The Met Office winter testbed 2020/2021: Experimenting with an on-demand 300-m ensemble in a real-time environment

C. L. Bain | M. Bush | J. C. Buxmann | P. Davies | A. Finnenkoetter | D. L. A. Flack | K. Hanley | R. W. Jones | H. W. Lean | H. Lewis | A. McCabe | J. C. Petch | A. Porson | C. Sanchez | E. K. Stone | D. Walters | S. Webster | S. Willington

Met Office, Exeter, UK

Correspondence
C. L. Bain, Met Office, Exeter, UK.
Email: caroline.bain@metoffice.gov.uk

Abstract
The Met Office held a testbed over winter 2020/2021 where a new numerical weather prediction (NWP) sub-km ensemble was set up on-demand in response to interesting weather phenomena in the United Kingdom. The domain for the model was chosen in real time by a community of Met Office Research Scientists and Operational Meteorologists and over a 4-month period the ensemble was triggered for nine events. The purpose of the testbed was to investigate whether a real-time weather regime-based enhancement in NWP capability was feasible, to understand what benefits a testbed environment might give, and to explore the practicalities of running such an event. Case studies from the testbed demonstrated that forecast ensembles at 2.2 km and 300 m grid spacing were able to capture observed winter weather, with greater spatial detail apparent, especially over complex orography, in the 300-m model. Ensemble spread appeared less influenced by resolution, potentially due to the size of the domains tested or the weather regimes of the case studies. The testbed also showcased underutilized observations and additional radiosonde ascents were conducted. All the testbed meetings were conducted virtually due to COVID-19 restrictions, and decisions were made about when to trigger the event using an online message board. The winter 2020/2021 testbed provides ideas for how on-demand weather-dependent testbeds might be conducted in the future. However, several recommendations are made that would enhance testbed benefits further, including more dedicated resource, stronger technology and data visualization and greater participation from both academia and weather information users.

Keywords
ensemble, NWP, sub-km, testbed, winter

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 Crown copyright. Meteorological Applications published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society. This article is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.
1 | INTRODUCTION

Testbeds have been long used as a practical way to bring Research Scientists and Operational Meteorologists (forecasters) together in an operational-like environment to test new scientific products, build communities and identify gaps in capability. Testbed activities are more commonly used in the United States, with an annual budget and dedicated staff from the National Oceanic and Atmospheric Administration (NOAA) or the National Centers for Environmental Prediction responsible for coordinating and supporting these activities (Ralph et al., 2013). Testbeds have also been a regular fixture in Europe (Groenemeijer, 2012). For several years the Met Office has taken part in the USA Hazardous Weather Testbed (HWT) Spring Experiment (Kain et al., 2017), coordinated by the Storm Prediction Center and NOAAs National Severe Storms Laboratory, which takes place every May and involves many US and international Scientists and Operational Meteorologists. Historically there have been fewer testbed-like events in the United Kingdom, the best examples being CSIP in 2004, 05 (Browning et al., 2006), which was mainly a field campaign and academic-based activity and the Met Office testbeds in Autumn 2013 (Semple, 2014) and Summer 2015.

During Met Office Forecast Experiments conducted in 2013 and 2015, Operational Meteorologists and Scientists were based in a room for a week with dedicated resource. Semple (2014) describe the experiment, which focused on United Kingdom forecasting with deterministic km-scale (UKV) and sub-km (London Model) systems. The report commented on how the intensive forecasting activities enhanced relationships between Scientists and Operational Meteorologists, though it was recommended that more time should be spent on verification in future testbeds. Similar commitment is experienced at the USA HWT testbed where weather discussions are held daily, visualizations of numerical weather prediction (NWP) model output are consistent across the whole period, and structured questionnaires are completed, with participants committing their time for up to 5 weeks (A. J. Clark et al., 2012; Gallo et al., 2017; Kain et al., 2003).

Testbeds are a safe and experimental environment for operational centres to test new capabilities in real time, with a significant number of people evaluating and discussing their use. Where new capability is deemed to offer value to operations, testbeds offer a pathway to accelerate their introduction to operations and adoption. In winter 2020/2021, the Met Office held a testbed-on-demand over a period of 3 months, December to February with a preliminary month (November) used to trial the experimental set-up. The main aims were to discover what testbed good practice might look like in an on-demand environment in the United Kingdom and experiment with some new capabilities in real time.

It is likely that with increases in computer power, greater use will be made of very high-resolution (sub-km) NWP models in future operations. Sub-km models have better representation of meteorological processes, such as convection, which is explicitly represented and better resolved than in models with km-scale grid spacing (Hanley et al., 2021; Stein et al., 2015; Woodhams et al., 2021). Features such as orographic enhancement, cold pools in valleys (Vosper et al., 2014), fog (Smith et al., 2021), gust fronts and downdrafts are better simulated as the grid scale becomes closer to the length scales of meteorological processes in the real world. Experiments with small regional models with approximately 300-m grid spacing have been ongoing over the last decade. For example, the non-operational London Model (Boutle et al., 2015), with 0.003 degree grid spacing in a rotated latitude–longitude projection, is assessed routinely by forecasters for Heathrow airport. The Met Office is running an ‘Urban-scale’ modelling project to develop the capability for models in the range of grid lengths 25–300 m. This work is expected to lead to more effective configurations of the sub-km models than those used in the testbed reported here.

One element that limits the effectiveness of sub-km models is the lack of fine-scale representative observations to constrain the downscaled analysis. This means that there is always an element of error in the forecast due to the uncertainty of the initial conditions. Ensemble models should provide some mitigation to this by accounting for the inherent uncertainty and providing a realistic representation of the possible scenarios, recognizing that no ensemble system is perfect and sometimes none of the possible simulated scenarios are close to reality. The highest resolution model available to Operational Meteorologists over the United Kingdom in 2020/2021 was the UKV model (Tang et al., 2013), which has a grid spacing of 1.5 km over land. The highest resolution ensemble was MOGREPS-UK (Porson et al., 2020), which has a grid spacing of 2.2 km. The winter 2020/2021 testbed was a first step into exploring the added benefits of sub-km ensemble systems in an operational (and case study only) context as well as testing the on-demand capability.

The downside of sub-km ensembles is the high cost of running them operationally. An 18-member ensemble of $500 \times 500$ point domain ($150 \times 150$ km) for the $300$ m ($0.003^\circ$) model is more computationally expensive than running a global operational model with latitude-longitude grid spacing of about 10 km in mid-latitudes. This is because a sub-km model needs a shorter timestep of 10 s, whereas a global model at 10 km has a timestep of about 5 min. Therefore, it was suggested that rather
than running the sub-km ensemble constantly over a fixed domain, it could be triggered to run on-demand on a roaming location depending on the weather type and likely impacts. This would enable a ‘regime-based capability response’ (Done et al., 2006; Flack et al., 2018) that would add additional tools for Operational Meteorologists to provide stakeholders with more details and greater understanding of uncertainty for high-impact weather events. Trialling this capability in a testbed environment allows demonstration of benefits and gains understanding as to the practicality of being introduced to operations. This shelters operations and end customers from exposure to experimental and yet to be fully tested new capabilities. The winter testbed provided the first opportunity to trial this potential new capability in real time, with a greater effort of Scientists and Operational Meteorologists to scrutinize and discuss the outcomes.

In addition to enhanced NWP capability during high-impact events, there is also the possibility of enhanced observation capability. The UK observation network is among the more advanced in the world, with full coverage radar of the country, extensive ground observation stations and four operational radiosonde stations launching two times daily. The UK observation network also includes some additional sonde stations, which can be triggered to launch on demand by the on-duty Chief Meteorologist. This capability was built to provide additional upper-level information in times of scenario uncertainty, to complement existing sources of information and to enhance situational awareness for high-impact weather events. However, the effectiveness of the capability had not been extensively trialled, and the testbed provided an excellent opportunity to combine extra observation capability with enhanced modelling and scientific analysis.

Here, we set out how the UK winter testbed 2020/2021 was organized and report some of the key findings. The aim of this article is to describe the interpreted differences between the real-time performance of the 300-m and 2.2-km ensembles alongside the description of the practicalities of running the testbed and documentation of lessons learned. The description of the cases is primarily drawn from our in-person conversations during the testbed. The overall purpose of the testbed was to investigate whether a real-time weather regime-based enhancement in NWP capability was feasible, to understand what benefits a testbed environment might give, and to explore the practicalities of running such an event.

Section 2 describes the practical functioning of the testbed and introduces the data sets. Section 3 shows some of the case studies representative of the overall findings from the period. Section 4 discusses the lessons learned, both scientifically and practical recommendations for running testbeds in the United Kingdom. Section 5 provides a brief summary and conclusions.

2 | METHOD

In winter 2020/2021, the world was facing extensive challenges due to the COVID-19 pandemic, with significant ‘lockdown’ in force across the United Kingdom. Most Met Office staff, including Operational Meteorologists, were working from home and as schools were temporarily closed many staff had to flex their days around personal responsibilities. It was recognized at an Executive level in the Met Office that delivery of non-essential projects may be disrupted. Given these circumstances, there was a unique focus on making sure that the testbed did not add undue additional pressure to workloads.

The Met Office testbed approach for this winter would therefore be different from a traditional intensive testbed environment, with the NWP experiments run and extra observations provided on-demand and dependent on weather conditions. The testbed was held in a completely virtual environment, with no in-person meetings. The virtual nature meant there was no barrier to participation across different locations. This also facilitated academic collaborators to take part with minimal commitment, and staff from the University of Leeds were present for many of the testbed calls.

This approach presented two questions—how to keep momentum going over the longer period when there would not always be an ‘event’ to study and what would happen when there was a triggered testbed ‘event’?

2.1 | Extended testbed period

In summer 2020, as part of the Short-Range Numerical Weather Prediction—Ensemble Prediction Systems (SRNWP-EPS) project (Callado-Pallarès et al., 2021)—a ‘mini-testbed’ occurred in the United Kingdom. In this ‘mini-testbed’, MOGREPS-UK forecasts were examined, daily, over the three summer months (June, July and August) and cases that were deemed forecast busts were subjectively examined in greater detail. These case studies were used to understand the benefits of multi-model ensembles at this scale (personal communication from Porson and Flack). Simple plotting tools (pre-producing plots using a python script and indexing in a webpage) were used to examine the cases, and regular drop-in meetings were held between Scientists and Operational Meteorologists. These meetings kept momentum and gained insight from Operational Meteorologists on their view of the ensembles’ behaviour. Further to these
fortnightly meetings, a digital notebook was set up for the period and cases of interest were discussed in more detail in slower time. The notebook was also used to generate ideas for new plots for visualizing the ensemble and ideas of why the busts occurred.

The structure of the winter 2020/2021 testbed extended period followed a similar blueprint, and a weekly drop-in meeting was held between Scientists and Operational Meteorologists to maintain momentum and build the community. The purpose of this meeting was to review the existing operational models’ performance for the previous week and look for initial signs in the coming weeks forecast for a high-impact event that could trigger a surge in modelling and observational effort.

The meeting was an effective way to share community understanding of Met Office capability and inspect operational NWP performance in a real-time environment, something that is not always done by Scientists. The meeting was led by a host who commented over weather events and model performance, but it had an informal format, so interjection was encouraged to generate more discussion and wider understanding across expert areas.

In addition, and to take into account shift working patterns, online open discussion and decision-making was conducted on a messaging board facility rather than over email. This ensured that the whole testbed community had equal opportunity to discuss and influence the triggering of additional capability. This transparency in decision-making was intended to further build the sense of community and questions and answers were available for all members to view.

2.2 Triggered testbed events

It was agreed between participants that triggering of the sub-km ensemble and extra observations would be driven by weather phenomenon. There was a stronger weighting towards weather type than population impacts in this testbed as it was felt that the core purpose was to gain feedback on modelling capability rather than stress-test services to customers. Future testbeds may have a stronger impacts-focus and include external stakeholders (discussed further in Section 4). The priority weather-types were discussed prior to the testbed and included fog, winter weather (with a focus on identification of the rain/snow boundary) and high-impact weather events.

The triggering of the ensemble was manual with only four or five Scientists able to execute the set-up. Therefore, when there was a possibility of an interesting event occurring, usually a discussion was held on the messaging board for a go/no-go collective decision. It should be noted that there was no single person responsible for final call on the decision, so a consensus was sought if possible. Overall, there were approximately 15 regular contributors to the testbed with a further 10 people somewhat involved. Thus, gaining agreement about which cases to run was generally successful given the constraints of the working practices (fitting this in around business-as-usual work) and the small numbers involved in the testbed.

2.3 UM sub-km (300 m) ensemble

At the beginning of the testbed, the decision to trigger the sub-km ensemble needed to occur at least 3 days before the weather event. This gave time for the ancillary files to be generated, the model to be reconfigured to the required domain, the model to run on the supercomputer (in the non-operational queue) and the plotting routines to pre-produce the static images that were then used in the discussion. By the end of the testbed in February, the process had been refined to being able to trigger at less than 24 h notice (dependent on Scientist availability), a considerable improvement and achievement over such a short period.

The flexibility of the testbed environment allowed different domain sizes and science configurations to be tested in parallel. Table 1 describes the science configurations used for the various models that were inspected in the testbed. The operational MOGEPS-UK in 2020/2021 did not output the necessary fields to drive the 300-m ensemble, so an intermediate model was needed to drive the 300 m. A new 18-member 2.2-km ensemble was run inside MOGREPS-G before the 18-member 300-m ensemble was nested within the larger 2.2-km domain (note: due to upgrades since the testbed sub-km ensembles can now be nested directly inside MOGREPS-UK). The testbed 2.2-km ensemble was not identical to MOGREPS-UK. A relocatable domain was used for the testbed 2.2-km ensemble, and it was chosen to best cover the area of interest. The 2.2-km domain was different (slightly smaller for computational constraints) than that of MOGREPS-UK and so the initial and lateral boundary conditions for the 300-m domain were slightly different than what would have been used if nested directly in MOGREPS-UK.

A consequence of re-running the 2.2-km ensemble and using a relocatable domain was the lack of MURK aerosol scheme (P. A. Clark et al., 2008) and MORUSES urban scheme (Porson et al., 2010) within the testbed ensembles (both 2.2 km and 300 m), due to technical limitations with the ancillary generation capability. The orography ancillaries were also generated from different sources to those in MOGREPS-UK. These small differences in configuration should be considered when
comparing any results from the 300-m and 2.2-km testbed ensembles to operational MOGREPS-UK. All ensemble forecasts were run out to a forecast lead time of $T + 54$ h for all testbed events.

To achieve timely forecasts when running on the Met Office supercomputer, the domain size was limited to a maximum of 500 grid points. This allowed the running of up to three or four separate regions (domains) of interest for the same forecast cycle. This was relatively successful and shows there is scope for running multiple embedded sub-km ensembles simultaneously, for example, in the November orographic rainfall event (see Table 2) domains were run across, north Wales, northwest England and Scotland.

In early 2021, new capabilities introduced in the ensemble system meant that experiments could be run on the faster ECMWF X40 supercomputer. This enabled the use of larger domains such as the $1250 \times 1250$ grid point 300-m ensemble for Storm Christoph in January 2021, and $1000 \times 700$ grid point domains for the February 2021 events.

Throughout the 2020/2021 winter testbed, the RAL2-M Regional Atmosphere & Land v2 (mid-latitude) science configuration was run as the standard configuration for the testbed 2.2-km and 300-m ensembles (see Bush et al., 2020 for a description of RAL1-M).

The additional supercomputer resource from ECMWF towards the end of the testbed allowed additional science configurations to be tested in comparison with the baseline RAL2-M. Two case studies in February trialled a real-time development version of RAL configuration at both 300-m and 2.2-km horizontal resolution (see Table 2). The development of RAL moved on such that these early prototypes have not been largely adopted, but the experiments helpfully demonstrated the computational cost of running the simulations was very similar to the costs of the control RAL2-M simulations.

### 2.4 On-demand observations

The Met Office has several on-demand observation capabilities that allow the triggering of extra observations at times when the Chief Meteorologist on duty requires additional situational information. To trigger extra radiosonde observations, the Chief Meteorologist has to make a request to the staffed observatories to enable additional sonde launches for the following day. This happens routinely when the Met Office Guidance Unit believes additional sonde data will provide further insight into the meteorology. It is also possible to request data from the Upper Atmospheric Research Satellite and to request partnerships teams within technical services to increase the number of Aircraft Meteorological Data Relay (AMDAR) ascents from UK airports—though this is reliant on their being flights from those airports with AMDAR equipped aircraft.

During the period of this testbed all additional observation mechanisms had added restrictions due to COVID-19. An increased frequency of sonde launches requires additional engineering resource to visit the auto-sonde sites to refill the systems, limiting the scope for requesting additional radiosonde ascents. COVID-19 also had a significant impact on the number of flights leaving UK airports, and the AMDAR team had been trying to

---

**Table 1:** Science configurations for the models tested

|                        | MOGREPS-UK ensemble (2.2-km resolution) | 2.2-km ensemble | 300-m ensemble |
|------------------------|-----------------------------------------|----------------|----------------|
| Initial conditions     | UKV anomalies                           | UKV anomalies  | Regridded UKV anomalies |
| Lateral boundary       | MOGREPS-G                               | MOGREPS-G      | 2.2-km ensemble |
| conditions             |                                         |                |                 |
| Cycling                | Hourly following Porson et al. (2020)   | Hourly following Porson et al. (2020) | Hourly following Porson et al. (2020) |
| Domain                 | As Hagelin et al. (2017)                | Subsection of MOGREPS-UK on demand | 150 x 150 km on demand |
| Science configuration  | As Porson et al. (2020)                 | RAL2-M         | RAL2-M (with additional dev-RAL experiments) |
| Aerosol scheme         | MURK                                    | None           | None           |
| Urban scheme           | MORUSES                                 | None           | None           |
| Orography ancillaries  | DTED 100-m (aggregated to 1-km) dataset | SRTM 90-m resolution dataset (Farr et al., 2007) | SRTM 90-m resolution dataset (Farr et al., 2007) |

*Note: MOGREPS-UK is classed as an operational model and runs continuously at the Met Office, whereas the 2.2-km and the 300-m models were set up on-demand for the testbed only.*
TABLE 2  Summary of testbed triggered events for the winter 2020/2021 Met Office testbed

| Dates                  | Case study description                        | Location of 300 m ensemble(s)                                                                 | Notes                                                                 | Model configurations |
|-----------------------|-----------------------------------------------|-----------------------------------------------------------------------------------------------|----------------------------------------------------------------------|----------------------|
| 23–25 November 2020   | Orographic precipitation in NW Scotland and Cumbria | Cairngorms, NW Scotland (Fort William), NW Wales, Cumbria                                      | Test case to understand ensemble set up and test plotting routines    | RAL2-M               |
| 25–30 November 2020   | Fog in Midlands, Yorkshire and Eastern England | Devon, Wiltshire, Corby (Midlands), York(shire)                                               | Test case to understand ensemble set up and test plotting routines    | RAL2-M               |
| 3–5 December 2020     | Precipitation boundaries Scotland and Northern England (rain/sleet/snow event) | NE Scotland, NW Scotland, Peak District, Gloucestershire                                        |                                                                      | RAL2-M               |
| 12–13 December 2020   | Precipitation banding in Devon, Gloucestershire and Snowdon | Devon, Gloucestershire, Peak District                                                        |                                                                      | RAL2-M               |
| 15–16 January 2021    | Snow/Freezing Rain SE, NE England snow        | SE England, NE England                                                                        | See Section 3.1 for more details Additional radiosondes at Herstmonceux 06 and 09 Z on 16 January (Albermarle was requested but not available) | RAL2-M               |
| 19–21 January 2021    | Storm Christoph Flooding in North Wales and NW England | Large 300-m domain centred over Chester (1250 by 1250 grid points), covering North Wales and NW England | See Section 3.2 for more details                                      | RAL2-M               |
| 6–11 February         | Frontal snowfall in Southeast England         | 500 (E–W) by 700 (N–S) grid point domain over East Anglia at first.                            |                                                                      | RAL2-M, Dev-RAL; Dev-RAL 140 levels tests |
| 9–11 February         | Winter showers in Eastern United Kingdom      | 1000 × 700 grid point domain centred over Midlands—North-east England.                         |                                                                      | RAL2-M, Dev-RAL      |
| 19–20 February        | Orographic enhancement South Wales and Southwest England | Large 300 m domain 1000 (N–S) × 700 (E–W) centred in Southwest England                        |                                                                      | RAL2-M, Dev-RAL      |

Note: The case studies discussed in this article are highlighted in bold.
optimize the aircraft used to ensure the smallest impact to the number of observations. However, the impact on observations had still been significant (a reduction of up to 80%), so there was no capacity to ask for an increased number of AMDAR observations.

A testbed environment presents an opportunity to look at other observation products, which are not routinely used by Operational Meteorologists or are still in development. In the winter of 2020/2021, this included products such as observations from the Weather Observation Website (WOW), which are shared by members of the public as well as other operational observation data that were generally underutilized (e.g., freezing/melting layer information from Laser Cloud Base Recorder (LCBR) backscatter, and radiosonde descent data) and ad-hoc products (e.g., air mass mapping from interpolated Mode-S data).

The objectives for including additional observations in the testbed were generally to explore new capabilities rather than scientific: to extract more value from current observations, showcase products and get feedback from Scientists and Operational Meteorologists about how to make observations more useful and useable. There was also an aim to understand how additional observations might be integrated into future testbeds (processes, triggers, etc.).

3 | RESULTS

The relocatable sub-km model was triggered nine times over the winter testbed, with two of those being tests carried out in November 2020 to experiment with the model set-up and plotting routines. Table 2 lists the case studies and their characteristics. Here, four of the case studies are discussed and initial results presented, before discussion of a cross-testbed analysis of ensemble characteristics.

3.1 | 16 January 2021: Snow precipitation boundaries

After the passing of a transient ridge on 15 January cyclonic flow from a low centred south of Iceland, with trailing fronts reaching the United Kingdom dominated the large-scale situation (Figure 1a). The warm front located over England was predicted to traverse eastward across the country, and given the very cold air ahead of the front, it was likely that there could be a mix of rain/sleet/snow/freezing rain occurring between 03 and 15 UTC on 16 January 2021. The operational models (UKV, UM Global Model, MOGREPS-G and MOGREPS-UK, not shown) had been suggesting that snow was likely for the south-east of England and that freezing rain was possible in the north-east of England, thus during and shortly after the weekly drop-in meeting on 14 January, it was decided that two domains would be run with the 300-m ensemble to examine these boundaries: a south-east England domain, which included the radiosonde station at Herstmonceux (Figure 1b), and a north-east England domain, which included the radiosonde station at Albermarle (Figure 1b). Further to this, additional radiosonde ascents were requested for 16 January and two additional ascents at 06 and 09 UTC were obtained, both from Herstmonceux.

Figure 1. (a) 06 UTC 16 January 2021 surface analysis showing a warm front passing over England. (b) Location of UK radiosonde launch sites for sonde launches available at 00 UTC and 12 UTC daily (from north-to-south, L, Lerwick; A, Albermarle; CB, Castor Bay; W, Watnall; H, Herstmonceux, C, Camborne)
The snowfall probabilities for the two different regions in both the 2.2-km and the 300-m ensembles are shown in Figure 2 and the radar observations of precipitation accumulation are shown in Figure 3. Snow was observed at many of the weather stations covered by the model domains (Figure 3 white circles). Note that the weather station observation of ‘snow’ is obtained from an automated sensor (CS125 Present Weather and Visibility Sensor, Campbell Scientific, 2022) and it is recognized that it can be difficult to discern the difference between rain/sleet and snow. On manned stations, observations of snow are confirmed by a human observer.

The plots in Figure 2 show similar spatial coverage in the 2.2-km and 300-m ensembles, but the probabilities of snow accumulation are greater in the 300-m model, and these are especially tied to regions of higher orography such as to the north-west of London. This enhancement is realistic and demonstrates there is a benefit in the 300-m ensemble in both domains in relation to its representation of orographic processes, though it is difficult to confirm such detail due to the model grid being more detailed than observation station density (circles in Figure 3). The spatial envelope of probabilities remains broadly the same for the 300-m and 2.2-km models, likely suggesting that most spread in the 300-m ensemble is coming from the driving (2.2 km) ensemble. The impact of the size of the 300-m domain on ensemble spread is further discussed in Section 3.3. Another potential limitation to ensemble spread is the weather regime, which was a synoptically forced frontal system rather than locally driven weather.

Given the focus for this case was precipitation-type boundaries, there was interest in how closely the models compare to radiosonde profiles. Figure 4 shows the model ascents against observations at Herstmonceux and Albermarle at 06 UTC and 00 UTC on 16 January, respectively. An assessment of the sensitivity to the method used to compare sonde with model vertical profile data was conducted. This showed that while horizontal balloon drift had some impact (Laroche & Sarrazin, 2013), qualitatively the impact on model—observation errors and spread between ensemble members were sufficiently small (not shown) that a simple comparison between radiosonde data and the vertical model column nearest the sonde release location at the surface was used.

The ascent at Hestmonceux (Figure 4a) shows both temperature and dew point observations are generally

![Probability of 3-h snowfall accumulation exceeding 2-mm water equivalent](image)

**FIGURE 2** Probability of 3-h snowfall accumulation exceeding 2-mm water equivalent: (left column): In 3 h 03:00 to 06:00 on 16 January in NE England domain—Note the image has been clipped to focus on land points; (right column): In 3 h 06:00 to 09:00 on 16 January in SE England domain. Top row shows results from 2.2 -km model, bottom row shows the 300-m model.
FIGURE 3  Shading represents radar precipitation accumulation from 03:00–09:00 UTC on 16 January 2021. Weather Station locations are shown as dots with those reporting snow on 16 January 2021 marked as white and those reporting rain-only marked black. Albermarle (north; Figure 1b) and Herstmonceux (south coast; Figure 1b) are marked by blue rings around the stations. RAF Northolt station has a red ring. The location of central London is marked by a cross. The northern 300-m model domain is shown in red, and the southern domain is in purple.

FIGURE 4  (a) Herstmonceux ascent for 06 UTC and (b) Albermarle ascent for 00 UTC both on 16 January 2021 (observation in black), model grid point ascents for 6 of the 18 members of the 2.2-km (red) and 300-m (blue) ensembles for the 05:00 Z 15 January 2021 cycle. Solid lines are temperature and dashed are dew point. Model dew point values are calculated from relative humidity and temperature between ±35°C following Bolton (1980).

captured by the ensemble simulations. Note that only six members are used in the comparison due to data archiving. The ensemble members are plotted as the ensemble mean would likely wash out any signal. The model profiles for temperature show errors within 5 K of that observed in the sounding but there are times the dew
point field was not as well represented. Closer analysis of model errors showed there were several times in the profiles where many of the members had dew points larger than the (dry bulb) temperature, implying the air was super-saturated throughout much of the troposphere. Given the surface temperatures at this time, this implies that either sleet or snow would be the dominant form of precipitation. This agrees with the probabilities shown in the south of the SE domain in Figure 2. There was no statistically significant difference found in the errors between the model and radiosonde temperature at different resolutions, although there may be a small difference in the dew point (and hence humidity) and most errors were within the spread of the ensemble.

At Albermarle, both 2.2-km and 300-m ensembles produced similar ascents (Figure 4b). The radiosonde station is located more centrally in the model domain, so the models are less likely to suffer boundary effects at this location. Again, temperature was reasonably represented by the models; however, all members lacked the defined inversion at the top of the boundary layer such that the model was too warm beneath the inversion and too cool above. There was more spread in the dew points with several members producing very dry ascents aloft and there was a reduction in the number of super-saturated points compared with the Herstmonceux ascent in the moist areas.

For both Herstmonceux and Albermarle, all ascents from the period were compared against observations (8 radiosonde ascents in total during the case) to understand the differences between the resolutions. Figure 5 shows the errors for the 2.2-km and 300-m ensembles for Albermarle showing that the errors for models compared with observations are significantly larger than the differences between the 2.2-km and 300-m models, with limited changes with resolution. The consistency in the errors for the different resolutions is particularly strong above 600 hPa in both fields. This is to be expected given that weather phenomena at higher levels are synoptically driven and these larger scales are well-resolved by global scale models. This consistency in model errors above 600 hPa likely suggests that the 300-m domain is not large enough to allow perturbations to spin-up in this area and develop differences compared with the 2.2-km ensemble at higher levels. As discussed further in Section 3.3, errors on the kilometre/sub-kilometre scale, in these small domains, would not have enough time (or space) to spin-up and impact on the upper levels and the error is dominated by the driving global model.

![Figure 5](image-url)
Confirmation of snow events is challenging using the observation network. The United Kingdom has limited snow depth recorders due to the rarity of snow events, and while other SYNOP and METAR observations are available, the density is low, and many are provided by automated systems that have difficulty identifying marginal snow events. Arguably the easiest way to spatially diagnose observed snow is by satellite images (to see lying snow), but this is only appropriate for clear skies and daytime. The resolution of satellites is also not comparable to models, meaning that geographic details of snow/rain boundaries, especially over orography, are not able to be confirmed. Other observation types were consulted to gain more information about the case study and diagnose snow versus rain.

There is some evidence that an LCBR (a type of low power LIDAR instrument also known as ceilometer) may be useful for distinguishing liquid layers, ice clouds, fog and precipitation as well as phase transitions occurring within the melting layers (Di Girolamo et al., 2012; Sassen et al., 2005; Sassen & Chen, 1995).

In the United Kingdom, there is a network of about 40 LCBR, which provide backscatter data to a central processing system and to a Europe-wide network (https://e-profile.eu/#/cm_profile). Figure 6a shows the uncalibrated range corrected backscatter data for 16 January 2020 at RAF Northolt. This site is in the middle-left of the SE England domain and declared ‘snow falling’ and ‘snow lying’ in the station records for the day (see Figure 3).

In Figure 6a the black dots indicate cloud base heights, blue/green colours indicate boundary layer aerosol and weaker backscatter not associated with cloud and falling hydrometeors are shown as red/yellow colours. From 03:00 to 04:30 UTC, evaporating rain is observed under the cloud base. During this period, the boundary layer height can be determined by a step change in signal strength at about 750 m altitude. This precipitation begins to reach the ground at approximately 05:00 UTC.

Between 06:00 and 08:00 The LCBR image shows some more complex structure (e.g., green yellow banding at 1 km altitude indicating a phase transition) and given the

![Figure 6](image_url)

**Figure 6** (a) Laser Cloud Base Recorder time series for 16 January 2021 at RAF Northolt (b) Northolt weather station observations of minutely temperature (black line, values on right axis) and rainfall (grey bars, values on left axis)
temperature drop (Figure 6b) and human observation of snow, we might conclude that at this time the precipitation fell as snow.

An additional challenge that this case highlighted during the testbed was the time coincidence of the prediction and observations in real time. They were often available only at different intervals, data were stored in completely different locations and there was huge dependency on locally created plotting scripts rather than a coordinated technical strategy or visualization system. The logistics of comparing the data sources side by side was a limiting success factor in the assessment of the case.

### 3.2 High-impact weather case study 19–21 January 2021: Storm Christoph

Storm Christoph was unusual in that it was named due to the potential for large amounts of rainfall (and associated impacts) rather than the more traditional wind related hazards. With the near-stationary front across northern England (Figure 7), an Amber National Severe Weather Warning Service (NSWWS, see Neal et al., 2014 for an explanation) was issued on 18 January 2021, mentioning the potential to see up to 200 mm of rain across high ground during the 48-h period.

Due to both the large area that was forecast to see significant impacts from the storm and the new capability allowing use of the ECMWF supercomputer, the pragmatic decision was taken to run a larger 300-m ensemble. A 1250 (N–S) by 1250 (E–W) grid point domain was created centred on north-west England, covering the high ground of the Pennine hills (the hills running the spine of England), Wales and Cumbria/NW England. This was, at the time, the largest domain sub-km ensemble ever run by the Met Office.

Figure 8 shows the postage stamp precipitation accumulation plots from the ensembles at 300 m and 2.2 km valid from 09 to 12 UTC on 20 January 2021. Postage stamps over short time periods are useful for identifying variations in short-lived high-intensity events, whereas a longer time period would smear out differences between ensemble members. Note that two of the 300-m ensemble members failed to run, and this made real-time comparisons between individual run members in the postage

![Analysis charts showing Storm Christoph across the United Kingdom: Top left, 00:00 Z 19 January 2021; top right, 12:00 Z 19 January; bottom left, 00:00 Z 20 January 2021; bottom right, 12:00 Z 20 January 2021. The frontal system stalled across northern England bringing a prolonged period of heavy rain](image)
We have chosen not to re-plot but display the run as seen in real time in Figure 8 to illustrate the issue.

Probability maps such as Figure 9 were also available in real time. The probability plots are shown over a longer time period (48 h) as individual tracks of rainbands in a shorter period would under-display the event, for example, similar members could produce rainbands but not at coincident times, so looking over a short time period might underestimate the probability of the event. Figure 9 shows more widespread probabilities of chance of high precipitation accumulations in the 300-m ensemble, but the trade off with displaying the longer time period means that details of the event are smeared out, making targeted information on impacts difficult to discern.

This case highlights the need for better ways of visualizing ensemble information such that impacts and geographic information of greatest interest to users is amplified. Roberts et al. (2019) also discuss this issue and summarize some novel ways to display relevant weather phenomena from ensemble data for the United States.

**FIGURE 8** Postage stamps of 3-h precipitation accumulation (mm) for ensemble members of the 300 m (top) and 2.2 km (bottom) from 09:00 to 12:00 UTC on 20 January 2021 ($T + 28$ to $T + 31$ h)
is clear from our work that the United Kingdom would also benefit from research by displaying ensemble diagnostics that highlight important weather phenomena relevant for the United Kingdom.

Figure 9 shows that the ensembles picked up the risk of large precipitation totals, particularly across high ground of N Wales and NW England. The 300-m ensemble simulated higher probabilities of 48-h accumulations more than 100 mm (bottom row) across the higher ground in Wales. The 300-m ensemble also indicated much stronger probabilities for rainfall greater than 100 mm in parts of N/NW England with the differences mainly tied to higher elevation areas. The event was detected in fewer of the MOGREPS-UK ensemble members, resulting in a reduced probability of extreme weather (Figure 9 left panel), generally under-representing the extremity of the event.

Storm Christoph resulted in flooding across parts of NW England and N Wales, the rainfall totals at some locations were more than 100 mm for the event total (Figure 10) and was one of the wettest 3-day periods for these regions on record (Kendon, 2021). Focusing on NW England, the regions highlighted in purple on Figure 9 (>100 mm observed) are remarkably similar to those highlighted as likely to see greater than 100 mm in the 300 m ensemble. The probabilities for >100 mm increased from approximately 20% in MOGREPS-UK to approximately 60%-70% in the 300-m ensemble, suggesting the 300-m ensemble would have been able to add confidence for the warnings in this extreme event had it been operational. Greater Manchester, located in the centre of the domain, narrowly avoided more major flooding, but many rural areas suffered severe impacts, with the peak of the flooding occurring on 21 January. Fortunately, the event was broadly well predicted by operational models, and Operational Meteorologists were able to detect this was a severe event and issue warnings accordingly. This case study highlights that the experimental 300-m model would have added further detail and created more targeted geographic warnings. In the actual event, the Met Office and Flood Forecasting Centre worked directly with the central government and local authorities to prepare for the storm and the subsequent flooding, reducing the overall harm and resulting in no lives lost.

3.3 | Cross-testbed assessment of ensemble spread in atmospheric profiles

In addition to the near real-time assessments of model simulations and observations during the testbed, the experiments conducted have provided a resource for further scientific evaluation, examining whether more
robust conclusions can be better drawn through assessment across a number of cases. The specific example of ensemble spread in the vertical atmospheric profiles is discussed here.

A total of 17 radiosonde ascents were compared with multiple ensemble forecasts during the cases of 15–16 January, 19–21 January and 19–20 February 2021 (Table 2), resulting in 202 coincident model profiles. Figure 11 shows the absolute model minus observation differences in vertical temperature and dew point temperature in 2.2-km and 300-m RAL2-M ensembles. The model dew point temperature was derived from relative humidity following Bolton (1980), and is thereby only shown for temperatures within the range of ± 35°C.

In common with the single case study results shown in Section 3.1 (Figure 5), the mean errors for temperature are statistically indistinguishable between 2.2-km and 300-m resolutions across multiple cases. The largest errors are associated with a single ascent at Watnall on 20 January 2021 (see Figure 1 for location). Differences between model resolutions are consistently an order of magnitude smaller than the errors to observations of around 1 K on average.

Consistent error profiles are also found for dew point temperatures. Larger errors are apparent, linked to occasions when the model is dry but should be saturated, and to times when model profiles are super-saturated in cloudy regions. Unlike for temperature, the mean errors are found to be statistically different at 5% confidence level when considered for multiple cases, with an average improvement of 0.2 K in the sub-km ensemble compared with the 2.2-km ensemble.

As a first step to assessing the spread–skill relationship of 2.2-km and 300-m ensembles, Figure 12 directly compares the model to observation vertical temperature error (the ‘skill’) with member to member differences for unique member–member pairs (defined as the ‘spread’ here).

For both 2.2-km and 300-m ensembles, the relationship between spread and skill with height is qualitatively similar, with similar SD of the spread and errors, particularly near the surface and at earlier lead times. This suggests that perturbations applied in the regional ensembles are likely to be of correct magnitude. Larger differences occur in regions of sharp inversions such as the top of the boundary layer and tropopause, where these are not reliably captured in the same location as observed and the model error lies outside the spread (Hanley & Lean, 2021). By later lead times, the spread–skill relationship changes in structure, with the spread remaining more Gaussian around the mean, centred near zero and relatively constant with height, while the mean error grows and becomes less constant with height. Similar conclusions can be drawn, albeit with larger errors (Figure 11) and larger spread from assessment of the vertical dew point profiles across multiple cases from the testbed (not shown).

The consistency of the spread–skill results from this initial sample of km-scale and sub-km ensemble simulations suggests a hypothesis of there being more limited impact of higher resolution due to the limited domain size. The maximum domain size of sub-km ensembles considered here is 420 km, compared with 925 km horizontal extent of the MOGREPS-UK operational domain. This suggests that only the initial perturbation growth (Selz & Craig, 2015) can take place in the domain, but that when these small-scale errors begin to upscale they are quickly removed from the model domain and hence have limited influence on the forecast. Illustrative examples of error refresh rates for the different ensemble domain sizes are provided in the Annex for reference. This implies that as computational resources increase, allowing for use of larger sub-km ensemble domains, it should be expected that spread from the sub-km ensemble will improve and become more representative through both downstream and upscale error growth. However, further work into physically based (P. A. Clark et al., 2021) and physically related perturbations should be explored.

The on-demand testbed provided only a limited number of cases from the same season and constraints on output data volumes in order to complete forecasts in a timely manner that reduced the potential sample size of

FIGURE 10 Map showing estimated rainfall totals during Storm Christoph 09:00 UTC 18 January to 09:00 UTC 21 January 2021. Areas highlighted in purple indicate that event totals exceeded 100 mm. Note the time period is slightly different from the probability maps in Figure 9 to cover the whole of the event rather than just the most intense period.
data available for analysis. However, it has proven possible to draw a more robust indication of cumulative statistics, such as spread–skill relationships, from assessment across multiple testbed results. This suggests that there is value in repeating similar model experiments through future testbed activities, including studies during summer, for example.

4 | DISCUSSION

The on-demand winter 2020/2021 testbed was the first of its kind in the United Kingdom. The aim of the testbed was to understand the logistics of running a virtual testbed at the Met Office, what benefits it could give, and to trial a new sub-km ensemble capability using a regime-based triggering approach. The testbed was also used to subjectively evaluate the model in near real time through Scientist and Operational Meteorologist interactions. This was the first time that a sub-km ensemble had been tested in a near real-time context over the United Kingdom and this gave some insight into where benefits of this type of modelling system may come. However, all the findings shown need to be strongly caveated with the limitations of the experimental set-up, not least the very sparse number of case studies and the constrained type of weather event and season.

It appeared that the sub-km ensemble benefits were most noticeable when local forcing was an important driver of weather conditions. For example, wintry showers seemed well captured as did weather over complex orography, for which the 300-m scale was more able to resolve compared with the 2.2-km ensemble (such as during the Storm Christoph case study).
Sub-km models might be expected to do well at representing surface-forced phenomena due to the small-scale heterogeneities inherent in these types of weather features. There may even be a statistical benefit for some aspects of frontal structure (Hanley et al., 2016). Future work might consider whether an objective measure of large-scale versus local-scale weather drivers would enable a more complete analysis and evaluation of the ensembles in the testbed environment. Such a diagnostic would also hold the potential to identify weather regimes where sub-km modelling is more beneficial. More objective measures could build a pathway towards efficient use of operational sub-km ensemble model triggering in the future.

Sub-km models might be expected to do well at representing surface-forced phenomena due to the small-scale heterogeneities inherent in these types of weather features. There may even be a statistical benefit for some aspects of frontal structure (Hanley et al., 2016). Future work might consider whether an objective measure of large-scale versus local-scale weather drivers would enable a more complete analysis and evaluation of the ensembles in the testbed environment. Such a diagnostic would also hold the potential to identify weather regimes where sub-km modelling is more beneficial. More objective measures could build a pathway towards efficient use of operational sub-km ensemble model triggering in the future.

There is an important avenue to explore on understanding how performance gains from the sub-km model forecast are translated into benefits for users of weather information. This could be tested more rigorously in future testbeds where ways to exploit the new information, essentially experimental co-production of products, could be investigated with expert and industry multi-sector users (Calhoun et al., 2021; Fletcher et al., 2021; Nkiaka et al., 2020).

Testbeds are an ideal place to try new model configurations that are yet to be fully developed. Trialling experimental science in real time can be a fast way to either accelerate or quickly dismiss a particular idea. During the testbed, we compared a development physics version of the model (dev-RAL) with the current operational version (RAL2-M), as briefly described in Table 2. Although the results are not shown here, the February case studies showed that both 2.2-km and 300-m resolution dev-RAL had a greater tendency to produce bands of snow showers that were aligned with the easterly flow. Results from the case studies prompted discussion around increasing the
vertical resolution to use 140 levels to revisit the early February case studies. This test configuration was computationally very expensive and the difference it made to the ensemble output was relatively small (for this snowfall case study), and no further investigation was pursued. However, this demonstrates the advantage of pursuing NWP research alongside a testbed environment as being immersed in case studies and real-time model output can throw up new avenues for research (or close off avenues) that might not otherwise have developed.

From an observation perspective, the testbed highlighted some strengths and weaknesses in the real-time availability of data in the United Kingdom. Radiosondes provided useful data for the testbed, and this utilization could be expanded further. It is recommended that future testbeds consider more direct collaboration with operational observation providers such that a parallel and clearly non-operational protocol can be accessed to request additional radiosondes. This is important for clearly demarking or sanitizing the testbed process away from the operational process such that requests for additional observations do not need to go through the on-duty Chief Meteorologist.

More surface observations were available for after event analysis and the access and plotting of the data in real time (such as the LCBR) was sometimes restricted to operational systems not widely available to Scientists. Many other observations are available in the UK network, for example, radiosonde descent data and surface observations, which were not used in this testbed due to the accessibility of the data. While some of the data are made available to Operational Meteorologists in real time, it is not easy to retrieve, understand and visualize the data and challenging to compare these measurements directly with the model. For testbed environments, real-time access to observation (and model) data is essential and it is recognized that creating the technical pathways to satisfy demand is not trivial, especially given that the volumes of data, for even single fields, are significant. Visualization of observations, ideally in the same visualization system as the experimental model data, which is open to everyone on the testbed, is also a significant need.

The testbed showcased automated snow falling observations but highlighted the current dependence on UK ‘citizen science’ to report snow depth as the Met Office observation network does not have widespread coverage for this variable. Observations that proved useful during real-time event analysis were WOW snow depth reports and the commercial site #uksnowmap that scrapes Twitter for any mentions of snow reports in the United Kingdom tagged with the hashtag. Unfortunately, at present, the #uksnowmap wipes its data on snow reports after a set number of hours, and future work might benefit from capturing the crowd-sourced Twitter observations for use in model evaluation. There are longer term options to further explore and exploit crowd-sourced observations. For example, for case studies, citizen volunteers could upload photos and GPS information, altitude, location, etc. from mobile phones, either to Twitter or to WOW. In WOW, entering snow observations is possible but not widely advertised and possibly not well known by regular users due to the rarity of the weather events.

Other types of Met Office observations did show some potential further use, the most promising of which was identifying freezing layer/melting layer from LCBR backscatter data. Further investigation is needed to confirm snow observations from the LCBR before it is more widely used. This may include using the staffed observation stations along with radiosondes to provide some verification to the LCBR data.

### 4.1 Benefits of testbed approach and lessons learned

The winter 2020/2021 testbed was the first testbed of its kind to be held in the United Kingdom and it is intended that testbeds will now be used more regularly, forming part of the research to operations process at the Met Office. The design of the testbed and the approach used were mainly a result of necessity (due to the extensive COVID-19 lockdown) rather than design. However, there were some clear advantages from its virtual approach. Most prominently, having the virtual event meant greater participation from disparate teams—such that Met Office staff based in Reading and Exeter locations were able to work together seamlessly. Decision-making was also more transparent as there was more digital information in addition to verbal communication. The message-board discussion group and the weekly virtual meetings were harmonious, though it should be noted that the number of ‘active’ participants was small (often less than 10 people) and almost exclusively Scientists. Through this, an interested community has begun to build. It should be noted that there are some disadvantages to virtual testbeds, and the online meetings could sometimes lack engagement, depending more on a strong host commenting on the weather situation. It is possible that some participants were intimidated by not seeing others on the call or knowing them (the ‘not wanting to say something stupid’ factor). This is the advantage that an in-person testbed may have, as friendships and relationships are built in both informal and formal environments, ultimately cultivating increased collaboration between participants.

It is recommended that future testbeds need clearer roles and responsibilities with more participation from Operational Meteorologists and partners from academic institutions as well as representatives from users of
weather information such as civil authorities or private industry. Including users is beneficial for shaping the scientific questions to be answered and understand tolerances for forecast error. For example, it became clear that the 300-m model could provide much stronger geographic detail during events, but it is unclear how much value this detail is for different sector users. Including users of this information in a test environment might shape how scientific results are interpreted and ultimately whether the science then evolves into operational practice.

Funding should be a high priority for future testbeds such that dedicated time and resource can be deployed. Future testbeds might also consider whether structured daily assessments with online forms would be beneficial, as seen in the HWT (A. J. Clark et al., 2022); although these are resource-heavy, they do allow for further investigation of how weather information is interpreted by human users.

The on-demand nature of the experiments was designed to minimize resource requirements, and although this worked reasonably, it was very dependent on a few dedicated Scientists. Predefined experiments would have reduced the logistical burden and facilitated resource management. However, the significant advantage of the on-demand testbed was that effort was only put in when the weather situation was worthy of note. Technical advances in ancillary generation since the testbed mean that some of the logistical challenges in launching the on-demand model have been improved. Even so, the advantages and disadvantages of both on-demand or predetermined approaches, along with science needs and resourcing considerations need to be considered when planning future testbeds.

The purpose of the experiment was partly scientific but also partly about logistics exploration. The testbed was the first real-time exploration of a regime-based response capability, where extra model runs are initiated based on weather events. It became clear from this testbed that more automated ways of triggering enhanced NWP or perhaps having a pre-set list of domains to choose from, is required if this capability were to become operational.

In addition to the logistical questions around triggering enhanced NWP are the scientific questions around the added value of on-demand relocatable sub-km ensembles. During the winter 2020/2021 testbed, it was often unclear as to if certain weather situations were worth triggering the sub-km model, again pointing to a more automated way to objectively identify weather regimes where the local scale dominates, and high resolution is likely to be advantageous. An example of a potential method for this in convective regimes is shown in fig. 6.2 of Flack (2017). Interaction with end-users to probe their information requirements for decision-making would also likely identify the weather situations of greatest interest where uncertainty information or details on particular parameters are highly desirable.

For winter 2020/2021 testbed the models were visualized using python code written by Scientists, which produced plots of popular variables and ensemble diagnostics. These images were then indexed with metadata and displayed on a website with drop-down menus to choose which plots to view. This plotting took time as thousands of static images were produced for each run. Improvements suggested for future testbeds might include better dashboard and visualization technology so that pre-plotting is not required, and the data are visualized as soon as they are available. It should also be easier to find case studies and interact with the plots (e.g., zoom and pan). This may give additional benefit, especially given the high resolution nature of the 300-m model. In general, it would be an advantage if the testbed visualization were in the same format and used the same colours as the operational tools, smoothing the route from research to operations should testbed capabilities prove successful. This needs to be balanced with security and accessibility requirements so that testbed participants both inside and outside of the Met Office have access to the same information.

Finally, writing up the case studies and bringing together data from operational observation systems and non-operational modelling systems was sometimes challenging. This article was strengthened by the breadth of expertise of the authors, but it also exposed difficulties in data discovery, data formatting and data storage. In the future, it would be advantageous if there were more consistencies in the technology underpinning research and operations such that both real-time analysis and post-event analysis were simpler to execute.

5 | SUMMARY

The UK winter testbed of 2020/2021 focused on trialling new modelling capability using an on-demand sub-km ensemble, triggered manually depending on weather conditions. The aim of the testbed was to investigate the logistics of triggering models on-demand as well as understand what advantage a testbed environment could give to stimulate science to operations testing and pull through. The testbed provided some interesting science but much of the discovery was about the mechanisms of running the testbed itself. An extensive number of lessons learned have been captured to improve future testbeds.

AUTHOR CONTRIBUTIONS

Caroline L. Bain: Conceptualization (lead); formal analysis (equal); investigation (equal); methodology (equal);
project administration (lead); writing – original draft (lead); writing – review and editing (lead). **Mike Bush:** Conceptualization (supporting); formal analysis (supporting); investigation (supporting); methodology (equal); supervision (equal). **Joelle Buxmann:** Data curation (supporting); visualization (supporting). **Paul Davies:** Conceptualization (lead); methodology (equal); project administration (equal); resources (equal); supervision (equal). **Anke Finnenkoetter:** Data curation (supporting); formal analysis (supporting); investigation (supporting); writing – original draft (supporting). **David Flack:** Conceptualization (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Humphrey William Lean:** Conceptualization (equal); investigation (equal); methodology (equal); project administration (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Kirsty Hanley:** Data curation (equal); formal analysis (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Richard W. Jones:** Data curation (lead); formal analysis (equal); investigation (equal); methodology (equal); software (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Jon Petch:** Conceptualization (lead); funding acquisition (lead); methodology (lead); project administration (lead); resources (equal); writing – original draft (equal); writing – review and editing (equal). **Aurore Porson:** Formal analysis (equal); investigation (equal); methodology (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Claudio Sanchez:** Data curation (equal); methodology (equal); software (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Edmund Stone:** Data curation (equal); formal analysis (equal); investigation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **David Walters:** Formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Stuart Webster:** Data curation (equal); investigation (equal); writing – original draft (equal); writing – review and editing (equal). **Steve Willington:** Investigation (equal); writing – original draft (equal); writing – review and editing (equal).

**ACKNOWLEDGMENTS**

The authors would like to thank all the participants at the UK winter testbed for useful discussions during the weekly meetings, including Prof. Doug Parker, Prof. John Marsham and others from the University of Leeds. Images from radar and maps of station data were provided from an archive map service created by Dan Hollis at the National Climate Information Centre (NCIC). We would also like to thank the Met Office Upper Air and Remote Sensing Network Management team for providing additional radiosonde ascents during the testbed. The authors thank the two anonymous reviewers whose comments greatly improved the manuscript.

**ORCID**

C. L. Bain [https://orcid.org/0000-0002-2993-9308](https://orcid.org/0000-0002-2993-9308)

D. L. A. Flack [https://orcid.org/0000-0001-7262-4937](https://orcid.org/0000-0001-7262-4937)

K. Hanley [https://orcid.org/0000-0002-1002-7667](https://orcid.org/0000-0002-1002-7667)

H. Lewis [https://orcid.org/0000-0003-3581-6459](https://orcid.org/0000-0003-3581-6459)

J. C. Petch [https://orcid.org/0000-0003-1476-5103](https://orcid.org/0000-0003-1476-5103)

A. Porson [https://orcid.org/0000-0002-5023-8522](https://orcid.org/0000-0002-5023-8522)

D. Walters [https://orcid.org/0000-0003-0383-7378](https://orcid.org/0000-0003-0383-7378)

**REFERENCES**

Bolton, D. (1980) The computation of equivalent potential temperature. *Monthly Weather Review*, 108, 1046–1053. [https://doi.org/10.1175/1520-0493](https://doi.org/10.1175/1520-0493)

Boulte, I.A., Finnenkoetter, A., Lock, A.P. & Wells, H. (2015) The London Model: forecasting fog at 333m resolution. *Quarterly Journal of the Royal Meteorological Society*, 142(694), 360–371.

Browning, K., Morcrette, C., Blyth, A., Bennett, L., Clarke, P., Corsmeier, U. & co-authors (2006) A summary of the convective storm initiation project intensive observation periods. Met Office Forecasting Research Technical Report no. 474, 153. Available at [https://library.metoffice.gov.uk/Portal/Default/en-GB/RecordView/Index/251895](https://library.metoffice.gov.uk/Portal/Default/en-GB/RecordView/Index/251895) [Accessed 3 Jan 2022].

Bush, M., Allen, T., Bain, C., Boutle, I., Edwards, J., Finnenkoetter, A. et al. (2020) The Met Office unified model/JULES regional atmosphere and land configurations (RAL) - version 1. *Geoscientific Model Development*, 13(4), 1999–2029. [https://doi.org/10.5194/gmd-13-1999-2020](https://doi.org/10.5194/gmd-13-1999-2020)

Calhoun, K.M., Berry, K.L., Kingfield, D.M., Meyer, T., Kroack, M.J., Smith, T.M. et al. (2021) The experimental warning program of NOAA’s hazardous weather testbed. *Bulletin of the American Meteorological Society*, 1–51. [https://doi.org/10.1175/BAMS-D-21-0017.1](https://doi.org/10.1175/BAMS-D-21-0017.1)

Callado-Pallarés, A., Marsigli, C. & Marcucci, F. (2021) EUMETNET convection-permitting ensemble databases hosted at ECMWF. *ECMWFRestletter*, 166, 6–7.

Campbell Scientific. (2022) CS125 present weather and visibility sensor. Available at: [https://www.campbellsci.com/cs125](https://www.campbellsci.com/cs125) [Accessed 3 Jan 2022].

Clark, A.J., Jirak, I.L., Gallo, B.T., Knopfmeier, K.H., Roberts, B., Kroack, M. et al. (2022) The second real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, 103(4), E1114–E1116.

Clark, A.J., Weiss, S.J., Kain, J.S., Jirak, I.L., Coniglio, M., Melick, C.J. et al. (2012) An overview of the 2010 hazardous
weather testbed experimental forecast program spring experiment. Bulletin of the American Meteorological Society, 93(1), 55–74.

Clark, P.A., Halliwell, C.E. & Flack, D.L.A. (2021) A physically based stochastic boundary layer perturbation scheme. Part I: formulation and evaluation in a convection-permitting model. Journal of the Atmospheric Sciences, 78, 727–746. https://doi.org/10.1175/JAS-D-19-0291.1

Clark, P.A., Harcourt, B., Macpherson, B., Mathison, C.T., Cusack, S. & Naylor, M. (2008) Prediction of visibility and aerosol within the operational Met Office Unified Model. I: model formulation and variational assimilation. Quarterly Journal of the Royal Meteorological Society, 134, 1801–1816. https://doi.org/10.1002/qj.318

Di Girolamo, P., Summa, D., Cacciani, M., Norton, E.G., Peters, G. & Dufournet, Y. (2012) Lidar and radar measurements of the melting layer: observations of dark and bright band phenomena. Atmospheric Chemistry and Physics, 12, 4143–4157. https://doi.org/10.5194/acp-12-4143-2012

Done, J., Craig, G., Gray, S., Clark, P. & Gray, M. (2006) Mesoscale simulations of organized convection: importance of convective equilibrium. Quarterly Journal of the Royal Meteorological Society, 132, 737–756. https://doi.org/10.1256/qj.04.84

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S. et al. (2007) The shuttle radar topography mission. Reviews of Geophysics, 45, RG2004. https://doi.org/10.1029/2005RG000183

Flack, D.L.A. (2017) Environmental controls on convective-scale perturbation growth, PhD Thesis, University of Reading, p. 146. Available at: https://centaur.reading.ac.uk/74387

Flack, D.L.A., Gray, S.L., Plant, R.S., Lean, H.W. & Craig, G.C. (2018) Convective-scale perturbation growth across the spectrum of convective regimes. Monthly Weather Review, 146, 387–405. https://doi.org/10.1175/MWR-D-17-0024.1

Fletcher, J.K., Parker, D.J., Hartley, A., Nkiaka, E., Adefisian, E., Clarke, S. et al. (2021) GCRF African SWIFT testbed 1 report. Leeds: University of Leeds. https://doi.org/10.5518/100/73

Gallo, B.T., Clark, A.J., Jirak, I., Kain, J.S., Weiss, S.J., Coniglio, M. et al. (2017) Breaking new ground in severe weather prediction: the 2015 NOAA/hazardous weather testbed spring forecasting experiment. Weather and Forecasting, 32(4), 1541–1568.

Groenemeijer, P. (2012) Testing of DWD products at the ESSL Testbed 2012. ESSL Technical Report 2012–04. Available at: https://www.essl.org/media/publications/essl-tech-rep-2012-04.pdf [Accessed 3 Jan 2022].

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. & Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPS-UK. Quarterly Journal of the Royal Meteorological Society, 143(708), 2846–2861.

Hanley, K. E., Pirret, J. S. R., Bain, C. L., Hartley, A., Lean, H. W. & Woodhams, B. J. (2021). Assessment of the Tropical Africa model over the Lake Victoria basin region. Quarterly Journal of the Royal Meteorological Society, 147(736), 1642–1660. https://doi.org/10.1002/qj.3988

Hanley, K.E. & Lean, H. (2011) Elucidating the causes of errors in 2.2 km Met Office Unified Model simulations of a convective case over the US Great Plains. Quarterly Journal of the Royal Meteorological Society, 147, 2741–2759. https://doi.org/10.1002/qj.4049

Hanley, K.E., Barrett, A.I. & Lean, H.W. (2016) Simulating the 20 May 2013 Moore, Oklahoma tornado with a 100-metre grid-length NWP model. Atmospheric Science Letters, 17(8), 453–461.

Kain, J.S., Paul, J., Weiss, S., Baldwin, M., Schneider, R. & Brooks, H. (2003) Collaboration between forecasters and research scientists at the NSRL and SPC: the spring program. Bulletin of the American Meteorological Society, 84, 1797–1806. https://doi.org/10.1175/BAMS-84-12-1797

Kain, J.S., Willington, S., Clark, A.J., Weiss, S.J., Weeks, M., Jirak, I.L. et al. (2017) Collaborative efforts between the United States and United Kingdom to advance prediction of high-impact weather. Bulletin of the American Meteorological Society, 98(5), 937–948.

Kendon, M. (2021) Storm Christoph 18 to 20 January 2021, Met Office report on past weather events. Available at: https://www.metoffice.gov.uk/weather/learn-about/past-uk-weather-events#y2021 [Accessed 3 Jan 2022].

Laroch, S. & Sarrazin, R. (2013). Impact of radiosonde balloon drift on numerical weather prediction and verification. Weather and Forecasting, 28, 772–782. https://doi.org/10.1175/WAF-D-12-00114.1

Neal, R.A., Boyle, P., Grahame, N., Mylne, K. & Sharpe, M. (2014) Ensemble based first guess support towards a risk-based severe weather warning service. Meteorological Applications, 21(3), 563–577. https://doi.org/10.1002/met.1377

Nkiaka, E., Taylor, A., Dougill, A.J., Antwi-Agyei, P., Adefisian, E.A., Ahiataku, M.A. et al. (2020) Exploring the need for developing impact-based forecasting in West Africa. Frontiers in Climate, 2, 565000. https://doi.org/10.3389/fclim.2020.565000

Porson, A.N., Clark, P.A., Harman, I.N., Best, M.J. & Belcher, S.E. (2010) Implementation of a new urban energy budget scheme in the MetUM. Part I: description and idealized simulations. Quarterly Journal of the Royal Meteorological Society, 136(651), 1514–1529. https://doi.org/10.1002/qj668

Porson, A.N., Carr, J.M., Hagelin, S., Darvell, R., Walters, D. 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/qj3844

Ralph, F.M., Intrieri, J., Andra, D., Jr., Atlas, R., Boukabara, S., Bright, D. et al. (2013) The emergence of weather-related test beds linking research and forecasting operations. Bulletin of the American Meteorological Society, 94(8), 1187–1211.

Raynaud, L. & Bouttier, F. (2016) Comparison of initial perturbation methods for ensemble prediction at convective scale. Q.J.R. Meteorol. Soc., 142, 854–866. https://doi.org/10.1002/qj.2686

Roberts, B., Jirak, I.L., Clark, A.J., Weiss, S.J. & Kain, J.S. (2019) Post processing and visualization techniques for convection-allowing ensembles. Bulletin of the American Meteorological Society, 100(7), 1245–1258.

Sassen, K., Campbell, J.R., Zhu, J., Kollias, P., Shupe, M. & Williams, C. (2005) Lidar and triple-wavelength Doppler radar measurements of the melting layer: a revised model for dark- and brightband phenomena. Journal of Applied Meteorology, 44(3), 301–312.

Sassen, K. & Chen, T. (1995) The lidar dark band: an oddity of the radar bright band analogy. Geophysical Research Letters, 22(24), 3505–3508.

Selz, T. & Craig, G.C. (2015) Upscale error growth in a high-resolution simulation of a summertime weather event over Europe. Monthly Weather Review, 143, 813–827. https://doi.org/10.1175/MWR-D-14-00140.1
APPENDIX: ESTIMATING ERROR REFRESH RATES FOR TESTBED ENSEMBLES

The speed it takes for an error to leave a model domain is termed the error refresh rate, and can be calculated based on speed, distance and time relation. Table A1 shows the error refresh rate computed for different relocatable sub-km domain sizes considered in the 2020/2021 testbed and the operational MOGREPS-UK variable resolution domain. Values are shown for different wind speeds, representative of different surface conditions or varying error refresh rates with height.

TABLE A1  Theoretical error refresh rates calculated for different domain sizes and wind speeds

| Model                              | Refresh rate (h) | $U = 5\text{ m s}^{-1}$ | $U = 10\text{ m s}^{-1}$ | $U = 40\text{ m s}^{-1}$ | $U = 50\text{ m s}^{-1}$ |
|-----------------------------------|------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Sub-km ensemble (165-km domain width) | 9.2              | 4.6                      | 1.1                      | 0.9                      |
| Sub-km ensemble (420-km domain width) | 23.3             | 11.7                     | 2.9                      | 2.3                      |
| MOGREPS-UK ensemble (925-km domain width) | 51.4             | 25.7                     | 6.4                      | 5.1                      |

Note: The refresh rates have been calculated with time = distance/speed.

The error refresh rates for typical wind speeds demonstrate why when sub-km scale domains are small, they produce limited additional spread compared with the current operational ensemble (MOGREPS-UK), despite being anticipated to better capture small-scale error structures. Raynaud and Bouttier (2016) suggested that for down-scaled convective-scale forecasts, the small-scale perturbation spin-up period is about 10 h. Shorter error refresh rates in the smaller sub-km ensemble domain indicate limited downstream influence of spread produced by the sub-km model.