IMPROVER

The New Probabilistic Postprocessing System at the Met Office

Nigel Roberts, Benjamin Ayliffe, Gavin Evans, Stephen Moseley, Fiona Rust,
Caroline Sandford, Tomasz Trzeciak, Paul Abernethy, Laurence Beard,
Neil Crosswaite, Ben Fitzpatrick, Jonathan Flowerdew, Tom Gale, Leigh Holly,
Aaron Hopkinson, Katharine Hurst, Simon Jackson, Caroline Jones, Ken Mylne,
Christopher Sampson, Michael Sharpe, Bruce Wright, Simon Backhouse, Mark Baker,
Daniel Brierley, Anna Booton, Clare Bysouth, Robert Coulson, Sean Coultas,
Ric Crocker, Roger Harbord, Kathryn Howard, Teresa Hughes, Marion Mittermaier,
Jon Petch, Tim Pillinger, Victoria Smart, Eleanor Smith, and Mark Worsfold

ABSTRACT: The Met Office in the United Kingdom has developed a completely new probabilistic postprocessing system called IMPROVER to operate on outputs from its operational numerical weather prediction (NWP) forecasts and precipitation nowcasts. The aim is to improve weather forecast information to the public and other stakeholders while better exploiting the current and future generations of underpinning kilometer-scale NWP ensembles. We wish to provide seamless forecasts from nowcasting to medium range, provide consistency between gridded and site-specific forecasts, and be able to verify every stage of the processing. The software is written in a modern modular framework that is easy to maintain, develop, and share. IMPROVER allows forecast information to be provided with greater spatial and temporal detail and a faster update frequency than previous postprocessing. Independent probabilistic processing chains are constructed for each meteorological variable consisting of a series of processing stages that operate on predefined grids and blend outputs from several NWP inputs to give a frequently updated, probabilistic forecast solution. Probabilistic information is produced as standard, with the option of extracting a most likely or yes–no outcome if required. Verification can be performed at all stages, although it is only currently switched on for the most significant stages when run in real time. IMPROVER has been producing real-time output since March 2021 and became operational in spring 2022.

KEYWORDS: Ensembles; Forecasting techniques; Operational forecasting; Probability forecasts/models/distribution; Postprocessing

AFFILIATIONS: Roberts, Ayliffe, Evans, Moseley, Rust, Sandford, Trzeciak, Abernethy, Beard, Crosswaite, Fitzpatrick, Flowerdew, Holly, Hopkinson, Hurst, Jackson, Jones, Mylne, Sampson, Sharpe, Wright, Backhouse, Baker, Brierley, Booton, Bysouth, Coulson, Coultas, Crocker, Harbord, Howard, Hughes, Mittermaier, Petch, Pillinger, Smart, Smith, and Worsfold—Met Office, Exeter, United Kingdom; Gale—Bureau of Meteorology, Melbourne, Victoria, Australia

"Postprocessing" is the term used here to describe the additional processing that is applied to output from numerical weather prediction (NWP) forecasts. It is an essential component of an optimal NWP forecast system and various types of postprocessing are applied to NWP output in national meteorological services (Hamill et al. 2017; Ylinen et al. 2020). It has two principal aims. The first is to improve the skill of the forecasts using statistical (Wilks 2011; Vannitsem et al. 2021) or physical methods (Moseley 2011; Howard and Clark 2007; Sheridan et al. 2010) to adjust for errors or deficiencies in the NWP forecasts. The second is to exploit the information contained in the forecasts in a more useful way for the users of the forecasts, which may include deriving new outputs, such as "feels-like" temperature (Steadman 1984; Osczevski and Bluestein 2005, 2008) or blending forecasts or utilizing ensemble information. To successfully meet these two aims, postprocessing must continually adapt to changes in NWP modeling and user requirements.

Recent years have seen a shift to kilometer-scale NWP and ensembles, which has changed the requirements for postprocessing. A clear benefit of going to finer resolution is that gridcell values become more equivalent to point locations leading to a reduction in representativeness errors (Göber et al. 2008; Ben Bouallègue et al. 2020). If the grid is sufficiently fine, it makes more sense to utilize the full gridded output directly rather than postprocess for selected locations and interpolate to other locations. The latter has been the preferred option for coarser resolutions. However, although representativeness and realism are improved, the unpredictable nature of small scales means that explicitly represented weather elements such as convective showers are very often misplaced in space and time, even in short-range forecasts (Mass et al. 2002; Roberts and Lean 2008; Clark et al. 2010). A forecast will often be wrong in detail even if the weather is generally well represented. As a result, kilometer-scale ensembles have been introduced by national meteorological services to use probabilities to account for local uncertainty (Gebhardt et al. 2011; Clark et al. 2016; Hagelin et al. 2017; Clark et al. 2018; Zhang 2018; Porson et al. 2020; Bouttier and Marchal 2020). If an ensemble is not available, spatial methods such as neighborhood processing (Roberts 2003; Theis et al. 2005; Schwartz and Sobash 2017) can be used to generate probabilities instead. However, even if a kilometer-scale ensemble is available, the number of members is restricted by computational cost and the small-scale uncertainty is still undersampled; therefore, neighborhood methods remain essential to address that undersampling (Clark et al. 2011; Golding et al. 2016; Schwartz and Sobash 2017; Roberts et al. 2019).

High-resolution models/ensembles are typically run for shorter forecast lengths than their coarser-resolution counterparts because of computational expense, which creates a jump in resolution during the forecast period. To transition seamlessly from one forecast to another, requires suitable blending, possibly including prior calibration of each forecast. Blending should involve minimal loss of forecast information, which is most effectively

achieved by blending probability distributions rather than physical quantities, lending further weight to a probabilistic approach to postprocessing.

Thinking more widely than purely NWP performance, the proliferation of frequently updated high-resolution forecasts may result in information overload for operational meteorologists, leaving the possibility of key signals being missed. Unless the most salient information can be presented at a glance, some ensemble information is likely to be ignored. It is therefore helpful to consolidate the large number of forecasts into a single probability distribution. A further challenge is adaptation to the way forecasts are disseminated and used in the modern world. Weather forecasts were once provided via television, radio, and newspapers for general geographical areas every few hours. Nowadays the public expect hourly forecasts (at least) for their location of interest on their mobile phones, along with timely warnings of severe weather. The Met Office is required to produce frequently updated automated weather forecasts for any location in the United Kingdom along with warnings of severe weather via operational meteorologists. There is also a requirement for targeted outputs for specific customers and for forecasts around the globe.

In the Met Office, postprocessing has evolved as needs have arisen, which has meant independent systems being developed to meet the needs of different users or to exploit new or evolving NWP outputs. Our previous postprocessing generated automated weather forecasts for a large set of site locations. These were produced independently of gridded outputs for meteorologists and the media, which were separate again from ensemble processing. Running separate systems in that way is not only expensive, in terms of maintenance and duplication of effort, but leads to inconsistencies that can confuse the forecast message.

We are served better by having a single postprocessing framework that allows new postprocessing methods to be added as the science evolves. Such a framework needs a software infrastructure that enables code to be easily plugged in or taken out and can adapt to changing IT environments, such as the need for highly parallel processing. The postprocessing should operate primarily on the grid, be fully probabilistic, and be able to employ both spatial neighborhood and calibration methods. It should also be capable of handling both deterministic and ensemble forecasts with differing resolutions, and seamlessly blend them together. A verification capability is required to measure the benefit of each individual component.

Based on the reasoning above, a strategy for future postprocessing in the Met Office was endorsed by the Met Office Scientific Advisory Committee (MOSAC) in late 2015. The new probabilistic postprocessing system called IMPROVER (Integrated Model post-PROcessing and VERification) has been under development at the Met Office since 2016 and an operationally supported demonstration system was introduced in March 2021. IMPROVER became operational in April 2022.

The rest of this article will describe IMPROVER as configured for postprocessing in the Met Office before outlining the generic processing steps and some of the methods used.

What is IMPROVER?

IMPROVER is the new probabilistic postprocessing system running operationally at the Met Office, in which postprocessing steps are applied in sequence to individual nowcasts and NWP forecasts on a common grid before blending to provide a seamless probabilistic forecast for each processed variable (Fig. 1). IMPROVER ingests forecasts as soon as they are available and is designed to exploit the latest high-resolution models and ensembles. It is comprised of an open-source Python code repository (Evans et al. 2020) that contains a collection of postprocessing algorithms. The postprocessing algorithms are run within an internal "suite" workflow orchestration infrastructure (Oliver et al. 2018, 2019). This chains together the postprocessing tasks and schedules when they are run and allows both a real-time running configuration and trial mode assessments. The configuration discussed in this article only
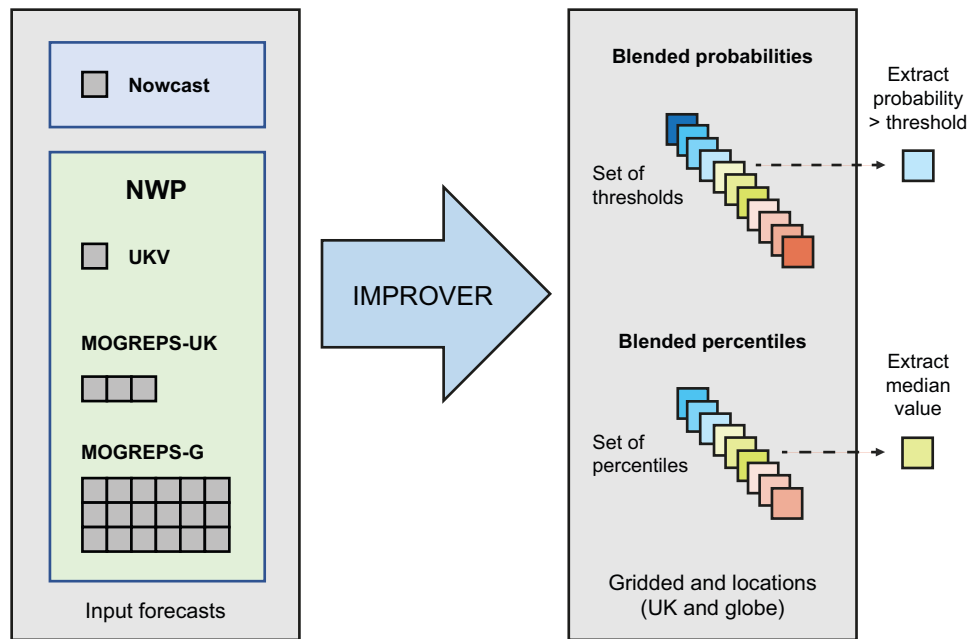
Fig. 1. Diagram to depict how IMPROVER ingests nowcasts and NWP forecasts and produces a set of probabilities and percentiles for each variable from which single values such as the median or 90th percentile can be extracted to give a best estimate or plausible high value. See Table 1 for information about input forecasts.

operates on outputs from the forecast models run at the Met Office, but forecasts from external models can be included.

A set of key principles has been followed when developing the code and constructing the suites:

1) A flexible modular framework with independent processing chains is used for each variable.
2) Verification is incorporated at every processing stage.
3) Processing is done on common grids with forecasts at site locations extracted at the end, ensuring consistency in forecast outputs.
4) Meteorological variables are represented and postprocessed as probability distributions.
5) New forecasts are processed as they become available and forecasts from different models are probabilistically blended to provide a rapidly updating seamless forecast.
6) Outputs are probabilistic by default, consisting of a set of probabilities and percentiles, from which single outcomes can be extracted if needed.

The modular framework provides the flexibility to apply physical and statistical adjustments in the order that best suits each variable. It is highly adaptable, enabling it to orchestrate the processing across source models, variables, and lead times as required. For instance, each chain groups together a number of steps, which correspond to plug-ins in the open-source repository. These chains can then be reused for multiple lead times and models, and the steps can be combined in different ways to make up new processing chains. When a new step/chain is added, a pull request is set up on the suite repository and a review process operates in line with the open-source repository policy of having at least two reviewers before any code is merged. Each pull request made to either repository passes through the continuous integration (CI) workflow, which runs unit and acceptance tests and, in the suite repository, runs through the full graph orchestration for every model. Additionally, each night, all the suites are run for 6 h using code in the master branch and any errors are recorded in a dashboard.

Postprocessing stages

When configured for operational postprocessing at the Met Office, each of the postprocessing chains follows a sequence of processing stages, shown by Fig. 2, that are applied to each individual meteorological variable. A sequential approach to postprocessing has also been applied in other centers (Kober et al. 2012; Bouttier and Marchal 2020). The following sections will outline the purpose of each stage and briefly describe some of the postprocessing methods used within them.

The IMPROVER postprocessing stages shown in Fig. 2 will now be described.

Inputs. IMPROVER takes as input gridded NWP forecasts from the UKV (U.K. deterministic model) (Tang et al. 2013), MOGREPS-G (global ensemble) (Inverarity et al. 2023), and MOGREPS-UK (U.K. ensemble) (Hagelin et al. 2017; Porson et al. 2020), as well as gridded observations-based precipitation nowcasts using optical flow (Bowler et al. 2004), as described in Table 1. Further NWP configurations or models can be added in future.

Fig. 2. Flow diagram of the IMPROVER processing stages applied to each variable.

All the forecasts used by IMPROVER have first been standardized to a predetermined set of grids and forecast times. This acts to decouple the NWP models from downstream systems allowing both to be upgraded independently. The grids and forecast times were pragmatically determined by choosing resolution, domain and forecast times close to the NWP outputs that are manageable in terms of data volumes and processing speed. At present, two different map projections are used to cover the globe and the U.K. area. The global grid is an equirectangular latitude–longitude projection with a grid spacing of approximately 20 km over midlatitudes. The U.K. grid uses a Lambert azimuthal equal-area (LAEA) projection with a uniform grid spacing of 2 km. IMPROVER needs to regrid the coarser-resolution MOGREPS-G ensemble forecasts to the 2 km U.K. grid, to support subsequent blending between models for U.K. forecasts. This is done using nearest-neighbor interpolation and includes the option to choose the nearest point of matching surface type to be coastline aware.

New variables and physical adjustments. Any new variables required by users or needed for physical adjustments of other variables are computed for each ensemble member. This includes parameters such as the temperature lapse rate, the feels-like temperature, cloud texture diagnosis, and the snow melting level. We also introduce a common unsmoothed orography, which is closer to the true orography than the smoothed versions used in the NWP models, allowing the same adjustments across models. Physical adjustments are then applied to improve realism or accuracy which are distinct from later statistical adjustments
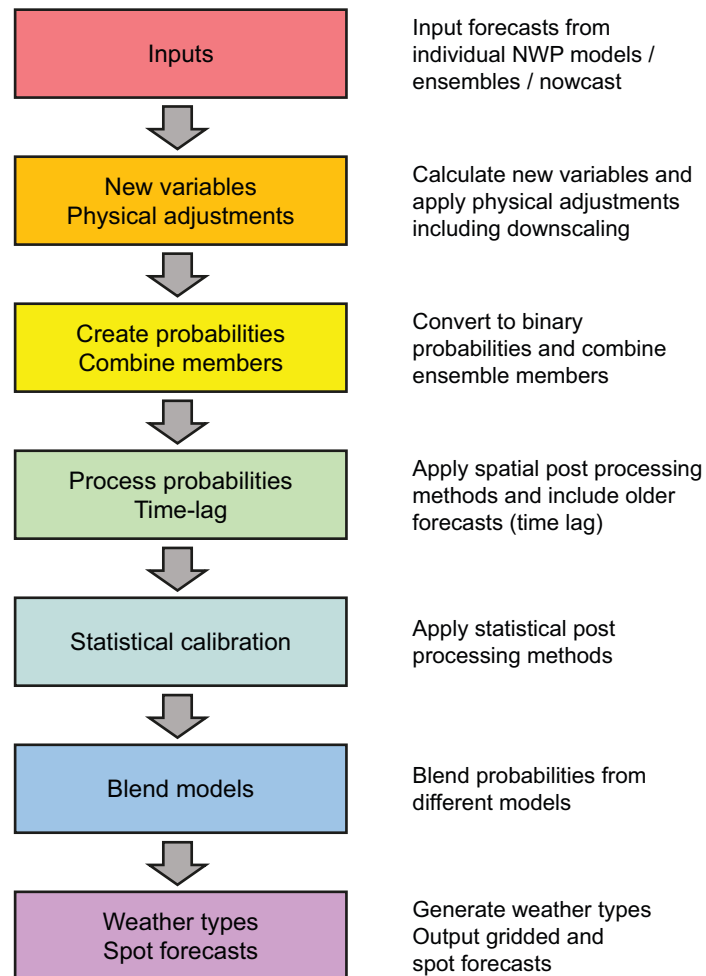
Table 1. Table showing the models providing input data for IMPROVER. Note that there are some longer UKV forecasts not used.

Models (end of 2022)	Grid spacing	Forecast frequency	Members	Forecast length
Radar extrapolation (precipitation only)	2 km	15 min	1	6 h
UKV (U.K. deterministic)	Raw 1.5 km	1 h	1	12 h
	Reprojected 2 km			
MOGREPS-UK (U.K. ensemble)	Raw 2.2 km	1 h	3	5 days
	Reprojected 2 km			
MOGREPS-G (global ensemble)	Raw ~20 km	6 h (0000, 0600, 1200, 1800 UTC)	18	7 days
	Reprojected ~20 km			

in that they follow physical principles rather than learned statistical relationships. At present we have the capability to adjust the temperature, wind speed, wind gust, and precipitation phase (rain, mixed phase, snow). The adjustments of temperature and precipitation phase are briefly outlined below.

TEMPERATURE ADJUSTMENT. To match the unsmoothed orography, temperatures are adjusted according to altitude on the U.K. grid using the local lapse rate (Sheridan et al. 2010). This entails calculating a local mean vertical temperature gradient at each point in each forecast by sampling all data points in a seven-gridpoint radius and computing the regression of the variation of 1.5 m temperature with altitude. This is constrained to be within the dry adiabatic lapse rate (-0.01 K m^{-1}) and an inverted rate 3 times greater ($+0.03$ K m^{-1}). The temperature gradient is then used to adjust the 1.5 m temperature to the unsmoothed orography.

PRECIPITATION PHASE ADJUSTMENT. IMPROVER outputs probabilities of the precipitation phases; rain, mixed phase, and snow, requiring knowledge of what the phase would be at each grid square even where there is no precipitation. Because that information is not directly available from the forecasts, it needs to be derived.

We have implemented a scheme that finds the altitude of the phase change levels from snow to mixed phase and mixed phase to rain based on the effective melting depth defined as an integral of the difference between the wet-bulb temperature and the melting point over atmospheric depth. These levels are used to produce continuous surfaces across the postprocessing domain from which probabilities of rain, mixed phase, and snow are derived.

Create binary probabilities and combine members. Once the necessary adjustments have been made to the physical quantities in each forecast, we convert to probabilities. A cumulative probability distribution is constructed, by obtaining binary probabilities (1 or 0) of exceeding a set of thresholds for each variable, at every grid square and every forecast time. For temperature, we set thresholds at 1 K intervals over most of the temperature range and for wind speed it is every 1 m s^{-1}. For precipitation, the thresholds are spaced logarithmically (powers of 2), with the addition of a few high-impact values.

For some variables, we also generate binary probabilities based on whether a threshold is exceeded within a specified vicinity of each grid square. This provides the binary neighborhood probability (BNP) for each threshold defined by Schwartz and Sobash (2017).

The ensemble probabilities are then computed in the conventional way by calculating the mean of the binary probabilities from each ensemble member (hence creating an ensemble cumulative distribution). The same can be done to the BNPs. When applied to the BNPs, the neighborhood maximum ensemble probability (NMEP) is obtained

(Schwartz and Sobash 2017), which can be thought of as the probability of an occurrence within a square vicinity of each grid square. The NMEP is more appropriate for rarer, more localized, or extreme weather, for which gridscale probabilities may be very low. The probabilities for the nowcast and UKV remain binary at this stage because they are deterministic forecasts.

Process probabilities including neighborhood methods and time-lagging. Various forms of spatial and temporal neighborhood processing, spatial filtering, and time-lagging are applied to the probabilities produced in the previous step.

The basic neighborhood processing (Roberts 2003; Theis et al. 2005; Roberts and Lean 2008; Schwartz et al. 2010; Clark et al. 2011; Schwartz and Sobash 2017) generates a new probability at each grid square by computing the mean of the probabilities within a square neighborhood of predefined length (km) surrounding each grid square. Schwartz and Sobash (2017) refer to this as the neighborhood probability (NP) when applied to the binary probabilities from a single forecast, or the neighborhood ensemble probability (NEP) when applied to ensemble probabilities. When considering a deterministic forecast, the purpose of neighborhood processing is to account for positional error brought about by the lack of predictability at small scales. For an ensemble, it effectively increases the ensemble size, enabling a reduction of the undersampling and noisiness at small scales that might otherwise give false confidence in small-scale forecast outcomes. Given an appropriate neighborhood size, the skill of both deterministic and ensemble forecasts can be improved (Theis et al. 2005; Roberts and Lean 2008; Clark et al. 2011; Ben Bouallègue and Theis 2014). Variables with a typically higher spatial correlation, such as temperature and wind speed, need smaller neighborhoods than less spatially correlated variables such as precipitation, cloud, visibility, and lightning. Highly related variables such as cloud and precipitation require the same sized neighborhood. The same mean-in-neighborhood method can also be applied to the NMEPs to provide smoothing and account for larger-scale positional uncertainty if the NMEP neighborhood is small.

The probabilities of precipitation falling as rain, mixed phase, or snow are found by multiplying the neighborhood processed probabilities of precipitation by the conditional probabilities of rain, mixed phase, or snow at the grid-square altitude derived from the phase change levels. This multiplication is applied to each individual member of an ensemble to retain the multivariate dependency between wet-bulb temperature and precipitation in each realization, such as the simulated relationship between the melting level and precipitation intensity. Consequently, the neighborhood processing of precipitation, unlike for most other variables, is performed on each member individually before the ensemble probabilities are computed.

EXTENSIONS TO SPATIAL NEIGHBORHOOD PROCESSING. The mean-in-neighborhood calculation assumes that an occurrence is equally likely anywhere in the neighborhood, which may not be appropriate for variables like temperature, visibility, and wind speed that vary considerably with altitude. Hence, we have introduced a topographic neighborhood approach that gives more weight to grid squares of similar altitude to the central grid square to make the neighborhood more representative. Figure 3 shows the effect of incorporating topography into the neighborhood processing.

Square neighborhoods are used for computational speed and simplicity. A weighted circle or Gaussian kernel may be more physically justifiable, but they are much slower to compute. As an alternative, we introduce a recursive filter (Hayden and Purser 1995; Roberts 2003) in addition to the square neighborhood. This is fast and the result is similar to applying a Gaussian kernel for smoothing. Further topographic awareness can be incorporated by applying less filtering across steep topographic gradients or across coastlines. The result is a Gaussian-kernel alternative that does not smear out topographic variations and can be used in conjunction with the topographically weighted NEP neighborhood.
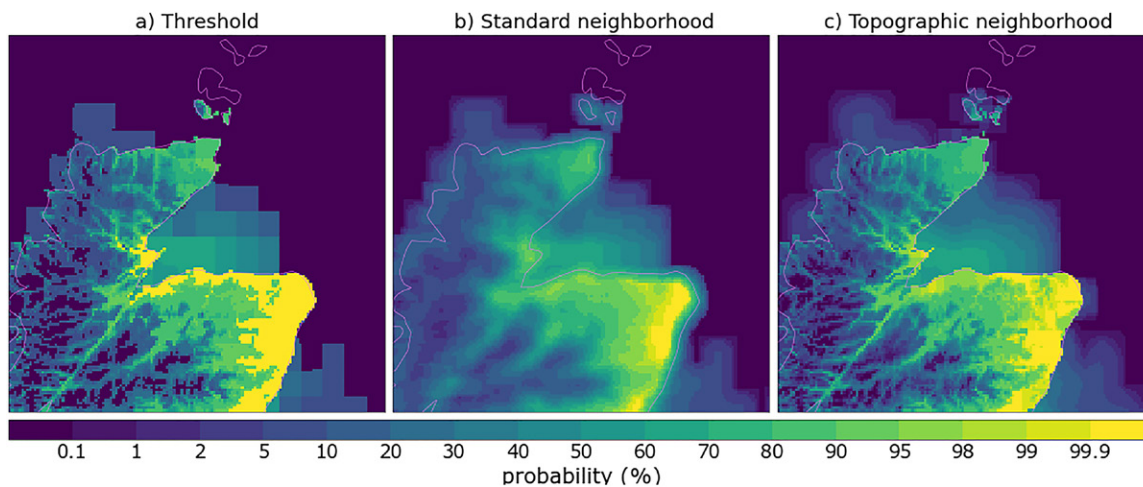
Fig. 3. An example of neighborhood processing (NEP) for a temperature threshold with and without the inclusion of topography. (a) Probabilities after downscaling MOGREPS-G to the 2 km U.K. grid, (b) the effect of standard neighborhood processing smearing out topographic variation, and (c) the ability of the topographic neighborhood to retain topographic variation while smearing out blockiness elsewhere.

TEMPORAL NEIGHBORHOOD PROCESSING. Probabilistic forecasts should take account of the forecast uncertainty in time as well as space (Theis et al. 2005; Duc et al. 2013). We have introduced the capability to include temporal neighborhoods (time windows) that are analogous to their spatial counterparts and can be implemented sequentially alongside them. The time window symmetrically spans the time of interest and is used to update the probability at the central point by taking the average over the probabilities in the time window, with the option of linearly weighting more toward the center. The maximum value within a time window can also be found, providing the probability of occurrence sometime within the time window. Different neighborhood methods or combinations provide different time–space definitions to the resulting probabilities, which must be carefully defined from the outset.

TIME-LAG BLENDING. Time-lagging is the process of blending forecasts with different initiation times. It has been shown to improve the skill of both deterministic and ensemble forecasts (Mittermaier 2007; Osinski and Bouttier 2018) and reduces "flip-flopping" (Griffiths et al. 2019) between forecast cycles. It is essential for MOGREPS-UK, which was designed as a time-lagged ensemble constructed of several cycles of three members per hour (Porson et al. 2020). Time-lagging is also applied to the UKV and MOGREPS-G. It can be applied to the full set of probabilities before or after neighborhood processing because these steps are commutative. Options are available for applying different weights to each forecast length, although at present, equal weighting is used.

Statistical calibration. Statistical calibration uses information about the performance of past forecasts to try to correct systematic errors in the current forecast (Wilks 2011; Hamill and Scheuerer 2018; Vannitsem et al. 2021). These corrections aim to improve characteristics such as bias, ensemble spread, or the reliability component of an ensemble forecast (Richardson 2000). Applying these corrections to the full ensemble, rather than individual ensemble members, avoids artificially deflating ensemble spread (Gneiting 2014). Ensemble calibration is applied after process probabilities and time-lagging to utilize the resulting fuller probability distribution. Two ensemble calibration methods have been introduced so far and are described below. These calibrations are performed on the grid using UKV analyses as the best available gridded truth. Residual errors arising from the UKV analysis not being a perfect "truth" will therefore remain after calibration (Feldmann et al. 2019; Allen et al. 2021a). The statistical calibration uses historic forecasts that have undergone the

same temporal and spatial processing as the forecasts being corrected. Similarly, the truth is representative of the forecast, so the meaning is not altered.

EMOS. EMOS, otherwise known as nonhomogeneous regression (NR), following Gneiting et al. (2005), is a widely used technique for calibrating ensemble forecasts using the assumption that the future state of a weather variable can be represented using a single parametric distribution. It involves estimating four coefficients, when using the ensemble mean by minimizing or maximizing a scoring rule such as the Continuous Ranked Probability Score (CRPS; Gneiting et al. 2005), along with an appropriate assumption for the distribution of the variable. EMOS has been applied to MOGREPS-UK temperature and wind speed forecasts using a normal (Gneiting et al. 2005) and truncated normal (Thorarinsdottir and Gneiting 2010) distribution, respectively. For gridded forecasts, EMOS is currently configured to operate on all grid points at once or use a split between land and sea based on ancillary information. This land–sea localization yields better results as demonstrated in Fig. 4, showing the potential benefit of further localization (Thorarinsdottir and Gneiting 2010). We use a rolling training period of 15 days, which can be adjusted. The ability to use a short training period is a benefit of operating on the large number of gridded values and applying the same coefficients to large groupings of points. The predictors used for EMOS are the ensemble mean and variance, which are extracted from the cumulative distribution function defined by the probability forecasts. The calibrated distribution produced by EMOS is sampled at a set of thresholds to produce calibrated probability forecasts.

RELIABILITY CALIBRATION. We have also introduced the reliability calibration technique presented in Flowerdew (2014), similar to Zhu et al. (1996) for precipitation and cloud. In this technique, an aggregated "reliability table" is populated that relates the forecast probabilities to the truth event frequencies for each grid square, lead time, threshold, and forecast probability bin over a sample of previous forecasts. The objective is to adjust the future forecast probabilities so that they agree with the previous event frequencies. For example, whenever the forecast probability of an event is 60%–70% the event should have occurred 60%–70% of the time. If it only occurred 30%–40% of the time in the training data, the future forecast probabilities should be adjusted accordingly. Treating each lead time separately facilitates the capturing of diurnal variability with the expectation that the reliability tables vary relatively smoothly between lead times. We use a rolling training period of 30 days, with a short training period favored from an operational perspective. As in Flowerdew (2014),
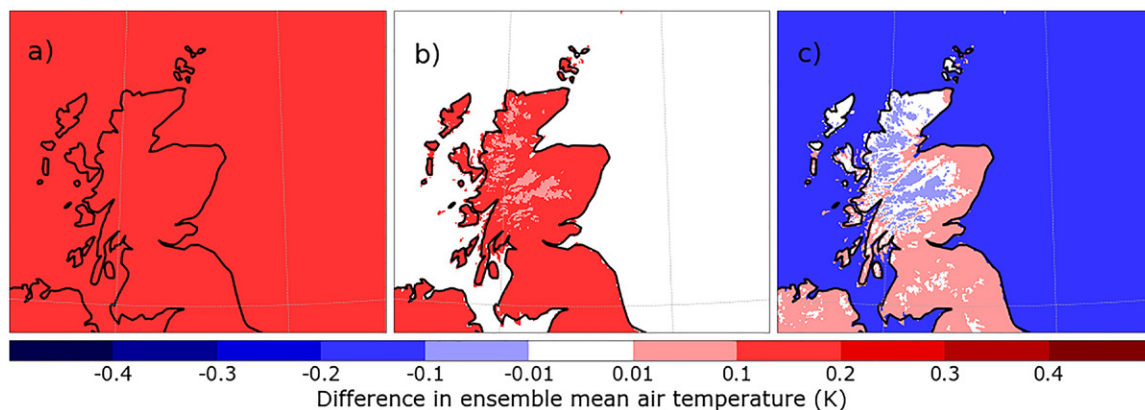

Fig. 4. Example of differences in the ensemble-mean screen temperature for a reduced area around Scotland when EMOS is computed using (a) all grid points within the domain (EMOS-all) and (b) only land points within the domain (EMOS-land only), and (c) the difference between the two.

a minimum sample count for the forecast probability bins prevents calibration being applied for events at extreme, poorly sampled thresholds.

Reliability calibration is a conceptually simple, nonparametric technique that is adept at correcting variables with particularly non-Gaussian distributions. Since it naturally operates in probability space, it is a particularly useful correction technique for probabilistic forecasts that can be used in conjunction with neighborhood processing (Kober et al. 2012; Johnson and Wang 2012; Bouttier and Marchal 2020). Improvement to the reliability of cloud forecasts is demonstrated in Fig. 5.

Since reliability calibration makes the probability distributions from two different models/ensembles more reliable and hence more alike, the later blending between them becomes more seamless. This is particularly beneficial when two ensembles have very different behaviors such as ~2 km MOGREPS-UK and the ~20 km MOGREPS-G.
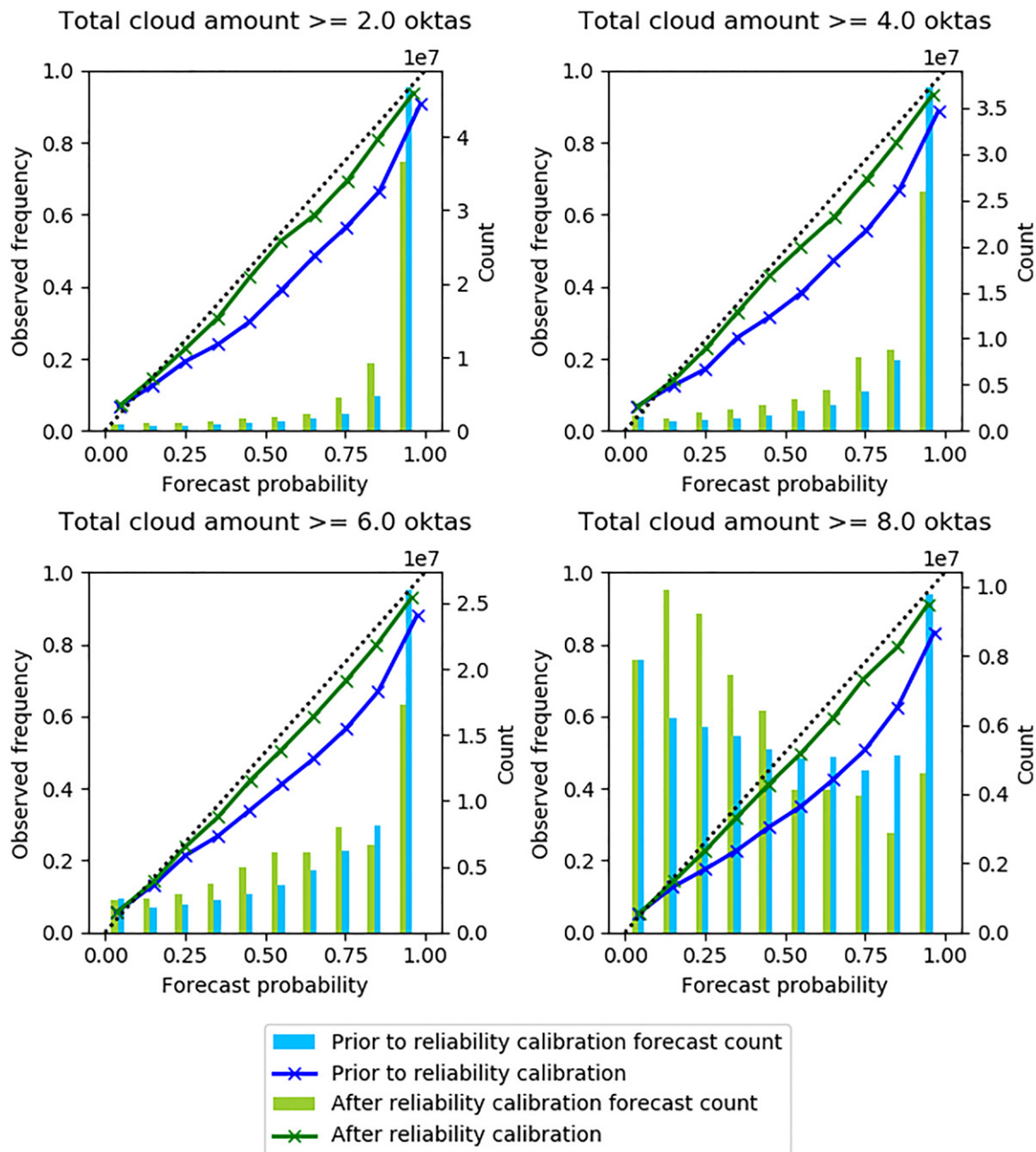

Fig. 5. An example of reliability diagrams prior to and after the application of reliability calibration for NEP total cloud amount forecasts from MOGREPS-G for *T* + 48 h for a selection of thresholds. The verification period is February and March 2020. Lines closer to the diagonal show better reliability.
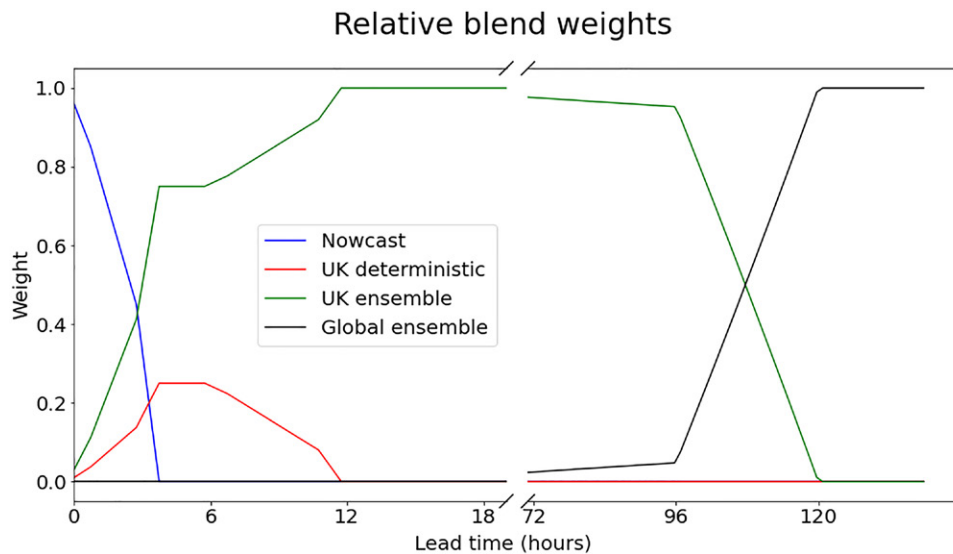
Fig. 6. Schematic of the blending weights at time of writing. The global ensemble starts being introduced from $T + 48$ h (not shown).

Model blending. After calibration, the models/ensembles are blended. This is done to improve skill (Beck et al. 2016) and to allow the final blended forecast to transition between different models or ensembles. The blending is done using probabilities to transition seamlessly between different models used for different forecast periods and preserve the full distribution of physical values that may otherwise be averaged out. Figure 6 shows the forecast periods over which the models/ensembles are currently blended on the U.K. domain, along with the weights applied. The weights are chosen to represent the relative skill of each model for key variables (precipitation, temperature, wind speed and direction, cloud cover, and visibility) and to seamlessly deal with the transitions between available models imposed by different forecast horizons. These weights can be adjusted as new models are included, or as model skill evolves. The weights are given by $P_{blended} = w_{m1}P_{m1} + w_{m2}P_{m2} + \cdots + w_{mn}P_{mn}$, where $P_{blended}$ is the blended probability, w_{m1} is the weighting for model 1 applied to P_{m1} (probability for model 1) through to w_{mn} (weight for model n) applied to P_{mn} (probability for model n). The weights add up to 1 such that $w_{m1} + w_{m2} + \cdots + w_{mn} = 1$.

For precipitation, we combine a radar-based extrapolation nowcast with the UKV and MOGREPS-UK over the first few hours. Each has had NEP neighborhood processing applied, with a smaller neighborhood size used for the radar extrapolation at very short lead times because of its greater spatial accuracy. The benefit of blending probabilities rather than physical values is schematically demonstrated in Fig. 7 along with an example of the nowcast blending. Later in the forecast MOGREPS-UK is blended with downscaled and calibrated MOGREPS-G. The combination of neighborhood-processed, time-lagged, and blended probabilities reduces spurious jumpiness from one time to the next and overconfidence in the weather at a particular place.

Here, the nowcast is weighted the highest giving the sharpest blended probabilities in the areas where it is available and this seamlessly blends into the other parts of the domain where the MOGREPS-UK ensemble probabilities dominate. For other lead times, a similar blend is produced using the weights in Fig. 6.

Weather symbols and site extraction. Weather symbols are a prominent feature of automated weather forecasts provided to the public. Their purpose is to succinctly communicate the important weather affecting a location (Reed and Senkbeil 2020, 2021). IMPROVER uses a decision tree to generate a most probable weather type (e.g., sunny or heavy rain shower) from the probabilities for every 2 km grid square over the United Kingdom, and each grid cell
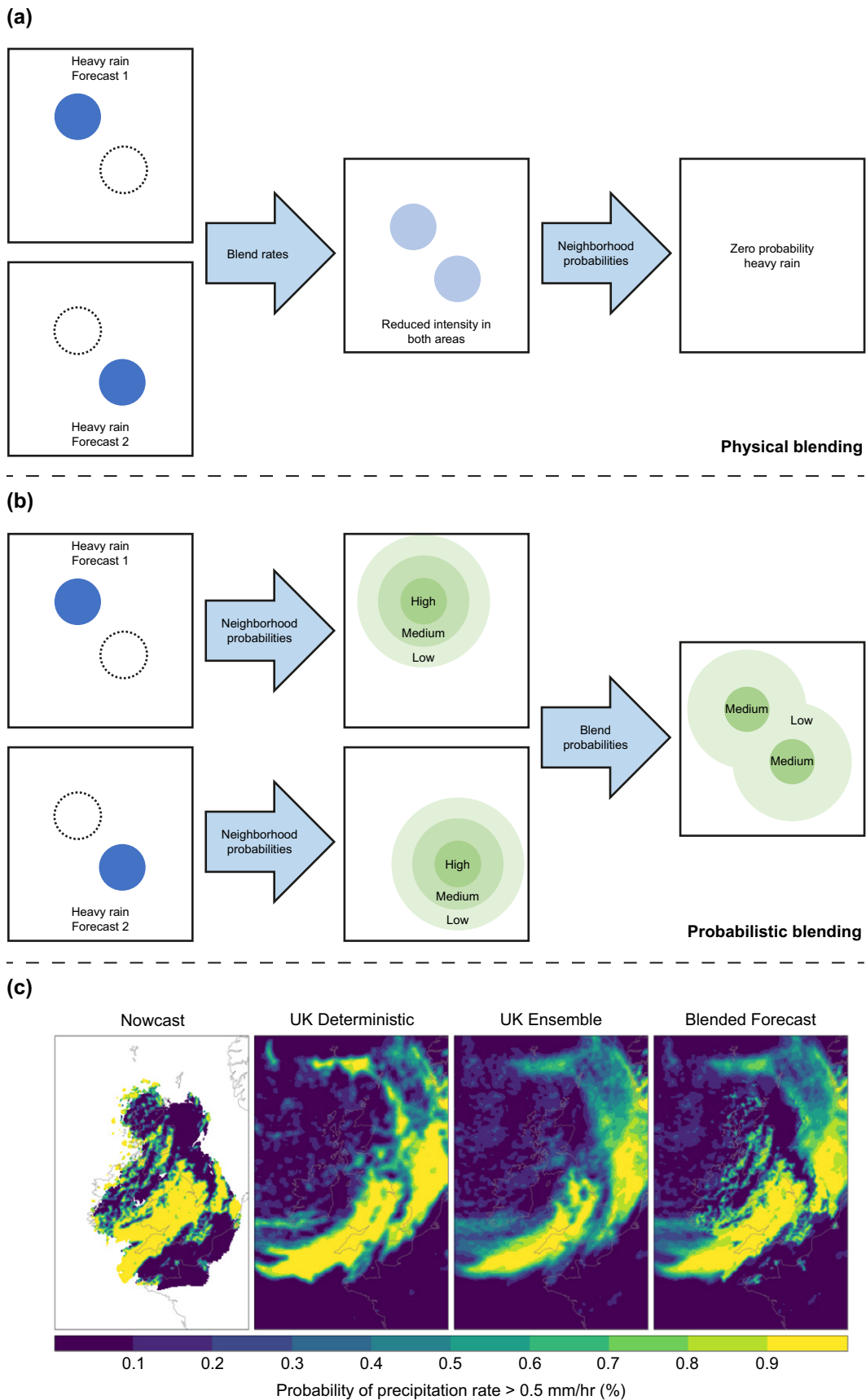
Fig. 7. (a) A schematic depiction of how two forecasts with heavy rain in different locations (dark blue areas) can lead to reduced rainfall intensity when blended giving a zero probability of heavy rain. (b) When the heavy rain is turned into probabilities (described as high, medium, and low) and then blended, probabilities of heavy rain are retained. (c) A real example of blending the most recent NEP precipitation probability fields for a given validity time. The rightmost panel shows the resulting blended forecast produced by IMPROVER when combining all three available forecasts for a *T* + 1 h lead time.
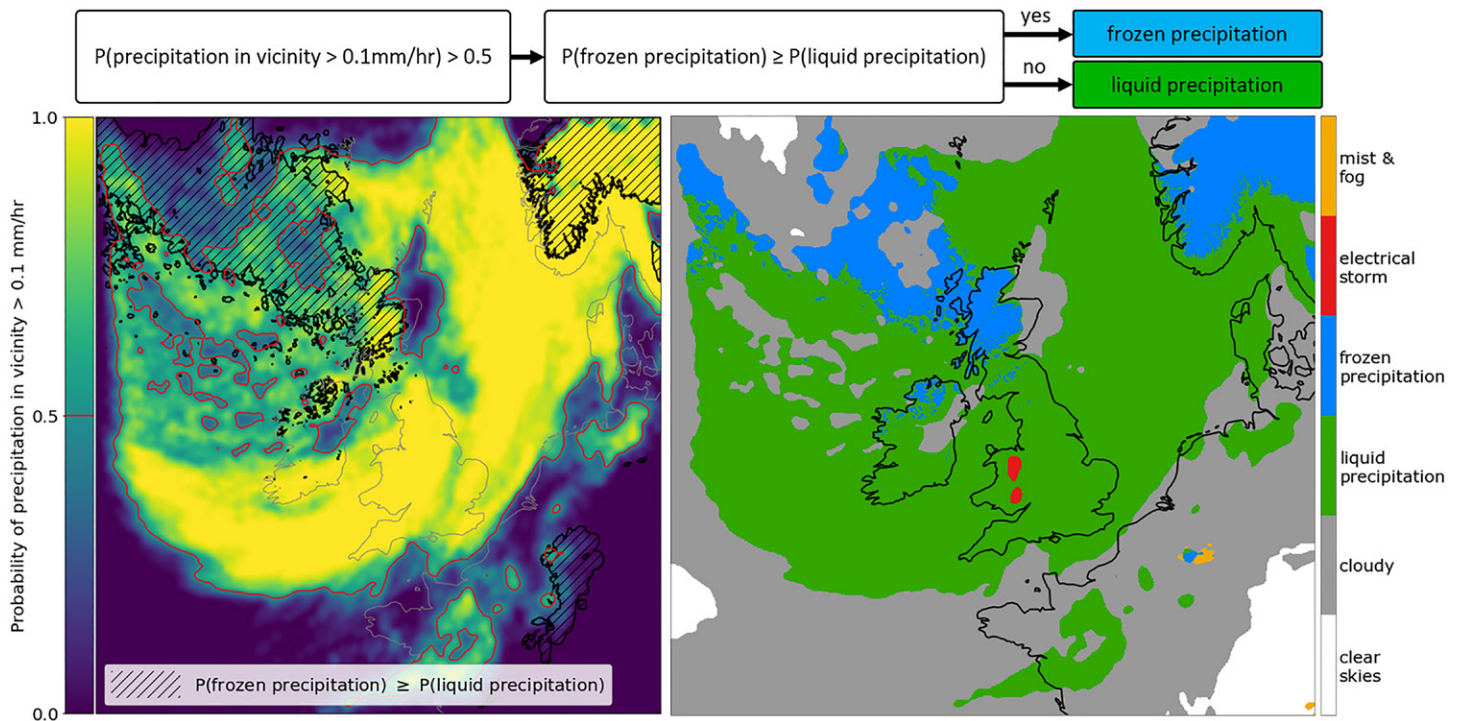
Fig. 8. (left) The model blended probability of precipitation in a 10 km vicinity (NMEP) exceeding 0.1 mm h⁻¹, with the 50% probability contour shown as a red dashed line. The hatched area shows where the probability of frozen precipitation (sleet or snow) is higher than the probability of rain. The probabilities provide inputs to a decision tree, as depiction in a simplified way above the panels, to determine the weather type shown in the right-hand panel. In this example, the areas inside the hatched region and red contour have a frozen precipitation weather type; elsewhere within the red contour the weather type is liquid precipitation. (right) A subset the 32 available classifications for ease of visualization. The other weather types such as cloud, fog, and lightning are also derived within the larger decision tree from the relevant probabilities.

over the globe. The decision tree compares the probabilities of different weather elements to arrive at the weather type that best represents each combination of probabilities. These can be displayed as weather symbols. This generation of weather types directly from the grid enforces consistency across spatial areas without the need for interpolation. An example of the gridded weather types is shown in Fig. 8. They are verified by comparison with SYNOP weather codes, allowing for refinement to help improve the forecast.

A deficiency of traditional weather symbols is that they only provide a single symbol showing one type of weather without alternative possibilities. The use of probabilities will make it possible in future to generate multiple weather symbols, with for example, primary and secondary symbols used to describe the most likely weather (e.g., sunny) along with other less likely possibilities (e.g., a thunderstorm).

Site extraction. To ensure a consistent forecast is presented in both gridded and site-specific products, site-specific forecasts are generated at the end of the chain from the gridded probabilities, percentiles and weather types. IMPROVER utilizes the intelligent gridpoint selection (IGPS) technique (Moseley 2011) by choosing the most representative grid square from either; the nearest grid square, the nearest land-point grid square, the nearest minimum-height-error grid square, or the minimum-height-error land-point grid square. The nearest grid square is the most appropriate for variables less directly influenced by the topography such as cloud and precipitation. The other options are more suitable for more topographically constrained variables such as visibility and temperature.

Temperature adjustments are made to account for the altitudes of the spot locations using the lapse rates derived on the grid. Additionally, EMOS as described in the statistical calibration section, is applied to calibrate spot temperatures. This includes an additional

predictor within EMOS to account for the difference in altitude between the nearest grid point and the site location. The use of both lapse rate and EMOS corrections gives more accurate temperatures for hilltops or valleys that are not well resolved even on the 2 km grid. Altitude adjustment is not yet used for other variables, with precipitation type the highest priority for future work.

Verification and trialing

IMPROVER incorporates verification at every stage for both real-time running and examining periods in the past. Verification allows components to be tuned, alternative methods to be compared, and priorities for future work identified. The verification of each science component means that they can be independently assessed and optimized, retaining only those that improve forecast performance. At present, the verification supports a range of standard ensemble verification metrics computed within the Met Office verification system (VER). Additional deterministic measures of forecast performance at site locations are also included. Figure 9 gives examples of verification graphs showing changes in skill coming from the use of a temporal neighborhood and different calibration methods. Several trials have shown that the steps we currently include improve forecast skill to differing degrees depending on the variable. Before initial operational implementation, we set a condition that, for the majority of locations, the final IMPROVER outputs should at least match the previous postprocessing. We do not provide comprehensive verification results here because of the limited space.

Usage and future plans

Usage. The initial implementation of the IMPROVER system has been designed with generic functionality to meet a range of user requirements. The first operational release has targeted two major generic use cases: to provide seamless forecasts of surface weather for automated public forecasts (as displayed on the Met Office website and app), and to deliver probabilistic detection of hazardous or extreme weather for operational meteorologists. The first of these will be met when IMPROVER feeds into the automated forecasts provided to the public. This will initially require that a single value such as the median is taken from the probabilistic information to give a most likely outcome to ensure compatibility with existing feeds, which can then evolve into a greater use of the probabilistic information. The second use case is an ongoing process involving dialogue with meteorologists and others interested in hazard management to request and develop IMPROVER outputs to best meet their needs. An internal testbed in 2022 allowed IMPROVER outputs to be assessed by operational meteorologists that provided useful feedback on future requirements.

The aim is for IMPROVER to be the source of new and innovative probabilistic forecast information to better exploit our convective-scale ensemble forecasts for a wide variety of users. It is recognized that IMPROVER outputs will not meet the bespoke needs of every user but can support the downstream generation of more bespoke products, for example, for industry use. IMPROVER is now being used by the Australian Bureau of Meteorology, who are collaborating on research toward applications that meet their user needs.

Future plans. The most immediate scientific focus is the incorporation of ECMWF ensemble forecasts to extend the forecast range from 7 to 14 days. We will continue enhancing the calibration methods to improve forecast skill and blending. Although the processing of variables individually is effective for our purposes, some users may require more coupling between variables. This would mean creating more joint-variable probabilities or using methods such as ensemble copula coupling (ECC) (Schefzik et al. 2013) to reinstate multivariate relationships that are lost in a univariate approach. Neighborhood processing and other spatial
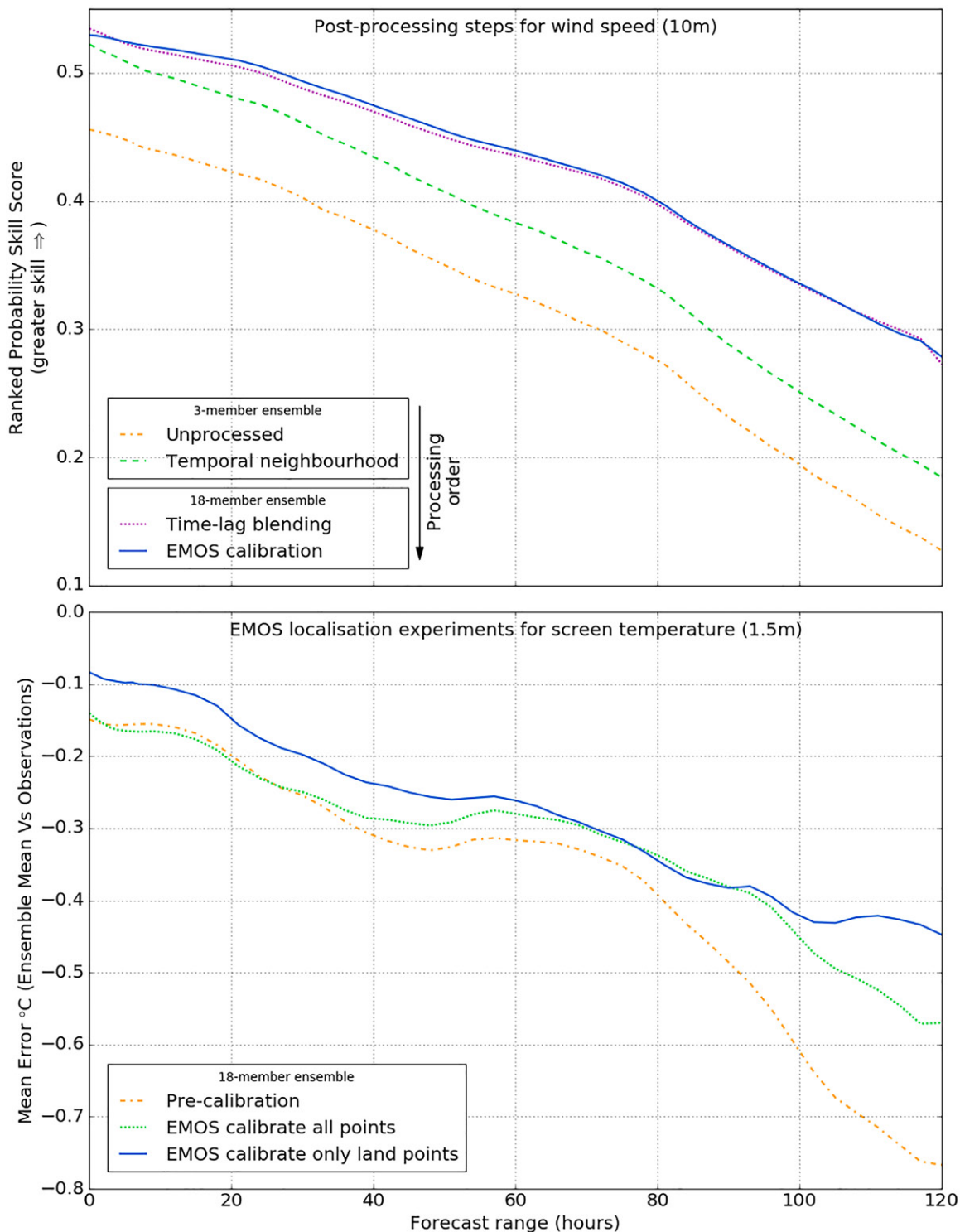
Fig. 9. Two example verification plots demonstrating assessment of processing chains and different calibration methods. (top) The RPSS performance of NEP 10 m wind speed forecasts from their ingestion into IMPROVER (unprocessed), through successive processing steps to the final EMOS-calibrated postprocessed product. (bottom) The mean error of NEP temperature forecasts prior to calibration, and subsequent to two calibration configurations, allowing the best configuration to be found.

methods can be further enhanced to improve skill. The IMPROVER framework lends itself to the introduction of methods for diagnosing phenomena or regimes to improve calibration or provide more useful outputs (Allen et al. 2021b) as well as the introduction of newer forms of machine learning, which has already started (Rasp and Lerch 2018). Increasing the exposure of probabilistic information from IMPROVER through the use of testbeds or surveys will allow more feedback on its future development needs.

We have an unprecedented opportunity within the IMPROVER framework to collaborate with partners, including the Australian Bureau of Meteorology, on new postprocessing methods and broader applications. On the technical side, work will continue to improve infrastructure efficiency, design, and optimization. We need the infrastructure to be able to incorporate the next generation of NWP models, including city-scale models, and to be able to operate on multiple platforms, including making use of the code modularity to move the processing to the cloud.

References

Allen, S., G. R. Evans, P. Buchanan, and F. Kwasniok, 2021a: Accounting for skew when postprocessing MOGREPS-UK temperature forecast fields. *Mon. Wea. Rev.*, **149**, 2835–2852, https://doi.org/10.1175/MWR-D-20-0422.1.

——, ——, ——, and ——, 2021b: Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. *Quart. J. Roy. Meteor. Soc.*, **147**, 1403–1418, https://doi.org/10.1002/qj.3983.

Beck, J., F. Bouttier, L. Wiegand, C. Gebhardt, C. Eagle, and N. Roberts, 2016: Development and verification of two convection-allowing multi-model ensembles over western Europe. *Quart. J. Roy. Meteor. Soc.*, **142**, 2808–2826, https://doi.org/10.1002/qj.2870.

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, https://doi.org/10.1002/met.1435.

——, T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148**, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1.

Bouttier, F., and H. Marchal, 2020: Probabilistic thunderstorm forecasting by blending multiple ensembles. *Tellus*, **72A**, 1–19, https://doi.org/10.1080/16000870.2019.1696142.

Bowler, N. E. H., C. E. Pierce, and A. Seed, 2004: Development of a precipitation nowcasting algorithm based on optical flow techniques. *J. Hydrol.*, **288**, 74–91, https://doi.org/10.1016/j.jhydrol.2003.11.011.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

——, and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, https://doi.org/10.3402/tellusa.v65i0.18171.

Evans, G. R., and Coauthors, 2020: metoppv/IMPROVER: IMPROVER: A library of algorithms for meteorological post-processing (Version 0.10.0). Zenodo, https://doi.org/10.5281/zenodo.3744431.

Feldmann, K., D. Richardson, and T. Gneiting, 2019: Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.*, **46**, 7744–7751, https://doi.org/10.1029/2019GL083189.

Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. *Tellus*, **66A**, 22662, https://doi.org/10.3402/tellusa.v66.22662.

Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Gneiting, T., 2014: Calibration of medium-range weather forecasts. ECMWF Tech. Memo. 719, 28 pp, https://doi.org/10.21957/8xna7glta.

——, A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Göber, M., E. Zsoter, and D. S. Richardson, 2008: Could a perfect model ever satisfy a naive forecaster? On grid box mean versus point verification. *Meteor. Appl.*, **15**, 359–365, https://doi.org/10.1002/met.78.

Golding, B., N. Roberts, G. Leoncini, K. Mylne, and R. Swinbank, 2016: MOGREPS-UK convection-permitting ensemble products for surface water flood forecasting: Rationale and first results. *J. Hydrometeor.*, **17**, 1383–1406, https://doi.org/10.1175/JHM-D-15-0083.1.

Griffiths, D., M. Foley, I. Ioannou, and T. Leeuwenburg, 2019: Flip-Flop Index: Quantifying revision stability for fixed-event forecasts. *Meteor. Appl.*, **26**, 30–35, https://doi.org/10.1002/met.1732.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

Hamill, T. M., and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Hayden, C. M., and R. J. Purser, 1995: Recursive filter for objective analysis of meteorological fields: Applications to NESDIS operational processing. *J. Appl. Meteor.*, **34**, 3–15, https://doi.org/10.1175/1520-0450-34.1.3.

Howard, T., and P. Clark, 2007: Correction and downscaling of NWP wind speed forecasts. *Meteor. Appl.*, **14**, 105–116, https://doi.org/10.1002/met.12.

Inverarity, G. W., and Coauthors, 2023: Met Office MOGREPS-G initialisation using an ensemble of hybrid four-dimensional ensemble variational (En-4DEnVar) data assimilations. *Quart. J. Roy. Meteor. Soc.*, https://doi.org/10.1002/qj.4431, in press.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, https://doi.org/10.1175/MWR-D-11-00356.1.

Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack, 2012: Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 755–768, https://doi.org/10.1002/qj.939.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc.*, **133**, 1487–1500, https://doi.org/10.1002/qj.135.

Moseley, S., 2011: From observations to forecasts—Part 12: Getting the most out of model data. *Weather*, **66**, 272–276, https://doi.org/10.1002/wea.844.

Oliver, H., and Coauthors, 2019: Workflow automation for cycling systems: The Cylc workflow engine. *Comput. Sci. Eng.*, **21**, 7–21. https://doi.org/10.1109/MCSE.2019.2906593.

Oliver, H. J., M. Shin, and O. Sanders, 2018: Cylc: A workflow engine for cycling systems. *J. Open Source Software*, **3**, 737, https://doi.org/10.21105/joss.00737.

Osczevski, R., and M. Bluestein, 2005: The new wind chill equivalent temperature chart. *Bull. Amer. Meteor. Soc.*, **86**, 1453–1458, https://doi.org/10.1175/BAMS-86-10-1453.

——, and ——, 2008: Comments on "Inconsistencies in the 'new' windchill chart at low wind speeds." *J. Appl. Meteor. Climatol.*, **47**, 2737–2738, https://doi.org/10.1175/2008JAMC1827.1.

Osinski, R., and F. Bouttier, 2018: Short-range probabilistic forecasting of convective risks for aviation based on a lagged-average-forecast ensemble approach. *Meteor. Appl.*, **25**, 105–118, https://doi.org/10.1002/met.1674.

Porson, A. N., and Coauthors, 2020: Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Quart. J. Roy. Meteor. Soc.*, **146**, 3245–3265, https://doi.org/10.1002/qj.3844.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Reed, J. R., and J. C. Senkbeil, 2020: Perception and comprehension of the extended forecast graphic: A survey of broadcast meteorologists and the public. *Bull. Amer. Meteor. Soc.*, **101**, E221–E236. https://doi.org/10.1175/BAMS-D-19-0078.1

——, and ——, 2021: Modifying the extended forecast graphic to improve comprehension. *Wea. Climate Soc.*, **13**, 57–66, https://doi.org/10.1175/WCAS-D-20-0086.1.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Roberts, N. M., 2003: Precipitation diagnostics for a high resolution forecasting system. Met Office Forecasting Research Tech. Rep. 423, 45 pp., https://library.metoffice.gov.uk/Portal/Default/en-GB/RecordView/Index/251943.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, https://doi.org/10.1214/13-STS443.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

Sheridan, P., S. Smith, A. Brown, and S. Vosper, 2010: A simple height-based correction for temperature downscaling in complex terrain. *Meteor. Appl.*, **17**, 329–339, https://doi.org/10.1002/met.177.

Steadman, R. G., 1984: A universal scale of apparent temperature. *J. Climate Appl. Meteor.*, **23**, 1674–1687, https://doi.org/10.1175/1520-0450(1984)023<1674:AUSOAT>2.0.CO;2.

Tang, Y., H. W. Lean, and J. Bornemann, 2013: The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteor. Appl.*, **20**, 417–426, https://doi.org/10.1002/met.1300.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257, https://doi.org/10.1017/S1350482705001763.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Ylinen, K., O. Räty, and M. Laine, 2020: Operational statistical postprocessing of temperature ensemble forecasts with station-specific predictors. *Meteor. Appl.*, **27**, e1971, https://doi.org/10.1002/met.1971.

Zhang, X., 2018: Application of a convection-permitting ensemble prediction system to quantitative precipitation forecasts over southern China: Preliminary results during SCMREX. *Quart. J. Roy. Meteor. Soc.*, **144**, 2842–2862, https://doi.org/10.1002/qj.3411.

Zhu, Y., G. Iyengar, Z. Toth, S. M. Tracton, and T. Marchok, 1996: Objective Evaluation of the NCEP Global ensemble forecasting system. *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.