Exploiting the Benefits of Convection-Permitting Ensembles

Adam Gainford

Department of Meteorology

University of Reading

A thesis presented for the degree of

Doctor of Philosophy

September 2025

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Adam Gainford

Acknowledgements

First I would like to give a huge thanks to my supervisors Sue Gray, Tom Frame, Aurore Porson and Marco Milan for keeping the project headed in the right direction and helping me navigate the PhD process. I have learned so much from each of them, and I would not have reached this stage without their continued support, encouragement, and frequent patience. Thanks also go to my monitoring committee, Thorwald Stein and John Methven, for their encouragement and useful discussions.

Further thanks go to everyone at the Met Office who have made me feel welcome during visits, including Anne McCabe, David Flack, David Walters, Rebekah Hicks, and Rob Neal, among many others.

I would also like to thank everyone in the meteorology department that I have come to know over the years, especially the other PhD students going through similar journeys. Special thanks to Hannah Croad, Indrakshi Mukherjee, Juan Garcia-Valencia, Xueqing Ling, and anyone else who has called 1U07 their home for all the fun, gossip, and occasional game of office table tennis.

Thanks also to my friends in Cumbria, Chris Pepper, Grant Forsyth and Mark Page, for reminding me of the world outside of academia and keeping me company on my visits back north.

Also, a huge thanks to my mum Sheila, my sister Tracey, and my partner's family Darren, Tracey and Jodie for always being there to help during the tough times and to celebrate during the good. And finally, all my love to my partner Connor, who has been with me throughout this long, sometimes stressful, but often rewarding journey. I certainly wouldn't be here without his constant love and support and I'll be forever grateful.

Oh, and I can't forget the cats Tia and Bob, who I'm sure are causing mischief to the residents of Keswick as we speak!

Abstract

The development and use of convection-permitting ensembles (CPEs) within operational meteorological centres has improved the quality of guidance available for decision making, especially for precipitation. The explicit representation of km-scale details provides better quantitative predictions of hazardous convection than models that must rely on parametrizations. Despite these benefits, long-standing spread deficiencies limit the usefulness of CPE outputs, restricting their full potential. Therefore, efforts to elucidate CPE characteristics can be highly beneficial for both operational and research purposes. Here, three studies are presented that link the behaviour of CPEs run over the UK to the synoptic regime, which provides a source of predictability, using the Met Office's weather forecast models. These studies focus on analysing precipitation forecasts, which we expect to benefit the most from operating at convective scales.

It has long been understood that CPEs do not possess enough spread in the placement of precipitation. In the first study, the effect of blending the more accurate large-scale initial conditions of the global model into the CPE analysis is assessed, demonstrating that the spatial spread-skill relationship is improved by a modest but significant amount. Similarly, it is also well known that the synoptic-scale variability introduced through the driving ensemble can have a large influence on the evolution of nested CPE forecasts. The second study quantifies the driving-ensemble influence, which is strongest under mobile regimes and weakest under conditionally unstable regimes. To facilitate these studies, new spatial analysis methods, including the parent-child Fractions Skill Score, have been developed that provide broader insights into the behaviour of CPE forecasts. In the final study, an experimental post-processing system driven by spatial methods is applied to CPE forecasts to assess the value of clustering members based on the co-location of precipitation features. CPE clustering is shown to be a reliable tool, and provides the most value when targeted over regions that will be impacted by hazardous convection.

Contents

[The radar] said 40% chance of rain, but that was wrong. 100% chance of weather, that's all you can say.

David Croft, 2012 Brazilian GP

Chapter 1

Introduction

Many studies have demonstrated the benefits of running numerical weather prediction models at scales that are fine enough to represent convection explicitly (e.g., Cafaro et al. 2019; Clark et al. 2009; Hanley et al. 2011; Lean et al. 2008; Roberts et al. 2009; Schwartz et al. 2010; Woodhams et al. 2018). These models provide more realistic forecasts of convection than models that must instead rely on convective parametrizations, which improves the forecasts of their impacts. For instance, wintery showers that develop over warm seas can realistically advect inland within convection-permitting models, whereas they remain confined to the areas of instability in convection-parametrized models (Lean et al. 2008). However, this enhanced realism does not necessarily translate to enhanced accuracy, as the skill of these models is limited by errors that grow faster than on coarser grids (Lorenz 1969). As such, determining the precise locations that will be impacted by extreme convection can usually only be done once the system has developed and its subsequent evolution becomes more dependent on predictable mechanisms, such as steering flows. Therefore, convection-permitting ensembles (CPEs) are now run routinely at meteorological centres around the world to quantify the uncertainty associated with short-range, high-impact events (e.g., Frogner et al. 2019b; Porson et al. 2020; Reinert et al. 2025).

Developing and maintaining these systems requires a large commitment of research and computational resources. From the assimilation of more observations and the development of convective-scale physics schemes, to the hardware that underpins the day-to-day operations, meteorological centres expend large effort to understand and improve the performance of CPEs. There is therefore a strong desire to exploit the wealth of information produced daily by these ensembles, as well as to address long-standing deficiencies with their outputs. For instance, convection is still considered to be under-resolved, which leads to delays in initiation and a lack of organisation of individual cells into more complex structures (Lean et al. 2024). It has also long been understood that CPEs are underspread, and that objective evaluations show that CPEs around the world follow the control member too closely (e.g., Beck et al. 2016; Cafaro et al. 2021; Clark et al. 2009; Frogner et al. 2019a; Klasa et al. 2018; Schwartz et al. 2014). These biases can have a significant impact, as the verified outcome may not lie within the distribution predicted by the ensemble, and the spread may not be a good indicator of eventual skill.

Many techniques have been employed to introduce more spread within CPEs, such as staggered initialisation schedules (Porson et al. 2020; Raynaud and Bouttier 2017) and improved model physics schemes (Flack et al. 2021; McCabe et al. 2016). However, it is also important to have good knowledge of these biases and the situations in which they may be more or less impactful. Hence, this thesis focuses on improving our understanding of the conditions in which synoptic-scale variability, usually introduced through the driving ensemble, affects CPEs. For instance, it is well known that the driving ensemble plays an important role in the evolution of each CPE member through the inheritance of information via lateral boundary conditions (Clark et al. 2010). This synoptic information is likely to evolve in a similar fashion in each model once these forcings are consistent between the two ensembles. However, the potential for the CPE to respond to this synoptic forcing at smaller scales is the key to evaluating the added value from running the CPE. After all, if the outputs from the CPE simply resemble that of the driving ensemble, just on a finer grid, this arguably does not yield enough benefit for operating at these scales. Therefore, it is important to quantify the influence of the driving ensemble to understand the conditions under which the CPE provides more value. These trends can then be linked to the spatial spread of precipitation to explore the situations and regimes in which the CPE also provides value through additional spread, as well as through different forecast scenarios.

It is also known that the limited area data assimilation schemes will have deficiencies when initialising forecasts at scales similar to the model domain size (Milan et al. 2023). These deficiencies could be corrected by blending the accurate small-scale information with the accurate large-scale information from the global model. This change should yield improved CPE skill, but it is currently unknown whether this will also improve the spread-skill relationship by reducing the disparity with spread.

These questions will be addressed in subsequent research chapters, with a particular focus on assessing precipitation outputs. While a lot of work has been done to understand the spread of other variables, understanding the spread of precipitation forecasts is likely to be one of the most useful due to the improved representation of convection within CPEs compared to global models. Hazardous convection is still very difficult to model but can bring large impacts from flash flooding. However, verifying precipitation fields requires the use of neighbourhood-based metrics for robust evaluation that avoids the double penalty problem (Gilleland et al. 2009; Roberts and Lean 2008). Therefore, fewer studies have been conducted which focus on the spatial spread of precipitation patterns (Dey et al. 2014). There is also the potential for these techniques to be exploited further. For instance, extensions of these neighbourhood methods have been used to estimate the physical distance between forecasts and verification (Skok 2022; Skok and Roberts 2018), and to provide localised analysis of skill (Woodhams et al. 2018).

Finally, there is the need to develop tools that can use these techniques for more intelligent summaries of ensemble information. Forecasters regularly face information overload from the amount of data available, meaning that the potential of the ensemble is not being fully realised for decision making. One such answer to this problem is the development of clustering techniques that can identify groups of similar members within an ensemble and provide distinct forecast storylines for each of those groups (Boykin 2022). This tool has shown promise during trials at the (UK) Met Office (UKMO) when run on global ensembles, but it is currently an open question whether there is additional value to be gained from running clustering on CPEs.

In summary, this thesis presents a collection of studies to improve our understanding of the role that synoptic-scale information plays in the performance of CPE precipitation forecasts. In particular, the research questions posed by this thesis are:

RQ1 How can the global ensemble be used to better understand and improve spatial precipitation spread within CPEs?

RQ2 How dependent is spatial spread and skill of precipitation in CPEs on the synoptic-scale flow?

RQ3 How can the benefits of spatial verification methods be utilised to understand CPE behaviour?

RQ4 How can spatial techniques be further exploited for operational purposes?

These questions are addressed using convection-permitting ensembles run at the UKMO covering the UK domain. To answer these questions, chapter 3 explores the impact that a new data assimilation scheme developed at the UKMO has on the spatial spread-skill relationship of precipitation during the early periods of a CPE. This scheme constrains the large scales within the convective-scale analysis to follow those of the more accurate global analysis. Additionally, a new spatial verification method is used to assess the local impacts of this scheme on a case study. Then, in chapter 4, spatial-spread statistics are compared between a CPE and its driving ensemble to understand the leadtimes and regimes when the CPE inherits its spread from the driving ensemble, and the situations when the CPE adds more value. This study is aided by the development of another novel technique that directly compares driving-nested member pairs between the two ensembles. It is further demonstrated that this technique can be useful operationally by highlighting periods within precipitation forecasts when the ensembles tell the same story, and periods when the ensembles diverge. Finally, in chapter 5, the value of applying feature-based clustering techniques to CPEs is assessed and compared against running clustering on coarser global ensembles over the same region. This study is also the first to assess the reliability of clustering over an extended period. Then, an impactful case study is analysed in detail to highlight the operational benefits that can be expected from running clustering on CPEs.

Chapter 2

Literature Review

2.1 Uncertainty in Numerical Weather Prediction

2.1.1 Origins of Ensemble Forecasting

The ultimate goal of any numerical weather prediction (NWP) model is to provide decision-makers with a more informed estimate of the upcoming weather than could be achieved through more rudimentary methods. For example, in some regions, the weather rarely deviates from its climatalogical average, and predictions can easily be made by hand. In most regions of the Earth, though, accurately predicting the upcoming weather can only be achieved by dedicating computational resources to running NWP simulations. These simulations are performed by taking the best guess of the current state of the atmosphere (the "analysis"), as provided by data-assimilation techniques, and evolving this state based on governing dynamical principles. Such problems are only tractable using supercomputers, and thus advances in NWP only came after the invention of the transistor. Lewis Fry Richardson attempted to do this manually in 1922 (Richardson 1922), and while his final prediction implied an unphysically large change in pressure, his failure was born from imbalanced initial conditions, not the fundamental technique used (Kalnay 2002).

The "numerical" term in NWP refers to the method used to evolve the atmospheric state. All dynamical equations that govern the time-evolution of the atmosphere derive from the Navier-Stokes equations of fluid dynamics, which are unsolvable analytically except in simple cases. NWP, therefore, operates by constructing an atmospheric analogue, where physical values are discretised onto a grid with finite spacings and evolved using finite timesteps. This spatio-temporal discretisation is an unavoidable source of uncertainty within NWP. However, it has long been recognised that greater sources of uncertainty can contribute to forecast errors. Even during the first successful NWP experiments, Charney 1951 acknowledged the contributions of subgrid physical processes (in particular, condensation, radiation, and fluxes of heat, momentum, and moisture) to forecast accuracy. Rather than being handled by the model, such processes must be accounted for using separate parametrization schemes. Even with modern methods of representing these processes, including the increasing use of machine learning, the simplifications used to make these schemes efficient are an additional source of model error.

Each of these model approximations, as well as the errors associated with using imperfect initial and boundary conditions, introduce uncertainty into NWP forecasts. In the early days of NWP, however, the problems introduced by these uncertainties were not viewed as so insurmountable that NWP must deviate from its purely deterministic framework. In his exhaustive study into the origins of ensemble forecasting, Lewis 2005 discusses the experiments and realisations that that led to the adoption of uncertainty as a core facet of NWP. In particular, it was the work of Eric Eady and Philip Thomson in the 1950s that raised the alarm about the need to move away from the "pleasingly deterministic" view of NWP. Then, in his pair of seminal papers, Edward Lorenz reported on NWP predictability limits due to flow instability (Lorenz 1963a,b). This work was inspired by a failed simulation that was restarted using a truncated output, whereby Lorenz noticed that the growth of truncation errors had overwhelmed the main signal within a few simulated months. Later, Lorenz showed how these error growth rates depended on the scale of the flow (Lorenz 1969). The inevitable conclusion from these studies is that there are fundamental predictability limits governed by the chaotic nature of the atmosphere that limit the usefulness of a single deterministic forecast, even if all sources of model error can be eliminated.

Following these discoveries, it was clear that more advanced models were needed that could quantify forecast uncertainty with leadtime. Edward Epstein's stochastic-dynamic view was the first to propose the use of multiple deterministic "members" that were each perturbed by a small amount, rather than using just a single deterministic forecast (Epstein 1969). But the practical tests of these systems were severely limited by the computational power of the time. It wasn't until the start of the 1990s, when parallel processing had become sufficiently advanced and more work had been done on the best perturbation methods, that ensembles as we know them started running operationally.

2.1.2 Ensemble Generation

Even with a perfect model, it is expected that forecast accuracy degrades with increasing leadtime as the initial conditions become a less accurate predictor of future events due to atmospheric chaos. To anticipate the skill of a given forecast, an ensemble can be used to understand the predictability of that forecast. In the most basic form, an ensemble is a collection of deterministic model runs whereby

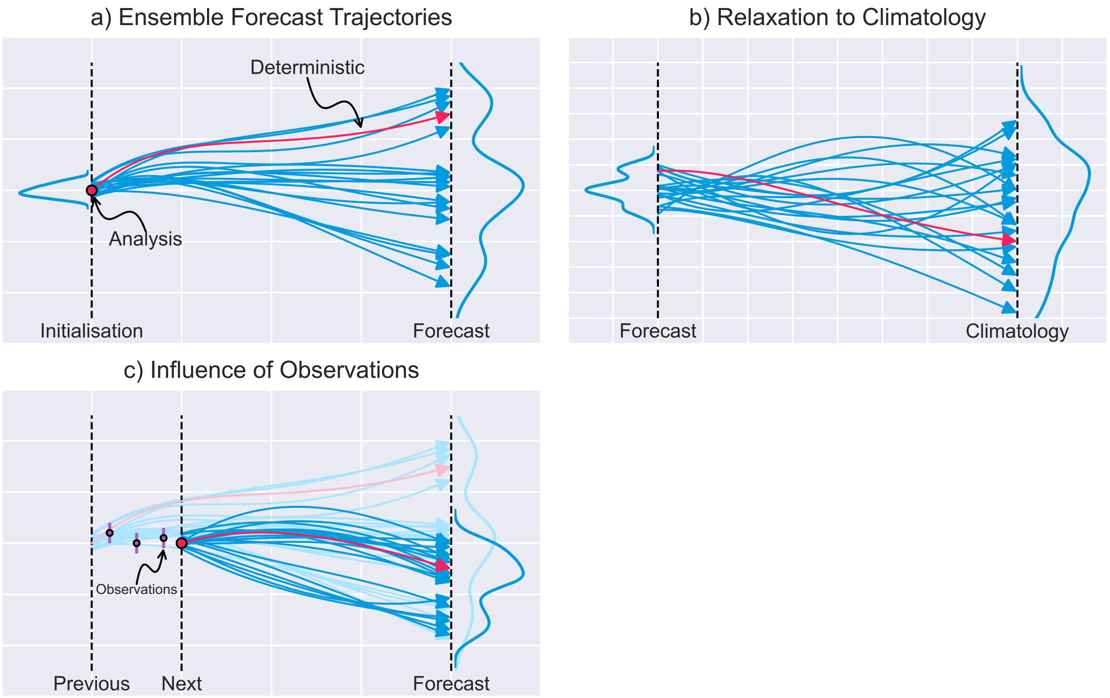

Figure 2.1: Ensemble trajectories throughout the course of a forecast. Panel a) shows the transition from the initial, sharp Gaussian distribution to a multimodal distribution at a later time. Panel b) then shows this multimodal distribution relaxing to a broader climatalogical Gaussian. Panel c) shows the influence that observations have on subsequent forecasts.

each member is initially perturbed by an amount consistent with the analysis error and then integrated forwards in time. Large agreement among ensemble members reflects high forecast confidence and an expectation of high accuracy. Low agreement suggests uncertainty and less accuracy. This link between ensemble agreement (spread) and skill will be discussed further in Section 2.1.3.

Usually, ensembles also include a control member which uses the unperturbed analysis as its starting state. Figure 2.1 shows the typical evolution of an ensemble forecast. It may be useful to consider these trajectories as representing the change in a particular meteorological value at a particular location. In Figure 2.1a), the ensemble is initialised with a small spread of values normally distributed about the analysis. As each member is integrated forwards in time, the spread of values grows. At a later time, it may be possible to distinguish multiple different modes in the probability density function (pdf), where members have become grouped into distinct regimes. Each regime may represent a broadly different forecast scenario, and each member represents subtle changes in the exact outcomes of those regimes. Eventually, though, this multimodal distribution will relax into the base climatalogical state, as represented in Figure 2.1b). At this stage, we assert that there is little use in continuing to run the simulation since each member behaves as if it is drawn from the climatology. Therefore, to make predictions at further instances in the future, it is necessary to constrain each member using recent observations, as shown in Figure 2.1c). In this example, the observations favour an evolution towards the lower end of the distribution, which results in the next forecast initialisation also being shifted to lower values. The distribution of this updated forecast is narrower than for the previous forecast due to the more recent initialisation.

The basic principle driving the construction of ensembles is that of equiprobability - that each ensemble member can be regarded as an equally likely future outcome (Epstein 1969). This principle allows for the convenient estimation of uncertainty simply by finding the proportion of members that predict a given event. While recent studies argue that this interpretation is too strict, and that ensemble members should merely be considered exchangeable rather than equiprobable (that is, that the joint distribution of a given parameter remains unchanged under permutation), this refinement does not meaningfully change the processes by which we should handle ensemble data (Bröcker and Kantz 2011).

The initialisation of an ensemble requires the use of an effective perturbation strategy. In an ideal world with a perfect model and a large, well-sampled ensemble, one could routinely produce an accurate representation of the initial condition error at all times. In practise, models are not perfect and samples are small, so the initial condition error is not well constrained. Therefore, perturbation strategies are designed to produce a spread of outputs that represent either the analysis uncertainty or the range of potential atmospheric evolutions. Despite the potential implications of the exchangeability principle, performing Monte Carlo style random sampling of a given uncertainty distribution was found to produce an insufficient sampling for running an ensemble (Epstein 1969; Leith 1971, 1974). Instead, it was recognised that sampling the fastest growing modes was a larger priority for the ensemble, and thus the need for more advanced perturbation techniques became clear. By the late 1990s, two schools of thought had emerged for perturbing global ensembles: The singular vector method developed at the European Centre for Medium-Range Weather Forecasts (ECMWF, Buizza and Palmer 1995; Molteni et al. 1996), and the Lyapunov or breeding vector method developed at the National Centers for Environmental Prediction (NCEP, Toth and Kalnay 1997). The singular vector method applies Singular Value Decomposition to a simplified atmospheric representation of the current state to find the errors after a 48 h period. The fastest growing errors from this decomposition are then applied as perturbations for running the full ensemble. Meanwhile, the breeding vector method attempts to maintain existing differences between members by rescaling their analysis errors for use as perturbations for the next cycle. By directly incorporating the analysis into the perturbation strategy, future ensemble cycles benefit from more realistic evolutions compared to singular vectors, although this is at the expense of slower growth rates (Magnusson et al. 2008).

The singular vector and breeding vector methods are not the only perturbation strategies, with more recent developments iterating and improving on these early designs. The perturbed observation method developed by the Canadian Meteorological Center applies semi-random noise to each member, where the statistics of the perturbations are consistent with the error covariance of the analysis. This scheme has been found to improve ensemble statistics compared to those that were initialised using singular or bred vector approaches (Hamill et al. 2000). Similarly, the ensemble Kalman filter method (and the computationally cheaper ensemble transform Kalman filter) is designed to improve spread growth at the earliest leadtimes compared to the singular vector method, which is targetted more at the medium range (Bishop et al. 2001). The ensemble transform Kalman filter is similar to the breeding vector method, but rather than being derived from a single member-control forecast difference, each perturbation is derived as a linear combination of all member-control differences, which also ingests observations (Bowler et al. 2008). This mixing allows perturbations to grow faster during earlier leadtimes. Aided by advancements in compute power and efficiency, more recent techniques apply data assimilation to each individual ensemble member, which helps improve the accuracy of each member at short leadtimes (Inverarity et al. 2023).

Initial error is just one source of uncertainty to consider, however. An effective ensemble should aim to represent all sources of uncertainty, not just those intrinsic to the flow or associated with initial conditions (Buizza et al. 1999). Model error is another source of uncertainty, but is much harder to quantify. Model uncertainties arise from the imperfect representation of the atmosphere in models due to discretisation and the reliance on parametrization schemes to represent subgrid processes. Three main techniques are used to account for model error: multi-physics schemes, stochastic physics schemes, and multi-model ensembles. Multi-physics schemes are developed on the basis that no single parametrization of a given process will perform better than another over the full range of conditions that it is expected to represent and thus subgrid uncertainty can be represented by tweaking the physics scheme parameters used in each member (e.g., Berner et al. 2011; Charron et al. 2010).

Stochastic physics schemes may also aim to represent the uncertainty in subgrid processes, but usually do so by perturbing the resolved parameters of the model rather than the unresolved ones. The motivation for this stems from the interpretation of a parameter as a mean of the unresolved processes. If these unresolved processes operate at faster timescales than the resolved processes, the use of a mean is appropriate for the timestep as dictated by the resolved parameters. In practise, there is not always a large separation of timescales, so stochastic physics schemes aim to represent this additional variability. Among stochastic physics schemes, there are four that are most commonly used. Firstly, Stochastic Kinetic Energy Backscatter schemes inject additional energy into resolved parameters that is dissipated during the advection of information from one timestep to the next (e.g., Berner et al. 2009; Tennant et al. 2011). Secondly, Stochastically Perturbed Physics Tendencies apply adjustments to the outputs of all parametrization schemes rather than targetting specific schemes, and is used as a more general method for representing subgrid uncertainty (e.g., Buizza et al. 1999). Thirdly, Stochastically Perturbed Parameter Schemes apply stochastic perturbations to individual physics parametrizations (e.g., Bowler et al. 2008; Christensen et al. 2015; McCabe et al. 2016). Lastly, additive inflation applies random noise drawn from a known distribution to the resolved variables of each ensemble members which helps increase spread during early leadtimes (Yang et al. 2015). It is common for ensembles to use multiple stochastic physics schemes, since each will represent a different source of model uncertainty (Berner et al. 2011; Charron et al. 2010).

2.1.3 Traditional Ensemble Verification

Whether using a single or multi-model approach, it is important that ensembles fulfil certain criteria such that forecasters can trust their outputs and have confidence in the guidance they produce. For NWP models to be useful for forecasters, it is important that they perform to a high standard as evaluated using certain metrics. The most obvious aspect of model performance that developers of any NWP system try to maximise is accuracy. Even with a perfect model, it is expected that forecast accuracy degrades with increasing leadtime as the initial conditions become a less reliable predictor of future events due to atmospheric chaos. Forecast accuracy is typically measured by comparing the difference in values at each grid point to observation and averaging the differences over all grid points. The simplest comparison metrics are the Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) which all ensure differences between model and verification do not cancel during grid point summation.

As well as forecast accuracy, it is also useful to estimate the added value that has been provided when compared to a simple baseline (e.g., a persistence forecast or climatology). Metrics which account for these factors are called skill scores. To motivate the distinction between accuracy and skill, consider an example of a precipitation forecast generated over the Sahara Desert. It is highly likely that this would be an unremarkable forecast, with dry conditions predicted throughout the region. Upon verification, the forecast is found to be extremely accurate since no precipitation developed. However, it is difficult to justify running an expensive

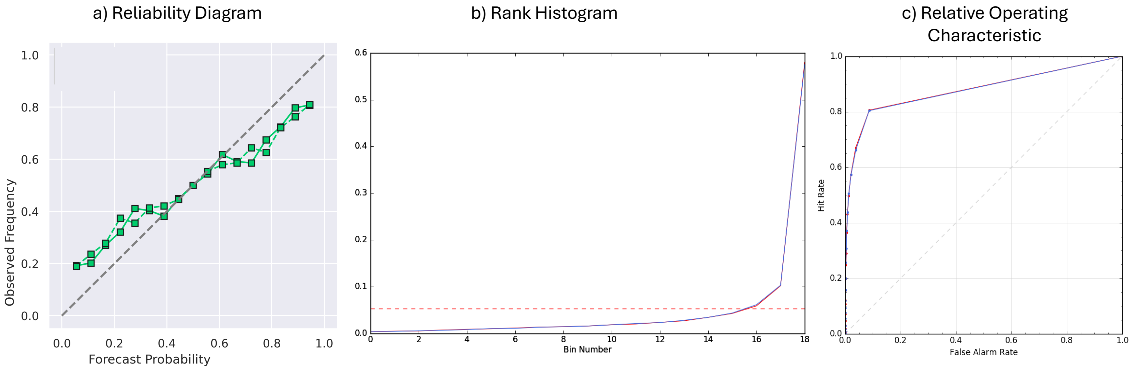

Figure 2.2: Examples of three metrics commonly used to assess ensemble performance. The reliability diagram in panel a) was calculated using cluster based metrics on the 90th centile of precipitation for two different models (see Section 5.4. The rank histogram in panel b) was calculated on hourly precipitation accumulations from 16 June to 07 July 2019 for two different MOGREPS-UK configurations. Red dashed line shows the ideal histogram. The ROC curve in panel c) was calculated on hourly accumulation exceeding 2 mm over the same period and using the same configurations as the rank histogram in panel b).

NWP model to predict an outcome that is essentially no different from climatology. Hence, while the accuracy of this forecast is large, running this forecast has not provided any meaningful information compared to our prior estimate, and its skill is therefore low. A common method for assessing skill is the Anomaly Correlation Coefficient, which estimates the correlation between forecast and observation after accounting for climatalogical trends (Murphy and Epstein 1989).

For ensembles, forecasters also expect that the probabilistic outputs are providing appropriate guidance, which is dictated by the ensemble spread. Probabilistic predictions first require the user to define an “event”, which can be as simple as defining the exceedance of a given threshold in a given parameter (e.g., temperatures above 20°C). The probability that an event occurs is then found by finding the fraction of ensemble members where this parameter threshold is reached or exceeded. There is the argument that the probabilistic prediction of a given event that is not completely certain one way or another (i.e., probabilities that aren’t 0% or 100%) can never be wrong, since the true outcome that the forecast is compared against can only ever occur once. This is a rather unhelpful perspective, however, as it offers a very limited framework for assessing probabilistic performance. Therefore, unlike skill-based metrics, probabilistic verification often requires us to evaluate ensemble performance over many forecasts rather than just a single forecast.

Ensemble reliability asserts that the predicted probability of a given event occurring should match the frequency with which that event verifies over multiple forecasts (Johnson and Bowler 2009). As an example, we might say that each time the forecast predicts a 30% chance of precipitation above 5 mm/hr over Reading, these predictions are reliable if this occurs approximately 30% of the time this prediction is made. The ensemble would instead be considered over- or under-confident if the verified frequency of occurrence is smaller or larger than the predicted frequency. This relationship can be plotted on a reliability diagram to understand the dependence on probability, with a perfectly reliable ensemble distributing scores along the 1:1 line. If the ensemble shows systematic unreliability, the predicted probabilities can be calibrated using simple statistical corrections before being used operationally. Figure 2.2a) shows an example of a reliability diagram, whereby the ensemble is slightly overconfident at lower probabilities and slightly underconfident at higher probabilities.

Along with reliability, another desirable property that ensembles should possess is that of "sharpness", or the ability to make confident predictions (Buizza et al. 1999). If we assume that climatalogical probabilities are approximately 50%, then an ensemble is sharp if it makes predictions that are close to 0% or 100%. Of course, these predictions must also be reliable for a sharp ensemble to be considered useful. An ensemble should also have good "resolution", which is the ability of the ensemble to make probabilistic predictions that differ from climatology. As with ensemble sharpness, this is subject to the constraint that these probabilities are also reliable.

Another approach to evaluate the suitability of ensemble spread is to expect that the verified outcome will lie within the range of values predicted by the ensemble. This property can be interrogated by binning and ranking the observation within the ensemble members and inspecting the distribution of the ranked verification over many forecasts (Candille et al. 2007; Hamill 2001). These rank histograms will be flat for a well-calibrated ensemble, indicating that the observation acts as if it was drawn as a random sample from the ensemble distribution. If instead the rank histogram displays non-uniformity, this indicates a deficiency within the ensemble. U-shaped histograms occur when the verification lies towards the extremes of the predicted values, suggesting the ensemble is underspread. Likewise, an n-shaped histogram occurs if the ensemble is overspread. A slope in the histogram indicates the presence of model bias. Figure 2.2b) shows an example of a rank histogram whereby the ensemble is noticeably underspread, and that neither of the model configurations include better spread than the other.

As well as visualising ensemble reliability, probabilistic verification scores can also be produced. The Brier Score (BS) acts like a probabilistic analogue of the Mean Square Error, which averages the mean square difference between the predicted probability (p) and observed outcome (o, either 0 or 1) of an event ($BS = (1/T)\Sigma_{t=1}^{T}\overline{(p_t - o_t)^2}$) over all forecasts evaluated (T), with values closer to 0 indicating good performance, and values closer to 1 indicating poor performance

(Brier 1950). The Brier Score can be decomposed into terms representing ensemble reliability, resolution and uncertainty in observation (Murphy 1973). The Brier Score can also be adjusted to account for a baseline reference to produce the Brier Skill Score, $BSS = 1 - (BS_{fcst}/BS_{ref})$, where scores of unity indicate perfect skill and scores of zero or less indicate forecasts that are no better or worse than the baseline (Weigel et al. 2007).

It is more common for ensemble evaluations to use the BSS over the BS due to the explicit comparison to the reference forecast. Of course, this strategy raises the question of the best reference forecast to use for normalisation. Ideally, reference forecasts would be constructed from the climatalogical average of the same ensemble that generated the verifying forecast. However, this approach is only appropriate if an extended archive of forecasts exists to construct this climatology. For complete consistency, this archive would also need to be regenerated after the implementation of model updates, at extensive computational expense. In principle, archives from other models could be used at the cost of this consistency. In general, the only requirement of the reference forecast is that it represents a model state possessing no skill. So, instead, it is most common to construct an in-sample climatology which uses data from the verifying forecast itself to construct the reference forecast. While this approach may sound somewhat tautological, it is the cheapest and most convenient method available. Such a state can be generated by comparing the observed outcome at each time (o_t) to the average outcome across all times ($\overline{o}$) as $BSS_{ref} = (1/T)\Sigma_{t=1}^{T}\overline{(\overline{o} - o_t)^2}$.

The BS and BSS requires the selection of an appropriate threshold of interest to define events. This choice of threshold may be entirely arbitrary, but can have a strong impact on the final score. For instance, the quality of the score can suffer severely if the chosen threshold corresponds to events that occur infrequently, where many more ensemble members would be needed to properly account for sampling uncertainty. The Ranked Probability Score removes this threshold dependence by calculating the BS for a set of discrete thresholds and averaging the result (Weigel et al. 2007). The Continuous Ranked Probability Score, CRPS, further generalises this method by replacing the discrete threshold series with a continuous series through the use of cumulative distribution functions (Hersbach 2000). Similar to the BSS, the Continuous Ranked Probability Skill Score can also be evaluated by normalising the forecast CRPS by the CRPS calculated for a low-skill reference forecast.

Another common method of evaluating model performance is the Relative Operating Characteristic (ROC), which plots the ensemble hit rate against the false alarm rate (Buizza et al. 1999; Mason and Graham 1999). The 1:1 line represents a forecast with no skill and just as many hits as false alarms. Ideal models would maximise the number of hits, resulting in more points that exist towards the top of the graph. The ROC performance is summarised by calculating the area under the curve, with larger areas indicating better performance. Figure 2.2c) shows an example of a ROC curve whereby the ensemble has a reasonably good hit-rate, but that neither of the model configurations shown perform better than the other.

Finally, the spread-skill relationship verifies the statement that ensemble spread can be considered a reliable estimate of its eventual skill (Buizza 1997; Houtekamer 1993; Whitaker and Loughe 1998). In this instance, spread is estimated by averaging the standard deviation between ensemble members across multiple forecasts. Spread is then compared to the average RMSE between each member and the verification over the same forecasts. We should expect the spread and skill to be approximately equal for a well calibrated ensemble. If the average standard deviation is smaller (larger) than the average RMSE, the ensemble is underspread (overspread).

The performance of any model depends solely on the choices made during development. As well as the ensemble initialisation and uncertainty quantification strategies, as previously discussed, the other two major factors are the model resolution and the number of ensemble members, each of which have a substantial impact on the quality of forecasts produced.

2.1.4 How Suboptimal is Less Than Infinity?

It has long been recognised that ensemble performance is strongly dependent on the nature of the event being evaluated. Climatologically common events are typically well represented in current ensembles, while rare, high impact events are more likely to suffer from sampling uncertainty (e.g., Craig et al. 2022; Leutbecher 2019; Mullen and Buizza 2002). In other words, more ensemble members are needed to properly represent the shape of the ensemble pdf, especially for non-Gaussian distributions with multiple modes or distorted tails. However, it is just as important for NWP systems to accurately represent these rare and potentially damaging events as it is for them to represent the day-to-day weather. An ideal ensemble that did not need to worry about uncertainties related to limited sampling would have the opportunity to achieve perfect reliability regardless of event extremity. But even running a few ensemble members at the same resolution and complexity of an equivalent deterministic system can be very computationally expensive. Meteorological services must therefore balance the costs associated with ensemble size, resolution and complexity when designing their systems. To this end, previous studies have been conducted to offer guidance about the effects of ensemble size on performance.

Early work assessing initial configurations of the ECMWF ensemble demonstrated that RPS, ROC and BSS scores saturate at approximately eight members (Buizza et al. 1998; Ebert 2001; Mullen and Buizza 2002). Eight members was also previously found to be the minimum number required to sufficiently approximate stochastic dynamic forecasts (Leith 1974). More recent work using higher-resolution ensembles found score saturation for 14 members, but that the presence of a "knee" in the score curves was also evident at around eight members (Marsigli et al. 2014). Clearly, then, there is reasonably strong agreement that an ensemble should not be run with fewer than eight members. One could also use these results to argue that running an ensemble with more than eight members would be wasted computational effort, given the apparently marginal score increases that would be obtained from adding more members. Most authors of the aforementioned studies do not make this argument, though, as they recognise that the impact of ensemble size depends strongly on the specific verification measure. For instance, Buizza et al. 1998 assert that the benefits of running 32 members on the spread-skill relationship and other metrics based on the "best member" is enough justification for the increase in computational cost, and even speculate that adding more members (perhaps over 100) would provide further benefits. But it has only been recently that investigations with large ensembles containing hundreds of members could be performed.

In the paper that motivated the title of this section, Leutbecher 2019 investigated the dependence of probabilistic skill scores with ensemble size for a 29 km grid spacing, 200 member ensemble. As with previous studies, score convergence was found to depend on the verification metric. The BS, RPS, and CRPS all converged with ensemble size (N) as $1 + N^{-1}$, as did scores calculated using median values. Scores calculated with larger percentiles converged much slower. Studies using a higher resolution 3 km grid spacing ensemble consisting of 1,000 members (Craig et al. 2022; Tempest et al. 2024) found mean, standard deviation, and sampling uncertainty convergence followed the $N^{-1/2}$ rate expected from the Central Limit Theorem (Fischer 2011). Once again, however, scores which sampled the tail ends of the pdf converged much slower, with as many as 100,000 members needed to achieve sampling uncertainty convergence for the 1st and 99th centiles (Tempest et al. 2023). Of course, these studies recognise that running an operational ensemble with 100,000 members is completely infeasible, and therefore also aimed to test methods that could compensate for the limited operational membership. For instance, in the previously mentioned study, Leutbecher 2019 demonstrates that an ensemble of only 4-8 members is sufficient for testing and designing a ≈ 50 member ensemble, provided additional statistical techniques are used to compensate for the small number of members.

The most popular method for artificially inflating ensemble sizes is "neighbourhooding", which treats values in the neighbourhood surrounding a particular grid point as additional samples for that grid point (Gilleland et al. 2009, 2010; Theis et al. 2005). In other words, if an event occurs within a specified vicinity of the central grid point, it is assumed that another realisation exists in which that event instead occurs at the central grid point and that this realisation has not been captured by the current model. Larger neighbourhoods provide larger pseudo-samples, but also weaken the assumption that all points in the neighbourhood are representative of the central point (particularly for regions with large topographic variability). Using surrounding grid points as a pseudo-ensemble also heavily depends on the homogeneity of the surrounding land cover; terrain-aware neighbourhood schemes are now used to improve pseudo-sampling in regions with variable orography (Roberts et al. 2023). While neighbourhooding was initially proposed as a means of producing probabilistic-style products from deterministic outputs (Theis et al. 2005), it has also demonstrated its effectiveness at compensating for limited ensemble sizes (Craig et al. 2022; Flack et al. 2021; Frogner et al. 2019a; Raynaud and Bouttier 2017; Schwartz et al. 2010; Tempest et al. 2024).

As well as neighbourhooding, so-called “fair” versions of the CRPS and BSS have been developed that are designed to estimate the scores that would be obtained if the verifying ensemble contained infinite members (Ferro 2014). Fair scores are maximised when, on average, the verification behaves as if it is drawn from the same distribution as the ensemble members. As long as the ensemble demonstrates exchangeability, fair scores can be considered to be independent of ensemble size (Ferro et al. 2008), and can provide a more appropriate verification method for ensembles with limited members.

At the time of writing, current operational ensembles use anywhere from 15 members at the Centre for Weather Forecasting and Climate Studies (CPTEC) (Cunningham et al. 2015), to 101 members at the ECMWF (ECMWF 2022) (which was upgraded in May 2023 from the 51 members in use since the 1990s). However, this section has conveniently ignored another large factor that can impact the choice of ensemble size. Lorenz 1969 showed that errors grow faster on smaller grids and it is therefore necessary to use different physics schemes, uncertainty representations and even verification metrics depending on the scales represented within the model. Indeed, it has long been recognised that model resolution represents a similar barrier to skill improvement as ensemble size, and that efforts to increase resolution can yield significant performance gains (Buizza et al. 1999; Raynaud and Bouttier 2017). This chapter has focussed on global ensembles with grid sizes that are too large to explicitly resolve convection. The next sections will describe the additional considerations needed to construct convective-scale models.

2.2 Convective-Scale Modelling

2.2.1 Explicitly Represented vs Parametrized Convection

The advances in computing power from the 1960s to the 1990s were largely used to improve the resolution and complexity of NWP models, and later to develop the first ensembles (e.g., Buizza 1997). From the turn of the 21st century, further advancements allowed researchers to experiment with models using grids that were fine enough to represent convection explicitly. Termed the "next generation of NWP" (Done et al. 2004) and a "step-change in rainfall forecasting" (Clark et al. 2016), these high-resolution models could now resolve the largest-scale processes involved in the formation of convective cells.

The general wisdom is that a model can be considered "convective-scale" if it has a grid size of 4 km or smaller (Done et al. 2004; Lean et al. 2008). At these scales, it is common to turn off the parametrization schemes that simply switch convection on or off in the relevant grid boxes. Instead, the model can explicitly develop convective structures and realistically advect these structures through subsequent timesteps (Lean et al. 2008). The representation of winter showers is a good example: in parametrized models, they only occur over the sea whereas in convection-permitting models they can advect inland (Lean et al. 2008). Additionally, representing convection explicitly provides more realistic predictions of mesoscale features like squall lines, sea-breeze fronts and mescoscale convective systems (Done et al. 2004; Lean et al. 2008; Speer and Leslie 2002). Figure 2.3 shows a comparison between a convective-scale and a convection-parametrized model output. There is an overall lack of less intense rain in the convection-permitting model but this is a known deficiency of the physics scheme used in this model (Bush et al. 2023). Both models place the precipitation associated with an advancing occlusion in approximately the same area over the south west of the UK and southern Wales, but there is considerably more detail in the convective-scale model. An organised band of pre-frontal convection is also easily identifiable in the convective-scale model, despite being somewhat displaced compared to the radar. The same general pattern is also present in the convection-parametrized output but the grid length is too coarse to distinguish any perceivable structure. However, in this instance, the convection-parametrized model has placed the convection more accurately compared to the convection-permitting model, despite the enhanced realism. This event will be discussed more thoroughly in section 4.8.2.

The benefits of running models at convection-permitting scales can only be realised if the model fields contain structures at convective scales. In other words, there is little use in running these high-resolution models if they cannot develop

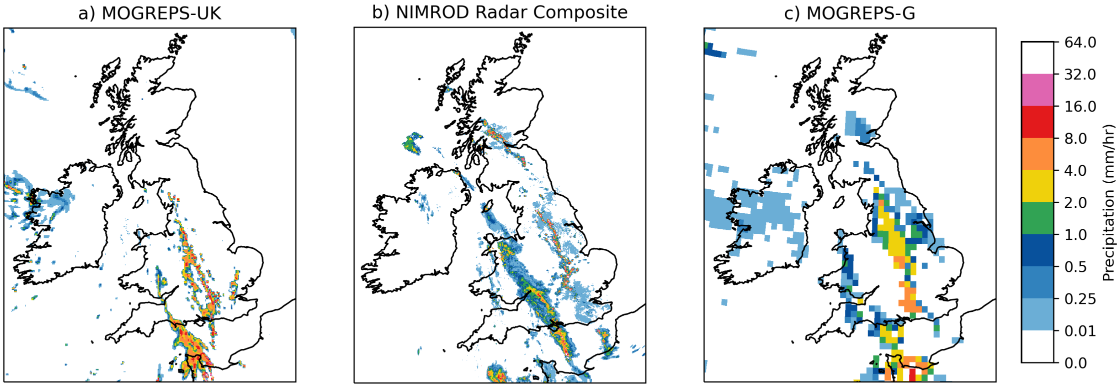

Figure 2.3: Comparison of a precipitation event between a given convective-scale MOGREPS-UK member, NIMROD radar and the same convection-parametrized MOGREPS-G member. An organised band of convection is advecting to the north east ahead of an approaching occlusion. This event occurred on 8 July 2023 at 0900Z, and is studied in more detail in section 4.8.2.

features at scales smaller than their global counterparts. Convective-scale detail can either be introduced during initialisation, or be left to develop. The most common way to introduce this detail is with a data-assimilation scheme that can produce convective-scale analyses (e.g., Dixon et al. 2009; Hagelin et al. 2017; Milan et al. 2020). However, such schemes are costly and time-consuming to develop, since they require extensive testing and knowledge of the error covariance characteristics. Recent techniques have been developed that can carry over the small-scale detail from previous model runs by selectively blending these scales with the larger scales updated via fresh analyses (Short and Petch 2022). This "warm starting" technique ameliorates issues at short leadtimes when convective-scale detail is not present. In contrast, so-called "cold starting" models are initialised as a downscaler of the low-resolution analyses (i.e., the low-resolution data is simply projected onto the high-resolution grid), and smaller-scale circulations are left to spin up naturally (e.g., through interactions with the surface) (Lean et al. 2008). While this is the easiest method, and sometimes the only option available, precipitation fields present undesirable behaviour during this spin-up period. It takes finite time for convective-scale structure to develop explicitly, during which an absence of precipitation may be noticed. Then, the model may develop convective structure by producing too much rain in a short timeframe as it releases its instability. Hence, an overshoot in precipitation totals is often observed before relaxation to a more consistent state (Lean et al. 2008). This spin-up contamination can take multiple simulation hours to complete, and precipitation outputs are generally disregarded until this occurs.

The other choice to make when developing convective-scale models concerns the domain. Given the time available for operational forecasting, it is generally too expensive to run models at this level of detail covering the entire Earth. Since numerical stability is governed by the Courant-Friedrichs-Lewy criteria, any reduction in grid size must also be accompanied by a reduction in the timestep of the simulation (Courant et al. 1928). For example, a halving of the grid size in each spatial dimension would require a timestep that is reduced by at least eight times to maintain stability, which adds a considerable amount of additional processing on top of that required for using the finer grid. Therefore, various strategies have been developed to save computational costs when running at convection-permitting scales. The most common method is to construct a limited area model nested within a lower-resolution model that provides lateral boundary conditions (LBCs). For convective-scale ensembles, these models may also provide initial condition perturbations to ensure consistency with the LBCs. In this configuration, it is also common to surround the high-resolution grid with a coarser-resolution "sponge" region, giving time for convective detail to spin-up in the boundary information before it is advected into the area of interest (Hagelin et al. 2017; Hanley and Lean 2024; Milan et al. 2020). Another option for running convective-scale models is to use a full global model that tapers to the required grid size over the desired region. The Model for Prediction Across Scales (MPAS) is one such model, which uses an unstructured Voronoi mesh to achieve this grid flexibility (Heinzeller et al. 2016). The key advantage of using a model like MPAS over the conceptually simpler nesting approach is the representation of a broader range of scales. Limited area models can only adequately represent processes with scales as large as the domain size before aliasing effects reduce accuracy (Guidard and Fischer 2008; Milan et al. 2023). By using a single model covering the Earth, MPAS can more accurately capture interactions between scales, which is particularly useful over regions where small-scale processes can potentially feedback onto the synoptic-scale dynamics (e.g., Fowler et al. 2020; Núñez Ocasio and Rios-Berrios 2023; Schwartz 2019).

2.2.2 Model Improvements from the Explicit Representation of Convection

It is now well established that running models at convection-permitting scales provides a broad range of benefits over coarser models. Higher-resolution models can utilise higher-resolution inputs, from observations for use in a convective-scale data-assimilation schemes (Milan et al. 2020), to more detailed land cover and orography maps (Barrett et al. 2016). The interaction of synoptic-scale features with these detailed land cover datasets, together with the associated smaller-scale flows that develop, can also produce more realistic representations of larger-scale patterns than are present in coarser models (Barrett et al. 2015; Gowan et al. 2018; Schellander-Gorgas et al. 2017).

Convection-permitting models also improve the representation of extreme precipitation. This improvement can be expected simply from a geometric standpoint since each grid box represents a smaller area than in coarser models. Convective parametrizations are designed to represent the area-average precipitation rate (Arakawa and Schubert 1974), which smooths extremes. For convection-permitting models with smaller grid sizes, extreme values are more likely to be retained. From a meteorological standpoint, parametrizations must attempt to capture the processes involved in convection that occur across many scales of motion (Wilson and Ballard 1999), and while they are reasonably useful at identifying the broad regions of instability from synoptic conditions, they often struggle to produce realistic distributions. Convective-scale models are more successful at producing heavier precipitation (Clark et al. 2009; Marsigli et al. 2005; Schellander-Gorgas et al. 2017; Schwartz et al. 2010), as well as placing it in the correct locations (Duc et al. 2013; Frogner et al. 2019a; Marsigli et al. 2008).

Of course, there are different processes that generate extreme precipitation, but we expect that running models at convective scales will have a beneficial impact on most of these processes. For instance, convection-permitting models improve orographically enhanced precipitation through the use of more detailed surface maps and the resolution of smaller-scale flows (Barrett et al. 2015; Gowan et al. 2018; Hanley et al. 2011; Schellander-Gorgas et al. 2017). Convection-permitting models also improve the diurnal cycle of precipitation in tropical regions, which can cause intense flash flooding events (Glazer et al. 2025; Woodhams et al. 2018). In parametrized models, convection tends to initiate too early since schemes will usually act to trigger convection as soon as atmospheric instability is present (Clark et al. 2016; Lean et al. 2008). Explicitly-resolved convection better represents the physical processes and timescales that lead to the development of diurnal convective storms (Clark et al. 2009; Woodhams et al. 2018). For instance, in a recent study of the climatology over Lake Victoria, intense nighttime thunderstorms were linked to the emergence of cold pools over the basin, which require grid sizes of the order 3 km to resolve (Glazer et al. 2025).

Other studies have also focussed on the improved representation of specific events within convection-permitting models:

- Tornado forecasts were found to be improved when updraught helicity fields produced from convection-permitting models were combined with other environmental information important for tornadogenesis (Gallo et al. 2016.
- Squall lines (Hanley et al. 2013), stationary convective bands (Barrett et al. 2016), and flash flooding events (Frogner et al. 2019a; Roberts et al. 2009) were better represented in convection-permitting models, but their performance was

shown to be more dependent on the accuracy of the upstream synoptic conditions provided by the parent model.

- Mesoscale convective systems are better represented since convection-permitting models can do a reasonable job of organising individual cells into coherent structures (Clark et al. 2009; Clark et al. 2016).
- Sting jets (localised and transient regions of intense wind gusts occasionally found at the cloud hook of extratropical cyclones) typically require good vertical resolution to accurately capture the dynamics of the descending airstream (Manning et al. 2022; Martínez-Alvarado et al. 2010) although can be diagnosed from coarser models (Gray et al. 2021).
- Greater vertical resolution also helps to forecast the transition between rain and snow or freezing rain. More vertical levels helps to accurately predict the latent heat release and associated in-situ atmospheric cooling from melting snow. The exact type and composition of hydrometeors determines associated hazards, but is very sensitive to atmospheric conditions (Young and Grahame 2024b).
- Finally, convection-permitting models have also demonstrated their benefit in correctly predicting sea breezes compared not just to a coarser model, but to a statistical model trained on both high- and low-resolution outputs (Cafaro et al. 2019). This study is especially illuminating as it reveals that the explicit representation of the mesoscale interactions necessary to produce sea breezes cannot reliably be inferred from coarser models, even using learned statistical relationships.

While these studies have demonstrated the benefits of running models with explicitly resolved convection, it is also recognised that km-scale grids are still too coarse to resolve all of the important motions involved. A 4 km model, for instance, will be able to capture larger storms but not any accompanying smaller showers. This limitation necessitates only a partial removal of convective parametrization schemes for these coarser models, such that the larger scales are left to evolve explicitly while the smaller scales must still be parametrized (Arakawa and Wu 2013; Gerard et al. 2009; Yu and Lee 2010). This complication is known as the convective gray zone problem. However, even the explicitly resolved components of convection are not always well represented in coarse convective-scale models. At 4 km grid spacing, convection is still underresolved (Petch 2006), which produces too few showers that are too large, too intense (Schellander-Gorgas et al. 2017), and that are initialised too late (Lean et al. 2008). The move to finer grids in recent years partially alleviates these problem by allowing the models to explicitly represent a greater range of scales (Lean et al. 2008). However, even these km-scale models cannot represent the smallest-scale turbulent processes that occur within clouds, meaning that some biases with cell size, intensity, and organisation remain (Hanley et al. 2015; Stein et al. 2015b). Experiments with models at hectometre scales have shown promise at correcting these biases and producing more detailed fields (Lean et al. 2024).

2.2.3 Scale-Dependent Error Growth

It is not immediately obvious that the growth of errors in NWP models should be tied to the length scales of the processes involved. The first paper to make this connection was Lorenz 1969, that examined the effect of perturbations applied at two different length scales. Lorenz shows that errors originating at smaller scales grow much faster in amplitude and scale than those originating at larger scales. He argues that the time horizon for accurate forecasts is unavoidably limited by the faster growth of these smaller errors. In other words, even if researchers could reproduce the current atmospheric state perfectly down to a particular scale, the remaining uncertainties in smaller scales would negate these benefits through the fast growth of errors in time and amplitude.

Durran and Weyn 2016 contest the severity of the conclusions from Lorenz 1969, and assert that since perfect initial conditions on any scale aren't possible, relatively minor errors at larger spatial scales can be just as important as relatively large errors at smaller scales. Their experiments reproduce those of Lorenz 1969, albeit with a more advanced model than the original, and the results are reproduced here in Figure 2.4. These findings show that large-scale errors of the same amplitude as small-scale errors induce qualitatively similar growth patterns. Initially, the large-scale error rapidly downscales, but then grows from this smaller scale in a similar manner as if the error had originated from that smaller scale. Then, errors in both the large- and small-scale perturbation experiments saturate at similar scales after the first day. Despite both perturbations having the same initial amplitude, when compared to the background mean, the smaller-scale error is initially fully saturated, while the larger-scale error is only a trivial fraction of its maximum saturation amplitude. The authors use this result to argue that Lorenz 1969 was somewhat misguided in their assertion that small-scale imperfections are the largest barrier to forecast predictability, since the only situation where this would be important is if the *relative* size of the large-scale error is unattainably small. The authors then use this logic to demonstrate that for forecasts of thunderstorms and squall lines, meteorological services should attempt to improve the *larger-scale* initial conditions, which are easier to measure than the smaller scales but can introduce similarly damaging

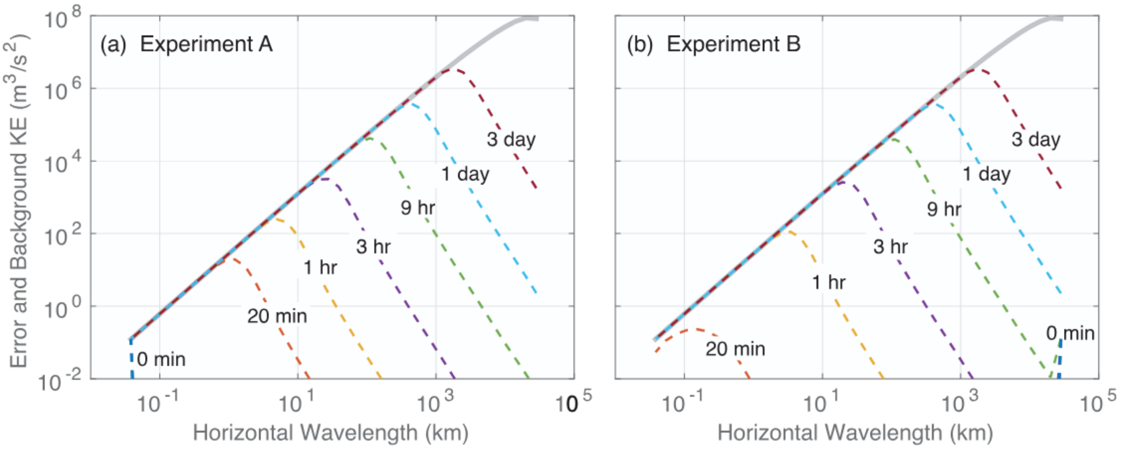

Figure 2.4: Lorenz model experiment showing error growth characteristics for an initial small-scale error (Experiment A) and an initial large-scale error with the same amplitude (Experiment B). Solid grey lines show background kinetic energy of ensemble mean. Figure reproduced from Durran and Weyn 2016.

errors.

Understanding error growth characteristics is not only important for targeting improvements to initial conditions, it is also necessary for understanding the processes that contribute the most to forecast uncertainty. Hohenegger and Schar 2007 demonstrated that error growth rates for convective systems were ten times faster than for synoptic-scale systems. Zhang et al. 2007 built upon this idea to describe a three-stage error growth mechanism explaining convective-scale error growth. During stage 1, errors at convective scales grow within the region of convection, as uncertainty in moist processes drives large disparities between model runs. As such, faster error growths occur when instability is larger. This error growth quickly saturates as the errors reach amplitudes similar to the background mean state (see Fig 1 of Durran and Weyn 2016). Then, during stage 2, the error growth transitions from a purely convective-scale error into an error that starts to feed back onto the synoptic scales. Not all of this error is retained, however, as some of it is dispersed through imbalanced motions such as gravity wave activity. During stage 3, the error retained in the balanced component of the flow grows through the expected baroclinic instability mechanism, whereby potential vorticity anomalies at different vertical levels interact to amplify perturbations (Eady 1949). This three-stage growth mechanism has been observed in multiple studies (Selz and Craig 2015; Surcel et al. 2015; Zhang et al. 2007).

There is also evidence to suggest that the impact of small-scale perturbations depends on the convective regime (Done et al. 2006). Here, convective regime refers to the scale of the forcing that initiates convective activity. Under strongly-forced convection, the generation of convective-available potential energy (CAPE) is in quasi-equilibrium with its release, such that the synoptic conditions responsible for the conditional instability are the dominant factor controlling convection. In contrast, weakly-forced convection is characterised by a large build up of CAPE due to the presence of a stable layer further aloft, such that this instability cannot easily be released. Instead, turbulent interactions with surface features induce smaller-scale CAPE variability, until a particular region develops enough CAPE to erode the stable layer and release the built-up instability. Due to this dependence on local factors, weakly-forced convection is associated with much larger uncertainty than strongly-forced convection (Keil and Craig 2011; Keil et al. 2014; Tempest et al. 2024). Climatalogical studies have shown that approximately 85% of convective events are strongly forced over the UK (Flack et al. 2016), while 66% of events are strongly forced over Germany (Zimmer et al. 2011).

Given the challenges associated with accurately predicting the development and evolution of small-scale features like convection, it is clear that deterministic approaches can only offer so much. Many studies have shown that convection-permitting models can certainly produce more realistic outputs and provide far better guidance than coarser models. But, as demonstrated in this section (particularly Fig 2.3), this realism does not necessarily translate to *accuracy*. Indeed, we have moved from an era where we should not interpret individual precipitation forecasts too literally because the resolution is too low, to an era where we should still not interpret these forecasts too literally despite the enhanced realism. It is now common, therefore, for met services to run convection-permitting ensembles that are designed to target different predictability limitations to global ensembles and use different tools to characterise this uncertainty.

2.3 Convection-Permitting Ensembles

2.3.1 Configuring Convection-Permitting Ensembles

In recent years, more meteorological services have begun implementing convection-permitting ensembles (CPEs) designed to quantify the uncertainty of small-scale processes at shorter time frames. Given the computational costs of running even a single model at convection-permitting scales, it is common to nest each CPE member inside a corresponding driving member of the global ensemble (i.e., the driving member provides LBCs to the nested member). These nested members may run with a slight reduction in resolution and/or model complexity compared to similar deterministic models to further save costs. It is also common for each global member to provide initial condition perturbations to the corresponding CPE member, although this is not the only strategy available. For instance, each member of the ICON-D2 model operated by Deutscher Wetterdienst is initialised directly from a local DA scheme and does not use information from the global ensemble (Reinert et al. 2025). This approach also ensures that each member is initialised with high-resolution detail such that the ensemble outputs will not suffer from spin-up issues and will be useful as soon as the ensemble begins running. Other centres instead introduce this high-resolution detail by blending the large-scale perturbations with convective-scale analyses (Ono et al. 2021; Porson et al. 2019; Seity et al. 2011).

Another difference between global and driving ensembles is the number of members needed. Given the shorter predictability limits discussed in the previous section, it is implied that more members would be needed to adequately sample the uncertainty of these processes. And yet, since each member of a nested CPE must be paired with a corresponding driving member providing LBCs, the maximum number of CPE members is limited by the global ensemble it is nested within. Furthermore, the computational costs of running CPEs are large, even within limited-area configurations. As such, some centres have run configurations that select a limited number of global members that are expected to have the most diversity for use in driving CPEs (Montani et al. 2011; Weidle et al. 2013) These representative members are selected using clustering algorithms that group the ensemble into distinct subsets, and then choose the members that are most central within those subsets. The computational savings from running fewer CPE members could then allow developers to run those members at an increased resolution. This trade-off between resolution and ensemble membership (plus other model complexities like physics and DA schemes) is an important choice to make, and depends on the aims of the user. For instance, a study of the AROME CPE found that running at a higher resolution produced better probabilistic outputs at shorter leadtimes, while running more members at a lower resolution was beneficial at longer leadtimes when there was larger forecast uncertainty (Raynaud and Bouttier 2017). This result is supported by Clark et al. 2011 which demonstrated that, at longer leadtimes, more members were needed to reach an asymptote in ROC scores than at shorter leadtimes.

The use of a model uncertainty scheme is another choice that developers can make when designing CPEs. These schemes may be adapted from similar schemes used in global ensembles. For instance, some studies have tested SPPT (Bouttier et al. 2012; Romine et al. 2014), SKEB (Romine et al. 2014) and other stochastic physics schemes (Baker et al. 2014; McCabe et al. 2016). These studies typically increase ensemble spread, but can also introduce biases. CPEs may also employ additional schemes that are either not present in global ensembles or have been heavily refined. For instance, stochastic boundary layer schemes aim to represent the uncertainty of processes that are still unresolved at convective-scales, such as the motions of turbulent eddies (Flack et al. 2021; Kober and Craig 2016; Leoncini et al. 2013). These schemes can improve the timing of convection by capturing the effects of the initial, smaller-scale flows that are thought to be important for accurate initiation (Lean et al. 2008).

In general, there are few concrete rules for deciding the model configuration, and developers often need to perform extensive package testing to decide on the combination of resolution, members, and model uncertainty schemes that best suit their needs. Even the methods of testing these different configurations require their own strategies and decisions. For instance, is it better to test a particular configuration on a limited set of case studies with full complexity, or over a longer period using reduced complexity? The case study approach may give an early indication of potential problems (and benefits) with the configuration change, but cannot be used to assess the broader aspects of ensemble performance like reliability. These evaluations instead require an extended period of running for proper assessment. Therefore, once a configuration update has been decided, it is common to run both new and old versions in parallel for a time. This allows developers to properly assess updates and identify issues that were not previously known. At the UK Met Office, a parallel suite of improvements runs alongside the operational suite used for delivering products. After a year of trials, the parallel suite becomes operational and development on the next set of updates begins (Aurore Porson, Pers. comm.).

2.3.2 Operational Convection-Permitting Ensembles Around the World

Table 2.1 lists the CPEs that are currently in operation in meteorological centres around the world. Each ensemble covers the respective country. In this subsection, these CPEs are briefly discussed in turn.

In the UK, the Met Office (UKMO) Global and Regional Ensemble Prediction System (MOGREPS) consists of a global (MOGREPS-G) and convection-permitting (MOGREPS-UK) ensemble. The MOGREPS-G model runs four times per day with 18 members, and has grid spacing of approximately 20 km over the UK. These members are used to drive an equal number of 2.2 km grid spacing MOGREPS-UK members but on a different schedule to that of the global ensemble. Rather than initialising each member every six hours, MOGREPS-UK employs a "time-lagged" approach, whereby three members are initialised every hour and the full 18-member set is constructed by combining these members with those from the previous five cycles (Porson et al. 2020). This initialisation strategy improves the spread of short-range convective events, since each cycle of three members is initialised using the latest convective-scale analysis, produced hourly. Introducing a greater diversity of starting states improves the spatial spread of convective initiation and produces better probabilistic guidance (Porson et al. 2020; Raynaud and Bouttier 2017). Within the next couple of years, the UKMO will retire its deterministic systems and transition to an ensembles-only approach. This transition will be accompanied by a resolution increase for both the global ensemble and the CPE to match the resolutions of the retiring deterministic models (grid spacings of 10 km and 1.5 km respectively). Following these change, the Unified Model that powers these ensembles will be replaced by the LFRic model (named after Lewis Fry Richardson) which, among other changes, replaces the regular latitude-longitude grid with a cubed sphere arrangement (Adams et al. 2019). This new grid alleviates problems associated with gridlines converging at the poles.

Météo France run the Applications of Research to Operations at Mesoscale (AROME) CPE over France (Brousseau et al. 2016; Seity et al. 2011), but also provide the foundation for many other limited-area models run in countries throughout Europe as part of the High Resolution Limited Area Model (HIRLAM) and Aire Limitée Adaptation Dynamique Développment International (ALADIN) schemes (Bengtsson et al. 2017; Frogner et al. 2019b). These models fall under the "HARMONIE-AROME" (and, for ensembles in particular, the "HARMON-EPS") umbrella to distinguish them from the AROME model run over France. These models can be separately configured by each meteorological agency by are all driven by the ARPEGE global ensemble developed by Météo France and ECMWF. A list of HARMON-EPS models is presented in Table 2 of Frogner et al. 2019b.

The Icosahedral Nonhydrostatic (ICON) model is run by Deutscher Wetterdienst (DWD) in Germany and MeteoSwiss in Switzerland (Reinert et al. 2025). ICON uses a cubed-triangular grid to eliminate polar singularities. The ICON-D2 CPE is a separate configuration from the ICON-EU model that provides LBCs, and employs its own 4D-LETKF DA scheme to initialise its members. The ICON framework recently replaced the GME global ensemble, the COSMO-EU European nested ensemble, and the COSMO-D2 Germany/Switzerland nested ensemble. As well as ICON-D2, which cycles every three hours out to a 48 h leadtime, DWD have also recently introduced the rapidly updating ICON-D2-RUC which cycles every hour out to 14 h leadtime (Reinert et al. 2025). The more frequent cycles allows the model to ingest observations sooner into the NWP process.

The 5 km ensemble run at the Japanese Meteorological Agency is technically not a CPE given the previous definitions of "convective-scale", but does run over a limited region cover Japan. Initial perturbations are derived from a combination of singular vectors from 40 km and 80 km mesoscale models, as well as a global spectral model with 63 wavenumbers and 40 vertical levels (T63L40) (Ono et al. 2021). These perturbations are then interpolated to the 40 km grid before being applied to the limited area ensemble. The T63L40 model also provides LBCs to each member of the limited area ensemble. The Korean Meteorological Agency (KMA) runs a 2.2 km ensemble that is cold-started from the MOGREPS-G ensemble (Kim et al. 2015). Recently, however, the KMA has transitioned away from using MOGREPS-G and has instead developed their own Korean Integrated Model (KIM). The CPE nested in KIM now runs to 120 h leadtime rather than 72 h as previously, although it does also run with a slightly reduced resolution of 3 km (Shim et al. 2025).

The Naval Research Laboratory (NRL) Coupled Ocean-Atmosphere Mesoscale Prediction System-Tropical Cyclones (COAMPS-TC) is not a traditional CPE in that it does not run at fixed times or over fixed locations. Instead, COAMPS-TC is used to track the development of tropical storms, and the convection-permitting 4 km grid follows the tracks of those storms (Komaromi et al. 2021). Running this ensemble at convective-scales improves the accuracy of storm tracks and structures compared to global ensembles, and is especially important for providing accurate predictions of storm intensities (Komaromi et al. 2021). Since COAMPS-TC is designed for a different purpose than most other CPEs, the model uncertainty scheme is also slightly different, and perturbs the drag coefficient at high winds in the momentum exchange parametrization scheme. Each member is also initialised with a perturbed initial vortex strength.

While there appears to be no routinely running CPE over the USA, plenty of research has been done on running experimental models on a more ad-hoc basis (Schwartz 2019; Schwartz et al. 2015, 2019). CPEs are also initialised over regions of the USA that are expected to experience severe weather. There is also the High Resolution Ensemble Forecast (HREF) which, despite the name, is actually just a multi-model collection of deterministic forecasts (Roberts et al. 2020). These setups are often called "poor-man's ensembles" since they do not require any additional numerical processing, just the post-processing needed to standardise each of the individual outputs. Poor-man's ensembles are also not generally considered to be equiprobable. However, in lieu of an actual CPE, these ensembles can still provide forecasters with reasonable probabilistic guidance.

As processing power increases and the price-per-flop gets cheaper, it is expected that more meteorological services will consider running CPEs, while those that already run CPEs will continue to introduce improvements. In particular, there is a general need across all CPEs to improve spread.

Centre & Ensemble	Grid Spacing	Members	Forecast Range	Initial Conditions	Boundary Conditions	Model Uncertainty
UK Met Office (MOGREPS, Porson et al. 2020)	2.2 km	18 time-lagged (3 per hour)	120 h	4D-Var DA + global ensemble perts	Global ensemble (MOGREPS-G 20 km)	Stochastic physics (RP2)
Météo France (AROME, Brousseau et al. 2016), (HARMON-EPS Frogner et al. 2019b)	1.3 km	25	51 h	3D-EnVar DA + 3.2 km ensemble perts	Selected from global ensemble (ARPEGE 5-24 km)	SPPT
Deutscher Wetterdienst (DWD), Germany (ICON-D2, Reinert et al. 2025)	2.1 km	20	48 h	4D Local Ensemble Transform Kalman Filter	Europe ensemble (ICON-EU 13 km)	Random parameter perts
Japan Meteorological Agency (JMA), (ASUCA MEPS, Ono et al. 2021)	5 km	21	39 h	4D-Var DA + 40 km global ensemble perts	Global ensemble (JMA)	SPPT
Korean Meteorological Agency (Kim et al. 2015)	2.2 km	13	72 h	Cold-started from global ensemble	Global ensemble (MOGREPS-G 20 km)	Random Parameters
Naval Research Laboratory (NRL), USA (COAMPS-TC, Komaromi et al. 2021)	4 km	11	120 h	Cold-started from 36 km global ensemble	36 km global ensemble	Perturbed drag coefficients

Table 2.1: Operational Convection-Permitting Ensembles in use at meteorological centres around the world. References for each model are based on the latest known publications, but in some cases do not reflect the current configurations that are described in the table.

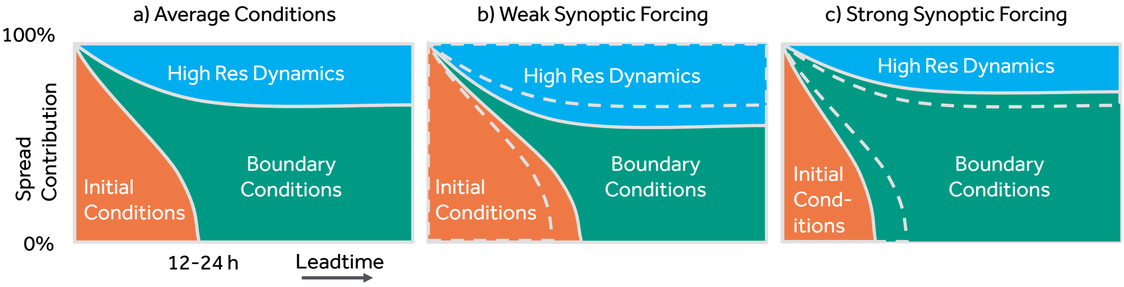

Figure 2.5: Schematic showing attribution of spread by leadtime and regime in a nested CPE. Dashed lines in panels b) and c) show the reference partitions from panel a).

2.3.3 Sources of Convective-Permitting Ensemble Spread

Spread is the most important factor that determines the value of running a CPE compared to a deterministic alternative. Fortunately, along with the improvements to skill that comes with running at the convective scale (see Section 2.2.2), studies have also shown that CPEs improve probabilistic guidance compared to global ensembles (e.g., Cafaro et al. 2019; Clark et al. 2009; Frogner et al. 2019a; Marsigli et al. 2005; Marsigli et al. 2008) which can partly be attributed to increased spread. Both global ensembles and CPEs are typically underspread (see Section 2.3.4) so any increase in spread will improve probabilistic outputs (provided skill is unchanged). Given equivalent conditions, we should expect a priori that CPEs will have more spread than global ensembles due to the representation of smaller-scale flows and use of higher-resolution surface maps driving faster error growth. However, the magnitude of these improvements is highly dependent on the conditions being predicted due to the different drivers of CPE spread.

To understand the mechanisms that drive CPE spread, Fig. 2.5 shows a schematic of the three main contributors as a function of leadtime and regime, developed using insights gained from completing this PhD. Note that this schematic shows the *contribution* to the total spread, rather than the magnitude of spread, which will typically increase with leadtime. Additionally, this figure is mostly illustrative and is only partly informed by data (though the results in section 4.5 could help solidify these relationships). Figure 2.5a) shows the partitioning during typical conditions. At the start of the forecast, the only contributor to spread will be differences in the initial conditions. Then, as each member is integrated forwards in time, these initial conditions will be advected out of the domain and replaced by information arriving from the global ensemble via the boundaries. At the same time, internal model processes will spin-up additional features from the initial and boundary information which are likely to further contribute to differences between members. Eventually, the initial information will have advected out of the domain entirely, leaving just the partition between boundary conditions and model dynamics. The time taken for this to occur depends on the size of the model and the strength of the prevailing winds, but from previous studies this can occur from 12–24 h (Gebhardt et al. 2011; Hohenegger et al. 2008; Kühnlein et al. 2014; Porson et al. 2019; Vié et al. 2011; Zhang et al. 2023).

Figures 2.5 b) and c) show the changes that would be expected under slack and mobile cases, or strong and weak forcing. Under slack conditions, we would expect weaker wind strength and therefore more persistent initial conditions. The lighter winds will also give the model more time to spin-up small scale features, thereby increasing the contribution that internal dynamics makes to the overall spread. Hence, under convective setups, we would expect the evolution of each CPE member to diverge more rapidly from that of the corresponding driving member compared to typical conditions. Conversely, during more strongly-forced mobile regimes, the boundary conditions become even more dominant than average, reducing the opportunity for the high-resolution dynamics to spin-up meaningful differences compared to the driving ensemble.

This picture of CPE spread growth is supported by previous findings. For instance, Clark et al. 2010 examined the relative influence of LBC/IC spread and model-physics spread both within and between convection-permitting and convection-parametrized ensembles. They found that model physics schemes had a stronger impact on the spread of low-level fields such as 2m temperature and three-hourly precipitation than upper-level "mass" fields which were dominated by advection. The contribution of model physics to the overall spread increased with leadtime (out to T+33 h) but was always much smaller than the spread provided by LBCs and ICs. Similarly, Flack et al. 2018 and Flack et al. 2021 demonstrated the importance of boundary layer perturbations for initiating and developing convection, particularly in cases of nonequilibrium (i.e., weakly-forced) convection.

The differences between the expected sources of spread can also produce differences in the spread magnitude. Frogner et al. 2019a conducted an in-depth study into the added value provided by a CPE compared to a global driving ensemble and compared to its own control member (a stand-in for a deterministic system). While the CPE was more useful in all situations tested, they found that the value added by running a CPE also depends on the situation being predicted. The authors highlighted this relationship by showing that the CPE provided more useful probabilistic outputs when predicting flash flooding caused by a Mesoscale Convective System (MCS), and provided slightly less useful outputs when predicting the impacts of a named storm. The authors then extended these findings to show that the CPE provided more value during summer (when weakly-forced regimes are more common) and during shorter leadtimes (when the small-scale information in the ensemble has been derived from convective-scale DA, rather than spun-up from the larger-scale conditions common to both ensembles).

However, the typical dominance of the boundary conditions in determining CPE spread underscores the importance of the synoptic information. Consider, for example, an instance where the driving ensemble misplaces a region of instability. The CPE may then develop this instability in an entirely accurate way for the synoptic conditions it was provided, yet the overall verification will still depend entirely on the magnitude of the initial positioning error. The dependence of smaller-scale processes on synoptic accuracy has been well documented in previous studies analysing sensitivity in squall lines (Hanley et al. 2013), stationary convective bands (Barrett et al. 2015, 2016) and MCSs (Frogner et al. 2019a). These studies suggest that improvements to the synoptic scales entering CPEs (either by correcting those scales in the CPE itself, or by correcting the driving ensemble) can also provide substantial downscale improvements. Hence the motivation of this thesis to focus on the understanding the impacts of these synoptic scales on CPE spread across a range of scales, leadtimes and situations.

Overall, probabilistic guidance produced by CPEs consistently outperforms that produced by driving ensembles due to the more accurate representation of small-scale extreme events and the additional spread produced by running at higher resolutions (Cafaro et al. 2019; Clark et al. 2009; Frogner et al. 2019a; Marsigli et al. 2005; Marsigli et al. 2008). Importantly, it has also been demonstrated that the additional detail generated by CPEs cannot be replicated using statistical models trained on driving-nested member pairs, indicating the value of operating at the convective-scale (Cafaro et al. 2019). However, many studies analysing CPE and global ensemble spread also report issues with the spread-skill relationships, finding that they are often underspread.

2.3.4 Convection-Permitting Ensemble Evaluation and The Underspread Problem

Lack of spread between members is a systemic issue facing operational and research CPEs. Underspread ensembles are more likely to provide overconfident predictions which has a deleterious impact on probabilistic metrics like reliability and Brier Scores. Given the desirable properties of ensembles outlined in section 2.1.3, this problem is reflected in a broad range of CPE performance aspects. The spread-skill relationship, for instance, provides the most obvious indicator of the underspread problem. On average, CPE spread (measured using the standard deviation or the variance between each member at each grid point) should be a reliable estimate of the eventual skill (measured using the RMSE or MSE between each member and verification at each grid point). Many studies have reported that average spread scores are smaller than average skill scores for parameters such as 2 m temperature (Beck et al. 2016; Frogner et al. 2019a; Klasa et al. 2018), 2 m relative humidity (Beck et al. 2016; Klasa et al. 2018), wind speed (Beck et al. 2016; Klasa et al. 2018), mean sea-level pressure (Frogner et al. 2019a), and precipitation accumulated over hourly (Clark et al. 2009; Schwartz et al. 2014), three-hourly (Clark et al. 2009), six-hourly (Beck et al. 2016; Clark et al. 2009, 2011; Raynaud and Bouttier 2017), and 12-hourly periods (Frogner et al. 2019a,b). Additionally, rank histogram analysis shows that the verification often falls outside the range of values predicted by the CPE (Beck et al. 2016; Clark et al. 2011; Vié et al. 2011).

However, it is important to acknowledge a particular limitation in some of the aforementioned studies. Performing spread and skill comparisons at each grid point is the traditional method of evaluating forecasts on coarser grids, and while this can also be appropriate for CPEs, it is an overly punitive approach when validating fields possessing large spatial gradients or noisy characteristics. Consider an example of a small area of precipitation that is misplaced in a forecast such that there is no overlap between the predicted and verified coverage. If this displacement is small and there are no other external factors that should preferentially influence the placement of this precipitation, a user performing a subjective evaluation would likely score the forecast reasonably highly. We should then expect that a useful objective metric will reproduce the broad trends of the subjective evaluation. However, instead, an objective gridscale metric would doubly punish this forecast for having both a forecast bust (missing the verified area of precipitation) and a false alarm (predicting precipitation in an area that did not occur) (Gilleland et al. 2009). This is known as the double penalty problem, and is most problematic when comparing or verifying high-resolution fields possessing large spatial gradients such as precipitation, cloud cover, or fog.

A number of techniques have been developed to mitigate the impacts of the double penalty problem. These techniques are well summarised in Gilleland et al. 2009, and the two most popular ones will be discussed here. One approach is to identify features of interest in each field under consideration and then match features to compare size and displacement. The Method for Object-based Diagnostic Evaluation (MODE) is a popular tool for feature-based evaluations (Davis et al. 2009). However, there is considerable subjectivity as to the best techniques for identifying and matching features, the choices of which are likely to impact the resulting analysis (Gilleland et al. 2009). Another popular technique is the SAL (structure, amplitude, location) method, which measures differences in the size and shape (structure), area-averaged precipitation (amplitude) and centre of mass (location) between objects in two fields (Wernli et al. 2008). A score is assigned to each of these aspects, with an

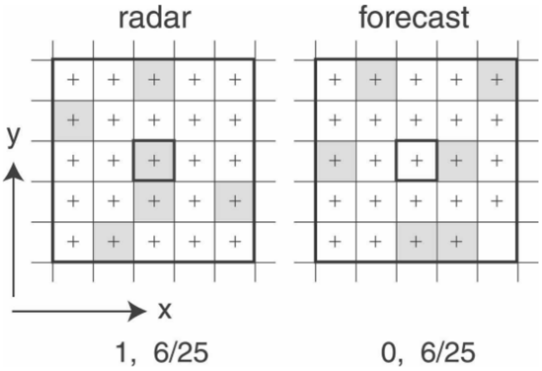

Figure 2.6: Schematic demonstrating the utility of the Fractions Skill Score (FSS). A grid point containing a feature (for example, precipitation above a given threshold) is assigned a binary value of 1 (shaded), which grid points not containing a featuer are assigned 0 (unshaded). Values below each grid represent the number of features found at the central grid point, and in the 5 × 5 neighbourhood surrounding that grid point respectively. Reproduced from Roberts and Lean 2008.

ideal forecast giving zero for all three.

Alternatively, neighbourhood, or "fuzzy", verification methods apply spatial smoothing to each input feature field, which relaxes the constraint that the inputs must match at the gridscale. The most popular neighbourhood method is the Fractions Skill Score (Roberts and Lean 2008) which performs a simple mean-square difference between the two smoothed inputs and normalises by a low skill baseline. Figure 2.6 shows an example FSS calculation for a single grid point. It is not uncommon to find spatially noisy fields similar to those presented in this figure when verifying precipitation. The central grid point does not agree on the presence of a feature in the radar and the forecast, meaning that a metric that evaluated skill at the gridscale would score this forecast poorly. However, because there is an equal number of grid points containing features in the 5 × 5 neighbourhood surrounding this grid point, the FSS would score this forecast highly.

Users can select the desired level of smoothing by changing the number of grid points used within each neighbourhood. Here, the term neighbourhood is simply a 2D moving average, or convolution, applied to each input field. The larger the neighbourhood used, the more smoothing is applied, resulting in more similarity and larger scores. This choice of neighbourhood size allows the user to analyse performance across a range of scales. So, if the user is interested in comparing convective-scale differences with synoptic-scale differences, they can choose appropriate neighbourhood sizes that will probe those scales. Alternatively, because scores typically increase with neighbourhood size, the user could instead find the scale at which an acceptable score is reached. In Roberts and Lean 2008, this size is defined as the skillful scale (when it is used to compare a model with verification), and is set to be the neighbourhood size at which the FSS reaches approximately 0.5 (plus a correction for feature coverage).

The FSS can also be used to assess spread in ensembles. By calculating the FSS between each ensemble member-member pair and averaging the result, the dispersion FSS (dFSS) measures the total similarity between members at a particular scale (Dey et al. 2014). As a measure of spread, the dFSS is negatively oriented, so scores of 1 mean there is no spread, and scores of 0 means there is maximum spread. The dFSS can then be compared with the error FSS (eFSS), which calculates the FSS between each ensemble member and verification, and then averages the result. Since the FSS is primarily a measure of spatial similarity, the difference between the dFSS and the eFSS measures the spatial spread-skill relationship. The use of neighbourhood smoothing in the FSS means it is a more appropriate measure of the spread-skill relationship than the standard RMSE/standard deviation approach when evaluating high-resolution precipitation forecasts. Not accounting for the double penalty problem could explain the large discrepancies found in some studies when precipitation evaluations are compared with other metrics (Frogner et al. 2019a,b). However, other studies which do use the FSS also find that ensembles are spatially underspread (Cafaro et al. 2021; Ferrett et al. 2021), meaning that precipitation is too colocated in each member given the skill of those members.

In a few of the aforementioned studies, CPE performance is directly compared with global ensemble performance evaluated over the same regions and over the same periods. Despite being evaluated as underspread in each of these studies, CPEs were consistently found to be less underspread than the global ensembles that drive them (Cafaro et al. 2021; Clark et al. 2009, 2010; Ferrett et al. 2021; Frogner et al. 2019a; Klasa et al. 2018; Vié et al. 2011). We should expect these improvements given the increase in skill (Section 2.2.2) and spread (Section 2.3.3) from running ensembles at convection permitting resolutions. While these findings are encouraging, recall from Fig. 2.5 that we may only expect to observe these benefits in the situations when the CPE can develop this additional, skillful detail.

Additionally, some studies have interrogated the use of multi-model blends for improving ensemble spread, with mixed results. In Beck et al. 2016, the multi-model ensemble produced more appropriate spread-skill relationships and rank histograms than any of the component ensembles. Additionally, Marsigli et al. 2014 found that driving a limited-area ensemble using inputs from a diversity of sources improved probabilistic verification measures. However, Porson et al. 2019 found the opposite, that the multi-model blend only possessed as much spread as the component model with the largest spread at a given time.

Overall, insufficient spread implies that there is a source of uncertainty that is not being represented within ensembles. After all, the foundational principles underpinning ensemble efficacy require all sources of uncertainty be accounted for, whether they are related to intrinsic, flow-dependent errors or to model imperfections. However, another often overlooked source of error present during validation is observation uncertainty. It is often assumed that verification datasets represent perfect measurements of the atmospheric state at the time the measurements were taken. This assumption is clearly quite broad and rarely holds. Radar datasets, for instance, become a less reliable indicator of surface precipitation at greater distances due to beam attenuation, the presence of clutter, impeding objects, and surface curvature (Golding 1998). While some of these issues can be corrected within radar composites, some errors will inevitably remain. Whether verification is being made for precipitation or for another parameter, observation uncertainty can be accounted for by perturbing ensemble members using an estimate of the observation error, as sampled using a known error distribution (Hamill 2001). Applying these adjustments can improve the underdispersion observed in ensemble forecasts as assessed using rank histograms (Bouttier et al. 2016; Frogner et al. 2019b) and the spread-skill relationship (Bouttier et al. 2016). However, even accounting for observational errors, CPEs remain somewhat underspread.

While it is always possible to increase the overall spread by fine-tuning other ensemble parameters, this should not come at the expense of reduced reliability or sharpness. Fortunately, post-processing methods provide an inexpensive method for compensating for this lack of spread. In particular, neighbourhood-based methods can artificially increase the sample size by treating values in surrounding grid points as additional samples (see Sect. 2.1.4). Given the previous discussion in this section, it is important to clarify that neighbourhooding for post-processing is different than neighbourhooding for model evaluation. Post-processing requires searching surrounding grid points for the maximum value within the neighbourhood, while model evaluation requires averaging surrounding values to smooth the field. Despite some apparent confusion in the literature (Schwartz and Sobash 2017), these processes produce very different results and should not be used interchangeably. It is clear from previous studies that post-process neighbourhooding is practically required to obtain sufficient probabilistic guidance that is not impacted as strongly by the underlying lack of spread (Craig et al. 2022; Flack et al. 2021; Frogner et al. 2019a; Raynaud and Bouttier 2017; Schwartz et al. 2010; Tempest et al. 2024). In the next section, further post-processing techniques and CPE uses will be described.

2.4 Exploiting the Benefits of Convection-Permitting Ensembles

Modelling at convection-permitting scales, combined with the application of ensembles, can provide substantial benefits for forecasters. For instance, as part of ongoing ensemble assessments, the UKMO regularly runs testbeds to subjectively evaluate model performance and trial new services. A common activity within these testbeds is to assess the added value provided by the operational CPE, MOGREPS-UK, through forecast denial experiments. These experiments task a group of participants with producing warnings for upcoming weather, but with access to only a limited set of NWP outputs. In the 2023 testbed, the group with access to MOGREPS-UK data produced more accurate warning areas for convection than the group that only had access to deterministic data (which also included a "poor-man's" ensemble of the four most recent deterministic outputs produced every six hours), showing the benefits of running the CPE (Pearson et al. 2023, internal report). Similar activities are conducted at each testbed to track the changing value as the CPE is updated and improved.

Meteorological centres in different regions can have different uses for CPEs. In Southeast Asia, for instance, running CPEs can provide more accurate probabilistic guidance on the tracks and intensities on tropical storms, which provides better assessments of likely impacts and reasonable worst-case scenarios (Met Office 2025). In the United States, CPEs can enhance the prediction of tornadoes and other impactful convective features at daily leadtimes (Gallo et al. 2016; Schwartz 2019). Additionally, global climate models are starting to adopt convection-permitting grids to provide more robust statistical studies of changes in extreme weather (Fosser et al. 2024). These are just a few examples of the broader use of CPEs, but given the focus of this thesis on understanding the behaviour of ensembles over the UK at hourly and daily timescales, this section will also primarily discuss the uses of CPEs within the UK Met Office (UKMO) in an NWP setting.

2.4.1 Post-Processing Convection-Permitting Ensemble Outputs

In most aspects, CPEs can be used in the same way as any other ensemble. In theory, each member of a well-tuned ensemble can be interpreted as an equally-likely realisation of the upcoming weather that can be inspected without further processing. It is therefore common to produce postage stamp plots showing each of these realizations. This procedure can help with the modification of analyses, where operational meteorologists use their knowledge and understanding of model biases, informed by uncertainties from EPS, to modify analysis charts (Young and

Grahame 2024a). However, forecasters rarely have the time to thoroughly inspect the outputs of each field in each member and at each timestep, necessitating the use of methods that can summarise this information. The simplest consolidation method is to produce deterministic-like products which uses data from all ensemble members to produce mean fields. Features common to all members are retained in the mean field while uncommon features are outweighed, ultimately producing an output that is often more skilful than a similar output from a deterministic run (Molteni et al. 1996; Whitaker and Loughe 1998). Alternatively, probabilistic fields are also used to highlight the regions where members agree or disagree on the exceedance of a specified threshold. While these methods are conceptually simple, they are typically only useful with spatially smooth fields, and only when members are not split into multiple regimes.

Recently, there has been a desire to produce methods that can summarise ensemble information more intelligently and for different purposes. One example is the use of clustering methods for the objective identification of weather patterns (Fereday et al. 2008; Ferranti and Corti 2011). These methods have been very successful at categorising and communicating synoptic-scale uncertainty out to leadtimes of multiple weeks, with separate schemes in use covering Europe at the ECMWF (Ferranti and Corti 2011; Ferranti et al. 2015) and the UKMO (Neal et al. 2016, 2024), as well as over North America (Lee and Messori 2024; Lee et al. 2023). Given the medium-range timescales that these schemes target, they are often produced using coarser global ensembles. While regime-based clustering is successful at concisely summarising forecast uncertainty at these timescales, it is too broad to classify differences between individual features within forecasts, which limits its usefulness in the short term. As such, recent attention has focussed on applying clustering techniques in a more dynamic way, where groups of members are found directly from the ensemble and do not need to be compared to a predetermined climatology (Boykin 2022).

In general, there are two broad methods used for clustering. In hierarchical clustering, clusters are formed either by successively grouping the nearest data points (agglomerative) or by splitting the furthest data points from a single initial group (divisive) until the requested number of clusters have been reached (Omran et al. 2007). Previous studies which choose a hierarchical approach typically use Ward's method, which forms clusters by combining data points such that the total within-cluster distance is minimised at each step (Atger 1999; Bouttier and Raynaud 2018; Branković et al. 2008; Marsigli et al. 2001; Molteni et al. 2001; Nakaegawa and Kanamitsu 2006; Nuissier et al. 2012; Weidle et al. 2013). While this method is conceptually simple and more effective at distinguishing global minima from local minima, it is also more restrictive with its placement of members within clusters

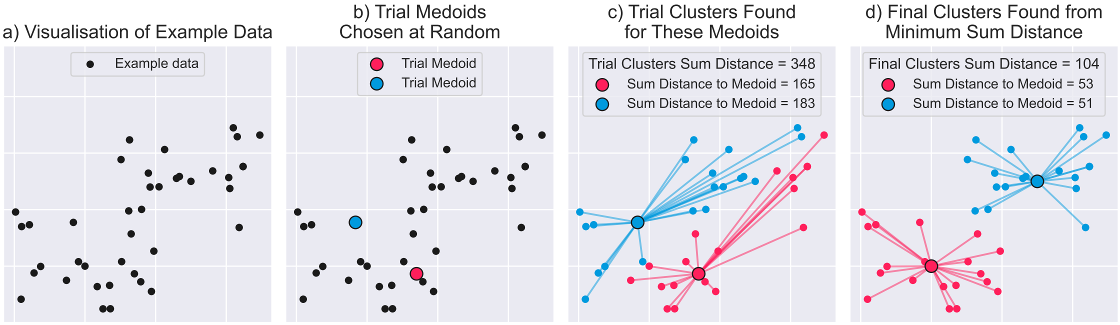

Figure 2.7: Schematic demonstrating an example of the K-medoids procedure. Units are arbitrary.

than partitional clustering methods (Omran et al. 2007). Once a member has been assigned to a given cluster, it is unable to switch to a different cluster in subsequent iterations, even if another cluster may become a better fit. This drawback is particularly problematic for ‘weakly-clustered’ data, where there may be a large number of alternative cluster arrangements that are similarly valid. Alternatively, partitional methods determine clusters by finding the optimal points about which each cluster is centred. The number of clusters requested by the user determines the number of central points. These central points may either be a member of the cluster itself (K-medoids (Brusco et al. 2019)) or the mean distance between all members of the cluster (K-means, (Omran et al. 2007)). Most previous studies that choose a partitional approach use K-means due to its popularity in other fields (Ferranti et al. 2015; Serafin et al. 2019) and capacity to produce ‘fuzzy’ clusters, where members are assigned a coefficient to each cluster rather than a binary classification (Harr et al. 2008; Keller et al. 2011; Lamberson et al. 2023; Zheng et al. 2017). However, if the user is interested in finding the central members within each cluster, this is not automatically provided using K-means. A separate algorithm must therefore be employed, and the central members that are chosen will likely be sensitive to the decisions made in constructing this algorithm.

Figure 2.7 shows an example of the K-medoids procedure for some test data. In this example, the test data shown in the Fig. 2.7a) will be clustered into two groups. First, two data points are chosen at random as trial medoids, as demonstrated in Fig. 2.7b). Then, the distance between each other data point and the trial medoids is calculated and used to form clusters. A simple Cartesian distance is used in this example, but clustering higher-dimensional data will require a more complicated method of estimating distance. The clusters found for these trial medoids is shown in Fig. 2.7c). After iterating over all permutations of data points as trial medoids, the final clusters are decided based on the minimum distance between each data point and its corresponding medoid, as shown in Fig. 2.7d).

Since clustering methods can identify the distinct forecast scenarios that contain the strongest support from other ensemble members, they could also be used to identify members that are more likely to be outliers. In theory, any members that are placed into their own separate cluster, or small clusters, are less likely to verify compared to members in larger clusters. This theory motivates the exploration of ensemble sub-setting, whereby ensemble members that are least likely to verify are ejected from the ensemble. This would refine the focus of the ensemble onto the members that are most supported, thereby providing a cheap method of improving ensemble performance. However, this approach also has the potential to falsely discount low-likelihood, high-impact events. Therefore, it is important to thoroughly test this technique to understand the limitations.

Ensemble Sensitivity Analysis (ESA) is another novel method of exploiting ensemble data. ESA analyses the sensitivity of a forecast within a target region to differences in the conditions in the upstream flow (Ancell and Hakim 2007; Torn and Hakim 2008). For instance, the forecast of precipitation associated with a developing MCS will likely be dependent on the initial trajectory of the ingredients required for this MCS to develop, such as a region of high wet-bulb potential temperature coinciding with a region of large wind shear. ESA can link the initial behaviour of these parameters to the resulting weather using simple linear relationships between variables. It is important to note, however, that this method merely highlights correlations within parameters which may not necessarily be causally related. Nevertheless, ESA has proved to be a popular tool within research contexts (Bednarczyk and Ancell 2015; Hanley et al. 2013; Necker et al. 2020), and has also been explored as a tool to provide operational value (Pearson et al. 2023). For instance, ESA may be used to improve forecasts by targetting observations in areas most likely to impact the resulting weather. Alternatively, if conducting additional observations is not possible, any observations that have been made can be used to subselect ensemble members that are most accurate within the upstream sensitivity area (Ancell and Hakim 2007).

Given the number of NWP models running operationally at meteorological centres, it can be difficult to know which model to use over another, especially when there is overlap between their most accurate operating windows. Indeed, there can often be useful signals present in one model but not another. Producing multi-model blends can capture the broader spread between ensembles which can help produce more appropriate guidance (Candille 2009). At the UKMO, the IMPROVER system produces probabilistic blends that incorporate nowcasting (short-range statistical advection of current radar fields, accurate up to 6 h leadtime), deterministic, CPE and global ensemble outputs, with weights applied to each model based on the leadtime (Roberts et al. 2023). For instance, at the time of writing, the blend at T+6 h comprised 25% deterministic and 75% CPE input (Roberts et al. 2023), though this can easily be changed. IMPROVER also includes additional postprocessing techniques such as EMOS (Gneiting et al. 2005) and neighbourhooding to improve spread (Theis et al. 2005), which are implemented with intelligent land-sea and orography masks.

Given the wealth of ensemble information produced each day, there is large potential for this data to be used in downstream models. The next section will describe some of these models.

2.4.2 Using Convection-Permitting Ensemble Outputs to Drive Downstream Models

One of the first downstream models to use MOGREPS-UK data was the internal MOGREPS-W ('W' standing for 'Warnings') system, a flexible first-guess warning tool whereby probabilistic CPE outputs for a selection of parameters were combined with thresholds to assess weather-related risk (Neal et al. 2014). These risk areas could be defined at a county level or for individual grid points. Additionally, the impact thresholds could be defined and updated regionally to account for local factors and antecedent conditions. This system has provided value for forecasters when producing warnings under the National Severe Weather Warning Service (NSWWS), as demonstrated by the automated warnings produced when predicting the impacts of strong wind gusts associated with post-tropical storm Katia in 2011 (Neal et al. 2014).

However, the emphasis with MOGREPS-W was on "first guess" warnings as and aid for forecasters to produce more informed guidance. More recently, other bespoke tools have been developed which use CPE data to make more specific predictions. Following widespread flooding in the UK in 2007, the Flood Forecasting Centre (FFC) and Scottish Flood Forecasting Service (SFFS) were established as partnerships between the UKMO and the Environment Agency and Scottish Environmental Protection Agency, respectively (Cabinet Office 2008). These agencies were established to predict and communicate the hazards associated with surface water flooding through the estimations of reasonable worst-case scenarios. As such, these agencies have developed tools for estimating the likelihood and impacts of surface water flooding. Initial research showed that CPEs outputs were a useful source of data for predicting the regions most likely to be impacted by surface water flooding associated with summertime convection (Golding et al. 2016). Neighbourhood postprocessing is also required to account for limited ensemble membership and inflate the probabilities to an appropriate size (Speight et al. 2021). Flood forecasting was successfully trialled over Glasgow for the 2014 Commonwealth Games, where

CPE and nowcasting outputs were used in conjunction to provide guidance across a range of timescales (Speight et al. 2018). Other recent models compare reasonable worst-case scenarios generated from CPE and nowcasting outputs to pre-simulated hydrological impact scenarios to further refine the areas most at risk (Maybee et al. 2024). Recent advances in surface water forecasting, including the role of CPEs, the strategies involved with constructing a flood model, and the use of products generated by these models is discussed in the review article by Speight et al. 2021.

In South-East Asia, the risks associated with high-impact weather are being studied and quantified using CPEs. Here, large-scale weather features such as the Madden-Julian Oscillation, tropical waves, and the formation of tropical storms can bring hazardous conditions to this part of the world. Recently, studies have shown that CPEs provide better forecasts over this region than global ensembles (Ferrett et al. 2021; Porson et al. 2019). These convection-permitting forecasts can then be improved by utilising the enhanced predictability associated with the aforementioned large-scale weather processes via hybrid statistical-dynamical models (Wolf et al. 2024). The value demonstrated by running CPEs over this region has led to the production of MOGREPS-SEA ('SEA' standing for 'South-East Asia'), a set of nested ensembles covering different domains within the region.

Other sectors also use probabilistic NWP outputs for decision making, although few require these outputs to contain convective-scale detail. For instance, ensemble data is incredibly useful for forecasting energy production from wind power (Baran and Lerch 2015; Phipps et al. 2022; Al-Yahyai et al. 2012; Zhang et al. 2014) and the broader demands placed on the energy grid (Yang and Kleissl 2023). In Sub-Saharan Africa, sporadic outbreaks of Meningitis have been linked to periods of high temperature, high dust concentration, and low humidity. As such, a Meningitis early warning system has been produced to assess the risk of outbreak up to two weeks ahead using global ensemble data (Dione et al. 2022).

Ultimately, the production of these warning systems is only useful if the predicted risks can be effectively communicated to the intended audience. In the UK, NSWWS weather warnings are issued based on an impact matrix which also accounts for event likelihood to determine the severity of the warning (Suri and Davies 2021). These warnings can then be communicated across TV broadcasts, social media platforms and journalistic outlets. For localised hazards associated with more spatially unpredictable weather like severe convection, there may also be the need for more targetted and timely warnings driven by observations and nowcasting procedures (Golding 2022). Such systems are already commonplace in the USA, especially in the South and Mid-West, where tornadoes can produce significant risks (Craven et al. 2020). While some indication of hazardous weather can be communicated a few days in advance, the specific impacts can only accurately be predicted after the particular system has started to develop. Recent tragedies like the flash flooding event in Valencia highlight the need for robust early warning systems that can accurately communicate severity, impacts, and suggested actions to the correct users (World Meteorological Organization 2024).

Chapter 3

Spread–Skill Relationship Improvements through Blending

This chapter has been published in the Quarterly Journal of the Royal Meteorological Society with the following reference:

> 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, 150(762), 3146–3166. Available from: https://doi.org/10.1002/qj.4754

The roles of the other authors of this paper in relation to the project are as follows: S. L. Gray (supervisor: academic), T.H.A. Frame (supervisor: academic), A.N. Porson (supervisor: Met Office), M. Milan (supervisor: Met Office). The study was designed in collaboration with my supervisors, with the new data assimilation technique and ensemble experiments designed by A.N. Porson and M. Milan, and the verification strategy discussed among all paper authors. I performed the formal analysis and validation using code provided by A.N. Porson, and I adapted it for the purposes of the study. Guidance on this project was provided from all supervisors through weekly meetings. I wrote the first draft of the paper, prepared all figures, and had overall control of the submitted paper. All authors contributed to reviewing and editing the published manuscript. Approximately 70% of the paper was my work, and 30% was contributions from other authors.

Abstract

Convective-scale ensembles are routinely used 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.

3.1 Introduction

Convective-scale ensembles have been used for over a decade to quantify the uncertainty in convective-scale weather forecasts (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 skillful 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 (Beck et al. 2016; Cafaro et al. 2021; Ferrett et al. 2021; Porson et al. 2019, 2020; Raynaud and 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 and Fischer 2008). Therefore, it is expected that a regional model DA scheme will represent scales approaching its own domain size less accurately compared to 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 similar or larger rate as 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 and 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. But 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 to 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 is initialised using 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 to 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 consistently demonstrated the benefits of using blending schemes in regional models. For instance, blending has been shown to remove large systematic biases effecting 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 hours 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 have dependence 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 – 1,000 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 further improve model performance compared to 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" (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 recentering 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 ROC area (not shown), which suggests that the predominant benefit will instead be observed spatially. Additionally, we would expect LSB to impact other variables than just precipitation, particularly those that undergo blending directly (Milan et al. 2023), but we do not analyse this here.

This paper 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, the results section (Section 3) 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 against 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 paper by discussing limitations and future work.

3.2 Methods

This section starts by outlining the ensemble configuration used in this work (section 3.2.1), before describing how LSB is implemented within this ensemble (section 3.2.2). After this, the metrics used to assess the spread-skill relationship are presented (sections 4.2.2 and 3.2.4) before concluding with a discussion on our significance estimation approach (section 3.2.5).

3.2.1 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 3.1 shows a schematic of the initialisation procedure. Note that this is updated from figure 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 3 new members run out to 120 hours, 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 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 6 hours at 0000, 0600, 1200, and 1800 UTC

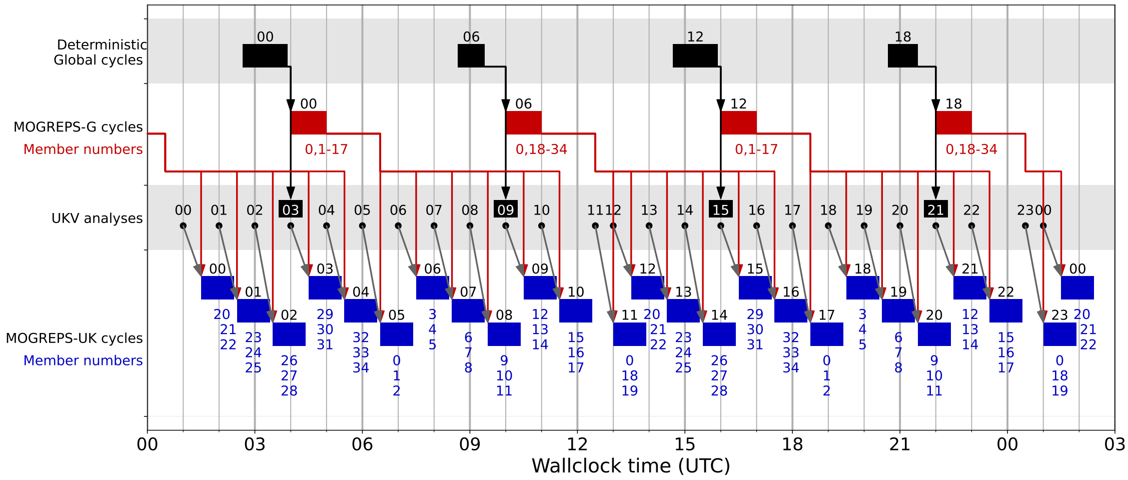

Figure 3.1: Schematic showing the data flow for initialising time-lagged MOGREPS-UK ensemble. Top, black boxes show the six-hourly deterministic global model cycling frequency. The red boxes, arrows and numbers show the MOGREPS-G members that provide initial condition perturbations and lateral boundary conditions. The black dots and grey arrow show the UKV analyses around which a given MOGREPS-UK cycle is centered. 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).

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 3.1). Note that there is no relation between members with the same labels initialised 12 hours 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 mid-latitudes (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 which were an order of magnitude smaller than for the summer and are therefore not discussed further in this work.

3.2.2 Large-Scale Blending in MOGREPS-UK

Large-Scale Blending (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 fully integrate blending 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) with 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 figure 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 section 2.2 and the appendix of Milan et al. 2023).

The only difference between the "blended" and "reference" MOGREPS-UK configurations used in this study are 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 where 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:

- LSB is only directly applied to the initial conditions of the members initialised at 0300, 0900, 1500, and 2100 UTC (UKV analyses with black boxes in figure 3.1). 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 which have used blended backgrounds, i.e., LSB has been applied during a *prior* cycle (UKV analyses without black boxes in figure 3.1). 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 the 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 which 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 briefly inspect the member fields. Figure 3.2 shows a comparison of hourly precipitation accumulations for a selection of members from the reference and blended ensembles. This period occurs 10 hours before the case study considered in section 3.3.4 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 3.3.2. There are, however, subtle differences in the spatial patterns which we hypothesise to be, on average, more accurate in the blended ensemble. One of the focuses of this study is to verify this statement, which is achieved using the Fractions Skill Score.

3.2.3 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 and Lean 2008), a neighbourhood-based metric designed to calculate the difference between two fields

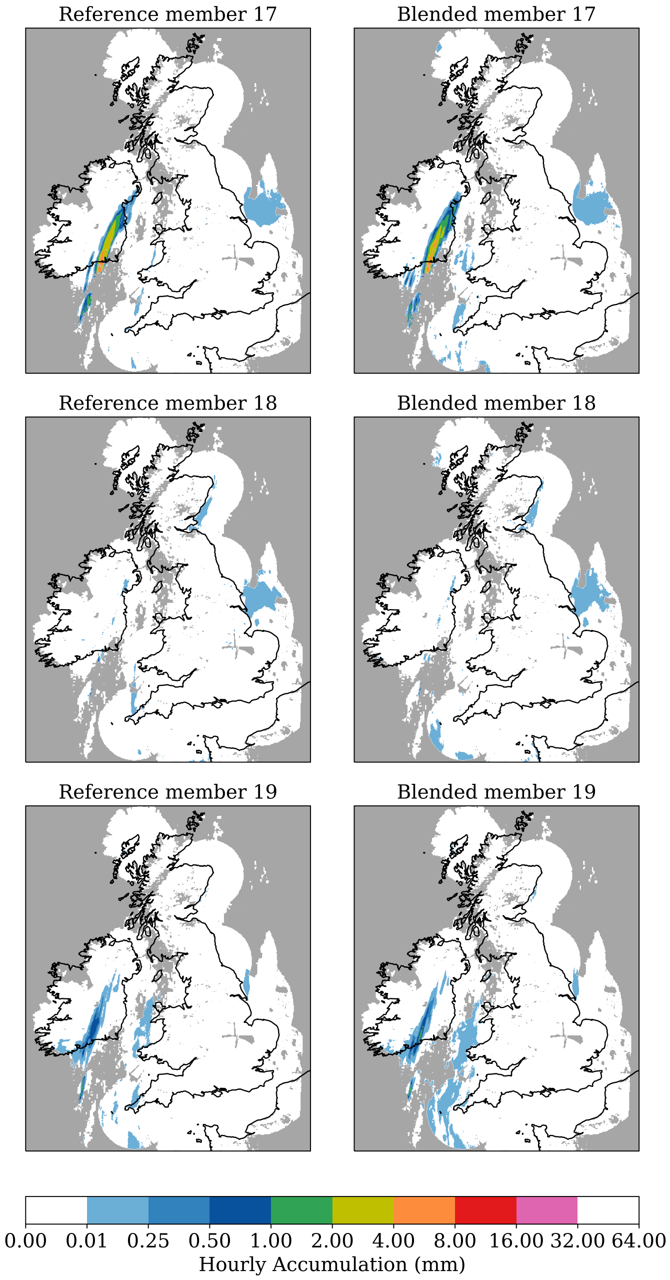

Figure 3.2: A selection of postage stamps from the reference and blended ensembles for the 29 June 2019, 0300 UTC ensemble forecast, leadtime T+4 h. Members 17 were initialised at 2200 UTC the previous day (leadtime T+9 h), 1 hour after direct blending. Members 18 and 19 were initialised at 2300 UTC the previous day (leadtime T+8 h), 2 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 4.2.2.

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 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 which 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:

$$\mathrm{MSD}_{(n)}\left(A,B\right) = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left[A_{(n)i,j} - B_{(n)i,j}\right]^2 , \tag{3.1}$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A,B\right) = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left[A^2_{(n)i,j} + B^2_{(n)i,j}\right] , \tag{3.2}$$

$$\mathrm{FSS}_{(n)}\left(A,B\right) = 1 - \frac{\mathrm{MSD}_{(n)}\left(A,B\right)}{\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A,B\right)} , \tag{3.3}$$

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 which is used to calculate the fractions fields regrids the data onto a stage grid with spacing 2,327 m (Roberts et al. 2023), hence the grid point to km conversion is slightly different than the expected 2.2 km for MOGREPS-UK.

Typically, the FSS is used to understand the scales at which a deterministic forecast becomes skillful by comparing the forecast to 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 to 100 km (see figure 3.7). 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 skillful 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 fractions 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 which 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:

$$\mathrm{dFSS}_{(n)} = \frac{1}{M\left(M-1\right)} \sum_{M_a=1}^{M} \sum_{\substack{M_b=1 \\ M_a \neq M_b}}^{M} \mathrm{FSS}_{(n)}\left(M_a, M_b\right) \ , \tag{3.4}$$

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

$$\mathrm{eFSS}_{(n)} = \frac{1}{M} \sum_{M_a=1}^{M} \mathrm{FSS}_{(n)}\left(M_a, O\right) \ , \tag{3.5}$$

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.

3.2.4 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 5.4–5.6, but instead uses summations over time to obtain a spatial field of scores at each grid point, i, j. The LFSS is calculated as

$$\mathrm{MSD}_{(n,i,j)}\left(A,B\right) = \frac{1}{T}\sum_{t=1}^{T}\left[A_{(n,i,j)t} - B_{(n,i,j)t}\right]^2 \ , \tag{3.6}$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n,i,j)}\left(A,B\right) = \frac{1}{T}\sum_{t=1}^{T}\left[A^2_{(n,i,j)t} + B^2_{(n,i,j)t}\right] \ , \tag{3.7}$$

$$\mathrm{LFSS}_{(n,i,j)}\left(A,B\right) = 1 - \frac{\mathrm{MSD}_{(n,i,j)}\left(A,B\right)}{\mathrm{MSD}^{\mathrm{ref}}_{(n,i,j)}\left(A,B\right)} \ , \tag{3.8}$$

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:

$$\mathrm{dLFSS}_{(n,i,j)} = \frac{1}{M\left(M-1\right)}\sum_{M_a=1}^{M}\sum_{\substack{M_b=1\\ M_a\neq M_b}}^{M}\mathrm{LFSS}_{(n,i,j)}\left(M_a, M_b\right) \ , \tag{3.9}$$

$$\mathrm{eLFSS}_{(n,i,j)} = \frac{1}{M}\sum_{M_a=1}^{M}\mathrm{LFSS}_{(n,i,j)}\left(M_a, O\right) \ . \tag{3.10}$$

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 leadtimes, 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 leadtimes over a single cycle using integration periods of 24 hours (Woodhams et al. 2018) and 3 hours (Ferrett et al. 2021). Our results use 12-hour periods sequencing hourly precipitation fields from leadtimes T+2 to T+13 h.

3.2.5 Significance Estimation

The differences between the precipitation fields of different members of the same ensemble configuration is much larger than the differences between the same member of the two configurations, as can be seen in figure 3.2. 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 which 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 which occurs purely due to LSB, not due to mixing members from sub-ensembles with different distributions. From this, we construct confidence limits which 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 3.5.

3.3 Results

3.3.1 Trial Period Characteristics

The UK was under a southwesterly flow at the beginning of the trial period (16 June 2019) which encouraged a number of convective storms to develop over southern

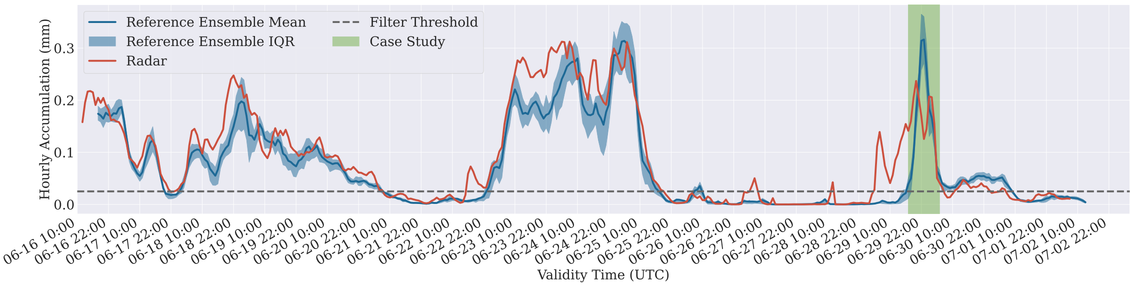

Figure 3.3: Time series of the domain-averaged hourly precipitation in the reference ensemble and radar. Ensemble data are calculated for leadtime T+8 h. The blended ensemble data have been omitted for clarity, but they largely follow the reference ensemble.

England. High pressure and settled conditions then moved in from 21 June and persisted until 24 June. Additional thunderstorms developed over southern and central England from 24 June, with slow moving light rain clearing from the north east in the early hours of 26 June. 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 29 June. Scattered showers persisted across the UK and Ireland until the end of the trial period on 2 July (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 3.3. Typically, the ensemble mean underestimates the precipitation across the domain compared to the radar. This is especially noticeable towards the end of 22 June 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 29 June, with a majority of members forecasting little or no rain even at short leadtimes (see figure 3.2). 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 on figure 3.3 and are studied in more detail in section 3.3.4.

Also highlighted on figure 3.3 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 means that it returns low scores when dry events are correctly forecast (as discussed in Mittermaier 2021). Moreover, the FSS behaves far more sensitively with low fractional precipitation coverages (Roberts and Lean 2008), which we have found can have a large effect on the average FSS. Low

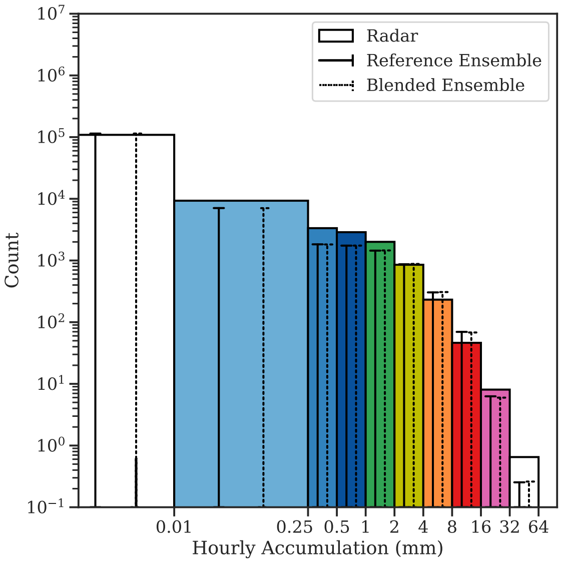

Figure 3.4: Domain-wide precipitation distributions from the radar data (bars) and both ensembles (stalks, representing the height of the equivalent bars in the ensemble histograms). Radar bars use the same logarithmic colourbar as in figure 3.2 for consistency. Ensemble distributions are averaged across all cycles, leadtimes and ensemble members.

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 3.3.3 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/hr by the American Meteorology Society Glossary of Meteorology (AMS 2023). Upon inspection, the filtered periods are primarily dominated by the high pressure conditions of 21–23 June, and 25–29 June.

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 the 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.

3.3.2 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

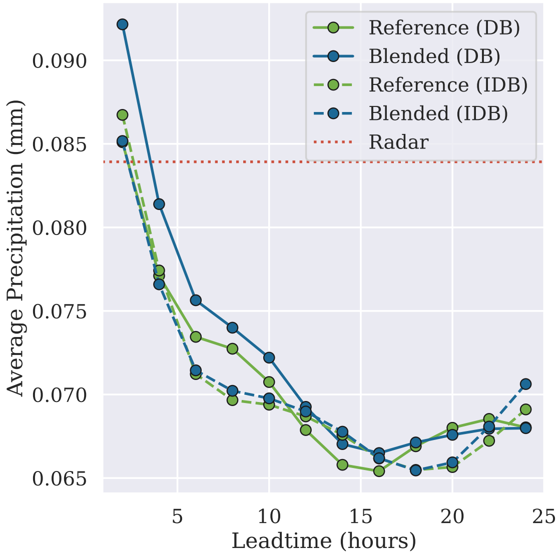

Figure 3.5: Domain-wide precipitation averaged across all cycles for the directly blended members (DB, solid lines) and a selection of indirectly blended members that were initialised five hours after blending (IDB, dashed lines). Radar value was calculated by averaging across all events in the trial period and does not have leadtime dependence.

of the overall accumulation. The postage stamps presented in figure 3.2 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 leadtimes, and all members.

Figure 3.4 shows that both ensembles under-represent the lightest and heaviest rain compared to 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 leadtime: averaged over leadtimes 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 leadtimes 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 3.5 shows domain- and cycle-average precipitation as a function of leadtime for two sets of sub-ensembles from both configurations. In the blended ensemble, the set of members labelled "DB" have been 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 effected by blending. The reference DB and IDB sets use these same selections of members, although no blending takes place in either set.

Figure 3.5 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 when averaged over all members. Note that this spindown behaviour is atypical when compared to 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 3.2.2 that the DB set is also the set which 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 is having 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.

3.3.3 Impact of LSB on spatially integrated spread-skill relationship

The variation in FSS across the trial period is shown in figure 3.6 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 leadtime cutoff of 24 hours based on previous LSB work with the deterministic UKV model, which demonstrated that the blending signal persists for approximately 18 hours (Milan et al. 2023). Additional work presented later in this section supports this cutoff leadtime. 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 3.7. Figure 3.6a) shows the dFSS (spread) scores for the reference ensemble, where higher values mean more confi-

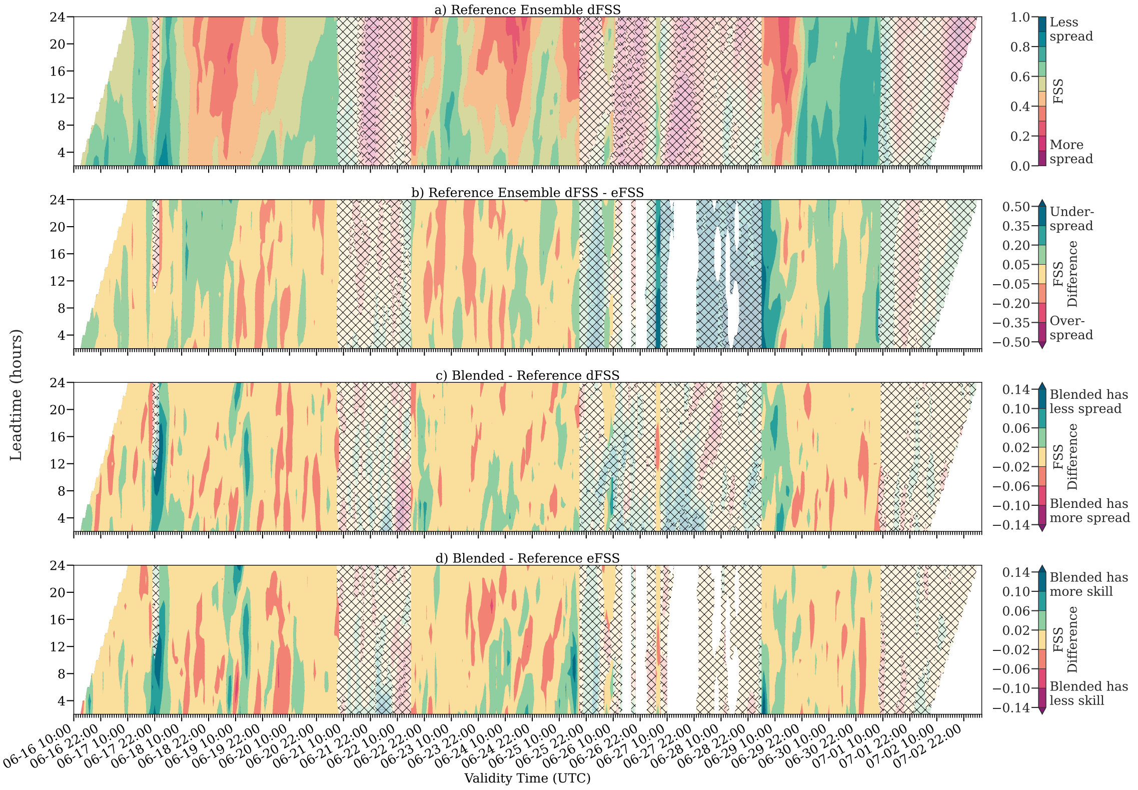

Figure 3.6: Contour plots showing evolution of 90th centile FSS with a 44 km (19 × 19 grid points) neighbourhood as a function of leadtime and validity time over the trial period. A single cycle is along the diagonal. Hatched sections are dry events which have been filtered out of the dataset in subsequent 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.

dence 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 leadtimes, as expected.

Next, figure 3.6b) 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 fully meet the requested centile. Typically the lower skill regions occur during the extended dry period between 26–29 June which has been filtered out. There is no clear dependence of the correctness-of-spread with leadtime, with some events becoming more correctly spread at shorter leadtimes, 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.

Figures 3.6c) and d) show the difference between the blended and reference

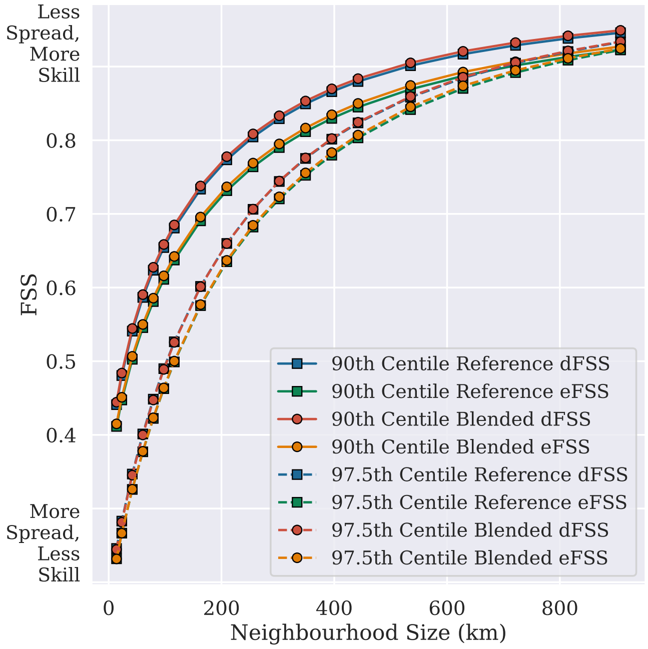

Figure 3.7: Scale-dependent dFSS (spread) and eFSS (skill) curves obtained by averaging FSS values similar to those presented in figure 3.6 a) over the trial period and over leadtimes T+2 to T+24 h. The reference and blended curves are difficult to distinguish, especially for the 97.5th centile.

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 3.6b), which is expected given the large similarity between fields shown in figure 3.2. One notable exception occurs towards the end of 17 June and start of 18 June, when the blended ensemble is both more skillful and more confident. These score increases are persistent across all leadtimes 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 3.7–3.10 show the averaged values across the filtered (non-hatched, full-opacity) space of figure 3.6, 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 separately aggregating the score components (Mittermaier 2021). The sensitivity of the scores to this choice will be explored in future work.

To start, figure 3.7 shows the expected increases in FSS with neighbourhood size (Roberts and 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

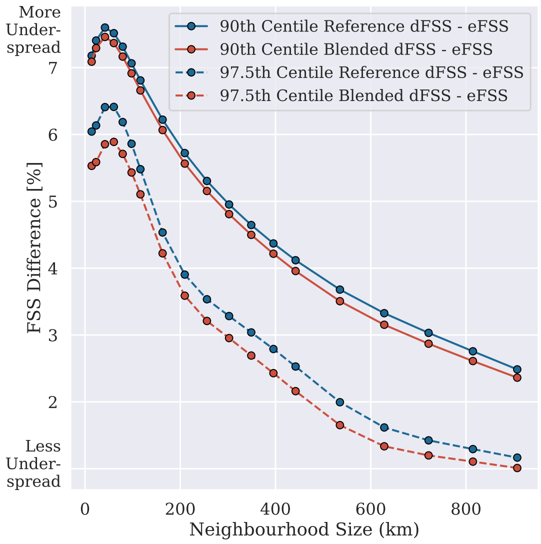

Figure 3.8: Scale-dependent percentage differences between dFSS and eFSS averages presented in figure 3.7. eFSS averages are used for normalisation.

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 3.8 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 at 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 3.7 and 3.8 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 3.7 increasing when LSB is applied. To see this explicitly, the solid lines of figure 3.9 show the difference between the blended and reference ensembles for the dFSS (spread) and eFSS (skill) for the 90th and 97.5th centiles. Larger values on 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 on these figures as dashed lines are the mixed-

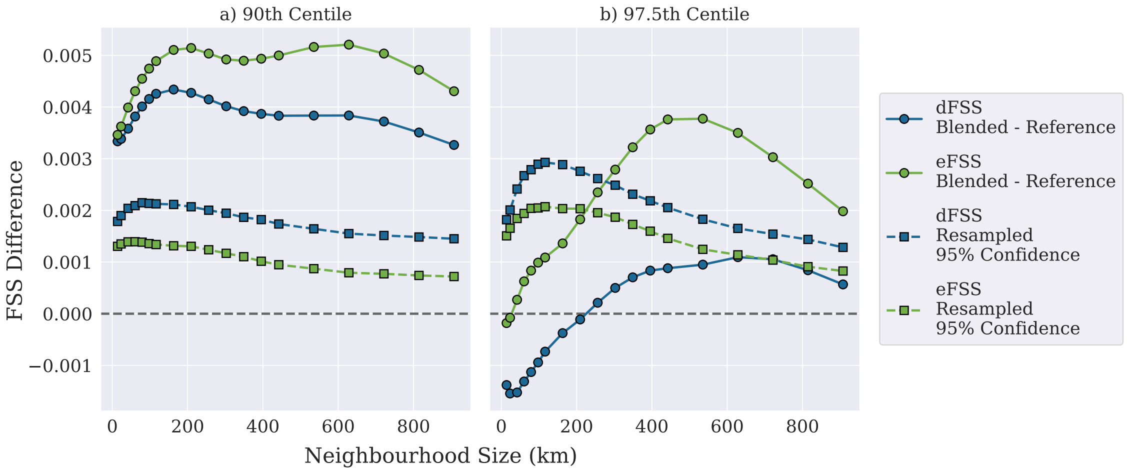

Figure 3.9: Scale-dependent FSS difference between the blended and reference ensembles for the spread (dFSS) and skill (eFSS) averages presented in figure 3.7. Averages are performed over leadtimes T+2 to T+24 h. Solid line, circle markers: difference between the blended and reference FSS averages. Dashed line, square markers: upper limit of the 95% confidence estimated through constrained resampling technique.

member 95% confidence estimations as described in section 3.2.5 and appendix 3.5.

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 3.7), 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 with 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

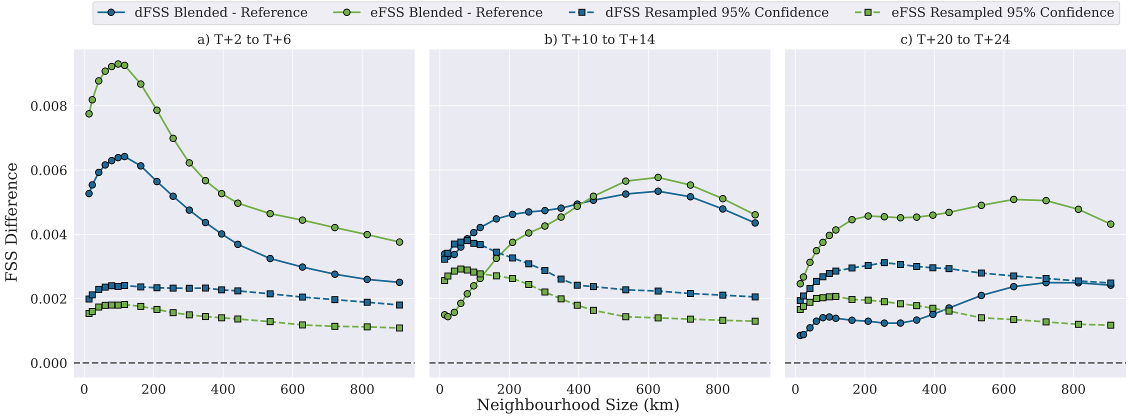

Figure 3.10: As with figure 3.9 but for the 90th centile only, separated by leadtime.

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 3.9 are averages across all leadtimes and validity times of the trial period, but it is also instructive to interrogate the leadtime dependence of the LSB response. Figure 3.10 has the same format as figure 3.9 but shows the results using the 90th centile for specific leadtime ranges. Given the fact that blending only modifies the initial conditions, we expect it to have the largest impact at early leadtimes, and this is indeed verified in this figure with maximum skill score differences approaching 0.01 between leadtimes T+2 to T+6 h. This difference diminishes with increasing leadtime for both spread and skill scores. The spread score ensemble differences become insignificant across all neighbourhoods for leadtimes longer than T+20 h. There is also a scale dependence observed with these results where score differences are larger towards the grid scale than synoptic scales at very short leadtimes (figure 3.10 a)), before decaying and becoming less significantly different compared to larger neighbourhood sizes at longer leadtimes (figure 3.10 b), c)). We speculate that this is related to the inherent growth rate of small-scale errors being much faster than large-scale ones (Lorenz 1969). We also notice a leadtime dependence to the skill improvements provided by LSB, with scores deteriorating at the grid scale between leadtimes T+12–T+18 h before recovering and becoming significant again at leadtimes longer than T+20 h. This dependence can be partly observed in figure 3.10b) 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 that:

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 3.9). 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 3.9). 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 3.8).

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 3.8).

5. The impact of LSB on ensemble spread persists for approximately 18–20 hours. The ensemble is made more skillful until at least 24 hours after initiation, although smaller improvements are noted between leadtimes T+12–T+18 h for smaller neighbourhood sizes (Figure 3.10).

3.3.4 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 29 June 1700 UTC to 30 June 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 leadtimes 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 leadtimes). Longer leadtime studies were considered to be less informative given the

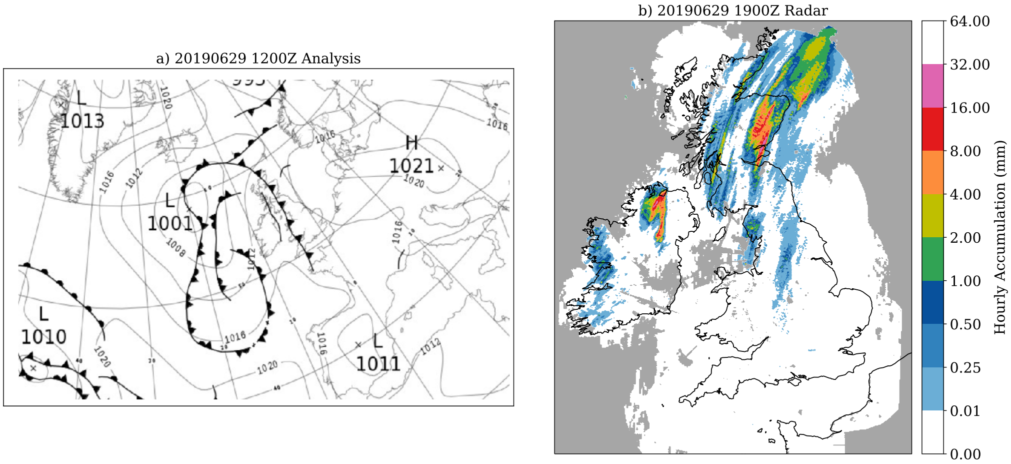

Figure 3.11: Synoptic overview of the case study period: a) synoptic chart from the Met Office daily weather summary (UKMO 2019) for the closest period before the integration window used in the LFSS case study. Contours show mean sea level pressure in 4 hPa intervals. b) Hourly rain radar valid at 29 June 2019 1900 UTC, chosen to show the time and location of the strongest period of elevated convection over Ireland.

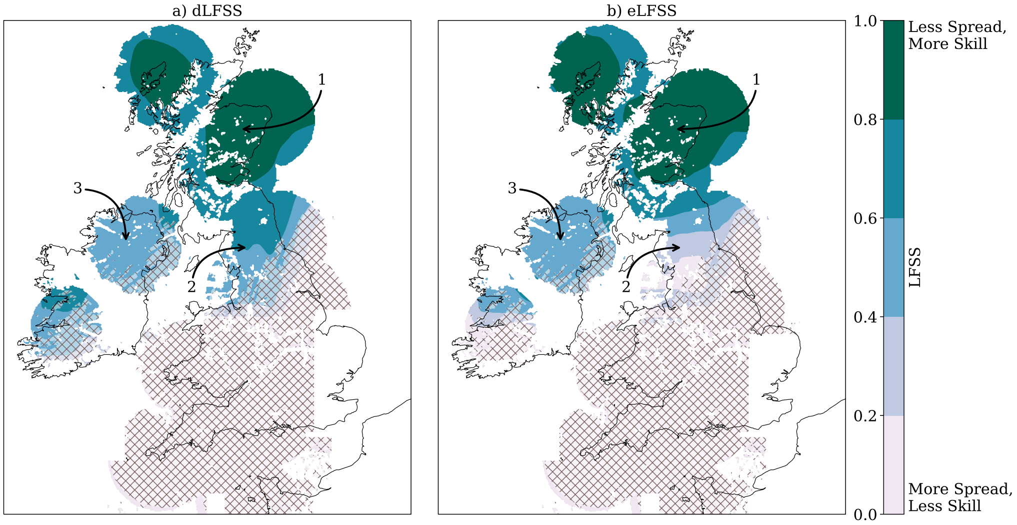

Figure 3.12: 97.5th-centile, 260-km neighbourhood size (112 × 112 grid points) dLFSS (left) and eLFSS (right) for the reference ensemble calculated over the case study period (1700 UTC 29 June to 0500 UTC 30 June 2019). If data are missing for a grid point in any of the fractions fields used to create 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 leadtimes 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 leadtimes, or there was 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-hour period is less than 1 mm in all three datasets (both ensembles and radar). Annotated regions are the focus of analysis in the text.

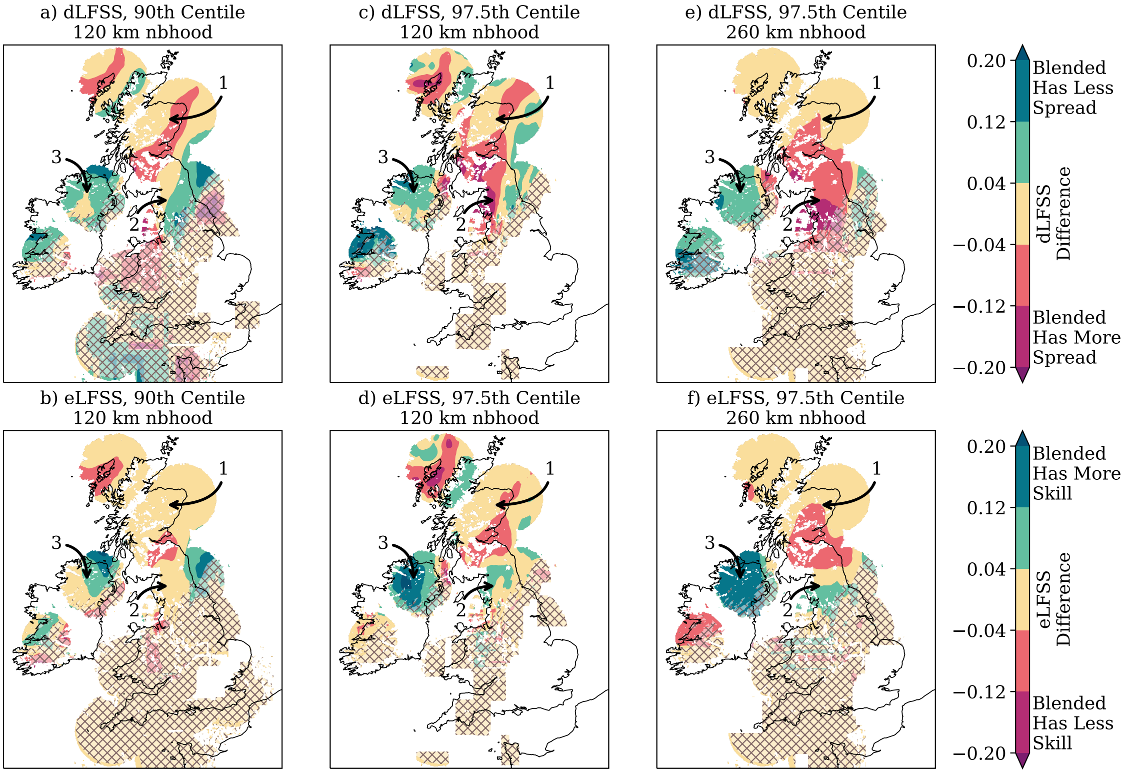

Figure 3.13: Blended − Reference LFSS differences for a selection of centiles and 120 km (51×51 grid points) and 260 km (112×112 grid points) neighbourhoods. For the dLFSS difference maps, higher scores show regions where the blended ensemble has larger dLFSS (lower spread) than the reference ensemble. 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.

short persistence of the LSB signal (section 3.3.3).

The case study conditions are shown in figure 3.11. At the start of the period, a band of thunderstorms 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 30 June 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 29 June 2019 1600 UTC, moving north-eastwards. 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 (pers. comm. David Flack), 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 3.12 largely meets this requirement for this case. Greater confidence and skill is shown over eastern Scotland (region 1) than 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 skillful 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-hour 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 3.13 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 overconfident forecast of convection. This is especially clear in the 97.5th-centile, 260-km neighbourhood size dLFSS map (figure 3.13e)), 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 3.13 of increased skill and decreased spread. Figures 3.13a) and b) show 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 more accurately positioned the broad precipitation envelope. 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 figures 3.13c)–f). For instance, figure 3.13d) shows maximum skill score improvements of more than 0.2, while panel 3.13f) 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 north east compared to 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.

3.4 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 to 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 further interrogated, 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 significantly affected both skill and spread scores across all neighbourhood sizes, but skill scores improved by an average of 0.56% in 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 Localised Fractions Skill Score 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 to reliability curves, rank histograms and ROC area (not shown). We also note a large seasonal dependence to impacts of LSB. This work only presents findings from the summer trial since the results of the winter trial showed minimal impacts. 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 to 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 to the FSS with LSB applied, the results presented in figure 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 which 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 as 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 was 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 hours from forecast initiation. This is in broad agreement with other work which has investigated the persistence of blending (Wang et al. 2014) and the persistence of initial conditions 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 leadtimes 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 that LSB is applied to MOGREPS-UK should be aware of the effect that this may have to the other ensemble members which currently only inherit the effect of blending through the ingestion of blended background fields. It may be possible to further improve skill 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 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 would like to 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 would also like to thank David Walters and Joanne Carr for providing the code used to adapt the MOGREPS-UK schematic in figure 3.1. Finally, we would like to 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

The authors declare no potential conflicts of interest.

Data Availability

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

3.5 Appendix: 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 skillful 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 to 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 subensembles which can each be considered an i.i.d. sample. Therefore, we seek to isolate the response which occurs purely due to LSB, not due to mixing members from sub-ensembles with different distributions. This requires the use of constraints

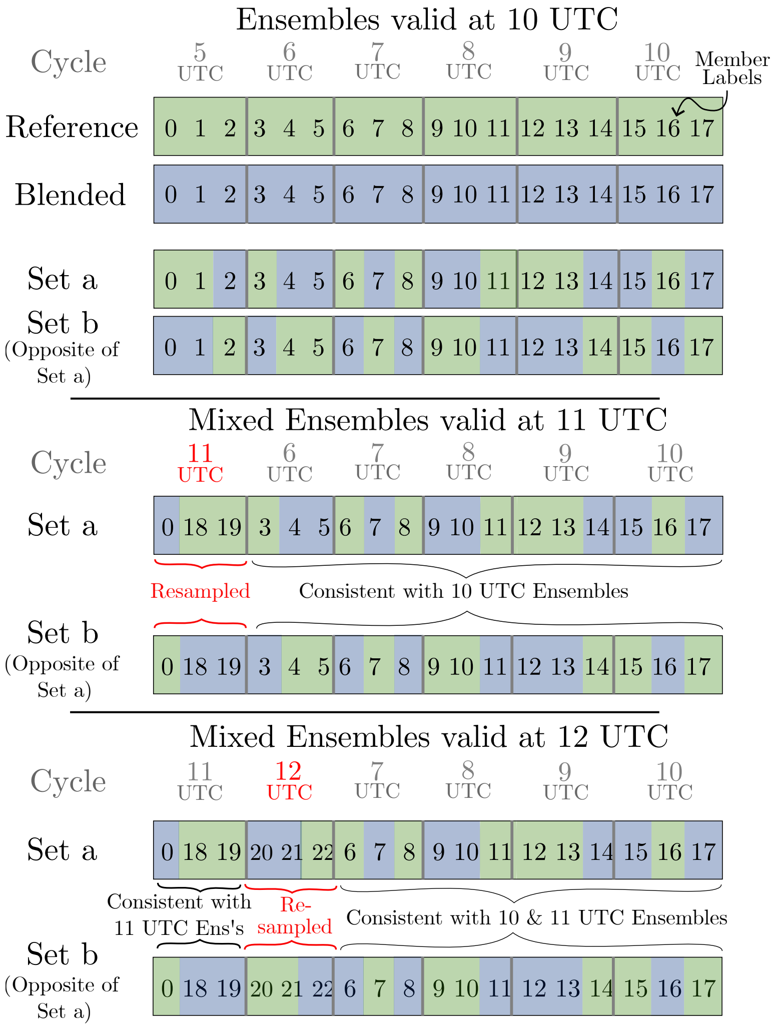

Figure 3.14: Schematic demonstrating an example of the ensemble member resampling process. The number within each box is the member label inherited from the corresponding MOGREPS-G member. Each hourly, time-lagged ensemble is comprised of the three-member sub-ensemble initialised during that cycle, along with the five previous the sub-ensembles. To create the first mixed-member ensembles (“Set a” and “Set b” at 10 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 (11 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 of an equal mix of reference and blended members.

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 which create the fairest comparison between mixed ensembles are the following:

1. Only members which are initialised during matching cycles are permuted;
2. Only the newly updated members at each hour are resampled; and
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 as outlined in Wilks 1997, however, since we already have knowledge of the data structure and their correlations, we do not need to anticipate some of the more user-dependent aspects of this method. Because these criteria were designed to most closely replicate the construction of the lagged ensembles themselves, they necessarily take into account additional variance which 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 3.14, 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 3 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 leadtime 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.

Chapter 4

Quantifying Driving Ensemble Influence on Convection-Permitting Ensemble Spread

This chapter has been through two rounds of revisions in the Quarterly Journal of the Royal Meteorological Society. It is expected that this chapter will be accepted for publication shortly following the submission of this thesis.

The roles of the other authors of this paper in relation to the project are as follows: S. L. Gray (supervisor: academic), T.H.A. Frame (supervisor: academic), A.N. Porson (supervisor: Met Office), M. Milan (supervisor: Met Office). The study was designed in collaboration with my supervisors, with the research questions discussed among all paper authors. I conceptualised, developed, and tested the methodology used in the paper with guidance from all authors. I retrieved all data and performed formal analysis, with guidance and interpretation provided from all supervisors through weekly meetings. I wrote the first draft of the paper, prepared all figures, and had overall control of the submitted paper. All authors contributed to reviewing and editing the manuscript. Approximately 80% of the paper was my work, and 20% was contributions from other authors.

Abstract

Convection-permitting ensembles (CPEs) are a common short-range forecasting tool designed to quantify the uncertainty in convective-scale processes, but their usefulness is limited by insufficient spread. While most efforts to improve spread have targeted the CPE itself, previous studies have shown that the “parent” driving ensemble can exert a strong influence over the “child” CPE. Few studies have examined the parent-child relationship for precipitation patterns, which are important for forecast guidance production but require the use of neighbourhood-based metrics for robust evaluation. By comparing spatial statistics between an operational CPE and the global ensemble used to drive it, we investigate the leadtimes and regimes under which the CPE diverges from the driving ensemble and link this to the spread-skill relationship. As a compliment to existing methods, we introduce the Parent-Child Fractions Skill Score (pcFSS) that directly compares precipitation patterns between the two ensembles. Both ensembles are similarly underspread under mobile regimes and at leadtimes when the boundaries dominate. Under convective regimes, the CPE shows the potential for larger forecast differences compared to the global ensemble. CPE spread is also larger in these regimes compared to others, but also suffers from lower skill. Ultimately, we show that using the pcFSS in conjunction with existing methods provides a broader understanding of CPE behaviour by highlighting instances of stronger and weaker driving ensemble influence.

4.1 Introduction

Global ensembles are used in operational centres around the world to estimate the growth of forecast uncertainty of synoptic and mesoscale phenomena with leadtime (e.g., Inverarity et al. 2023; Palmer 2019; Zhou et al. 2022). Recently, convection-permitting ensembles (CPEs) have been introduced that instead quantify the uncertainty in short-range, small-scale processes that global models are too coarse to resolve explicitly (e.g., Gebhardt et al. 2011; Hagelin et al. 2017; Raynaud and Bouttier 2017; Schwartz et al. 2015). Due to the computational cost of operating at the convective scale (with grid lengths 1–4 km), these ensembles are typically run over a limited domain and nested inside lower-resolution models which can provide both lateral boundary conditions and initial condition perturbations.

Many studies have shown that convection-permitting models produce far superior precipitation forecasts than those which must rely on convective parametrizations. Convective-scale models improve the diurnal cycle of precipitation (Clark et al. 2009; Woodhams et al. 2018, orographically enhanced precipitation (Barrett et al. 2015; Gowan et al. 2018; Schellander-Gorgas et al. 2017), the modelling of extreme precipitation (Clark et al. 2009; Duc et al. 2013; Frogner et al. 2019a; Marsigli et al. 2005; Marsigli et al. 2008; Schellander-Gorgas et al. 2017; Schwartz et al. 2010), probabilistic guidance such as event/non-event discrimination (Cafaro et al. 2019; Frogner et al. 2019a; Marsigli et al. 2008), and the forecasting of specific events (Barrett et al. 2016; Gallo et al. 2016; Hanley et al. 2011, 2013; Sobash et al. 2016; Trier et al. 2015).

Within CPEs, developers can attempt to introduce spread at multiple stages: by perturbing the initial conditions (ICs), by using different lateral boundary conditions (LBCs) from the driving model, by introducing stochastic perturbations to the physics schemes used in each member (Flack et al. 2021; McCabe et al. 2016), by combining members initialised around earlier analyses into a time-lagged group (Ben Bouallègue et al. 2013; Mittermaier 2007; Porson et al. 2020; Raynaud and Bouttier 2017), or by combining members from different configurations into a multimodel ensemble (Beck et al. 2016; Porson et al. 2019), to name just a few of the available methods. However, a well-known issue with CPEs is the lack of spread between members when compared to the eventual verification, especially for precipitation patterns (e.g., Beck et al. 2016; Cafaro et al. 2021; Ferrett et al. 2021; Porson et al. 2019, 2020; Raynaud and Bouttier 2017; Schwartz et al. 2014; Tennant 2015). This lack of spread leads to overconfident forecasts and reduces forecaster trust in these computationally expensive outputs.

Ideally, ensemble spread should be a reliable indicator of skill as measured through the spread-skill relationship (Buizza 1997; Dey et al. 2014; Hopson 2014;

Whitaker and Loughe 1998). Previous studies have found smaller spread-skill discrepancies within CPEs than the corresponding driving ensemble when evaluated over common regions (Cafaro et al. 2021; Clark et al. 2009, 2010; Frogner et al. 2019a; Klasa et al. 2018; Montani et al. 2011). Separately examining each metric typically reveals that improvements in both spread and skill contribute to the better spread-skill relationship. Larger skill is produced through the aforementioned mechanisms. Larger spread is a natural consequence of increasing model resolution which leads to smaller-scale flows being resolved and allows for more detailed topography to be used (Clark et al. 2010; Klasa et al. 2018). Despite these findings, though, CPEs still struggle with providing reasonable spread.

To better contextualise the underspread problem, many studies have examined the sources and growth of spread. In systems like those used at the UK Met Office (UKMO), IC perturbations are provided by the corresponding member of the driving ensemble so as not to introduce discrepancies with the LBCs. These IC perturbations, along with the central ensemble state, dominate spread growth at the start of the integration before differences in the the LBCs take over (Warner et al. 1997). The transition timing between these two sources of spread is highly dependent on the time taken for LBC information to transmit across the domain. For smaller domains, LBCs dominate spread in precipitation-based metrics after 6–12 h (Gebhardt et al. 2011; Kühnlein et al. 2014; Vié et al. 2011; Zhang et al. 2023). For slightly larger domains similar to the size of those used at the UKMO, LBCs typically dominate after 24 h (Hohenegger et al. 2008; Porson et al. 2019). Despite the importance of IC perturbations and LBCs, internally produced spread can still be the dominant source under the right conditions. Model physics and IC perturbations—which can influence the state at smaller scales—can provide a greater spread contribution under weaker synoptic conditions where the interaction of the initial state with local processes is more important for the developing weather (Clark et al. 2010; Flack et al. 2021; Keil et al. 2014; Kühnlein et al. 2014; Vié et al. 2011; Zhang et al. 2023). Conversely, LBC perturbations—which only contain larger-scale information—provide a greater spread contribution under stronger, more mobile synoptic conditions (Klasa et al. 2018; Nielsen and Schumacher 2016).

While much work has been done to improve spread in the CPE itself (Beck et al. 2016; Ben Bouallègue et al. 2013; Gainford et al. 2024; McCabe et al. 2016; Mittermaier 2007; Porson et al. 2019, 2020; Raynaud and Bouttier 2017), relatively little work has been done to explore the culpability of the driving ensemble for the lack of spread. Driving ensembles can exert a strong influence over the evolution of each CPE member through the inheritance of IC perturbations and LBCs. Conceptually, this dynamic is similar to the traditional nature vs nurture debate in psychology, which describes behavioural traits as either being inherited from the parent (nature) or developed in response to a person's environment (nurture). In our case, the "child" ensemble can either inherit spread from the driving ensemble, or can develop its own spread through interaction with the surroundings (e.g., through the upscale growth of convective-scale errors). In the nurture scenario, the source of spread is related to CPE model physics and dynamics, and there is the potential for large differences to emerge between the spread of the global ensemble and the CPE. This study explores which of these scenarios is more appropriate for different leadtimes and under different weather regimes, as well as the connection that these scenarios may have to the spread-skill relationship. In particular, our parent-child analysis concentrates on understanding the spatial spread and similarity of precipitation patterns, which has not been the focus of other studies. Additionally, we perform this analysis on outputs from an operational ensemble rather than an idealised setup, which allows us to understand the behaviour of an ensemble that is regularly used for decision making.

The rest of this paper outlines the ensemble model configurations (Sect. 5.2.1) and evaluation methods (Sect. 4.2.2) used in this study, before introducing a new method, the parent-child FSS, for making direct comparisons between the nested and driving ensemble (Sect. 4.2.3). After this, the characteristics of the trial period used in this study are discussed (Sect. 4.3). The results start by showing comparisons of the spread-skill relationship for each model (Sect. 4.4). Next, the new comparison method is used to show leadtime trends in ensemble behaviour (Sect. 4.5). Finally, the regime dependence using this method is explored in more detail and related to the spread-skill relationship (Sect. 4.6).

4.2 Methods

4.2.1 MOGREPS

For this study, we use data from the Met Office Global and Regional Ensemble Prediction System (MOGREPS): an operational ensemble configuration run at the UKMO comprised of a global ensemble, MOGREPS-G, and a nested ensemble run over the UK, MOGREPS-UK. The use of an operational configuration introduces more complexities between ensembles than using a more direct driving/nested configuration. However, these are also the ensembles that are used for day-to-day decision making, and it is arguably more important to understand the behaviour of these ensembles than it is to understand the behaviour of ensembles that are purely used for research purposes. Indeed, forecasters will have to account for similar considerations when interpreting outputs from any routinely running ensemble, whether it be production delays between the output of the driving and nested ensembles, or

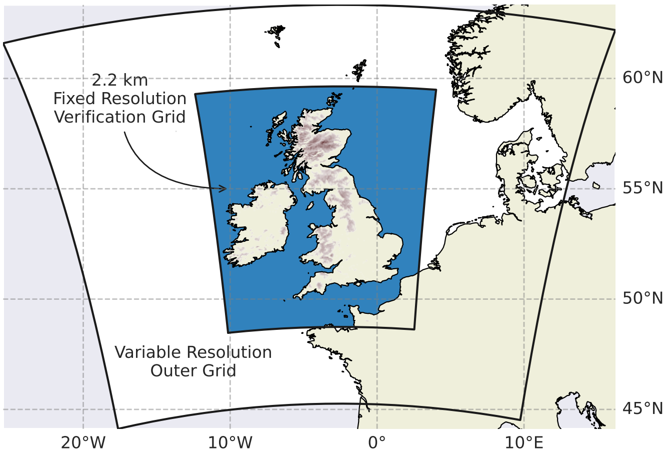

Figure 4.1: MOGREPS-UK domain showing the boundary locations of the variable resolution grid and the inner grid used for verification. Elevation above 200 m is shaded light to dark within the inner grid.

.

configuration differences designed to maximise performance over different periods. In the interests of clarity, we will remind the reader of any configuration differences between ensembles that are relevant for the results being presented.

MOGREPS-G has a grid spacing of approximately 20 km in the midlatitudes and uses a parametrization scheme to represent convection. MOGREPS-G cycles every 6 h at 00Z, 06Z, 12Z, and 18Z producing 17 perturbed members plus a control member from a global analysis, with each perturbed member separately initialised using a hybrid 4D ensemble variational data assimilation system (i.e., each member uses a different background but assimilates the same observations). Each perturbed member is partially recentered on the deterministic analysis (which itself uses the ensemble to estimate flow-dependent uncertainty) before several inflation methods are applied (Inverarity et al. 2023). Model error is represented through additive inflation and stochastic physics schemes. MOGREPS-G outputs three-hourly precipitation accumulations, which we use as the accumulation window for both ensembles throughout this work.

MOGREPS-UK is an 18 member lagged ensemble with 2.2 km grid spacing (Hagelin et al. 2017). An inner, fixed-resolution grid of 421×546 points is surrounded by a variable resolution grid with total size 740×753 points, as shown in figure 4.1. This setup allows convection to spin-up away from the fixed-resolution region that is used for verification. Unlike MOGREPS-G, which is fully coupled to a 0.25 degree ocean model, MOGREPS-UK only represents the atmosphere. Both deep and shallow convection is represented explicitly within MOGREPS-UK. The production schedule of MOGREPS-UK members is shown in Fig. 4.2: MOGREPS-UK cycles every hour producing three new members run out to 120 h which are combined with

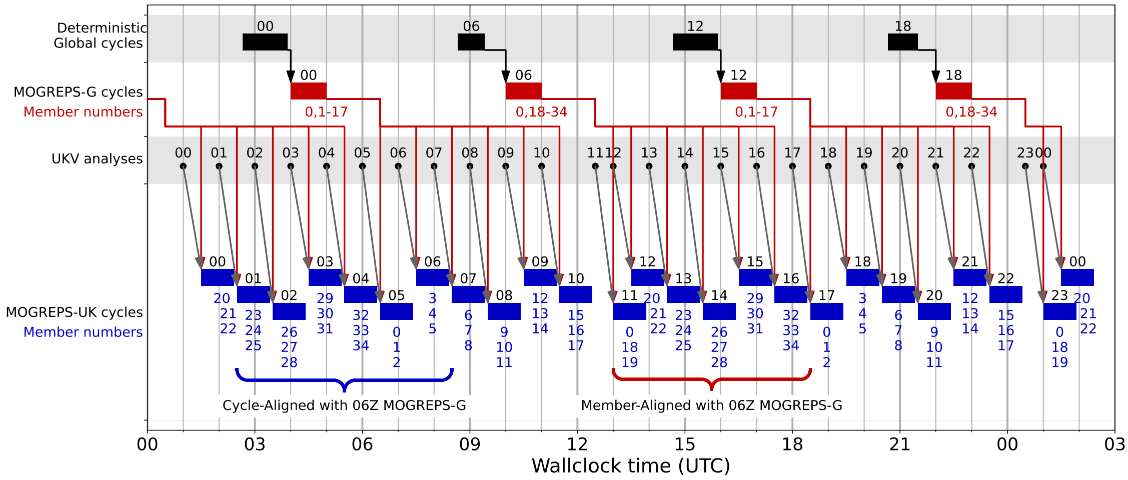

Figure 4.2: Schematic showing the data flow for initialising time-lagged MOGREPS-UK ensemble. Top, black boxes show the six-hourly deterministic global model cycling frequency. The red boxes, arrows and numbers show the MOGREPS-G members that provide initial condition perturbations and lateral boundary conditions. The black dots and grey arrows show the UKV analyses around which a given MOGREPS-UK cycle is centered. 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. An example of the cycle-aligned and member-aligned comparison choices are shown with reference to the MOGREPS-G ensemble initialised at 06Z. The highlighted cycle-aligned MOGREPS-UK members are comprised of the three members initialised at 06Z lagged with the previous five cycles. These comparison choices are explained further in the text. Figure adapted from (Porson et al. 2020), though note that the operational scheduling of these members has changed since publication in 2020.

.

the 15 members from the previous five cycles to produce the full 18-member set (Porson et al. 2020). For brevity and convenience, when referring to MOGREPS-UK leadtimes, we will ignore the different initialisation times between members and instead only quote the leadtime of the members from the most recent initialisation.

Every hour a high resolution analysis with 1.5 km grid spacing is produced over the UK domain using convective-scale 4D-Var data assimilation (Milan et al. 2020). This analysis is used for initialising the deterministic UKV forecast and provides the base state around which the new MOGREPS-UK members are recentered. To produce the three new 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 the high-resolution analysis. These same MOGREPS-G members also provide LBCs to the corresponding MOGREPS-UK members. Due to the production time required for the MOGREPS-UK ensemble, there is a 5–

10 h offset between the generation of perturbations in MOGREPS-G and their use in MOGREPS-UK. For example, the members of the MOGREPS-G ensemble initialised at 06Z are used to drive members of the MOGREPS-UK ensemble initialised hourly between 11Z–16Z, as shown by the red lines in Fig. 4.2.

The timing discrepancy between the generation of MOGREPS-G members and their inheritance by MOGREPS-UK members means there is a misalignment between the ensembles. It is not possible to compare both ensembles with the same sets of members recentered around analyses produced at the same time. Therefore, an appropriate comparison choice must be made for each experiment. If it is more important to ensure that forecasts of the same event are evaluated at the same initialisation times and leadtimes, then the ensembles should be compared using a "cycle-aligned" setup. If, instead, it is more important to ensure forecasts of the same event are evaluated using the same member information (IC perturbations and LBCs), then the ensembles should be compared using a "member-aligned" setup. For illustration, both alignment choices are shown in Fig. 4.2 with reference to the 06Z MOGREPS-G cycle. For member-aligned comparisons, the initialisation times of the MOGREPS-UK ensemble used will be 5–10 h later than that of MOGREPS-G ensemble (11Z–16Z in MOGREPS-UK). To ensure comparisons are made at the same verification time, there must be a corresponding 5–10 h decrease in the leadtimes of the forecasts used from MOGREPS-UK. For consistency with quoting time-lagged leadtimes mentioned earlier, we will refer to this leadtime offset as being 10 h throughout this paper.

While this leadtime offsetting may introduce bias, this study is not the first to intentionally misalign model initialisation times for comparison purposes: Weusthoff et al. 2010 used the latest cycles from each model rather than ensuring leadtime consistency since this is more closely aligned to the data availability schedule for forecasters. We assert that member-aligned comparisons are appropriate as long we maintain awareness of the leadtime biases introduced.

4.2.2 Fractions Skill Score

The Fractions Skill Score (FSS) is a neighbourhood-based metric most commonly used to evaluate the spatial skill of precipitation in high-resolution forecasts (Roberts and Lean 2008). The FSS requires a choice of threshold, which is used to create a binary field based on the exceedance of that threshold, as well as a choice of neighbourhood over which the binary fields are smoothed. The use of a smoothing filter relaxes the constraint that both fields must match at the gridscale, which often leads to overly punitive score penalties for small displacement errors. The comparison of binary fields ensures that the magnitude at each grid point is not taken into account, and thus the FSS only measures the spatial similarity of each input.

A fractions field is produced by calculating the fraction of points in the $n \times n$ neighbourhood surrounding each grid point that exceed the specified criteria. Two fractions fields, A_n and B_n, can be compared to produce the FSS, calculated using the mean-squared difference, MSD, as:

$$\mathrm{MSD}_{(n)}\left(A,B\right) = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left[A_{(n)i,j} - B_{(n)i,j}\right]^2 \ , \tag{4.1}$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A,B\right) = \frac{1}{N_x N_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\left[A^2_{(n)i,j} + B^2_{(n)i,j}\right] \ , \tag{4.2}$$

$$\mathrm{FSS}_{(n)}\left(A,B\right) = 1 - \frac{\mathrm{MSD}_{(n)}\left(A,B\right)}{\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A,B\right)} \ , \tag{4.3}$$

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.

The spatial spread-skill relationship can be estimated by comparing the dispersion FSS (dFSS) to the error FSS (eFSS) (Dey et al. 2014). The dFSS averages the FSS between all member-member pairs of an M-member ensemble to measure ensemble similarity, and is calculated as

$$\mathrm{dFSS}_{(n)} = \frac{1}{M\left(M-1\right)}\sum_{M_a=1}^{M}\sum_{\substack{M_b=1 \\ M_a \neq M_b}}^{M}\mathrm{FSS}_{(n)}\left(M_a, M_b\right) \ , \tag{4.4}$$

where M_a and M_b are the fractions fields for the members being compared. Larger dFSS values mean there is greater similarity between members, and therefore lower spread.

Similarly, the eFSS averages the FSS between each member and an observation field, O, to measure ensemble skill. We use NIMROD radar data (Golding 1998) as our "truth", acknowledging that there are many sources of uncertainty associated with this system, even despite the processing used to create the composite product. As with other studies, we do not account for this observational uncertainty. For comparison with model data, each radar field is interpolated to the corresponding model grid using a nearest-neighbour algorithm that masks extrapolated points. In tests compared to a linear interpolation scheme, the nearest-neighbour algorithm

MOGREPS-G Nbhood	MOGREPS-UK Nbhood
3 × 3 (58.4 km)	25 × 25 (58.2 km)
5 × 5 (97.3 km)	41 × 41 (95.4 km)
7 × 7 (136.2 km)	59 × 59 (137.3 km)
9 × 9 (175.1 km)	75 × 75 (175.5 km)
11 × 11 (214.0 km)	91 × 91 (211.8 km)
13 × 13 (252.9 km)	109 × 109 (253.6 km)

Table 4.1: Neighbourhood sizes in number of grid points and physical side length for MOGREPS-G and MOGREPS-UK.

was judged to produce more representative downscaled fields since it was better at retaining extremes. The eFSS is calculated as

$$\mathrm{eFSS}_{(n)} = \frac{1}{M} \sum_{M_a=1}^{M} \mathrm{FSS}_{(n)}\left(M_a, O\right) \ , \tag{4.5}$$

where larger values means higher skill. A useful spread-skill relationship should show no bias between the eFSS or dFSS (Dey et al. 2014). An ensemble is underspread if the dFSS is consistently larger than the eFSS, or overspread if the dFSS is consistently smaller than the eFSS.

When comparing scores calculated between different ensembles, a choice must be made about whether to first coarse-grain the CPE onto the driving grid or whether to evaluate both models on their native grids. Ultimately, we choose to evaluate on both native and driving model grids since both are useful in different situations: evaluating on the native grid preserves information at the convective-scale, while coarse-graining enables direct comparison between driving-nested member pairs. In the situations where evaluation is performed on the native grids, we chose to compare scores evaluated over similar neighbourhood sizes, as this produced similar dFSS and eFSS trends between the ensembles over a one week sensitivity test. This approach is also most commonly used in other studies (Duc et al. 2013; Hagelin et al. 2017; Weusthoff et al. 2010). The list of neighbourhoods used for these comparisons is shown in Table 4.1.

A further choice must be made when preparing FSS data accumulated across the trial period. Some authors suggest that aggregating the individual MSD and MSD$^{\mathrm{ref}}$ components is most appropriate given the similarity of the underlying methods to a 2 × 2 contingency table (Mittermaier 2021). Other authors suggest that simply averaging each FSS value is more appropriate since this will measure the overall similarity of precipitation patterns each time a forecast is made (Cafaro et al. 2021; Skok 2016). In a sensitivity test of dFSS and eFSS accumulations, we found an expected increase in scores when aggregating components rather than averaging, but ultimately found that both methods produced similar trends. Necker et al. 2024 also found that pooled FSS values were relatively insensitive to the choice of aggregation or averaging. Given the relative ease of use we use averaging for the rest of the study.

The more important factor determining score quality was the forecast selection criteria. Most FSS studies focus on short periods of precipitation which are carefully chosen to produce robust values. It is less common to produce seasonal FSS statistics which will likely include both wet and dry periods. Since the FSS is more sensitive at lower fractional coverages, seasonally-averaged FSS values tend to be biased towards the large fluctuations produced by dry periods. Therefore, we apply a filter to only include FSS values of forecasts with domain average precipitation greater than 0.1 mm/3 h. This threshold was chosen as the smallest value at which the average FSS becomes largely insensitive to further threshold increases, and is similar to the value used in a previous work (Gainford et al. 2024).

4.2.3 Parent-Child FSS (pcFSS)

Analysing the correlation of spread metrics between different ensembles offers a simple yet indirect method of assessing similarity. More direct methods have previously shown clear links between the accuracy of corresponding driving-nested member pairs (Barrett et al. 2016). Here, we introduce a new metric which directly assesses the similarity of precipitation patterns of driving-nested member pairs using the FSS framework.

The Parent-Child FSS (pcFSS) is calculated by:

1. Ensuring both ensembles are evaluated on the same grid (a requirement for any FSS calculation);
2. Ensuring both ensembles use consistent sets of driving information (i.e., each nested member pairs with a corresponding driving member).

To ensure the first requirement is satisfied, we regrid the nested ensemble onto the driving model grid and extract the common domains. To ensure the second requirement is satisfied, we use member-aligned comparisons as described in section 5.2.1. A matrix of pcFSS values is then produced by calculating the FSS between each driving-nested member pair. As with the FSS, scores of unity indicate fields that are identical, while scores of 0 indicate completely dissimilar fields.

The utility of this metric comes from comparing the same-member and different-member elements of the pcFSS matrix, since only the same-member elements have been calculated using driving-nested pairs that share member information. This approach will be the basis of the results shown in section 4.5.

4.2.4 Model Alignment Overview

In the following sections, we will present our studies analysing the FSS over the summer 2023 trial period. In section 4.4 and the second half of section 4.6, we compare the spread-skill relationships between MOGREPS-UK and MOGREPS-G. Since this analysis is only useful if there is no leadtime bias between the two ensembles, these results are produced using a cycle-aligned setup. In section 4.5 and the first half of section 4.6, we present the pcFSS calculated between the two ensembles. This method is only useful if the same members are present in both ensembles, thus it is necessary to perform these calculations using a member-aligned setup. Both alignment methods are compared against each other in Section 1 of the supplementary material.

4.3 Trial Period

This work analyses operational MOGREPS-UK and MOGREPS-G data from forecasts run from 1 June to 31 August 2023. This summer period was chosen due to the overlap with other studies and field campaigns (UKMO 2023b). This period also includes a greater frequency of convective activity compared to climatology, which provides a large sample of events that have the potential to produce broad differences between the two ensembles. Figure 4.3 compares the occurrence of each Decider regime in summer 2023 to its climatalogical baseline. Decider is a probabilistic weather pattern forecasting tool used by the UKMO which clusters the most common pressure patterns observed over the UK and Europe into 30 individual regimes, as well as a reduced set of 8 representative regimes (Neal et al. 2016). These 8 sets are clustered based on the similarity of the individual regimes, and are a useful way of selecting data by weather type: for instance, representative set 5 is most associated with convective activity. The 3 individual regimes shown in the right panel of Fig. 4.3 have previously been identified as the most likely to include thunderstorms (Wilkinson and Neal 2021). For context, regime 11 belongs to representative set 1 (Blocked, negative North-Atlantic Oscillation (NAO-)), and regimes 16 and 22 belong to representative set 5 (high pressure over Scandinavia bringing southerly or south-easterly winds).

The start of June 2023 was largely fine and dry due to a persistent block over

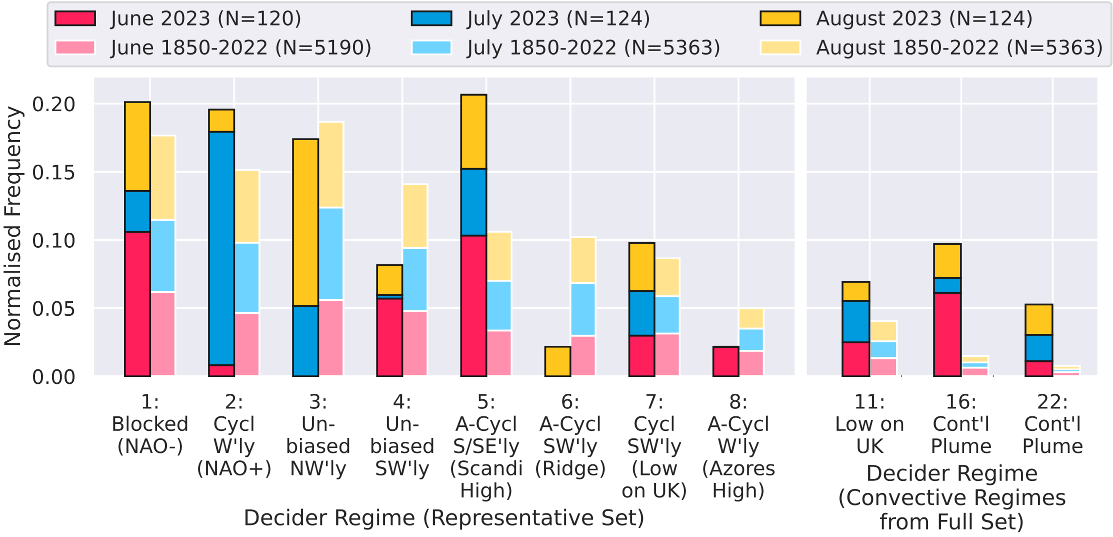

Figure 4.3: Overview of the Decider pressure pattern regime frequency for summer 2023 compared to the 1850-2022 climatology. Each bar is segmented to show monthly variability of regime occurrence. Left panel shows the representative set of 8 regime groups, where “Cycl/A-Cycl” refers to the presence of cyclonic or anti-cyclonic pressure patterns centered over the UK within a majority of the individual regimes of each set. “Unbiased” means that neither cyclonic nor anti-cyclonic pressure patterns are dominant within the set. Right panel shows a selection of three regimes from the full set of 30 which have previously been identified as having the greatest likelihood of thundery activity. “Cont’l Plume” refers to a typical continental plume setup.

the UK. A switch to more unsettled conditions occurred around the middle of June, with warmer and more humid air bringing a large number of thunderstorms into the region. These blocked (set 1) and southerly (set 5) regimes account for most of the month's weather and this was classed as the warmest June on record. In contrast, July 2023 was classed as one of the wettest on record, with a westerly setup bringing a succession of weather systems from the Atlantic. The cyclonic westerly (set 2) regime was dominant for most of July, although some blocked and Scandinavian high periods were also observed. August 2023 was more mixed than June and July, with wet periods interspersed with more settled conditions.

Three named storms occurred during summer 2023. On 4 July, Storm Poly began rapid cyclogenesis off the south-west coast of Ireland and brought large impacts to parts of the Netherlands after travelling through the English channel. In August, Storms Antoni and Betty brought unseasonably wet and windy weather to the UK, although the intensity and impacts were more limited (UKMO 2023a).

4.4 Spread-Skill Relationship Comparison between MOGREPS-UK and MOGREPS-G

Previous studies have shown that CPEs are more likely to have a better spread-skill relationship than those of the corresponding driving ensemble, since spread and skill both separately benefit from the increase in resolution. Here, we investigate this relationship for MOGREPS-UK and MOGREPS-G evaluated over the summer 2023 trial period.

Figure 4.4 shows heatmaps of the differences between 90th centile spread scores and skill scores for both ensembles over a range of neighbourhood sizes and leadtimes. Each value is an average of up to 260 scores from across the trial period (with some variance from the dry-event filtering mentioned in section 4.2.2). Note that these are cycle-aligned comparisons which ensures that each ensemble was evaluated using common initialisation- and leadtimes. Significant differences between spread and skill scores were found from the 95th centile confidence limits estimated through bootstrapping with 10,000 resamples, and are shown by black hatches in the figure.

Overall, both models are underspread (dFSS greater than eFSS) over a large combination of neighbourhoods and leadtimes, however, MOGREPS-UK appears to have a slightly better spread-skill relationship. MOGREPS-UK is not evaluated as significantly underspread over any combination of neighbourhoods or leadtimes, whereas MOGREPS-G is significantly underspread at the earliest and latest leadtimes. Moreover, during the first 15 h when MOGREPS-G is significantly underspread, MOGREPS-UK is close to being correctly spread. This trend is high-

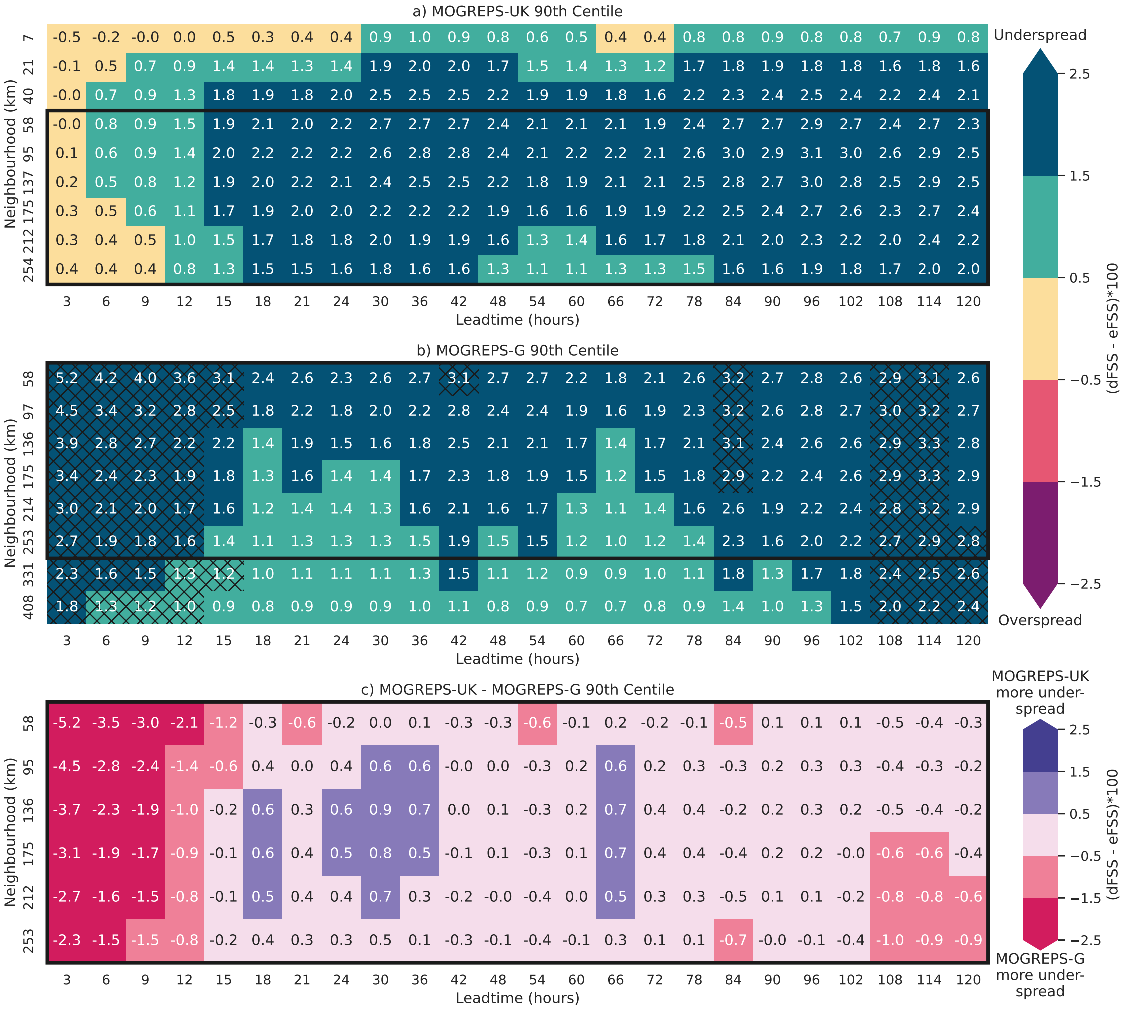

Figure 4.4: Heatmaps of the difference in 90th centile dFSS (spread) and eFSS (skill) scores for a) MOGREPS-UK, b) MOGREPS-G, and c) the MOGREPS-UK – MOGREPS-G difference for common neighbourhoods. Scores are multiplied by 100 for plot clarity. Ensembles were evaluated using a cycle-aligned setup. Values with hatched backgrounds show significant score differences compared to 95% confidence limits estimated by bootstrapping scores over the trial period. The black rectangular outline highlights the neighbourhoods common between the models shown in Table 4.1.

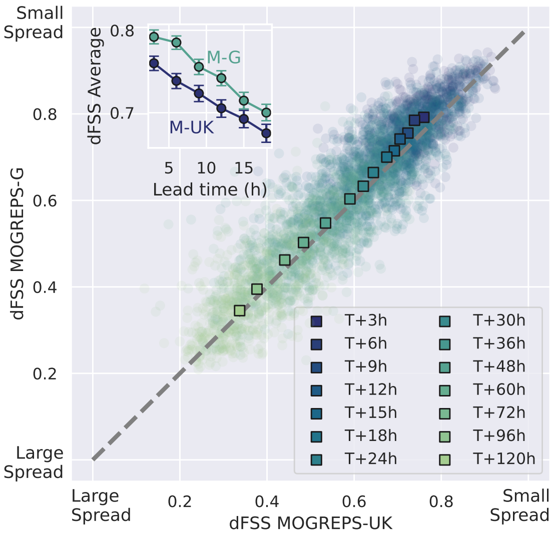

Figure 4.5: Scatter plot of spread (dFSS) scores between MOGREPS-UK and MOGREPS-G using cycle-aligned comparison choice for a range of leadtimes. Faded circles show each event, solid squares show the leadtime average. Scores were calculated using 90th centile and 58 km neighbourhood. Forecasts with equal spread in both ensembles lie on the 1:1 line. Any forecast in the upper-left half of the space is one in which MOGREPS-UK had more spread (lower dFSS values) than MOGREPS-G, and vice versa for forecasts in the lower-right half. Inset shows the average dFSS for each ensemble over the first 18 hours.

lighted by panel 4.4c), which shows the differences between the common neighbourhoods (Table 4.1) bordered in panels 4.4a) and 4.4b). Panel 4.4c) demonstrates that MOGREPS-G is considerably more underspread than MOGREPS-UK in the first 15 h, particularly for smaller neighbourhoods. After this period, however, both models show comparable spread-skill relationships.

There are two potential causes for the substantial discrepancy in the spread-skill relationship within the first 15 h: either the spread of MOGREPS-G ensemble members is smaller than required, or the skill of those solutions is worse than required (or a combination of both). To understand the contribution made by spread to this situation, Fig. 4.5 shows a scatter plot of the dFSS for each event in the trial period with a 58 km neighbourhood and at a selection of leadtimes. The average value for each leadtime is shown by the shaded squares, and highlights the result that MOGREPS-UK has slightly more spread over all leadtimes. This is most apparent at the earliest leadtimes, and particularly for T+6 h. To further highlight this trend, the inset of Fig. 4.5 shows the leadtime series of average dFSS scores for both models in the first 18 h. Confidence limits are estimated by the same bootstrapping analysis used in Fig. 4.4. MOGREPS-G has significantly smaller spread than MOGREPS-UK over these earlier leadtimes, demonstrating that the spread itself is at least part of the reason for the discrepancy in the spread-skill relationship between the two ensembles.

While we do not show it here for concision, a similar examination of eFSS (skill) scores does not reveal a significant difference between ensembles. Therefore, the discrepancy in the early spread-skill relationships of precipitation patterns is solely attributed to the lack of spread in MOGREPS-G, which has been previously demonstrated (Inverarity et al. 2023). We should not be surprised that MOGREPS-UK demonstrates slightly larger spread throughout the longest leadtimes since it is driven using members from multiple MOGREPS-G forecasts. We should also not be surprised that MOGREPS-UK is initialised with better spread given the additional processes used to enhance spread at earlier leadtimes, such as time-lagging. Additionally, the lack of early spread in MOGREPS-G is not necessarily an issue for operational UK weather forecasting since the convective-scale ensemble is typically used for such short-range forecasting. The lack of spread does, however, imply a lack of reliability when used for forecasts over other regions which do not run well-tuned CPEs.

The final point to highlight from Fig 4.4 is the time at which the convective-scale spread transitions from being largely well spread to largely underspread: approximately 15 h. If this transition is being driven by the growing influence of LBCs on ensemble spread, the timing is noticeably earlier than the 24 h timescale seen in other works with similar domain sizes (Hohenegger et al. 2008; Porson et al. 2019). The inset of Fig 4.5 shows that spread scores remain statistically different between the two ensembles until at least T+18 h, indicating that the spread-skill trend observed in Fig 4.4 is driven mostly by subtle changes in skill. Regardless of the exact LBC transition time indicated in these figures, we assert there is little need to precisely define this transition time for discussing the impacts of LBCs on spread, provided we use leadtimes that are sufficiently far from this transition period. Either way, it is clear that spread is worse in the convective-scale ensemble when LBCs dominate the evolution of each member.

Ultimately, while the methods presented here have been successful at discriminating between the IC and LBC spread regimes, they are indirect methods of evaluating similarity, and the conclusions are somewhat muddied by the complexity of these ensemble configurations. Therefore, to complement this analysis, we have produced a new metric which provides a more direct and intuitive approach to evaluate the similarity of these models while still using the FSS framework.

4.5 Parent-Child FSS Study

As described in section 4.2.3, the Parent-Child FSS (pcFSS) is a method for comparing the similarity of precipitation patterns between pairs of members in different

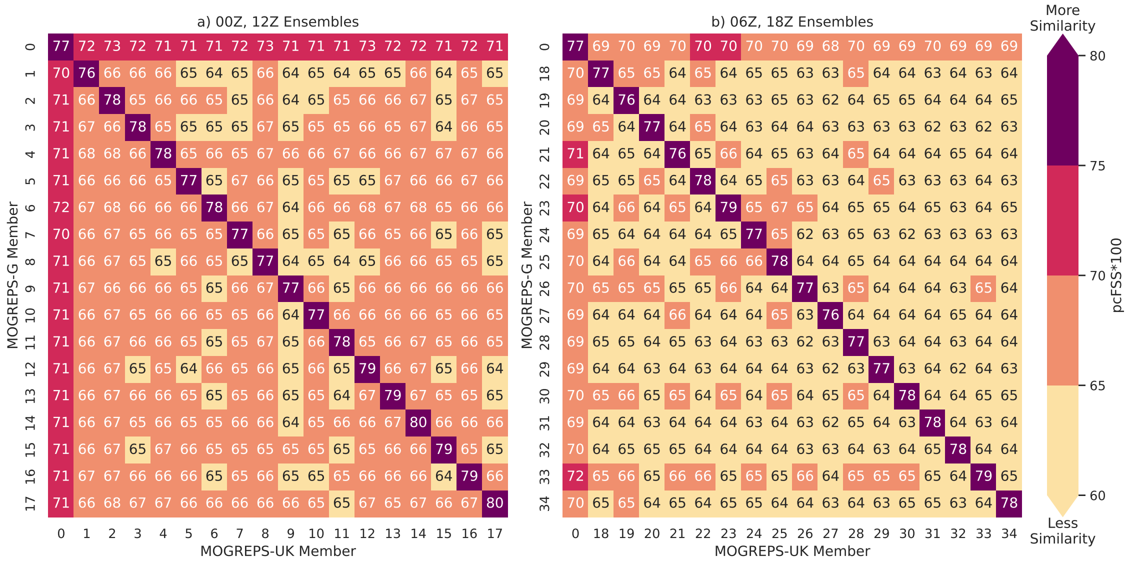

Figure 4.6: Heatmap of average pcFSS values between each MOGREPS-UK/MOGREPS-G member pair. Scores were calculated with 90th centile, 58 km neighbourhood and leadtime T+15 h in MOGREPS-G (T+5 h in MOGREPS-UK). The figure is split into two panels for the two alternating sets of driving MOGREPS-G members used every 6 h.

ensembles. In this section, we will investigate the behaviour of this metric over the full summer 2023 trial period.

Figure 4.6 shows the average score between each member-member pair throughout the trial period. These calculations were performed using the 90th centile and results are shown for the 58 km neighbourhood, although the trends are consistent across all neighbourhood sizes considered. This plot shows only the data for a leadtime of T+15 h in MOGREPS-G, and since the pcFSS requires member-aligned comparisons, this equates to a leadtime of T+5 h in MOGREPS-UK. This leadtime was chosen to show the behaviour before LBCs dominate in the CPE.

The diagonal elements show larger similarities than any other pairs of members, highlighting the important role that the driving ensemble plays in positioning precipitation in corresponding nested members. The next most similar pairings are the control members since these are the base state around which large-scale initial perturbations are calculated. This interpretation is supported by clustering work showing that the control member is the most likely to be the central member within the MOGREPS-UK distribution (unshown). There are no other obvious patterns within the remaining off-diagonal elements. In fact, this consistency is a good demonstration that the each perturbed member is considered an equally likely realisation, i.e., they are exchangeable (Bröcker and Kantz 2011). However, the scores for the forecasts starting at 00Z and 12Z are slightly larger than for the forecasts starting at 06Z and 18Z. This difference is due to a small diurnal variation in

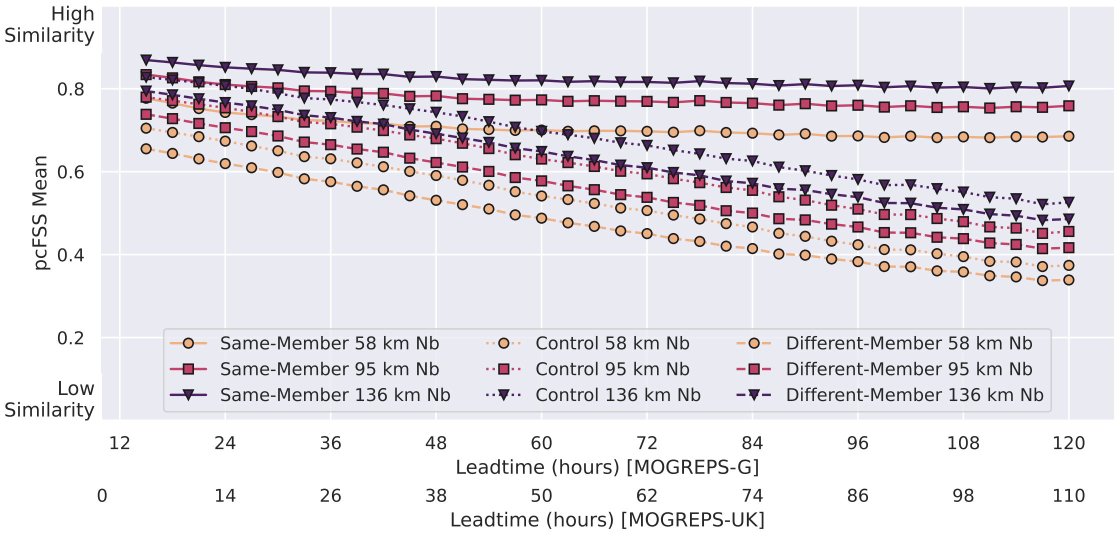

Figure 4.7: Leadtime series of average pcFSS scores for same-member, control and different-member sets for the three smallest common neighbourhoods considered in this work (see Table 4.1). Scores were calculated using 90th centile. Two x-axes are included as a reminder that member-aligned comparisons require a 10 h offset between MOGREPS-UK and MOGREPS-G initialisation times.

the FSS, whereby precipitation patterns are evaluated as being more similar during periods of larger coverage. At a leadtime of 15 h, this effect preferentially impacts the forecasts initialised at 00Z since convection will be most likely at these validity times (mid-afternoon local time). This diurnal variation is mostly mitigated when using a centile threshold for the FSS rather than an absolute threshold.

To summarise, the pcFSS has identified three sets of paired members with distinct behaviours:

1. Same-member pcFSS – diagonal elements;
2. Control-perturbed pcFSS – pairs of members including a control member and a perturbed member from either MOGREPS-UK or MOGREPS-G (first row and column excluding upper left);
3. Different-member pcFSS – All off-diagonal elements (including those from the control set mentioned above).

The broader utility of the pcFSS comes from analysing the behaviour of these sets, especially the same-member and different-member sets. Figure 4.7 shows the leadtime pcFSS series for the three smallest neighbourhoods used in this work. While the different-member and control-perturbed sets become increasingly dissimilar with leadtime, the same-member sets asymptote towards a constant similarity. The behaviour of the same-member sets is expected from parent-child pairs that share information. At the earliest leadtimes, the similarity between all member-member pairs decreases as IC differences and model dynamics drive differing evolutions between members. Eventually, information arriving through the boundaries dominates the evolution of the nested members, and the similarity of same-member scores stabilises as a result. Meanwhile, the different-member and control-perturbed sets continue to become less similar with leadtime due to the increasing spread within both ensembles. These sets are both well-fit by exponential decay curves with e-folding time of 6.6×10^{-3} h.

The same-member sets do not fit an exponential profile in the same way that the different-member sets do. Instead, a brief period of exponential decay followed by a linear relation with a very shallow negative gradient appears to be the optimal fit. This behaviour is further suggestive of the existence of two different spread growth regimes which dominate at different leadtimes. While it is difficult to determine the exact timing of the transition between these regimes, visual inspection places this time between T+18-30 h, which is consistent with previous studies and the results from Fig 4.5. We anticipate that more data would provide stronger bounds on this transition time.

It is also worth noting that even at the leadtimes when we expect the boundaries to dominate the evolution of each member, the same-member pcFSS is always less than one, and in fact is never larger than at the earliest leadtimes. One interpretation is that this represents the additional convective-scale detail that is spun-up within the variable resolution grid before being advected into the inner, fixed-resolution region used for verification. This explanation is consistent with the additional CPE spread shown in Fig. 4.5.

Given that the leadtime series presented here are just averages, there will be periods when the nested ensemble shows more and less similarity to the driving ensemble, and we expect that these periods will be linked to the forecast weather. The ability of the pcFSS to show degeneracy and divergence within the ensemble pair is explored further through case study analysis in Section 2 of the supplementary material. However, we can also learn more about the behaviour of these ensembles by analysing the average pcFSS response under different weather regimes.

4.6 Regime Study

Arguably the most significant benefit from running a model at the convective-scale is the improved representation of convection. Global models typically rely on subgrid parametrizations to estimate rainfall totals from convective storms, whereas convective-scale models can partly resolve the features involved in producing these storms. We therefore hypothesise that the differing representations of these smaller-scale convective features will lead to less parent-child similarity. The Scandinavian

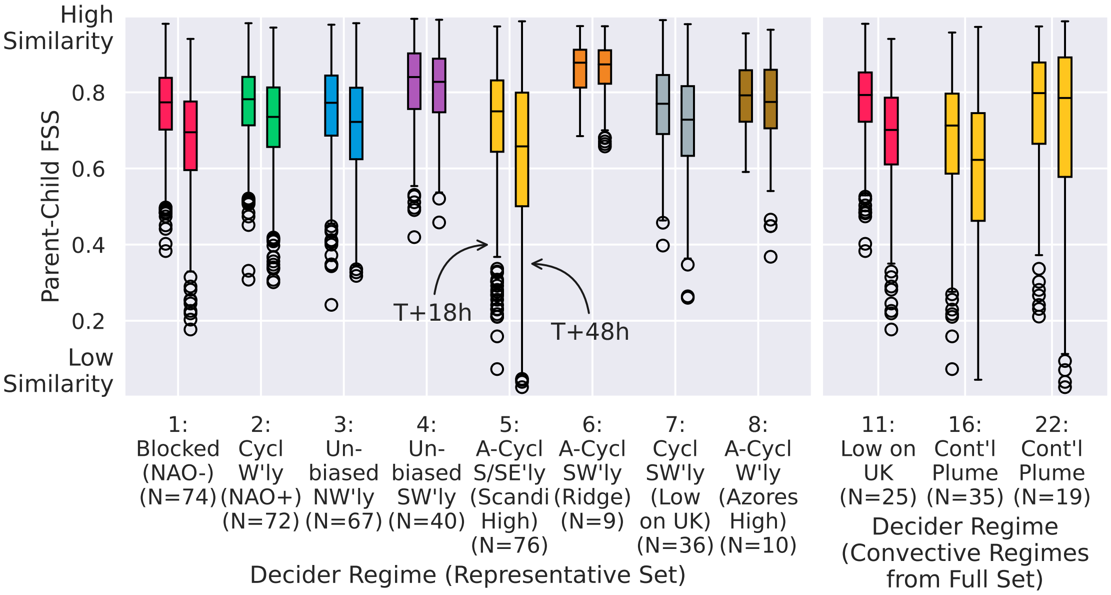

Figure 4.8: Same-member Parent-Child FSS distributions partitioned using the Decider classifications shown in Fig. 4.3. Values were calculated with 90th centile and 58 km neighbourhood. Distributions are shown for MOGREPS-G (MOGREPS-UK) leadtimes of T+18 h (T+8 h) and T+48 h (T+38 h), showing the periods before and after LBC forcings become the dominant spread mechanism in the convective-scale ensemble. Boxes represent upper and lower quartiles, middle line represents the median, and the whiskers either represent these quartiles extended by the inter-quartile range or the limit of the data, depending on which is closer to the median. Data outlying the whiskers are marked by open circles.

high Decider regime is most associated with convection, where southerly winds bring air with high wet-bulb potential temperature into the MOGREPS-UK domain. The summer 2023 trial period included a far greater frequency of Scandinavian high regimes than expected from climatology (see Fig. 4.3), which provides a more than sufficient sample. In this section, we investigate the regime dependence of pcFSS as well as any connection these regimes may have to the spread-skill relationship.

Here, we are only interested in the pcFSS between members which share common forcings. Therefore, we do not consider the control-perturbed or different-member pcFSS in this section, since their contribution to the analysis is less clear. Figure 4.8 shows the same-member pcFSS distributions categorised by different regime sets. Most distributions presented in this figure are robust, only regimes 6 and 8 of the representative set include a low number of events. Of the other regimes, the distributions for sets 1, 2, 3, and 7 are all comparable.

Two sets of distributions are presented for each regime showing the change as LBC forcings become more important. For both leadtimes presented, set 4 (unbiased southwesterly) shows larger overall similarity, which suggests a stronger influence from the driving ensemble on the convective-scale spread. This result is not surprising given that this set is comprised of highly mobile regimes that direct weather systems arriving from the Atlantic into the domain. We may also have expected to see this trend in set 2 (NAO+) which is commonly associated with strong westerlies, but this set is balanced by a number of regimes with weaker flows.

As expected, the Scandinavian high set (set 5) shows the broadest distributions of all regimes and for both leadtimes. Not only does this regime set include smaller lower whiskers and medians, there are more scores which lie outside the whisker range than for any other set. In fact, it is clear that most of these outliers are the same as those shown for regimes 16 and 22 in the right plot, meaning that the most extreme pcFSS events are being driven by regimes most associated with convective activity. Clearly, then, the continental plume setups have the greatest potential to provide a different evolution between the nested and driving ensembles.

To determine the impact of the ICs on pcFSS results, it is also instructive to compare the differences in distributions before and after LBCs become the dominant source of convective-scale spread. Most distributions show a shift towards lower pcFSS at the longer leadtime, consistent with Figure 4.7. Noticeably, however, the distributions for regime set 4 remain largely identical for both leadtimes, which reaffirms that convective-scale processes add little to the forecast under these strong synoptic conditions.

While the mobile regime shows little leadtime dependence, the Scandinavian high set 5 shows a much starker contrast between distributions. At T+48 h, the median pcFSS is approximately as small as the lower quartile at T+18 h, a pattern only seen

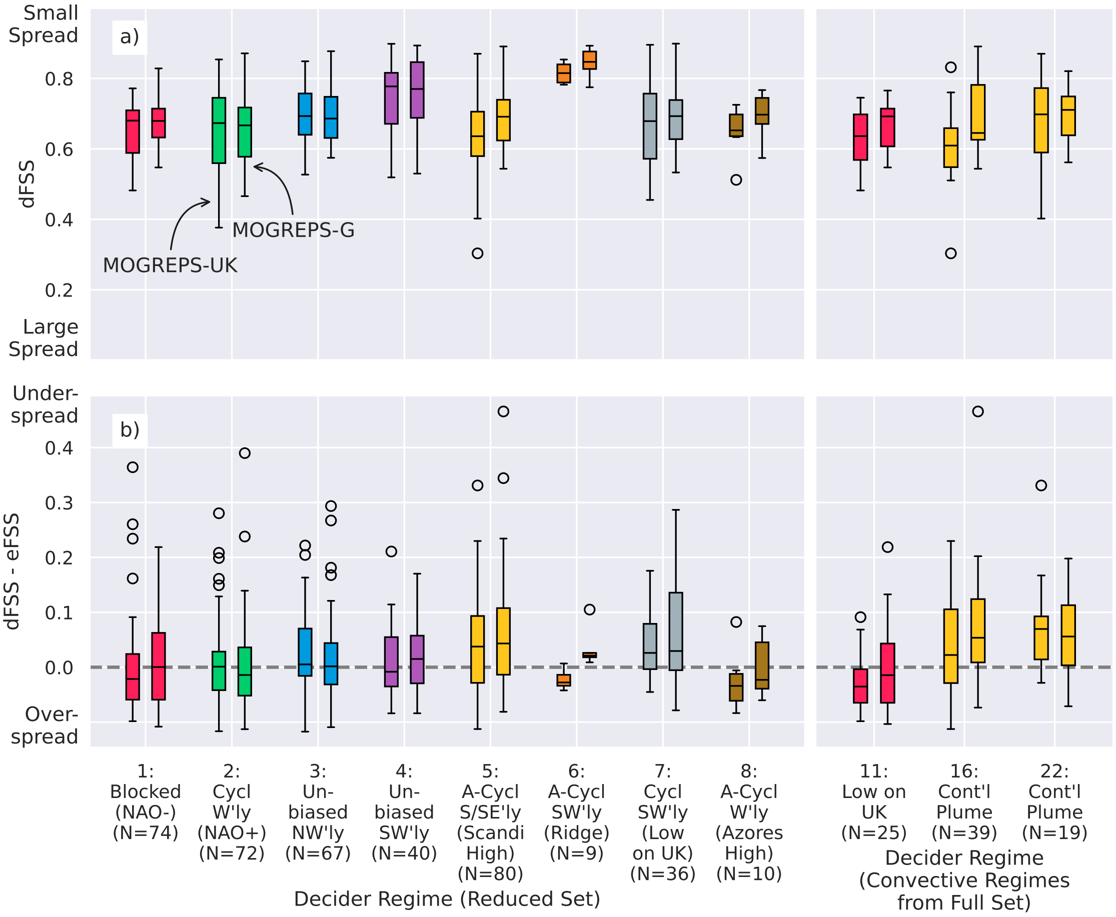

Figure 4.9: As with Fig 4.8 but for a) dFSS and b) dFSS–eFSS, and with panels showing data for T+18 h in both ensembles. Left boxes show distributions for MOGREPS-UK, right boxes show MOGREPS-G.

in one other regime set (set 1). The whiskers and outliers are also noticeably smaller at T+48 h, with some events approaching null scores and a complete mismatch over all members. Of course, at these leadtimes, there is no guarantee that the forecasts of convection produced by the nested ensemble will be more accurate or reliable than those of the global ensemble. However, this result does demonstrate that the LBCs are not as dominant for producing precipitation forecasts within this regime even when they are the source of synoptic-scale information.

Finally, to understand whether the lower similarity between driving and nested ensembles under convective regimes is connected to their spread-skill relationships, Fig. 4.9 shows a) the dFSS and b) the dFSS–eFSS at T+18 h categorised by the same regime sets as the pcFSS in Fig. 4.8. The dFSS data shows that MOGREPS-UK has marginally more spread across all regimes than MOGREPS-G, but that both ensembles have the most spread when forecasting regimes from representative sets 2, 5, and 7. The clearest difference between the spread of the two ensembles is observed for the Scandinavian High set 5, with lower dFSS extremes being over 0.2 smaller in MOGREPS-UK than MOGREPS-G. These differences are even more noticeable in the distributions for the convective regimes, with regimes 16 and 22 showing particularly strong differences. The additional spread displayed by MOGREPS-UK under convective setups is not surprising given the greater potential for upscale error growth.

As well as the clear spread implications, the connection between convective regimes and the spread-skill relationship in Fig. 4.9b) shows that Scandinavian high and convective regimes are also slightly more likely to be underspread than other regimes. Median values for these regimes are well above zero in both ensembles, despite being close to zero for most other regimes. There is also little difference between the two ensembles: MOGREPS-G may have slightly more outliers than MOGREPS-UK in both the Scandinavian high and convective regimes, but the rest of the distribution appears relatively consistent. However, regime set 7 also appears to be just as underspread as set 5, showing that the underspread problem is not solely attributed to convective setups.

Overall, the Scandinavian high, and particularly the continental plume subsets of this regime, are more likely to produce lower pcFSS scores, and thus larger disparities between ensemble forecasts. These sets are also associated with larger spread in the convective-scale ensemble compared to both the driving ensemble and compared to other sets. There is also a noticeable, though slightly less convincing, signal that these convective regimes are more underspread than others, indicating that this increased spread is still not large enough compared to verification. In contrast, the more mobile regime (set 4) shows larger pcFSS than other regimes, and therefore a stronger dependence on the driving ensemble to provide spread. This set shows no clear difference in spread or spread-skill relationships between the two ensembles, as expected given the larger similarity.

4.7 Discussion and Conclusions

In this study, we analysed the similarity of precipitation patterns between a nested, convection-permitting ensemble (CPE) and its driving ensemble over their common domains to identify the situations (leadtimes and regimes) when CPE spread is simply inherited from the driving ensemble and the situations when the ensembles diverge. Conceptually, this is similar to the nature vs nurture debate in psychology, which describes behavioural traits as either being inherited from the parent or developed in response to a person's environment. In particular, we assess the spatial similarity of precipitation patterns using the FSS framework as this has not been the focus of previous studies considering the role of the driving ensemble on CPE spread. Additionally, our focus on operational ensembles allows us to evaluate the performance of models that are regularly used for decision making, but this requires us to implement constraints during inter-comparison. These constraints account for the production delay between global and nested ensembles, meaning we must choose between leadtime consistency or forcing consistency depending on which factor is more appropriate.

We find that CPE spread is least influenced by the driving ensemble in two situations: during short leadtimes (Figs 4.5 and 4.7), and when forecasting convective events driven by continental plume setups (Fig 4.8). In both cases, the Parent-Child FSS (pcFSS) was used to make these assessments by directly comparing the similarity of precipitation patterns between the two ensembles. The pcFSS showed consistently lower distributions for convective regimes than any other regime. Given the plethora of studies showing the differing (i.e., more skillful) representation of convection within convection-permitting models (e.g. Barrett et al. 2016; Cafaro et al. 2019; Schwartz et al. 2010), we should expect to see stronger divergence from the driving ensemble under these conditions.

Meanwhile, by using the same-member pcFSS to identify two distinct periods of ensemble behaviour, we have shown that CPEs exhibit an initial period of growing disparity from the driving ensemble before the similarity between them stabilises (Fig 4.7). From inspection, this transition occurs between T+18-30 h, the period that we expect the lateral boundary conditions (LBCs) to become the dominant source of spread for a CPE domain of this size (Hohenegger et al. 2008; Porson et al. 2019). While other results presented in this work could also be used to infer an LBC transition time (e.g., Figs 4.5 and S1), the pcFSS provides the most direct measure. More data would allow us to better constrain the transition time using the pcFSS.

By linking these findings to the spread-skill relationship (Figs 4.4, 4.5, 4.9), this study also provides important context for tackling the underspread problem faced by CPEs. Spread is larger in the CPE than the driving ensemble when the driving ensemble influence is weakest (at early leadtimes and during convective events). This is consistent with the results of Clark et al. 2010 given their study used a larger nested domain and only considered leadtimes up to 33 h. Our results also demonstrate that the short-leadtime spread-skill relationship is better in the CPE than in the driving ensemble (Fig 4.4), reflecting the efficacy of the time-lagging methods used to generate additional spread (Porson et al. 2020). However, during forecasts of convective events, the increase in CPE spread is still not large enough to offset the lower skill found on verification, leading to forecasts which are similarly underspread in both ensembles (Fig 4.9).

In contrast to the above, the scenarios where the driving ensemble has a strong influence over the CPE do not show consistent differences in spread between ensembles. We find that this occurs at medium-to-long leadtimes (Fig 4.4) and under mobile regimes (Fig 4.9), which is consistent with the boundary arguments mentioned previously. Note, however, that our designation of mobile regimes is based on a single group of Decider regimes identified as having consistently stronger pressure patterns (representative set 4). A more thorough examination of regime mobility would be needed to confirm these results.

As well as understanding leadtime behaviour, case study analysis presented in Section 2 of the supplementary materials shows that the pcFSS could be a useful day-to-day guidance production tool for signalling periods of divergence and degeneracy within the ensemble pair. However, care must be taken when interpreting low pcFSS events, as model biases can contribute to forecast differences.

To conclude, we have shown that the pcFSS can be a complimentary tool for understanding CPE behaviour by highlighting instances of stronger and weaker driving-ensemble influence. If the driving ensemble is underdispersive, we should expect any ensembles nested within this to also be underdispersive once this information has propagated into the CPE domain. Within the current MOGREPS configuration, convective-scale ensemble spread is of a much higher quality during early leadtimes, when the additional spread introduced through time-lagging is still present within the model. While the variable resolution grid surrounding the inner verification region is useful at spinning up additional detail arriving from the driving ensemble, its impact on the spread is limited to those situations when convection occurs and this error growth can upscale faster. Even then, the spread-skill relationship remains similarly poor in both ensembles, indicating a lack of spatial skill. These conclusions provide a number of avenues for finding spread improvements in the nested ensemble. Using a larger nested domain, for instance, would likely prolong the benefits of the internally produced spread. Alternatively, implementing a boundary perturbation scheme may help offset the underdispersion of information arriving from the driving ensemble, provided the scheme was consistent in leadtime and could be applied without shocking the model.

It is clear that the exact nesting configuration used has a strong impact on the spread-skill relationship. This factor is especially important to consider given recent experiments with hectometric-scale ensembles which are themselves nested within convective-scale ensembles (Barrett et al. 2021; Blunn et al. 2024; Hanley and Lean 2024). With multiple layers of nesting, the coupling between each ensemble layer is likely to be complex. Our study provides the foundation for investigating these complex interactions in more detail.

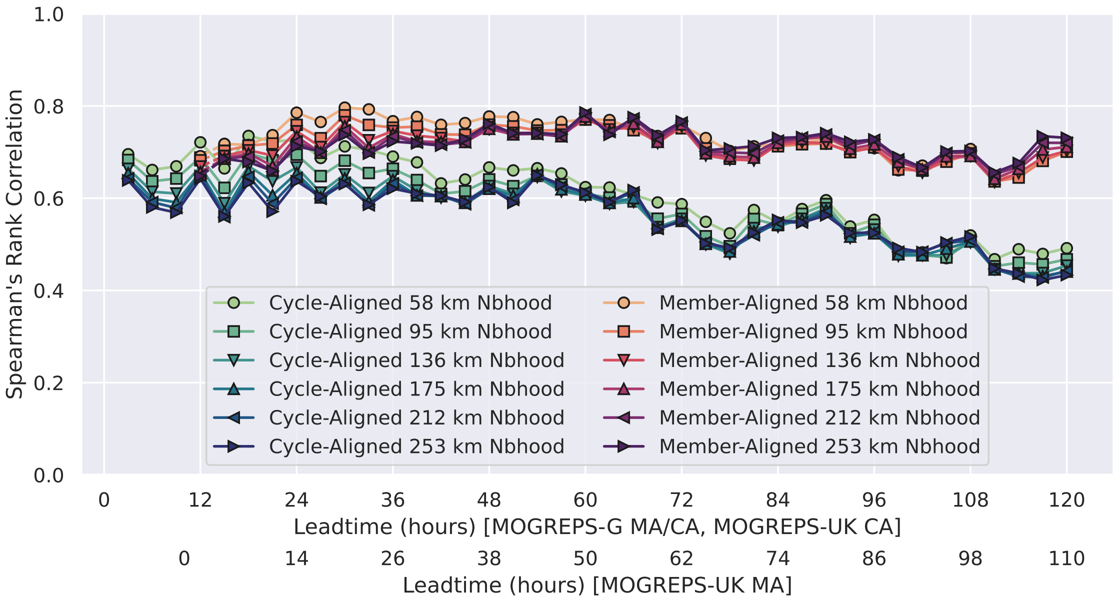

Figure 4.10: Leadtime series of Spearman's rank correlation coefficients between MOGREPS-UK and MOGREPS-G spread (dFSS) scores for both cycle-aligned (CA) and member-aligned (MA) comparison choices. Scores were calculated using 90th centile. Two x-axes are included as a reminder that member-aligned comparisons require a 10 h leadtime offset between MOGREPS-UK and MOGREPS-G initialisation times.

4.8 Supplement

In the main text, we used cycle-aligned comparisons to evaluate the spread-skill relationship of each ensemble, while member-aligned comparisons were reserved for calculating the parent-child FSS (pcFSS). In Section 1 of this supplement, we directly compare cycle-aligned and member-aligned statistics as a means to understand which method provides stronger correlations between ensembles at different leadtimes. Then in Section 2, we demonstrate the utility of the pcFSS to highlight periods of stronger and weaker ensemble agreement, which can be helpful for decision making.

4.8.1 Cycle- and Member-Aligned Spread Comparison between MOGREPS-UK and MOGREPS-G

When evaluating outputs between operational driving and nested ensembles, the appropriate choice of comparison is not always clear given the production delays that exist between the two. Is it more appropriate to compare ensembles with analyses valid at the same time ("cycle-aligned") or to compare ensembles with the same sets of member information ("member-aligned")? This choice is especially important when undertaking subjective evaluations of ensemble performance as this may influence the perceived value of higher resolution outputs. Therefore, in this section, we present comparisons between cycle-aligned and member-aligned ensemble statistics to help provide some guidance about the most appropriate framework for use at different leadtimes.

In particular, we calculate the Spearman's Rank correlation coefficient between MOGREPS-UK and MOGREPS-G dFSS (spread) scores. Other methods of estimating correlation—such as the coefficient of determination from linear regression—show similar trends. Correlations of unity mean that MOGREPS-G dFSS values are perfectly monotonically related to MOGREPS-UK dFSS values, while correlations of zero mean there is no rank relationship between the two dFSS sets. Since both the MOGREPS-UK and MOGREPS-G dFSS datasets are large, we can broaden this interpretation to state that larger correlations indicate better predictors of spread (i.e., it is easier to predict MOGREPS-UK spread given knowledge of MOGREPS-G spread). While some variance in spread is expected given the different grids and model configurations (as demonstrated in Fig 5 of the main text), we also expect that the specific dFSS values in each ensemble will be most sensitive to the forecast weather. As such, the differences in correlations between the two alignment methods indicate the relative importance of the parameter that is held consistent within that alignment. If the cycle-aligned comparison shows a stronger correlation than the member-aligned comparison, the common information in the initial states of both models (i.e., the large-scale initial conditions) is more important in determining spread than the member information. Conversely, if the member-aligned comparison shows a stronger correlation, the shared information between common sets of members is more important in determining spread.

Figure 4.10 shows the leadtime series of correlation coefficients between MOGREPS-UK and MOGREPS-G dFSS values for cycle-aligned and member-aligned comparisons over the set of six common neighbourhoods shown in Table 1 of the main text. The earliest member-aligned scores were evaluated at T+2 h in MOGREPS-UK (T+12 h in MOGREPS-G), and only include 12 members since the latest 6 members have not run for long enough to produce a three-hour accumulation. The statistics at all other times were produced with the full 18 member sets. Two distinct regimes are clear from inspecting Fig. 4.10. Over much of the leadtime range, member-aligned correlations are stronger than cycle-aligned correlations, with the difference becoming more substantial at later leadtimes. This difference reflects the decreasing relevance that the leadtime offset imposes on member-aligned correlations as leadtime increases, and highlights the strong connection between members with consistent boundary information. At earlier leadtimes, though, the correlations are broadly similar between each alignment choice. The fact that the member information is less important at early leadtimes may sound counter-intuitive given that the initial spread of most nested ensembles is determined largely by the member perturbations inherited from the driving model. However, MOGREPS-UK is a time-lagged ensemble which uses analyses over a range of times to recentre its members, and this contributes strongly to its initial spread (Porson et al. 2020).

Additionally, it may seem odd that scores from larger neighbourhoods tend to show smaller correlations than those from smaller neighbourhoods, given that the neighbourhood size controls the amount of smoothing. To understand this behaviour, refer to Fig. 5 in the main text and note the fact that the dFSS is bounded between 0 and 1. Larger neighbourhoods mean larger similarity between different members, and larger dFSS as a result. These larger scores are more likely to be confined to the upper right part of the space than the scores for smaller neighbourhoods, which makes the data less monotonic and produces smaller rank correlations. We therefore find that it is not instructive to draw conclusions from comparisons between different neighbourhood sizes. The same behaviour occurs for dFSS produced at shorter leadtimes, hence we will also not attempt to interpret the leadtime trends of these correlations. We only use this data to compare the relative correlation between cycle-aligned and member-aligned statistics.

To summarise, comparing ensembles with common forcings produces consistently stronger correlations at later leadtimes than comparing ensembles with common analysis validity times. This is expected when LBC information is the dominant source of convection-permitting ensemble spread. Therefore, member-aligned comparisons are likely to be more appropriate at these extended periods. Within the first day, however, both alignment options produce similar spread correlations, meaning that either method could be appropriate and the choice should depend on the specific needs of the user.

4.8.2 pcFSS Case Study

In the main text, we showed the utility of the parent-child FSS (pcFSS) for understanding the relationship between driving and nested ensembles at different leadtimes and under different regimes. However, these composites mask day-to-day score variability which signal periods of divergence and degeneracy within the ensemble pair. By highlighting periods when the driving-nested relationship is stronger or weaker, the user is able to quickly conclude whether the nested ensemble forecast is providing additional value in the placement of precipitation compared to the driving ensemble. Allowing the user to focus their attention on these periods can be incredibly useful given the vast quantities of information available for decision making. Therefore, we posit that the pcFSS could be a valuable operational tool for streamlining guidance production. In this supplement, we demonstrate this addi-

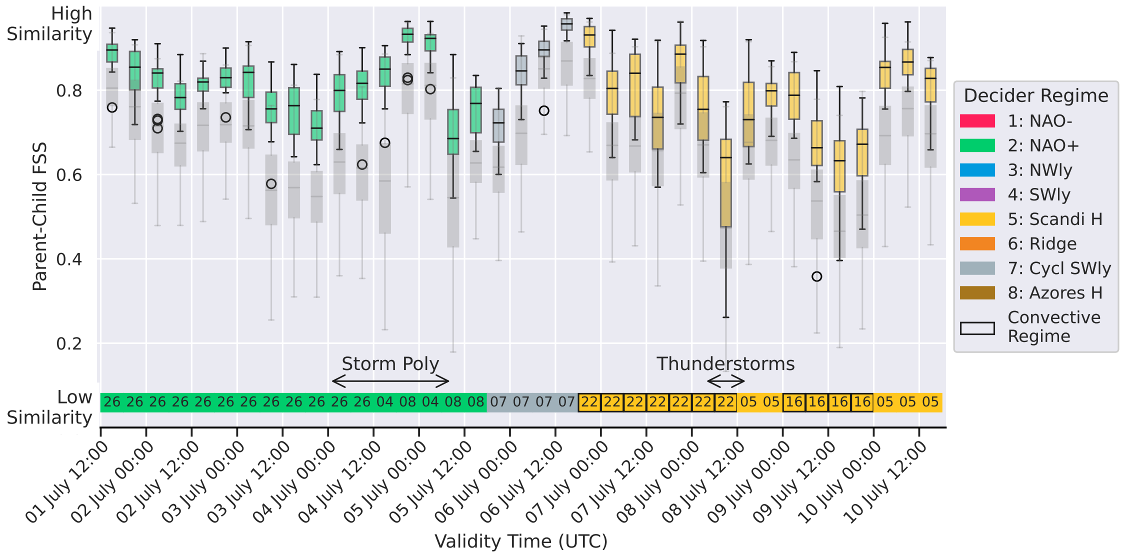

Figure 4.11: Time series of same-member (colours) and different-member (grey) pcFSS distributions for a nine-day period in July 2023. Scores were calculated using 90th centile, 58 km neighbourhood and leadtime T+15 h in MOGREPS-G (T+5 h in MOGREPS-UK). Boxes represent upper and lower quartiles, whiskers represent either these quartiles extended by the inter-quartile range or the limit of the data depending on which is closer to the median. Data outlying the whiskers are marked by circles. The Decider regimes for each six hour period are displayed by the number at the bottom of the plot, and the group each regime belongs to is indicated by the legend.

tional utility by examining pcFSS data over a selected case study period and show that the trends can be linked to the forecast weather.

Figure 4.11 shows a nine-day time series of pcFSS scores from the start of July 2023. These scores were produced using the 90th centile and 58 km neighbourhood, consistent with the values in the main text. The T+15 h leadtime was chosen to highlight a period when we expect the boundary information will not have had time to become the dominant source of spread. Additionally, these scores were calculated using a member-aligned setup, consistent with the other pcFSS studies in the main text. Filled, full opacity boxes show the distribution of same-member pcFSS while the grey boxes show different-member pcFSS. Given the larger sample size of different-member pcFSS compared to same-member pcFSS at any given time, we generally expect to see wider distributions for the different-member pcFSS sets.

For reference, the typical scenarios signalled by the pcFSS are:

- If the same-member pcFSS has a narrow distribution centered on large values, alongside
  - different-member pcFSS with similar distribution – all driving and nested members are similar to each other, and spread in both ensembles is low. Therefore, large similarities between corresponding pairs of nested and

    driving members are likely to be caused by inherent predictability of forecast weather.
  - different-member pcFSS with a broad distribution and lower scores – corresponding driving and nested members are similar despite large differences between different-member pairs. Large similarities are caused by driving members having a strong influence over nested members, i.e., the nested ensemble is filling in the details of the driving ensemble rather than providing radically different outcomes.
- If same-member pcFSS distribution is broad or centred on lower values – there is likely to be large spread in both ensembles and large differences between correlated driving-nested member pairs.

This case study period starts with westerly winds bringing a succession of weather systems to northern areas of the UK. Scores remain relatively stable over the first couple of days, although the progression of a thundery system on 3 July causes a slight drop in both same- and different-member scores. In the early hours of 4 July, Storm Poly begins rapid cyclogenesis in the south west of the MOGREPS-UK domain. This storm continues to intensify as it moves through the English channel until it makes landfall in the Netherlands, bringing severe impacts to both coastal and inland areas (ECMWF 2023). At 15Z, Storm Poly is bringing heavy precipitation to southern areas of England, and this is reflected in the pcFSS trend. Compared to the previous pcFSS data, the different-member distribution is quite broad, indicating large ensemble spread. Meanwhile, the same-member pcFSS data shows a slight upward trend, suggesting that precipitation in MOGREPS-UK is becoming increasingly dominated by the large-scale features which are also represented in MOGREPS-G. At 21Z, the storm is centered off the southeast coast of England and the low pressure centre has become well-developed. At this leadtime, there is now much less uncertainty and larger agreement between the ensembles. The rain associated with Storm Poly exits the verification region in the morning of 5 July and is replaced by a less widespread showery regime, explaining the sudden drop in scores.

After Storm Poly clears, showery precipitation with lower scores gives way to frontal, higher-similarity precipitation, and back again. A transition to the Scandinavian high regime brings a change in wind direction from westerly to southerly/southeasterly. This regime is more favourable for the development of convection over the UK, and indeed, the same-member pcFSS scores reflect the reduced similarity between the ensembles during these times.

Most noticeably, same-member pcFSS scores are far more variable at 09Z 8 July than at any other time. Met Office weather warnings for thunderstorms were in

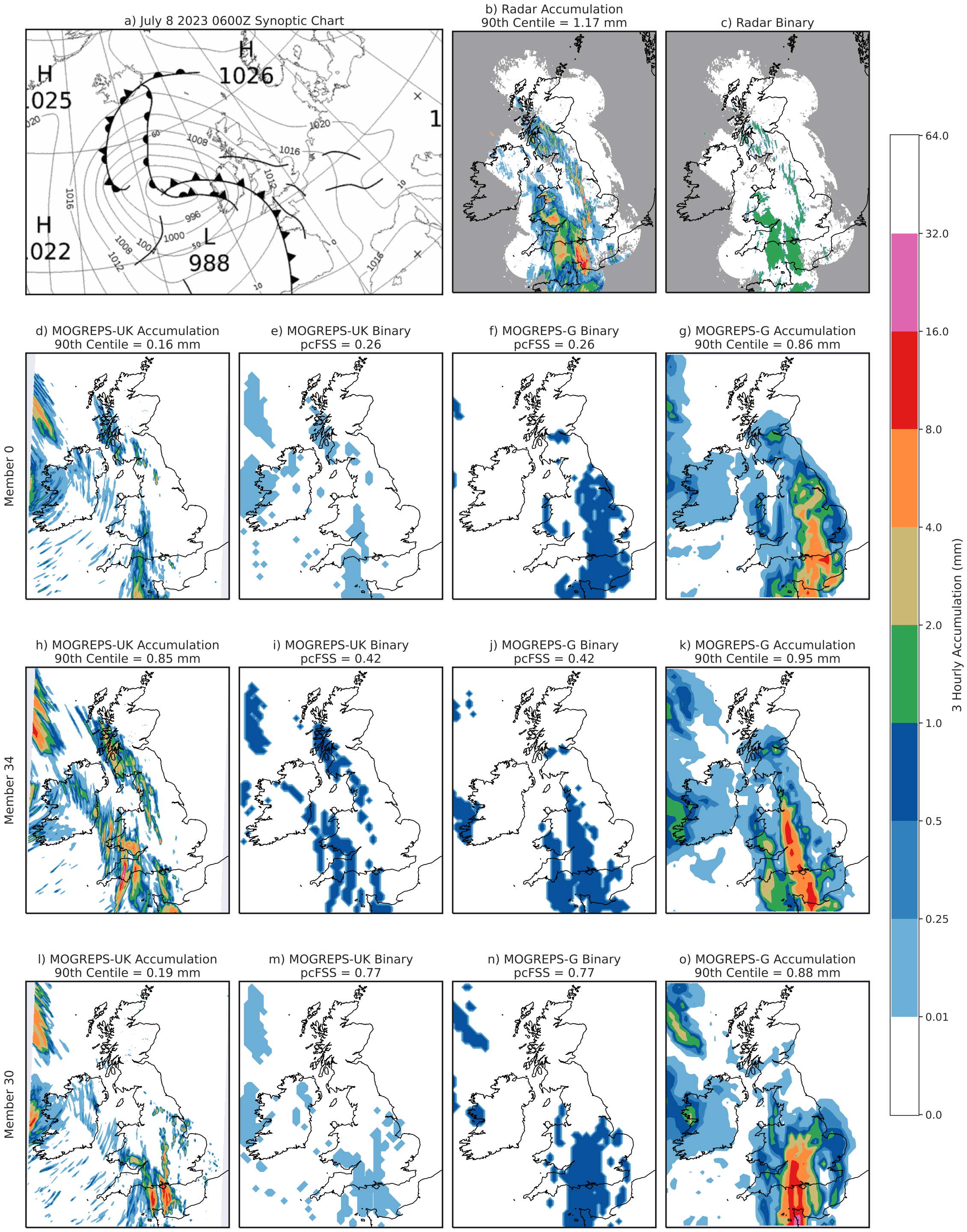

Figure 4.12: Case study for the thunderstorm event valid 09Z 8 July 2023 labelled on figure 4.11. Panel a) shows synoptic chart valid three hours before the event. All other panels show three-hourly precipitation accumulations valid 06Z–09Z 8 July 2023. Panel b) shows radar accumulation, while panel c) shows the location of 90th centile values from radar accumulation. The remaining panels show ensemble forecast data from MOGREPS-UK (Initialised 04Z 8 July 2023, T+5 h) and MOGREPS-G (Initialised 18Z 7 July 2023, T+15 h). Panels d), h), l) show the accumulation from a selection of members in MOGREPS-UK, while panels e), i), m) show the associated 90th centile locations used to calculate pcFSS (after coarse-graining to the MOGREPS-G grid). Panels g), k), o) show the accumulation from the corresponding members in MOGREPS-G, while panels f), j), n) show the associated 90th centile locations used to calculate pcFSS.

place across large parts of the country during the early hours of 8 July. These thunderstorms were associated with a plume of high 850 hPa wet-bulb potential temperature and high convective-available potential energy (CAPE) arriving from the continent (based on HYSPLIT trajectories (Stein et al. 2015a), not shown). Model uncertainty in the location and timing of convective outbreaks is often large under this continental plume setup, and the broad distribution of different-member pcFSS scores suggests that this is also true here.

This case is shown in more detail in Fig. 4.12. The synoptic chart in panel 4.12a) is valid 3 h before this case study, and shows the location of a pre-frontal trough ahead of an approaching cold front. The three-hourly radar accumulation (from 06Z to 09Z) in panel 4.12b) shows two distinct bands of rain, with thundery outbreaks occurring from southern England through into Scotland, and frontal rain pushing in from the west. The rest of the panels in Fig. 4.12 show the precipitation from a selection of three members from MOGREPS-UK and MOGREPS-G, along with the 90th centile binary fields used to compute the pcFSS.

Panels 4.12 d), e), f) and g) show the comparison of the control members of each ensemble and have the lowest scoring similarity of all member-member pairs in this period. However, this low similarity is not because MOGREPS-UK has been more successful than MOGREPS-G in initiating convection. Instead, the large differences in the threshold associated with the 90th centile, combined with the easterly bias of totals in MOGREPS-G, means that there is an almost perfect mismatch between the two binary fields. Most noticeably, the band of rain towards the western edge of the domain is included in the MOGREPS-UK binary but is largely absent in the MOGREPS-G binary. So, while neither member has accurately predicted all aspects of the verified weather, the low pcFSS member score is caused more by model biases than MOGREPS-UK providing additional forecast value.

However, this is not to say that differences in precipitation thresholds are always the cause for low pcFSS member scores. Panels 4.12 h), i), j) and k) show that member 34 has similar thresholds associated with the 90th centile in each model. In spite of this similarity, the pcFSS score is still small due to MOGREPS-UK placing more of the top 10% of rain within the region of the pre-frontal trough. Additionally, MOGREPS-G once again shows an eastern bias in the frontal precipitation compared to MOGREPS-UK, which also contributes to the lower similarity. Overall, however, this MOGREPS-UK member has been more successful in capturing the two distinct bands of precipitation, and the low pcFSS member score is reflective of this additional skill.

Finally, member 30 in panels 4.12 l), m), n) and o) shows a situation where the pcFSS can be large even despite large absolute threshold differences. The members from both models place the most extreme rain in approximately the same position, hence the large agreement between binary fields. MOGREPS-UK member 30 does a poor job of getting the correct convective coverage in much the same way that all MOGREPS-G members do, hence, the high pcFSS member score.

Overall, this case study has shown that low pcFSS scores can emerge both from the threshold biases that exist between models of such different resolutions, but also from the different, and more skillful, representation of smaller-scale features like convection. Therefore, care should be taken when interpreting low pcFSS events; it is up to the user to determine which of these causes are appropriate. The choice of T+15 h leadtime in MOGREPS-G (T+5 h in MOGREPS-UK) for this study shows the situation when MOGREPS-UK can make the best use of its high-resolution initial conditions, and before the boundary conditions dominate member evolution. Indeed, the same study performed with a leadtime of T+39 h in MOGREPS-G showed much stronger similarities between the two ensembles, and correspondingly larger scores. However, a 10 h offset in data assimilation cycles can be significant for forecasts of such unpredictable weather at these short leadtimes. Deconstructing the impact of the fresher MOGREPS-UK data assimilation cycles from the benefits provided by the higher resolution will be the subject of future work.

This supplement has shown that the pcFSS is responsive to the dominant weather regime, with convective regimes showing a greater tendency to produce smaller pcFSS values than other regimes.

Chapter 5

Assessing the Value of Clustering Convection-Permitting Ensemble Forecasts

This chapter has been submitted for publication in Meteorological Applications.

The roles of the other authors of this paper in relation to the project are as follows: T.H.A. Frame (supervisor: academic), S. L. Gray (supervisor: academic), R. Neal (Advisor: Met Office), A.N. Porson (supervisor: Met Office), M. Milan (supervisor: Met Office). The study was designed in collaboration with my supervisors, with the research questions discussed among all paper authors. The clustering method was initially developed by Kristine Boykin as part of her PhD Thesis entitled "Extracting Likely Scenarios from Ensemble Forecasts in Real-time". The iteration used in this study was developed by Robert Neal, with additional improvements implemented by myself. I conceptualised the research questions and strategy, and performed formal analysis with guidance and interpretation provided from all supervisors through weekly meetings. I wrote the first draft of the paper, prepared all figures, and had overall control of the submitted paper. All authors contributed to reviewing and editing the submitted manuscript. Approximately 80% of the paper was my work, and 20% was contributions from other authors.

Abstract

Ensembles provide a wealth of information to aid forecasters in their day-to-day operations, but with increasing ensemble size and complexity, there is rarely time to fully interrogate their outputs. Clustering ensemble members into distinct scenarios based on the co-location of hazardous weather features has previously shown promise when applied to global ensemble outputs. However, it is currently unclear whether further value can be gained when applying clustering to convection-permitting ensemble (CPE) outputs. This study compares precipitation clusters between the operational MOGREPS-G driving ensemble and the nested MOGREPS-UK CPE run at the (UK) Met Office during summer 2023. When applied over the UK domain, CPE clustering does not provide clear value compared to global ensemble clustering. Instead, clusters become increasingly similar with leadtime, strongly indicating that CPE clusters are most sensitive to the synoptic forcing common between the two ensembles, and that the presence of convective-scale detail has little influence. However, when focussed on a region impacted by hazardous convection, CPE clustering identified distinct precipitation scenarios and provided improved probabilistic value compared to driving-ensemble clustering. Finally, by comparing clusters with radar observations, it is demonstrated that the fraction of members supporting a particular scenario is a reliable quantitative prediction of the probability that the given scenario will be the most accurate. We recommend that global ensemble clustering is sufficient over larger domains, while CPE clustering is most useful when applied at regional scales.

5.1 Introduction

Ensembles are commonly used to quantify forecast uncertainty by running repeated simulations with different initial conditions and model parameters (e.g., Inverarity et al. 2023; Palmer 2019; Zhou et al. 2022). In theory, each member of a well-tuned ensemble can be interpreted as an equally-likely realisation of the upcoming weather that could be inspected without further processing. But with the strict deadlines imposed on forecasters, and the common production of additional convection-permitting ensemble (CPE) datasets, forecasters rarely have the time to perform these individual member examinations (Pagano et al. 2024; Young and Grahame 2024a). These restrictions motivate the need for methods that can intelligently summarise forecast outputs. While some benefits can be provided using common aggregation methods like ensemble means, these smoothed fields represent unphysical outcomes that can mask important spatial variability. It is therefore desirable to produce methods that can extract sets of unmodified members that represent the distinct forecast scenarios contained within the ensemble. While these methods have been previously trialled on global ensemble outputs (Atger 1999; Boykin 2022; Brill et al. 2015; Lamberson et al. 2023), only a few studies have examined the utility of these methods with CPEs (Branković et al. 2008; Johnson et al. 2011a), and none have performed a systematic comparison between the two ensemble types. Here, we perform such a comparison for precipitation forecasts using the operational ensemble from the (UK) Met Office.

In the early days of limited-area model design, clustering methods were explored as a way of selecting driving members that could provide the most spread for running a reduced number of computationally-expensive simulations (Marsigli et al. 2001; Molteni et al. 2001). These trials showed that the probability density function (pdf) of the high-resolution forecasts driven by cluster-informed representative members was a faithful recreation of the pdf from the driving ensemble. More recent studies have further confirmed the spread benefits when representative members are selected using clustering techniques over random subsampling (Bouttier and Raynaud 2018; Nuissier et al. 2012; Serafin et al. 2019; Weidle et al. 2013), such that this method has been used for driving the COSMO-LEPS (Montani et al. 2011), ALADIN-LAEF (Weidle et al. 2013) and HARMON-EPS (Frogner et al. 2019b) CPEs. It is noted, however, that these spread benefits typically only manifest when clustering is applied after leadtimes of approximately 48 h (termed T+48), when driving-ensemble pdfs are more likely to be multimodal.

Clustering methods are also used for the objective identification of weather patterns at the medium range (Fereday et al. 2008; Ferranti and Corti 2011). Here, climatological sets of mean sea-level pressure or geopotential height regimes are produced based on the occurrence of those regimes over multi-decadal timescales. Each regime represents a distinct circulation pattern and weather type. New forecasts are analysed using the same decomposition and assigned to the closest matching regime, providing a broad overview of the upcoming weather and its historical occurrence. These methods have been very successful at categorising and communicating synoptic-scale uncertainty out to multiple weeks, with separate schemes in use covering Europe at the European Centre for Medium-Range Weather Forecasts (ECMWF, Ferranti and Corti 2011; Ferranti et al. 2015) and the UK Met Office (UKMO, Neal et al. 2016, 2024), as well as over North America (Lee and Messori 2024; Lee et al. 2023).

Until recently, regime-based clustering was too broad to classify differences between individual features within high-resolution forecasts, limiting its usefulness for short leadtimes. Machine learning methods can now efficiently and accurately categorise regimes in CPEs, and can effectively reduce the dimensionality of their skewed precipitation distributions in a way that other statistical methods struggle with (**mounier˙rainfall˙2025**). Other work has focussed on applying clustering techniques in a more dynamic way, whereby groups of members are found directly from the ensemble and do not need to be compared to predetermined climatological clusters. Case study analysis has shown promise by successfully identifying distinct forecast scenarios when applied to limited areas from global ensemble outputs (Boykin 2022; Brill et al. 2015; Lamberson et al. 2023). Cluster verification has been more mixed, however, as the largest clusters have been found to be both more skillful (Lamberson et al. 2023) and less skillful (Brill et al. 2015) than the ensemble mean. It is likely that the performance of any clustering method depends strongly on the modality of the ensemble pdf. For instance, attempting to classify a Gaussian distribution (as would be anticipated from the ensemble at early leadtimes) into multiple sets will likely yield weakly-defined clusters that are ambiguous and of limited use (Atger 1999). It will likely be more instructive to perform clustering after the pdf has deviated significantly from Gaussianity, which may be more common in summertime convective cases (e.g., Hohenegger and Schar 2007; Lean et al. 2008). These pdf transitions have been well demonstrated in recent work analysing sampling uncertainties in large ensembles (Craig et al. 2022; Tempest et al. 2023, 2024). Therefore, one of the important aspects to consider concerns the timing of this pdf transition: is there any use applying clustering to CPEs, especially in the short term? Or, does the CPE pdf maintain Gaussianity until the lateral boundary conditions become dominant, after which the ensemble is likely to follow a similar trajectory to the global ensemble providing the boundary information (e.g., Gebhardt et al. 2011; Kühnlein et al. 2014; Zhang et al. 2023).

The results from a few existing studies can shed some light on this question.

Branković et al. 2008 compared clusters between a CPE and its driving ensemble to assess the added value from running ensembles at the convective scale. They found large differences between the ensemble clusterings: only a third of driving ensemble and CPE clusters possessed common representative members, and only half of the CPE clusters were closest to the expected driving ensemble cluster. Additionally, Johnson et al. 2011a,b developed clustering techniques using object tracking and neighbourhood-based smoothing, that account for the double-penalty problem commonly experienced when verifying high-resolution precipitation fields (Gilleland et al. 2009). Neighbourhood techniques provided more appropriate clusters than those that used raw model outputs. Finally, Boykin 2022 showed that clustering using a distance metric that directly incorporated neighbourhood smoothing could identify distinct frontal development scenarios and provide value to operational forecasters. The method used by Boykin 2022 is particularly useful since it is a purely spatial method, and can be applied to any input field regardless of heterogeneity, while also not relying on separate object-tracking algorithms to compute displacements.

In this study, we build upon the feature-based clustering work of Boykin 2022 to explore the potential benefits of applying clustering to CPEs. Of all the parameters that CPEs represent more accurately than global ensembles, precipitation forecasts benefit the most from the increase in resolution due to the explicit representation of convection (e.g., Cafaro et al. 2019; Clark et al. 2016; Hanley et al. 2011; Woodhams et al. 2018). As such, we focus on analysing cluster differences that emerge when applied to precipitation accumulations, and emphasise that the methods used here classify by spatial similarity, not intensity. We investigate the potential benefits of CPE clustering by answering three research questions. Firstly, do the bulk cluster statistics (cluster sizes, medoids, cluster memberships) demonstrate differences between convection-permitting and driving ensembles? Secondly, are clusters produced by the CPE more reliable than those produced by the global ensemble? Lastly, does CPE clustering provide better guidance in specific cases? Given the expected dependence of cluster quality on ensemble modality, it is likely that the answers to these questions will display some sensitivity to leadtime and regime.

The rest of the manuscript is organised as follows. Section 2 describes the ensembles and clustering methods used in the study. Additional analysis presented in the first section of the supplementary material compares different spatial methods for estimating distances between members in each ensemble. Then, Section 3 compares statistics between the cluster sets generated by the two ensembles. Section 4 then assesses cluster accuracy and reliability for both sets of clusters. Section 5 discusses the performance of clustering for a case of hazardous convection, and Section 6 summarises the main findings and recommendations.

5.2 Methods

In Section 5.2.1, the models and trial period used in this study are described. Then, in Section 5.2.2, the feature-based clustering procedure is described.

5.2.1 MOGREPS and Trial Period

For this study, we use data from the Met Office Global and Regional Ensemble Prediction System (MOGREPS): an operational ensemble configuration run at the (UK) Met Office comprised of a global ensemble, MOGREPS-G, and a nested ensemble run over the UK, MOGREPS-UK.

MOGREPS-G has a grid spacing of approximately 20 km in the midlatitudes, with 70 hybrid height vertical levels and a parametrization scheme to represent convection. It has initialisation cycles every 6 h at 00Z, 06Z, 12Z, and 18Z producing 17 perturbed members plus a control member from a global analysis, with each perturbed member separately initialised using a hybrid 4D ensemble variational data assimilation system (Inverarity et al. 2023). MOGREPS-G runs out to eight days and outputs three-hourly precipitation accumulations, and so we use three hours as the accumulation window for both ensembles throughout this work.

MOGREPS-UK is an 18-member lagged ensemble with 2.2 km grid spacing that runs out to five days (Hagelin et al. 2017). MOGREPS-UK has initialisation cycles every hour producing three new members that are combined with the 15 members from the previous five cycles to produce the full time-lagged 18-member set (Porson et al. 2020). For brevity and convenience, when referring to MOGREPS-UK leadtimes, we will ignore the different initialisation times between members and instead only quote the leadtime of the members from the most recent initialisation. This lagged setup allows each hourly three-member set to be recentred around the latest convective-scale analysis, with perturbations and boundary conditions provided by corresponding MOGREPS-G members.

We use operational ensembles in this study since the clustering tool is designed to facilitate forecast guidance production. Each ensemble is clustered on its native grid since we aim to understand the potential value provided by the inclusion of convective-scale detail. However, with any operational ensemble system, there are production delays between running the driving ensemble and using these outputs to drive the nested ensemble. In other words, the driving and nested ensemble forecasts that share a common initialisation time do not share common boundary conditions and/or perturbations. To properly compare clustering outputs between the two ensembles, consistent forcings must be used between the ensembles. Therefore, we use a “member-aligned” comparison setup which offsets the initialisation times

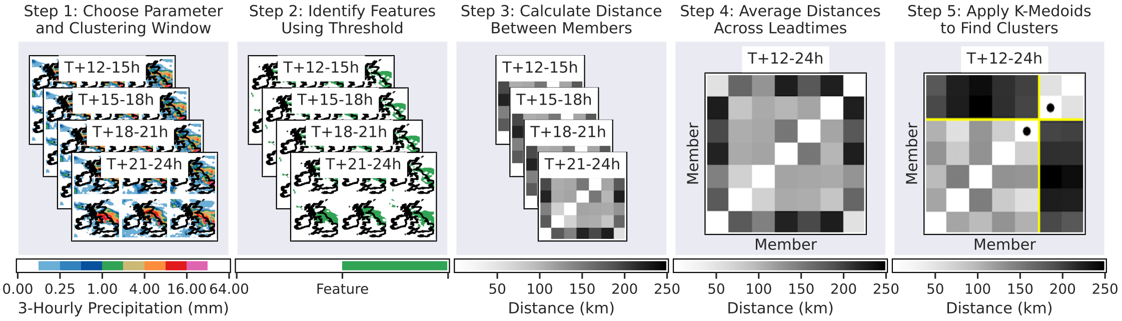
Step 1: Choose Parameter and Clustering Window
Step 2: Identify Features Using Threshold
Step 3: Calculate Distance Between Members
Step 4: Average Distances Across Leadtimes
Step 5: Apply K-Medoids to Find Clusters
T+12-15h
T+15-18h
T+18-21h
T+21-24h
T+12-24h
Member
Member
0.00 0.25 1.00 4.00 16.00 64.00
3-Hourly Precipitation (mm)
Feature
50 100 150 200 250
Distance (km)

Figure 5.1: Schematic showing clustering workflow. Step 1 shows the choice of parameter used in this work and the first clustering window mentioned in Figure 5.2. Step 2 shows the identification of features using a threshold. Step 3 shows the production of distance matrices between each member at each leadtime. Step 4 averages these distance matrices across each leadtime, before being used in K-Medoids clustering in Step 5 (which shows members reordered by their assigned clusters, where yellow borders denote clusters and black dots denote medoids). Each step is explained further in the text.

between the ensembles to ensure the same sets of members are being compared between both ensembles. In the MOGREPS system, the MOGREPS-UK forecast that includes the same members as the driving ensemble is initialised 10 h after the MOGREPS-G forecast (Gainford et al. 2024; Porson et al. 2020). So, a MOGREPS-G forecast initialised at 00 Z on a given day will be compared with the lagged MOGREPS-UK forecast with the most recent members initialised at 10 Z on that same day. By construction, each member is used exactly once. However, it is also important that the same events are being compared in each cluster set. Therefore, the clusters for a given MOGREPS-G forecast are compared to the clusters of a MOGREPS-UK forecast initialised 10 h later and with 10 h shorter leadtimes (see Fig. 5.2 for further details).

We apply clustering to operational forecasts run from June to August 2023. This period included a greater frequency of convective activity compared to climatology (UKMO 2023a), which provided a large sample of events that have the potential to produce broad differences between the two ensembles. The start of June 2023 was largely fine and dry due to a persistent block over the UK. A switch to more unsettled conditions occurred around the middle of June, with frequent thundery activity recorded. July 2023 was one of the wettest on record, with a predominantly westerly, mobile flow bringing a succession of weather systems from the Atlantic. August 2023 was more mixed than June and July, with wet periods interspersed with more settled conditions.

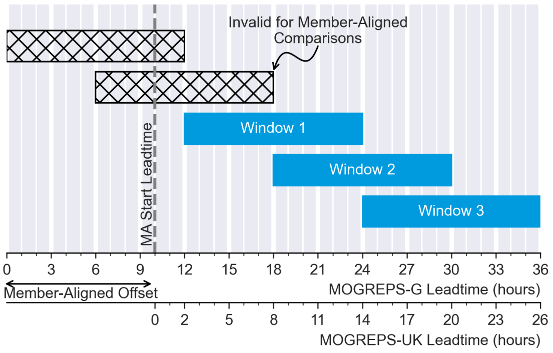

Figure 5.2: Leadtime window structure and comparison between ensembles accounting for 10 h member-aligned leadtime offset. Clustering is applied over each leadtime window to produce a set of consistent clusters valid for that window.

5.2.2 Clustering Procedure

The clustering workflow uses K-Medoids clustering with a distance metric that quantifies the average spatial displacement between features in different ensemble members. These methods have been integrated into an experimental tool run at the Met Office and are largely based on those developed in Boykin 2022, however two important differences have been implemented:

- The clustering window is now a user choice or free parameter, rather than defining it diagnostically,
- Clustering is now applied to all leadtimes within the clustering window, rather than to each leadtime separately.

Additionally, here, we use the Precipitation Smoothing Distance rather than the Fractions Skill Score Displacement to estimate spatial displacement, since this has been shown to provide more accurate estimates in idealised and real-world tests (Skok 2022). This is described further in Section 5.2.2

An example of the steps involved in the clustering workflow is depicted in Fig. 5.1 and described in the following subsections.

Steps 1 and 2: Leadtime Window Selection and Feature Identification

Firstly, a spatial field (e.g., gridded precipitation data), region, and clustering window is chosen. The spatial field can in principle be any meteorological parameter: here we choose three-hourly precipitation accumulations. We cluster over the MOGREPS-UK domain and extract this region from the MOGREPS-G fields. For the clustering windows, Fig. 5.2 shows an overview of the leadtime window structure used in this work. We use a smaller 12 h leadtime window rather than the 48 h window used by Boykin 2022 since we wish to cluster on timescales similar to those of convective storms. As mentioned in Section 5.2.1, we use member-aligned comparisons between the ensembles to ensure that a common set of members is available for clustering. This alignment choice imposes a leadtime offset between the windows, as demonstrated by the inclusion of multiple axes in the figure. Thus, window 1 of MOGREPS-G will always be compared to window 1 of MOGREPS-UK, but the MOGREPS-UK window will use a 10 h earlier leadtime range.

Spatial fields at each leadtime are then converted to a binary feature field by setting values above and below a chosen threshold to 1 and 0 respectively. Clustering on the feature field ensures that distances in the next step are calculated purely based on spatial displacements, and do not consider intensities. The threshold used can either be an absolute value or a centile value. We use a centile value since this accounts for coverage bias between members, and is the recommended approach for neighbourhood-based evaluation (Mittermaier 2021; Roberts and Lean 2008). We choose a 90th centile for use throughout this work, since initial sensitivity tests showed the clustering produced more consistent results with more populated feature fields. For context, the 90th centile corresponds to 0.74 mm/3h on average for MOGREPS-UK, and 0.72 mm/3h on average for MOGREPS-G. In an operational setting, this choice of centile may not result in clusters that are focussed on the areas of hazardous weather, and we would generally recommend a larger value be used provided it produced sufficient feature coverage.

Steps 3 and 4: Member Distance Calculation

Once features have been identified, a matrix of member-member distances is constructed at each leadtime and then averaged across those leadtimes. We use the Precipitation Smoothing Distance (PSD) to quantify member-member distances, which has been shown to estimate displacements more accurately compared to other spatial distance metrics (Skok 2022). The PSD operates first by normalising each input field, A, B, (here, the thresholded three-hourly accumulation) by the area average to remove biases. Then, the similarity between input fields is assessed using the Precipitation Smoothing Score, PSS, calculated as:

$$\mathrm{PSS}_{(r)}\left(A, B\right) = 1 - \frac{2}{Q N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left| A_{(r)} - B_{(r)} \right| \ , \tag{5.1}$$

where Q is the fraction of non-overlapping points between input fields, N_x, N_y are the number of grid points in x and y directions, and r is a smoothing radius. For the initial calculation, no smoothing is applied and $r = 0$. However, if the PSS does not exceed a score of 0.5, the input fields are smoothed using circular kernels of successively larger radii until the score condition is reached. Additionally, for comparisons at scales larger than the gridscale (i.e., $r > 1$), overlapping points between each input field are removed in $A_{(r)}$ and $B_{(r)}$, since these can lead to severe underestimations (Skok and Roberts 2018). The radius at which the PSS exceeds 0.5, $r_{PSS \geq 0.5}$, is then used to calculate PSD as:

$$\mathrm{PSD}_{(r)} = 0.808 \cdot Q \cdot r_{PSS \geq 0.5} \ . \tag{5.2}$$

Tests presented in the first section of the supplementary materials demonstrate that the estimated distances between MOGREPS-G members are much larger than those for MOGREPS-UK members, and that this bias occurs for all smoothing-based displacement metrics. This bias is likely reflective of the additional convective-scale detail in MOGREPS-UK, since the floor for feature distances is smaller than for the coarser global ensemble. These distance biases may contribute to clustering differences in subsequent results.

Step 5: K-Medoids Clustering

For the final clustering step, the leadtime-averaged distance matrix is grouped using K-medoids clustering. K-medoids is a partitional clustering method that determines clusters by first finding a set of k distinct central medoid members, where k is the desired number of clusters chosen by the user. We impose a maximum of 4 clusters, since this number was found to explain approximately 95% of the explained variance within the ensemble (Branković et al. 2008; Serafin et al. 2019). The medoids are found by iterating over all possible combinations of members as trial medoids. Each other member is then assigned to the closest trial medoid to create a set of trial clusters. The set of trial clusters that minimises the distance between each member and its medoid is chosen as the optimal set. Since the distance matrices are small in our case (18x18), each search process is exhaustive and finds the global minimum, giving the greatest likelihood that each medoid provides a distinct forecast scenario. Compared to other clustering techniques such as K-means, K-medoids has a distinct advantage that the central point is itself a physical solution, and can therefore be considered a suitable representation of all members within the cluster.

Occasionally, some members may have empty features fields at a particular leadtime (e.g., no precipitation exceeding the threshold value at any point) which requires special consideration. In such cases, we assign a distance of 0 km between two members that both do not have features, since they have identical fields. We also assert that it is impossible to estimate a physical distance between members

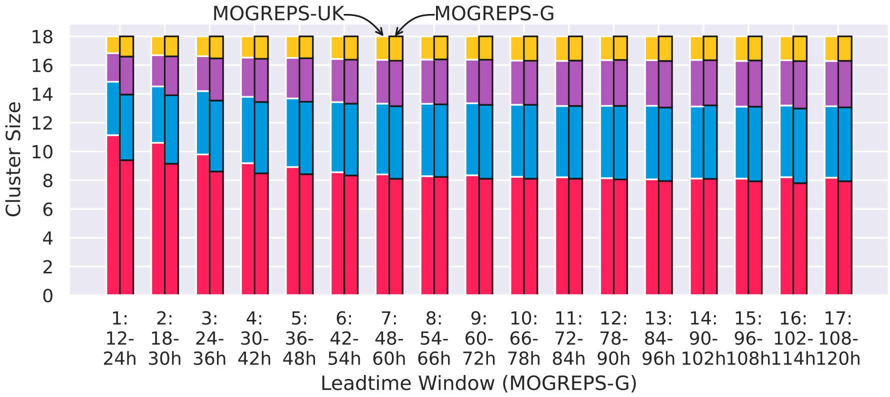

Figure 5.3: Histogram of average MOGREPS-UK and MOGREPS-G cluster sizes for $k = 4$

with and without features at a given leadtime, and therefore treat such distances as undefined so they do not contribute to the leadtime-averaged value. If the distance is undefined across all leadtimes in the 12 h window, the leadtime-average is then undefined. For the clustering step, any undefined leadtime-averaged distances are replaced by an arbitrarily large value of 9999 km so that all members without features are separated into an isolated cluster. For context, undefined values occur at least once in 7.9% of MOGREPS-G cluster windows, and at least once in 1.8% of MOGREPS-UK windows.

An example of the clustering outputs is shown in Figs 5.10, 5.11, and 5.12

5.3 Cluster Similarity

To understand the similarity between MOGREPS-UK and MOGREPS-G clusters, the following three subsections presents findings comparing trends in the size distributions, medoids, and cluster memberships from the two ensembles.

5.3.1 Cluster Sizes

Inspecting the cluster size distributions produced by each ensemble highlights the degree of heterogeneity within those members. For instance, a set of clusters in which each cluster contains a similar number of members can be interpreted as the forecast providing more diverse outcomes, since each forecast scenario has support from multiple members. Conversely, clusters with large disparities in size usually indicate that one solution is more strongly preferred than the others.

Figure 5.3 shows the average cluster sizes across the trial period at different leadtimes. Four clusters are enforced in this analysis, but the trends are broadly similar with fewer clusters. The clearest differences between cluster sizes occur during the first leadtime window, with the largest MOGREPS-UK cluster containing approximately two more members on average than the largest MOGREPS-G cluster. Consequently, the other three MOGREPS-UK clusters at this leadtime window contain slightly fewer members than MOGREPS-G. The difference in cluster sizes between the two ensembles diminishes with increasing leadtime and the sizes become largely equivalent by the sixth leadtime window (T+42-54 h in MOGREPS-G). Clusters typically remain at a consistent size for all subsequent leadtimes, with eight, five, three and two members.

These distributions indicate that MOGREPS-UK members are initially slightly more homogenous than MOGREPS-G members. This difference is not caused by the leadtime offset used in the member-aligned comparison setup, it is also present when leadtime consistency is enforced between the two ensembles. This finding is somewhat counter-intuitive given the time-lagged construction of the MOGREPS-UK ensemble which promotes larger spread compared to MOGREPS-G during early periods (Porson et al. 2020). Instead, it is likely that this trend emerges from a combination of two factors. Firstly, ensemble pdfs are typically unimodal at these early leadtimes, and the medoid associated with the largest cluster is often the most central member within the Gaussian. Secondly, there is a substantial reduction in the member displacements when evaluation is performed on finer grids, as discussed in the first section of the supplementary material. These displacement biases, combined with the modality argument, favours the production of more homogenous members in MOGREPS-UK, since each member is evaluated as being closer to the Gaussian medoid. This behaviour also explains the transition to more consistent cluster sizes from T+42-54 h, as more distinct ensemble modes are likely to develop after this period.

Interrogating the behaviour of the clustering through the lens of ensemble modality can also provide insight into the frequency of singleton clusters observed across the datasets. Figure 5.4a) shows the frequency that at least one cluster is produced containing only a single member: the medoid. Likewise, Fig. 5.4b) shows the frequency that at least two singleton clusters are produced, while Fig. 5.4c) shows the frequency that exactly three singleton clusters are produced (forming a 15-1-1-1 cluster structure). Note that it is not possible for two singleton clusters to exist for $k = 2$ (or three singletons to exist for $k = 3$), since all members must be assigned to a cluster. As with the size distributions presented in Fig. 5.3, there is a clear difference in the number of singleton clusters produced from the two ensembles. MOGREPS-UK produces substantially more singleton clusters than MOGREPS-G at early leadtimes, especially at the earliest T+12-24 h window. Almost 60% of $k = 3$ MOGREPS-UK clusters include a singleton at this leadtime, compared to only 38% of $k = 3$ MOGREPS-G clusters. In fact, MOGREPS-UK produces three singleton clusters ten times more often than MOGREPS-G within this earliest window. By T+42-54 h, however, both ensembles typically produce singletons at a consistent rate, but also at a much lower frequency than during earlier leadtimes.

These trends, including the reduction in the number of singletons with leadtime, can be understood by considering the clustering method in more detail. There are two main mechanisms that can generate singleton clusters. Intuitively, we would expect a singleton to emerge when one ensemble member provides a drastically different forecast to all other members such that it does not belong with any other grouping. However, this scenario is *less* likely to occur at earlier leadtimes when ensemble members are still normally distributed about the control member. Instead, it is likely that the presence of singleton clusters at early leadtimes arises from a sub-optimal number of clusters being forced onto the datasets. Consider an example of an ensemble pdf containing three distinct modes. When this dataset is clustered with $k = 3$, the outputs will ideally reflect these distinct modes. If this dataset is instead forced into $k = 4$, the new set of clusters that minimises the total member-medoid distance is simply the optimal set produced using $k = 3$ but with the member that is furthest from its medoid placed into its own cluster. By ‘peeling away’ this single member, the clustering retains the optimal minimisation produced using $k = 3$ as much as possible. Hence, the presence of singleton clusters can either indicate an inappropriate number of clusters, or can identify unique forecast scenarios.

The larger number of singleton clusters at early periods in MOGREPS-UK compared to MOGREPS-G is consistent with the previous interpretation of cluster size distributions. These findings demonstrate that MOGREPS-UK members are more homogenous at early leadtimes (up to T+54 h), but become similarly diverse at the leadtimes when we typically expect clustering to be more useful to forecasters. We also note that the signal at early leadtimes is not an effect of the different leadtimes used in the member-aligned comparison setup; it is also present when leadtime consistency is enforced between the two ensembles (not shown). However, these findings do not tell us about the similarity of the clusters themselves (i.e., medoids and membership), which is the focus of the next subsections.

5.3.2 Cluster Medoids

To understand the similarity of the cluster medoids found by clustering MOGREPS-UK and MOGREPS-G data, it is first instructive to inspect the typical distribution of medoids for particular leadtime windows. Figure 5.5 shows the frequency with which each member is chosen as a $k = 4$ medoid for an early and late leadtime window. Member 0 is the control in each set, and colours on the MOGREPS-UK panel indicate the members that were initialised in the same time-lagged cycle.

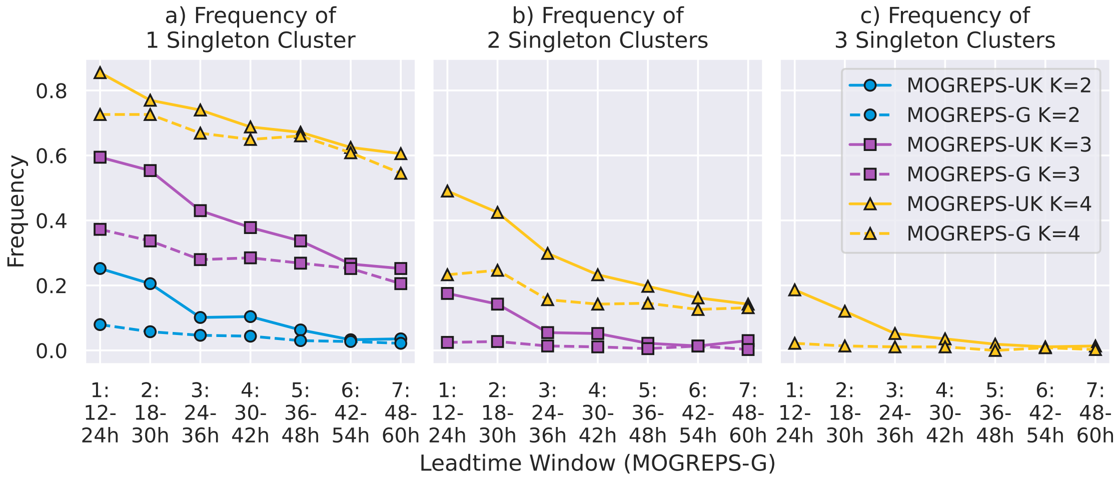

Figure 5.4: Frequency with which a) 1 singleton cluster, b) 2 singleton clusters, and c) 3 singleton clusters occur in MOGREPS-UK (solid) and MOGREPS-G (dashed) cluster sets for the first seven leadtime windows.

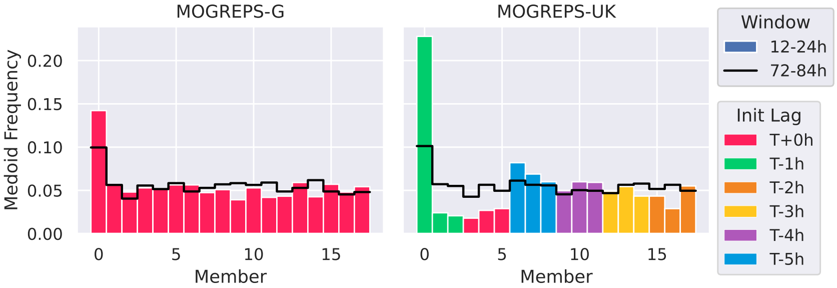

Figure 5.5: Frequency with which each member in MOGREPS-UK and MOGREPS-G clusters is chosen as a $k = 4$ medoid, displayed by filled bars for the first leadtime window (T+12-24 h in MOGREPS-G, T+2-14 h in MOGREPS-UK) and by outline for the eleventh window (T+72-84 h in MOGREPS-G, T+62-74 h in MOGREPS-UK). MOGREPS-UK members are coloured by the time-lagged initialisation cycle, as explained further in text.

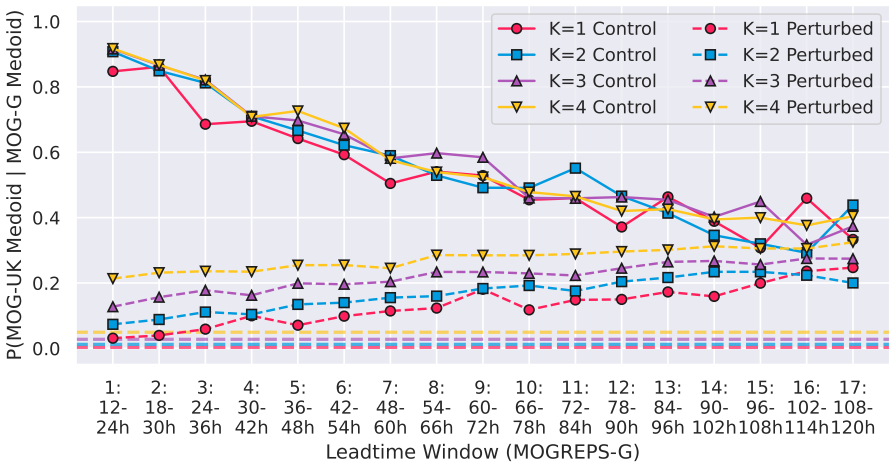

Figure 5.6: Conditional probability of finding a given medoid in MOGREPS-UK given it is also a medoid in MOGREPS-G. Perturbed trends are average over probabilities for each perturbed member individually. Dashed lines without markers are the probabilities of finding the same medoid in each ensemble by chance (e.g, for $k = 1$ this is $1/18$, for $k = 2$ this is $1 - (17/18 * 16/17)$ etc.).

For the earliest leadtime window, the control member is much more likely to be chosen as the medoid than any other member. This preference is to be expected given the fact that the perturbed members should still be centered around the control at this earliest leadtime window, so it is encouraging to see this trend in the data. It is also encouraging, though slightly more unexpected, to observe structural differences between the perturbed medoid distributions of the two ensembles. In MOGREPS-G, each perturbed member is approximately equally likely to be chosen as a medoid, while there is a clear bias towards certain perturbed MOGREPS-UK members being chosen. In the time-lagged construction of a MOGREPS-UK ensemble, members 6, 7, and 8 (blue) are consistently the oldest and consequently are more likely to be chosen as representing distinct outcomes. In general, the more recently a MOGREPS-UK member is initialised, the less likely it will be chosen as a medoid for this early leadtime window. At the later leadtime window, the disparity between control and perturbed members has substantially reduced, although is not completely eliminated.

This analysis motivates the need to consider the medoid similarity for control and perturbed members separately. To understand the degree of similarity between medoids, Fig. 5.6 shows the conditional probabilities of finding a given medoid in MOGREPS-UK clusters given its existence as a medoid in MOGREPS-G clusters. There is a decreasing chance of finding a control member medoid in MOGREPS-UK given its existence in MOGREPS-G as leadtime increases, reflecting the lower frequency with which control members are selected as medoids as spread develops in each ensemble. Conversely, the probability that a given perturbed member is

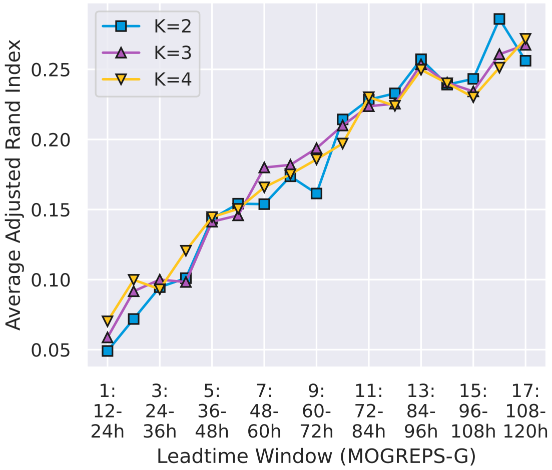

Figure 5.7: Average Adjusted Rand Index between ensemble clusters evaluating similarity of ensemble membership.

chosen as a medoid in each ensemble increases with leadtime, and at all times is larger than expected by random chance. This finding is reflective of the ensembles falling into distinct modes, but also suggests that these modes are consistent between the two ensembles. This consistency is perhaps partly explained by the influence of the lateral boundary conditions, which largely determine the evolution of each MOGREPS-UK member after the first day, and are provided by corresponding members of MOGREPS-G. By the final leadtime window, there is large parity between control and perturbed medoid probabilities.

These findings demonstrate a large degree of similarity in the central members chosen in each ensemble. Notably, this similarity increases with leadtime as the ensembles are more likely to develop distinct modes within the distribution. Hence, we may also expect to find larger similarity between cluster memberships in each ensemble as leadtime progresses.

5.3.3 Cluster Memberships

While simple methods can be used to compare cluster sizes and medoids, understanding similarities between cluster membership requires the use of slightly more involved methods. A popular choice for comparing two different cluster sets (here from MOGREPS-G and MOGREPS-UK) is the Adjusted Rand Index (ARI, Rand 1971; Vinh et al. 2010. The ARI operates by selecting a pair of members (e.g., members 1 and 5) and determining whether they are in the same cluster or different clusters in both sets by classifying each pair comparison into one of four categories:

- N_{11}: The number of pairs in the same cluster in both sets (e.g., member 1 and 5 are both in cluster 1 in MOGREPS-G and both in cluster 2 in MOGREPS-UK, or both in cluster 1 in both ensembles),

- N_{00}: The number of pairs in different clusters in both sets (e.g., member 1 and 5 are in clusters 1 and 2 respectively in MOGREPS-G, but are in clusters 2 and 1 in MOGREPS-UK),
- N_{10}: The number of pairs in the same cluster in the first set, but in different clusters in the second set (e.g., member 1 and 5 are both in cluster 1 in MOGREPS-G, but are in clusters 1 and 2 in MOGREPS-UK),
- N_{01}: The number of pairs in different clusters in the first set, but in the same cluster in the second set (e.g., member 1 and 5 are are in clusters 1 and 2 in MOGREPS-G, but are both in cluster 1 in MOGREPS-UK).

The ARI is then calculated as:

$$\mathrm{ARI} = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11}) + (N_{00}+N_{10})(N_{10}+N_{11})}\,, \tag{5.3}$$

where scores close to 0 indicate that clusters are no more similar than a random permutation of labels, and scores of 1 indicate perfect agreement between cluster sets. Negative scores indicate large dissimilarity.

Figure 5.7 shows the ARI calculated between the two ensembles for different k. As with the conditional medoid probabilities in Fig. 5.7, there is a clear leadtime trend whereby clusters are initially evaluated as being similar only by chance, but progressively become more similar for later periods. Once again, this evolution is likely a reflection of the modality of the ensembles at these times, where clusters applied to normally distributed members are less likely to be similar than clusters applied to multimodally distributed members. However, even by the final leadtime window, these scores are still relatively small, indicating that there are still large differences in the exact memberships between ensembles. Whether these differences are meaningful, or just reflective of large sensitivity to specific clustering parameters (domain, feature threshold etc.) is difficult to determine from this data. After all, it is unlikely that distinct clusters will always be present, and so we should expect some degree of uncertainty about the optimal cluster arrangements associated with the choice of inputs to the tool (Brill et al. 2015). Also note that there is little dependence on the value of k within these scores, which arises due to the normalisation of the Rand Index when accounting for random chance.

5.3.4 Cluster Similarity Summary

Taken together, the results in this section indicate that the similarity between clusters is most responsive to the modality of the ensemble distributions. While there

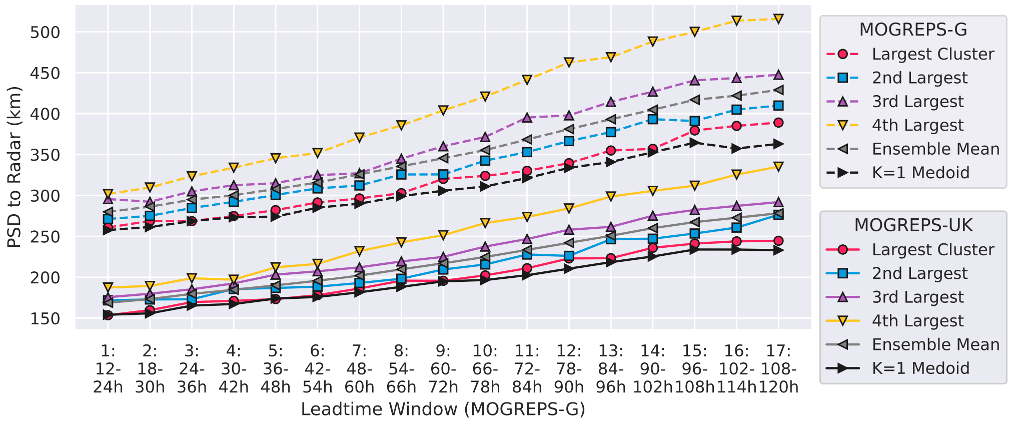

Figure 5.8: Average PSD between radar and ensemble medoids for each $k = 4$ cluster. Ensemble mean represents the mean PSD between each ensemble member and the radar.

is never complete agreement between the clusters, the clear trends with leadtime suggests that the clustering in both ensembles is largely being driven by the modes that exist at the common scales within the domains. Any clustering differences can be readily explained by fundamental uncertainties in the placement of members, caused by the fact that the ensemble modes may not always be entirely distinct, or that a given member may be an appropriate fit for multiple clusters. This uncertainty is even more pronounced when clustering over leadtime *windows*, rather than clustering on each leadtime separately. However, within this uncertainty there is the potential for a given set of clusters to provide better guidance than the other. The next section will explore whether this is the case by evaluating typical cluster skill and reliability for both ensembles.

5.4 Cluster Skill and Reliability

In this section, we determine the extent to which cluster size acts as a predictor of the likelihood of verification, with larger clusters indicating a more likely event. To investigate this, we calculate the PSD (eq. 5.1 and 5.2) between each medoid and the NIMROD radar (Golding 1998) three-hourly accumulations across the trial period. We focus on medoid skill in this section rather than cluster average skill to alleviate sampling differences that may occur with clusters of different sizes. To enable these comparisons, each radar field is interpolated to the corresponding model grid using a nearest-neighbour algorithm that masks extrapolated points. For each ensemble cycle, we then average the radar-medoid PSD across each leadtime window, consistent with the main clustering procedure.

Figure 5.8 shows the average PSD between the radar and the cluster medoids (using $k = 4$). For comparison, Fig. 5.8 also plots the mean PSD between the radar and each ensemble member, as well as the PSD between the radar and the $k = 1$ medoid, as two representations of the average distance from the full ensemble to the radar. The first trend to note is the segmentation between the two ensembles. We observe this same trend when using any smoothing-based displacement measure, and studies in the first section of the supplementary material link this to the grid resolution. Therefore, Fig 5.8 should not be interpreted as evidence that MOGREPS-UK is drastically more skillful than MOGREPS-G.

However, there is a clear separation in each ensemble between the medoid-radar PSD associated with different cluster sizes. The most populated cluster medoid is consistently closer to the radar than other medoids. In fact, all medoid distances are ranked by the size of the cluster they represent. Additionally, there is a notable offset between the medoid distances of the smallest cluster and the distances of all other medoids, especially at later leadtimes. Indeed, the least populated cluster medoid can be as much as 50% further from the radar than the most populated cluster medoid. This result is not too surprising given the frequency with which this smallest cluster is singleton (Figs. 5.3, 5.4a)), as well as the associated singleton arguments discussed in Section 5.3.1.

In comparison with the ensemble average, the largest and second largest cluster medoids are both typically closer to the radar than the ensemble mean. For context, Fig. 5.3 shows that these two cluster medoids combined typically represent 13–15 of the 18 members included in the ensembles, depending on the leadtime window. However, when compared to the $k = 1$ medoid (the member which has the smallest total distance from all other members), the largest cluster medoid is usually slightly further from the radar. So, despite the impressive separation of medoids by skill, the technique for finding the most likely ensemble mode selects a representative member that is less accurate compared to just finding the central state of the ensemble. Indeed, further interrogation reveals that the largest cluster medoid for $k = 4$ is the same as the $k = 1$ medoid approximately 70–80% of the time at early windows, but falls to under 50% of the time at the latest leadtime windows, explaining the growing disparity between the two.

Overall, these findings demonstrate that the medoids associated with larger clusters are consistently more skillful than those associated with smaller clusters. However, these findings do not provide insight into the reliability of the clusters, i.e., does the size of a cluster provide a useful quantitative estimate of the likelihood that the verifying observation will be closer to that clusters' medoid than any other medoid. Note that this verification is not estimating the *absolute* skill of the cluster medoid, only the probability that it is closest to the observation compared to all other medoids. To determine this, we use the radar-medoid PSDs to assign the radar to a cluster. There are two ways that this process can be implemented. One

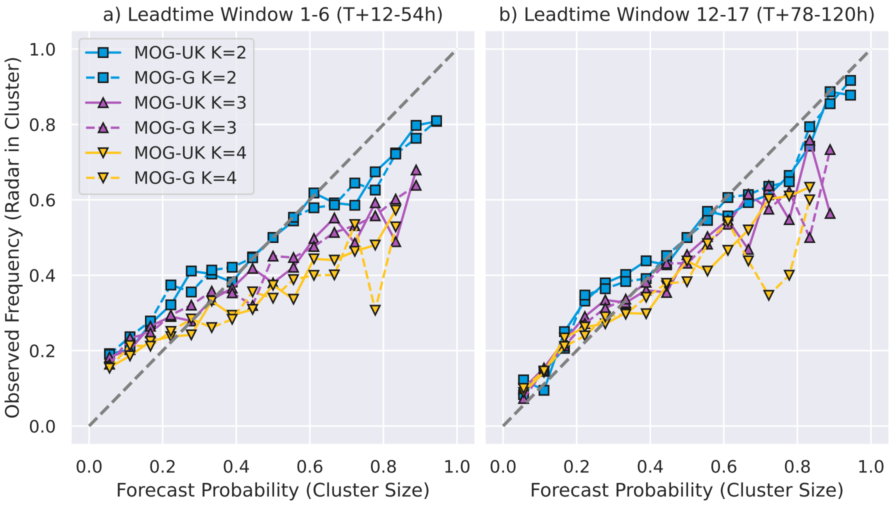

Figure 5.9: Cluster reliability diagrams for a) the first 6 leadtime windows and b) the last 6 leadtime windows. Forecast probability is the cluster size, observed frequency is the frequency with which the radar is placed into a cluster of that size.

approach is to include the radar as an 'extra ensemble member' and apply the full K-medoids workflow to these 18+1 members. However, due to the underspread nature of these ensembles, this often leads to the radar being placed into its own separate cluster and does not provide information about the cluster reliability. Therefore, we instead manually assign the radar to the cluster with the minimum medoid-radar PSD, thereby ensuring that the radar is placed into a cluster containing at least one ensemble member.

Figure 5.9 shows cluster reliability diagrams averaged over the first and last six leadtime windows. Here, forecast probability is determined by the radar cluster size normalised by the total number of ensemble members (18), while the observed frequency is determined by the fraction of instances that the radar is placed into a cluster of that size. As an example, if clustering is a reliable tool, we should expect the radar to placed into a cluster of size 12 approximately two-thirds of the time. It is also worth emphasising at this stage that we assign the radar to a cluster based on the closest *medoid*, not based on the closest member. This approach ensures that the radar is not preferentially placed into larger clusters by chance, and is also consistent with the K-medoids procedure.

Broadly speaking, the data in Fig. 5.9 follows the 1:1 perfect reliability line reasonably well, especially at later leadtime windows. The reduced reliability during early periods is likely reflective of members being distributed more normally at these leadtimes, which does not favour robust classification into distinct groups. Across all leadtimes, however, clusters with smaller k are typically more reliable than larger k. These differences may be related to symmetries in the reliability curves that emerge from the designation of the radar to a particular cluster. For instance, for $k = 2$, if the radar is placed into a cluster of size 12 approximately 75% of the time (as opposed to two-thirds of the time for perfect reliability), by construction, this necessitates the radar being placed into a cluster of size 6 only 25% of the time. Hence, any displayed underconfidence at one end of the 1:1 line and will be reflected as overconfidence at the other end. Following the same logic, we should expect to find perfect reliability for $k = 2$ at 50% probability, and indeed this is observed. We might anticipate that these arguments could be extended to larger values of k, and in general, the reliability curves appear to cross the 1:1 line at approximately $1/k$. However, the neatness of these symmetries will be unavoidably broken compared to $k = 2$ by the addition of more clusters for the radar to be placed into.

In summary, we have found that clustering is a reliable tool, and the number of members that supports each medoid is a useful measure of the probability that the medoid will verify most accurately. While there is certainly scope for improvements at the extreme ends, this is likely reflective of the underspread nature of the ensembles. However, these findings have also demonstrated that neither ensemble is more reliable than the other. Together with the results from the previous section, we are forced to conclude that clustering on the CPE does not add value compared to clustering on the driving ensemble, at least over these scales. Therefore, it is likely that the tool is most sensitive to the synoptic-scale variability that exists across the UK domain, and is not affected by the smaller-scale detail included in the CPE. This conclusion is supported by findings in Section 3 of the supplementary materials, showing a case where large-scale variability is well represented in clusters at the expense of smaller-scale variability. However, it is still possible that CPE clustering can provide value when used in a more ad-hoc basis, by isolating specific regions that will be impacted by extreme weather. Therefore, the final section of this study analyses the clustering performance in each ensemble for an impactful event within the trial period.

5.5 Convective Case Study

The event discussed in this section concerns a case of hazardous convection that impacted Wales and central England on 12 June 2023. This event was characterised by an area of high wet-bulb potential temperature over western areas of the UK with strong diurnal forcing providing the initiation. Slack pressure and slow winds prolonged the potential hazards, and an amber weather warning was issued over the effected regions. Impacts from surface-water flooding, hail, and thunderstorms were reported (UKMO 2023a).

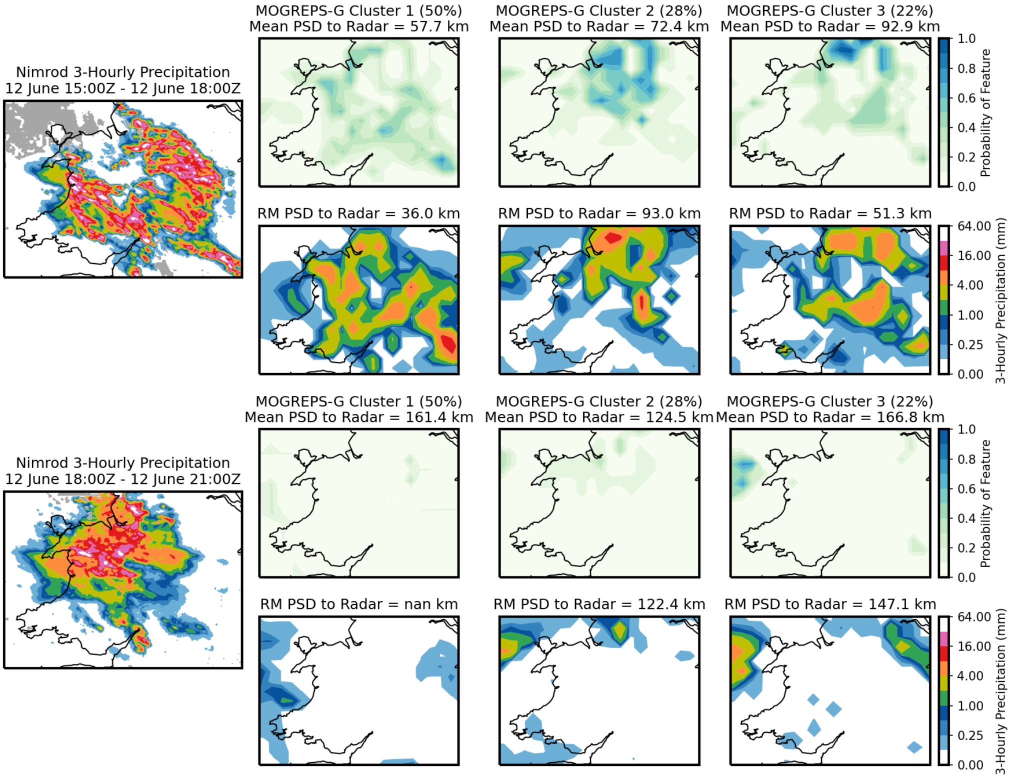
MOGREPS-G Cluster 1 (50%)
Mean PSD to Radar = 57.7 km
MOGREPS-G Cluster 2 (28%)
Mean PSD to Radar = 72.4 km
MOGREPS-G Cluster 3 (22%)
Mean PSD to Radar = 92.9 km
Nimrod 3-Hourly Precipitation
12 June 15:00Z - 12 June 18:00Z
RM PSD to Radar = 36.0 km
RM PSD to Radar = 93.0 km
RM PSD to Radar = 51.3 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00
MOGREPS-G Cluster 1 (50%)
Mean PSD to Radar = 161.4 km
MOGREPS-G Cluster 2 (28%)
Mean PSD to Radar = 124.5 km
MOGREPS-G Cluster 3 (22%)
Mean PSD to Radar = 166.8 km
Nimrod 3-Hourly Precipitation
12 June 18:00Z - 12 June 21:00Z
RM PSD to Radar = nan km
RM PSD to Radar = 122.4 km
RM PSD to Radar = 147.1 km

Figure 5.10: MOGREPS-G clusters for case study accumulation periods 2 (top) and 3 (bottom). NIMROD three-hourly verification is shown in the left column, using the same scale as the other precipitation plots but with grey regions indicating areas of insufficient returns (more than 10 minutes of missing data in a one hour period). Other columns show MOGREPS-G clusters. Feature density plots show the cluster-wide agreement of finding a feature at that location. Representative member (RM) plots beneath this show the three-hourly accumulation for the medoid of that cluster. The PSD between a member without features and the radar is undefined (NaN).

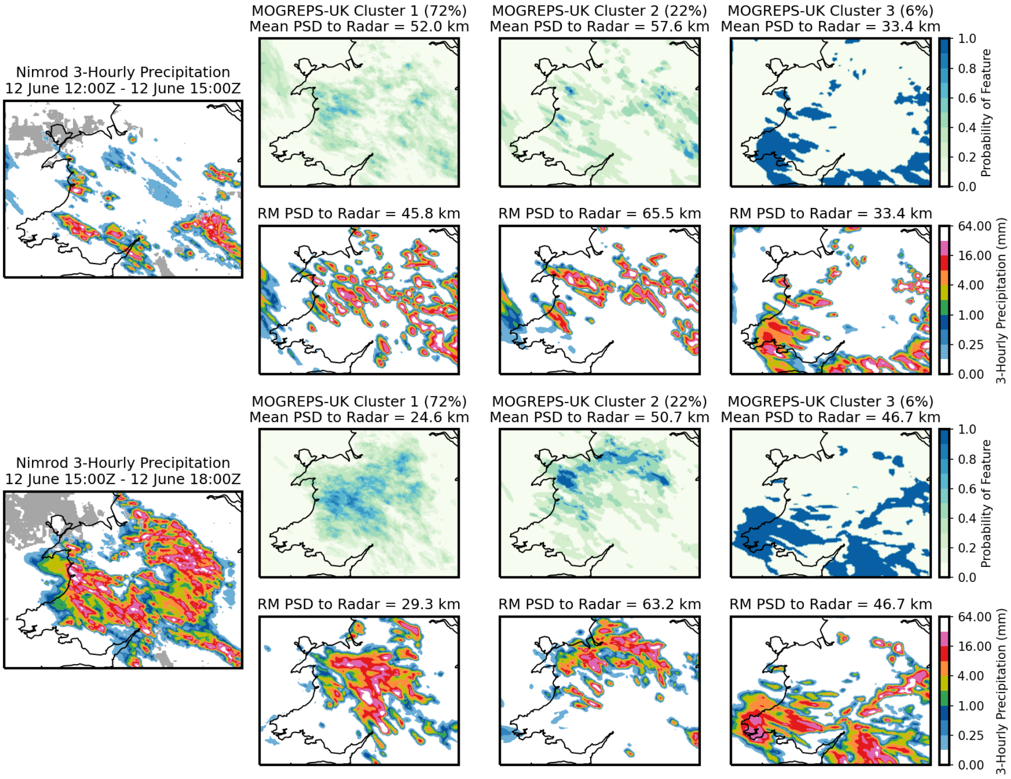
Nimrod 3-Hourly Precipitation
12 June 12:00Z - 12 June 15:00Z
MOGREPS-UK Cluster 1 (72%)
Mean PSD to Radar = 52.0 km
MOGREPS-UK Cluster 2 (22%)
Mean PSD to Radar = 57.6 km
MOGREPS-UK Cluster 3 (6%)
Mean PSD to Radar = 33.4 km
RM PSD to Radar = 45.8 km
RM PSD to Radar = 65.5 km
RM PSD to Radar = 33.4 km
Nimrod 3-Hourly Precipitation
12 June 15:00Z - 12 June 18:00Z
MOGREPS-UK Cluster 1 (72%)
Mean PSD to Radar = 24.6 km
MOGREPS-UK Cluster 2 (22%)
Mean PSD to Radar = 50.7 km
MOGREPS-UK Cluster 3 (6%)
Mean PSD to Radar = 46.7 km
RM PSD to Radar = 29.3 km
RM PSD to Radar = 63.2 km
RM PSD to Radar = 46.7 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00

Figure 5.11: As with Fig. 5.10 but for MOGREPS-UK clusters showing the first two case study accumulation periods.

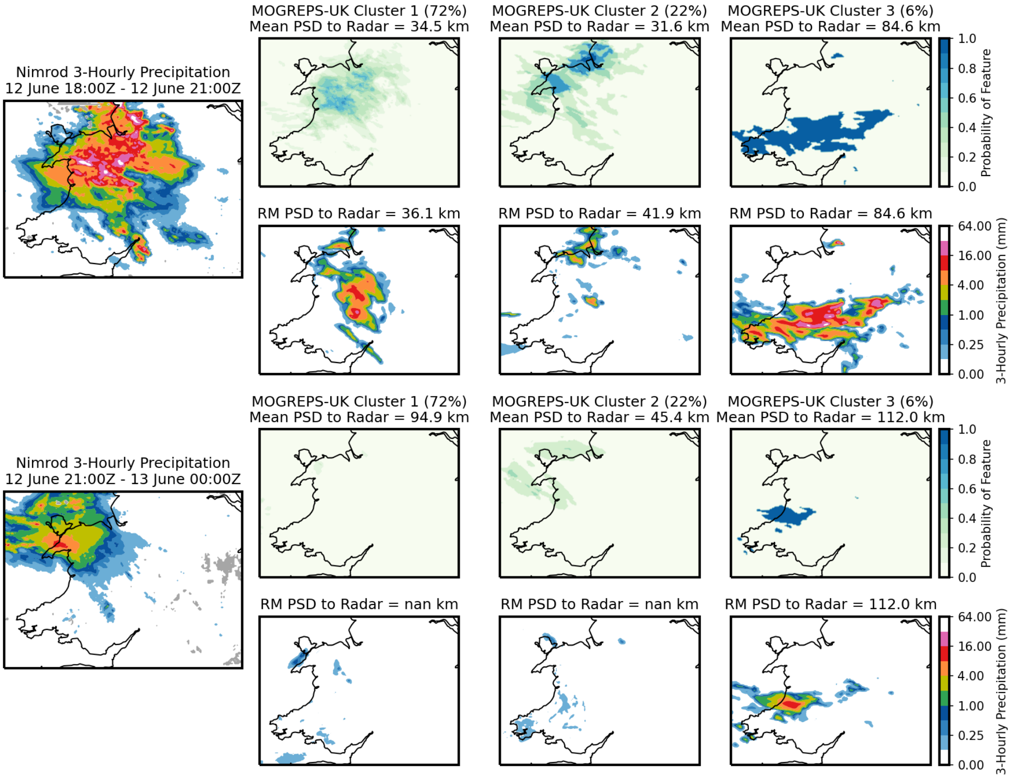
Nimrod 3-Hourly Precipitation
12 June 18:00Z - 12 June 21:00Z
MOGREPS-UK Cluster 1 (72%)
Mean PSD to Radar = 34.5 km
MOGREPS-UK Cluster 2 (22%)
Mean PSD to Radar = 31.6 km
MOGREPS-UK Cluster 3 (6%)
Mean PSD to Radar = 84.6 km
RM PSD to Radar = 36.1 km
RM PSD to Radar = 41.9 km
RM PSD to Radar = 84.6 km
Nimrod 3-Hourly Precipitation
12 June 21:00Z - 13 June 00:00Z
MOGREPS-UK Cluster 1 (72%)
Mean PSD to Radar = 94.9 km
MOGREPS-UK Cluster 2 (22%)
Mean PSD to Radar = 45.4 km
MOGREPS-UK Cluster 3 (6%)
Mean PSD to Radar = 112.0 km
RM PSD to Radar = nan km
RM PSD to Radar = nan km
RM PSD to Radar = 112.0 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00

Figure 5.12: As with Fig. 5.11 but for the final two case study accumulation periods.

To assess clustering performance, each ensemble is clustered over the region identified by forecasters as most at risk in guidance produced that day. The 12 h clustering period runs from 1200Z 12 June to 0000Z 13 June, which covers the formation and dissipation of convection. Clusters were produced using the shortest leadtimes (Window 1) available with the setup outlined in Fig. 5.2. Therefore, the MOGREPS-G forecast used here was initialised at 0000Z 12 June using leadtimes T+12-24 h, while the MOGREPS-UK forecast was initialised at 1000Z 12 June using leadtimes T+2-14 h. For each ensemble, four sets of three-hourly precipitation accumulations are used to produce clusters. As with the rest of the study, features are selected using the 90th centile, which corresponds here to 2.00 mm in MOGREPS-G and 1.47 mm in MOGREPS-UK. The outputs from using $k = 3$ are shown for each ensemble, as these were subjectively evaluated as giving the best clusters (all forecast scenarios represented without any being repeated).

Figure 5.10 shows clusters from MOGREPS-G for the second and third accumulation periods used for clustering, which are chosen to highlight the main trends for this ensemble (data from the entire period is shown in section 2 of the supplement for completeness). From 1500Z to 1800Z, the radar shows a peak of precipitation intensity as outbreaks of convection continued across Wales and central England. At the same time, MOGREPS-G presents much lower intensities, as is typical of these coarse grids. However, even accounting for the differences in resolution, MOGREPS-G clusters do not offer useful guidance for forecasting the locations that will be impacted by convection. Clusters 2 and 3 both predict the impacts will be largest across northern areas of the domain, while cluster 1 does not show a clear signal anywhere. Then, in the next three-hour period, precipitation in MOGREPS-G has largely dissipated, despite heavy radar returns being recorded across north Wales for the same period. In summary, MOGREPS-G has produced a poor forecast for this event, and while the accuracy of the underlying data will inevitably limit the potential of the clustering to add value, it is also clear the clustering has had limited success in distinguishing different outcomes.

Figures 5.11 and 5.12 between them show MOGREPS-UK clusters for all four accumulation periods used in clustering. In contrast to the MOGREPS-G clusters, each MOGREPS-UK cluster shows a distinct outcome, with clusters 2 and 3 showing northerly and southerly shifts in the impacted areas, and cluster 1 being between the two. MOGREPS-UK cluster sizes are also more unequal, with the scenario presented in cluster 1 being favoured by 13 members, while cluster 3 is only a singleton. In terms of forecast evolutions, most MOGREPS-UK members initialise convection too early and clear it too quickly. For the first accumulation period, the medoid for the cluster which shows a southerly bias (cluster 2) verifies closest to the radar. However, this southerly bias remains throughout all accumulation periods in this cluster, despite impacts pushing further to the north-west at later times. Subsequently, at the times when convection is heaviest (the second and third periods), cluster 2 does not verify as well at these times. Instead, the medoid representing the largest cluster verifies most accurately. Additionally, the probabilistic guidance from feature density plots is subjectively a better fit to the radar for cluster 1 than clusters 2 and 3, and the mean cluster PSD largely reflects this. For the final accumulation period, members in cluster 1 have all dissipated the convection too quickly, while some members from other clusters do a better job of retaining impacts for this period.

It is clear, then, that MOGREPS-UK clustering has provided more appropriate guidance than MOGREPS-G clusters for this event. While MOGREPS-G clustering was hampered by a poor forecast, it is also the case that clustering did not successfully highlight distinct scenarios within this poor forecast. Conversely, even though no individual MOGREPS-UK member fully resembled the verified event across all periods, clustering revealed useful probabilistic trends. Additionally, the medoids chosen for each cluster were representative of the trends highlighted by those clusters. Further, the medoid for the largest cluster verified most accurately of all medoids when all periods were taken into account. Inspecting other members within the ensemble revealed that one member from the largest cluster verified more accurately throughout all four accumulation periods than the largest cluster medoid. Apart from this member, the largest cluster medoid provided the best forecast for this event.

5.6 Discussion and Conclusions

Ensembles are becoming an ever more important part of a forecaster's toolkit, such that some meteorological services are retiring their deterministic models entirely and transitioning to an ensembles-only approach. With increasing ensemble importance, complexity, and size comes the need to produce methods that can intelligently summarise these large datasets. Feature-based clustering has previously shown value for identifying distinct frontal development areas in global ensembles (Boykin 2022). Here, we determine whether there is additional value to be gained from systematically applying clustering to convection-permitting ensembles (CPEs) compared to the global ensembles that drive them. We use the operational MOGREPS-G driving ensemble and MOGREPS-UK CPE for these comparisons and apply clustering to the 90th centile of three-hourly precipitation accumulations over a three month period. Note also that the tool used in this study clusters only on positional similarity of precipitation features, it does not consider magnitude differences.

In a routinely-running configuration, with both ensembles set up to cluster over the UK in 12-hourly windows, CPE clustering did not add clear value compared to driving-ensemble clustering. The leadtime trends of the representative member and cluster membership statistics strongly indicates that clusters are most sensitive to large-scale features. A separate case study presented in section 3 of the supplementary materials reinforces this conclusion by highlighting a situation where large-scale variability is well represented within the clusters while small-scale variability is largely neglected. This finding is consistent with previous interpretations of the behaviour of spatial verification methods (Roberts and Lean 2008).

Additionally, it is expected that clustering will perform more reliably and predictably when multiple distinct modes are present in the ensemble pdf. Here, we see that ensemble clusters are more similar at the leadtimes that are more likely to present multiple synoptic-scale modes than at earlier leadtimes, when ensemble members are still normally distributed about the control. Some differences between cluster sets are evident (for example, there is only approximately a one-third chance of finding the same medoid in both cluster sets at the longest leadtimes tested). This is due in part to unavoidable sensitivity to the clustering parameters when the ensembles do not fully capture the distributions they are attempting to represent (Brill et al. 2015).

This study also performs a systematic verification of feature-based clustering to determine the reliability of identified forecast scenarios. In each ensemble, the medoids representing the largest and second largest clusters are typically more skillful than the ensemble average. Further, the medoid representing the smallest cluster (when forced into four clusters) can be substantially less skillful than other medoids. However, when analysed from a reliability perspective, the smallest cluster can occasionally verify more accurately than other clusters. In fact, clustering demonstrated reasonable reliability in each ensemble, particularly for later leadtimes. Forecasters should therefore be confident that the number of ensemble members supporting a particular outcome is a reliable quantitative prediction of the probability that the given outcome will verify most accurately compared to the other identified outcomes. Of course, within underspread ensembles, this outcome may still be reasonably far from the verification, but this is not an issue that clustering can address.

While CPE clustering did not demonstrate consistent value when used at synoptic scales, it did demonstrate clear value when targeted over a region impacted by hazardous convection for a case study. While no CPE member fully resembled the event across all three-hourly accumulation periods contained in the 12 h window, clustering revealed distinct scenarios and useful probabilistic trends. Additionally, the medoid representing the largest cluster verified most accurately compared to the other cluster medoids. In contrast, the driving ensemble performed poorly, and clustering was not able to identify distinct scenarios. This case study reveals that CPE clustering is most useful when applied on an ad-hoc basis over more targeted domains. Therefore, a fully on-demand process would greatly enhance the appeal of the tool for use with forecasting mesoscale features.

When issuing guidance, it is also common practise for forecasters to compare outputs from other meteorological centres to judge the broader multi-model agreement. Given the persistent problem of underdispersion in ensembles, multi-model distributions can provide a wider range of possible outcomes. This technique is driving efforts to formalise these processes into methods that produce a consistent probabilistic output, whether it be at the short-to-medium range (Roberts et al. 2023) or at the medium-to-extended range (Neal et al. 2024). It may also be useful to apply feature-based clustering to multi-model ensembles, where there has previously been limited success in testing methods that are willing to mix members from different ensembles (Alhamed et al. 2002; Brill et al. 2015; Lamberson et al. 2023; Yussouf et al. 2004). Additionally, it may also be useful to apply clustering to multiple parameters at once to identify self-consistent, multi-hazard scenarios, such as those associated with freezing temperatures and heavy precipitation.

Finally, the clustering process described in this study requires the user to decide ahead of time on the desired number of clusters, k, which may not always be known. In an operational setting, a forecaster is likely only concerned with the number of clusters needed to provide the best guidance, i.e., the clusters that display all of the possible scenarios without any of those scenarios being repeated between clusters. In such cases, k is more useful as an *indication* of the number of distinct modes contained in the ensemble, rather than as a free parameter. Therefore, it is desirable to produce additional processing methods that can decide on a “suggested” or “optimal” k to present to the user. Developing a method that can reliably identify the optimal outputs will require extensive testing and verification.

5.7 Comparison of Spatial Distance Methods

In Sec. 2.2.2 of the main text, we used the Precipitation Smoothing Distance (PSD, eq. 1 and 2) to measure the displacements between fields of different members, but we also mentioned the possibility of using other methods. Since the PSD shows large biases between the average displacements for each ensemble, it is instructive to understand whether these biases are a common feature for all smoothing-based displacement metrics.

The simplest of these methods, the FSS Displacement (Skok and Roberts 2018), operates by finding the neighbourhood at which the FSS (Roberts and Lean 2008) reaches 0.5, where the FSS is calculated as:

$$\mathrm{MSD}_{(n)}\left(A, B\right) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[A_{(n)i,j} - B_{(n)i,j}\right]^2 \ , \tag{5.4}$$

$$\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A, B\right) = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[A^2_{(n)i,j} + B^2_{(n)i,j}\right] \ , \tag{5.5}$$

$$\mathrm{FSS}_{(n)}\left(A, B\right) = 1 - \frac{\mathrm{MSD}_{(n)}\left(A, B\right)}{\mathrm{MSD}^{\mathrm{ref}}_{(n)}\left(A, B\right)} \ , \tag{5.6}$$

where MSD is the mean-square difference between features fields, $A_{(n)}$ and $B_{(n)}$ that have been smoothed by a neighbourhood of grid size n, MSD$_{ref}$ is the low-skill climatalogical baseline, and N_x, 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. Then, the neighbourhood at which the FSS reaches 0.5, $n_{\mathrm{FSS}=0.5}$, is used to calculate the FSS Displacement (Skok and Roberts 2018), $DFSS$ as:

$$\mathrm{DFSS} = \left(1 - \mathrm{FSS}_{(n_{\mathrm{FSS}=0.5})}\right) \cdot n_{\mathrm{FSS}=0.5} \ , \tag{5.7}$$

which can often be approximated to simply DFSS $= n_{\mathrm{FSS}=0.5}/2$ since the FSS is close to 0.5. In this form, the FSS Displacement is the average grid-point distance between features in fields $A_{(n)}$ and $B_{(n)}$, and can easily be converted to a physical distance by multiplying by the grid size.

However, Skok and Roberts 2018 found that the displacement between overlapping features could be severely underestimated with the formulation outlined in equation 5.7. It was found that a simple overlap-adjustment corrects this underestimation by setting any overlaps between members equal to 0 in both fields when calculating distances using any applied smoothing (i.e., for $n > 1$). The overlap-adjusted FSS Displacement is then calculated as:

$$\mathrm{DFSS}_{\mathrm{OA}} = (1 - \mathrm{FSS}_{(n=1)}) \cdot \mathrm{DFSS}\ , \tag{5.8}$$

where $\mathrm{FSS}_{(n=1)}$ is the FSS applied at the gridscale (where overlaps must be preserved to ensure nonzero scores). The DFSS in Equation 5.8 has been calculated as in equation 5.7 but with overlaps removed in the binary feature fields used to create the smoothed $A_{(n)}$ and $B_{(n)}$ fields. A full procedure and set of considerations for using the overlap-adjusted DFSS can be found at the end of Skok and Roberts 2018.

While all methods are functionally similar, the main differences between the two DFSS formulations and the PSD are:

1. Unlike the DFSS, the PSD does not require binary data to calculate displacements. Instead, each field is normalised by its area-averaged value to account for intensity biases. However, using non-binary data introduces intensity weighting to the calculation, meaning that the score can no longer be interpreted as a purely spatial displacement. Using the PSD with non-binary fields can be advantageous since the score will provide greater weight to displacements between areas of more impactful weather. Here, however, we are only interested in offset between *features*, and therefore we will continue to use binary input fields.
2. The PSD uses a circular neighbourhood rather than a (typically) square one as used in the DFSS. While there is no fundamental reason why the DFSS cannot also use a circular neighbourhood, square neighbourhoods are easier to implement. Either way, it is not believed that the neighbourhood shape has a meaningful impact on scores (Roberts and Lean 2008).
3. Not only are the input fields adjusted to remove overlaps as with the $\mathrm{DFSS}_{\mathrm{OA}}$, the PSD explicitly takes account of the overlap fraction, Q, within both the similarity score and the distance estimations.
4. The PSD does not normalise its similarity score by a low-skill climatology as with the DFSS.
5. The PSD compares fields using a mean-absolute difference rather than a mean-squared difference, since this was found to produce more accurate estimates of displacement.

Figure 5.13 shows the average distances provided by all three methods over 12 h leadtime windows. For references, MOGREPS-UK grid spacing is 2.2 km

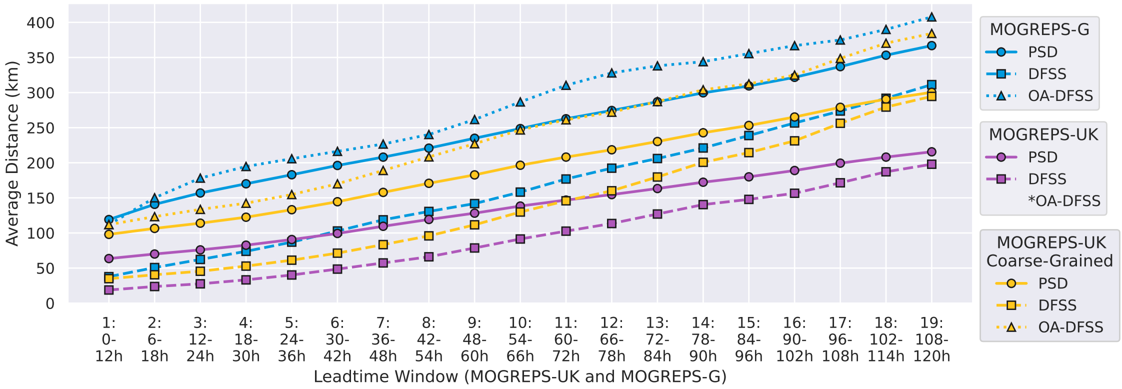

Figure 5.13: Average distances between all methods for 12 h leadtime windows for MOGREPS-G, MOGREPS-UK and MOGREPS-UK coarse-grained onto the MOGREPS-G grid. Coarse-graining was performed using a nearest-neighbour algorithm that masked extrapolated points. Overlap-Adjusted FSS Displacements were not able to be calculated for MOGREPS-UK on its native grid due to the increased time complexity of using this method. Averages were calculated over all forecasts initialised in July 2023.

while MOGREPS-G grid spacing is approximately 20 km over the region tested. All methods show that the distances estimated by MOGREPS-G are larger than those for MOGREPS-UK on their native grids. This bias is likely reflective of the additional convective-scale detail introduced in MOGREPS-UK, where the floor for feature distances is smaller than for the coarser global ensemble. The increased similarity between MOGREPS-G trends and those of MOGREPS-UK coarse-grained onto the same grid is supportive of this conclusion, although some differences remain.

Comparisons of the different methods shows that the FSS Displacement displays consistently smaller distances, consistent with the expectation that including overlapping features reduces estimates (Skok and Roberts 2018). Note that the overlap-adjusted FSS Distance for MOGREPS-UK was not calculated due to the additional computational time complexity required to reach the FSS threshold. While procedures do exist to reduce the computational costs associated with neighbourhooding (e.g., Faggian et al. 2015), we did not anticipate this would yield important insights.

In summary, it is clear there are large distance biases between the ensembles no matter the method used. These differences are important to consider when comparing cluster outputs between the two ensembles, as well as for future work that may wish to use these methods. Since all methods tested here show the same biases, the choice was made to continue with the PSD in the main article.

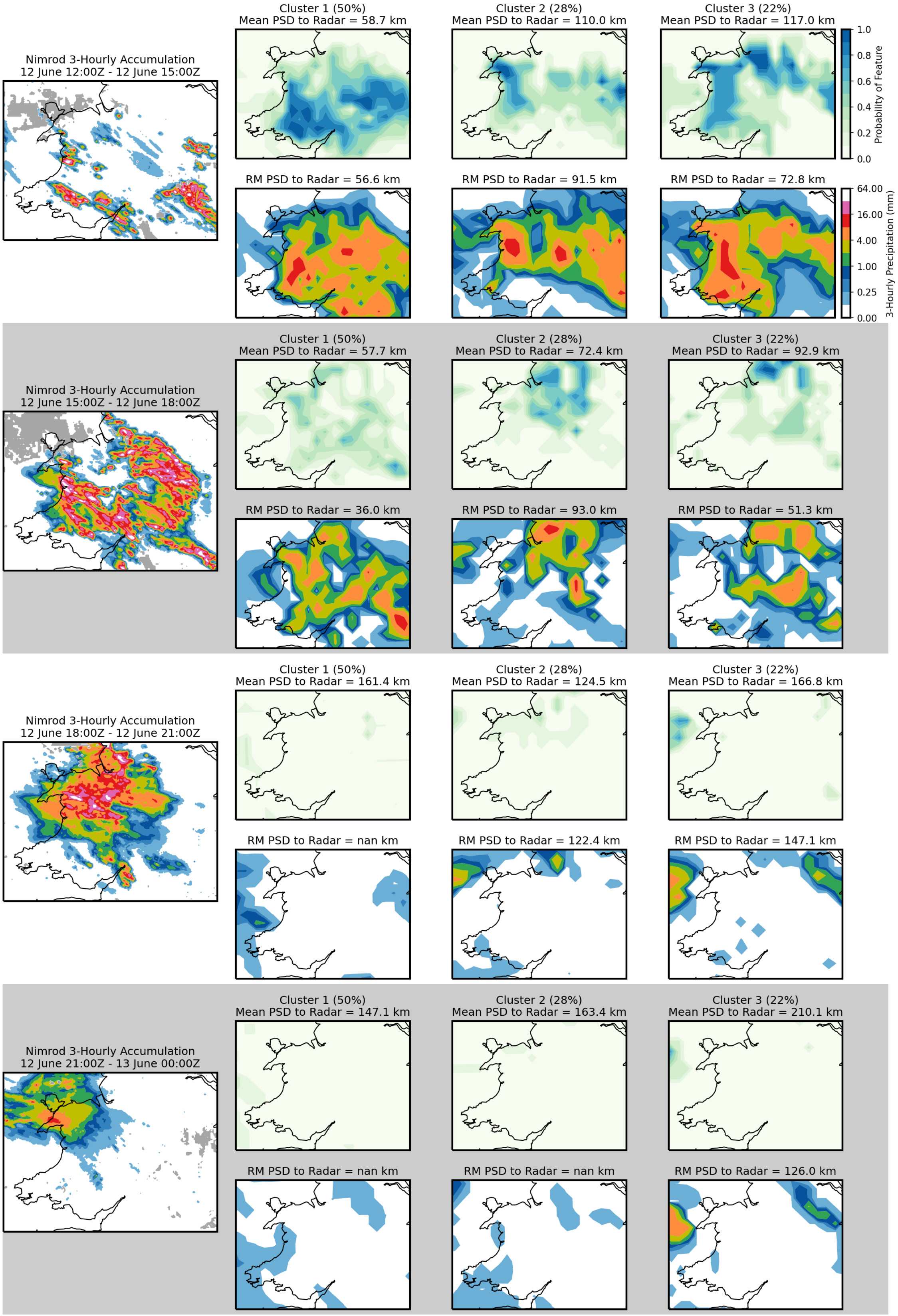
Nimrod 3-Hourly Accumulation
12 June 12:00Z - 12 June 15:00Z
Cluster 1 (50%)
Mean PSD to Radar = 58.7 km
Cluster 2 (28%)
Mean PSD to Radar = 110.0 km
Cluster 3 (22%)
Mean PSD to Radar = 117.0 km
RM PSD to Radar = 56.6 km
RM PSD to Radar = 91.5 km
RM PSD to Radar = 72.8 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00
Nimrod 3-Hourly Accumulation
12 June 15:00Z - 12 June 18:00Z
Cluster 1 (50%)
Mean PSD to Radar = 57.7 km
Cluster 2 (28%)
Mean PSD to Radar = 72.4 km
Cluster 3 (22%)
Mean PSD to Radar = 92.9 km
RM PSD to Radar = 36.0 km
RM PSD to Radar = 93.0 km
RM PSD to Radar = 51.3 km
Nimrod 3-Hourly Accumulation
12 June 18:00Z - 12 June 21:00Z
Cluster 1 (50%)
Mean PSD to Radar = 161.4 km
Cluster 2 (28%)
Mean PSD to Radar = 124.5 km
Cluster 3 (22%)
Mean PSD to Radar = 166.8 km
RM PSD to Radar = nan km
RM PSD to Radar = 122.4 km
RM PSD to Radar = 147.1 km
Nimrod 3-Hourly Accumulation
12 June 21:00Z - 13 June 00:00Z
Cluster 1 (50%)
Mean PSD to Radar = 147.1 km
Cluster 2 (28%)
Mean PSD to Radar = 163.4 km
Cluster 3 (22%)
Mean PSD to Radar = 210.1 km
RM PSD to Radar = nan km
RM PSD to Radar = nan km
RM PSD to Radar = 126.0 km

Figure 5.14: MOGREPS-G clusters for the full convective case study period discussed in Section 5 of the main text, and following the same format as Figs 5.10, 5.11, 5.12.

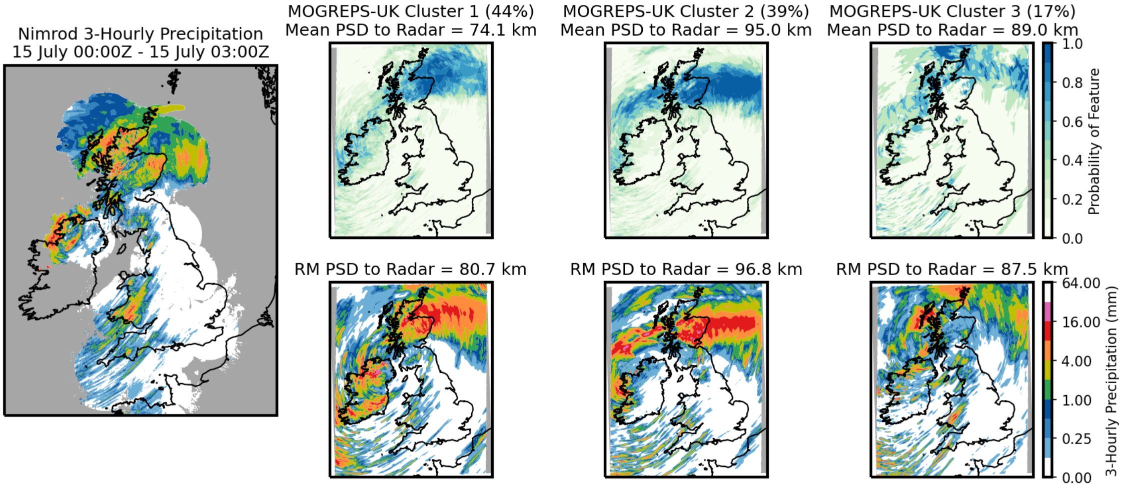
Nimrod 3-Hourly Precipitation
15 July 00:00Z - 15 July 03:00Z
MOGREPS-UK Cluster 1 (44%)
Mean PSD to Radar = 74.1 km
MOGREPS-UK Cluster 2 (39%)
Mean PSD to Radar = 95.0 km
MOGREPS-UK Cluster 3 (17%)
Mean PSD to Radar = 89.0 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
RM PSD to Radar = 80.7 km
RM PSD to Radar = 96.8 km
RM PSD to Radar = 87.5 km
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00

Figure 5.15: MOGREPS-UK clusters for a case demonstrating the sensitivity of clustering to different spatial scales. Follows the same format as Fig. 5.14 and other figures presented in Section 5 of the main text. Grey regions of the radar are masked due to insufficient returns in at least one hour of the three-hourly period (where insufficient returns is defined as less than 50 minutes of available data). This case was clustered using a 2 mm threshold and with a single 3 h timestep, initialised on 13 July 2023 with leadtime T+48-51 h, valid 15 July 2023 00:00Z - 03:00Z.

5.8 Convective Case Study

For completeness, Fig. 5.14 shows the full case study period as discussed in Section 5 of the main text. Grey backgrounds are used to visually delineate different leadtimes.

5.9 Multi-Scale Case Study

Statistical analysis from the main text show that there is little difference in the composition and skill of clusters from convection-permitting (CPE) and global ensembles when applied over the UK domain. These findings indicate that the presence of convective-scale detail in the CPE has little effect on the construction of clusters, meaning that the clusters are mostly sensitive to the synoptic-scale variability common between the two ensembles. To show this sensitivity explicitly, this section presents a case study using a forecast containing hazardous weather across synoptic and convective scales.

The event presented in this section is from 15 July 2023 0000Z-0300Z. The forecast of this event was initialised 13 July 2023 0000Z, with leadtimes T+48-51 h. For ease of interpretation, this section will only consider a cluster window of 3 h, comprising of a single three-hourly accumulation timestep. During this period, heavy precipitation associated with an active south-westerly front impacted parts of Scotland and Ireland. At these leadtimes, there is some uncertainty in the position and orientation of this front as well as its associated impacts. In the wake of this front, convection pushed in from the southwest and brought more localised impacts to parts of Wales and southwest England. As with most forecasts of convection, there is uncertainty in the exact positioning of these convective-scale features, with some members showing large totals over Wales, and others keeping the convection from reaching land (not shown). Since there is uncertainty in the location of both the frontal and convective-scale features, this case provides an opportunity to demonstrate the scale sensitivity of the clustering procedure.

Figure 5.15 shows MOGREPS-UK clusters for this event using $k = 3$, which was subjectively identified as representing all possible scenarios without any scenario being repeated. We do not consider MOGREPS-G clusters for this case study, since we are interested in understanding the influence of convective-scale detail on clustering. For this reason, we also use a 2 mm absolute threshold to cluster the ensemble, since the threshold associated with the 90th centile was too large (4.5 mm) to provide sufficient coverage for convective-scale features. To capture all of the hazardous weather, clustering was run over the full UK domain.

Both the feature probability density and representative member plots in Fig. 5.15 show that the large-scale variability is well represented in these clusters. Frontal precipitation is noticeably less zonal over Scotland in cluster 1 than in cluster 2. These clusters also differ in their prediction of impacts across Ireland, with cluster 1 showing a greater likelihood than cluster 2. In contrast, cluster 3 shows the potential for some impacts across Northern Ireland, but with a much weaker frontal band of precipitation and located further north. Therefore, all three clusters show distinct scenarios in the positioning of large-scale features.

In contrast, the clusters have not represented the variability in the smaller-scale convection affecting the south-west of the UK. Each feature probability plot shows a similar chance of convection affecting similar regions of Wales, despite broader uncertainty existing within the ensemble. The representative members for these clusters demonstrate slightly more diversity, as may be anticipated from physical fields rather than probabilistic fields. However, these members do not contain the full range of solutions observed in the ensemble, since none of the representative members show a scenario that keeps the convection away from land. Therefore, the clustering workflow appears to be less sensitive to differences in the placement of convection.

In summary, this case has demonstrated that the small-scale convective detail contained in MOGREPS-UK has little effect on the clusters compared to the dominating influence of the variability in frontal positioning. This finding reinforces the statistical results found in the main text, and highlights the need to consider the scale of the hazardous weather when deciding on the clustering domain. This finding is also consistent with previous interpretations of the behaviour of spatial verification methods, which shows that they are most sensitive to the largest-scales contained in the domain (Roberts and Lean 2008).

Chapter 6

Discussion and Conclusions

6.1 Overview

This thesis presents three studies to improve our understanding of the role that synoptic-scale information plays in the performance of CPE precipitation forecasts. As a reminder, the research questions posed in this thesis are:

RQ1 How can the global ensemble be used to better understand and improve spatial precipitation spread within CPEs?

RQ2 How dependent is spatial spread and skill of precipitation in CPEs on the synoptic-scale flow?

RQ3 How can the benefits of spatial verification methods be utilised to understand CPE behaviour?

RQ4 How can spatial techniques be further exploited for operational purposes?

In addressing **RQ1**, chapter 3 demonstrates that blending the synoptic-scales provided by the global analysis into the CPE analysis improves the spatial precipitation spread-skill relationship by a modest but significant amount. These improvements persist for approximately 24 h, until the synoptic-scale information from the driving ensemble introduced through the lateral boundaries dominates the evolution of each CPE member. Given the improvements observed when these synoptic-scale corrections are applied directly to only three of the 18 CPE members, it would be useful to explore any further improvements that could be made when corrections are applied to all members instead. The importance of the lateral boundaries is then further interrogated in chapter 4, where it is shown that the similarity between pairs of members in the driving and nested ensembles that share consistent forcings stabilises from T+18-30 h. This is attributed to the initial condition information being replaced by lateral boundary information. Before the boundary transition period, the CPE is evaluated as being correctly spread owing to the lagged initialisation schedule used, while the driving ensemble is significantly underspread.

After the boundary transition period, both the CPE and the driving ensemble are evaluated as being similarly underspread, highlighting the important role that the driving ensemble plays in determining spread after the first day. Finally in chapter 5, clustering techniques are used as another method for interrogating the similarity between the driving ensemble and the CPE. Clustering is most sensitive to the number of distinct modes in the ensemble pdf, therefore the leadtimes at which the two ensembles produce similar outputs is longer than in chapter 4. At the synoptic scales used in the majority of this study, ensemble clustering shows increasing similarity with leadtime as distinct synoptic modes are more likely to be present. The similarity of clustering outputs therefore reflects the similarity of the large-scale detail that exists between the two ensembles.

In addressing **RQ2**, chapter 4 demonstrates that particular synoptic regimes are more prone to producing larger spread within the CPE than the driving ensemble. In the regimes that favour convective activity, CPE spread is typically larger than in the driving ensemble, and these CPE forecasts show larger departures from the corresponding global forecast than in other regimes. However, the skill of these CPE forecasts can still be poor, which offsets the benefits of increased spread and means the spread-skill relationship is still suboptimal. Conversely, within mobile regimes, spread is largely similar between the two ensembles, and the similarity between ensemble forecasts is larger than for other regimes. Since these similarities highlight the importance of steering winds in transmitting information through the limited CPE domain, further studies could attempt to link the similarity of precipitation forecasts between the driving and nested ensembles to wind speed at upper levels. Additionally, it may also be useful to link driving-nested similarity to the verified convective regime, where quasi-equilibrium events forced by stronger synoptic conditions may be more likely to show larger similarities than non-equilibrium events driven by regional variability. Results presented in chapter 5 show that clustering is more reliable in both CPEs and driving ensembles from approximately day 3, when distinct synoptic regimes are more likely to be present. While this finding does not relate directly to ensemble spread or skill, it does demonstrate that the number of members supporting a particular synoptic outcome is a good indicator of its likelihood to verify, and therefore its skill.

In addressing **RQ3**, chapter 3 describes an extension of the Localised Fractions Skill Score (LFSS), originally designed to highlight regional skill variation in deterministic forecasts. Chapter 3 explores a modified formulation of the LFSS for use with ensembles to identify regional spread and skill variation. This method is used to highlight the impact of large-scale blending on the spread and skill of particular weather events, and shows that correcting the large-scale initial conditions can introduce more spread in areas that are otherwise lacking. While this LFSS extension was useful in this particular case study, it is also anticipated that these methods could provide insights when applied to an extended period of data. Rather than assessing individual weather events, the ensemble LFSS would instead identify areas that are more likely to suffer from spread or skill deficiencies if applied to a season or more of data. In chapter 4, the Fractions Skill Score is used to develop a new method that identifies the conditions in which the driving ensemble exerts more influence over the CPE by measuring the similarity of driving-nested member pairs. This method can reproduce expected behaviours like the lateral boundary transition timing, and is then used to show the regimes in which CPE provides more value than the driving ensemble through different precipitation forecasts. In chapter 5, an experimental post-processing system driven by spatial methods is applied to CPE precipitation forecasts to assess the value of clustering members based on the co-location of hazardous weather features. A case study is examined which demonstrates that clustering can help to distinguish common trends between different groups of members and provide value for assessing the regions most likely to be impacted by extreme weather. Even though no individual member resembles the verified outcome throughout the full clustering window, the ensemble contained enough spread such that there are multiple distinct forecast storylines, and the storyline most supported by other members is that which verified most accurately.

In addressing **RQ4**, the introduction of multiple comparison frameworks between the driving ensemble and CPE in chapter 4 formalises a common choice that model developers have to make when evaluating the performance of operational CPEs. Because the CPE relies on the driving ensemble for boundary information (as well as initial perturbations in some ensembles), there is a delay between the production of driving ensemble and CPE outputs. In this time, further observations will have been taken and used to construct fresher analysis for the CPE. Therefore, the choice between maintaining initialisation consistency and maintaining forcing or member consistency is important when undertaking subjective evaluations of ensemble performance as this may influence the perceived value of higher resolution outputs. The first section of the supplement for 4 directly compares both of these comparison choices to help provide some guidance about the most appropriate framework for use at different leadtimes. Additionally, the second section of the supplement demonstrates the utility of the new FSS method developed in the main text as a tool for aiding the production of forecast guidance. By displaying time series of similarity scores, users can quickly distinguish between the periods when the CPE diverges more from the driving ensemble, and therefore provides a distinct forecast outcome that may be worth investigating in more detail, from the periods when both ensembles produce a consistent story. In chapter 5, the clustering tool used throughout the study is designed to be implemented operationally once there is sufficient confidence in the techniques. Part of building this confidence involves verifying that the tool produces reliable clusters, whereby the number of members in a given cluster supporting a given outcome can be used as a prediction of the probability that the outcome will verify most accurately. The work presented in 5 confirms that clustering is a reliable tool through the development and application of objective verification metrics. It will also be useful to subjectively evaluate this iteration of the clustering tool over an extended period to further build confidence.

6.2 Novel Contributions

This section lists the novel contributions provided by each research chapter of this thesis.

Chapter 3 performs analysis on the effects of a new data-assimilation scheme that constrains the large-scale initial conditions of the CPE to follow those of the global analysis, which is more accurate at those scales. This scheme was tested and developed at the UK Met Office and was demonstrated to improve skill during early periods of deterministic forecasts. While other studies have analysed the effects of similar schemes on CPEs (Keresturi et al. 2019), our study uses a different framework, different blending configuration, and different analysis methods:

- The construction of the method for estimating uncertainty was carefully considered by taking account of the initialisation and blending procedures specific to the MOGREPS-UK ensemble. This method is then used to demonstrate the periods within which the impacts on spread and skill are significant.
- This chapter also introduces a novel extension of the Localised Fractions Skill Score, which had been used in other studies to assess regional variations in deterministic skill. Here, we extend the use of the LFSS to highlight regional variability in CPE performance. This extension is used to show that a forecast of elevated convection contained more spread when the large-scale initial conditions were corrected in the CPE, and an area of deficient spread was also improved.

Chapter 4 describes the processes and methods for comparing operational outputs between a CPE and its driving ensemble:

- Due to production delays between the two ensembles, it is not possible to compare CPE outputs that are initialised with analyses produced at the same time and that contain the same sets of members. This work discusses these complications and makes recommendations for the relative importance of these comparison frameworks when performing subjective and objective evaluations.

These comparison frameworks are then used throughout the rest of the study to identify the periods when CPE precipitation forecasts provide the most value compared to driving ensembles.

- Many studies have compared outputs between CPEs and driving ensembles (e.g., Clark et al. 2009, 2010; Klasa et al. 2018), but none have focussed on spatial precipitation spread due to the additional complexity of accounting for the double penalty problem via the FSS. This research gap then facilitates the development of a new method using FSS that directly compares outputs between the two ensembles, and it is shown that this can be a useful tool both operationally and in research contexts.
- This work is also the first to link the spatial spread-skill relationship of precipitation patterns to distinct forecast regimes, and shows that the CPE produces larger spread than the driving ensemble in conditionally unstable setups.

Chapter 5 investigates the value of running feature-based clustering on a CPE over running clustering just on the driving ensemble. This research builds upon the work of Boykin 2022 who developed the initial version of this technique and showed that it was a useful tool for extracting distinct frontal development zones within global ensembles:

- While other studies have also used clustering with CPEs (e.g., Ferranti and Corti 2011; Lamberson et al. 2023; Molteni et al. 2001), none have systematically compared feature-based clustering for post-processing purposes between a CPE and its driving ensemble.
- This is also the first study to produce objective verification of feature-based clusters across an extended period of data, showing that feature-based clustering is a reliable tool and that the skill of a given forecast scenario is linked to the number of members supporting that scenario.
- This work makes clear recommendations about the best uses of driving ensembles and CPEs for clustering in an operational context.

6.3 Future Work

6.3.1 Applications of the Localised Fractions Skill Score to Assess Regional CPE Biases

Previous research has demonstrated the utility of the Localised Fractions Skill Score (LFSS) for highlighting regions of enhanced or diminished skill within deterministic forecasts (Ferrett et al. 2021; Woodhams et al. 2018). The work presented in 3 expands on this and develops a method for using the LFSS to analyse regions of enhanced and diminished spread within ensembles. This approach can also identify regions that may be more responsible for deficiencies in the spread-skill relationship. The study presented in 3 applied these ensemble-based LFSS extensions to an individual case study to show the impact that a configuration update made to ensemble spread and skill. However, there is also the potential to apply these methods to a larger period of data to understand systematic regional biases in ensemble spread and skill. For instance, during a typical summer over the UK, we might expect regions that are more prone to experience extreme convection, such as the southeast of England, to show larger spread and lower skill than regions further to the north where the weather is more dictated by synoptic-scale weather patterns. It would also be useful to link variations in spread and skill to the orography of the surrounding area. Previous work has shown that more mountainous regions predict precipitation more skilfully as assessed using the LFSS (Woodhams et al. 2018), but it is not currently clear whether this would have a meaningful impact on local spread.

Given a sufficient sample of data, it may also be useful to further decompose this LFSS analysis by regime type, similar to the approach used in 4. This decomposition could accentuate the regional signal observed from analysing data across the full period, and provide confidence that the trends are linked to the forecast weather. Additionally, a further extension of the LFSS that compares corresponding parent-child members may also be useful for highlighting the regions in the CPE that are most likely to follow the driving ensemble, and those that are more likely to provide a different forecast outcome. Again, these trends will likely be regime dependent, and additional decomposition using the synoptic flow could provide broader insights.

6.3.2 Investigating the Impact of Observations in a Case Study of Convection

In the second section of the supplementary material for 4, it was demonstrated that the parent-child FSS could be a useful tool for signalling periods of divergence and degeneracy between the CPE and the driving ensemble. This analysis was aided by a case study of a period that the pcFSS identified as showing particularly large differences between the two ensembles. Inspecting the forecasts for each ensemble revealed that these differences were largely driven by biases in the location of pre-frontal convection, with the CPE providing better guidance than the driving ensemble. However, since the parent-child FSS requires the use of consistent forcings at the expense of consistent initialisation timings, the CPE used fresher observations for this forecasts than the driving ensemble. Forecasts of convection are particularly sensitive to the initial state, therefore it is possible that the CPE may have performed better purely due to the assimilation of more recent observations.

Therefore, it would be useful to conduct an experiment to understand the degree to which the improved CPE guidance can be explained by using fresher observations, or whether operating at the convective-scale also provides a meaningful contribution to this particular forecast. This experiment could be performed by running the CPE as a downscaler of the driving ensemble, which does not recentre CPE members around observations. Analysing changes in the FSS between the operational and downscaler runs would highlight the impact that the lack of observation makes to the forecast. If the pcFSS is similar in the downscaler compared to the operational runs, then the observations make little impact and the CPE performance can be attributed mostly to the high-resolution grid. If the pcFSS is different (i.e., larger) in the downscaler, then this shows that the observations played an important role in this forecast. However, this experiment is only useful if the driving ensemble provides a consistent forecast between operational and downscaled runs. If there are small differences in the model configurations (or even differences in the architecture that runs the code), this could produce a slightly different set of driving ensemble forcings that would limit the strength of the comparison.

6.3.3 Investigating New Methods for Producing a Data-Informed Spread Attribution Diagram

Figure 2.5 presented in Section 2.3.3 shows a useful schematic for understanding the typical contributions to CPE spread (intitial conditions, boundary conditions and high-resolution model processes) and the variation with synoptic forcing. However, this schematic is simply an estimate of the partitioning between different spread contributions based on previous results in the literature, and is not informed by data. Therefore, it would be useful to produce a version of this figure that uses analysis from CPE outputs to rigorously define typical spread contributions. Further, it would be especially useful if this figure was produced using spatial verification methods and could provide insight into the factors governing spatial precipitation spread within CPEs.

Fortunately, one such method already discussed in this thesis is a promising candidate for producing such analysis. The parent-child FSS (pcFSS) discussed in chapter 4 directly compares the similarity of precipitation accumulations between driving-nested member pairs. In fact, Figs. 4.7 and 4.8 show that the pcFSS can quantify the differences between the CPE and corresponding driving members that emerge from high-resolution model processes. Additionally, Fig. 4.7 and the associated discussion demonstrates that the pcFSS can also be used to estimate the period when the initial conditions are replaced by information arriving from the boundaries. Attempts to analyse the gradient of the pcFSS leadtime series to constrain this transition time were limited by the a lack of data and general noise of the gradient. However, with a sufficient quantity of data, there is the potential for the gradient of pcFSS to give a quantifiable estimate of the rate at which initial conditions are replaced by boundary information. This, combined with pcFSS leadtime series as an estimate of the contribution from high-resolution model processes, could provide the full partitioning expected from the schematic.

For this approach to be convincing, more research into these interpretations would be needed. After all, the method described above would essentially be manipulating the same dataset in two different ways to estimate the contribution from two different sources of spread. Further theoretical work or idealised tests would be needed to demonstrate that the pcFSS can be decomposed and interpreted in this way. Alternatively, a different FSS formulation may be needed to estimate the lateral boundary transition timing in a manner consistent with the pcFSS.

6.3.4 Objective Determination of Optimal Cluster Sets to Present to Users

The study presented in chapter 5 shows that clustering is a reliable tool and that it can provide genuine value to forecasters. However, the process described in this chapter requires the user to decide ahead of time on the desired number of clusters, k, which may not always be known. In an operational setting, for instance, a forecaster is likely only concerned with the number of clusters needed to provide the best guidance, i.e., the clusters that display all of the possible scenarios without any of those scenarios being repeated between clusters. In such cases, k is more useful as an *overview* of the number of distinct modes contained in the forecast pdf, rather than as a free parameter. Therefore, it is desirable to produce additional processing methods that can decide on a “suggested” or “optimal” k to present to the user. There are many existing techniques for determining the optimal cluster size from a given set of clusters: the elbow method, for instance, chooses the value of k where the percentage of the total-member variance represented by the clusters demonstrates diminishing returns with further increases of k, which is observed as a noticeable “elbow” on a distance vs k plot. Alternatively, the silhouette method (and the similar representivity index (Molteni et al. 2001)) assesses the cohesion and separation of clusters by comparing distances between members within the same cluster to distances in the nearest cluster. If a given member has a silhouette score close to 1, it is both well matched to its cluster and poorly matched to other clusters. If many members have negative silhouette scores, there are either too many or too few clusters.

Recent attempts at using these methods to select optimal NWP clusters have shown some promise, but are currently too inconsistent to be used reliably. We expect that any successful optimal cluster algorithm would satisfy three main criteria:

1. The algorithm should be simple to understand, such that users can easily follow the train of logic and understand the reasons behind a given output being chosen.
2. Evaluated over many forecasts, the optimal cluster size should show sensible leadtime distributions. In other words, the algorithm should preferentially choose smaller cluster sizes at earlier leadtimes and larger cluster sizes at later leadtimes. However, this property should not be so rigidly enforced that the optimal cluster algorithm is unresponsive to the forecast weather. For instance, a predictable, low-spread event should be assigned a smaller cluster size regardless of whether it occurs at a shorter or longer leadtime.
3. Related to the above example, in the cases where the number of clusters is obvious from visual inspection, the optimal cluster algorithm should match that number. A handful of real-world cases should be chosen which represent a broad range of situations, from low-spread, low-cluster events, to forecasts with larger spread containing an obvious number of distinct scenarios. In most cases, the optimal cluster size should agree with intuition, or at the very least should not depart too far from it.

Previous attempts at using the elbow method have been unsuccessful since many forecasts lack the distinct elbow needed to make a confident assessment (Boykin 2022). In other tests, the silhouette method was moderately successful at choosing the correct cluster size from a limited set of events, but is also more complicated than other methods. In a routinely running experimental tool at the UKMO, the optimal cluster size is found simply as the smallest cluster size at which all representative members are evaluated as being more than 250 km apart. This threshold was chosen as an informed estimate based on the size of the domain and typical distance at which we expect the clusters to become distinct. A more intelligent threshold would need to account for changes in the domain size, and would also likely vary with parameter and time of year. Additionally, there has been little research into the distributions that this choice of algorithm and threshold produces. However, from a handful of subjective evaluations, it performs well at choosing an appropriate number of clusters in various scenarios.

To help determine the best strategy for choosing the optimal cluster size, more work should be conducted to rigorously define the desired criteria stated above.

For instance, we do not fully understand the typical leadtime distributions that the algorithm should produce – at what point should we expect the algorithm to preferentially choose $k = 2$ over $k = 1$, for example, or choose $k = 3$ over $k = 2$? And how often should we expect the algorithm to choose $k = 4$ at leadtimes approaching T+120 h when the percentage of explained variance does not change by much compared to $k = 3$? Additionally, to build confidence in the third criteria, a reasonable number of cases should be evaluated by a diverse group of users. A testbed using clustering outputs from the most recent model runs would be an ideal setting for collecting a large group of responses that is likely to provide both clear and uncertain cases. The relative frequency of cases that are subjectively evaluated as clear or uncertain will also be a useful point of information for judging the potential efficacy of the ideal optimal cluster algorithm.

Bibliography

Adams, S. V. et al. (2019). “LFRic: Meeting the challenges of scalability and performance portability in Weather and Climate models”. *Journal of Parallel and Distributed Computing* 132, pp. 383–396. DOI: 10.1016/j.jpdc.2019.02.007 (cit. on p. 27).

Alhamed, A., S. Lakshmivarahan, and D. J. Stensrud (2002). “Cluster Analysis of Multimodel Ensemble Data from SAMEX”. *Monthly Weather Review* 130.2, pp. 226–256. DOI: 10.1175/1520-0493(2002)130<0226:CAOMED>2.0.CO;2 (cit. on p. 139).

AMS (2023). *Drizzle - Glossary of Meteorology American Meteorology Society.* URL: https://glossary.ametsoc.org/wiki/Drizzle (visited on 09/26/2023) (cit. on p. 60).

Ancell, B. and G. J. Hakim (2007). “Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting”. *Monthly Weather Review* 135.12, pp. 4117–4134. DOI: 10.1175/2007MWR1904.1 (cit. on p. 41).

Arakawa, A. and W. H. Schubert (1974). “Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I”. *Journal of the Atmospheric Sciences* 31.3, pp. 674–701. DOI: 10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2 (cit. on p. 20).

Arakawa, A. and C.-M. Wu (2013). “A Unified Representation of Deep Moist Convection in Numerical Modeling of the Atmosphere. Part I”. *Journal of the Atmospheric Sciences* 70.7, pp. 1977–1992. DOI: 10.1175/JAS-D-12-0330.1 (cit. on p. 21).

Atger, F. (1999). “Tubing: An Alternative to Clustering for the Classification of Ensemble Forecasts”. *Weather and Forecasting* 14.5, pp. 741–757. DOI: 10.1175/1520-0434(1999)014<0741:TAATCF>2.0.CO;2 (cit. on pp. 39, 115, 116).

Baker, L. H. et al. (2014). “Representation of model error in a convective-scale ensemble prediction system”. *Nonlinear Processes in Geophysics* 21.1, pp. 19–39. DOI: 10.5194/npg-21-19-2014 (cit. on p. 26).

Baran, S. and S. Lerch (2015). “Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting”. *Quarterly Journal of the Royal Meteorological Society* 141.691, pp. 2289–2299. DOI: 10.1002/qj.2521 (cit. on p. 43).

Barrett, A. I. et al. (2015). “Synoptic versus orographic control on stationary convective banding”. *Quarterly Journal of the Royal Meteorological Society* 141.689, pp. 1101–1113. DOI: 10.1002/qj.2409 (cit. on pp. 19, 20, 33, 81).

Barrett, A. I. et al. (2016). “The Utility of Convection-Permitting Ensembles for the Prediction of Stationary Convective Bands”. *Monthly Weather Review* 144.3, pp. 1093–1114. DOI: 10.1175/MWR-D-15-0148.1 (cit. on pp. 19, 20, 33, 81, 89, 103).

Barrett, P. et al. (2021). *WesCon 2023: Wessex UK Summertime Convection Field Campaign.* EGU21-2357. Copernicus Meetings. DOI: 10.5194/egusphere-egu21-2357 (cit. on p. 104).

Beck, J. et al. (2016). “Development and verification of two convection-allowing multi-model ensembles over Western Europe”. *Quarterly Journal of the Royal Meteorological Society* 142.700, pp. 2808–2826. DOI: 10.1002/qj.2870 (cit. on pp. 2, 34, 36, 47, 48, 81, 82).

Bednarczyk, C. N. and B. C. Ancell (2015). “Ensemble Sensitivity Analysis Applied to a Southern Plains Convective Event”. *Monthly Weather Review* 143.1, pp. 230–249. DOI: 10.1175/MWR-D-13-00321.1 (cit. on p. 41).

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt (2013). “Enhancing COSMO-DE ensemble forecasts by inexpensive techniques”. *Meteorologische Zeitschrift*, pp. 49–59. DOI: 10.1127/0941-2948/2013/0374 (cit. on pp. 48, 81, 82).

Bengtsson, L. et al. (2017). “The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System”. *Monthly Weather Review* 145.5, pp. 1919–1935. DOI: 10.1175/MWR-D-16-0417.1 (cit. on pp. 27, 47).

Berner, A. J. et al. (2009). “A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system”. *Journal of the Atmospheric Sciences* (cit. on p. 10).

Berner, J. et al. (2011). “Model Uncertainty in a Mesoscale Ensemble Prediction System: Stochastic versus Multiphysics Representations”. *Monthly Weather Review* 139.6, pp. 1972–1995. DOI: 10.1175/2010MWR3595.1 (cit. on pp. 9, 10).

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2001). “Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects”. *Monthly Weather Review* 129.3, pp. 420–436. DOI: 10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2 (cit. on p. 9).

Blunn, L. P. et al. (2024). “The influence of resolved convective motions on scalar dispersion in hectometric-scale numerical weather prediction models”. *Quarterly Journal of the Royal Meteorological Society*, qj.4632. DOI: 10.1002/qj.4632 (cit. on p. 104).

Bouttier, F. and L. Raynaud (2018). “Clustering and selection of boundary conditions for limited-area ensemble prediction”. *Quarterly Journal of the Royal Meteorological Society* 144.717, pp. 2381–2391. DOI: 10.1002/qj.3304 (cit. on pp. 39, 115).

Bouttier, F. et al. (2012). “Impact of Stochastic Physics in a Convection-Permitting Ensemble”. *Monthly Weather Review* 140.11, pp. 3706–3721. DOI: 10.1175/MWR-D-12-00031.1 (cit. on p. 26).

Bouttier, F. et al. (2016). “Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX”. *Quarterly Journal of the Royal Meteorological Society* 142 (S1), pp. 390–403. DOI: 10.1002/qj.2622 (cit. on p. 37).

Bowler, N. E. et al. (2008). “The MOGREPS short-range ensemble prediction system”. *Quarterly Journal of the Royal Meteorological Society* 134.632, pp. 703–722. DOI: 10.1002/qj.234 (cit. on pp. 9, 10).

Boykin, K. A. (2022). “Extracting Likely Scenarios from Ensemble Forecasts in Real-time”. PhD thesis (cit. on pp. 3, 39, 115–117, 120, 137, 151, 155).

Branković, Č. et al. (2008). “Downscaling of ECMWF Ensemble Forecasts for Cases of Severe Weather: Ensemble Statistics and Cluster Analysis”. *Monthly Weather Review* 136.9, pp. 3323–3342. DOI: 10.1175/2008MWR2322.1 (cit. on pp. 39, 115, 117, 122).

Brier, G. W. (1950). “VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY”. *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2 (cit. on p. 13).

Brill, K. F., A. R. Fracasso, and C. M. Bailey (2015). “Applying a Divisive Clustering Algorithm to a Large Ensemble for Medium-Range Forecasting at the Weather Prediction Center”. *Weather and Forecasting* 30.4, pp. 873–891. DOI: 10.1175/WAF-D-14-00137.1 (cit. on pp. 115, 116, 129, 138, 139).

Bröcker, J. and H. Kantz (2011). “The concept of exchangeability in ensemble forecasting”. *Nonlinear Processes in Geophysics* 18.1, pp. 1–5. DOI: 10.5194/npg-18-1-2011 (cit. on pp. 8, 96).

Brousseau, P. et al. (2016). “Improvement of the forecast of convective activity from the AROME-France system”. *Quarterly Journal of the Royal Meteorological Society* 142.699, pp. 2231–2243. DOI: 10.1002/qj.2822 (cit. on pp. 27, 30).

Brusco, M. J., D. Steinley, and J. Stevens (2019). “K-medoids inverse regression”. *Communications in Statistics - Theory and Methods* 48.20, pp. 4999–5011. DOI: 10.1080/03610926.2018.1504076 (cit. on p. 40).

Buizza, R. and T. N. Palmer (1995). “The Singular-Vector Structure of the Atmospheric Global Circulation”. *Journal of the Atmospheric Sciences* 52.9, pp. 1434–1456. DOI: 10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2 (cit. on p. 8).

Buizza, R. et al. (1998). “Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System”. *Quarterly Journal of the Royal Meteorological Society* 124.550, pp. 1935–1960. DOI: 10.1002/qj.49712455008 (cit. on pp. 14, 15).

Buizza, R. et al. (1999). “Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System”. *Weather and Forecasting* 14.2, pp. 168–189. DOI: 10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2 (cit. on pp. 9, 10, 12, 13, 16).

Buizza, R. (1997). “Potential Forecast Skill of Ensemble Prediction and Spread and Skill Distributions of the ECMWF Ensemble Prediction System”. *Monthly Weather Review* 125.1, pp. 99–119. DOI: 10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2 (cit. on pp. 14, 17, 47, 81).

Bush, M. et al. (2023). “The second Met Office Unified Model–JULES Regional Atmosphere and Land configuration, RAL2”. *Geoscientific Model Development* 16.6, pp. 1713–1734. DOI: 10.5194/gmd-16-1713-2023 (cit. on pp. 17, 51, 61).

Cabinet Office (2008). *Learning Lessons from the 2007 Floods: An Independent Review by Sir Michael Pitt* (cit. on p. 42).

Cafaro, C. et al. (2019). "The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble". *Quarterly Journal of the Royal Meteorological Society* 145.721, pp. 1780–1798. DOI: `10.1002/qj.3531` (cit. on pp. 1, 21, 31, 33, 81, 103, 117).

Cafaro, C. et al. (2021). "Do Convection-Permitting Ensembles Lead to More Skillful Short-Range Probabilistic Rainfall Forecasts over Tropical East Africa?" *Weather and Forecasting* 36.2, pp. 697–716. DOI: `10.1175/WAF-D-20-0172.1` (cit. on pp. 2, 36, 47, 81, 82, 89).

Candille, G. et al. (2007). "Verification of an Ensemble Prediction System against Observations". *Monthly Weather Review* 135.7, pp. 2688–2699. DOI: `10.1175/MWR3414.1` (cit. on p. 12).

Candille, G. (2009). "The Multiensemble Approach: The NAEFS Example". *Monthly Weather Review* 137.5, pp. 1655–1665. DOI: `10.1175/2008MWR2682.1` (cit. on p. 41).

Caron, J.-F. (2013). "Mismatching Perturbations at the Lateral Boundaries in Limited-Area Ensemble Forecasting: A Case Study". *Monthly Weather Review* 141.1, pp. 356–374. DOI: `10.1175/MWR-D-12-00051.1` (cit. on p. 48).

Charney, J. G. (1951). "Dynamic Forecasting by Numerical Process". In: *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology.* Ed. by H. R. Byers et al. Boston, MA: American Meteorological Society, pp. 470–482. DOI: `10.1007/978-1-940033-70-9_40` (cit. on p. 5).

Charron, M. et al. (2010). "Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System". *Monthly Weather Review* 138.5, pp. 1877–1901. DOI: `10.1175/2009MWR3187.1` (cit. on pp. 9, 10).

Christensen, H. M., I. M. Moroz, and T. N. Palmer (2015). "Stochastic and Perturbed Parameter Representations of Model Uncertainty in Convection Parameterization". *Journal of the Atmospheric Sciences* 72.6, pp. 2525–2544. DOI: `10.1175/JAS-D-14-0250.1` (cit. on p. 10).

Clark, A. J. et al. (2009). "A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles". *Weather and Forecasting* 24.4, pp. 1121–1140. DOI: `10.1175/2009WAF2222222.1` (cit. on pp. 1, 2, 20, 21, 31, 33, 34, 36, 81, 82, 151).

— (2010). "Growth of Spread in Convection-Allowing and Convection-Parameterizing Ensembles". *Weather and Forecasting* 25.2, pp. 594–612. DOI: `10.1175/2009WAF2222318.1` (cit. on pp. 2, 32, 36, 82, 103, 151).

Clark, A. J. et al. (2011). "Probabilistic Precipitation Forecast Skill as a Function of Ensemble Size and Spatial Scale in a Convection-Allowing Ensemble". *Monthly Weather Review* 139.5, pp. 1410–1418. DOI: `10.1175/2010MWR3624.1` (cit. on pp. 26, 34, 47).

Clark, P. et al. (2016). "Convection-permitting models: a step-change in rainfall forecasting". *Meteorological Applications* 23.2, pp. 165–181. DOI: 10.1002/met.1538 (cit. on pp. 17, 20, 21, 117).

Courant, R., K. Friedrichs, and H. Lewy (1928). "On the Partial Difference Equations of Mathematical Physics". *Mathematische Annalen* (cit. on p. 19).

Craig, G. C. et al. (2022). "Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble". *Quarterly Journal of the Royal Meteorological Society* 148.746, pp. 2325–2343. DOI: 10.1002/qj.4305 (cit. on pp. 14–16, 37, 116).

Craven, J. P., D. E. Rudack, and P. E. Shafer (2020). "National Blend of Models: A Statistically Post-Processed Multi-Model Ensemble". *Journal of Operational Meteorology*, pp. 1–14. DOI: 10.15191/nwajom.2020.0801 (cit. on p. 43).

Cunningham, C., J. P. Bonatti, and M. Ferreira (2015). "Assessing improved CPTEC probabilistic forecasts on medium-range timescale". *Meteorological Applications* 22.3, pp. 378–384. DOI: 10.1002/met.1464 (cit. on p. 16).

Davis, C. A. et al. (2009). "The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program". *Weather and Forecasting* 24.5, pp. 1252–1267. DOI: 10.1175/2009WAF2222241.1 (cit. on p. 34).

Dey, S. R. A. et al. (2014). "A Spatial View of Ensemble Spread in Convection Permitting Ensembles". *Monthly Weather Review* 142.11, pp. 4091–4107. DOI: 10.1175/MWR-D-14-00172.1 (cit. on pp. 3, 36, 56, 81, 87, 88).

Dey, S. R. A. et al. (2016). "Assessing spatial precipitation uncertainties in a convective-scale ensemble". *Quarterly Journal of the Royal Meteorological Society* 142.701, pp. 2935–2948. DOI: 10.1002/qj.2893 (cit. on p. 56).

Dione, C. et al. (2022). "Improved sub-seasonal forecasts to support preparedness action for meningitis outbreak in Africa". *Climate Services* 28, p. 100326. DOI: 10.1016/j.cliser.2022.100326 (cit. on p. 43).

Dixon, M. et al. (2009). "Impact of Data Assimilation on Forecasting Convection over the United Kingdom Using a High-Resolution Version of the Met Office Unified Model". *Monthly Weather Review* 137.5, pp. 1562–1584. DOI: 10.1175/2008MWR2561.1 (cit. on p. 18).

Done, J. M. et al. (2006). "Mesoscale simulations of organized convection: Importance of convective equilibrium". *Quarterly Journal of the Royal Meteorological Society* 132.616, pp. 737–756. DOI: 10.1256/qj.04.84 (cit. on p. 23).

Done, J., C. A. Davis, and M. Weisman (2004). "The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model". *Atmospheric Science Letters* 5.6, pp. 110–117. DOI: 10.1002/asl.72 (cit. on p. 17).

Duc, L., K. Saito, and H. Seko (2013). "Spatial-temporal fractions verification for high-resolution ensemble forecasts". *Tellus A: Dynamic Meteorology and Oceanography* 65.1, p. 18171. DOI: 10.3402/tellusa.v65i0.18171 (cit. on pp. 20, 81, 88).

Durran, D. R. and J. A. Weyn (2016). “Thunderstorms Do Not Get Butterflies”. *Bulletin of the American Meteorological Society* 97.2, pp. 237–243. DOI: `10.1175/BAMS-D-15-00070.1` (cit. on pp. 22, 23).

Eady, E. T. (1949). “Long Waves and Cyclone Waves”. *Tellus* 1.3, pp. 33–52. DOI: `10.1111/j.2153-3490.1949.tb01265.x` (cit. on p. 23).

Ebert, E. E. (2001). “Ability of a Poor Man’s Ensemble to Predict the Probability and Distribution of Precipitation”. *Monthly Weather Review* 129.10, pp. 2461–2480. DOI: `10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2` (cit. on p. 14).

ECMWF (2022). *The next extended-range configuration for IFS Cycle 48r1*. ECMWF. URL: `https://www.ecmwf.int/en/newsletter/173/earth-system-science/next-extended-range-configuration-ifs-cycle-48r1` (visited on 05/15/2025) (cit. on p. 16).

— (2023). *Windstorm Poly - Forecast User - ECMWF Confluence Wiki*. URL: `https://confluence.ecmwf.int/display/FCST/202307+-+Windstorm+-+Poly` (visited on 09/23/2024) (cit. on p. 109).

Epstein, E. S. (1969). “Stochastic dynamic prediction”. *Tellus* 21.6, pp. 739–759. DOI: `10.1111/j.2153-3490.1969.tb00483.x` (cit. on pp. 6, 8).

Faggian, N. et al. (2015). “Fast calculation of the Fractions Skill Score”. *Mausam* 66, pp. 457–466. DOI: `10.54302/mausam.v66i3.555` (cit. on p. 142).

Feng, J., J. Sun, and Y. Zhang (2020). “A Dynamic Blending Scheme to Mitigate Large-Scale Bias in Regional Models”. *Journal of Advances in Modeling Earth Systems* 12.3, e2019MS001754. DOI: `10.1029/2019MS001754` (cit. on p. 49).

Feng, J. et al. (2021). “An Implementation of Full Cycle Strategy Using Dynamic Blending for Rapid Refresh Short-range Weather Forecasting in China”. *Advances in Atmospheric Sciences* 38.6, pp. 943–956. DOI: `10.1007/s00376-021-0316-7` (cit. on p. 49).

Fereday, D. R. et al. (2008). “Cluster Analysis of North Atlantic–European Circulation Types and Links with Tropical Pacific Sea Surface Temperatures”. *Journal of Climate* 21.15, pp. 3687–3703. DOI: `10.1175/2007JCLI1875.1` (cit. on pp. 39, 115).

Ferranti, L. and S. Corti (2011). “New clustering products”. DOI: `10.21957/LR3BCISE` (cit. on pp. 39, 115, 116, 151).

Ferranti, L., S. Corti, and M. Janousek (2015). “Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector”. *Quarterly Journal of the Royal Meteorological Society* 141.688, pp. 916–924. DOI: `10.1002/qj.2411` (cit. on pp. 39, 40, 116).

Ferrett, S. et al. (2021). “Evaluating Convection-Permitting Ensemble Forecasts of Precipitation over Southeast Asia”. *Weather and Forecasting* 36.4, pp. 1199–1217. DOI: `10.1175/WAF-D-20-0216.1` (cit. on pp. 36, 43, 47, 58, 81, 151).

Ferro, C. a. T. (2014). “Fair scores for ensemble forecasts”. *Quarterly Journal of the Royal Meteorological Society* 140.683, pp. 1917–1923. DOI: `10.1002/qj.2270` (cit. on p. 16).

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel (2008). "On the effect of ensemble size on the discrete and continuous ranked probability scores". *Meteorological Applications* 15.1, pp. 19–24. DOI: `10.1002/met.45` (cit. on p. 16).

Fischer, H. (2011). *A History of the Central Limit Theorem: From Classical to Modern Probability Theory.* New York, NY: Springer New York. DOI: `10.1007/978-0-387-87857-7` (cit. on p. 15).

Flack, D. L. A. et al. (2021). "A Physically Based Stochastic Boundary Layer Perturbation Scheme. Part II: Perturbation Growth within a Superensemble Framework". *Journal of the Atmospheric Sciences* 78.3, pp. 747–761. DOI: `10.1175/JAS-D-19-0292.1` (cit. on pp. 2, 16, 26, 32, 37, 81, 82).

Flack, D. L. A. et al. (2016). "Characterisation of convective regimes over the British Isles". *Quarterly Journal of the Royal Meteorological Society* 142.696, pp. 1541–1553. DOI: `10.1002/qj.2758` (cit. on p. 24).

Flack, D. L. A. et al. (2018). "Convective-Scale Perturbation Growth across the Spectrum of Convective Regimes". *Monthly Weather Review* 146.1, pp. 387–405. DOI: `10.1175/MWR-D-17-0024.1` (cit. on p. 32).

Flack, D. L. A. et al. (2023). "Characteristics of Diagnostics for Identifying Elevated Convection over the British Isles in a Convection-Allowing Model". *Weather and Forecasting* 38.7, pp. 1079–1094. DOI: `10.1175/WAF-D-22-0219.1` (cit. on pp. 71, 73).

Fosser, G. et al. (2024). "Convection-permitting climate models offer more certain extreme rainfall projections". *npj Climate and Atmospheric Science* 7.1, p. 51. DOI: `10.1038/s41612-024-00600-w` (cit. on p. 38).

Fowler, L. D., M. C. Barth, and K. Alapaty (2020). "Impact of scale-aware deep convection on the cloud liquid and ice water paths and precipitation using the Model for Prediction Across Scales (MPAS-v5.2)". *Geoscientific Model Development* 13.6, pp. 2851–2877. DOI: `10.5194/gmd-13-2851-2020` (cit. on p. 19).

Frogner, I.-L. et al. (2019a). "Convection-permitting ensembles: Challenges related to their design and use". *Quarterly Journal of the Royal Meteorological Society* 145 (S1), pp. 90–106. DOI: `10.1002/qj.3525` (cit. on pp. 2, 16, 20, 31–34, 36, 37, 81, 82).

Frogner, I.-L. et al. (2019b). "HarmonEPS—The HARMONIE Ensemble Prediction System". *Weather and Forecasting* 34.6, pp. 1909–1937. DOI: `10.1175/WAF-D-19-0030.1` (cit. on pp. 1, 27, 30, 34, 36, 37, 115).

Gainford, A. et al. (2024). "Improvements in the spread–skill relationship of precipitation in a convective-scale ensemble through blending". *Quarterly Journal of the Royal Meteorological Society* n/a (n/a). DOI: `10.1002/qj.4754` (cit. on pp. 82, 89, 119).

Gallo, B. T., A. J. Clark, and S. R. Dembek (2016). "Forecasting Tornadoes Using Convection-Permitting Ensembles". *Weather and Forecasting* 31.1, pp. 273–295. DOI: `10.1175/WAF-D-15-0134.1` (cit. on pp. 20, 38, 81).

Gebhardt, C. et al. (2011). "Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries". *Atmospheric Re-*

search 100.2, pp. 168–177. DOI: 10.1016/j.atmosres.2010.12.008 (cit. on pp. 32, 81, 82, 116).

Gerard, L. et al. (2009). "Cloud and Precipitation Parameterization in a Meso-Gamma-Scale Operational Weather Prediction Model". *Monthly Weather Review* 137.11, pp. 3960–3977. DOI: 10.1175/2009MWR2750.1 (cit. on p. 21).

Gilleland, E. et al. (2009). "Intercomparison of Spatial Forecast Verification Methods". *Weather and Forecasting* 24.5, pp. 1416–1430. DOI: 10.1175/2009WAF2222269.1 (cit. on pp. 3, 15, 34, 55, 117).

Gilleland, E. et al. (2010). "Verifying Forecasts Spatially". *Bulletin of the American Meteorological Society* 91.10, pp. 1365–1376. DOI: 10.1175/2010BAMS2819.1 (cit. on p. 15).

Glazer, R. H., E. Coppola, and F. Giorgi (2025). "Understanding Nocturnally-driven Extreme Precipitation Events over Lake Victoria in a Convection-Permitting Model". *Monthly Weather Review* -1 (aop). DOI: 10.1175/MWR-D-22-0339.1 (cit. on p. 20).

Gneiting, T. et al. (2005). "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation". *Monthly Weather Review* 133.5, pp. 1098–1118. DOI: 10.1175/MWR2904.1 (cit. on p. 42).

Golding, B. W. (1998). "Nimrod: a system for generating automated very short range forecasts". *Meteorological Applications* 5.1, pp. 1–16. DOI: 10.1017/S1350482798000577 (cit. on pp. 37, 55, 87, 130).

Golding, B., ed. (2022). *Towards the "Perfect" Weather Warning: Bridging Disciplinary Gaps through Partnership and Communication.* Cham: Springer International Publishing. DOI: 10.1007/978-3-030-98989-7 (cit. on p. 43).

Golding, B. et al. (2016). "MOGREPS-UK Convection-Permitting Ensemble Products for Surface Water Flood Forecasting: Rationale and First Results". *Journal of Hydrometeorology* 17.5, pp. 1383–1406. DOI: 10.1175/JHM-D-15-0083.1 (cit. on p. 42).

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz (2018). "Validation of Mountain Precipitation Forecasts from the Convection-Permitting NCAR Ensemble and Operational Forecast Systems over the Western United States". *Weather and Forecasting* 33.3, pp. 739–765. DOI: 10.1175/WAF-D-17-0144.1 (cit. on pp. 19, 20, 81).

Gray, S. L. et al. (2021). "Development of a prototype real-time sting-jet precursor tool for forecasters". *Weather* 76.11, pp. 369–373. DOI: 10.1002/wea.3889 (cit. on p. 21).

Guidard, V. and C. Fischer (2008). "Introducing the coupling information in a limited-area variational assimilation". *Quarterly Journal of the Royal Meteorological Society* 134.632, pp. 723–735. DOI: 10.1002/qj.215 (cit. on pp. 19, 47).

Hagelin, S. et al. (2017). "The Met Office convective-scale ensemble, MOGREPS-UK". *Quarterly Journal of the Royal Meteorological Society* 143.708, pp. 2846–2861. DOI: 10.1002/qj.3135 (cit. on pp. 18, 19, 81, 84, 88, 118).

Hamill, T. M. (1999). "Hypothesis Tests for Evaluating Numerical Precipitation Forecasts". *Weather and Forecasting* 14.2, pp. 155–167. DOI: 10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2 (cit. on p. 75).

— (2001). “Interpretation of Rank Histograms for Verifying Ensemble Forecasts”. *Monthly Weather Review* 129.3, pp. 550–560. DOI: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2 (cit. on pp. 12, 37).

Hamill, T. M., C. Snyder, and R. E. Morss (2000). “A Comparison of Probabilistic Forecasts from Bred, Singular-Vector, and Perturbed Observation Ensembles”. *Monthly Weather Review* 128.6, pp. 1835–1851. DOI: 10.1175/1520-0493(2000)128<1835:ACOPFF>2.0.CO;2 (cit. on p. 9).

Hanley, K. E. et al. (2011). “Ensemble predictability of an isolated mountain thunderstorm in a high-resolution model”. *Quarterly Journal of the Royal Meteorological Society* 137.661, pp. 2124–2137. DOI: 10.1002/qj.877 (cit. on pp. 1, 20, 81, 117).

Hanley, K. E. et al. (2013). “Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble”. *Monthly Weather Review* 141.1, pp. 112–133. DOI: 10.1175/MWR-D-12-00013.1 (cit. on pp. 20, 33, 41, 81).

Hanley, K. E. and H. W. Lean (2024). “The performance of a variable-resolution 300-m ensemble for forecasting convection over London”. *Quarterly Journal of the Royal Meteorological Society* 150.763, pp. 3737–3756. DOI: 10.1002/qj.4794 (cit. on pp. 19, 104).

Hanley, K. E. et al. (2015). “Mixing-length controls on high-resolution simulations of convective storms”. *Quarterly Journal of the Royal Meteorological Society* 141.686, pp. 272–284. DOI: 10.1002/qj.2356 (cit. on p. 22).

Harr, P. A., D. Anwender, and S. C. Jones (2008). “Predictability Associated with the Downstream Impacts of the Extratropical Transition of Tropical Cyclones: Methodology and a Case Study of Typhoon Nabi (2005)”. *Monthly Weather Review* 136.9, pp. 3205–3225. DOI: 10.1175/2008MWR2248.1 (cit. on p. 40).

Heinzeller, D., M. G. Duda, and H. Kunstmann (2016). “Towards convection-resolving, global atmospheric simulations with the Model for Prediction Across Scales (MPAS) v3.1: an extreme scaling experiment”. *Geoscientific Model Development* 9.1, pp. 77–110. DOI: 10.5194/gmd-9-77-2016 (cit. on p. 19).

Hersbach, H. (2000). “Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems”. *Weather and Forecasting* 15.5, pp. 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2 (cit. on p. 13).

Hohenegger, C. and C. Schar (2007). “Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales”. *Bulletin of the American Meteorological Society* 88.11, pp. 1783–1794. DOI: 10.1175/BAMS-88-11-1783 (cit. on pp. 23, 116).

Hohenegger, C. et al. (2008). “Cloud-resolving ensemble simulations of the August 2005 Alpine flood”. *Quarterly Journal of the Royal Meteorological Society* 134.633, pp. 889–904. DOI: 10.1002/qj.252 (cit. on pp. 32, 82, 95, 103).

Hopson, T. M. (2014). “Assessing the Ensemble Spread–Error Relationship”. *Monthly Weather Review* 142.3, pp. 1125–1142. DOI: 10.1175/MWR-D-12-00111.1 (cit. on pp. 47, 81).

Houtekamer, P. L. (1993). "Global and Local Skill Forecasts". *Monthly Weather Review* 121.6, pp. 1834–1846. DOI: 10.1175/1520-0493(1993)121<1834:GALSF>2.0.CO;2 (cit. on p. 14).

Hsiao, L.-F. 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* 30.3, pp. 754–770. DOI: 10.1175/WAF-D-14-00047.1 (cit. on p. 48).

Inverarity, G. W. 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* n/a (n/a). DOI: 10.1002/qj.4431 (cit. on pp. 9, 48, 50, 81, 84, 95, 115, 118).

Johnson, A. et al. (2011a). "Hierarchical Cluster Analysis of a Convection-Allowing Ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the Object-Oriented Cluster Analysis Method for Precipitation Fields". *Monthly Weather Review* 139.12, pp. 3673–3693. DOI: 10.1175/MWR-D-11-00015.1 (cit. on pp. 115, 117).

Johnson, A. et al. (2011b). "Hierarchical Cluster Analysis of a Convection-Allowing Ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble Clustering over the Whole Experiment Period". *Monthly Weather Review* 139.12, pp. 3694–3710. DOI: 10.1175/MWR-D-11-00016.1 (cit. on p. 117).

Johnson, C. and N. Bowler (2009). "On the Reliability and Calibration of Ensemble Forecasts". *Monthly Weather Review* 137.5, pp. 1717–1720. DOI: 10.1175/2009MWR2715.1 (cit. on p. 11).

Kalnay, E. (2002). *Atmospheric Modeling, Data Assimilation and Predictability.* Higher Education from Cambridge University Press. DOI: 10.1017/CBO9780511802270. URL: https://www.cambridge.org/highereducation/books/atmospheric-modeling-data-assimilation-and-predictability/C5FD207439132836E85027754CE9BC1A (visited on 04/22/2025) (cit. on p. 5).

Keil, C. and G. Craig (2011). "Regime-dependent forecast uncertainty of convective precipitation". *Meteorologische Zeitschrift* 20, pp. 145–151. DOI: 10.1127/0941-2948/2011/0219 (cit. on p. 24).

Keil, C., F. Heinlein, and G. C. Craig (2014). "The convective adjustment time-scale as indicator of predictability of convective precipitation". *Quarterly Journal of the Royal Meteorological Society* 140.679, pp. 480–490. DOI: 10.1002/qj.2143 (cit. on pp. 24, 82).

Keller, J. H. et al. (2011). "Characteristics of the TIGGE multimodel ensemble prediction system in representing forecast variability associated with extratropical transition". *Geophysical Research Letters* 38.12. DOI: 10.1029/2011GL047275 (cit. on p. 40).

Keresturi, E. et al. (2019). "Improving initial condition perturbations in a convection-permitting ensemble prediction system". *Quarterly Journal of the Royal Meteorological Society* 145.720, pp. 993–1012. DOI: 10.1002/qj.3473 (cit. on pp. 47, 48, 73, 150).

Kim, S. et al. (2015). “Development and Evaluation of the High Resolution Limited Area Ensemble Prediction System in the Korea Meteorological Administration”. *Atmosphere* 25.1, pp. 67–83. DOI: `10.14191/Atmos.2015.25.1.067` (cit. on pp. 28, 30).

Klasa, C. et al. (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, pp. 744–764. DOI: `10.1002/qj.3245` (cit. on pp. 2, 34, 36, 47, 82, 151).

Kober, K. and G. C. Craig (2016). “Physically Based Stochastic Perturbations (PSP) in the Boundary Layer to Represent Uncertainty in Convective Initiation”. *Journal of the Atmospheric Sciences* 73.7, pp. 2893–2911. DOI: `10.1175/JAS-D-15-0144.1` (cit. on p. 26).

Komaromi, W. A. et al. (2021). “The Naval Research Laboratory’s Coupled Ocean–Atmosphere Mesoscale Prediction System-Tropical Cyclone Ensemble (COAMPS-TC Ensemble)”. *Weather and Forecasting* 36.2, pp. 499–517. DOI: `10.1175/WAF-D-20-0038.1` (cit. on pp. 28, 30).

Kühnlein, C. et al. (2014). “The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation”. *Quarterly Journal of the Royal Meteorological Society* 140.682, pp. 1552–1562. DOI: `10.1002/qj.2238` (cit. on pp. 32, 82, 116).

Lamberson, W. S. et al. (2023). “The Use of Ensemble Clustering on a Multimodel Ensemble for Medium-Range Forecasting at the Weather Prediction Center”. *Weather and Forecasting* 38.4, pp. 539–554. DOI: `10.1175/WAF-D-22-0154.1` (cit. on pp. 40, 115, 116, 139, 151).

Lean, H. W. et al. (2008). “Characteristics of High-Resolution Versions of the Met Office Unified Model for Forecasting Convection over the United Kingdom”. *Monthly Weather Review* 136.9, pp. 3408–3424. DOI: `10.1175/2008MWR2332.1` (cit. on pp. 1, 17, 18, 20–22, 26, 116).

Lean, H. W. et al. (2024). “The hectometric modelling challenge: Gaps in the current state of the art and ways forward towards the implementation of 100-m scale weather and climate models”. *Quarterly Journal of the Royal Meteorological Society* n/a (n/a). DOI: `10.1002/qj.4858` (cit. on pp. 1, 22).

Lee, S. H. and G. Messori (2024). “The Dynamical Footprint of Year-Round North American Weather Regimes”. *Geophysical Research Letters* 51.2, e2023GL107161. DOI: `10.1029/2023GL107161` (cit. on pp. 39, 116).

Lee, S. H., M. K. Tippett, and L. M. Polvani (2023). “A New Year-Round Weather Regime Classification for North America”. *Journal of Climate* 36.20, pp. 7091–7108. DOI: `10.1175/JCLI-D-23-0214.1` (cit. on pp. 39, 116).

Leith, C. E. (1971). “Atmospheric Predictability and Two-Dimensional Turbulence”. *Journal of the Atmospheric Sciences* 28.2, pp. 145–161. DOI: `10.1175/1520-0469(1971)028<0145:APATDT>2.0.CO;2` (cit. on p. 8).

Leith, C. E. (1974). “Theoretical Skill of Monte Carlo Forecasts”. *Monthly Weather Review* 102.6, pp. 409–418. DOI: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2 (cit. on pp. 8, 15).

Leoncini, G. et al. (2013). “Ensemble forecasts of a flood-producing storm: comparison of the influence of model-state perturbations and parameter modifications”. *Quarterly Journal of the Royal Meteorological Society* 139.670, pp. 198–211. DOI: 10.1002/qj.1951 (cit. on p. 26).

Leutbecher, M. (2019). “Ensemble size: How suboptimal is less than infinity?” *Quarterly Journal of the Royal Meteorological Society* 145 (S1), pp. 107–128. DOI: 10.1002/qj.3387 (cit. on pp. 14, 15).

Lewis, J. M. (2005). “Roots of Ensemble Forecasting”. *Monthly Weather Review* 133.7, pp. 1865–1885. DOI: 10.1175/MWR2949.1 (cit. on p. 6).

Lorenz, E. N. (1963a). “Deterministic Nonperiodic Flow”. *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2 (cit. on p. 6).

— (1963b). “The Predictability of Hydrodynamic Flow”. *Transactions of the New York Academy of Sciences* 25.4, pp. 409–432. DOI: 10.1111/j.2164-0947.1963.tb01464.x (cit. on p. 6).

— (1969). “The predictability of a flow which possesses many scales of motion”. *Tellus* 21.3, pp. 289–307. DOI: 10.1111/j.2153-3490.1969.tb00444.x (cit. on pp. 1, 6, 16, 22, 67).

Magnusson, L., M. Leutbecher, and E. Källén (2008). “Comparison between Singular Vectors and Breeding Vectors as Initial Perturbations for the ECMWF Ensemble Prediction System”. *Monthly Weather Review* 136.11, pp. 4092–4104. DOI: 10.1175/2008MWR2498.1 (cit. on p. 9).

Manning, C. et al. (2022). “Extreme windstorms and sting jets in convection-permitting climate simulations over Europe”. *Climate Dynamics* 58.9, pp. 2387–2404. DOI: 10.1007/s00382-021-06011-4 (cit. on p. 21).

Marsigli, C., A. Montani, and T. Paccagnella (2014). “Perturbation of initial and boundary conditions for a limited-area ensemble: multi-model versus single-model approach”. *Quarterly Journal of the Royal Meteorological Society* 140.678, pp. 197–208. DOI: 10.1002/qj.2128 (cit. on pp. 15, 36).

Marsigli, C. et al. (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, pp. 2095–2115. DOI: 10.1002/qj.49712757613 (cit. on pp. 39, 115).

Marsigli, C. et al. (2005). “The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification”. *Nonlinear Processes in Geophysics* 12.4, pp. 527–536. DOI: 10.5194/npg-12-527-2005 (cit. on pp. 20, 31, 33, 81).

Marsigli, C., A. Montani, and T. Paccangnella (2008). “A spatial verification method applied to the evaluation of high-resolution ensemble forecasts”. *Meteorological Applications* 15.1, pp. 125–143. DOI: `10.1002/met.65` (cit. on pp. 20, 31, 33, 81).

Martínez-Alvarado, O., F. Weidle, and S. L. Gray (2010). “Sting Jets in Simulations of a Real Cyclone by Two Mesoscale Models”. *Monthly Weather Review* 138.11, pp. 4054–4075. DOI: `10.1175/2010MWR3290.1` (cit. on p. 21).

Mason, S. J. and N. E. Graham (1999). “Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels”. *Weather and Forecasting* 14.5, pp. 713–725. DOI: `10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2` (cit. on p. 13).

Maybee, B. et al. (2024). “FOREWARNS: development and multifaceted verification of enhanced regional-scale surface water flood forecasts”. *Natural Hazards and Earth System Sciences* 24.4, pp. 1415–1436. DOI: `10.5194/nhess-24-1415-2024` (cit. on p. 43).

McCabe, A. et al. (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, pp. 2897–2910. DOI: `10.1002/qj.2876` (cit. on pp. 2, 10, 26, 48, 51, 81, 82).

Met Office (2025). *WCSSP Southeast Asia*. Met Office. URL: `https://www.metoffice.gov.uk/research/approach/collaboration/wcssp/weather-and-climate-science-for-service-partnership-southeast-asia` (visited on 06/13/2025) (cit. on p. 38).

Milan, M. et al. (2020). “Hourly 4D-Var in the Met Office UKV operational forecast model”. *Quarterly Journal of the Royal Meteorological Society* 146.728, pp. 1281–1301. DOI: `10.1002/qj.3737` (cit. on pp. 18, 19, 49, 50, 85).

Milan, M. et al. (2023). “Large-scale blending in an hourly 4D-Var framework for a numerical weather prediction model”. *Quarterly Journal of the Royal Meteorological Society* n/a (n/a). DOI: `10.1002/qj.4495` (cit. on pp. 2, 19, 47, 49, 52, 62, 73).

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, pp. 1487–1500. DOI: `10.1002/qj.135` (cit. on pp. 48, 81, 82).

— (2021). “A “Meta” Analysis of the Fractions Skill Score: The Limiting Case and Implications for Aggregation”. *Monthly Weather Review* 149.10, pp. 3491–3504. DOI: `10.1175/MWR-D-18-0106.1` (cit. on pp. 59, 64, 88, 121).

Molteni, F. et al. (1996). “The ECMWF Ensemble Prediction System: Methodology and validation”. *Quarterly Journal of the Royal Meteorological Society* 122.529, pp. 73–119. DOI: `10.1002/qj.49712252905` (cit. on pp. 8, 39).

Molteni, F. et al. (2001). “A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments”. *Quarterly Journal of the Royal Meteorological Society* 127.576, pp. 2069–2094. DOI: `10.1002/qj.49712757612` (cit. on pp. 39, 115, 151, 154).

Montani, A. et al. (2011). "Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges". *Tellus A: Dynamic Meteorology and Oceanography* 63.3 (cit. on pp. 25, 82, 115).

Mullen, S. L. and R. Buizza (2002). "The Impact of Horizontal Resolution and Ensemble Size on Probabilistic Forecasts of Precipitation by the ECMWF Ensemble Prediction System". *Weather and Forecasting* 17.2, pp. 173–191. DOI: `10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2` (cit. on p. 14).

Murphy, A. H. (1973). "A New Vector Partition of the Probability Score". *Journal of Applied Meteorology and Climatology* 12.4, pp. 595–600. DOI: `10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2` (cit. on p. 13).

Murphy, A. H. and E. S. Epstein (1989). "Skill Scores and Correlation Coefficients in Model Verification". *Monthly Weather Review* 117.3, pp. 572–582. DOI: `10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2` (cit. on p. 11).

Nachamkin, J. E. and J. Schmidt (2015). "Applying a Neighborhood Fractions Sampling Approach as a Diagnostic Tool". *Monthly Weather Review* 143.11, pp. 4736–4749. DOI: `10.1175/MWR-D-14-00411.1` (cit. on p. 56).

Nakaegawa, T. and M. Kanamitsu (2006). "Cluster Analysis of the Seasonal Forecast Skill of the NCEP SFM over the Pacific–North America Sector". *Journal of Climate* 19.1, pp. 123–138. DOI: `10.1175/JCLI3609.1` (cit. on p. 39).

Neal, R. et al. (2016). "A flexible approach to defining weather patterns and their application in weather forecasting over Europe". *Meteorological Applications* 23.3, pp. 389–400. DOI: `10.1002/met.1563` (cit. on pp. 39, 90, 116).

Neal, R. et al. (2024). "A seamless blended multi-model ensemble approach to probabilistic medium-range weather pattern forecasts over the UK". *Meteorological Applications* 31.1, e2179. DOI: `10.1002/met.2179` (cit. on pp. 39, 116, 139).

Neal, R. A. et al. (2014). "Ensemble based first guess support towards a risk-based severe weather warning service". *Meteorological Applications* 21.3, pp. 563–577. DOI: `10.1002/met.1377` (cit. on p. 42).

Necker, T. et al. (2020). "A convective-scale 1,000-member ensemble simulation and potential applications". *Quarterly Journal of the Royal Meteorological Society* 146.728, pp. 1423–1442. DOI: `10.1002/qj.3744` (cit. on p. 41).

Necker, T. et al. (2024). *The fractions skill score for ensemble forecast verification.* DOI: `10.22541/au.169169008.89657659/v2` (cit. on p. 89).

Nielsen, E. R. and R. S. Schumacher (2016). "Using Convection-Allowing Ensembles to Understand the Predictability of an Extreme Rainfall Event". *Monthly Weather Review* 144.10, pp. 3651–3676. DOI: `10.1175/MWR-D-16-0083.1` (cit. on p. 82).

Nuissier, O. et al. (2012). "Uncertainty of lateral boundary conditions in a convection-permitting ensemble: a strategy of selection for Mediterranean heavy precipitation events". *Natural Hazards and Earth System Sciences* 12.10, pp. 2993–3011. DOI: `10.5194/nhess-12-2993-2012` (cit. on pp. 39, 115).

Núñez Ocasio, K. M. and R. Rios-Berrios (2023). "African Easterly Wave Evolution and Tropical Cyclogenesis in a Pre-Helene (2006) Hindcast Using the Model for Prediction Across Scales-Atmosphere (MPAS-A)". *Journal of Advances in Modeling Earth Systems* 15.2, e2022MS003181. DOI: `10.1029/2022MS003181` (cit. on p. 19).

Omran, M. G., A. P. Engelbrecht, and A. Salman (2007). "An overview of clustering methods". *Intelligent Data Analysis* 11.6, pp. 583–605. DOI: `10.3233/IDA-2007-11602` (cit. on pp. 39, 40).

Ono, K., M. Kunii, and Y. Honda (2021). "The regional model-based Mesoscale Ensemble Prediction System, MEPS, at the Japan Meteorological Agency". *Quarterly Journal of the Royal Meteorological Society* 147.734, pp. 465–484. DOI: `10.1002/qj.3928` (cit. on pp. 25, 28, 30).

Pagano, T. C. et al. (2024). "Challenges of Operational Weather Forecast Verification and Evaluation". *Bulletin of the American Meteorological Society* 105.4, E789–E802. DOI: `10.1175/BAMS-D-22-0257.1` (cit. on p. 115).

Palmer, T. (2019). "The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years". *Quarterly Journal of the Royal Meteorological Society* 145 (S1), pp. 12–24. DOI: `10.1002/qj.3383` (cit. on pp. 81, 115).

Pearson, C. et al. (2023). *Experiment: comparing forecasting with deterministic and ensemble output* (cit. on pp. 38, 41).

Petch, J. C. (2006). "Sensitivity studies of developing convection in a cloud-resolving model". *Quarterly Journal of the Royal Meteorological Society* 132.615, pp. 345–358. DOI: `10.1256/qj.05.71` (cit. on p. 21).

Phipps, K. et al. (2022). "Evaluating ensemble post-processing for wind power forecasts". *Wind Energy* 25.8, pp. 1379–1405. DOI: `10.1002/we.2736` (cit. on p. 43).

Porson, A. N. et al. (2019). "Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore". *Quarterly Journal of the Royal Meteorological Society* 145.724, pp. 3004–3022. DOI: `10.1002/qj.3601` (cit. on pp. 25, 32, 36, 43, 47, 48, 81, 82, 95, 103).

Porson, A. N. 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* 146.732, pp. 3245–3265. DOI: `10.1002/qj.3844` (cit. on pp. 1, 2, 27, 30, 47, 50, 51, 74, 81, 82, 85, 103, 107, 118, 119, 124).

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". *Journal of the American Statistical Association* 66.336, pp. 846–850. DOI: `10.1080/01621459.1971.10482356` (cit. on p. 128).

Raymond, W. H. (1988). "High-Order Low-Pass Implicit Tangent Filters for Use in Finite Area Calculations". *Monthly Weather Review* 116.11, pp. 2132–2141. DOI: `10.1175/1520-0493(1988)116<2132:HOLPIT>2.0.CO;2` (cit. on pp. 48, 52).

Raynaud, L. and F. Bouttier (2017). "The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts". *Quarterly Journal of the Royal Mete-*

orological Society 143.709, pp. 3037–3047. DOI: 10.1002/qj.3159 (cit. on pp. 2, 16, 26, 27, 34, 37, 47, 48, 81, 82).

Reinert, D. et al. (2025). *DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System.* 2.5.0. Deutscher Wetterdienst (cit. on pp. 1, 25, 27, 28, 30).

Richardson, L. F. (1922). *Weather Prediction by Numerical Process.* 2nd ed. Cambridge Mathematical Library. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511618291 (cit. on p. 5).

Roberts, B. et al. (2020). "What Does a Convection-Allowing Ensemble of Opportunity Buy Us in Forecasting Thunderstorms?" *Weather and Forecasting* 35.6, pp. 2293–2316. DOI: 10.1175/WAF-D-20-0069.1 (cit. on p. 28).

Roberts, N. et al. (2023). "IMPROVER: The New Probabilistic Postprocessing System at the Met Office". *Bulletin of the American Meteorological Society* 104.3, E680–E697. DOI: 10.1175/BAMS-D-21-0273.1 (cit. on pp. 16, 41, 42, 55, 75, 139).

Roberts, N. M. and H. W. Lean (2008). "Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events". *Monthly Weather Review* 136.1, pp. 78–97. DOI: 10.1175/2007MWR2123.1 (cit. on pp. 3, 35, 53, 59, 64, 86, 121, 138, 140, 141, 146).

Roberts, N. M. et al. (2009). "Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, north-west England". *Meteorological Applications* 16.1, pp. 23–34. DOI: 10.1002/met.94 (cit. on pp. 1, 20).

Romine, G. S. et al. (2014). "Representing Forecast Error in a Convection-Permitting Ensemble System". *Monthly Weather Review* 142.12, pp. 4519–4541. DOI: 10.1175/MWR-D-14-00100.1 (cit. on p. 26).

Schellander-Gorgas, T. et al. (2017). "On the forecast skill of a convection-permitting ensemble". *Geoscientific Model Development* 10.1, pp. 35–56. DOI: 10.5194/gmd-10-35-2017 (cit. on pp. 19–21, 81).

Schwartz, C. S. (2019). "Medium-Range Convection-Allowing Ensemble Forecasts with a Variable-Resolution Global Model". *Monthly Weather Review* 147.8, pp. 2997–3023. DOI: 10.1175/MWR-D-18-0452.1 (cit. on pp. 19, 28, 38).

Schwartz, C. S., G. S. Romine, and D. C. Dowell (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* 36.2, pp. 379–405. DOI: 10.1175/WAF-D-20-0110.1 (cit. on pp. 47, 48, 61, 74).

Schwartz, C. S. and R. A. Sobash (2017). "Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations". *Monthly Weather Review* 145.9, pp. 3397–3418. DOI: 10.1175/MWR-D-16-0400.1 (cit. on p. 37).

Schwartz, C. S. et al. (2010). "Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble

Membership". *Weather and Forecasting* 25.1, pp. 263–280. DOI: 10.1175/2009WAF2222267.1 (cit. on pp. 1, 16, 20, 37, 81, 103).

Schwartz, C. S. et al. (2014). "Characterizing and Optimizing Precipitation Forecasts from a Convection-Permitting Ensemble Initialized by a Mesoscale Ensemble Kalman Filter". *Weather and Forecasting* 29.6, pp. 1295–1318. DOI: 10.1175/WAF-D-13-00145.1 (cit. on pp. 2, 34, 47, 81).

Schwartz, C. S. et al. (2015). "NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System". *Weather and Forecasting* 30.6, pp. 1645–1654. DOI: 10.1175/WAF-D-15-0103.1 (cit. on pp. 28, 81).

— (2019). "NCAR's Real-Time Convection-Allowing Ensemble Project". *Bulletin of the American Meteorological Society* 100.2, pp. 321–343. DOI: 10.1175/BAMS-D-17-0297.1 (cit. on p. 28).

Schwartz, C. S. et al. (2022). "Comparing Partial and Continuously Cycling Ensemble Kalman Filter Data Assimilation Systems for Convection-Allowing Ensemble Forecast Initialization". *Weather and Forecasting* 37.1, pp. 85–112. DOI: 10.1175/WAF-D-21-0069.1 (cit. on pp. 47, 48, 74).

Seity, Y. et al. (2011). "The AROME-France Convective-Scale Operational Model". *Monthly Weather Review* 139.3, pp. 976–991. DOI: 10.1175/2010MWR3425.1 (cit. on pp. 25, 27).

Selz, T. and G. C. Craig (2015). "Upscale Error Growth in a High-Resolution Simulation of a Summertime Weather Event over Europe". *Monthly Weather Review* 143.3, pp. 813–827. DOI: 10.1175/MWR-D-14-00140.1 (cit. on p. 23).

Serafin, S., L. Strauss, and M. Dorninger (2019). "Ensemble reduction using cluster analysis". *Quarterly Journal of the Royal Meteorological Society* 145.719, pp. 659–674. DOI: 10.1002/qj.3458 (cit. on pp. 40, 115, 122).

Sharma, K. et al. (2023). "Adaptive selection of members for convective-permitting regional ensemble prediction over the western Maritime Continent". *Frontiers in Environmental Science* 11 (cit. on p. 64).

Shim, T. et al. (2025). "Development Status of KIM Ensemble Prediction System for Seamless Global Sub-Seasonal Prediction". In: 105th AMS Annual Meeting. AMS (cit. on p. 28).

Short, C. J. and J. Petch (2022). "Reducing the spin-up of a regional NWP system without data assimilation". *Quarterly Journal of the Royal Meteorological Society* 148.745, pp. 1623–1643. DOI: 10.1002/qj.4268 (cit. on p. 18).

Skok, G. (2016). "Analysis of Fraction Skill Score properties for a displaced rainy grid point in a rectangular domain". *Atmospheric Research* 169, pp. 556–565. DOI: 10.1016/j.atmosres.2015.04.012 (cit. on p. 89).

— (2022). "A New Spatial Distance Metric for Verification of Precipitation". *Applied Sciences* 12, p. 4048. DOI: 10.3390/app12084048 (cit. on pp. 3, 120, 121).

Skok, G. and N. Roberts (2018). "Estimating the displacement in precipitation forecasts using the Fractions Skill Score". *Quarterly Journal of the Royal Meteorological Society* 144.711, pp. 414–425. DOI: 10.1002/qj.3212 (cit. on pp. 3, 122, 140–142).

Sobash, R. A. et al. (2016). "Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System". *Weather and Forecasting* 31.1, pp. 255–271. DOI: 10.1175/WAF-D-15-0138.1 (cit. on p. 81).

Speer, M. S. and L. M. Leslie (2002). "The prediction of two cases of severe convection: implications for forecast guidance". *Meteorology and Atmospheric Physics* 80.1, pp. 165–175. DOI: 10.1007/s007030200023 (cit. on p. 17).

Speight, L. et al. (2018). "Developing surface water flood forecasting capabilities in Scotland: an operational pilot for the 2014 Commonwealth Games in Glasgow". *Journal of Flood Risk Management* 11 (S2), S884–S901. DOI: 10.1111/jfr3.12281 (cit. on p. 43).

Speight, L. J. et al. (2021). "Operational and emerging capabilities for surface water flood forecasting". *WIREs Water* 8.3, e1517. DOI: 10.1002/wat2.1517 (cit. on pp. 42, 43).

Stein, A. F. et al. (2015a). "NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System". *Bulletin of the American Meteorological Society* 96.12, pp. 2059–2077. DOI: 10.1175/BAMS-D-14-00110.1 (cit. on p. 111).

Stein, T. H. M. et al. (2015b). "The DYMECS Project: A Statistical Approach for the Evaluation of Convective Storms in High-Resolution NWP Models". *Bulletin of the American Meteorological Society* 96.6, pp. 939–951. DOI: 10.1175/BAMS-D-13-00279.1 (cit. on p. 22).

Surcel, M., I. Zawadzki, and M. K. Yau (2015). "A Study on the Scale Dependence of the Predictability of Precipitation Patterns". *Journal of the Atmospheric Sciences* 72.1, pp. 216–235. DOI: 10.1175/JAS-D-14-0071.1 (cit. on p. 23).

Suri, D. and P. Davies A. (2021). "A Decade of Impact-Based NSWWS Warnings at the Met Office". *The European Forecaster* (cit. on p. 43).

Tempest, K. I., G. C. Craig, and J. R. Brehmer (2023). "Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble". *Quarterly Journal of the Royal Meteorological Society* 149.752, pp. 677–702. DOI: 10.1002/qj.4410 (cit. on pp. 15, 116).

Tempest, K. I. et al. (2024). "Convergence of ensemble forecast distributions in weak and strong forcing convective weather regimes". *Quarterly Journal of the Royal Meteorological Society* 150.763, pp. 3220–3237. DOI: 10.1002/qj.4684 (cit. on pp. 15, 16, 24, 37, 116).

Tennant, W. (2015). "Improving initial condition perturbations for MOGREPS-UK". *Quarterly Journal of the Royal Meteorological Society* 141.691, pp. 2324–2336. DOI: 10.1002/qj.2524 (cit. on pp. 47, 48, 74, 81).

Tennant, W. J. et al. (2011). "Using a Stochastic Kinetic Energy Backscatter Scheme to Improve MOGREPS Probabilistic Forecast Skill". *Monthly Weather Review* 139.4, pp. 1190–1206. DOI: 10.1175/2010MWR3430.1 (cit. on p. 10).

Theis, S. E., A. Hense, and U. Damrath (2005). “Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach”. *Meteorological Applications* 12.3, pp. 257–268. DOI: 10.1017/S1350482705001763 (cit. on pp. 15, 16, 42).

Torn, R. D. and G. J. Hakim (2008). “Ensemble-Based Sensitivity Analysis”. *Monthly Weather Review* 136.2, pp. 663–677. DOI: 10.1175/2007MWR2132.1 (cit. on p. 41).

Toth, Z. and E. Kalnay (1997). “Ensemble Forecasting at NCEP and the Breeding Method”. *Monthly Weather Review* 125.12, pp. 3297–3319. DOI: 10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2 (cit. on p. 8).

Trier, S. B. et al. (2015). “Mesoscale Thermodynamic Influences on Convection Initiation near a Surface Dryline in a Convection-Permitting Ensemble”. *Monthly Weather Review* 143.9, pp. 3726–3753. DOI: 10.1175/MWR-D-15-0133.1 (cit. on p. 81).

UKMO (2019). *Met Office Daily Weather Summary June 2019* (cit. on pp. 59, 69).

— (2023a). *Met Office Daily Weather Summary June 2023* (cit. on pp. 92, 119, 133).

— (2023b). *The Wessex Convection Experiment (WesCon)*. Met Office. URL: https://www.metoffice.gov.uk/research/foundation/observational-studies/wessex-convection-experiment (visited on 09/23/2024) (cit. on p. 90).

Vié, B., O. Nuissier, and V. Ducrocq (2011). “Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions”. *Monthly Weather Review* 139.2, pp. 403–423. DOI: 10.1175/2010MWR3487.1 (cit. on pp. 32, 34, 36, 82).

Vinh, N. X., J. Epps, and J. Bailey (2010). “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. *Journal of Machine Learning Research* 11 (cit. on p. 128).

Wang, H. et al. (2014). “A scale-dependent blending scheme for WRFDA: impact on regional weather forecasting”. *Geoscientific Model Development* 7.4, pp. 1819–1828. DOI: 10.5194/gmd-7-1819-2014 (cit. on pp. 48, 74).

Wang, Y. et al. (2011). “The Central European limited-area ensemble forecasting system: ALADIN-LAEF”. *Quarterly Journal of the Royal Meteorological Society* 137.655, pp. 483–502. DOI: 10.1002/qj.751 (cit. on pp. 47, 48, 74).

Warner, T. T., R. A. Peterson, and R. E. Treadon (1997). “A Tutorial on Lateral Boundary Conditions as a Basic and Potentially Serious Limitation to Regional Numerical Weather Prediction”. *Bulletin of the American Meteorological Society* 78.11, pp. 2599–2618. DOI: 10.1175/1520-0477(1997)078<2599:ATOLBC>2.0.CO;2 (cit. on p. 82).

Weidle, F. et al. (2013). “Validation of Strategies using Clustering Analysis of ECMWF EPS for Initial Perturbations in a Limited Area Model Ensemble Prediction System”. *Atmosphere-Ocean* 51.3, pp. 284–295. DOI: 10.1080/07055900.2013.802217 (cit. on pp. 25, 39, 115).

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2007). “The Discrete Brier and Ranked Probability Skill Scores”. *Monthly Weather Review* 135.1, pp. 118–124. DOI: 10.1175/MWR3280.1 (cit. on p. 13).

Wernli, H., C. Hofmann, and M. Zimmer (2009). "Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique". *Weather and Forecasting* 24.6, pp. 1472–1484. DOI: 10.1175/2009WAF2222271.1 (cit. on p. 55).

Wernli, H. et al. (2008). "SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts". *Monthly Weather Review* 136.11, pp. 4470–4487. DOI: 10.1175/2008MWR2415.1 (cit. on p. 34).

Weusthoff, T. et al. (2010). "Assessing the Benefits of Convection-Permitting Models by Neighborhood Verification: Examples from MAP D-PHASE". *Monthly Weather Review* 138.9, pp. 3418–3433. DOI: 10.1175/2010MWR3380.1 (cit. on pp. 86, 88).

Whitaker, J. S. and A. F. Loughe (1998). "The Relationship between Ensemble Spread and Ensemble Mean Skill". *Monthly Weather Review* 126.12, pp. 3292–3302. DOI: 10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2 (cit. on pp. 14, 39, 81).

Wilkinson, J. M. and R. Neal (2021). "Exploring relationships between weather patterns and observed lightning activity for Britain and Ireland". *Quarterly Journal of the Royal Meteorological Society* 147.738, pp. 2772–2795. DOI: 10.1002/qj.4099 (cit. on p. 90).

Wilks, D. S. (1997). "Resampling Hypothesis Tests for Autocorrelated Fields". *Journal of Climate* 10.1, pp. 65–82. DOI: 10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2 (cit. on p. 77).

Wilson, D. R. and S. P. Ballard (1999). "A microphysically based precipitation scheme for the UK meteorological office unified model". *Quarterly Journal of the Royal Meteorological Society* 125.557, pp. 1607–1636. DOI: 10.1002/qj.49712555707 (cit. on p. 20).

Wolf, G. et al. (2024). "Comparison of probabilistic forecasts of extreme precipitation for a global and convection-permitting ensemble and hybrid statistical–dynamical method based on equatorial wave information". *Quarterly Journal of the Royal Meteorological Society* 150.759, pp. 877–896. DOI: 10.1002/qj.4627 (cit. on p. 43).

Woodhams, B. J. et al. (2018). "What Is the Added Value of a Convection-Permitting Model for Forecasting Extreme Rainfall over Tropical East Africa?" *Monthly Weather Review* 146.9, pp. 2757–2780. DOI: 10.1175/MWR-D-17-0396.1 (cit. on pp. 1, 3, 20, 57, 58, 64, 81, 117, 151, 152).

World Meteorological Organization (2024). *Devastating rainfall hits Spain in yet another flood-related disaster.* World Meteorological Organization. URL: https://wmo.int/media/news/devastating-rainfall-hits-spain-yet-another-flood-related-disaster (visited on 08/07/2025) (cit. on p. 44).

Al-Yahyai, S. et al. (2012). "Nested ensemble NWP approach for wind energy assessment". *Renewable Energy* 37.1, pp. 150–160. DOI: 10.1016/j.renene.2011.06.014 (cit. on p. 43).

Yang, D. and J. Kleissl (2023). "Summarizing ensemble NWP forecasts for grid operators: Consistency, elicitability, and economic value". *International Journal of Forecasting* 39.4, pp. 1640–1654. DOI: 10.1016/j.ijforecast.2022.08.002 (cit. on p. 43).

Yang, S.-C., E. Kalnay, and T. Enomoto (2015). “Ensemble singular vectors and their use as additive inflation in EnKF”. *Tellus A: Dynamic Meteorology and Oceanography* 67.1, p. 26536. DOI: 10.3402/tellusa.v67.26536 (cit. on p. 10).

Yang, X. (2005). “Analysis blending using spatial filter in grid-point model coupling”. *Hirlam Newslett.* 48 (cit. on p. 48).

Young, M. V. and N. S. Grahame (2024a). “The history of UK weather forecasting: the changing role of the central guidance forecaster. Part 7: Operational forecasting in the twenty-first century: graphical guidance products, risk assessment and impact-based warnings”. *Weather* 79.3, pp. 72–80. DOI: 10.1002/wea.4488 (cit. on pp. 38, 115).

— (2024b). “The history of UK weather forecasting: the changing role of the central guidance forecaster. Part 8. Operational forecasting in the twenty-first century: enhanced capabilities from nowcasting to extended range”. *Weather* 79.5, pp. 148–157. DOI: 10.1002/wea.4497 (cit. on p. 21).

Yu, X. and T.-Y. Lee (2010). “Role of convective parameterization in simulations of a convection band at grey-zone resolutions”. *Tellus A* 62.5, pp. 617–632. DOI: 10.1111/j.1600-0870.2010.00470.x (cit. on p. 21).

Yussouf, N., D. J. Stensrud, and S. Lakshmivarahan (2004). “Cluster Analysis of Multimodel Ensemble Data over New England”. *Monthly Weather Review* 132.10, pp. 2452–2462. DOI: 10.1175/1520-0493(2004)132<2452:CAOMED>2.0.CO;2 (cit. on p. 139).

Zhang, F. et al. (2007). “Mesoscale Predictability of Moist Baroclinic Waves: Convection-Permitting Experiments and Multistage Error Growth Dynamics”. *Journal of the Atmospheric Sciences* 64.10, pp. 3579–3594. DOI: 10.1175/JAS4028.1 (cit. on p. 23).

Zhang, H. et al. (2015). “Study on multi-scale blending initial condition perturbations for a regional ensemble prediction system”. *Advances in Atmospheric Sciences* 32.8, pp. 1143–1155. DOI: 10.1007/s00376-015-4232-6 (cit. on pp. 47–49, 74).

Zhang, L. et al. (2023). “The Lateral Boundary Perturbations Growth and Their Dependence on the Forcing Types of Severe Convection in Convection-Allowing Ensemble Forecasts”. *Atmosphere* 14.1, p. 176. DOI: 10.3390/atmos14010176 (cit. on pp. 32, 82, 116).

Zhang, Y., J. Wang, and X. Wang (2014). “Review on probabilistic forecasting of wind power generation”. *Renewable and Sustainable Energy Reviews* 32, pp. 255–270. DOI: 10.1016/j.rser.2014.01.033 (cit. on p. 43).

Zheng, M. et al. (2017). “Applying Fuzzy Clustering to a Multimodel Ensemble for U.S. East Coast Winter Storms: Scenario Identification and Forecast Verification”. *Weather and Forecasting* 32.3, pp. 881–903. DOI: 10.1175/WAF-D-16-0112.1 (cit. on p. 40).

Zhou, X. et al. (2022). “The Development of the NCEP Global Ensemble Forecast System Version 12”. *Weather and Forecasting* 37.6, pp. 1069–1084. DOI: 10.1175/WAF-D-21-0112.1 (cit. on pp. 81, 115).

Zimmer, M. et al. (2011). “Classification of precipitation events with a convective response timescale and their forecasting characteristics”. *Geophysical Research Letters* 38.5. DOI: 10.1029/2010GL046199 (cit. on p. 24).