Improvements in the spread–skill relationship of precipitation in a convective-scale ensemble through blending

Received: 11 October 2023 | Revised: 6 March 2024 | Accepted: 23 April 2024

DOI: 10.1002/qj.4754

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Improvements in the spread–skill relationship of precipitation in a convective-scale ensemble through blending

Adam Gainford[1] | **Suzanne L. Gray**[1] | **Thomas H. A. Frame**[1] | **Aurore N. Porson**[2] | **Marco Milan**[3]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]MetOffice@Reading, University of Reading, Reading, UK

[3]Met Office, Exeter, UK

Correspondence
Adam Gainford, Department of Meteorology, University of Reading, Brian Hoskins Building, Reading, Berkshire, RG6 6ET, UK.
Email: a.gainford@pgr.reading.ac.uk

Funding information
SCENARIO NERC Doctoral Training Partnership, University of Reading, Grant/Award Number: NE/S007261/1; Met Office CASE Studentship, Met Office

Abstract

Convective-scale ensembles are used routinely in operational centres around the world to produce probabilistic precipitation forecasts, but a lack of spread between members is providing forecasts that are frequently overconfident. This deficiency can be corrected by increasing spread, increasing forecast accuracy, or both. A recent development in the Met Office forecasting system is the inclusion of large-scale blending (LSB) in the convective-scale data assimilation scheme. This method aims to reduce the synoptic-scale forecast error in the analysis by reducing the influence of the convective-scale data assimilation at scales that are too large to be constrained by the limited domain. These scales are instead initialised using output from the global data assimilation scheme, which we expect to reduce the forecast error and thus improve the spread–skill relationship. In this study, we quantify the impact of LSB on the spread–skill relationship of hourly precipitation accumulations by comparing forecast ensembles with and without LSB over a 17-day summer trial period. This trial found modest but significant improvements to the spread–skill relationship as calculated using metrics based on the Fractions Skill Score. Skill is improved for a lower precipitation centile by an average of 0.56% at the largest scales, but a corresponding degradation of spread limits the overall correction. The spread–skill disparity is reduced the most in the higher centiles due to a more muted spread response, with significant reductions of up to 0.40% obtained at larger scales. Case-study analysis using a novel extension of the Localised Fractions Skill Score demonstrates how spread–skill improvements transfer to smaller-scale features, not just the scales that have been blended. There are promising signs that further spread–skill improvements can be made by implementing LSB more fully within the ensemble, and we encourage the Met Office to continue developing this technique.

KEYWORDS

convection-permitting, data assimilation, forecast skill, Fractions Skill Score

1 | INTRODUCTION

Convective-scale ensembles have been used for over a decade to quantify the uncertainty in convective-scale weather forecasts (e.g., Clark *et al.*, 2011; Klasa *et al.*, 2018; Wang *et al.*, 2011). Ideally, the spread between ensemble members should be equal to the expected error of the ensemble mean when verified over many forecasts (Buizza, 1997; Hopson, 2014). If this spread–skill relationship is well correlated, the spread can be used to predict the forecast skill, with small spread (large confidence) implying a skilful forecast and vice versa. However, meteorological centres around the world often report that convective-scale ensembles provide overconfident, and typically underspread, forecasts given the verified weather (e.g., Beck *et al.*, 2016; Cafaro *et al.*, 2021; Ferrett *et al.*, 2021; Porson *et al.*, 2019, 2020; Raynaud & Bouttier, 2017; Schwartz *et al.*, 2014; Tennant, 2015). This overconfidence can be addressed in two ways: either by increasing the spread between members, thereby decreasing the confidence, or by increasing the skill of the ensemble mean, thereby making the large confidence more appropriate. One way of improving the skill of convective-scale models throughout the early stages of the integration is to improve the accuracy of the initial state.

Due to the computational cost of running operational models at resolutions approaching the convective-scale, forecasts must be run over a limited region. By definition, the data assimilation (DA) schemes that initialise these regional models do not include information extending beyond the model domain, which limits the accuracy of features with scales exceeding the domain size (Guidard & Fischer, 2008). Therefore, it is expected that a regional model DA scheme will represent scales approaching its own domain size less accurately than the global host model within which it is nested (providing lateral boundary conditions). Recent studies have shown that nudging the synoptic-scale regional analysis of selected variables towards that of the host-model analysis improves skill in deterministic models (Bengtsson *et al.*, 2017; Milan *et al.*, 2023). Our work extends these findings to consider the impact of these blended analyses on the spread–skill relationship of a convective-scale ensemble. We posit that this ensemble will show the same improvement in spread and skill as other studies of this nature (Keresturi *et al.*, 2019; Schwartz *et al.*, 2021, 2022; Zhang *et al.*, 2015). In particular, we expect that the ensemble will benefit from the same increase in skill demonstrated in deterministic forecasts, while leaving the initial conditions in the convective-scale model to diverge at a rate similar to or larger than without blending. In this way, the spread–skill disparity will be reduced because the lack of spread between members will be more appropriate given the increase in skill.

Blending is just one of many methods being explored to improve the performance of convective-scale ensembles: time-lagging (Ben Bouallègue *et al.*, 2013; Mittermaier, 2007; Raynaud & Bouttier, 2017), stochastic physics schemes (McCabe *et al.*, 2016), and multi-model ensembles (Beck *et al.*, 2016; Porson *et al.*, 2019) have all shown promising spread improvements to varying degrees. However, there are also improvements being made to the more fundamental aspects of ensemble design, such as the perturbation and initial condition strategies. Recent upgrades to the Met Office Global and Regional Ensemble Prediction System–Global (MOGREPS-G) DA setup have produced large improvements in skill and modest increases in spread compared with the previous ensemble transform Kalman filter scheme (Inverarity *et al.*, 2023). However, it is unclear how much these spread improvements propagate through to the convective-scale ensembles that are nested within the global ensemble. This transfer of spread is likely to have some dependence on the method used to initiate the ensemble: that is, whether the ensemble is initialised as a simple downscaler of the global ensemble or whether it uses a separate, higher resolution DA scheme. Tennant (2015) has shown that using convective-scale analyses to initialise convective-scale ensembles increases skill and spread compared with a downscaled ensemble, and is therefore the preferred strategy for the operational Met Office convective-scale ensemble, MOGREPS-UK. However, the synoptic scales initialised using convective-scale DA may conflict with the synoptic scales arriving from the global model via the member perturbations or lateral boundary conditions, hence the desire to achieve a better balance in the initial state through blending (Caron, 2013).

Recent studies have demonstrated consistently the benefits of using blending schemes in regional models. For instance, blending has been shown to remove large systematic biases affecting typhoon tracks in the North Pacific Ocean (Hsiao *et al.*, 2015), correct mismatches between analysis and lateral boundary condition perturbations (Caron, 2013; Wang *et al.*, 2011), and reduce spin-up and wind errors in the first 24 h of integration (Wang *et al.*, 2014). These improvements all have positive impacts on model skill, but there is also evidence that synoptic-scale blending can introduce additional spread in convective-scale ensembles (Keresturi *et al.*, 2019; Schwartz *et al.*, 2021, 2022; Zhang *et al.*, 2015). In fact, Zhang *et al.* (2015) demonstrated that larger-scale perturbations are much more effective at generating ensemble spread than smaller-scale perturbations. However, these performance benefits have also been shown to depend on the specific blending technique used. One of the main

choices that must be made when implementing a blending scheme is the cutoff wavelength controlling the scale at which the host model begins to influence the regional model (Yang, 2005). Most studies choose a single wavelength, of between 500 and 1000 km, which defines the shape of a Raymond-like weighting profile (Raymond, 1988). Other studies have shown that introducing a dynamic cutoff wavelength, varying either by regime (Feng *et al.*, 2020, 2021) or by model variable and height (Zhang *et al.*, 2015), can improve model performance further compared with a static wavelength.

Despite these technical differences, there is large agreement that blending can improve the spread–skill disparity in convective-scale ensembles. Our study investigates this hypothesis by applying the "large-scale blending" (LSB: Milan *et al.*, 2023) formulation to the initial conditions of a convective-scale ensemble and measuring the associated response in spread and skill. LSB has recently been implemented into the Met Office's regional 4D-Variational DA scheme, and has been shown to reduce gravity-wave generation and improve skill in trials performed with the deterministic, convective-scale UK variable resolution (UKV) model (Milan *et al.*, 2023). Our work extends these findings to consider the spread–skill impact of recentring ensemble members around UKV background fields blended with LSB. Note that this work focuses only on assessing improvements to the spread–skill relationship of precipitation and does not consider any potential, broader ensemble quality improvements. In fact, blending has a negligible impact on probabilistic forecast metrics such as reliability curves, rank histograms, and relative operating characteristic (ROC) areas (not shown), which suggests that the predominant benefit will instead be observed spatially. Additionally, we would expect LSB to impact variables other than just precipitation, particularly those that undergo blending directly (Milan *et al.*, 2023), but we do not analyse this here.

This work presents results of a trial comparing two ensemble configurations: a reference ensemble where the initial state was updated using 4D-Var without blending as outlined in Milan *et al.* (2020), and a blended ensemble where LSB was included in the DA scheme. These ensemble forecasts were run in summer 2019 and include several convective events. After a description of MOGREPS-UK, the LSB method, and the diagnostic approaches in Section 2, Section 3 (the results section) begins with a discussion of the characteristics and climatology of the weather within the trial period. Then, precipitation distributions are analysed, which motivates the focus on assessing the LSB impacts in purely spatial terms. Next, the differences between the LSB and reference ensembles across the entire trial period are assessed, with a focus on evaluating the spread–skill response. The significance of these differences is considered by comparing them with similar statistics generated from mixing ensemble members of both trials. This technique allows us to quantify the extent to which the members composing the LSB ensemble can be considered a unique sampling of the underlying distribution, and not just another sampling of the reference distribution. After this, a case study is presented using a novel metric that locates areas of improved spread and skill within the domain. We show a case of elevated convection that has been predicted more accurately and more confidently with LSB included. Finally, Section 4 concludes the article by discussing limitations and future work.

2 | METHODS

This section starts by outlining the ensemble configuration used in this work (Section 2.1), before describing how LSB is implemented within this ensemble (Section 2.2). After this, the metrics used to assess the spread–skill relationship are presented (Sections 2.3 and 2.4), before concluding with a discussion on our significance estimation approach (Section 2.5).

2.1 | MOGREPS-UK

MOGREPS-UK is the Met Office's operational, 18-member, convective-scale ensemble run over the UK. The variable-resolution grid starts at 4-km grid spacing at the corners and tapers to 2.2-km grid spacing in the fixed-resolution inner mesh, where all subsequent analysis is performed. Figure 1 shows a schematic of the initialisation procedure. Note that this is updated from fig. 1 of (Porson *et al.*, 2020) to reflect the improved timeliness of the member initialisation, which was implemented shortly after the lagged configuration was introduced. MOGREPS-UK cycles every hour, producing three new members run out to 120 h, which are combined with the 15 members from the previous five cycles to produce an 18-member lagged ensemble. This time-lagged approach allows the model to utilise the hourly updates provided by the UKV convective-scale DA, and has the added benefit of improving spread between ensemble members (Porson *et al.*, 2020).

Every hour, a high-resolution analysis with 1.5-km grid spacing is produced over the UK domain using convective-scale 4D-Var DA (Milan *et al.*, 2020). To produce the three new perturbed high-resolution members, three members of the global MOGREPS-G ensemble are selected and perturbations about the 17-member ensemble mean (excluding the control member) are calculated. These perturbations are then added to

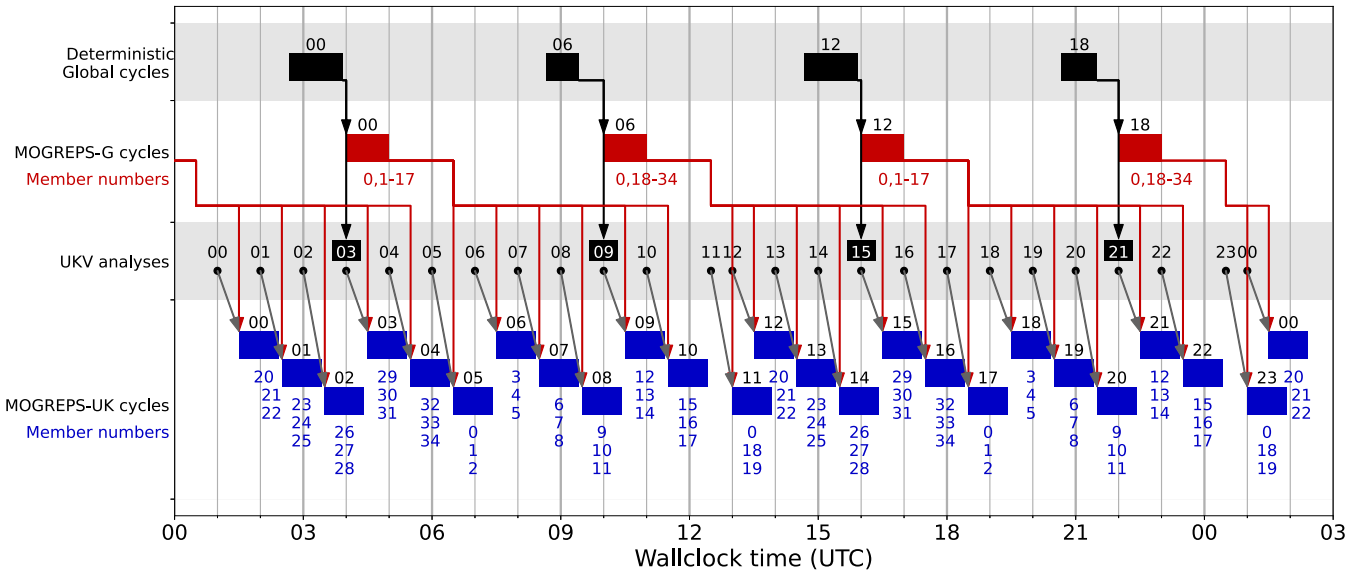

FIGURE 1 Schematic showing the data flow for initialising the time-lagged MOGREPS-UK ensemble. Top: black boxes show the six-hourly deterministic global model cycling frequency. The red boxes, arrows, and numbers show the MOGREPS-G members that provide initial-condition perturbations and lateral boundary conditions. The black dots and grey arrow show the UKV analyses around which a given MOGREPS-UK cycle is centred. The UKV analyses that are blended with the global model are highlighted by black backgrounds (only applies to the blended ensemble). Blue boxes show the run times of a single MOGREPS-UK cycle, while the blue numbers show the ensemble members initialised in that cycle. The 18-member lagged ensemble for a given hour is comprised of the three members initialised at that hour combined with the 15 members from the previous five cycles. Figure adapted from Porson *et al.* (2020). [Colour figure can be viewed at wileyonlinelibrary.com]

the high-resolution analysis to produce the three new high-resolution perturbed members. Due to the production time required for the MOGREPS-G ensemble, the perturbations do not always derive from the most recent analysis (e.g., the perturbed UK members produced from the 0900 UTC high-resolution analysis use perturbations from the 0000 UTC MOGREPS-G forecast rather than the 0600 UTC MOGREPS-G forecast, whereas the 1100 UTC MOGREPS-UK members are perturbed using the 0600 UTC MOGREPS-G).

Since December 2019, MOGREPS-G initialises each member separately using hybrid 4D ensemble variational data assimilation (hybrid 4DEnVar: Inverarity *et al.*, 2023). MOGREPS-G cycles every six hours at 0000, 0600, 1200, and 1800 UTC, producing 17 members + 1 control from global analysis. We retain the same member labelling from MOGREPS-G data assimilation for the corresponding MOGREPS-G and MOGREPS-UK members, meaning that there are 35 member labels despite the fact that only 18 members are included in a given forecast (see red text of Figure 1). Note that there is no relation between members with the same labels initialised 12 h apart.

This study analyses the effects of LSB within MOGREPS-UK by comparing forecasts from two ensemble configurations. The "reference" ensemble was run without LSB, while the "blended" ensemble implemented LSB as described in the next section. Each ensemble was run using the second Regional Atmosphere and Land science configuration for midlatitudes (RA2-M: Bush *et al.*, 2023) with additional stochastic physics perturbations introduced using the Random Parameter 2 scheme (McCabe *et al.*, 2016). Both ensembles were run for a 17-day period over summer and winter 2019. On average, though, the ensemble forecasts run over winter showed differences that were an order of magnitude smaller than for the summer and are therefore not discussed further in this work.

2.2 | Large-scale blending in MOGREPS-UK

LSB is the blending approach chosen by the Met Office to improve synoptic scales within regional model analyses. In general, blending schemes can choose to modify the synoptic scales of either the regional analysis post DA or the regional background within/prior to DA. Here, LSB opts to integrate blending fully into the DA process. Blended increments are obtained by finding the difference between the synoptic scales of the "host" model (the Met Office deterministic global model forecast downscaled onto the UKV grid) and the synoptic scales of the regional

(UKV) background. The incremental 4D-variational DA uses blended fields as background and in its formulation. Therefore, the increments after minimization are on the blended fields. For a full description of the LSB implementation at the Met Office, the reader is directed to section 2.1 of Milan *et al.* (2023).

In LSB, the synoptic scales are distinguished from the convective scales by a Raymond low-pass filter (Raymond, 1988) with wavelength cut-off of 700 km. For this choice of cut-off wavelength, blending begins to have an effect at scales above 400 km and reaches a maximum response at approximately 1100 km. At all scales larger than 1100 km, the blended field is composed of 75% host model background and 25% regional model background, where this choice of weights was found to maximise skill (Milan *et al.*, 2023). A schematic of this amplitude response is shown in fig. 4 of Milan *et al.* (2023). When LSB is applied, blended fields are obtained for the horizontal wind, potential temperature, pressure, and density. LSB is also applied to the total water-vapour content, but additional increments are added, which ensures the relative humidity field is nudged back towards the convective-scale DA state to avoid spurious precipitation spin-up (further details can be found in Milan *et al.* (2023), Section 2.2 and the appendix).

The only difference between the "blended" and "reference" MOGREPS-UK configurations used in this study is the UKV analyses providing the initial conditions. Both configurations receive the same lateral boundary conditions and member perturbations from MOGREPS-G. However, LSB is only applied to construct the blended analysis in one hour out of every six. This choice is made because of an observed effect in which synoptic-scale LSB and 4D-Var increments anti-correlate during cycles where LSB is applied without a corresponding update to the boundary conditions—this effect is explained in more detail in Milan *et al.* (2023). Therefore, because of the time-lagged configuration of MOGREPS-UK, the following holds.

- LSB is only applied directly to the initial conditions of the members initialised at 0300, 0900, 1500, and 2100 UTC (UKV analyses with black boxes in Figure 1). We refer to the members initialised during these cycles as being "directly blended" (members 12, 13, 14, 29, 30, 31).
- All other members are initialised around analyses that have used blended backgrounds, or in other words, LSB has been applied during a *prior* cycle (UKV analyses without black boxes in Figure 1). Even though blending has not been applied to the analyses of these cycles, the influence of LSB will feed through via a chain of backgrounds from the previous directly blended analysis. We refer to the members initialised during these cycles as being "indirectly blended."

As is the case for all lagged ensembles, the full 18-member ensemble does not form an independent and identically distributed (i.i.d.) sample of realisations, since we would expect the older three-member sub-ensembles of the 18-member lagged set to have larger variance than the fresher sub-ensembles. Moreover, since MOGREPS-G and MOGREPS-UK have different cycling frequencies, and because LSB is only applied to a single three-member sub-ensemble, there are structural differences in the production method that make each individual sub-ensemble distinct from another. These distinctions need to be taken into account in any statistical analysis aimed at determining the impact of blending on the ensemble.

The sporadic application of LSB implies limited divergence between the two ensemble configurations, so, to provide context, it is useful to inspect the member fields briefly. Figure 2 shows a comparison of hourly precipitation accumulations for a selection of members from the reference and blended ensembles. This period occurs 10 h before the case study considered in Section 3.4 and was chosen because of the large uncertainty in the development of a band of rain over Ireland, which illustrates the typical difference between the two ensembles. There is larger variation between members of the same ensemble than there is between the same member from the two ensembles. If the reference ensemble member did not evolve this rain band, the addition of blending did not cause a differing evolution. Similarly, if the band of rain did develop in an ensemble member, the intensity of the precipitation is similar in the same member in both ensembles. This observation suggests that the inclusion of blending does little to the distribution of precipitation, a hypothesis that is explored more thoroughly in Section 3.2. There are, however, subtle differences in the spatial patterns, which we hypothesise to be more accurate in the blended ensemble on average. One of the focuses of this study is to verify this statement, which is achieved using the Fractions Skill Score.

2.3 | Fractions Skill Score (FSS)

The effect of LSB on the spread–skill relationship is evaluated from the spatial improvements made to hourly precipitation accumulations. To measure these improvements we use the Fractions Skill Score (FSS: Roberts & Lean, 2008), a neighbourhood-based metric designed to calculate the difference between two fields over a prescribed scale. We use the FSS because it is not sensitive to the double penalty problem (Gilleland *et al.*, 2009; Wernli *et al.*, 2009) and allows us to easily assess the impact of LSB on ensemble spread and skill across a range of scales. This scale awareness is important, because

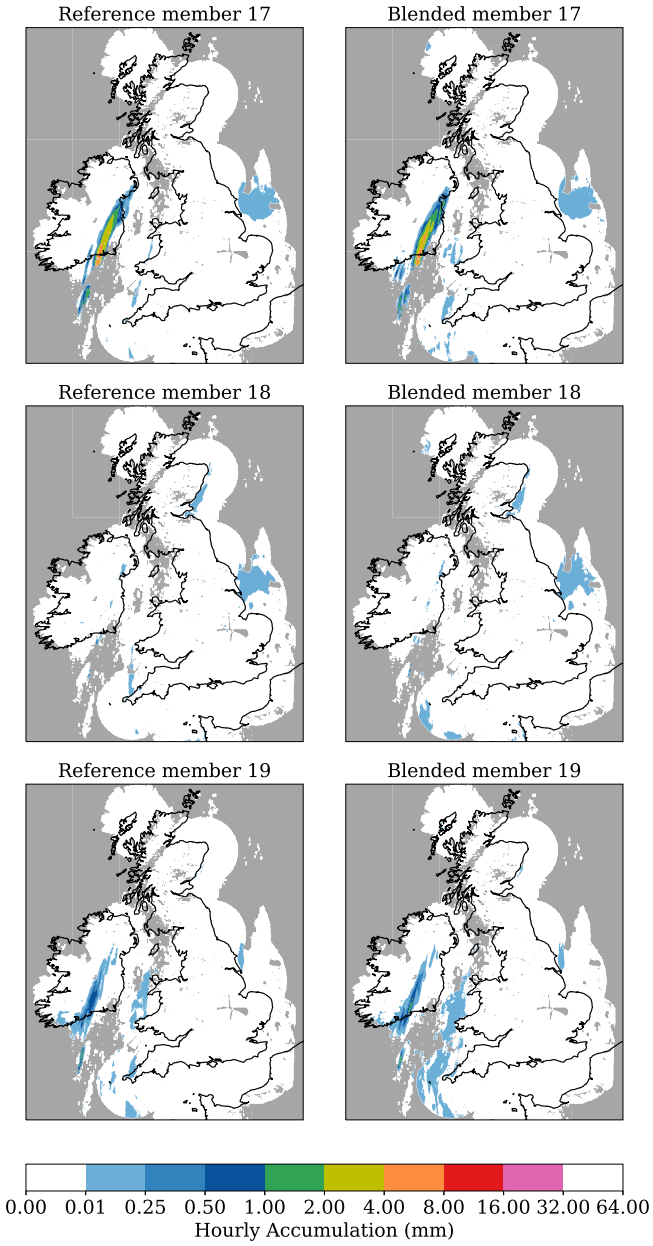

FIGURE 2 A selection of postage stamps from the reference and blended ensembles for the June 29, 2019, 0300 UTC ensemble forecast, lead time $T + 4$ h. Members 17 were initialised at 2200 UTC the previous day (lead time $T + 9$ h), one hour after direct blending. Members 18 and 19 were initialised at 2300 UTC the previous day (lead time $T + 8$ h), two hours after direct blending. There is large uncertainty within the ensemble about the development of the band of rain over Ireland. Mask applied from the radar as described in Section 2.3. [Colour figure can be viewed at wileyonlinelibrary.com]

we expect LSB to have a scale-dependent effect on the ensemble.

The FSS operates by first converting the forecast and observed precipitation hourly accumulations into binary fields that are equal to unity if the precipitation exceeds a specified threshold or zero otherwise. Observations are provided by the Nimrod radar system (Golding, 1998) and are interpolated to the MOGREPS-UK grid using a nearest-neighbour algorithm that masks any extrapolated points. We acknowledge that there are uncertainties associated with radar observations, especially with cases of elevated convection, but do not consider these uncertainties here. Regions that lie outside the radar envelope are masked out in both the observations and the forecast to ensure fair comparisons. To account for potential model bias in absolute precipitation amounts, the threshold used to create the binary field is a centile value and applied such that if, for example, the 90th percentile is used, 10% of grid points within the radar envelope have a value of one. These binary fields are then converted to fractions fields by averaging over a square neighbourhood of size $n \times n$ grid points, where n is also specified. Finally, two fractions fields, A and B, can be compared by calculating the mean squared difference ($\mathrm{MSD}_{(n)}$) between the two fields and benchmarking against a low-skill climatalogical baseline ($\mathrm{MSD}^{\mathrm{ref}}_{(n)}$) to produce the FSS:

$$\mathrm{MSD}_{(n)}(A,B) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[A_{(n)i,j} - B_{(n)i,j}\right]^2 , \quad (1)$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n)}(A,B) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[A^2_{(n)i,j} + B^2_{(n)i,j}\right] , \quad (2)$$

$$\mathrm{FSS}_{(n)}(A,B) = 1 - \frac{\mathrm{MSD}_{(n)}(A,B)}{\mathrm{MSD}^{\mathrm{ref}}_{(n)}(A,B)} , \quad (3)$$

where N_x and N_y are the number of grid points in the x and y directions. An FSS of unity indicates identical fractions fields, while a score of zero indicates fields that are completely mismatched. Note that the post-processing code that is used to calculate the fractions fields regrids the data onto a stage grid with spacing 2327 m (Roberts *et al.*, 2023), hence the grid point to km conversion is slightly different from the expected 2.2 km for MOGREPS-UK.

Typically, the FSS is used to understand the scales at which a deterministic forecast becomes skilful by comparing the forecast with a verification and recalculating the score for increasing neighbourhood sizes until an acceptable value has been reached (approximately 0.5). For MOGREPS-UK, this score usually occurs at neighbourhood sizes of between 50 and 100 km. For our purposes, we must extend the analysis to encompass larger scales, given that the cutoff wavelength for LSB is an order of magnitude larger than the typical skilful scale. However, for neighbourhood areas approaching the domain size, edge disparities can become increasingly important: a fractions value towards the boundary of the domain may be calculated with far fewer grid points in the surrounding large neighbourhood than a more central fractions value.

Nachamkin and Schmidt (2015) have shown that the FSS can be meaningfully impacted by the method to handle boundary-fraction values, especially for poor forecasts and small domains. For our study, we expect this effect to have a similar impact on both blended and reference ensembles, and thus the results comparing differences between the two ensembles will be largely insensitive to this handling method.

Dey *et al.* (2014) introduced two metrics that use the FSS to evaluate ensemble spread–skill relationships. For an M-member ensemble, the dispersion FSS (dFSS) is an average of the FSS between all member–member pairs to yield a single value representing the spread:

$$\mathrm{dFSS}_{(n)} = \frac{1}{M(M-1)} \sum_{M_a=1}^{M} \sum_{M_b 1=M_a \neq M_b}^{M} \mathrm{FSS}_{(n)}(M_a, M_b)\,, \quad (4)$$

where M_a and M_b are the fractions fields for the members being compared, as described previously. Larger dFSS values mean there is greater similarity between members, and therefore lower spread (and vice versa). As well as ensemble spread, the skill can be measured using the error FSS (eFSS), which averages the FSS between each ensemble member and a chosen verification field, O, as given by

$$\mathrm{eFSS}_{(n)} = \frac{1}{M} \sum_{M_a=1}^{M} \mathrm{FSS}_{(n)}(M_a, O)\,, \quad (5)$$

where higher eFSS values mean higher skill. A useful spread–skill relationship should show no bias between the eFSS or dFSS (Dey *et al.*, 2016). If the ensemble produces higher dFSS than eFSS values over many forecasts, it is underspread (and lower dFSS than eFSS values imply an overspread ensemble). Note that a single forecast cannot meaningfully be described as underspread or overspread, since these descriptors are only useful over multiple forecasts.

2.4 | Localised Fractions Skill Score (LFSS)

By design, skill scores such as the FSS produce a domain-averaged value that can be sequenced in time to understand the evolution of model performance. If, instead, we wish to understand the spatial distribution of model performance, we must modify this diagnostic to preserve spatial awareness. To achieve this, Woodhams *et al.* (2018) introduced the Localised Fractions Skill Score (LFSS), which uses an identical formulation to the FSS as presented in Equations (1)–(3), but instead uses summations over time to obtain a spatial field of scores at each grid point, i, j. The LFSS is calculated as

$$\mathrm{MSD}_{(n,i,j)}(A,B) = \frac{1}{T} \sum_{t=1}^{T} \left[A_{(n,i,j)t} - B_{(n,i,j)t}\right]^2\,, \quad (6)$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n,i,j)}(A,B) = \frac{1}{T} \sum_{t=1}^{T} \left[A^2_{(n,i,j)t} + B^2_{(n,i,j)t}\right]\,, \quad (7)$$

$$\mathrm{LFSS}_{(n,i,j)}(A,B) = 1 - \frac{\mathrm{MSD}_{(n,i,j)}(A,B)}{\mathrm{MSD}^{\mathrm{ref}}_{(n,i,j)}(A,B)}\,, \quad (8)$$

where T is the number of field snapshots included in the calculation. At a given grid point, an LFSS of unity means that all input fields are in agreement about the precipitation in the $n \times n$ neighbourhood surrounding the point, while a score of zero means there is complete mismatch.

In an analogous way to calculating the domain-averaged ensemble spread–skill relationship, we introduce a novel extension of the LFSS that can be used to generate fields that highlight areas of larger or smaller ensemble spread and skill. We define the "dispersion LFSS (dLFSS)" and "error LFSS (eLFSS)" for a given neighbourhood, n, as the following:

$$\mathrm{dLFSS}_{(n,i,j)} = \frac{1}{M(M-1)} \sum_{M_a=1}^{M} \sum_{M_b 1=M_a \neq M_b}^{M} \mathrm{LFSS}_{(n,i,j)}(M_a, M_b)\,, \quad (9)$$

$$\mathrm{eLFSS}_{(n,i,j)} = \frac{1}{M} \sum_{M_a=1}^{M} \mathrm{LFSS}_{(n,i,j)}(M_a, O)\,. \quad (10)$$

We expect an ensemble with a useful spread–skill relationship to colocate regions of similar dLFSS and eLFSS; however, we do not attempt to verify this here for concision.

This method does not mandate a particular choice of time coordinate, so in theory the LFSS could be calculated over different lead times, cycles, or a combination of both, depending on the aims of the user. However, in our experience (not shown here), iterating over multiple cycles introduced excessive noise, which made comparisons between the two ensembles difficult to interpret. Previous work using the LFSS has also restricted iteration to lead times over a single cycle using integration periods of 24 h (Woodhams *et al.*, 2018) and 3 h (Ferrett *et al.*, 2021). Our results use 12-h periods sequencing hourly precipitation fields from lead times $T + 2$ to $T + 13$ h.

2.5 | Significance estimation

The differences between the precipitation fields of different members of the same ensemble configuration are

much larger than the differences between the same member of the two configurations, as can be seen in Figure 2. As such, we expect the impacts of blending on ensemble spread and skill to be modest, especially when summarising the data by averaging over multiple cycles. The FSS does not include any built-in uncertainty estimation, so we seek a method that quantifies this uncertainty while respecting the statistical structure of the ensemble.

The full 18-member ensemble is not strictly speaking an i.i.d. sample of realisations and is most accurately described as a set of six three-member sub-ensembles, each of which can be considered an i.i.d. sample of realisations. To quantify the significance of any measured impact of blending, we use a null hypothesis that, for each three-member sub-ensemble, the blended ensemble and the reference ensemble are drawn from the same underlying distribution and use a resampling that exchanges members only between matched sub-ensembles. This approach ensures that we isolate the response that occurs purely due to LSB, not due to mixing members from sub-ensembles with different distributions. From this, we construct confidence limits that quantify the significance of the difference between the blended and reference configurations.

Details of this constrained resampling technique, including its implementation and use in generating confidence limits, are described in Appendix A.

3 | RESULTS

3.1 | Trial period characteristics

The UK was under a southwesterly flow at the beginning of the trial period (June 16, 2019), which encouraged a number of convective storms to develop over southern England. High pressure and settled conditions then moved in from June 21 and persisted until June 24. Additional thunderstorms developed over southern and central England from June 24, with slow-moving light rain clearing from the northeast in the early hours of June 26. Conditions then remained dry and settled under another area of high pressure until the arrival of an occlusion from the west triggered fresh thunderstorms over Ireland and southern Scotland on June 29. Scattered showers persisted across the UK and Ireland until the end of the trial period on July 2 (UKMO, 2019). Overall, the trial period was highly variable, with multiple convective and showery events interspersed with more dry and settled periods.

This regime variability has a noticeable impact on hourly precipitation accumulations across the domain, as seen in the time series presented in Figure 3. Typically, the ensemble mean underestimates the precipitation across the domain compared with the radar. This is especially noticeable towards the end of June 22, when both ensembles missed the timing of the convective initiation. The ensembles were also uncertain about the development of a strong band of thunderstorms over Ireland during the beginning of June 29, with a majority of members forecasting little or no rain even at short lead times (see Figure 2). The ensemble then becomes more accurate after this band clears Northern Ireland and regains strength over central Scotland, possibly due to the more predictable forcing provided by the orography. The events immediately succeeding this period are highlighted as the green shading in Figure 3 and are studied in more detail in Section 3.4.

Also highlighted in Figure 3 is a 0.025-mm domain-average threshold, which we use to filter out dry events that occur in both ensembles and the radar. Applying this filter ensures the average FSS results are not contaminated by an undesirable feature of the FSS design which causes it to return low scores when dry events are forecast correctly (as discussed in Mittermaier, 2021). Moreover, the FSS behaves far more sensitively with low fractional precipitation coverages (Roberts & Lean, 2008), which we have found can have a large effect on the average FSS. Low coverages can either be caused by isolated, but potentially impactful convective cells or by localised, scattered showers. The former is clearly more of a concern than the latter, and any filtering method used should distinguish between these two cases. Therefore, to ensure the average FSS is not biased towards these low-impact events, a domain-average filter was chosen to select only those periods with precipitation of note. The 0.025-mm threshold value was chosen as the smallest value at which the average results presented in Section 3.3 become largely insensitive to further threshold increases. For context, this domain-average value is equivalent to light drizzle occurring over approximately 10% of the domain, where light drizzle is defined as 0.3 mm/h by the American Meteorology Society Glossary of Meteorology (AMS, 2023). Upon inspection, the filtered periods are dominated primarily by the high-pressure conditions of June 21–23 and 25–29.

The FSS results presented in the following section focus on analysing the 90th and 97.5th centiles. The radar threshold values for these centiles when averaged across all data in the trial period are 0.20 and 0.80 mm, respectively. For the filtered trial period with dry events excluded, the thresholds are 0.31 and 1.23 mm, respectively.

3.2 | Precipitation distributions

LSB could impact the hourly precipitation field in two ways: it could change the position of precipitating points in the domain, or it could change the magnitude of the

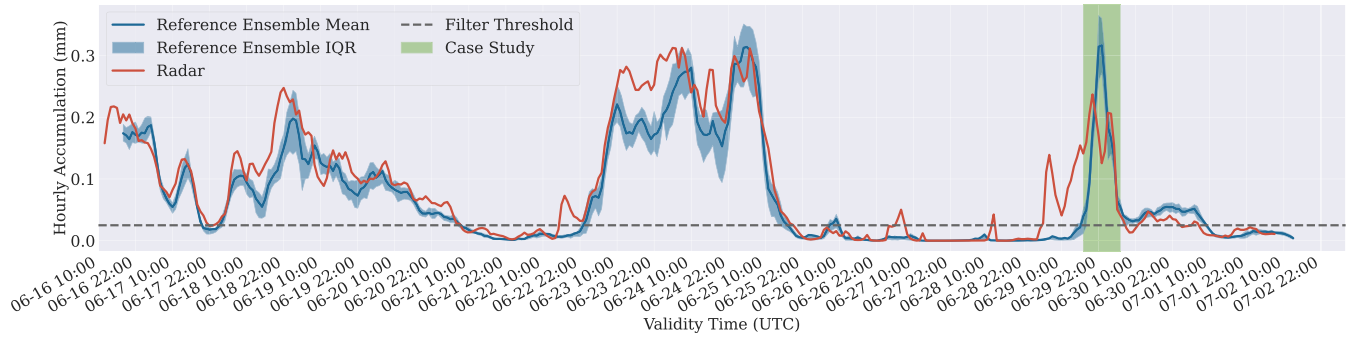

FIGURE 3 Time series of the domain-averaged hourly precipitation in the reference ensemble and radar. Ensemble data are calculated for lead time $T + 8$ h. The blended ensemble data have been omitted for clarity, but they largely follow the reference ensemble. [Colour figure can be viewed at wileyonlinelibrary.com]

overall accumulation. The postage stamps presented in Figure 2 suggest that LSB predominantly modifies the spatial location of precipitation, rather than the intensity. To examine this behaviour more thoroughly, the distribution of precipitation across the domain for both ensemble configurations was calculated and averaged over all cycles, all lead times, and all members.

Figure 4 shows that both ensembles under-represent the lightest and heaviest rain compared with the radar, and the addition of LSB has a negligible impact on the distributions when compared with the radar. For example, the percentage difference between the radar and the reference ensemble for the 0.25–0.50 mm bin is 0.834%, while the equivalent percentage difference between the blended and reference ensembles is −0.005%. This behaviour is largely insensitive to lead time: averaged over lead times $T + 2$, 4, 6 h, the radar-reference difference for the same bin is 0.582%, while the blended-reference difference is −0.020%. Similarly, for lead times $T + 20$, 22, 24 h, the radar-reference difference is 0.924%, while the blended-reference difference is −0.023%. The under-representation of light rain has been noted as a deficiency in the RA2-M physics package used for these ensembles, and is one of the targets for improvement in the RAL3 scheme (Bush *et al.*, 2023). This result is consistent with the differences being predominantly due to model biases rather than forecast initialisation.

The other concern with LSB is the generation of spurious precipitation at the start of the integration, which has been observed in other studies (Schwartz *et al.*, 2021). While we do not see this effect when analysing the ensemble as a whole, there is a much stronger signal when comparing blending between different sets of three-member sub-ensembles. Figure 5 shows domain- and cycle-average precipitation as a function of lead time for two sets of sub-ensembles from both configurations. In the blended ensemble, the set of members labelled "DB" have been

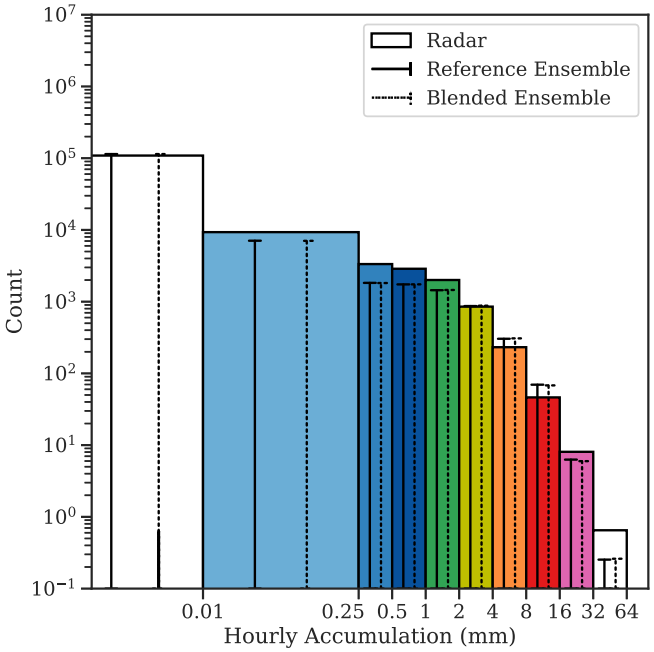

FIGURE 4 Domain-wide precipitation distributions from the radar data (bars) and both ensembles (stalks, representing the height of the equivalent bars in the ensemble histograms). Radar bars use the same logarithmic colour bar as in Figure 2 for consistency. Ensemble distributions are averaged across all cycles, lead times, and ensemble members. [Colour figure can be viewed at wileyonlinelibrary.com]

directly blended (members 12, 13, 14, 29, 30, 31). The set labelled "IDB" are the indirectly blended members initialised five hours after the most recent blending cycle (members 9, 10, 11, 26, 27, 28) and would therefore be the least affected by blending. The reference DB and IDB sets use these same selections of members, although no blending takes place in either set.

Figure 5 shows that both sets of members display significant spindown from $T + 2$ to $T + 6$ h before beginning to stabilise, which is broadly consistent with the behaviour

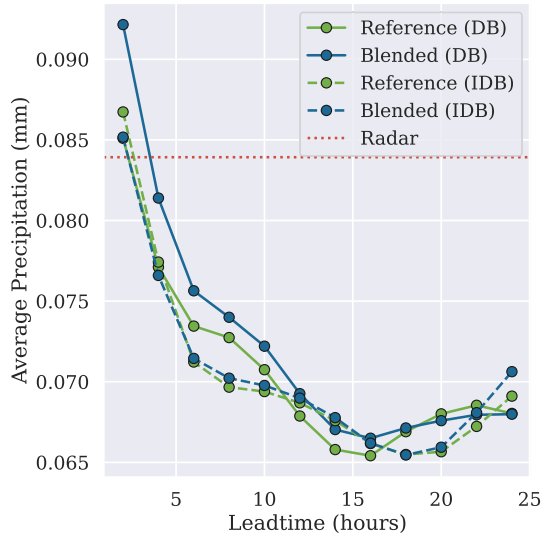

FIGURE 5 Domain-wide precipitation averaged across all cycles for the directly blended members (DB, solid lines) and a selection of indirectly blended members that were initialised five hours after blending (IDB, dashed lines). Radar value was calculated by averaging across all events in the trial period and does not have lead-time dependence. [Colour figure can be viewed at wileyonlinelibrary.com]

when averaged over all members. Note that this spindown behaviour is atypical when compared with operational outputs, possibly due to the effects of time lagging or the more limited amount of data in our trial. Regardless, the blended DB members show consistently larger average precipitation up to $T + 12$ h than any of the reference or IDB precipitations. Recall from Section 2.2 that the DB set is also the set that ingests new lateral boundary conditions from the global ensemble. Therefore, if there was a large disparity between the reference DB and reference IDB members, this would imply that the ingestion of new lateral boundary conditions was a predominant cause for the larger values. With the exceptions of $T + 6$ and $T + 8$ h, this is largely not the case. Therefore, LSB has a clear impact on the total accumulations for the directly blended members, meaning that spurious precipitation may be present. This effect has largely vanished five hours after blending occurs.

3.3 | Impact of LSB on spatially integrated spread–skill relationship

The variation in FSS across the trial period is shown in Figure 6 as a function of validity time and lead time. Each hourly cycle is included in these panels, with the forecast associated with a given cycle tracking along the diagonal. We choose a lead time cutoff of 24 h based on previous LSB work with the deterministic UKV model, which demonstrated that the blending signal persists for approximately 18 h (Milan *et al.*, 2023). Additional work presented later in this section supports this cutoff lead time. All panels show the FSS for the 90th centile and for a neighbourhood size of 44 km (width of 19 grid points), the neighbourhood size at which both ensembles exceed skill scores of 0.5 in Figure 7. Figure 6a shows the dFSS (spread) scores for the reference ensemble, where higher values mean more confidence and lower spread. Scores are variable over the trial period, with a notable period of higher confidence occurring towards the end of June and start of July. Typically, higher confidence is achieved at shorter lead times, as expected.

Next, Figure 6b shows the difference between the dFSS (spread) and eFSS (skill) scores for the reference ensemble. Some eFSS values are missing due to a pre-filter check, which ensures that at least 0.2% of the domain grid points contain precipitation above the percentile threshold. This check is typically failed with exceptionally small precipitation coverage, whereby there are far fewer grid points with nonzero precipitation to meet the requested centile fully. Typically, the lower skill regions occur during the extended dry period from June 26–29, which has been filtered out. There is no clear dependence of the correctness-of-spread on lead time, with some events becoming more correctly spread at shorter lead times and others becoming less correctly spread. It is also difficult to say from this representation of the data whether the reference ensemble overall is underspread or overspread.

Figure 6c,d shows the difference between the blended and reference ensembles for the dFSS and eFSS, respectively. Larger values mean the blended ensemble had higher scores (lower spread or larger skill). Score differences are much smaller than those in Figure 6b, which is expected given the large similarity between fields shown in Figure 2. One notable exception occurs towards the end of June 17 and start of June 18, when the blended ensemble is both more skilful and more confident. These score increases are persistent across all lead times included. The changes in dFSS and eFSS due to blending tend to have the same sign, but it is difficult to determine the overall effect.

While these contour plots are useful for understanding the broad variability of the FSS over the trial period, they only capture the influence of LSB for a single neighbourhood size and centile. Therefore, Figures 7–10 show the averaged values across the filtered (non-hatched, full-opacity) space of Figure 6, and the differences between these values, for neighbourhood sizes up to 900 km. Note that we chose to pool these FSS values by averaging the final scores, as in other studies (e.g., Sharma *et al.*, 2023; Woodhams *et al.*, 2018), rather than aggregating the

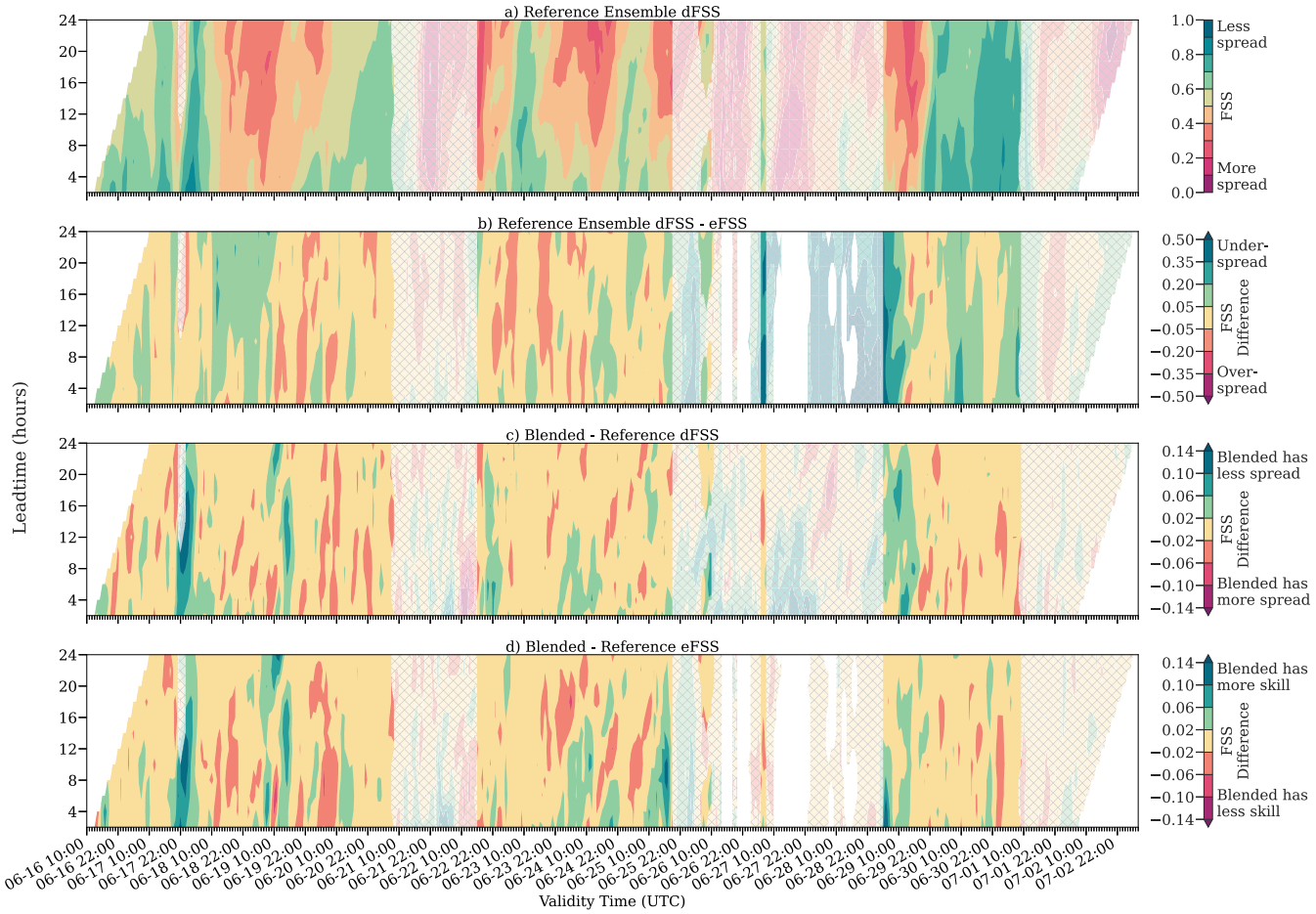

FIGURE 6 Contour plots showing evolution of 90th centile FSS with a 44-km (19 × 19 grid points) neighbourhood as a function of lead time and validity time over the trial period. A single cycle is along the diagonal. Hatched sections are dry events, which have been filtered out of the dataset in subsequent calculations (domain-averaged hourly accumulations less than 0.025 mm). (a) dFSS (spread) for the reference ensemble, (b) dFSS−eFSS (spread−skill) difference for the reference ensemble, (c) dFSS (spread) blended−reference difference, and (d) eFSS (skill) blended−reference difference. [Colour figure can be viewed at wileyonlinelibrary.com]

score components separately (Mittermaier, 2021). The sensitivity of the scores to this choice will be explored in future work.

To start, Figure 7 shows the expected increases in FSS with neighbourhood size (Roberts & Lean, 2008). The dFSS (spread) scores are higher than the eFSS (skill) scores for both the 90th and 97.5th centiles, hence both ensembles were underspread for both centiles considered. However, the ensembles are less underspread using the larger centile, suggesting that the biggest contributor to the overconfidence is in the lighter rain.

Differences between the two ensembles are difficult to distinguish from this presentation of the data, so Figure 8 shows the percentage differences between them. This percentage difference is calculated as 100(dFSS − eFSS)/eFSS, where larger percentages means more underspread. The underspread values peak at a neighbourhood size of approximately 50 km and steadily become more correctly spread for larger neighbourhoods. The blended ensemble is less underspread than the reference ensemble across all neighbourhoods and for both centiles. The 90th centile shows the smallest improvements to the spread from blending, with the blended ensemble being less underspread by only 0.2% at most. The largest sustained difference in the 97.5th centile approaches 0.4% at scales similar to the wavelength where LSB begins to blend fields, 400 km. There is also a large improvement in the spread–skill relationship closer to the grid scale at this higher centile.

We can infer from Figures 7 and 8 that LSB has had a larger impact on ensemble skill than spread, and in particular that LSB has decreased spread. This is because the blended ensemble is less underspread than the reference ensemble despite all dFSS and eFSS curves of Figure 7 increasing when LSB is applied. To see this explicitly, the solid lines of Figure 9 show the difference between the blended and reference ensembles for the dFSS (spread) and eFSS (skill) for the 90th and

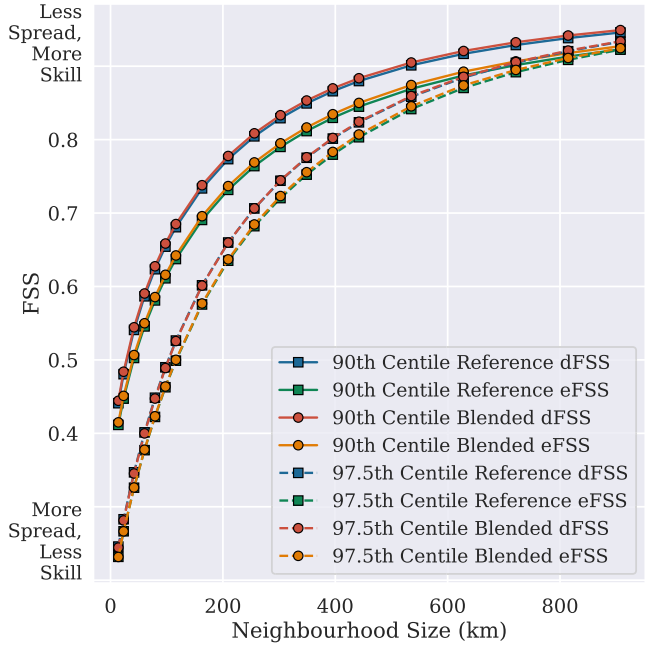

FIGURE 7 Scale-dependent dFSS (spread) and eFSS (skill) curves obtained by averaging FSS values similar to those presented in Figure 6 a over the trial period and over lead times $T + 2$ to $T + 24$ h. The reference and blended curves are difficult to distinguish, especially for the 97.5th centile. [Colour figure can be viewed at wileyonlinelibrary.com]

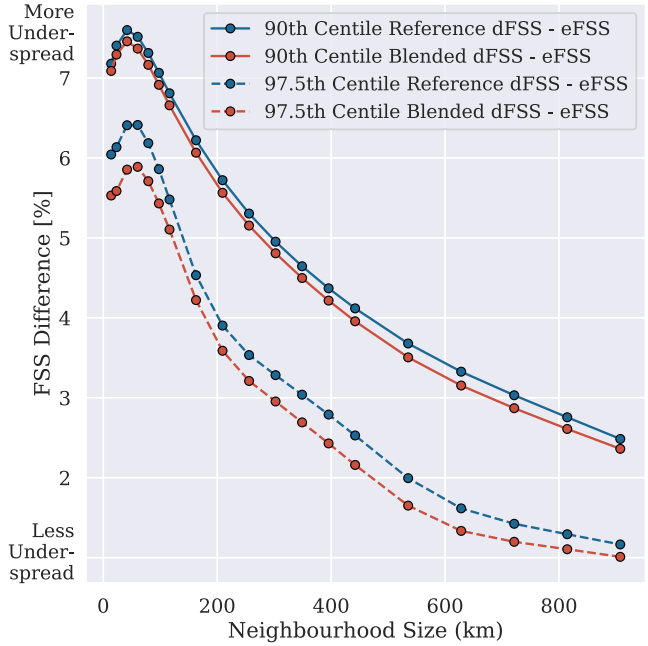

FIGURE 8 Scale-dependent percentage differences between dFSS and eFSS averages presented in Figure 7. eFSS averages are used for normalisation. [Colour figure can be viewed at wileyonlinelibrary.com]

97.5th centiles. Larger values in these plots show that the blended ensemble has larger skill scores (larger skill) or larger spread scores (smaller spread) than the reference ensemble. Also included in these figures as dashed lines are the mixed-member 95% confidence estimations as described in Section 2.5 and Appendix A.

Both spread and skill score differences between the blended and reference ensembles are comfortably larger than the 95% confidence level for the 90th precipitation centile data over all neighbourhood sizes, indicating significant results. In fact, skill score increases are larger than spread score increases across all neighbourhood sizes and centiles. Above 400-km neighbourhood size, LSB increases skill scores by an average of 0.56% in the 90th centile data, while spread scores only increase by 0.41%. Similarly, in the 97.5th centile data, skill scores above 400 km increase by an average of 0.37%, while spread scores only increase by 0.093%. Note, however, that these percentage increases are even larger towards the grid scale for the 90th centile due to the smaller normalisations (Figure 7), with a maximum skill increase of 0.84% observed at the smallest neighbourhood. Ultimately, though, the larger increase in skill scores shows that the correctness-of-spread improvements with blending are caused by increases in skill outweighing decreases in spread.

Interestingly, the dependence of LSB impact on neighbourhood size is different for the two centiles. Whereas blending leads to a fairly uniformly significant response across neighbourhood size for the 90th centile data, for the 97.5th centile the spread differences are never significant and the skill differences only significant for neighbourhood sizes larger than 200 km. Additionally, the blended ensemble has more spread than the reference ensemble at the grid scale using the 97.5th centile data, before becoming slightly less spread at neighbourhood sizes larger than 200 km. The smaller sample size for the 97.5th centile inherently lends itself to larger confidence intervals than the 90th centile, but, given the variability of the confidence intervals, is much smaller than that of the blended−reference profiles, this is not the predominant factor. Note that the 95th centile FSS curves were also investigated and were found to resemble a smoothly varying transition between the 90th and 97.5th centile data presented here.

The results presented in Figure 9 are averages across all lead times and validity times of the trial period, but it is also instructive to interrogate the lead-time dependence of the LSB response. Figure 10 has the same format as Figure 9 but shows the results using the 90th centile for specific lead time ranges. Given the fact that blending only modifies the initial conditions, we expect it to have the largest impact at early lead times, and this is indeed verified in this figure, with maximum skill score differences approaching 0.01 between lead times $T + 2$ and $T + 6$ h. This difference diminishes with increasing lead

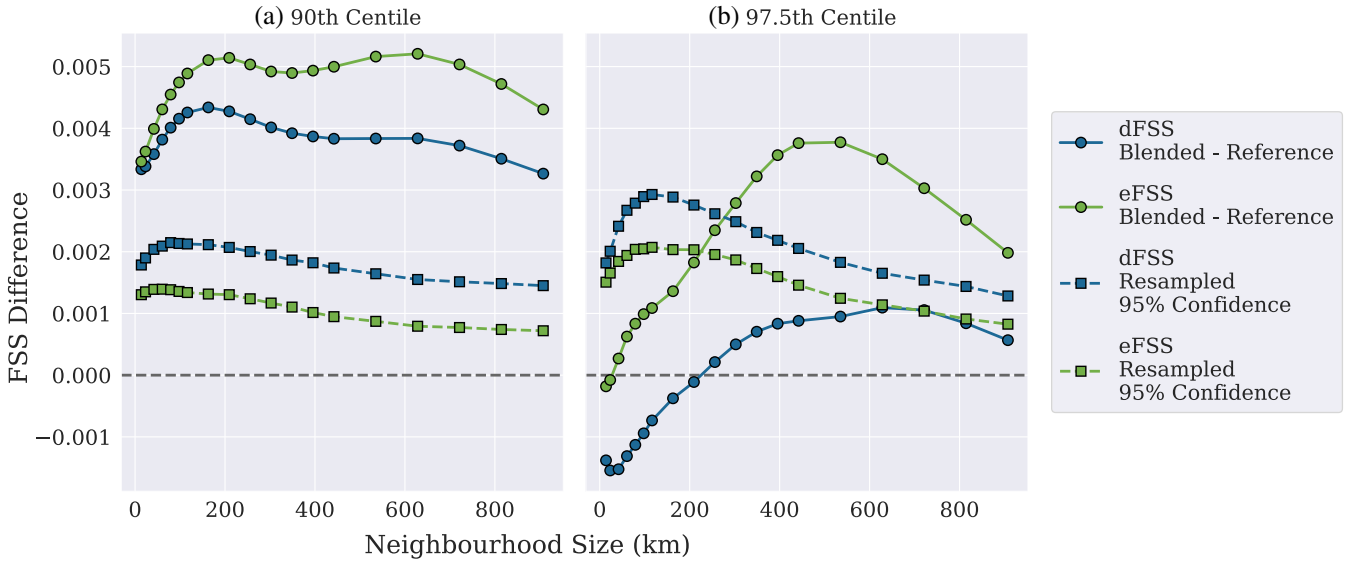

FIGURE 9 Scale-dependent FSS difference between the blended and reference ensembles for the spread (dFSS) and skill (eFSS) averages presented in Figure 7. Averages are performed over lead times $T + 2$ to $T + 24$ h. Solid line, circle markers: difference between the blended and reference FSS averages. Dashed line, square markers: upper limit of the 95% confidence estimated through constrained resampling technique. [Colour figure can be viewed at wileyonlinelibrary.com]

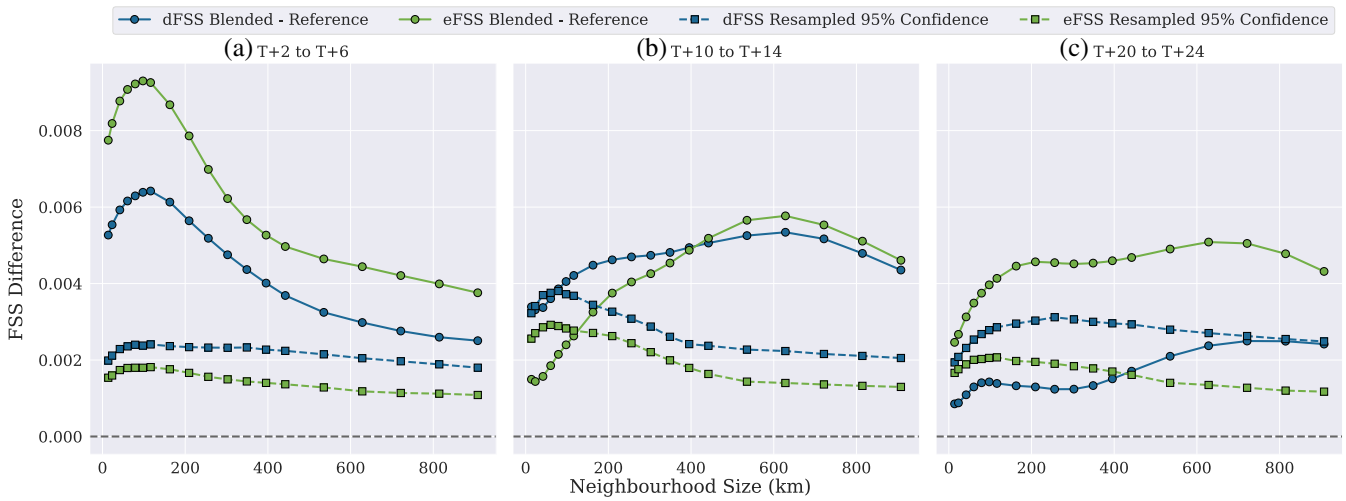

FIGURE 10 As Figure 9 but for the 90th centile only, separated by lead time. [Colour figure can be viewed at wileyonlinelibrary.com]

time for both spread and skill scores. The spread-score ensemble differences become insignificant across all neighbourhoods for lead times longer than $T + 20$ h. There is also a scale dependence observed in these results, where score differences are larger towards the grid scale than synoptic scales at very short lead times (Figure 10a), before decaying and becoming less significantly different compared with larger neighbourhood sizes at longer lead times (Figure 10b,c). We speculate that this is related to the inherent growth rate of small-scale errors being much faster than that of large-scale ones (Lorenz, 1969). We also notice a lead-time dependence in the skill improvements provided by LSB, with scores deteriorating at the grid scale between lead times $T + 12$ and $T + 18$ h, before recovering and becoming significant again at lead times longer than $T + 20$ h. This dependence can be partly observed in Figure 10b, with insignificant skill-score improvements observed below neighbourhood scales of approximately 120 km. We do not have a clear explanation for this behaviour, though it may simply be a consequence of the limited data used for this analysis.

In summary, the results from this section show the following.

1. On applying LSB, the skill of the lower hourly accumulation centile (including lighter and heavier precipitation) is improved by more than the skill of the higher centile (including just the heaviest precipitation) (Figure 9). This skill improvement is significant across all neighbourhood sizes for the lower centile, while the higher centile is only significantly improved for neighbourhood sizes above 200 km.
2. These skill improvements are accompanied by a significant decrease in spread for the lower centile, and a more modest, insignificant decrease in spread for the higher centile (Figure 9). This is true across all neighbourhood sizes of the lower centile and neighbourhood sizes above 200 km for the higher centile.
3. Given the context that the ensembles are underspread, these score increases show that the improvements in the spread–skill relationship in both centiles come from increases in skill scores outweighing increases in spread scores (degradations in spread). LSB corrects the spread–skill relationship for the higher centile more than for the lower centile (Figure 8).
4. The largest, consistent spread–skill improvements occur at the neighbourhood sizes where blending begins to modify fields, scales of approximately 400–500 km. Large improvements are also observed towards the grid scale for the higher centile (Figure 8).
5. The impact of LSB on ensemble spread persists for approximately 18–20 h. The ensemble is made more skilful until at least 24 h after initiation, although smaller improvements are noted between lead times $T+12$ and $T+18$ h for smaller neighbourhood sizes (Figure 10).

3.4 | Case study of the spatial impact of LSB

The previous section has shown that LSB improves the spread–skill relationship across all scales and precipitation intensities. The impacts at larger scales are expected, but the reasons for the spread–skill improvements towards the convective scale are not as immediately obvious. This case study shows an example of these downscale improvements and provides context to help interpret the previous FSS results. This case-study period runs from June 29, 2019, 1700 UTC to June 30, 2019, 0500 UTC, and was selected due to the presence of spatially separated synoptic-scale and regional-scale weather features. We chose to analyse the ensembles initialised at 1500 UTC (comprising cycles from 1000–1500 UTC), with lead times $T+2$ to $T+13$ h for the members initialised at 1500 UTC. This choice was made because the "directly blended" members are the freshest in these ensembles (i.e., have the shortest lead times). Longer lead-time studies were considered to be less informative, given the short persistence of the LSB signal (Section 3.3).

The case-study conditions are shown in Figure 11. At the start of the period, a band of thunderstorms

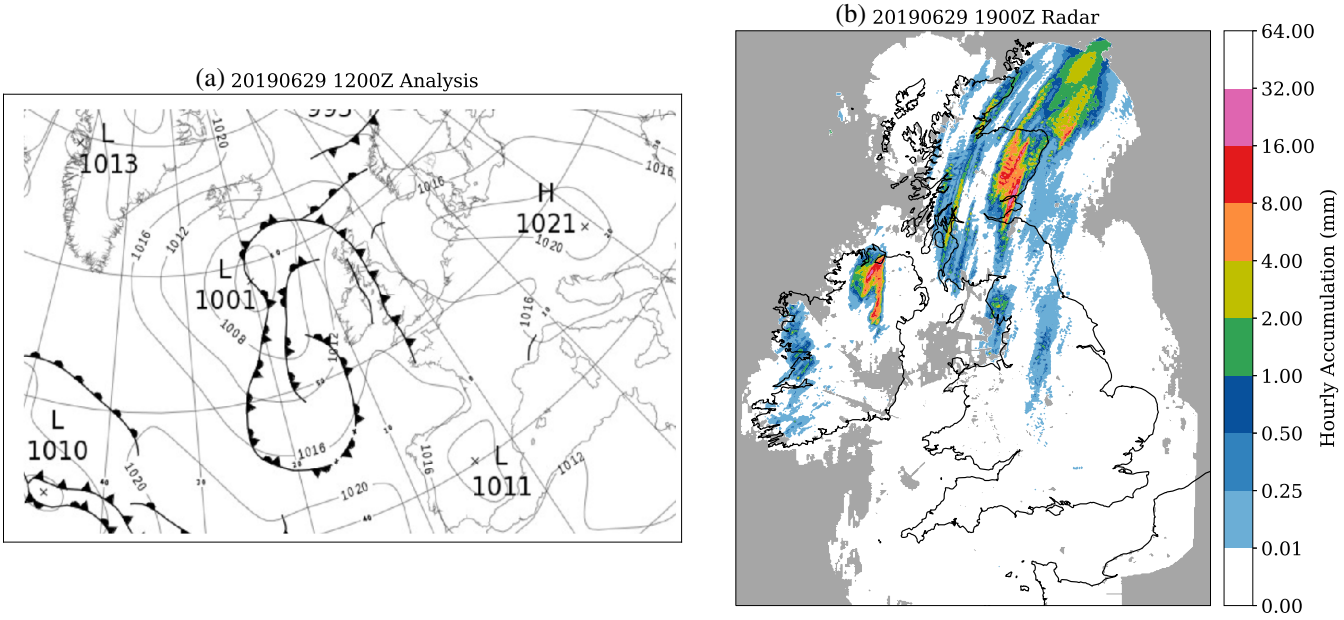

FIGURE 11 Synoptic overview of the case-study period: (a) synoptic chart from the Met Office daily weather summary (UKMO, 2019) for the closest period before the integration window used in the LFSS case study. Contours show mean sea-level pressure in 4-hPa intervals. (b) Hourly rain radar valid at June 29, 2019, 1900 UTC, chosen to show the time and location of the strongest period of elevated convection over Ireland. [Colour figure can be viewed at wileyonlinelibrary.com]

associated with a cold front was advecting over Scotland and northern England. The precipitation intensified as the band pushed into Scotland, and by midnight on June 30, 2019 all members correctly predicted maximum accumulations of more than 8 mm. Both ensemble confidence and skill increased as this band cleared into the North Sea. At the same time, a westerly moving occlusion brought warm, moist air aloft to the west coast of Ireland. Upper-level vorticity advection initiated a line of convection beginning on June 29, 2019 1600 UTC, moving northeastwards. Convective intensity reached a maximum at 1900 UTC over Northern Ireland, after which the forcing region overtook the convection and accumulations reduced. This convection was identified as elevated by forecasters (David Flack, personal communication, 2023), a situation that models have struggled with capturing in the past (Flack *et al.*, 2023). The rest of the United Kingdom was largely dry and settled during this period.

If we assert that a reliable ensemble should colocate regions of high dLFSS and eLFSS, the reference ensemble shown in Figure 12 largely meets this requirement for this case. Greater confidence and skill is shown over eastern Scotland (region 1) than in other areas of the domain, which is expected given the large-scale forcing driving precipitation in this region. Similarly, the ensemble was less confident and consequently less skilful in the location of elevated convection over Northern Ireland (region 3), which is expected given the lower predictability of this type of convection. Over southern Scotland and northern England (region 2), however, the ensemble is incorrectly confident about the convection in the trailing edge of this rain band. Most ensemble members initiated convection in northwest England, which was too intense and slightly too early. This mistiming caused spatial mismatch between model and radar fields within the 12-h window, leading to lower eLFSS scores over this region. Overall, however, the mistimed convection over region 2 is the only bust in an otherwise reliable forecast.

To assess the impact of LSB on this case study, Figure 13 shows the sensitivity of the difference between the blended and reference ensembles to the centile and neighbourhood size. Over differing parts of Scotland (region 1), the blended ensemble has increased spread and decreased skill for the centiles and neighbourhood sizes considered. However, given the already high scores associated with this region, this result suggests that LSB has had a minor impact on the spread–skill relationship of precipitation enhanced through predictable means. On the other hand, there is a notable increase in spread over northern England (region 2) in the data using the larger centile, indicating that the blending is somewhat correcting the

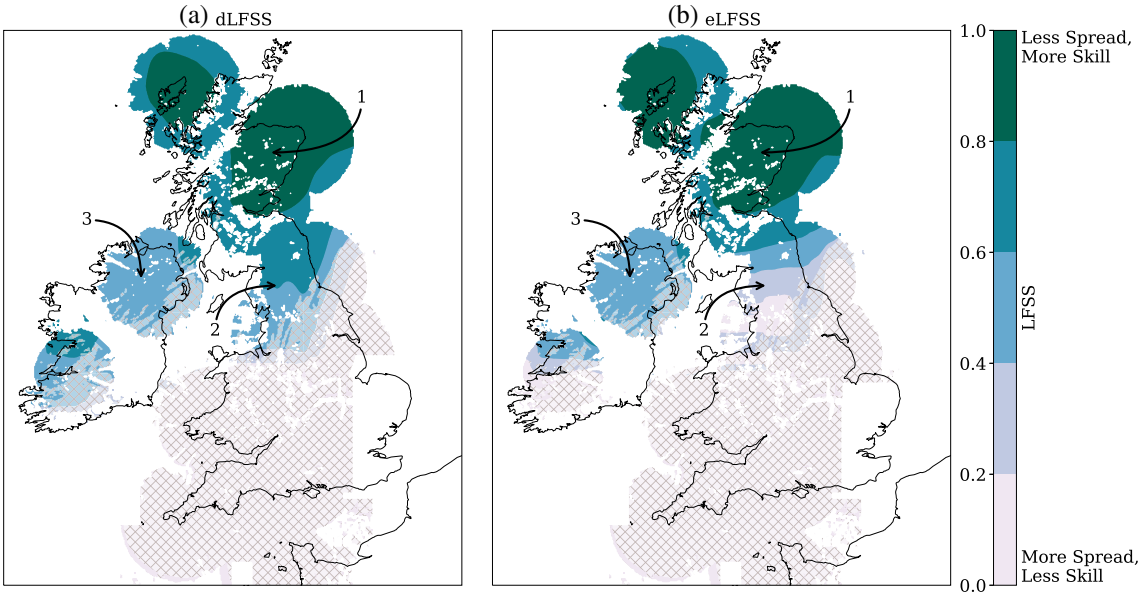

FIGURE 12 97.5th-centile, 260-km neighbourhood size (112 × 112 grid points) dLFSS (left) and eLFSS (right) for the reference ensemble calculated over the case-study period (June 29, 2019, 1700 UTC to June 30, 2019, 0500 UTC). If data are missing for a grid point in any of the fractions fields used to create these LFSS maps (which occurs when there are insufficient radar returns for a given hourly accumulation), that point is masked to ensure fairness. Darker areas in the plots indicate regions where the fractions fields were similar across all lead times and between all ensemble members (dLFSS) or between all members and radar (eLFSS). Lighter regions are where there was either large disagreement between members across all lead times or little precipitation. To make the distinction clear between the latter two cases, hatching is applied to any grid point where the total accumulation over the 12-h period is less than 1 mm in all three datasets (both ensembles and radar). Annotated regions are the focus of analysis in the text. [Colour figure can be viewed at wileyonlinelibrary.com]

overconfident forecast of convection. This is especially clear in the 97.5th-centile, 260-km neighbourhood size dLFSS map (Figure 13e), which shows blended scores that have decreased by over 0.12 in places. The areas with the largest increase in spread colocate with areas of improved skill scores, meaning the spread–skill relationship has improved.

The largest impact of LSB is observed for the case of elevated convection over Northern Ireland (region 3). There is a clear signal in all panels of Figure 13 of increased skill and decreased spread. Figure 13a,b shows increased scores for the spread and skill of the 90th centile (absolute threshold of approximately 0.5 mm over the integration period). Improvements in skill for this lower centile suggests that the blending has positioned the broad precipitation envelope more accurately. This improvement in skill is observed in all ensemble members, meaning they now have greater similarity and increased dLFSS scores (lower spread). However, the strongest impact of LSB is with the most intense rain, as can be seen in the 97.5th centile (2–5 mm absolute threshold) plots of Figure 13c–f. For instance, Figure 13d shows maximum skill-score improvements of more than 0.2, while Figure 13f shows a sustained improvement of over 0.12 using a neighbourhood size comparable to the east–west extent of Ireland. Blending has preferentially improved the location of the more intense precipitation. The spread scores have also increased over this area, keeping the spread–skill relationship broadly correct.

Taken together, these maps suggest that the blending has helped to represent smaller scale features over Ireland and northern England more accurately, which we attribute to improvements in the location of the synoptic-scale features providing the forcing. From inspection of the members, the blended ensemble has correctly shifted the elevated convection over Ireland to the northeast compared

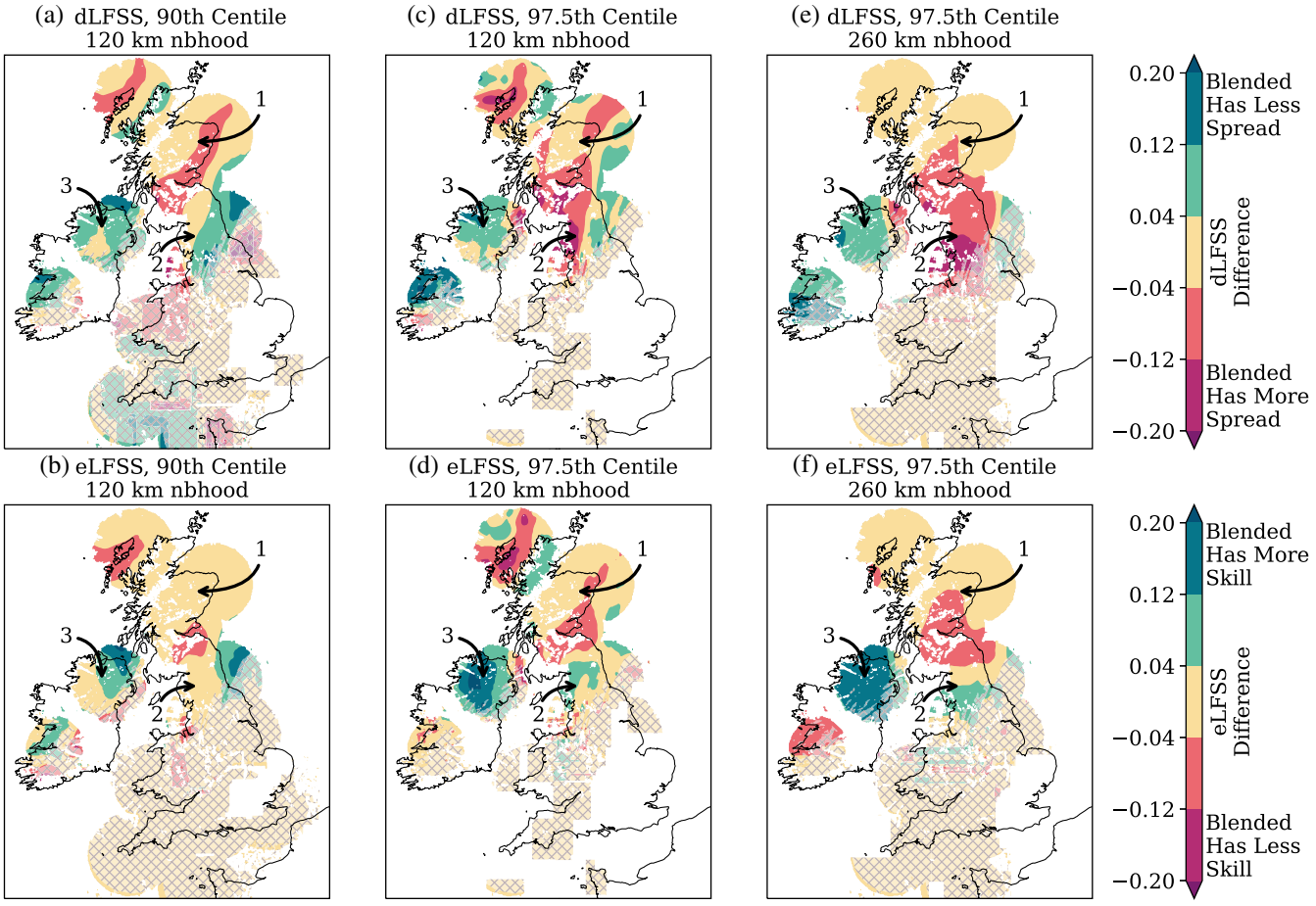

FIGURE 13 Blended−reference LFSS differences for a selection of centiles and 120-km (51 × 51 grid points) and 260-km (112 × 112 grid points) neighbourhoods. For the dLFSS difference maps, higher scores show regions where the blended ensemble has larger dLFSS (lower spread) than the reference ensemble. For the eLFSS difference maps, higher scores show regions of improved skill. Note that these metrics can produce sharp artifacts over areas of little precipitation, due to the discontinuous square neighbourhoods used when calculating fractions fields. These artifacts do not appear over precipitating regions. Hatching denotes areas that received less than 1 mm of precipitation in all three datasets (both ensembles and radar) over the integration window. [Colour figure can be viewed at wileyonlinelibrary.com]

with the reference ensemble. This is similar to the displacement made to the convection over northern England, which is associated with improvements in the timing of the convective initiation. This increase in skill is associated with an increase in spread over the deficient areas of northern England, but a decrease in spread over Ireland. Despite the existing spread–skill relationship of the reference ensemble being reasonable, blending has improved the most deficient areas and predicted elevated convection more confidently.

4 | CONCLUSIONS

This study investigated the impact of large-scale blending on the spread–skill relationship of hourly precipitation accumulations within the Met Office convective-scale ensemble, MOGREPS-UK. We hypothesised that LSB would improve the spread–skill relationship by preferentially increasing ensemble skill compared with spread. In a 17-day summer trial period, LSB improved the spread–skill relationship across all scales and precipitation thresholds, with the largest corrections of up to 0.4% noted for neighbourhood sizes above 400 km for the 97.5th centile threshold (note that 400 km is also the scale at which LSB begins to blend the host model into the regional model forecast). When interrogated further, these spread–skill corrections are caused by skill scores being improved (eFSS increased) by more than spread scores have deteriorated (dFSS increased). In the 90th-centile results, for instance, LSB affected both skill and spread scores significantly across all neighbourhood sizes, but skill scores improved by an average of 0.56% for the largest neighbourhood sizes, while spread scores only increased (i.e., spread was degraded) by an average of 0.41%. Spread scores in the 97.5th-centile results were not significantly different at any scale with LSB applied, while skill-score improvements were significant above 200-km neighbourhoods. Typically, LSB resulted in spread–skill improvements across all scales considered, not just the scales that had been blended. A novel extension of the LFSS demonstrated how these spread–skill improvements transfer to smaller scale features. By improving the synoptic-scale flow, the blended ensemble corrected an overconfident case of convection and improved performance with elevated convection. This is a particularly promising result given the historical difficulty of modelling elevated convection (Flack *et al.*, 2023).

This work has focused on assessing improvements to the spread–skill relationship only, since we found negligible impacts on reliability curves, rank histograms, and ROC area (not shown). We also note a large seasonal dependence to the impacts of LSB. This work only presents findings from the summer trial, since the results of the winter trial showed minimal changes. Inspection of the dominant regimes and precipitation totals within the winter trial period reveals that the weather was, on average, more vigorous and larger scale compared with the summer trial, and was therefore less sensitive to domain-scale corrections. This seasonal dependence is consistent with the previous deterministic LSB study, which showed much stronger improvements in the FSS results for summer than winter (Milan *et al.*, 2023). While the authors do not quote specific differences from the FSS with LSB applied, the results presented in fig. 16 of Milan *et al.* (2023) comparing forecasts with and without LSB appear to be similarly modest yet significant. Our work has shown that skill improvements in deterministic models extend to the convective-scale ensemble that recentres its members around these high-resolution analyses, though these improvements are still only modest.

Using LSB within convective-scale ensembles shows promising improvements to the spread–skill relationship, but these improvements are limited by a corresponding degradation in spread. Previous studies with convective-scale ensemble blending found similar skill improvements to this work but opposite spread responses (Keresturi *et al.*, 2019; Wang *et al.*, 2011; Zhang *et al.*, 2015). However, in these studies, blending was either incorporated alongside other model improvements or applied more holistically across the ensemble initiation. This work has been performed using an ensemble that only applied blending to the UKV background providing the initial conditions. Blending is not applied to the initial-condition perturbations or lateral boundary conditions provided by the host ensemble. While we should expect the synoptic scales of the host ensemble to be in better agreement with the blended analysis than the unblended analysis, some tension will inevitably remain, which may limit subsequent divergence between members.

Additionally, our study has shown that the impacts of LSB on ensemble spread persist for approximately 18–20 h from forecast initiation. This is in broad agreement with other work that has investigated the persistence of blending (Wang *et al.*, 2014) and the persistence of initial-condition perturbations in UK regional models (Porson *et al.*, 2020; Tennant, 2015). Other studies over the United States, however, have found that blending instead has a stronger response at later lead times than those seen in this work (Schwartz *et al.*, 2021, 2022). This difference may be partly due to the implementation of blending, with these studies applying blending to the analysis rather than integrating blending into the DA scheme itself. Additionally, we would expect the use of a much larger domain size to extend the persistence of blending, as it would take longer for the influence of the lateral boundary conditions

to become dominant over the initial conditions. Assessing the sensitivity of the LSB response to domain size is outside the scope of this work.

We have also observed signs of spurious precipitation spin-up within the members where LSB was applied directly, which is consistent with other works evaluating blending (Schwartz *et al.*, 2021). Any future work aiming to increase the frequency with which LSB is applied to MOGREPS-UK should be aware of the effect that this may have on the other ensemble members, which currently only inherit the effect of blending through the ingestion of blended background fields. It may be possible to improve skill further by applying LSB more frequently, but it is difficult to assess this using the currently available data because the six-hourly blending was applied at the same time as lateral boundary conditions were updated.

LSB has been shown to improve skill and the spread–skill relationship within this convective-scale ensemble in summer and we encourage the Met Office to continue developing this technique. Further improvements should focus on counteracting the associated reduction in spread, possibly by implementing LSB more frequently than every six hours or applying LSB more completely within the ensemble initiation process.

ACKNOWLEDGEMENTS

We thank the Prioritized Evaluation Group on Ensembles for kindly providing the FSS code used in this work. Specifically, we thank Anne McCabe, who developed and tested this code, with advice from Nigel Roberts. We also thank David Walters and Joanne Carr for providing the code used to adapt the MOGREPS-UK schematic in Figure 1. Finally, we express our appreciation to both reviewers for providing comments that improved the technical and presentation quality of this manuscript. Adam Gainford's contribution was funded through a Natural Environment Research Council (NERC) studentship with CASE support from the Met Office (NE/S007261/1).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

Model outputs are archived at the Met Office. Neighbourhood post-processing code is available through the open-source IMPROVER repository (Roberts *et al.*, 2023).

ORCID

Adam Gainford https://orcid.org/0000-0003-2484-8316
Suzanne L. Gray https://orcid.org/0000-0001-8658-362X
Thomas H. A. Frame https://orcid.org/0000-0001-6542-2173
Aurore N. Porson https://orcid.org/0000-0002-5023-8522
Marco Milan https://orcid.org/0000-0002-9309-5365

REFERENCES

AMS. (2023) Drizzle—glossary of meteorology American Meteorology Society. https://glossary.ametsoc.org/wiki/Drizzle

Beck, J., Bouttier, F., Wiegand, L., Gebhardt, C., Eagle, C. & Roberts, N. (2016) Development and verification of two convection-allowing multi-model ensembles over Western Europe. *Quarterly Journal of the Royal Meteorological Society*, 142, 2808–2826. Available from: https://doi.org/10.1002/qj.2870

Ben Bouallègue, Z., Theis, S.E. & Gebhardt, C. (2013) Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, 22, 49–59. Available from: https://www.schweizerbart.de/papers/metz/detail/22/79824/Enhancing_COSMO_DE_ensemble_forecasts_by_inexpensi?af=crossref

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., Rooy, W.d. et al. (2017) The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Monthly Weather Review*, 145, 1919–1935. Available from: https://journals.ametsoc.org/view/journals/mwre/145/5/mwr-d-16-0417.1.xml

Buizza, R. (1997) Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Monthly Weather Review*, 125, 99–119. Available from: https://journals.ametsoc.org/view/journals/mwre/125/1/1520-0493_1997_125_0099_pfsoep_2.0.co_2.xml

Bush, M., Boutle, I., Edwards, J., Finnenkoetter, A., Franklin, C., Hanley, K. et al. (2023) The second Met Office unified model–JULES regional atmosphere and land configuration, RAL2. *Geoscientific Model Development*, 16, 1713–1734. Available from: https://gmd.copernicus.org/articles/16/1713/2023/

Cafaro, C., Woodhams, B.J., Stein, T.H.M., Birch, C.E., Webster, S., Bain, C.L. et al. (2021) Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa. *Weather and Forecasting*, 36, 697–716. Available from: https://journals.ametsoc.org/view/journals/wefo/36/2/WAF-D-20-0172.1.xml

Caron, J.-F. (2013) Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: a case study. *Monthly Weather Review*, 141, 356–374. Available from: https://journals.ametsoc.org/view/journals/mwre/141/1/mwr-d-12-00051.1.xml

Clark, A.J., Kain, J.S., Stensrud, D.J., Xue, M., Kong, F., Coniglio, M.C. et al. (2011) Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Monthly Weather Review*, 139, 1410–1418. Available from: https://journals.ametsoc.org/view/journals/mwre/139/5/2010mwr3624.1.xml

Dey, S.R.A., Leoncini, G., Roberts, N.M., Plant, R.S. & Migliorini, S. (2014) A spatial view of ensemble spread in convection permitting ensembles. *Monthly Weather Review*, 142, 4091–4107. Available from: https://journals.ametsoc.org/view/journals/mwre/142/11/mwr-d-14-00172.1.xml

Dey, S.R.A., Plant, R.S., Roberts, N.M. & Migliorini, S. (2016) Assessing spatial precipitation uncertainties in a convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, 142, 2935–2948. Available from: https://doi.org/10.1002/qj.2893

Feng, J., Chen, M., Li, Y. & Zhong, J. (2021) An implementation of full cycle strategy using dynamic blending for rapid refresh short-range weather forecasting in China. *Advances in Atmospheric Sciences*, 38, 943–956. Available from: https://doi.org/10.1007/s00376-021-0316-7

Feng, J., Sun, J. & Zhang, Y. (2020) A dynamic blending scheme to mitigate large-scale bias in regional models. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001754. Available from: https://doi.org/10.1029/2019MS001754

Ferrett, S., Frame, T.H.A., Methven, J., Holloway, C.E., Webster, S., Stein, T.H.M. et al. (2021) Evaluating convection-permitting ensemble forecasts of precipitation over Southeast Asia. *Weather and Forecasting*, 36, 1199–1217. Available from: https://journals.ametsoc.org/view/journals/wefo/36/4/WAF-D-20-0216.1.xml

Flack, D.L.A., Lehnert, M., Lean, H.W. & Willington, S. (2023) Characteristics of diagnostics for identifying elevated convection over the British Isles in a convection-allowing model. *Weather and Forecasting*, 38, 1079–1094. Available from: https://journals.ametsoc.org/view/journals/wefo/38/7/WAF-D-22-0219.1.xml

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B. & Ebert, E.E. (2009) Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24, 1416–1430. Available from: https://journals.ametsoc.org/view/journals/wefo/24/5/2009waf2222269_1.xml

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

Guidard, V. & Fischer, C. (2008) Introducing the coupling information in a limited-area variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 134, 723–735. Available from: https://doi.org/10.1002/qj.215

Hamill, T.M. (1999) Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, 14, 155–167. Available from: https://journals.ametsoc.org/view/journals/wefo/14/2/1520-0434_1999_014_0155_htfenp_2_0_co_2.xml

Hopson, T.M. (2014) Assessing the ensemble spread–error relationship. *Monthly Weather Review*, 142, 1125–1142. Available from: https://journals.ametsoc.org/view/journals/mwre/142/3/mwr-d-12-00111.1.xml

Hsiao, L.-F., Huang, X.-Y., Kuo, Y.-H., Chen, D.-S., Wang, H., Tsai, C.-C. et al. (2015) Blending of global and regional analyses with a spatial filter: application to typhoon prediction over the Western North Pacific Ocean. *Weather and Forecasting*, 30, 754–770. Available from: https://journals.ametsoc.org/view/journals/wefo/30/3/waf-d-14-00047_1.xml

Inverarity, G.W., Tennant, W.J., Anton, L., Bowler, N.E., Clayton, A.M., Jardak, M. et al. (2023) Met Office MOGREPS-G initialisation using an ensemble of hybrid four-dimensional ensemble variational (En-4DEnVar) data assimilations. *Quarterly Journal of the Royal Meteorological Society*, 149, 1138–1164. Available from: https://doi.org/10.1002/qj.4431

Keresturi, E., Wang, Y., Meier, F., Weidle, F., Wittmann, C. & Atencia, A. (2019) Improving initial condition perturbations in a convection-permitting ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 145, 993–1012. Available from: https://doi.org/10.1002/qj.3473

Klasa, C., Arpagaus, M., Walser, A. & Wernli, H. (2018) An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quarterly Journal of the Royal Meteorological Society*, 144, 744–764. Available from: https://doi.org/10.1002/qj.3245

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307. Available from: https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

McCabe, A., Swinbank, R., Tennant, W. & Lock, A. (2016) Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting. *Quarterly Journal of the Royal Meteorological Society*, 142, 2897–2910. Available from: https://doi.org/10.1002/qj.2876

Milan, M., Clayton, A., Lorenc, A., Macpherson, B., Tubbs, R. & Dow, G. (2023) Large-scale blending in an hourly 4D-Var framework for a numerical weather prediction model. *Quarterly Journal of the Royal Meteorological Society*, 149, 2067–2090. Available from: https://doi.org/10.1002/qj.4495

Milan, M., Macpherson, B., Tubbs, R., Dow, G., Inverarity, G., Mittermaier, M. et al. (2020) Hourly 4D-Var in the Met Office UKV operational forecast model. *Quarterly Journal of the Royal Meteorological Society*, 146, 1281–1301. Available from: https://doi.org/10.1002/qj.3737

Mittermaier, M.P. (2007) Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society*, 133, 1487–1500. Available from: https://doi.org/10.1002/qj.135

Mittermaier, M.P. (2021) A "meta" analysis of the fractions skill score: the limiting case and implications for aggregation. *Monthly Weather Review*, 149, 3491–3504. Available from: https://journals.ametsoc.org/view/journals/mwre/149/10/MWR-D-18-0106.1.xml

Nachamkin, J.E. & Schmidt, J. (2015) Applying a neighborhood fractions sampling approach as a diagnostic tool. *Monthly Weather Review*, 143, 4736–4749. Available from: https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-14-00411.1.xml

Porson, A.N., Carr, J.M., Hagelin, S., Darvell, R., North, 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. Available from: https://doi.org/10.1002/qj.3844

Porson, A.N., Hagelin, S., Boyd, D.F., Roberts, N.M., North, R., Webster, S. et al. (2019) Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Quarterly Journal of the Royal Meteorological Society*, 145, 3004–3022. Available from: https://doi.org/10.1002/qj.3601

Raymond, W.H. (1988) High-order low-pass implicit tangent filters for use in finite area calculations. *Monthly Weather Review*, 116, 2132–2141. Available from: https://journals.ametsoc.org/view/journals/mwre/116/11/1520-0493_1988_116_2132_holpit_2_0_co_2.xml

Raynaud, L. & Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143, 3037–3047. Available from: https://doi.org/10.1002/qj.3159

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C. et al. (2023) IMPROVER: the new probabilistic postprocessing system at the Met Office. *Bulletin of the American Meteorological Society*, 104, E680–E697. Available from: https://journals.ametsoc.org/view/journals/bams/104/3/BAMS-D-21-0273.1.xml

Roberts, N.M. & Lean, H.W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97. Available from: https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml

Schwartz, C.S., Poterjoy, J., Carley, J.R., Dowell, D.C., Romine, G.S. & Ide, K. (2022) Comparing partial and continuously cycling ensemble Kalman filter data assimilation systems for convection-allowing ensemble forecast initialization. *Weather and Forecasting*, 37, 85–112. Available from: https://journals.ametsoc.org/view/journals/wefo/37/1/WAF-D-21-0069.1.xml

Schwartz, C.S., Romine, G.S. & Dowell, D.C. (2021) Toward unifying short-term and next-day convection-allowing ensemble forecast systems with a continuously cycling 3-km ensemble Kalman filter over the entire conterminous United States. *Weather and Forecasting*, 36, 379–405. Available from: https://journals.ametsoc.org/view/journals/wefo/36/2/WAF-D-20-0110.1.xml

Schwartz, C.S., Romine, G.S., Smith, K.R. & Weisman, M.L. (2014) Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Weather and Forecasting*, 29, 1295–1318. Available from: https://journals.ametsoc.org/view/journals/wefo/29/6/waf-d-13-00145_1.xml

Sharma, K., Lee, J.C.K., Porson, A., Chandramouli, K., Roberts, N., Boyd, D. et al. (2023) Adaptive selection of members for convective-permitting regional ensemble prediction over the western maritime continent. *Frontiers in Environmental Science*, 11, 1281265. Available from: https://doi.org/10.3389/fenvs.2023.1281265

Tennant, W. (2015) Improving initial condition perturbations for MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 141, 2324–2336. Available from: https://doi.org/10.1002/qj.2524

UKMO. (2019) *Met Office daily weather summary June 2019*. Technical report.

Wang, H., Huang, X.-Y., Xu, D. & Liu, J. (2014) A scale-dependent blending scheme for WRFDA: impact on regional weather forecasting. *Geoscientific Model Development*, 7, 1819–1828. Available from: https://gmd.copernicus.org/articles/7/1819/2014/

Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F., Kann, A. et al. (2011) The central European limited-area ensemble forecasting system: ALADIN-LAEF. *Quarterly Journal of the Royal Meteorological Society*, 137, 483–502. Available from: https://doi.org/10.1002/qj.751

Wernli, H., Hofmann, C. & Zimmer, M. (2009) Spatial forecast verification methods Intercomparison project: application of the SAL technique. *Weather and Forecasting*, 24, 1472–1484. Available from: https://journals.ametsoc.org/view/journals/wefo/24/6/2009waf2222271_1.xml

Wilks, D.S. (1997) Resampling hypothesis tests for autocorrelated fields. *Journal of Climate*, 10, 65–82. Available from: https://journals.ametsoc.org/view/journals/clim/10/1/1520-0442_1997_010_0065_rhtfaf_2.0.co_2.xml

Woodhams, B.J., Birch, C.E., Marsham, J.H., Bain, C.L., Roberts, N.M. & Boyd, D.F.A. (2018) What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa. *Monthly Weather Review*, 146, 2757–2780. Available from: https://journals.ametsoc.org/view/journals/mwre/146/9/mwr-d-17-0396.1.xml

Yang, X. (2005) Analysis blending using spatial filter in grid-point model coupling. *Hirlam Newsletter*, 48, 49–55.

Zhang, H., Chen, J., Zhi, X., Wang, Y. & Wang, Y. (2015) Study on multi-scale blending initial condition perturbations for a regional ensemble prediction system. *Advances in Atmospheric Sciences*, 32, 1143–1155. Available from: https://doi.org/10.1007/s00376-015-4232-6

How to cite this article: Gainford, A., Gray, S.L., Frame, T.H.A., Porson, A.N. & Milan, M. (2024) Improvements in the spread–skill relationship of precipitation in a convective-scale ensemble through blending. *Quarterly Journal of the Royal Meteorological Society*, 1–21. Available from: https://doi.org/10.1002/qj.4754

APPENDIX A. CONSTRAINED RESAMPLING

Our method for quantifying the uncertainty in the measured LSB response is based on the question of whether the addition of blending creates an ensemble that is just a different sampling of the same underlying distribution, or whether the blended ensemble samples a different, more skilful distribution. If the two ensembles are drawn from the same distribution, then we expect the differences between the statistics of the blended and reference ensembles to be no different from the statistics of two ensembles obtained by swapping half of the members of the blended and reference ensembles. We therefore use a resampling estimate of the null distribution based on these "mixed-member ensembles" to estimate the significance of the differences between the blended and reference ensembles.

Mixing forecasts has been shown to be an effective method for estimating uncertainty (Hamill, 1999). However, because MOGREPS-UK is lagged, it is more accurate to describe the 18-member ensemble as a set of six three-member sub-ensembles, which can each be considered an i.i.d. sample. Therefore, we seek to isolate the response that occurs purely due to LSB, not that due to mixing members from sub-ensembles with different distributions. This requires the use of constraints, which only permutes members between the two ensembles that would otherwise form a mixed i.i.d. sub-ensemble, were it not for the use of LSB. The constraints that create the fairest comparison between mixed ensembles are the following:

1. only members that are initialised during matching cycles are permuted;
2. only the newly updated members at each hour are resampled; and

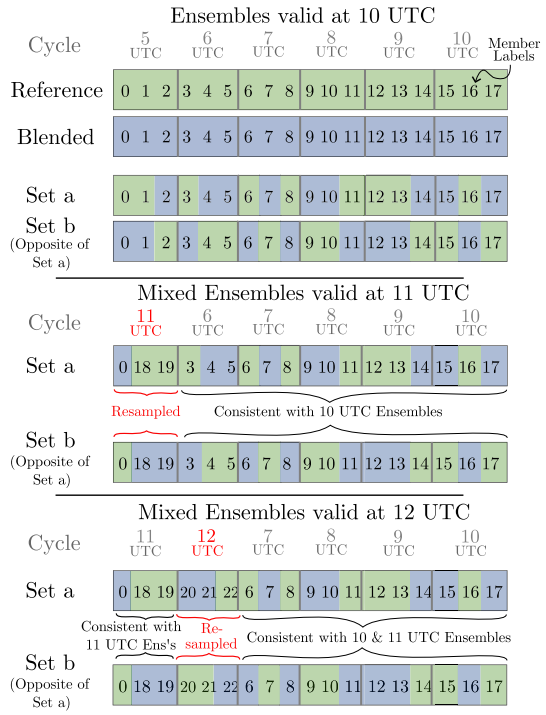

FIGURE A1 Schematic demonstrating an example of the ensemble member resampling process. The number within each box is the member label inherited from the corresponding MOGREPS-G member. Each hourly, time-lagged ensemble is comprised of the three-member sub-ensemble initialised during that cycle, along with the five previous the sub-ensembles. To create the first mixed-member ensembles ("set a" and "set b" at 1000 UTC), members are permuted between the three-member reference and blended sub-ensembles for each of the six sub-ensemble cycles. There is an alternating pattern of oversampling the reference or blended ensemble for each successive cycle, which ensures an equal mixing of reference and blended members. The mixed-member ensembles for the next hour (1100 UTC) are generated by fixing the permutations for those members common to the ensemble at the previous hour, resampling only the newly initialised members for that cycle. The resampling of the newly initialised members respects the oversample ordering, such that each 18-member mixed ensemble will always comprise an equal mix of reference and blended members. [Colour figure can be viewed at wileyonlinelibrary.com]

3. there are an equal number of members mixed between the reference and blended ensembles.

However, because three ensemble members are initialised each hour, criteria 2 and 3 cannot be compatible for each individual hour. Instead, an alternating pattern is applied, which oversamples the reference ensemble in one hour and then oversamples the blended ensemble in the next. This setup ensures that criterion 3 is met over the entire 18-member ensemble in the most fair way possible. Additionally, imposing criterion 2 ensures persistence between previously resampled sub-ensembles, which would introduce additional variance if otherwise neglected.

By stitching together sets of permutations between three-member sub-ensembles initialised in the same cycle, we are effectively performing a similar block bootstrap to that outlined in Wilks (1997); however, since we already have knowledge of the data structure and correlations, we do not need to anticipate some of the more user-dependent aspects of this method. Because these criteria were designed to replicate the construction of the lagged ensembles themselves most closely, they necessarily take into account additional variance that may be introduced through neglecting autocorrelations or through extra resampling, and ensure that the confidence limits are constructed by only considering the variance introduced by blending.

An example of the resampling process is shown graphically in Figure A1, where the numbers within each box represent the MOGREPS-UK member labels. Time-lagging of MOGREPS-UK members means that the ensemble valid at the next hour consists of 15 members that were in the previous hour, alongside three new members. This example represents just one of many possible ways of constructing a mixed ensemble using the outlined constraints. Therefore, the resampling process is repeated 1000 times to ensure robust confidence intervals can be constructed.

The confidence limits are estimated by performing the same filtered averages over validity time and lead time on the "set a" and "set b" mixed-member ensembles as for the blended and reference ensembles. Then, we calculate the difference between the two mixed-member ensemble averages. Finally, we take the upper 95th percentile across all 1000 resamples as our estimate of significance. Note that, by construction, we do not expect either "set a" or "set b" ensembles to include systematically larger values after averaging, so we present the absolute value of the averaged differences.