


RESEARCH ARTICLE

Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble

Aurore N. Porson¹  | Joanne M. Carr² | Susanna Hagelin^{2,3} | Rob Darvell² | Rachel North² | David Walters² | Kenneth R. Mylne² | Marion P. Mittermaier² | Steve Willington² | Bruce Macpherson²

¹MetOffice@Reading, Department of Meteorology, University of Reading, Reading, UK

²Met Office, Exeter, UK

³Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Correspondence

A.N. Porson, MetOffice@Reading, Department of Meteorology, University of Reading, Reading, RG6 7BE, UK.
Email: aurore.porson@metoffice.gov.uk

Abstract

In this article, we introduce a new configuration of the Met Office convective-scale ensemble for numerical weather prediction, for the Met Office Global and Regional Ensemble Prediction System over the United Kingdom (MOGREPS-UK). The new version, which became operational in March 2019, uses an hourly time-lagged configuration to take advantage of the hourly 4D-Var data assimilation run in the deterministic UK model with variable horizontal resolution, the UKV. An 18-member ensemble is created by running three members every hour and time-lagging these over a 6 hr window. This configuration is compared against the previous operational configuration, a 6-hourly convective-scale ensemble running 12 members. The main benefits of the time-lagged ensemble are to increase the ensemble size, to add small-scale uncertainties in the initial conditions and to generate more timely forecasts. The time-lagged configuration is shown to objectively improve the forecast at all lead times, with larger improvements in the first few hours. The improvement is seen in the ranked probability scores and is mainly associated with the improvements in the spread of the ensemble with an increase of about 5 to 10% in both summer and winter seasons. A larger ensemble size is necessary in the time-lagged configuration for it to outperform or maintain as good a performance against the previous 6-hourly configuration for all lead times. Alongside the update to an hourly configuration, the forecast length is more than doubled to 120 hr. Objective verification shows that the time-lagged configuration performs better than the high-resolution deterministic, UKV, and the global ensemble, MOGREPS-G, up to T + 120 hr. Increasing the size of the time-lagged ensemble through lagging over additional cycles leads to small but significant improvements, larger in most cases than those that can be obtained through neighbourhood processing.

KEYWORDS

Ensemble spread, time lagging, ensemble verification

1 | INTRODUCTION

Convective-scale ensembles are becoming increasingly important for operational numerical weather prediction (NWP) centres to help capture the uncertainty in forecasts of severe weather or in cases with weak synoptic forcing (Marsigli *et al.*, 2001; Gebhardt *et al.*, 2010; Kuhnlein *et al.*, 2014; Schwartz *et al.*, 2015; 2019; Golding *et al.*, 2016; Hagelin *et al.*, 2017; Raynaud and Bouttier, 2017; Klasa *et al.*, 2018; 2019; Frogner *et al.* 2019a; 2019b).

Recent studies have assessed the importance of different aspects of convective-scale ensemble design such as neighbourhood processing and increase in ensemble size (Raynaud and Bouttier, 2017; Schwartz and Sobash, 2017), generation of perturbations to the initial conditions (Kuhnlein *et al.*, 2014; Raynaud and Bouttier, 2016; Zhang, 2019), perturbations to the physics or stochastic physics (Schwartz *et al.*, 2010; Schumacher and Clark, 2014; McCabe *et al.*, 2016; Keil *et al.*, 2019; Zhang, 2019), increase in horizontal resolution (Hagelin *et al.*, 2017; Raynaud and Bouttier, 2017; Schwartz *et al.*, 2017), use of multi-model ensembles (Beck *et al.*, 2016) and use of multiple cycles via time-lagging (Raynaud and Bouttier, 2017).

The ideal perturbation strategy for convective-scale ensembles is not yet understood and this may depend on the synoptic forcing of the events (Flack *et al.*, 2018; Weyn and Durran, 2018; 2019; Keil *et al.*, 2019). Reviewing and intercomparing the characteristics of different convective-scale ensembles can help inform the most efficient and best performing system design for specific high-impact weather events (Clark *et al.*, 2018). However, the choice of ensemble design is likely to depend on the types of high-impact weather in which each NWP centre is most interested, as well as the resources available to run convective-scale ensembles.

Time-lagging of analyses has been seen as a theoretically grounded approach to generate ensembles from deterministic forecasts, either for medium-range forecasts (Brankovic *et al.*, 1990) or short-range forecasts (Mittermaier, 2007), the data assimilation (DA) cycles adding increments to the background that can be associated with ensemble perturbations.

However, when used on its own, this approach is limited with regard to the number of ensemble members it can generate by the cycling time.

Time-lagging of deterministic forecasts has been used successfully to create a small ensemble (Mittermaier, 2007; Yuan *et al.*, 2009; Kuchera and Rentschler, 2019; Xu *et al.*, 2019). Extension to time-lagging of ensemble forecasts has also been shown to be a valuable option (e.g. Raynaud and Bouttier, 2017), however it could benefit from a more frequent time-lagging option than a 6-hourly

cycle. A few centres, such as the Danish Meteorological Institute (DMI, Frogner *et al.*, 2019b) and the MetCoOp (personal communication, Ulf Andrae, 2019) are now starting to explore the benefits of hourly time-lagging within a convective-scale ensemble. Time-lagging introduces newer forecasts by replacing older forecasts. Older forecasts are important for the skill of the time-lagged ensemble particularly at early lead times, because they provide a larger sampling of the probability density distribution or pdf (Raynaud and Bouttier, 2017). Compared to a 6-hourly cycle, using an hourly cycle results in an ensemble of forecasts which are closer to their own data assimilation cycles and so potentially as skilful as each other, especially for fields which are less predictable such as clouds, visibility or precipitation. Hence, sampling an hourly pdf in comparison to a 6-hourly pdf should result in sampling more uncertainty in the forecasts whilst retaining the benefits of the latest data assimilation cycles. Including successive data assimilation cycles would also add different scales, that is, smaller scales, of perturbations into the ensemble, compared to large-scale perturbations from the parent ensemble, for example. Therefore, unlike the previous ensemble time-lagging study using 6-hourly time-lagging (Raynaud and Bouttier, 2017), the choice is made here to compare an hourly cycling of time-lagging directly against the operational non-time-lagged system.

An hourly time-lagged configuration is thus presented for the Met Office Global and Regional Ensemble Prediction System over the UK (MOGREPS-UK). MOGREPS-UK (Hagelin *et al.*, 2017) is centred on a deterministic analysis (Tennant, 2015) and takes its initial perturbations and lateral boundary conditions from the global ensemble MOGREPS-G (Bowler *et al.*, 2008; Flowerdew and Bowler, 2011; 2013; Brown *et al.*, 2012).

The new configuration is introduced here as an attempt to:

- Increase the ensemble size following Hagelin *et al.* (2017), in an operational context, distributing the required computational resources approximately evenly over a period of 6 hr.
- Take into account the hourly update in the 4D-Var analysis of the high-resolution deterministic UKV model (Milan *et al.*, 2020) and therefore to improve the spread in the initial conditions for the ensemble.
- Produce more timely forecasts.
- Release operational meteorologists from 6-hourly cycles in information availability, allowing them to spread production and produce guidance based on the latest information to hand – and not be constrained to await the next 6 hr cycle.

- Reduce the jumpiness of the forecasts (i.e. for when consecutive forecasts provide different information about an event) as each new cycle introduces a new small set of three ensemble members (thereby introducing slowly the impact of a new set of global perturbations and high-resolution analysis).
- Help operational meteorologists to build up a trend in the forecasts, for cases where the confidence about the timing and location of a specific event increases with the latest forecasts, helping to improve the definition of severe weather warning areas.

In addition to the time-lagging configuration changes, the forecast lead time is extended to 5 days. This is so that MOGREPS-UK can be compared against its parent ensemble (MOGREPS-G) and the deterministic model the UKV in their capability to predict severe weather at longer lead times.

In this article, our objectives are to describe the new hourly configuration and compare it with the previous operational system, to understand how an hourly time-lagging configuration compares to a 6-hourly configuration and to investigate the effect of different ensemble sizes on the performance to provide future directions for the development of MOGREPS-UK. Because the focus of this work is the comparison with the 6-hourly configuration, the full capability of the hourly cycling configuration (such as to reduce jumpiness) is not studied and this will be the subject of future work. In Section 2, we describe the new ensemble configuration and our methods for verifying a time-lagged ensemble. In Section 3, we focus on the objective verification to assess the skill, the reliability, the discrimination and the spread of the new time-lagged ensemble. In Section 4, we discuss the extension of the forecast range out to 5 days. To conclude, we summarise our work and suggest ideas for further research and improvements. A companion article will focus on case-study analysis to highlight the benefits of the new time-lagged configuration on product development and spread analysis for precipitation.

2 | DESCRIPTION OF THE HOURLY CONFIGURATION AND NEIGHBOURHOOD PROCESSING

2.1 | The new hourly configuration: Technical description

Up to March 2019, MOGREPS-UK was a 6-hourly ensemble running 12 members four times a day at the data times of 0300, 0900, 1500 and 2100 UTC (Hagelin *et al.*, 2017) up to 54 hr (i.e. $T + 54$ hr). MOGREPS-UK is a variable

resolution ensemble with 2.2 km grid space in the inner domain where the objective verification is calculated; the grid stretches to 4 km along the edges and the corners have a resolution of 4 km by 4 km (Hagelin *et al.*, 2017). In all the trials presented here, the model used is the midlatitude Regional Atmosphere 1 configuration (RA1-M) of the Met Office Unified Model (Bush *et al.*, 2020).

The parent ensemble to MOGREPS-UK, MOGREPS-G, is a global ensemble with approximately 20 km resolution at midlatitudes running 1 control and 17 perturbed members every 6 hr out to a forecast lead time of 7 days. For the atmosphere, MOGREPS-G uses the Global Atmosphere 6.1 configuration of the Unified Model (Walters *et al.*, 2017), whilst the surface exchange scheme is based on the Global Land 7 configuration of JULES described in Walters *et al.* (2019), but with additional changes to the growth of snow grains and the treatment of sea ice; it also uses the aggregate tile approach to surface exchange described in section 4.2.1 of Walters *et al.* (2017). Prior to December 2019, MOGREPS-G used an Ensemble Transform Kalman Filter (ETKF) scheme to generate its perturbations (Bowler *et al.*, 2008; Flowerdew and Bowler, 2011; 2013). Different sets of these perturbations are used every 12 hr (the first 17 members are used on the 0000 UTC and 1200 UTC, while the following 17 members are used on the 0600 UTC and 1800 UTC) in order to increase the sampling of the pdf. The members of MOGREPS-UK are also centred on the analyses from the DA system of the UKV high-resolution deterministic model (Tennant, 2015). The current 4D-Var DA system for the UKV is described in Milan *et al.* (2020). Additional perturbations come from stochastic physics via the Random Parameter 2 scheme (RP2: McCabe *et al.*, 2016).

Since the upgrade to the operational NWP models at the Met Office in March 2019, MOGREPS-UK has run three ensemble members every hour. Figure 1 represents the timeliness of the forecasts and how MOGREPS-UK depends on the cycle times of the parent ensemble MOGREPS-G (itself depending on the deterministic global cycles) and the UKV deterministic model.

On the 0300, 0900, 1500 and 2100 UTC cycles, as in the 6-hourly configuration, one of the members is a control member (member 0 in Figure 1) with unperturbed initial conditions and the RP2 stochastic physics scheme turned off. A given set of six consecutive hourly cycles therefore provides 17 perturbed members, with the RP2 scheme employed and initial-condition (IC) perturbations supplied by MOGREPS-G, and one control member. Whereas the IC perturbations for the 12 members of the 6-hourly configuration all came from the same MOGREPS-G forecast range (perturbed-member forecasts at $T + 3$ hr), a range of MOGREPS-G forecast times from $T + 3$ to $T + 8$ is required for the hourly cycling. For every hourly cycle, all

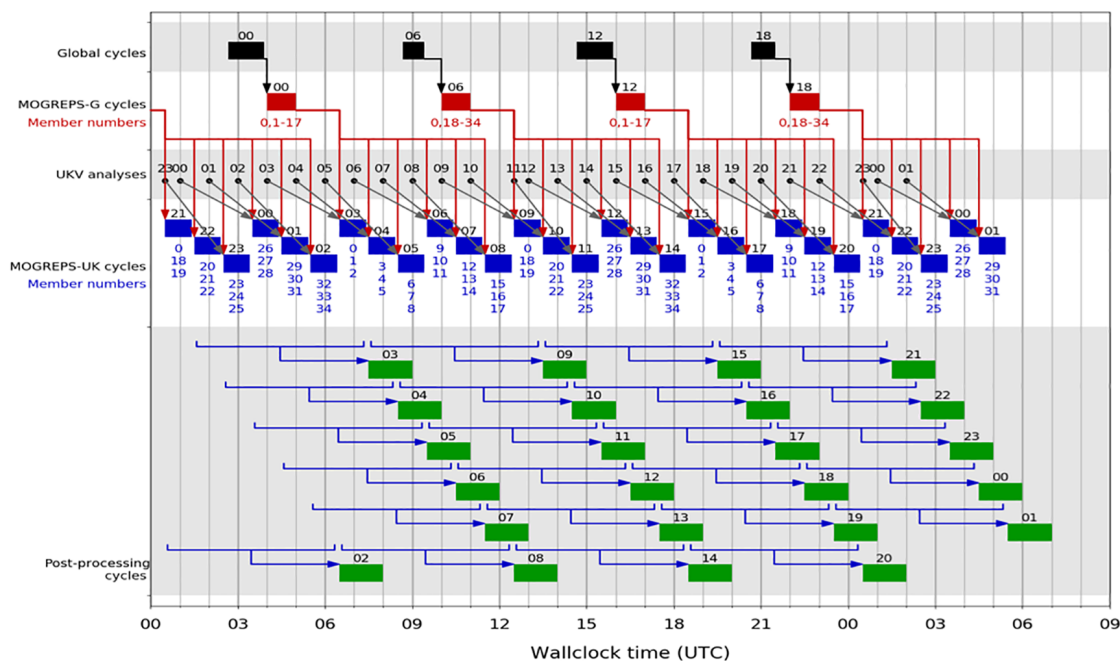


FIGURE 1 Diagram representing the timeliness of the time-lagged configuration of MOGREPS-UK. The blue boxes show the run times of a single MOGREPS-UK cycle; the blue numbers show the ensemble members of MOGREPS-UK (where member zero is the control). The black dots/grey arrows show the UKV analyses around which a given MOGREPS-UK cycle is centred. The red boxes, arrows and numbers show the MOGREPS-G members that provide the initial-condition perturbations and lateral boundary conditions. The blue arrows and green boxes show the members that are time-lagged into 18-member ensembles through post-processing

members are centred on the high-resolution deterministic analysis.

The new model configuration allows, among other possibilities, 18-member ensembles to be generated in post-processing by time-lagging over six cycles. When processing the ensemble, the forecast range and validity times align with the newest members. After each hourly cycle, a new time-lagged ensemble can be generated from the previous one by replacing the three oldest members with the three most recent. For example, the 0800 UTC ensemble comprises the three members from each of the 0800, 0700, 0600, 0500, 0400 and 0300 UTC model cycles, and is available by a wall-clock time of 1400 UTC. Then the 0900 UTC ensemble uses the 0900, 0800, 0700, 0600, 0500 and 0400 UTC model cycles and is available by a wall-clock time of 1500 UTC, and so on. Since our latest operational release (December 2019), these timings have now been improved to deliver the ensemble forecasts at $T + 4$ hr instead of $T + 6$ hr. In the context of this publication, whilst the timings are the same between the hourly time-lagged and the 6-hourly time-lagged configuration, it is worth keeping in mind that the hourly time-lagged ensemble is also run for longer lead times, albeit using more resources ($T + 54$ hr against $T + 120$ hr).

Compared with the 12-member sets from the 6-hourly configuration, a new 18-member time-lagged ensemble

incorporates information from five additional sets of high-resolution initial conditions. See the post-processing cycles in Figure 1 for more details. Also, an 18-member ensemble will have used driving data from six different MOGREPS-G forecast ranges, which for 20 out of every 24 such ensembles will have come from two different MOGREPS-G cycles. This configuration comes at the cost of using older MOGREPS-G cycles with longer forecast ranges. The forecast length of the individual members is such that an 18-member time-lagged ensemble covers a range of 120 hr (5 days).

An advantage of the time-lagging approach is that the size of the ensembles can be increased by adding older members into the post-processing without any additional computational costs in terms of the forecasts. For example, a 24-member ensemble can be created by adding two more sets of three members to the 18-member ensemble. For some cycles, this 24-member ensemble would contain two control forecasts (unperturbed members) from different MOGREPS-G cycles. Beyond 6 hr of time-lagging, any further time-lagging will use the same set of perturbations from MOGREPS-G but associated with different cycles of the global ensemble; we denote the newest unperturbed member as the control forecast of the ensemble. In the objective verification, 24 and 12 time-lagged members from the hourly configuration are tested here for

comparison against the operational 18-member ensemble and the 6-hourly cycling. Further time-lagging, while possible, was not tested for this study.

2.2 | Neighbourhood processing and verification metrics

Convective-scale models usually provide more realistic features than coarser-resolution models. However, verifying such nonlinear and sometimes rapidly evolving forecasts at a grid scale around the observation sites may be challenging. In this article, neighbourhood processing is used in objective verification within the “High Resolution Assessment” (HiRA) framework (Mittermaier, 2014). The HiRA framework was developed for verifying kilometre-scale models at observing locations, providing the ability to compare deterministic and ensemble forecasts (at different resolutions) using a single-observation-neighbourhood-forecast (SO-NF) approach. Even though neighbourhoods are used, HiRA represents an assessment of local forecast accuracy. It assumes that all grid points within a specified neighbourhood size can be considered as pseudo-ensemble members which are physically realistic and equiprobable outcomes for the observing location located at the centre of the neighbourhood. This method helps avoid the double penalty problem due to the errors in both the spatial and temporal distribution of the forecast fields (Ebert, 2008; Roberts and Lean, 2008; Mittermaier, 2014). When used with ensembles, the HiRA framework increases the ensemble size to provide more information at the small scales (Roberts and Lean, 2008; Dey *et al.*, 2014; Raynaud and Bouttier, 2017; Schwartz and Sobash, 2017). A neighbourhood size of 3 by 3 grid points (i.e. a scale of 6.6 by 6.6 km) is used here to overcome the problems of double penalty; such a scale was shown to perform best in Hagelin *et al.* (2017). Mittermaier and Csima (2017) also showed that the forecast skill of the ensemble was mostly sensitive to an increase in neighbourhood scale from 1×1 to 3×3 grid points. Beyond this, the forecast skill of the ensemble can deteriorate; for example, for temperature, cloud base height and visibility, where the local orography may be a strong influence or more generally the physical assumptions regarding the validity of a grid point at some distance from the observing location are rendered invalid.

A range of verification scores are used here as in Hagelin *et al.* (2017). The categorical scores used are the Ranked Probability Scores (RPS: Epstein, 1969), Continuous Ranked Probability Scores (CRPS: Hersbach, 2000), Hinton diagrams (Hinton and Shallice, 1991; Bremner *et al.*, 1994), ROC (Receiver Operating Characteristic) areas and reliability diagrams (Schwartz *et al.*, 2014; 2017;

Beck *et al.*, 2016; Mittermaier and Csima, 2017; Raynaud and Bouttier, 2017). RPS and CRPS measure the skill of the probability distribution and Hinton diagrams allow us to summarise these differences in skill between two ensemble forecasts for a range of fields such as temperature, wind speed, cloud amount, cloud base and precipitation for a given neighbourhood size. The lower the RPS or CRPS the better. The CRPS is chosen here for the temperature field as it is a continuous field and we are interested in the whole distribution of the parameter, while the RPS is used for other fields such as wind speeds, cloud amounts, cloud base heights, visibility and precipitation, which follow non-Gaussian distributions. The ROC area gives us an indication of the ability of the forecasts to discriminate between events and non-events. The higher its value, the better the discrimination. Reliability diagrams evaluate how close the probability of an event is to the observed frequency of this event so, the closer the points are to the diagonal, the better the reliability. All these scores are available in the HiRA framework. Finally, rank histograms are also used here to depict how the observations rank in comparison to the ensemble forecasts. A U-shaped rank histogram indicates that the observations fall too frequently outside the range/spread of ensemble members, and so can be interpreted as a lack of ensemble variance.

Continuous scores such as mean error and member deviation from ensemble mean are also used here. Statistical significance at the 0.05 level was determined using the Wilcoxon signed-rank test (Wilks, 2011, pp 163–164). All verification scores are produced using synoptic observations as truth, which come from a network of 119 quality-controlled stations distributed over land in the United Kingdom, that adhere to World Meteorological Organisation (WMO) technical regulation standards (WMO, 2018), referred to as LND SYN stations.

3 | COMPARISON OF THE 6-HOURLY AND HOURLY CONFIGURATIONS: OBJECTIVE VERIFICATION

3.1 | Description of the trials

Two trial periods for each of the 6-hourly and hourly configurations are used here to cover different types of weather conditions. The winter trial runs from 2 December 2017 to 2 January 2018. The summer trial runs from 2 July to 2 August 2017. In these trials, forecasts from the time-lagged system are produced out to T + 54 hr only, which allows a full comparison with the 6-hourly configuration but limits their computational cost. An additional pre-operational trial of the hourly configuration is also

used here as it provides information over a longer period from 1 January to 10 March 2019. We note that because the objective verification of the hourly time-lagged ensemble is made here for comparison with the 6-hourly ensemble, it is not the purpose here to demonstrate the skill of the new configuration in delivering a specific product using the most-up-to-date forecast of the hourly time-lagged ensemble. In this section, all verification results are equalised on the same lead times and validity times between the 6-hourly and hourly configurations, so we do only examine four cycles a day. To compare the different configurations, we have separated our analysis into different scores as ranked probability scores, biases, reliability and ensemble spread, as explained in Section 2.2. The analysis is detailed per score rather than per field because, firstly, each of these scores describes a particular feature of the ensemble verification and it is important to assess whether the new operational time-lagged configuration delivers some robust improvements for each of these different characteristics and, secondly, the same characteristics apply to most fields.

3.2 | Ranked probability scores

The Met Office uses Hinton diagrams to summarise the comparison between the two configurations using the RPS. A better (worse) performance for the trial vs. the control is depicted with green (purple) triangles. RPSs allow us to capture the quality of the ensemble in satisfying a set of specific thresholds. Being a distribution-based accuracy metric, it captures the performance of the complete ensemble distribution, rather than focusing on a small selection of fixed thresholds which usually only check for exceedance of a single threshold (and therefore do not check whether a forecast has just exceeded the threshold or is potentially an outlier). As a result, the RPS is a much stricter test for the ensemble distribution. In the Hinton diagrams the area of the triangles is proportional to the percentage of improvement/degradation in the scores presented. The area of the plotted triangle is normalised assuming a maximum difference between cases of 20%, that is, the largest triangle indicates a difference of 20% or greater. The black outlines on the triangles correspond to statistical significance. Other verification metrics comparing the reliability, discrimination and biases of the ensembles also add useful and specific information to this comparison, as described later.

The new configuration is compared against the 6-hourly configuration with 12 members and against a variant of the 6-hourly configuration run with 18 members to test the sensitivity of the time-lagging to ensemble size. Figure 2 illustrates these score cards for a summer month

and a winter month. Against the 12-member 6-hourly configuration (Figure 2a,b), the new hourly configuration described here shows good performance. A small negative impact is present in the CRPS for temperature up to $T + 1$ in the winter and $T + 3$ in the summer (however, note from the operational schedule in Figure 1 that forecast ranges out to approximately $T + 6$ have expired by the time the post-processed products become available). This is related to the sensitivity in the accuracy of the temperature forecast to the use of older initial conditions for the time-lagging configuration. This impact is more pronounced in the summer than in the winter and may be due to the benefits from data assimilation in correcting the latest cycles for the larger error in the more pronounced diurnal cycle in the summer (see bias results in Section 3.3). Mittermaier and Csima (2017) also found that MOGREPS-UK has poorer CRPS than the deterministic high-resolution UKV model at very short lead times and attributed this to the lack of DA in the ensemble (more specifically, in the perturbed members of the ensemble).

Another characteristic of this comparison is the fact that the differences appear to be largest at short lead times, but this varies with the variable and season. For example, the benefit of the time-lagged configuration at longer lead times (for example $T + 42$) is larger in the summer than the winter.

Compared to the 6-hourly configuration with 18 members, the hourly configuration is usually better up to $T + 15$, but at the end of the forecast range we see a more neutral or even negative performance for the winter season. The differences are, however, not statistically significant (i.e. not many triangles with a black outline, except for the CRPS for temperature). This suggests that for an equal ensemble size, the skill of the hourly configuration is mainly positive in the early lead times and similar or negative in the longer lead times due to the use of older forecasts. When the 18-member time-lagged ensemble is compared against the 12-member 6-hourly ensemble, the difference in ensemble size contributes to the improved performance at longer lead times.

Figure 3 displays a comparison of those different ensemble sizes for hourly precipitation in the winter trial, along with an additional hourly time-lagged configuration of 24 members (i.e. eight cycles of time-lagging). As a benchmark, the 18-member hourly configuration is used. This shows that the 18-member 6-hourly set-up can perform better than the hourly configuration with 18 members at some lead-times (differences again are not significant), but that the hourly configuration with 24 members is an even better configuration overall. These findings also apply to other seasons and most variables except for the CRPS for temperature: the skill of the hourly ensemble with 24 members shows a significantly better performance

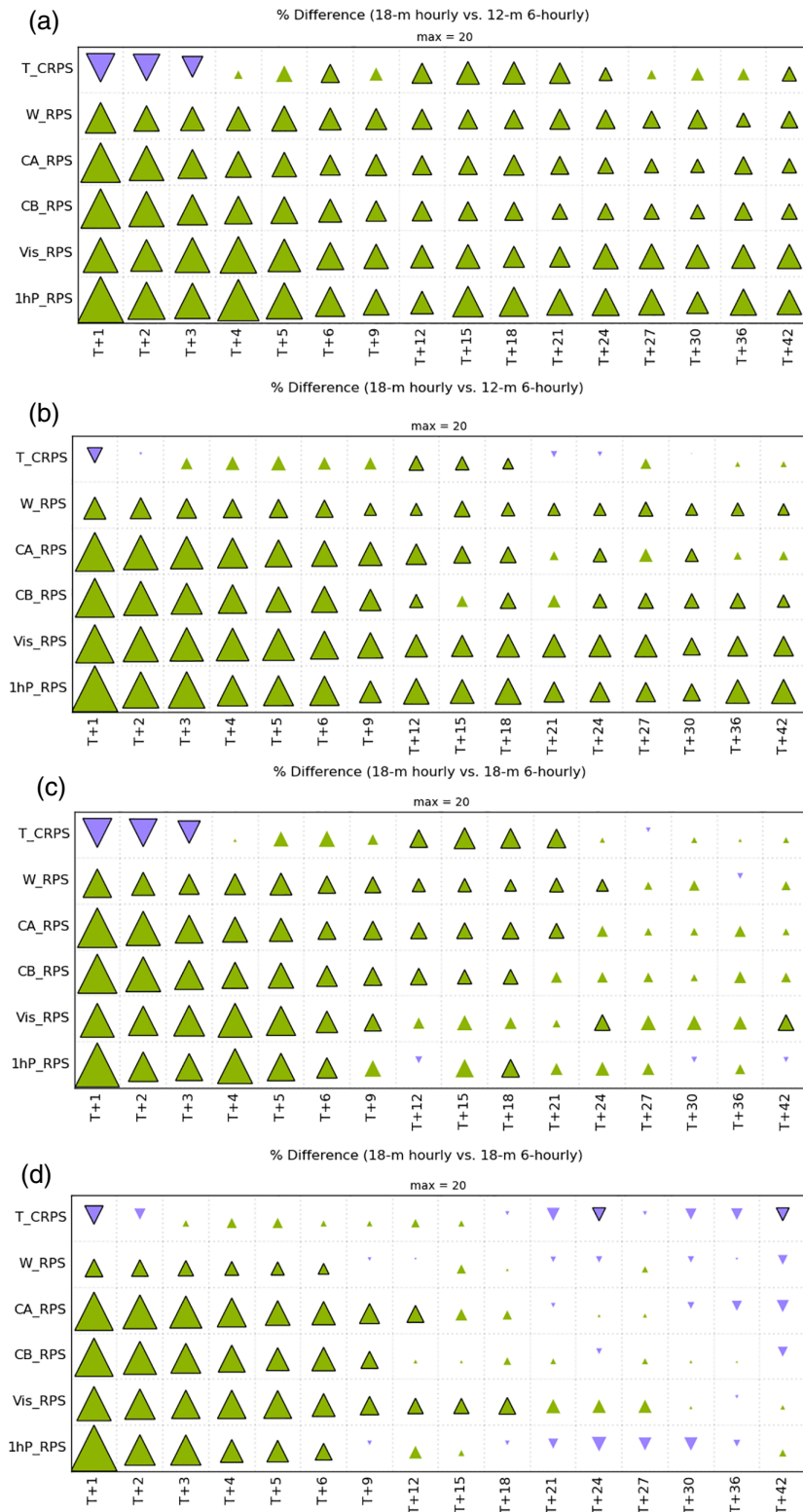


FIGURE 2 Hinton diagrams displaying the CRPS for temperature (T_CRPS) and RPSs for wind speed (W_RPS), cloud amount (CA_RPS), cloud base (CB_RPS), visibility (Vis_RPS) and precipitation (1hP_RPS) for a neighbourhood size of 3 by 3. (a) 2 July to 2 August 2017 18-m(ember) hourly against 12-m 6-hourly; (b) 2 December 2017 to 2 January 2018 18-m hourly against 12-m 6-hourly; (c) 2 July to 2 August 2017 18-m hourly against 18-m 6-hourly; and (d) 2 December 2017 to 2 January 2018 18-m hourly against 18-m 6-hourly. Statistical significance (at the 0.05 level) was determined using the Wilcoxon signed-rank test (Wilks, 2011); statistically significant results are represented by a black outline to the triangle. All data are equalised onto the 6-hourly ensemble validity and lead times.

than the hourly configuration with 18 members and is a better match to the 6-hourly configuration with 18 members at longer lead times. The fact that a 24-member ensemble (with its longer period of time-lagging) outperforms the 18-member ensemble agrees with Raynaud and Bouttier (2017) who showed that time-lagging over 12 hr

improves the skill of the ensemble mainly for short lead times.

The new finding here is that the performance of the 24-member time-lagged ensemble is useful to maintain statistical significance not just at early lead times, but mainly at the longer lead times. So, using older forecasts

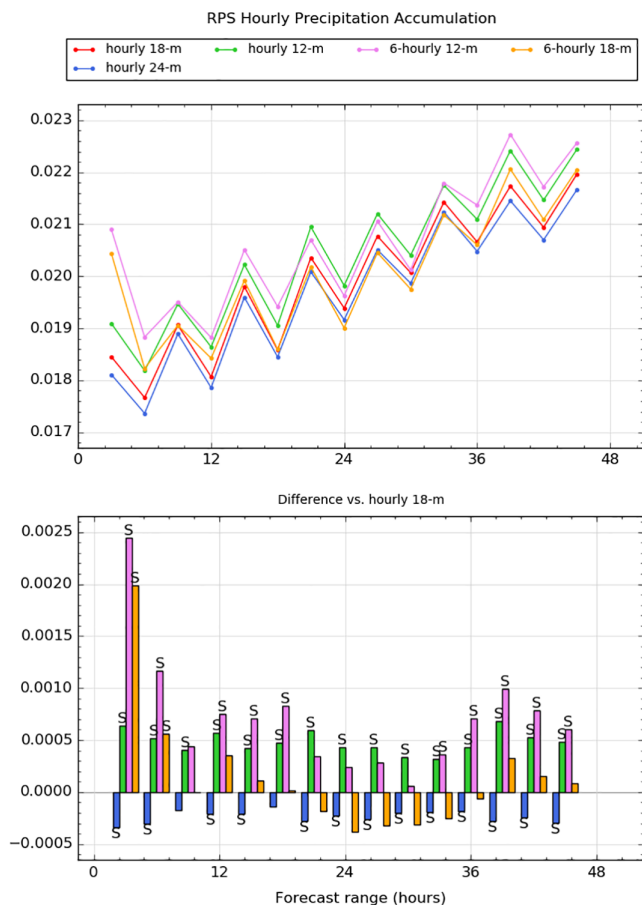


FIGURE 3 RPS for hourly precipitation from 2 December 2017 to 2 January 2018, showing the 6-hourly configuration for 12 and 18 members as well as the hourly configuration for 12, 18 and 24 members for a neighbourhood size of 3 by 3. A black “S” on the difference plot (lower panel) indicates statistical significance at the 0.05 level, calculated using the Wilcoxon signed-rank test (e.g. Wilks, 2011). All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.

and further time-lagging still improves the sampling of the pdf at all lead times. Whilst it is accepted that increasing ensemble sizes usually results in improvements in the skill of ensembles at all lead times (e.g. Hagelin *et al.*, 2017; Raynaud and Bouttier, 2017), this is less intuitive when the increase in ensemble size is associated with further time-lagging and thus older forecasts.

3.3 | Biases

The hourly configurations sometimes lead to better biases at very short lead times for fields like precipitation or clouds, but then, these biases either become worse or similar compared to the 6-hourly configurations at longer lead times. So, the improvements in the RPS Hinton diagrams,

as in Figure 2, are not strictly related to improvements in the biases. This is particularly true for the 18- and 24-member time-lagged ensembles against the 12-member 6-hourly ensemble for example (as these ensembles show a better performance at all lead times). The temperature bias follows a slightly different story because of a growing bias with lead times in the winter as well as a strong diurnal cycle bias particularly in the summer. We focus on the temperature problem in this section.

Figure 4 illustrates the impact of time-lagging on the winter temperature bias with HiRA using a neighbourhood scale of 3×3 grid points as for the RPSs. This provides a mean bias calculation over the ensemble size as well as the neighbourhood scale. As biases grow with forecast time, we have included the bias associated with the 3-member ensemble (i.e. the set of members with the earliest initialisation time, with no time-lagging). Following Figure 2 and the impact of time-lagging on the temperature at short lead times, we think that representing this 3-member ensemble provides useful information. Indeed, in comparison to the 6-hourly ensemble, it illustrates the impact of all the large-scale perturbations. In comparison to the hourly time-lagged ensemble, it illustrates the role of the DA as well as the impact from the older forecasts.

The ensemble has a cold bias overall, with oscillations showing the impact of the diurnal cycle (with a 6-hourly frequency because of the equalisation onto the 6-hourly cycles). The best performing group is the 3-member ensemble, followed by the 6-hourly configurations with first 18 members and then 12 members. The other hourly configurations, involving time-lagging, have even more negative biases, with the 18- and 24-member ensemble biases slightly better than in the 12-member ensemble. So, when using time-lagging over 4 hr (i.e. 12 members), the greater age of the forecasts leads to larger temperature biases. Beyond 4 hr, further time-lagging to 6 hr (i.e. 18 members) improves the bias whilst further time-lagging to 8 hr (i.e. 24 members) shows fewer improvements, suggesting a complicated feedback between the impact of time-lagging and the interaction of the DA cycle with the diurnal cycle bias. Interestingly, even though the hourly configurations have larger biases, the biases do not appear to affect the shape of the ensemble distribution sufficiently to have a detrimental impact on the CRPS and RPS (see Figure 2b against the 12-member 6-hourly system).

For the summer, as the diurnal oscillations are stronger, we focus here on a diurnal cycle plot (Figure 5). Figure 5 illustrates the impact of the diurnal cycle on the summer temperature bias for a 0300 UTC cycle only. This shows that the two configurations have different roles to play in this diurnal cycle: the 6-hourly configurations as well as the 3-member ensemble, with all members starting from a night-time DA analysis, have the smallest warm

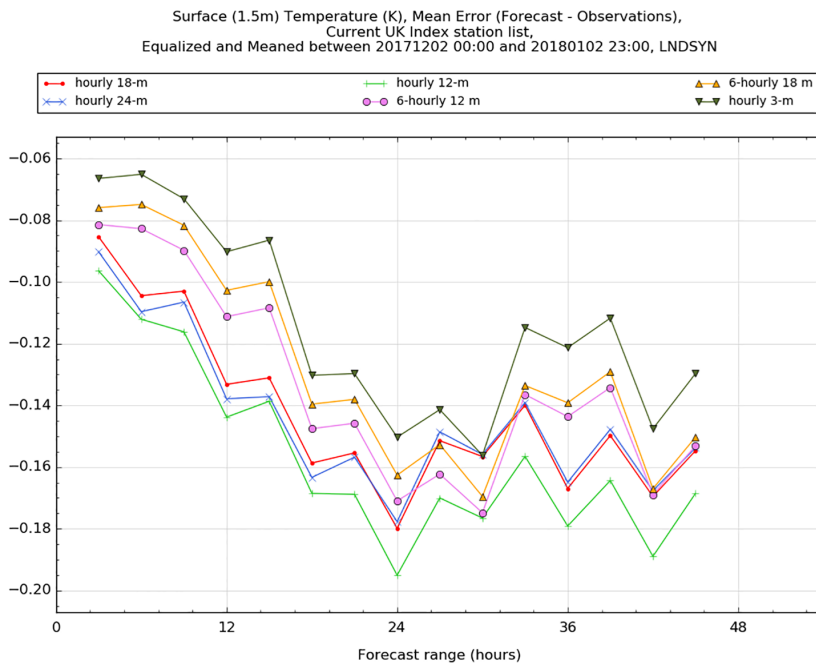


FIGURE 4 Temperature bias for different ensemble sizes from 2 December 2017 to 2 January 2018 as a function of forecast range, displaying the 6-hourly configuration for 12 and 18 members as well as the hourly configuration for 3, 12, 18 and 24 members. LND SYN is a label for the type of truth data used in the verification; in this case synoptic observations based at UK stations on land as defined in WMO technical regulations volume I (WMO, 2018). All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.

night-time bias, but the strongest cold daytime bias. The DA is able to address the temperature biases by removing heat from the atmosphere. All members of the 6-hourly and 3-member ensemble members are centred around the 0300 UTC analysis, where the DA is cooling the system to reduce the night-time cold bias. This reduction in heat, combined with too little warming through the day then leads to a larger daytime cold bias. The same plot at 1500 UTC shows that the DA is able to add heat to the system (not shown). The time-lagged ensembles behave in the opposite way. As for the CRPS results in Figure 2, the role of DA on the temperature bias and skill remains influential and perhaps more so in the summer than in the winter.

So, with the current model performance, we see different characteristics in the bias depending on time-lagging and on the season. As the ensembles have an underlying cold bias growing with lead time in the winter, the time-lagged ensembles are seen to perform worse (which may explain the worse performance in the winter CRPS at longer lead times). For the summer, there may be some links here between the bias in temperature and the CRPS results at short lead times (up to $T + 3$). These results do not explain the improvements in CRPS for the hourly configuration seen at other lead times in Figure 2 (i.e. between $T + 3$ and $T + 21$, for example, in comparison to the 12-member 6-hourly ensemble in both winter and summer).

3.4 | Reliability and discrimination

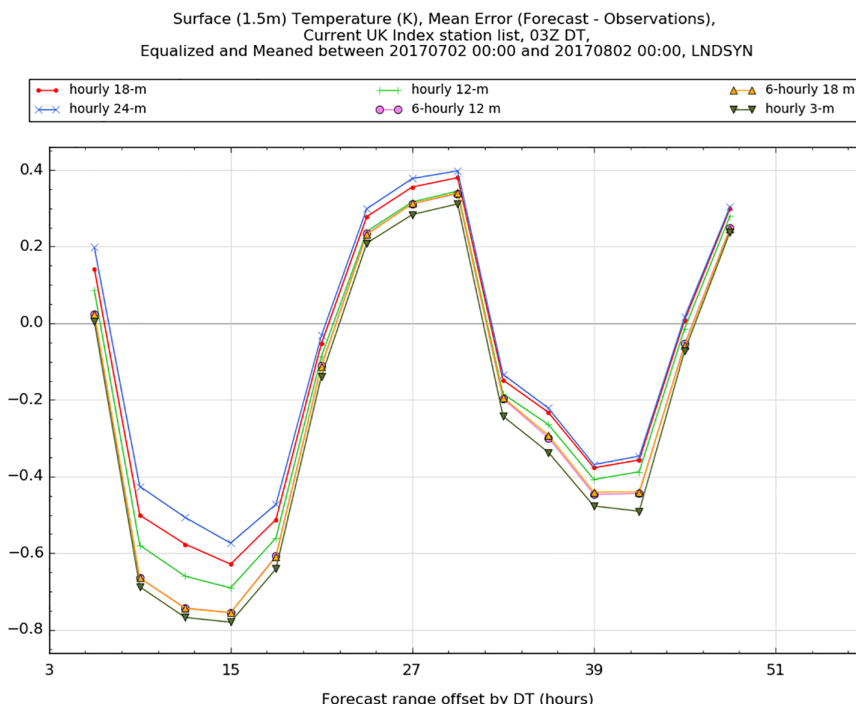
Reliability diagrams were examined at different lead times to compare the hourly and the 6-hourly configurations.

Overall, except at short lead times, we see little evidence of any improvement in the reliability between the two configurations in most fields and for both summer and winter trials. This may be due to a lack of events in the upper bins. Larger neighbourhood sizes than a scale of 3 by 3 grid points did not help in this regard.

Instead, the reliability part (Primo *et al.*, 2009; Flowerdew, 2014; Beck *et al.*, 2016) of the Brier score (Brier, 1950) for a neighbourhood scale of 3 by 3 points is used in Figure 6 to provide more information about the reliability of the ensemble configurations. When there is no sampling issue (i.e. lack of jumpiness with lead times), the hourly ensemble with 18 members has better reliability (mostly at short lead times but at all lead times for some fields) than the 6-hourly ensemble with 12 members. The hourly ensemble with 12 members has either comparable or better reliability than the 6-hourly ensemble with 12 members, depending on the field, threshold or season. The larger the time-lagged ensemble, the better the reliability. Note however that these differences in the reliability values of the Brier score are very small; except perhaps at short lead times, for most fields these are negligible, in agreement with the reliability diagrams.

Areas under the ROC curve were also examined (for forecast discrimination potential) between the two ensembles and for different ensemble sizes (not shown). This comparison is similar to that for the Ranked Probability and Brier scores. When sampling is not an issue, for all forecast lead times, the hourly configuration with 18 members performs better than the 6-hourly configuration with 12 members, with the 24-member ensemble performing best. In other details, for most thresholds, the 12-member time-lagged ensemble would show slightly higher areas

FIGURE 5 Diurnal cycle of the temperature bias, from a 0300 UTC cycle, for different ensemble sizes from 2 July to 2 August 2017, displaying the 6-hourly configurations for 12 and 18 members as well as the hourly configuration for 3, 12, 18 and 24 members. All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.



under the ROC than the 12-member 6-hourly ensemble for the shorter lead times and more similar values for the longer lead times, as in the RPSs. Again, these results apply to all fields.

3.5 | Ensemble spread and impact of initial conditions on growth of errors

Various metrics are available to analyse the spread of an ensemble depending on the field and application. Among those, rank histograms are used to determine the degree of dispersion in the ensemble pdf against the observations. Figure 7 compares the hourly and 6-hourly configurations for a size of 18 members using grid-point verification (rank histograms were not available within the HiRA framework). For most variables (except precipitation and the cloud fields) and for both seasons, the model is underspread and the use of the hourly configuration helps slightly to make the ensemble more dispersive. This result is one of the first indications that the hourly configuration increases the spread of the ensemble for variables such as temperature, wind speed, cloud amount, cloud base, and visibility. Note that this also applies when comparing the 12-member 6-hourly configuration with the 12-member hourly configuration (not shown). Rank histograms for precipitation show that the observations occur too frequently in the highest bins, indicating that the ensemble does not capture high enough rainfall, and that perhaps we still do not have enough members (note however that this relies on grid-point analysis).

The standard deviation between the ensemble members is assessed next for the hourly and 6-hourly configurations. As previously, Figure 8 illustrates the comparison between the two configurations and different ensemble sizes. For both the hourly and 6-hourly configurations, an increase in the ensemble size results in a small increase in the ensemble spread.

Between the 6-hourly configurations and the hourly configurations, we see a larger increase of almost 15% at the shorter lead times and approximately 5% at longer lead times. The key point to note here is that this difference is not related to the ensemble size, but to the difference in configurations: a 6-hourly ensemble with 18 members does not offer the same quality of spread as the hourly one with 12, 18 or 24 members. The ensemble is overall underspread, as illustrated in Figure 9a,b where the standard deviation to the observations is higher than the standard deviation between the members and the ensemble mean. The time-lagged configuration improves the ensemble spread and reduces the lack of dispersion of the ensemble; these improvements in the spread contribute to the improvements in the RPS presented in the Hinton diagrams in Figure 2. Note here that no observational error or post-processing technique (apart from the time-lagging itself) is considered in this overall assessment of the convective-scale ensemble. This will be the subject of future work.

Using the 24-member ensemble contributes to a further increase in the spread of the ensemble (as illustrated in Figure 8 for the standard deviation between the members) and a further reduction of the standard

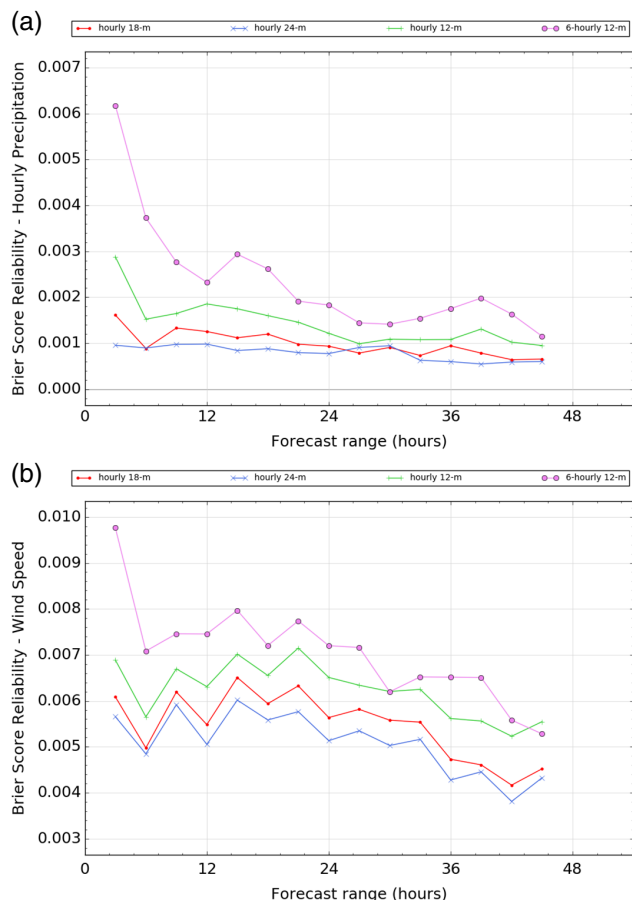


FIGURE 6 Reliability part of the Brier score for (a) a low precipitation threshold of 0.25 mm per hour from 2 July to 2 August 2017 and (b) a low wind threshold of 3.6 m·s⁻¹ from 2 December 2017 to 2 January 2018 for a neighbourhood scale of 3 by 3. All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.

deviation against the observations (not shown here), thereby further improving the dispersion issue of the convective-scale ensemble. This is perhaps explained by the systematic use of multiple global ensemble driving conditions in the 24-member time-lagged ensemble for each cycle, compared to the 18-member time-lagged ensemble. Indeed, as illustrated in Figure 9c,d, the cycles of the hourly configuration associated with a 50–50% mixture of MOGREPS-G cycles (e.g. 0500 UTC) do have additional spread throughout the forecasts, compared to the cycles of the hourly configuration with 100% forcing from the same MOGREPS-G conditions (e.g. 0800 UTC). This is slightly more pronounced in the winter than in the summer. So, mixing different cycles of large-scale perturbations further helps with the spread in the convective-scale ensemble. This suggests that further mixing of MOGREPS-G conditions could be beneficial to the spread.

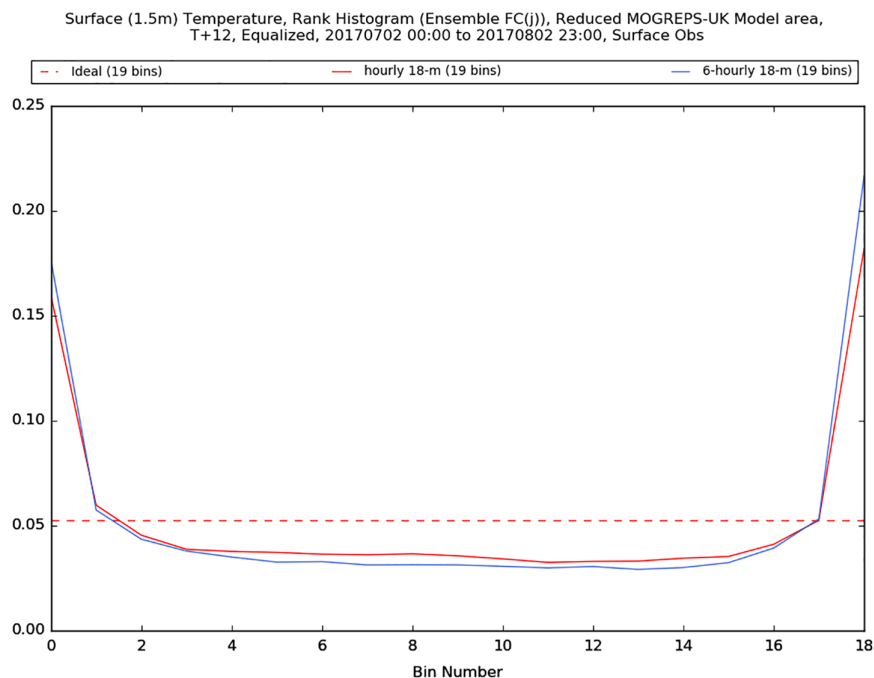
Despite these improvements, the ensemble is overall underspread when using these metrics of standard deviation, and this is a general issue for convective-scale ensembles (Gebhardt *et al.*, 2010; Schwartz *et al.*, 2014; Beck *et al.*, 2016; Dey *et al.*, 2016; McCabe *et al.*, 2016; Raynaud and Bouttier, 2017). Further efforts will need to assess the characteristics of the spread as well as developing more tools to understand and quantify the ensemble spread; metrics based on Fractions Skill Score (FSS), for example, show promise (Dey *et al.*, 2014; Weyn and Durran, 2018; 2019; Keil *et al.*, 2019; Porson *et al.*, 2019) or other metrics such as correspondence ratios (Gebhardt *et al.*, 2010). A step towards understanding this impact on the spread has been taken by separating the impact of the initial perturbations from the large-scale and those introduced by the hourly configurations, by sampling a range of hourly DA analyses.

To understand which aspects of the system are leading to the increase in performance at which time-scales, Figure 10 shows the decomposition of the error growth depending on the type of perturbations used to initiate the spread of the ensemble for the summer trial. One by one, the perturbations help to reduce the growth of errors or decay in quality by reducing the RPS (and thus improving the accuracy of the ensemble forecasts). These perturbations are: the boundary conditions from MOGREPS-G (LBCs), the initial-condition perturbations from MOGREPS-G (IPs) and the change from the 6-hourly to the hourly configuration. The red line (“6-hourly stph”) is an ensemble created only by the stochastic physics of the ensemble (McCabe *et al.*, 2016) as well as additional noise in the boundary layer to initiate the diurnal cycle (Bush *et al.*, 2020). This ensemble forecast has the largest error. Then, in the blue line (“6-hourly LBCs only”), the LBCs perturbations from MOGREPS-G only are introduced, and nothing from the initial perturbations. The impact on error growth dominates in the longer lead times as expected.

From the blue line to the green line (6-hourly with both IPs and LBCs), the IPs coming from MOGREPS-G, that is, from the large scale, are introduced. These perturbations dominate in the short lead times up to T + 30. By T + 30, the perturbations from the LBCs are responsible for the reduction in the error growth. Note that this impact of the LBCs at longer lead times is consistent with previous studies (Gebhardt *et al.*, 2010; Porson *et al.*, 2019).

The hourly configuration then adds another degree of initial perturbations by using a mixture of small-scale perturbations and (depending on the cycle and ensemble size) large-scale perturbations as well. The pink line (“hourly 12-m LBC only”) represents the error growth associated with the hourly configuration with no IPs from MOGREPS-G, but this configuration maintains the hourly re-centring of the analysis in the initial conditions coming

FIGURE 7 Rank histogram for surface temperature forecasts for 18 members for the hourly and 6-hourly configurations from 2 July to 2 August 2017 at T + 12. All data are equalised onto the 6-hourly ensemble validity and lead times.



from the UKV. So, this pink line includes the impact of time-lagged small-scale perturbations (i.e. four successive cycles here for the 12-member hourly ensemble).

The impact of including the multiple initial conditions from these successive cycles of the UKV is measured by comparing either the blue with the pink lines (for ensemble configurations without any large-scale initial perturbations) or the green with the yellow lines (representing respectively the 6-hourly and hourly configurations with all initial perturbations). The hourly time-lagged configuration reduces the error at short lead times with or without the initial perturbations from the large-scale ensemble, highlighting the importance of the small scales as in Weyn and Durran (2018; 2019).

Looking at the differences against the 6-hourly system with no perturbations to the driving conditions (i.e. “6-hourly stph”) in the bottom panel of Figure 10, the impact of the hourly configuration “hourly 12-m LBCs only” is nearly as large initially as the impact of the 6-hourly set-up with LBCs and IPs from MOGREPS-G. This suggests that unlike the 6-hourly system, the hourly configuration sees its error growth initially depending on the use of time-lagging itself, with or without large-scale perturbations.

All other variables follow the same pattern, except for the temperature, consistent with the CRPS results in Figure 2 and the bias analysis in Figure 4.

Overall, we have shown here that time-lagging contributes to an improvement in the quality of the ensemble (smaller RPS), due to the small scales entering the system from the hourly update in the high-resolution UKV deterministic model.

4 | SKILL OF THE HOURLY CONFIGURATION TO T + 120

It is reasonable to assume that higher-resolution convection-permitting ensembles will improve the realism of rainfall forecasts, compared to coarser-resolution ensembles. Kendon *et al.* (2012) have demonstrated this in a climate sense by comparing a 1.5 km model with a 12 km model. At the same time as introduction of the new hourly configuration, MOGREPS-UK was upgraded to run to T + 120 instead of T + 54. At these forecast lead times, MOGREPS-UK will then be compared to the other two NWP systems covering 5 days, namely its parent ensemble MOGREPS-G and the high-resolution deterministic UKV. The main objective is to improve the prediction of severe weather available from the global ensemble MOGREPS-G (about 20 km grid space at midlatitudes) and to model the uncertainty in location of severe weather predicted by our deterministic UKV model (1.5 km grid space).

We do not attempt here to provide a full analysis of the objective verification of these three NWP systems to T + 120, but instead assess the strengths and weaknesses of our convective-scale ensemble against the other systems already running to T + 120. These experiments may not provide robust results for high-impact weather. Sampling a large number of high-impact rainfall cases through subjective evaluation would be needed to fully complement this study, and this is ongoing work as the system is still in its infancy.

Here, we use operational data from 1 September to 1 November 2019 (dates between 16 and 20 September are missing) to analyse the performance

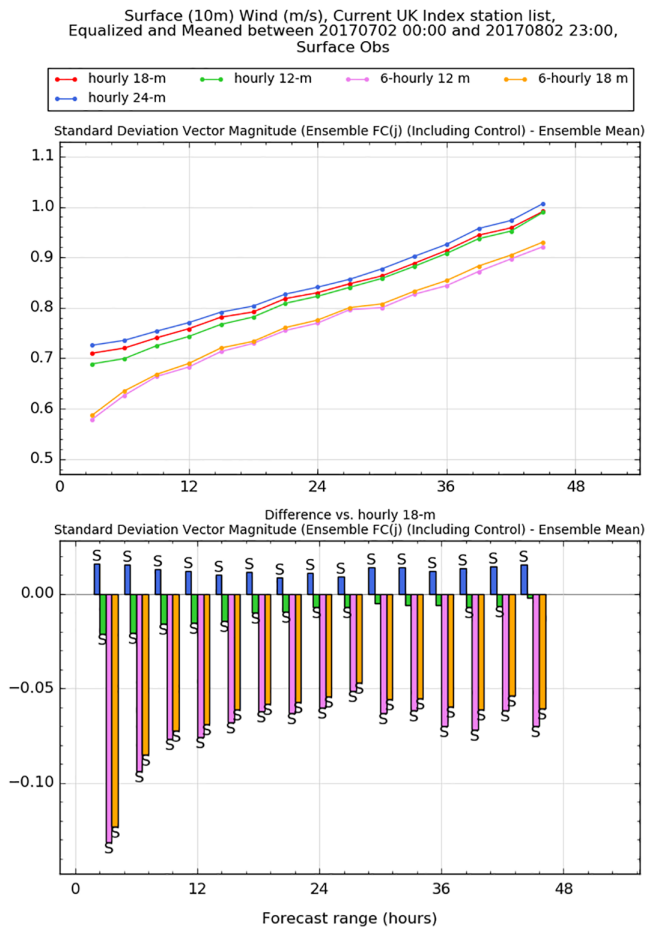


FIGURE 8 Standard deviation for wind vector magnitude between the ensemble members and the ensemble mean for the different ensemble configurations and sizes from 2 July to 2 August 2017. A black “S” on the difference plot (lower panel) indicates significance at the 0.05 level as calculated using the Wilcoxon signed-rank test (Wilks, 2011). All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.

of our convective-scale ensemble to T+120 against MOGREPS-G using the HiRA framework. A pre-operational set of data from 1 January to 10 March 2019 was also tested to compare the pre-operational MOGREPS-UK against the operational MOGREPS-G. We finish this article by testing how we could further improve MOGREPS-UK at these longer lead times over the same pre-operational data from 1 January to 10 March 2019 as well as additional trial data from 16 July to 16 August 2018.

4.1 | Comparison with existing NWP systems running to T + 120

Mittermaier and Csima (2017) presented a 3-year comparison of the UKV and the 6-hourly 12-member

MOGREPS-UK. They showed that there are two options when it comes to the choice of neighbourhoods for comparing models equitably: equalising on the physical neighbourhood size or on the approximate ensemble members. The former is appropriate when comparing deterministic forecasts (or when comparing the MOGREPS-UK control to the UKV for example, which tests the quality of the configurations). The latter is for comparing a deterministic model to an ensemble or one ensemble to another ensemble.

Mittermaier and Csima (2017) went on to show that, broadly speaking, there is an improvement in the 12-member MOGREPS-UK skill when using the smallest 3×3 neighbourhood (compared to using no neighbourhood), but negligible benefit from larger neighbourhoods. Hagelin *et al.* (2017) also showed that MOGREPS-UK has better accuracy compared to the UKV and that the increase in neighbourhood scale is more beneficial to the deterministic model than to MOGREPS-UK. A comparison was made here between the UKV using 11×11 (i.e. 121 pseudo-members) and 17×17 (i.e. 289 pseudo-members) against MOGREPS-UK with a 3×3 neighbourhood scale (i.e. $3 \times 3 \times 18 = 162$ pseudo-members) and this still shows better accuracy for MOGREPS-UK. This finding is consistent with previous works and extends the applicability out to T+120. Given the extensive previous research on this comparison, this is not illustrated here.

For MOGREPS-UK and MOGREPS-G, the comparison is based on an equalisation of ensemble members and a 3×3 neighbourhood scale for each ensemble.

MOGREPS-G data are only valid at 0000, 0600, 1200 and 1800 UTC and only 6-hourly forecast ranges are shown here. After T + 48, the data only cover the 0000 and 1200 UTC validity times. Using the RPS for wind speed, visibility and precipitation and CRPS for temperature, we plot the Hinton diagram summarising the comparison between MOGREPS-UK and MOGREPS-G in Figure 11. We see improvements in most fields, except for cloud amounts at longer lead times. This comparison depends, however, on the season. A comparison of the pre-operational data from MOGREPS-UK against the operational MOGREPS-G (not shown here) reveals that the cloud fields are not improved in MOGREPS-UK; the regional ensemble is worse for the cloud base heights, but again the differences are not statistically significant. Understanding the reasons for this would require a lot of additional work but currently the regional and global models do not use the same parametrizations, and this is contributing to the signals seen here. In future, there are plans to use consistent cloud parametrizations between the global and regional models.

MOGREPS-UK and MOGREPS-G also show different biases, changing with the seasons, as well as different reliability characteristics.

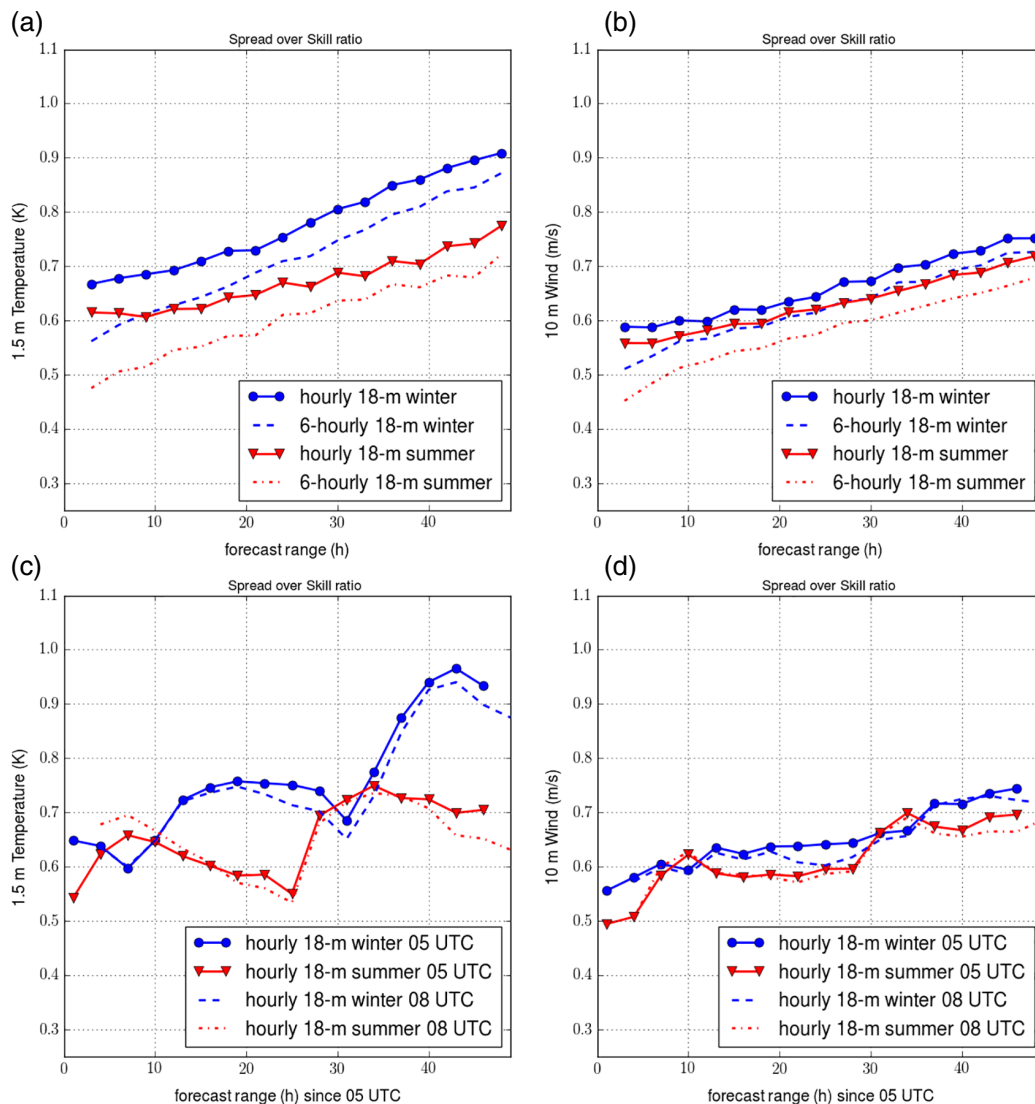


FIGURE 9 Ratio of ensemble standard deviation over standard deviation of ensemble mean against observations for (a) temperature and (b) wind speed, for the 6-hourly and hourly configurations. The summer trials run from 2 July to 2 August 2017 and the winter trials run from 2 December 2017 to 2 January 2018. All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hrs. Ratio of ensemble standard deviation over standard deviation of ensemble mean against observations for (c) temperature and (d) wind speed, for the 0500 UTC and 0800 UTC cycles for the hourly configuration only. The summer trials run from 3 July (as the 0500 cycle is not available on 2 July) to 2 August 2017 and the winter trials runs from 3 December 2017 (as the 0500 cycle is not available on 2 December) to 2 January 2018. All the spread and skill data are equalised on the same validity times for the individual cycles as well as between the two cycles for the common validity times. All data are plotted every 3 hr.

It is worth noting here that the biases from MOGREPS-UK are closer to the biases from the high-resolution UKV model than to the global ensemble MOGREPS-G, and this is likely to be due to differences in formulation between global and regional models.

Overall, MOGREPS-UK brings some improvements over MOGREPS-G in the scores analysed here, but these improvements become smaller with lead time (as in Schwartz, 2019). It is worth noting that the details of this comparison are likely to change with future upgrades in the science of either model, as well as future changes to the

initial conditions and data assimilation techniques in both ensembles. With this in mind, such a comparison has to be regularly re-assessed and repeated over different seasons.

4.2 | Future work on the design and processing of the hourly configuration

Here, we explore how we could improve the ensemble configuration to T + 120 hr. When comparing against the 6-hourly ensemble, we showed that a further increase in

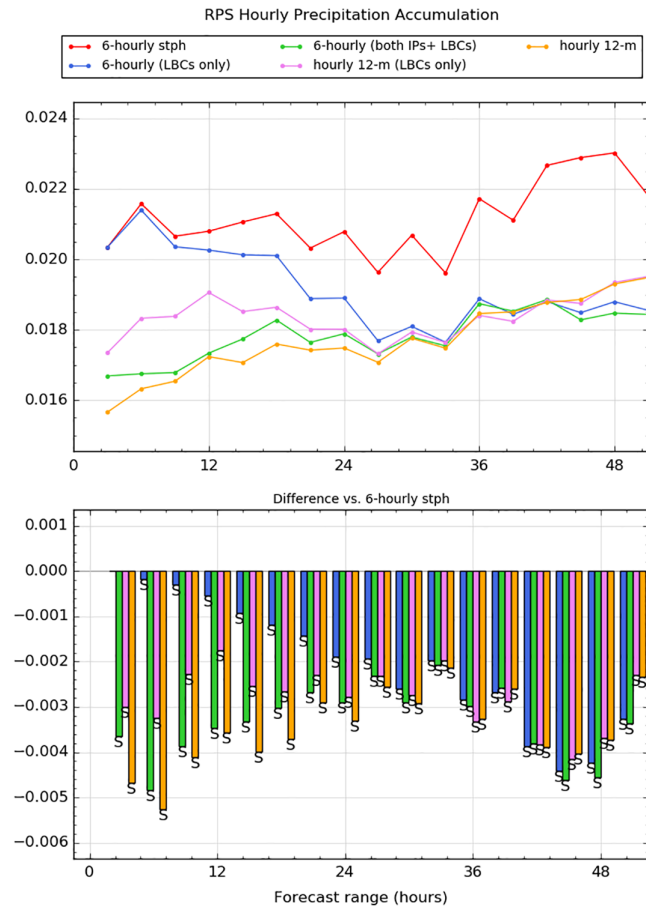


FIGURE 10 Decomposition of RPS for precipitation showing impact of initial conditions and lateral boundary conditions for both configurations for an ensemble size of 12 members from 2 July to 2 August 2017. A black ‘S’ on the difference plot (lower panel) indicates significance at the 0.05 level calculated using a Wilcoxon signed-rank test (Wilks, 2011). All data are equalised onto the 6-hourly ensemble validity and lead times. All data are plotted every 3 hr.

ensemble size from 18 to 24 members results in improvements in RPS, reliability, discrimination and spread. Now, we compare the impact of increasing the ensemble size directly against the impact of increasing the ensemble size using HiRA (and so a specific neighbourhood scale around the nearest grid-point in the verification) for lead times up to T + 120 hr. Whilst it would be reasonable to use adaptive neighbourhoods for post-processing to mitigate against the growth of forecast errors and decreasing predictability, this is not the case for verification of raw output, where there is an inherent interest in quantifying the growth in forecast error and loss of predictability. For this reason, a fixed neighbourhood size for verification is appropriate. For completeness it is worth noting here that using neighbourhoods in verification would be *inappropriate* if the forecasts had been post-processed using a neighbourhood technique prior to verification. We expect the results to

depend on the field (Mittermaier and Csima, 2017) and the season. These results may also be different if the type of neighbourhood processing used could adapt to the local terrain for example, but this is not investigated here.

The results here focus on two seasons: a winter pre-operational trial from 1 January 2019 to 10 March 2019 as well as a shorter summer trial from 16 July 2018 to 16 August 2018. The accuracy of the hourly ensemble up to T + 120 hr is studied using RPS as in Figure 12. Different sizes of neighbourhood scale are used to see whether an increase in the number of “pseudo” ensemble points can compete with a genuine increase in ensemble members to 24 and a neighbourhood no larger than 3 by 3. A 24-member 3 × 3 set-up provides 216 points. The 18-member 3 × 3 neighbourhood provides 162 points, whereas a 5 × 5 neighbourhood gives 450 points, 7 × 7 gives 882 points and 11 × 11 gives 2,178 points. The relative impact of increasing the ensemble size and neighbourhood scale not only depends on the field and season, but also on lead time. For example, Figure 12 shows that increasing to 5 × 5 for the 18-member ensemble provides some benefit at early lead times but increasing to 24 members with a 3 × 3 neighbourhood is better than the 18-member 3 × 3, that is, adding ensemble members is more beneficial. Further increase in neighbourhood size is not beneficial.

The impact of increasing the ensemble size is depicted in Hinton diagram scorecards in Figure 13a for the summer trial (the winter trial is not shown). For both the summer and winter trials, increasing the size of the ensemble leads to improvements at most lead times and for most fields. The improvements are small, however (of the order of 1%). We see again the impact of the DA with the age of the forecasts in the CRPS for temperature. Increasing the neighbourhood scale (Figure 13b,d) results in improvements for precipitation, cloud amounts, and wind speeds (although for the winter, this depends on the size of the neighbourhood as in Figure 12, perhaps this is due to the difference in predictability in wind patterns between the winter and the summer). Again, at longer lead times, these improvements are small. For visibility and temperature, fields sensitive to the local terrain, no type of neighbourhood processing improves the configuration, which suggests that an optimal ensemble size for MOGREPS-UK is obtained with a neighbourhood scale of 3 by 3 points (as in Mittermaier and Csima, 2017) at longer lead times as well. Again, any change here is very small. Finally, in Figure 13d, we show the comparison between increasing the neighbourhood size and increasing the ensemble size. Apart from very small detriments in cloud amount and wind speed in the summer, increasing the ensemble size is the best option overall.

This type of study reveals that future efforts should still focus on increasing the ensemble size as well as improving

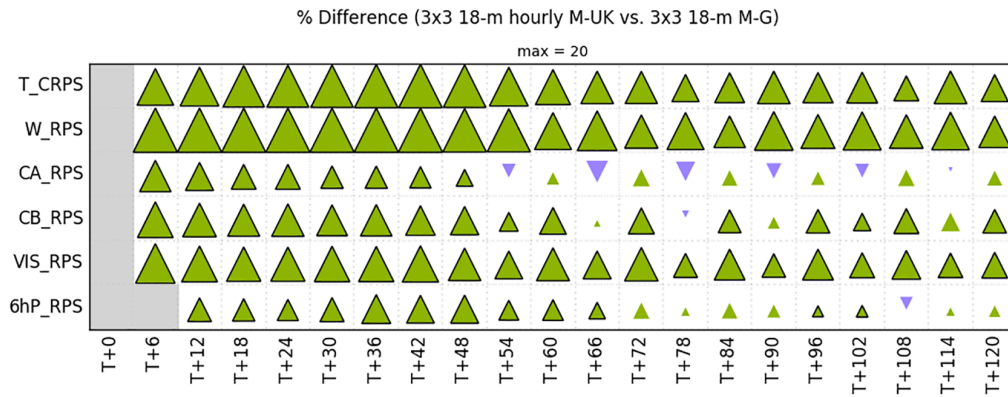


FIGURE 11 Hinton diagrams to display the CRPS for temperature (T_CRPS) and RPSs for wind (W_RPS), cloud amount (CA_RPS), cloud base (CB_RPS), visibility (VIS_RPS) and precipitation (6hP_RPS) from 1 September to 1 November 2019. Comparison of MOGREPS-UK 18-member hourly configuration for a neighbourhood scale of 3 by 3 points with MOGREPS-G for a neighbourhood scale of 3 by 3 grid points. Green triangles indicate 18-member hourly MOGREPS-UK is better. A black outline to the triangle indicates statistical significance at the 0.05 level, calculated using the Wilcoxon signed-rank test (Wilks, 2011). All data are equalised onto the validity and lead times for MOGREPS-G.

and developing further types of post-processing with the creation of smarter, more adaptive neighbourhood processing and statistical methods. Additional time-lagging may be beneficial for such long lead times as it increases the use of different MOGREPS-G driving conditions. Further investigations on the mechanisms controlling the growth of errors at these longer lead times may also be fruitful.

5 | CONCLUSIONS

Convective-scale ensembles are now widely used and developed in order to assess the uncertainty in the intensity, location and timing of high-impact weather. They increasingly constitute one of the primary sources of NWP forecast data.

Here, we study the performance of a new operational configuration for the Met Office's convective-scale ensemble: MOGREPS-UK. It relies on the hourly cycling of the high-resolution deterministic UKV model data-assimilation to generate a new set of initial conditions at each hour for the convective-scale ensemble. An hourly-updated 18-member ensemble now consists of six hourly time-lagged sets of three members and produces forecasts out to a lead time of 5 days.

This new hourly configuration is introduced to reduce the jumpiness between successive ensemble forecasts, to account for uncertainty in initial conditions by centring around consecutive hourly UKV analyses, to help in identifying forecast trends as well as to improve the spread of the ensemble.

Objective verification is used for a range of metrics (such as RPS, reliability, ROC areas, biases, standard

deviation between members and the mean and rank histograms) over different trial periods. Most of the results found here apply to all fields and both winter and summer seasons, except for the temperature which suffers from a strong diurnal cycle bias in the summer and a growing bias with lead times in the winter. For most fields, to ensure better performance at all lead times, a larger sized ensemble of the time-lagged configuration, compared to the 6-hourly configuration, should be used. This is due to the use of older driving conditions from MOGREPS-G as well as an increase in forecast age in the time-lagged configuration. We show that an 18-member time-lagged configuration performs better at mostly all lead times compared to a 12-member 6-hourly configuration. An ensemble size of 24 members is needed in the time-lagging configuration, in comparison to an ensemble size of 18 members in the 6-hourly configuration, in order to ensure better skill at all lead times. In general, a 24-member time-lagged ensemble also offers better discrimination and reliability than an 18-member time-lagged ensemble.

From the 6-hourly to the hourly configuration, the ensemble spread (i.e. standard deviation between the members and the mean) has increased by almost 15% at short lead times and approximately 5% at longer lead times. As well as increasing the ensemble spread, the hourly configuration contributes to a reduction in the standard deviation against observations, and so overall partially alleviates the lack of spread in the ensemble. Again, a larger ensemble size of 24 members contributes to a larger increase in ensemble spread as well as a reduction in the standard deviation against observations. It is important to note that, unlike the other metrics, the differences in spread do not depend on the size of the time-lagged ensemble, which gives better spread and better sampling of the

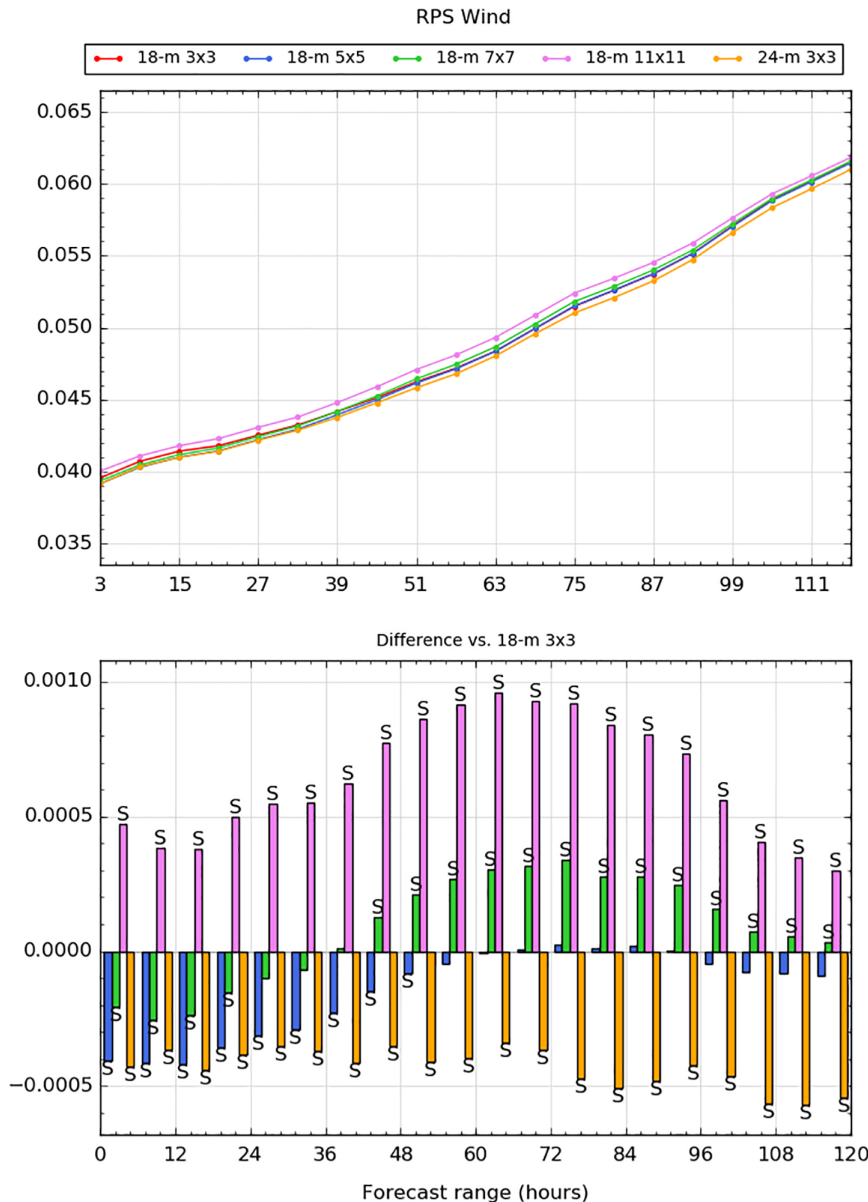


FIGURE 12 RPS with the hourly configuration to T + 120 hr for wind speed against synoptic observations from 1 January to 10 March 2019. Different sizes of neighbourhoods are used here (3 by 3 points, 5 by 5 points, 7 by 7 points, 11 by 11 points) for the 18-member configuration and a 3 by 3 neighbourhood is used in comparison for the 24-member ensemble. A black 'S' on the difference plot (lower panel) indicates statistical significance at the 0.05 level, calculated using the Wilcoxon signed-rank test (e.g. Wilks, 2011) All data are plotted every 6 hr.

pdf whether it uses the same number or a higher number of members than the 6-hourly configuration. However, the ensemble is still underspread and further efforts are needed to improve and assess its performance over a large sample of cases and trials.

In addition to the change from the 6-hourly to the hourly configuration, MOGREPS-UK was also upgraded to run to T + 120 hr. We therefore also assess the performance of the convective-scale ensemble against the Met Office's existing NWP systems running to T + 120 hr, the global ensemble MOGREPS-G and the high-resolution deterministic model UKV.

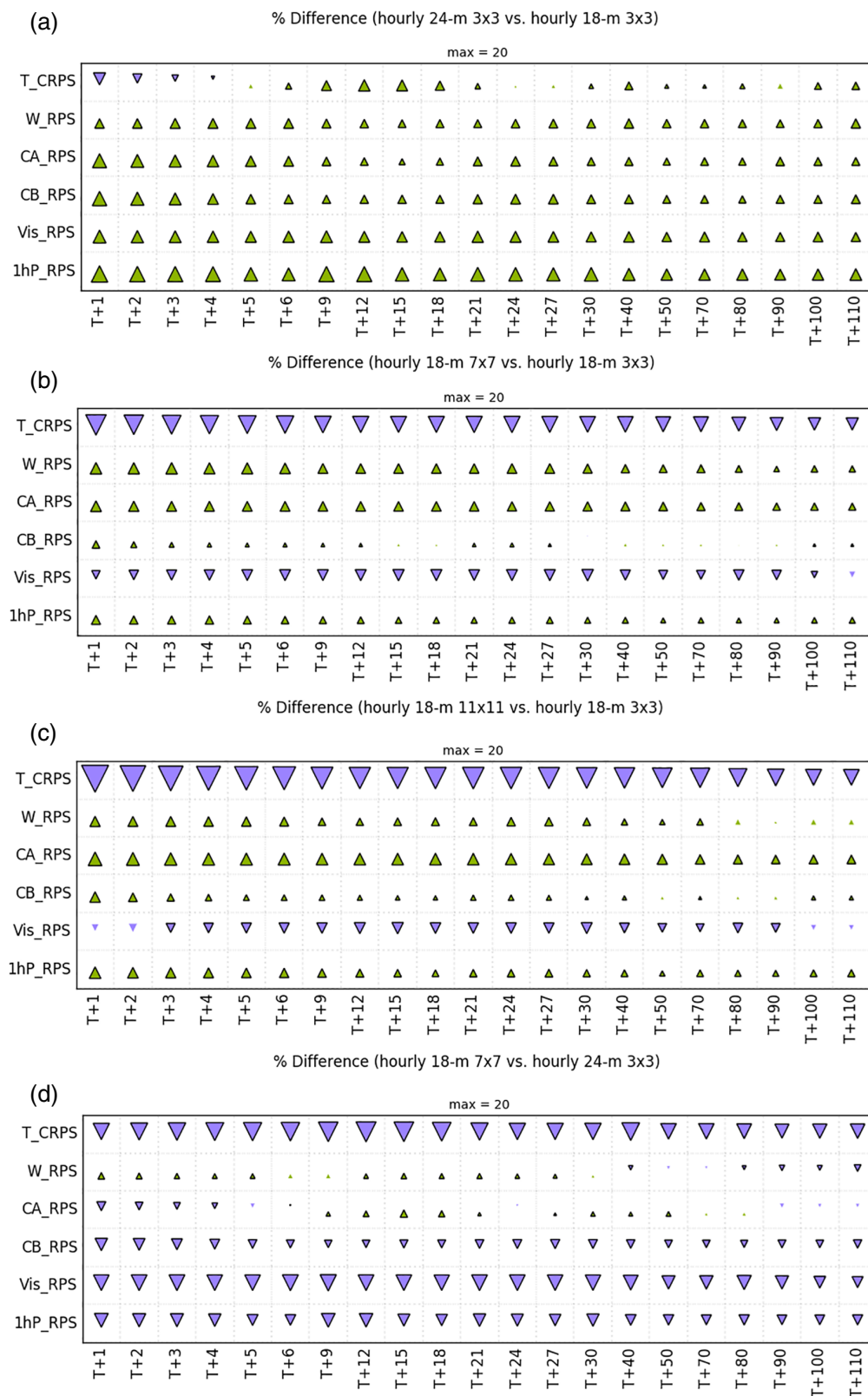
Against MOGREPS-G, MOGREPS-UK shows better skill, especially at short lead times, and especially for wind. For lead times as long as T + 80 hr, the improvements in the skill are smaller and lose statistical significance for some

fields. The benefits also relate more to improvements in discrimination than in reliability.

This article has highlighted some points for future research:

- More research is needed to quantify the spread of the ensemble for other variables such as clouds or precipitation using spatial verification methods (such as Dey *et al.*, 2014; 2016).
- While the hourly time-lagging configuration improves the spin-up of the forecasts, the same small-scale perturbations are applied to all three members running at each hour. Further work may be required in the data assimilation and stochastic physics to understand better the role of the small-scale perturbations per member on the spread and the skill of the ensemble.

FIGURE 13 Hinton diagrams to display the CRPS for temperature (T_CRPS) and RPSs for wind speed (W_RPS), cloud amount (CA_RPS), cloud base (CB_RPS), visibility (Vis_RPS) and precipitation (1hP_RPS) for the summer trial 16 July to 16 August 2018. (a) Impact of increasing the ensemble size to 24 members, (b) impact of increasing the neighbourhood scale to 7 by 7 grid points, (c) impact of increasing the neighbourhood scale to 11 by 11 grid points, and (d) relative impact of increasing the neighbourhood scale to 7 by 7 grid points against increasing the ensemble size



Whatever the scales of the perturbations, the impact of time-lagging will result in larger spread because of the variability provided by these older forecasts. However, the impact of time-lagging may also depend to a certain extent on the scales of the perturbations applied

to the ensemble: if only large-scale perturbations are used, the impact of time-lagging may be larger because of the introduction of small-scale perturbations by the older forecasts (which could provide a better spin-up of the active convection at early lead times). However,

if smaller-scale perturbations are used (Raynaud and Bouttier, 2016, for example), future work may need to address the impact of spin-up and the benefits of hourly time-lagging, depending on the variability and quality of the hourly cycling data assimilation against a less frequent cycling system.

- Further work is also required to assess the performance of the model for each successive cycle (i.e. jumpiness) in order to assess how the hourly refresh of the UKV analyses contributes to the performance of the ensemble forecast for high-impact cases. Weighting over ensemble members may need to be considered in this respect.
- Using objective verification, we show that MOGREPS-UK has better or similar skill to the other NWP systems running to T + 120 hr: the global ensemble MOGREPS-G and the deterministic UKV model. Future work is needed to sample a large set of case-studies to understand the performance of all these systems in high-impact weather.
- Compared to our operational configuration using six cycles of time-lagging with 18 members, we also showed that the skill of the ensemble is further improved by increasing the size of the time-lagged ensemble to eight cycles or 24 members. This may motivate the use of further time-lagging operationally in the future. The latest research at the grid scale shows that time-lagging to even 30 members (i.e. 10 hr) is still beneficial. Post-processing techniques are also valuable and could be further improved for fields like temperature and visibility.
- An investigation into the working practices of operational meteorologists would be useful. Whilst single model systems such as UKV and MOGREPS-UK continue to contain significant systematic errors, there is likely to be information missing that is available through an overview of multiple independent models, as are often used by operational meteorologists.
- Finally, whilst time-lagging improves the spread of the ensemble, in common with many convective-scale ensembles, the system is still believed to be under-dispersed and this is a high priority area of ongoing research.

ACKNOWLEDGEMENTS

The authors would like to thank Jonathan Flowerdew for valuable discussions about the verification metrics, the concept of time-lagging and the comparison of ensembles at different resolutions. The authors are also grateful to Anne McCabe for valuable inputs on ensemble spread. The authors would also like to thank the reviewers for their valuable suggestions to this work.

ORCID

Aurore N. Porson  <https://orcid.org/0000-0002-5023-8522>

REFERENCES

- Beck, J., Bouttier, F., Wiegand, L., Gebhardt, C., Eagle, C. and Roberts, N. (2016) Development and verification of two convection-allowing multi-model ensembles over Western Europe. *Quarterly Journal of the Royal Meteorological Society*, 142(700), 2808–2826. <https://doi.org/10.1002/qj.2870>.
- Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B. and Beare, S.E. (2008) The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134(632), 703–722. <https://doi.org/10.1002/qj.234>.
- Branković, Č., Palmer, T.N., Molteni, F., Tibaldi, S. and Cubasch, U. (1990) Extended-range predictions with ECMWF models: time-lagged ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, 116(494), 867–912. <https://doi.org/10.1002/qj.49711649405>.
- Bremner, F.J., Gotts, S.J. and Denham, D.L. (1994) Hinton diagrams: viewing connection strengths in neural networks. *Behavior Research Methods, Instruments, & Computers*, 26, 215–218. <https://doi.org/10.3758/BF03204624>.
- Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Brown, A., Milton, S., Cullen, M.J.P., Golding, B., Mitchell, J. and Shelly, A. (2012) Unified modeling and prediction of weather and climate: a 25-year journey. *Bulletin of the American Meteorological Society*, 93, 1865–1877. <https://doi.org/10.1175/BAMS-D-12-00018.1>.
- Bush, M., Allen, T., Boutle, I., Edwards, J., Finnenkoetter, A., Franklin, C., Hanley, K., Lean, H., Lock, A., Manners, J., Mittermaier, M., Morcrette, C., North, R., Petch, J., Short, C., Vosper, S., Walters, D., Webster, S., Weeks, M., Wilkinson, J., Wood, N. and Zerroukat, M. (2020) *The first Met Office Unified Model/JULES Regional Atmosphere and Land configuration, RAL1*, Vol. 13, Geoscientific Model Development, pp. 1999–2029. <https://doi.org/10.5194/gmd-13-1999-2020>.
- Clark, A.J., Jirak, I.L., Dembek, S.R., Creager, G.J., Kong, F.Y., Thomas, K.W., Knopfmeier, K.H., Gallo, B.T., Melick, C.J., Xue, M., Brewster, K.A., Jung, Y.S., Kennedy, A., Dong, X.Q., Markel, J., Gilmore, M., Romine, G.S., Fossell, K.R., Sobash, R.A., Carley, J.R., Ferrier, B.S., Pyle, M., Alexander, C.R., Weiss, S.J., Kain, J.S., Wicker, L.J., Thompson, G., Adams-Stelin, R.D. and Imy, D.A. (2018) The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Bulletin of the American Meteorological Society*, 99, 1433–1448. <https://doi.org/10.1175/BAMS-D-16-0309.1>.
- Dey, S.R.A., Leoncini, G., Roberts, N.M., Plant, R.S. and Migliorini, S. (2014) A spatial view of ensemble spread in convection permitting ensembles. *Monthly Weather Review*, 142, 4091–4107. <https://doi.org/10.1175/MWR-D-14-00172.1>.
- Dey, S.R.A., Plant, R.S., Roberts, N.M. and Migliorini, S. (2016) Assessing spatial precipitation uncertainties in a convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, 142(701), 2935–2948. <https://doi.org/10.1002/qj.2893>.

- Epstein, E.S. (1969) A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6), 985–987. Available at: <http://www.jstor.org/stable/26174707>.
- Flack, D.L.A., Gray, S.L., Plant, R.S., Lean, H.W. and Craig, G.C. (2018) Convective-scale perturbation growth across the spectrum of convective regimes. *Monthly Weather Review*, 146, 387–405. <https://doi.org/10.1175/MWR-D-17-0024.1>.
- Flowerdew, J. (2014) Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A*, 66(1), 22662. <https://doi.org/10.3402/tellusa.v66.22662>.
- Flowerdew, J. and Bowler, N.E. (2011) Improving the use of observations to calibrate ensemble spread. *Quarterly Journal of the Royal Meteorological Society*, 137(655), 467–482. <https://doi.org/10.1002/qj.744>.
- Flowerdew, J. and Bowler, N.E. (2013) On-line calibration of the vertical distribution of ensemble spread. *Quarterly Journal of the Royal Meteorological Society*, 139(676), 1863–1874. <https://doi.org/10.1002/qj.2072>.
- Frogner, I.-L., Andrae, U., Bojarova, J., Callado, A., Escribà, P., Feddersen, H., Hally, A., Kauhanen, J., Randriamampianina, R., Singleton, A., Smet, G., van der Veen, S. and Vignes, O. (2019b) HarmonEPS – the HARMONIE ensemble prediction system. *Weather and Forecasting*, 34, 1909–1937. <https://doi.org/10.1175/WAF-D-19-0030.1>.
- Frogner, I.-L., Singleton, A.T., Koltzow, M.Ø. and Andrae, U. (2019a) Convective-permitting ensembles: challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 90–106. <https://doi.org/10.1002/qj.3525>.
- Gebhardt, C., Theis, S.E., Paulat, M. and Ben Bouallègue, Z. (2010) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, 100, 168–177. <https://doi.org/10.1016/j.atmosres.2010.12.008>.
- Golding, B., Roberts, N.M., Leoncini, G., Mylne, K.R. and Swinbank, R. (2016) MOGREPS-UK convection-permitting ensemble products for surface water flood forecasting: rationale and first results. *Journal of Hydrometeorology*, 17, 1383–1406. <https://doi.org/10.1175/JHM-D-15-0083.1>.
- Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 143(708), 2846–2861. <https://doi.org/10.1002/qj.3135>.
- Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570. [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Hinton, G.E. and Shallice, T. (1991) Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review*, 98(1), 74–92. <https://doi.org/10.1037/0033-295X.98.1.74>.
- Keil, C., Baur, F., Bachmann, K., Rasp, S., Schneider, L. and Barthlott, C. (2019) Relative contribution of soil moisture, boundary-layer and microphysical perturbations on convective predictability in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 145(724), 3102–3115. <https://doi.org/10.1002/qj.3607>.
- Kendon, E.J., Roberts, N.M., Senior, C.A. and Roberts, M.J. (2012) Realism of rainfall in a very high-resolution regional climate model. *Journal of Climate*, 25, 5791–5806. <https://doi.org/10.1175/JCLI-D-11-00562.1>.
- Klasa, C., Arpagaus, M., Walser, A. and 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(712), 744–764. <https://doi.org/10.1002/qj.3245>.
- Klasa, C., Arpagaus, M., Walser, A. and Wernli, H. (2019) On the time evolution of limited-area ensemble variance: case studies with the convection-permitting ensemble COSMO-E. *Journal of the Atmospheric Sciences*, 76, 11–26. <https://doi.org/10.1175/JAS-D-18-0013.1>.
- Kuchera, E. and Rentschler, S. (2019) *Ensemble efforts for the US Air Force*. Presentation at the 8th NCEP Ensemble User Workshop. <https://ral.ucar.edu/events/2019/8th-ncep-ensemble-user-workshop>.
- Kühnlein, C., Keil, C., Craig, G.C. and Gebhardt, C. (2014) The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quarterly Journal of the Royal Meteorological Society*, 140(682), 1552–1562. <https://doi.org/10.1002/qj.2238>.
- Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F. and Buizza, R. (2001) A strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events. *Quarterly Journal of the Royal Meteorological Society*, 127(576), 2095–2115. <https://doi.org/10.1002/qj.49712757613>.
- McCabe, A., Swinbank, R., Tennant, W. and 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(700), 2897–2910. <https://doi.org/10.1002/qj.2876>.
- Milan, M., Macpherson, B., Tubbs, R., Dow, G., Inverarity, G., Mittermaier, M., Halloran, G., Kelly, G., Li, D., Maycock, A., Payne, T., Piccolo, C., Stewart, L. and Wlasak, M. (2020) Hourly 4D-Var in the Met Office UKV operational forecast model. *Quarterly Journal of the Royal Meteorological Society*, 146(728), 1281–1301. <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(627), 1487–1500. <https://doi.org/10.1002/qj.135>.
- Mittermaier, M.P. (2014) A strategy for verifying near-convection-resolving model forecasts at observing sites. *Weather and Forecasting*, 29, 185–204. <https://doi.org/10.1175/WAF-D-12-00075.1>.
- Mittermaier, M.P. and Csima, G. (2017) Ensemble versus deterministic performance at the kilometer scale. *Weather and Forecasting*, 32, 1697–1709. <https://doi.org/10.1175/WAF-D-16-0164.1>.
- Porson, A.N., Hagelin, S., Boyd, D.F.A., Roberts, N.M., North, R., Webster, S. and Lo, J.C.-F. (2019) Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Quarterly Journal of the Royal Meteorological Society*, 145(724), 3004–3022. <https://doi.org/10.1002/qj.3601>.
- Primo, C., Ferro, C.A.T., Jolliffe, I.T. and Stephenson, D.B. (2009) Calibration of probabilistic forecasts of binary events. *Monthly Weather Review*, 137, 1142–1149. <https://doi.org/10.1175/2008MWR2579.1>.
- Raynaud, L. and Bouttier, F. (2016) Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quarterly Journal of the Royal Meteorological Society*, 142(695), 854–866. <https://doi.org/10.1002/qj.2686>.

- Raynaud, L. and Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143(709), 3037–3047. <https://doi.org/10.1002/qj.3159>.
- Roberts, N.M. and Lean, H.W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97. <https://doi.org/10.1175/2007MWR2123.1>.
- Schumacher, R.S. and Clark, A.J. (2014) Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Monthly Weather Review*, 142, 4108–4138. <https://doi.org/10.1175/MWR-D-13-00357.1>.
- Schwartz, C.S. (2019) Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Monthly Weather Review*, 147, 2997–3023. <https://doi.org/10.1175/MWR-D-18-0452.1>.
- Schwartz, C.S., Kain, J.S., Weiss, S.J., Xue, M., Bright, D.R., Kong, F.Y., Thomas, K.W., Levit, J.J., Coniglio, M.C. and Wandishin, M.S. (2010) Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, 25, 263–280. <https://doi.org/10.1175/2009WAF2222267.1>.
- Schwartz, C.S., Romine, G.S., Smith, K.R. and 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. <https://doi.org/10.1175/WAF-D-13-00145.1>.
- Schwartz, C.S., Romine, G.S., Sobash, R.A., Fossell, K.R. and Weisman, M.L. (2015) NCAR's experimental real-time convection-allowing ensemble prediction system. *Weather and Forecasting*, 30, 1645–1654. <https://doi.org/10.1175/WAF-D-15-0103.1>.
- Schwartz, C.S., Romine, G.S., Fossell, K.R., Sobash, R.A. and Weisman, M.L. (2017) Toward 1-km ensemble forecasts over large domains. *Monthly Weather Review*, 145, 2943–2969. <https://doi.org/10.1175/MWR-D-16-0410.1>.
- Schwartz, C.S., Romine, G.S., Sobash, R.A., Fossell, K.R. and Weisman, M.L. (2019) NCAR's real-time convection-allowing ensemble project. *Bulletin of the American Meteorological Society*, 100, 321–343. <https://doi.org/10.1175/BAMS-D-17-0297.1>.
- Schwartz, C.S. and Sobash, R.A. (2017) Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: a review and recommendations. *Monthly Weather Review*, 145, 3397–3418. <https://doi.org/10.1175/MWR-D-16-0400.1>.
- Tennant, W. (2015) Improving initial condition perturbations for MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 141(691), 2324–2336. <https://doi.org/10.1002/qj.2524>.
- Walters D, Baran AJ, Boutle I, Brooks M, Earnshaw P, Edwards J, Furtado K, Hill P, Lock A, Manners J, Morcrette C, Mulcahy J, Sanchez C, Smith C, Stratton R, Tennant W, Tomassini L, Van Weverberg K, Vosper S, Willett M, Browse J, Bushell A, Carslaw K, Dalvi M, Essery R, Gedney N, Hardiman S, Johnson B, Johnson C, Jones A, Jones C, Mann G, Milton S, Rumbold H, Selrar A, Ujiie M, Whitall M, Williams K, Zerroukat M. (2019) The Met Office Unified Model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geoscientific Model Development*, 12, 1909–1963. doi: 10.5194/gmd-12-1909-2019.
- Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P.L., Wilkinson, J., Willett, M., Woolnough, S. and Xavier, P. (2017) The Met Office Unified Model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geoscientific Model Development*, 10, 1487–1520. <https://doi.org/10.5194/gmd-10-1487-2017>.
- Weyn, J.A. and Durran, D.R. (2018) Ensemble spread grows more rapidly in higher-resolution simulations of deep convection. *Journal of the Atmospheric Sciences*, 75, 3331–3345. <https://doi.org/10.1175/JAS-D-17-0332.1>.
- Weyn, J.A. and Durran, D.R. (2019) The scale dependence of initial-condition sensitivities in simulations of convective systems over the southeastern United States. *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 57–74. <https://doi.org/10.1002/qj.3367>.
- Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edition. Amsterdam: Elsevier, pp. 162–166.
- WMO. (2018) *Manual on codes - international codes, volume I.1, annex II to the WMO technical regulations: part A- alphanumeric codes*. Geneva, Switzerland: WMO Publication No. 306. https://library.wmo.int/doc_num.php?explnum_id=5708.
- Xu, M., Thompson, G., Adriaansen, D.R. and Landolt, S.D. (2019) On the value of time-lag-ensemble averaging to improve numerical model predictions of aircraft icing conditions. *Weather and Forecasting*, 34, 507–519. <https://doi.org/10.1175/WAF-D-18-0087.1>.
- Yuan, H., Mullen, S., Gao, X., Sorooshian, S., Du, J., Juang, H., Lu, C., McGinley, J.A., Schultz, P.J., Jamison, B.D., Wharton, L. and Anderson, C.J. (2009) Evaluation of shortrange quantitative precipitation forecasts from a time-lagged multimodel ensemble. *Weather and Forecasting*, 24, 18–38. <https://doi.org/10.1175/2008WAF2007053.1>.
- Zhang, X. (2019) Multiscale Characteristics of Different-Source Perturbations and Their Interactions for Convection-Permitting Ensemble Forecasting during SCMREX. *Mon. Wea. Rev.*, 147(1), 291–310. <https://doi.org/10.1175/MWR-D-18-0218.1>.
- Ebert, E. E. (2008) Fuzzy verification high-resolution gridded forecasts: a review and proposed framework. *Meteorol. Appl.*, 15, 51–64. <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.25>.

How to cite this article: Porson AN, Carr JM, Hagelin S, *et al.* Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Q J R Meteorol Soc.* 2020;1–21. <https://doi.org/10.1002/qj.3844>