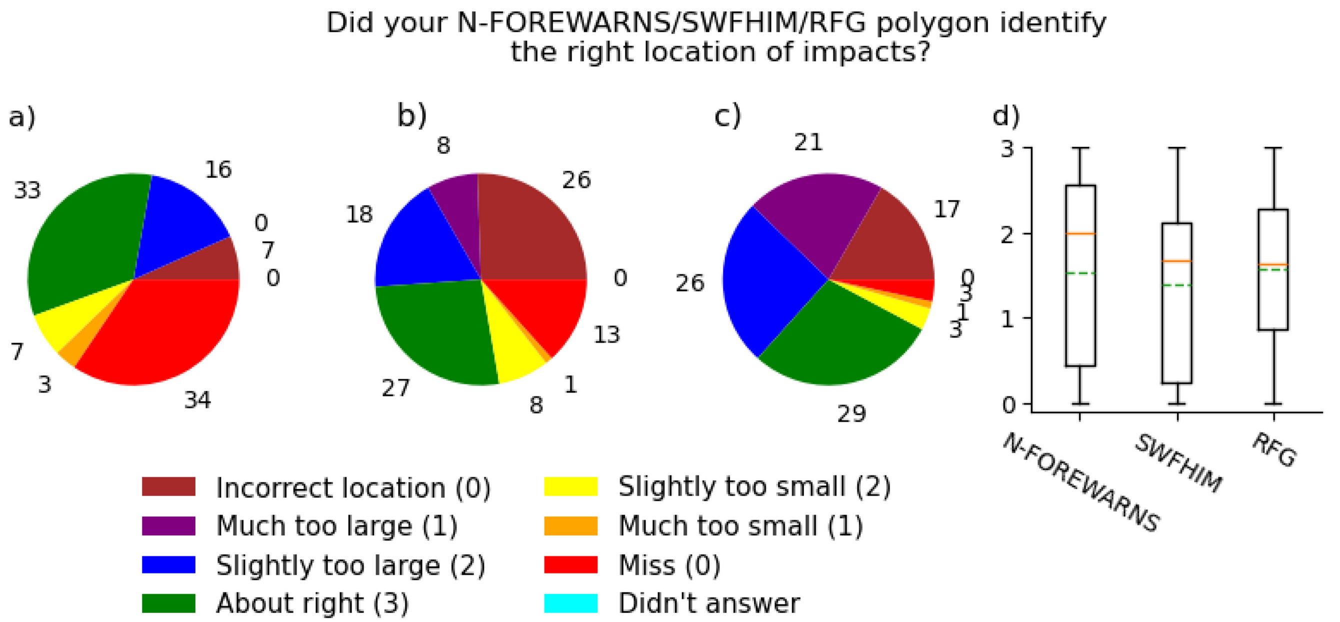


FIGURE 9 | Same as Figure 8 but for location evaluation categories.

terms of a hit, miss, false alarm or correct rejection. SWFHIM produces a greater proportion of votes for a hit and a false alarm, whereas N-FOREWARNS produces a greater proportion of votes for correct rejection and miss—a reflection of the tools behaving in opposing ways. SWFHIM produces a greater number of SWF predictions, some of which are correct—generating a higher hit score—but at the same time generating more false alarms. N-FOREWARNS on the other hand produces far less SWF predictions, resulting in a greater number of correct rejections, but also producing more misses.

3.2 | Quantitative N-FOREWARNS Analysis

The requirement to align the SWF nowcasts with the corresponding RFG warnings (Section 2.3.2) mean that, for 14 of the 20 testbed events, N-FOREWARNS (and SWFHIM) were assessed with lead times of 3–4 h. However, STEPS skill has been shown to reduce rapidly over the first 3 h, after which there is a transition from radar to NWP and skill flattens with lead time (Bowler et al. 2006). Table 3 shows the N-FOREWARNS output for the same five events in Table 2 but using a 1 h lead time

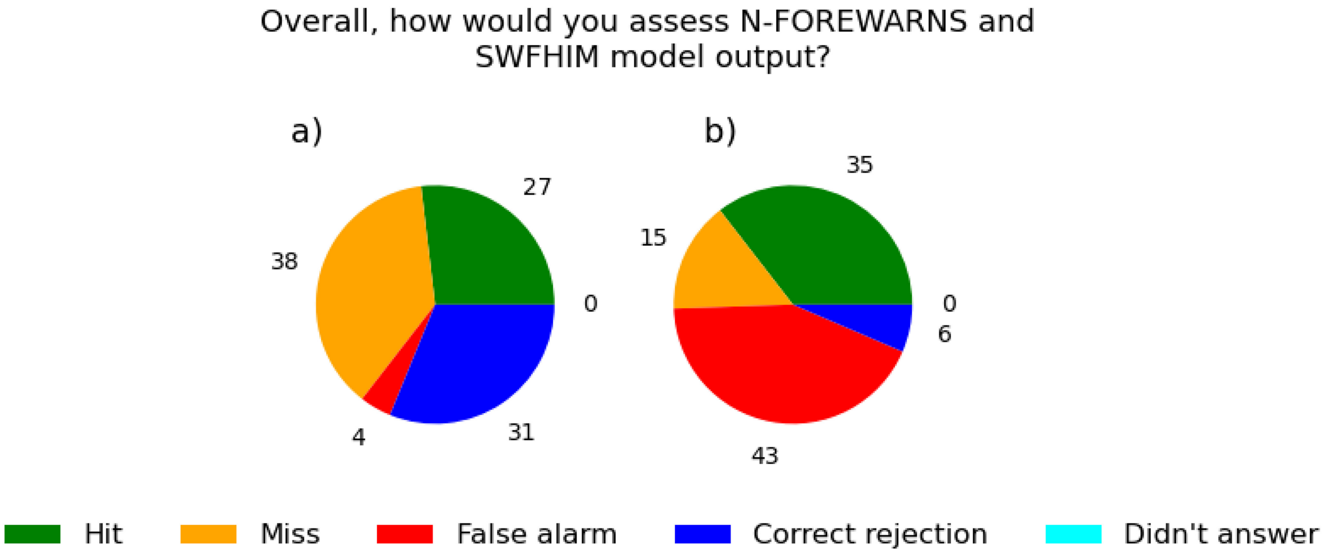


FIGURE 10 | Across all events, the percentage of participants who voted for each overall evaluation category for (a) N-FOREWARNS and (b) SWFHIM.

nowcast instead. For these cases, by driving it with a shorter lead time rainfall nowcast from STEPS (i.e., one with more skill), N-FOREWARNS is better able to capture the flood impact risk. Therefore, in order to have a better understanding of the performance of N-FOREWARNS, it is necessary to evaluate it across multiple lead times.

Figure 11a shows the distribution of H, 1-F and SR spatial scores for N-FOREWARNS over 1–6 h′ lead time, for the 70 most significant SWF events recorded in Northern England from social media reports between 2013 and 2022 (taken from Maybee et al. 2024), verified against R-FOREWARNS. Higher values denote higher skill for all three scores in Figure 11a. The likelihood of detecting a radar-driven event (i.e., mean H score) decreases from 0.37 to 0.1 across 1–3 h (as the skill of the STEPS nowcast reduces). The majority of scores then drop to 0 from 4 h onwards. Across all lead times, the likelihood of falsely predicting an event is low ($1-F$ score remains above 0.95), however, this result is to be expected due to the low H scores (i.e., N-FOREWARNS produces a high proportion of no-SWF event predictions). When N-FOREWARNS produces a prediction, the likelihood that it is a hit (i.e., mean SR score) steadily reduces from 0.41 to 0.05 across 1–5 h. The 0.05 increase in the mean SR score from 5 to 6 h lead time may be a reflection of the increasing skill in the NWP model as it spins up (Clark et al. 2016).

To gain insight into N-FOREWARNS's performance over a full summer season, rather than evaluating it solely at the time of known flood events as in Figure 11a, N-FOREWARNS was run every hour between 1 May 2024 and 30 September 2024 and verified against R-FOREWARNS. Figure 11b shows the distribution of the skill scores for 3740 issuances for lead times of 1–6 h. Overall, the results are similar to those for the 70 known flood events in Figure 11a. The mean H score decreases with lead time from 0.35 to 0.1, while the mean 1-F scores remain 1 across all lead times (93% of R-FOREWARNS issuances contained no events). The SR score decreases with lead time from 0.36 to 0.05 but, due to a slightly greater number of false alarms, the mean SR values are lower at each lead time compared to Figure 11a (by a difference of 0.05 for 1 h′ lead time). Testing on a dataset with a lower proportion of observed significant events means that false alarms are more likely to occur (compared to a dataset containing significant events only). Overall, the comparable results of Figure 11a,b shows that, even when it applied to a majority of cases with no SWF events, N-FOREWARNS still maintains its level of capability in capturing the rarer SWF events and does not have a high false alarm rate.

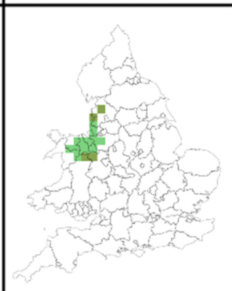
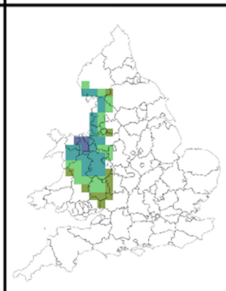
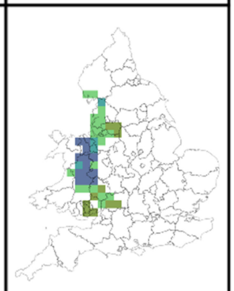
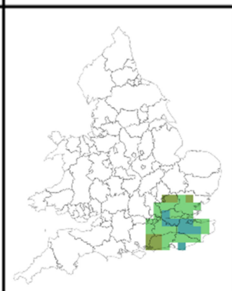
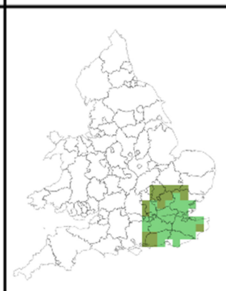
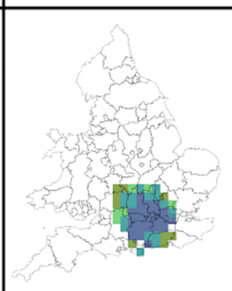
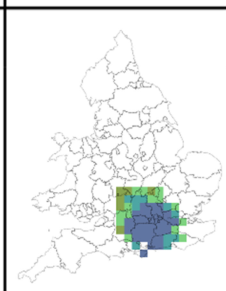
To understand the skill of N-FOREWARNS when input with more accurate rainfall fields, a comparison of the R-FOREWARNS output (Figure S3) with the corresponding reported SWF impact observation for every event in the testbed shows that the SWF prediction in R-FOREWARNS covers the location for approximately 75% of the reported events. Although this result is derived from testing over only 20 events, it suggests that, when accurate rainfall data is used (i.e., radar data), the N-FOREWARNS methodology can capture the majority of SWF events. Furthermore, it suggests that the main uncertainties come from the rainfall data that is input into N-FOREWARNS, rather than it is processing into a SWF prediction.

4 | Discussion

The results from this study provide both a qualitative evaluation of N-FOREWARNS and SWFHIM (an existing SWF prediction tool) by flood professionals during the summer 2024 testbed and a quantitative evaluation of N-FOREWARNS over a longer historical period.

The quantitative analysis across 70 historical, significant, observed flood events showed that N-FOREWARNS's ability to capture SWF events gradually decreases when moving from a 1 to 3 h lead time. There is a rapid drop in performance at lead times longer than 3 h, indicating the limit of predictability

TABLE 3 | Comparison of N-FOREWANRNS maps issued during the testbed (with 2–3 h lead times, see Table 2) against N-FOREWARNS with a 1 h lead time and R-FOREWARNS, for each event (rows).

	N-FOREWARNS issued in testbed	N-FOREWARNS (+1 hour lead time)	R-FOREWARNS
Event 1			
Event 2			
Event 3			
Event 4			
Event 5			

Return period/years

5 (= Minimal) 30 (= Minor) 1000 (= Severe)
10 (= Minor) 100 (= Significant)

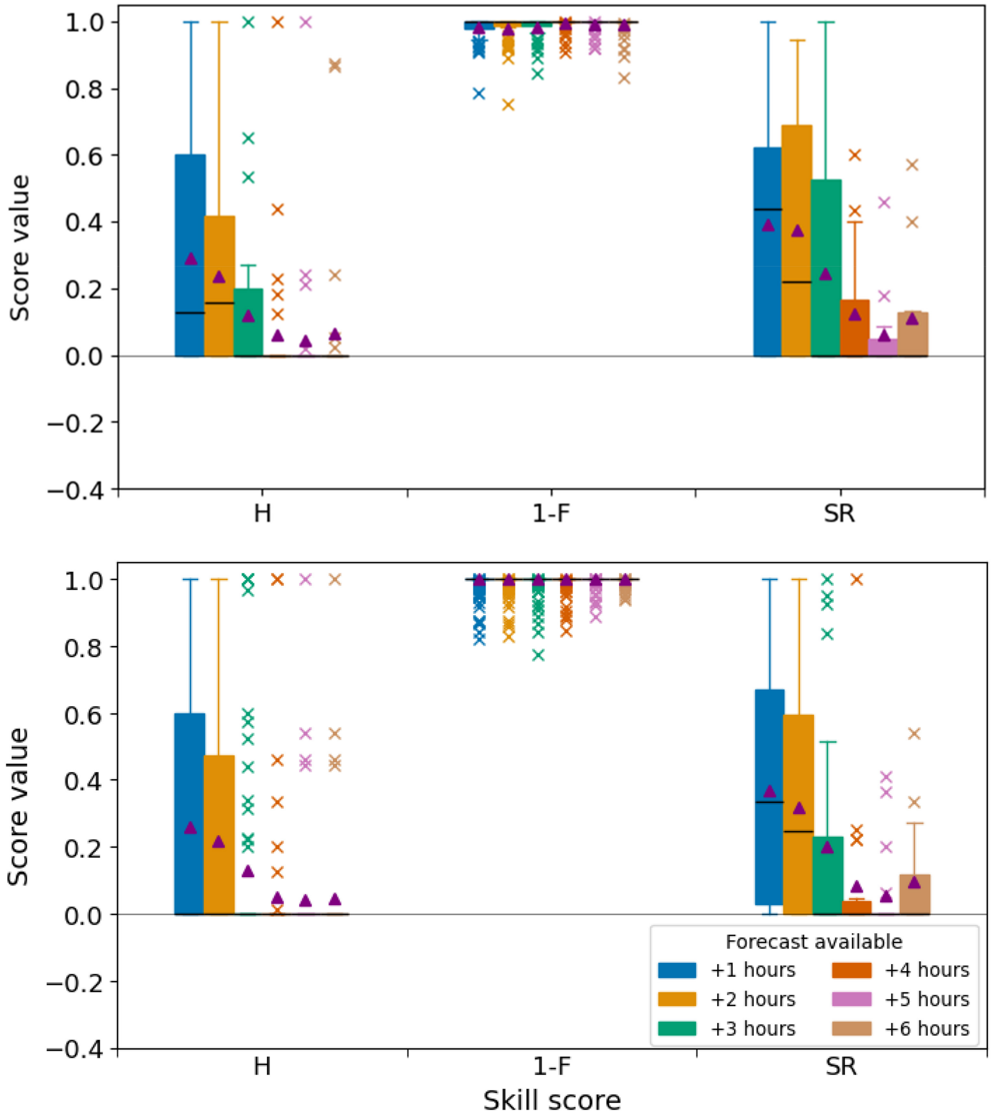


FIGURE 11 | H, 1-F and SR score distributions for (a) 70 flooding events between 2013 and 2022 (to match Maybee et al. 2024) and (b) 3470 issuances across the continuous time period of summer 2024, produced by N-FOREWARNS for 1–6 h lead times. Mean scores are represented by purple triangles, while the median scores are represented by horizontal black lines.

using N-FOREWARNS. The significant drop off is due to the transition from radar-extrapolated rainfall data to NWP rainfall data within the STEPS algorithm. At these short lead times the NWP input has poor skill on the high spatiotemporal resolution required (i.e., to capture rapidly developing storms) for input into N-FOREWARNS (Clark et al. 2016). This demonstrates the sensitivity of N-FOREWARNS to its input—a skilful N-FOREWARNS prediction is reliant on accurate rainfall data.

When evaluated over a continuous period between May–September 2024 (3740 predictions for each lead time), N-FOREWARNS produced a similar set of results, with higher H scores (chance of capturing observed SWF events) but reduced SR scores (chance of an observed SWF event, given that a SWF nowcast is produced) due to a greater number of false alarms. This performance consistency over every day in a summer season, including days when no flood events occurred, shows that N-FOREWARNS can capture the less frequent SWF events and maintain a low false alarm rate.

N-FOREWARNS is a modified version of FOREWARNS (Maybee et al. 2024), which uses rainfall inputs from ensemble NWP forecasts and produces predictions at lead times of 1–4 days. A direct comparison of the skill in FOREWARNS and N-FOREWARNS is not possible, but it is still useful to qualitatively do so for context. For the 70 recorded flood events used in this study, N-FOREWARNS (at a 1 h lead time) produces lower spatial H scores (by 0.1) but higher 1-F scores (F represents the proportion of observed no-SWF events that were nowcasted as a SWF event) (by 0.20) and higher SR scores (by 0.10) than the FOREWARNS predictions at 1 day lead time (evaluated in Maybee et al. 2024). This suggests that N-FOREWARNS produces a lower proportion of false alarms, but at the compromise of fewer hits. When comparing across continuous periods (Maybee et al. 2024 evaluated over a total of 725 days in May–October 2019–2022), N-FOREWARNS produces approximately the same H score but higher SR scores (by 0.20), which again can be attributed to a lower number of false alarms. Given that N-FOREWARNS is predicting with a shorter lead time than FOREWARNS (hours compared to days), it may be expected to produce better performance scores. However, the two methods have been set different predictive expectations—N-FOREWARNS is required to predict at a specific valid time (meaning that it needs to get the hour of the day correct), whereas FOREWARNS is predicting events that occur at any time within a 24-h period. Having to overcome these inherently different challenges has resulted in comparable scores for the two SWF prediction tools. Nevertheless, both tools show less than 50% chance of capturing their corresponding radar-driven SWF risks (*H* scores), highlighting that, despite the short lead times of N-FOREWARNS, skill scores for SWF prediction are still low.

Overall, the quantitative analysis in this study shows that significant improvements in the ability of NWP and nowcasting to predict the location of convective rainfall are required to improve SWF prediction.

A testbed hosted by the FFC and Met Office during summer 2024 focussed on nowcasting SWF capabilities with the aim to (1) compare the performance of N-FOREWARNS with SWFHIM (FFC's current SWF model) and (2) assess whether N-FOREWARNS and SWFHIM could add value to the RFG production process, which were being trialled operationally over summer 2024. Across 20 SWF events during summer 2024, N-FOREWARNS produced fewer flood predictions compared to SWFHIM and missed a larger proportion of the observed SWF events, particularly at the minor impact severity level. However, when N-FOREWARNS did produce a SWF prediction, it was generally accurate, with a low number of false alarms reported. By producing more predictions, SWFHIM achieved a greater number of hits than N-FOREWARNS; however, output from SWFHIM often has low spatial precision. This results in the SWF nowcast producing predictions in the incorrect location, triggering a larger number of false alarms. Furthermore, SWFHIM produces impact severity levels that are too high compared to the reported SWF observations that were available. The testbed also showed that the RFG is generally good at identifying the region of SWF impacts and the accompanying detail on rainfall amounts and impact likelihood provides useful information for end users. However, the RFG warnings tend to be too widespread and overestimate the level of impact severity. Furthermore, forecaster subjectivity provides a source of sensitivity for RFG polygon shapes, highlighting a key limitation of the warning process. The conclusion for aim 2, therefore, is that both N-FOREWARNS and SWFHIM can add value to the SWF forecasting process by providing operational forecasters with improved spatial accuracy and help distinguishing between impact severity levels.

Testbed participant responses on location and impact severity levels of SWFHIM and N-FOREWARNS produced a wide distribution of performances. One reason is that each tool showed better/worse performance under certain conditions, that is, N-FOREWARNS performed better during high rainfall accumulation events and SWFHIM performed better during minor impact rainfall events. Another reason is the differing opinions on what participants consider to be important characteristics of a SWF prediction. Some participants considered a high number of false alarms to be an acceptable compromise in order to get a hit, whereas others prioritised reducing the false alarms to achieve better reliability. These opposing opinions highlight the challenge of definitively assessing SWF nowcasts and demonstrates the importance of involving experts from different technical backgrounds to provide a well-rounded evaluation. Further testing with decision-makers (rather than technical experts) would provide more clarification on the key priorities.

When considering the operational utility of N-FOREWARNS, the testbed results showed that it should be used alongside the current set of FFC tools. N-FOREWARNS's single SWF risk map provides a forecaster with an easily interpretable set of information to support quick decision-making. As STEPS is run operationally at the Met Office, N-FOREWARNS can be run in real time by the FFC, with its risk maps presented alongside SWFHIM (and other operational SWF tools). When using N-FOREWARNS alongside SWFHIM, it is important for a forecaster to understand the behaviours that were reported during the testbed. For example, knowing that it performs better during high rainfall events, or performs poorer during lower rainfall events, will help a forecaster give weight to the SWF tools during different rainfall scenarios.

By predicting flood return periods, N-FOREWARNS produces a single, simple SWF risk map; however, this simplicity may not fulfil some user-needs that require specific details of impact severity. Furthermore, the quantitative analysis performed in this study provides an understanding of how N-FOREWARNS performs against a corresponding radar-driven version, that assumes a ‘perfect’ rainfall field. It is therefore an assessment of the rainfall fields, not the SWF impacts themselves. Ideally, N-FOREWARNS output would be verified against a reliable dataset of SWF observations. However, deficiencies in collating SWF observations (e.g., from unreliable and patchy media sources) mean that a robust evaluation is not currently possible. To overcome this, future work should focus on the improvement of SWF impact identification methodologies. The use of satellite data has shown promise for this application (Notti et al. 2018), specifically combined with machine learning tools for flood mapping (Portalés-Julià et al. 2023). Additionally, future work should assess the use of other rainfall nowcasting products (e.g., MONOW; Moseley and Sandford 2019) to drive N-FOREWARNS, which may produce improvements in SWF prediction skill.

5 | Conclusions

Access to accurate SWF predictions at short lead times is a key requirement in the flood response community. However, the chaotic nature of convection and the requirement to capture non-meteorological factors makes SWF prediction challenging. This study aims to address these issues by providing an evaluation of tools used for SWF prediction on nowcasting timescales. This study presents N-FOREWARNS, a new tool that produces SWF nowcasts at lead times of 1–6 h using probabilistic rainfall predictions from the STEPS nowcast system. Through expert evaluation and quantitative verification, this study shows that:

- N-FOREWARNS demonstrates useful SWF predictive skill at lead times of up to 3 h, beyond which skill declines sharply.
- Across both a set of significant historical SWF events and the continuous summer 2024 evaluation, N-FOREWARNS shows consistent behaviour, characterised by relatively low false-alarm rates and good spatial accuracy when predictions are issued.
- Comparison with FOREWARNS indicates that despite shorter lead times, N-FOREWARNS achieves comparable skill, though both systems capture fewer than half of radar-driven SWF predictions (i.e., our best indication of reality)—highlighting the inherent challenge of SWF prediction.
- Testbed evaluations show that N-FOREWARNS and SWFHIM each offer complementary strengths: N-FOREWARNS provides spatial precision but often misses minor events, while SWFHIM offers greater sensitivity but at the cost of more false alarms.
- Operationally, N-FOREWARNS can add value when used alongside existing FFC tools, providing forecasters with an easily interpretable short-lead-time risk map that aids rapid decision-making.
- N-FOREWARNS' skill is reliant on accurate rainfall input. Significant improvements in the ability of NWP and nowcasting to predict the location of convective rainfall are required to improve SWF prediction.
- Improving the quality, consistency and availability of SWF observations will also be essential, enabling more robust verification of model skill and more reliable development of future SWF prediction tools.

Author Contributions

Joseph Smith: writing – original draft, methodology, writing – review and editing, formal analysis, conceptualization. **Cathryn Birch:** conceptualization, writing – review and editing, supervision, methodology. **Julia Perez:** conceptualization, methodology, writing – review and editing. **Linda Speight:** conceptualization, writing – review and editing, methodology. **John Mooney:** writing – review and editing, methodology. **Ben Maybee:** writing – review and editing. **John Marsham:** writing – review and editing, supervision.

Acknowledgements

We are extremely grateful to all participants of the Met Office Summer 2024 nowcasting testbed for their invaluable contributions to this study. We would specifically like to thank Rosie Nation, Katie Norman, Matt Clark and Dan Suri for their work in organising the testbed. We have benefited enormously from wider collaborative discussions with the Met Office and Flood Forecasting Centre. This work predominantly used JASMIN and the Met Office MASS data archive and we would like to thank Richard Rigby of CEMAC, University of Leeds for his help with writing code.

Funding

This study was conducted as part of the Yorkshire Integrated Catchment Solutions Programme (iCASP, Natural Environment Research Council grant NE/P011160/1). The work in writing and performing the analysis for this study was funded by the NERC SENSE CDT (NE/T00939X/1).

Data Availability Statement

The code for FOREWARNS is publicly available at https://zenodo.org/records/10987887. The modification code for N-FOREWARNS is available upon request. The UK Nimrod radar data that was used in this study is available through the CEDA archive (https://archive.ceda.ac.uk/). The STEPS nowcast data was extracted from the Met Office's MASS archive system, which is not publicly available. Sharing the testbed survey data is not possible as consent was not given by the participants.

References

Acosta-Coll, M., F. Ballester-Merelo, M. Martinez-Peiró, and E. De La Hoz-Franco. 2018. "Real-Time Early Warning System Design for Pluvial Flash Floods—A Review." *Sensors* 18: 2255. https://doi.org/10.3390/s18072255.

Aldridge, T., O. Gunawan, R. J. Moore, S. J. Cole, G. Boyce, and R. Cowling. 2020. "Developing an Impact Library for Forecasting Surface Water Flood Risk." *Journal of Flood Risk Management* 13: e12641. https://doi.org/10.1111/jfr3.12641.

Ayzel, G., T. Scheffer, and M. Heistermann. 2020. "RainNet v1.0: A Convolutional Neural Network for Radar-Based Precipitation Nowcasting." *Geoscientific Model Development* 13: 2631–2644. https://doi.org/10.5194/gmd-13-2631-2020.

Baker, L. 2017. *A Summer of Floods!, Flooding From Intense Rainfall*. https://blogs.reading.ac.uk/flooding/2017/10/11/a-summer-of-floods.

BBC News. 2025. "Cornwall Floods: Clear-Up Follows Flood 'Devastation'." Accessed March 4, 2025. https://www.bbc.co.uk/news/uk-england-cornwall-40652594.

Birch, C. E., B. L. Rabb, S. J. Böing, et al. 2021. "Enhanced Surface Water Flood Forecasts: User-Led Development and Testing." *Journal of Flood Risk Management* 14: e12691. https://doi.org/10.1111/jfr3.12691.

Böing, S. J., C. E. Birch, B. L. Rabb, and K. L. Shelton. 2020. "A Percentile-Based Approach to Rainfall Scenario Construction for Surface-Water Flood Forecasts." *Meteorological Applications* 27: e1963. https://doi.org/10.1002/met.1963.

Bowler, N. E., C. E. Pierce, and A. W. Seed. 2006. "STEPS: A Probabilistic Precipitation Forecasting Scheme Which Merges an Extrapolation Nowcast With Downscaled NWP." *Quarterly Journal of the Royal Meteorological Society* 132: 2127–2155. https://doi.org/10.1256/qj.04.100.

Carlyon, H., J. Hitching, A. McNeill, and T. Dixon. 2012. *Northumberland County Council Sustainable Transport FCERM County Hall Morpeth NE61 2EF*. Northumberland County Council.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez. 2016. "Convection-Permitting Models: A Step-Change in Rainfall Forecasting: A Step-Change in Rainfall Forecasting." *Meteorological Applications* 23: 165–181. https://doi.org/10.1002/met.1538.

Cole, S. J., and R. J. Moore. 2009. "Distributed Hydrological Modelling Using Weather Radar in Gauged and Ungauged Basins." *Advances in Water Resources* 32: 1107–1120. https://doi.org/10.1016/j.advwatres.2009.01.006.

De Bruijn, J. A., H. De Moel, B. Jongman, M. C. De Ruiter, J. Wagemaker, and J. C. J. H. Aerts. 2019. "A Global Database of Historic and Real-Time Flood Events Based on Social Media." *Scientific Data* 6: 311. https://doi.org/10.1038/s41597-019-0326-9.

Dottori, F., M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen. 2017. "An Operational Procedure for Rapid Flood Risk Assessment in Europe." *Natural Hazards and Earth System Sciences* 17: 1111–1126. https://doi.org/10.5194/nhess-17-1111-2017.

Environment Agency. 2025. "Guidance Risk of Flooding From Surface Water—Understanding and Using the Map." Accessed May 3, 2015. https://www.gov.uk/government/publications/flood-risk-maps-for-surfa ce-water-how-to-use-the-map/risk-of-flooding-from-surface-water-understanding-and-using-the-map.

Falconer, R. H., D. Cobby, P. Smyth, G. Astle, J. Dent, and B. Golding. 2009. “Pluvial Flooding: New Approaches in Flood Warning, Mapping and Risk Management.” *Journal of Flood Risk Management* 2: 198–208. https://doi.org/10.1111/j.1753-318X.2009.01034.x.

Flack, D. L. A., C. J. Skinner, L. Hawkness-Smith, et al. 2019. “Recommendations for Improving Integration in National End-To-End Flood Forecasting Systems: An Overview of the FFIR (Flooding From Intense Rainfall) Programme.” *Water* 11: 725. https://doi.org/10.3390/w11040725.

Flood Forecasting Centre. 2024. “Guidance Rapid Flood Guidance Service 2025: User Guide.” 16 April 2024. https://www.gov.uk/guidance/rapid-flood-guidance-service-trial-user-guide.

Flood Forecasting Centre. 2025. “Guidance Flood Guidance Statement User Guide.” Accessed May 3, 2015. https://www.gov.uk/government/publications/flood-guidance-statement-fgs-user-guide/flood-guidance-statement-user-guide.

Flood Tags. 2025. “Real Stories. Real Impact. Real-Time.” Accessed April 7, 2015. https://www.floodtags.com/.

Golding, B. W. 1998. “Nimrod: A System for Generating Automated Very Short Range Forecasts.” *Meteorological Applications* 5: 1–16. https://doi.org/10.1017/S1350482798000577.

Green, D., D. Yu, I. Pattison, et al. 2017. “City-Scale Accessibility of Emergency Responders Operating During Flood Events.” *Natural Hazards and Earth System Sciences* 17: 1–16. https://doi.org/10.5194/nhess-17-1-2017.

Harrison, D. L., K. Norman, C. Pierce, and N. Gaussiat. 2012. “Radar Products for Hydrological Applications in the UK.” *Proceedings of the Institution of Civil Engineers: Water Management* 165: 89–103. https://doi.org/10.1680/wama.2012.165.2.89.

Hill, P. G., T. H. M. Stein, A. J. Roberts, J. K. Fletcher, J. H. Marsham, and J. Groves. 2020. “How Skilful Are Nowcasting Satellite Applications Facility Products for Tropical Africa?” *Meteorological Applications* 27: e1966. https://doi.org/10.1002/met.1966.

Javelle, P., D. Organde, J. Demargne, et al. 2016. “Setting up a French National Flash Flood Warning System for Ungauged Catchments Based on the AIGA Method.” *E3S Web of Conferences* 7: 18010. https://doi.org/10.1051/e3sconf/20160718010.

Lean, H. W., P. A. Clark, M. Dixon, et al. 2008. “Characteristics of High-Resolution Versions of the Met Office Unified Model for Forecasting Convection Over the United Kingdom.” *Monthly Weather Review* 136: 3408–3424. https://doi.org/10.1175/2008MWR2332.1.

Maybee, B., C. E. Birch, S. J. Böing, et al. 2024. “FOREWARNS: Development and Multifaceted Verification of Enhanced Regional-Scale Surface Water Flood Forecasts.” *Natural Hazards and Earth System Sciences* 24: 1415–1436. https://doi.org/10.5194/nhess-24-1415-2024.

Moseley, S., and C. Sandford. 2019. “Blended Probabilistic Nowcasting With the IMPROVER Postprocessing System.” Met Office. https://repositorio.aemet.es/bitstream/20.500.11765/10622/1/NTSP8_Sandford_3ENC2019.pdf.

Notti, D., D. Giordan, F. Caló, A. Pepe, F. Zucca, and J. P. Galve. 2018. “Potential and Limitations of Open Satellite Data for Flood Mapping.” *Remote Sensing* 10: 1673. https://doi.org/10.3390/rs10111673.

Pilling, C. 2016. “New Developments at the Flood Forecasting Centre: Operations and Flood Risk Guidance.” *Urban Water* 165: 237–248. https://doi.org/10.2495/UW160211.

Porson, A. N., J. M. Carr, S. Hagelin, et al. 2020. “Recent Upgrades to the Met Office Convective-Scale Ensemble: An Hourly Time-Lagged 5-Day Ensemble.” *Quarterly Journal of the Royal Meteorological Society* 146: 3245–3265. https://doi.org/10.1002/qj.3844.

Portalés-Julià, E., G. Mateo-García, C. Purcell, and L. Gómez-Chova. 2023. “Global Flood Extent Segmentation in Optical Satellite Images.” *Scientific Reports* 13: 20316. https://doi.org/10.1038/s41598-023-47595-7.

Pulkkinen, S., D. Nerini, A. A. Pérez Hortal, et al. 2019. “Pysteps: An Open-Source Python Library for Probabilistic Precipitation Nowcasting (v1.0).” *Geoscientific Model Development* 12: 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019.

Ravuri, S., K. Lenc, M. Willson, et al. 2021. “Skilful Precipitation Nowcasting Using Deep Generative Models of Radar.” *Nature* 597: 672–677. https://doi.org/10.1038/s41586-021-03854-z.

Roberts, A. J., J. K. Fletcher, J. Groves, et al. 2022. “Nowcasting for Africa: Advances, Potential and Value.” *Weather* 77: 250–256. https://doi.org/10.1002/wea.3936.

Smith, J., C. Birch, J. Marsham, S. Peatman, M. Bollasina, and G. Pankiewicz. 2024. “Evaluating pySTEPS Optical Flow Algorithms for Convection Nowcasting Over the Maritime Continent Using Satellite Data.” *Natural Hazards and Earth System Sciences* 24: 567–582. https://doi.org/10.5194/nhess-24-567-2024.

Speight, L., C. Birch, K. Self, and S. Brown. 2025. “Expert Perspectives on the Next Generation of UK Surface Water Flood Warning Services.” *Weather* 81: 28–34.

Speight, L., S. J. Cole, R. J. Moore, et al. 2018. “Developing Surface Water Flood Forecasting Capabilities in Scotland: An Operational Pilot for the 2014 Commonwealth Games in Glasgow.” *Journal of Flood Risk Management* 11: S884–S901. https://doi.org/10.1111/jfr3.12281.

Speight, L. J., M. D. Cranston, C. J. White, and L. Kelly. 2021. “Operational and Emerging Capabilities for Surface Water Flood Forecasting.” *WIREs Water* 8: e1517. https://doi.org/10.1002/wat2.1517.

Vesuviano, G. 2022. *The FEH22 Rainfall Depth- Duration-Frequency (DDF) Model*. Science Centre Technical Report.

Yu, D., and T. J. Coulthard. 2015. “Evaluating the Importance of Catchment Hydrological Parameters for Urban Surface Water Flood Modelling Using a Simple Hydro-Inundation Model.” *Journal of Hydrology* 524: 385–400. https://doi.org/10.1016/j.jhydrol.2015.02.040.

Zhang, Y., M. Long, K. Chen, et al. 2023. “Skilful Nowcasting of Extreme Precipitation With NowcastNet.” *Nature* 619: 526–532. https://doi.org/10.1038/s41586-023-06184-4.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Same as Table 1 but for all areas evaluated across the testbed. The term ‘areas’ is used instead of ‘Event’ here because some events were evaluated multiple times during their validity period. **Figure S2:** The survey questions that were given to participant groups during the testbed. **Figure S3:** R-FOREWARNS output for each area evaluated in the testbed. Each ‘Area’ corresponds with the ‘Area’ in Figure S1.