University of Reading

Assessing the value of clustering convection-permitting ensemble forecasts

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Gainford, A. ORCID: https://orcid.org/0000-0003-2484-8316, Frame, T. H. A. ORCID: https://orcid.org/0000-0001-6542-2173, Gray, S. L. ORCID: https://orcid.org/0000-0001-8658-362X, Neal, R. ORCID: https://orcid.org/0000-0003-2678-6016, Porson, A. N. and Milan, M. (2025) Assessing the value of clustering convection-permitting ensemble forecasts. Meteorological Applications, 32 (6). e70139. ISSN 1469-8080 doi: 10.1002/met.70139 Available at https://centaur.reading.ac.uk/127503/

It is advisable to refer to the publisher’s version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1002/met.70139

Publisher: Royal Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

University of Reading

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Meteorological Applications

RESEARCH ARTICLE OPEN ACCESS

Assessing the Value of Clustering Convection-Permitting Ensemble Forecasts

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

[1]Department of Meteorology, University of Reading, Reading, UK | [2]Met Office, Exeter, UK | [3]MetOffice@Reading, University of Reading, Reading, UK

Correspondence: Adam Gainford (adam.gainford@reading.ac.uk)

Received: 17 September 2025 | **Revised:** 14 November 2025 | **Accepted:** 29 November 2025

Keywords: clustering | forecasting | post-processing | spatial methods

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.

1 | Introduction

Ensembles are commonly used to quantify forecast uncertainty by running repeated simulations with different initial conditions and model parameters (e.g., Palmer 2019; Zhou et al. 2022; Inverarity et al. 2023). 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) data sets, forecasters rarely have the time to perform these individual member examinations (Young and Grahame 2024; Pagano et al. 2024). 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; Brill et al. 2015; Boykin 2022; Lamberson et al. 2023), only a few studies have examined the utility of these methods with CPEs (Branković et al. 2008; Johnson, Wang, Kong, and Xue 2011), 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.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 Crown copyright. Met Office and The Author(s). *Meteorological Applications* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society. This article is published with the permission of the Controller of HMSO and the King's Printer for Scotland.

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 (Molteni et al. 2001; Marsigli 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 (Nuissier et al. 2012; Weidle et al. 2013; Bouttier and Raynaud 2018; Serafin et al. 2019), 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. 2019) 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 et al. 2023; Lee and Messori 2024).

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 et al. 2025). Other work has focused 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 (Brill et al. 2015; Boykin 2022; Lamberson et al. 2023). Cluster verification has been more mixed however, as the largest clusters have been found to be both more skilful (Lamberson et al. 2023) and less skilful (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, Wang, Kong, and Xue 2011; Johnson, Wang, Xue, and Kong 2011) 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., Hanley et al. 2011; Clark et al. 2016; Woodhams et al. 2018; Cafaro et al. 2019). 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. First, do the bulk cluster statistics (cluster sizes, medoids, cluster memberships) demonstrate differences between convection-permitting and driving ensembles? Second, 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 Supporting Information compares different spatial methods for estimating distances between members in each ensemble. Then, Section 3 compares statistics between the cluster sets

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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.

2 | Methods

In Section 2.1, the models and trial period used in this study are described. Then, in Section 2.2, the feature-based clustering procedure is described.

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 8 days and outputs three-hourly precipitation accumulations, and so we use 3 h 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 5 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 recentered 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 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 (Porson et al. 2020; Gainford et al. 2024). 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 Figure 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 2023), 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 United Kingdom. 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.

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 2.2.2.

An example of the steps involved in the clustering workflow is depicted in Figure 1 and described in the following subsections.

2.2.1 | Steps 1 and 2: Leadtime Window Selection and Feature Identification

First, a spatial field (e.g., gridded precipitation data), region and clustering window are chosen. The spatial field can in principle be any meteorological parameter: here we choose

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

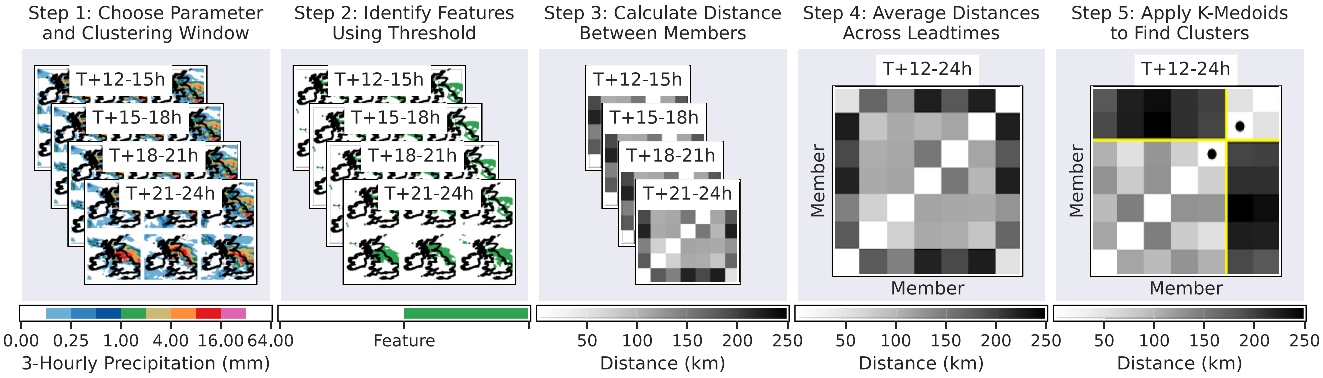

Step 1: Choose Parameter and Clustering Window
Step 2: Identify Features Using Threshold
Step 3: Calculate Distance Between Members
Step 4: Average Distances Across Leadtimes
Step 5: Apply K-Medoids to Find Clusters
T+12-15h
T+15-18h
T+18-21h
T+21-24h
T+12-24h
Member
Member
0.00 0.25 1.00 4.00 16.00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

FIGURE 1 | Schematic showing clustering workflow. Step 1 shows the choice of parameter used in this work and the first clustering window mentioned in Figure 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.

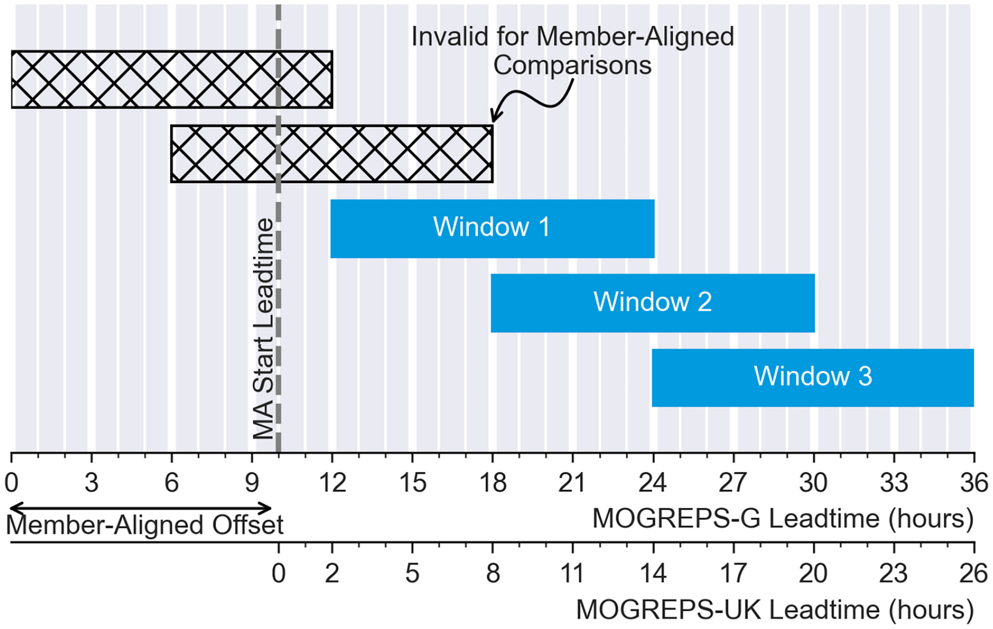


FIGURE 2 | Leadtime window structure and comparison between ensembles accounting for 10 h member-aligned leadtime offset. Clustering is applied over each leadtime window to produce a set of consistent clusters valid for that window.

three-hourly precipitation accumulations. We cluster over the MOGREPS-UK domain and extract this region from the MOGREPS-G fields. For the clustering windows, Figure 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 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 (Roberts and Lean 2008; Mittermaier 2021). We choose a 90th percentile 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/3 h on average for MOGREPS-UK and 0.72 mm/3 h on average for MOGREPS-G. In an operational

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

setting, this choice of centile may not result in clusters that are focused on the areas of hazardous weather, and we would generally recommend a larger value be used provided it produced sufficient feature coverage.

2.2.2 | 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:

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

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:

$$\mathrm{PSD}_{(r)} = 0.808 \cdot Q \cdot r_{PSS\geq 0.5}. \tag{2}$$

Tests presented in the first section of the Supporting Information 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.

2.2.3 | 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 (18 × 18), 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 feature 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 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 Figures 10, 11 and 12.

3 | Cluster Similarity

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

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 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

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 Supporting Information. These displacement biases, combined with the modality argument, favour 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 data sets. Figure 4a shows the frequency that at least one cluster is produced containing only a single member: the medoid. Likewise, Figure 4b shows the frequency that at least two singleton clusters are

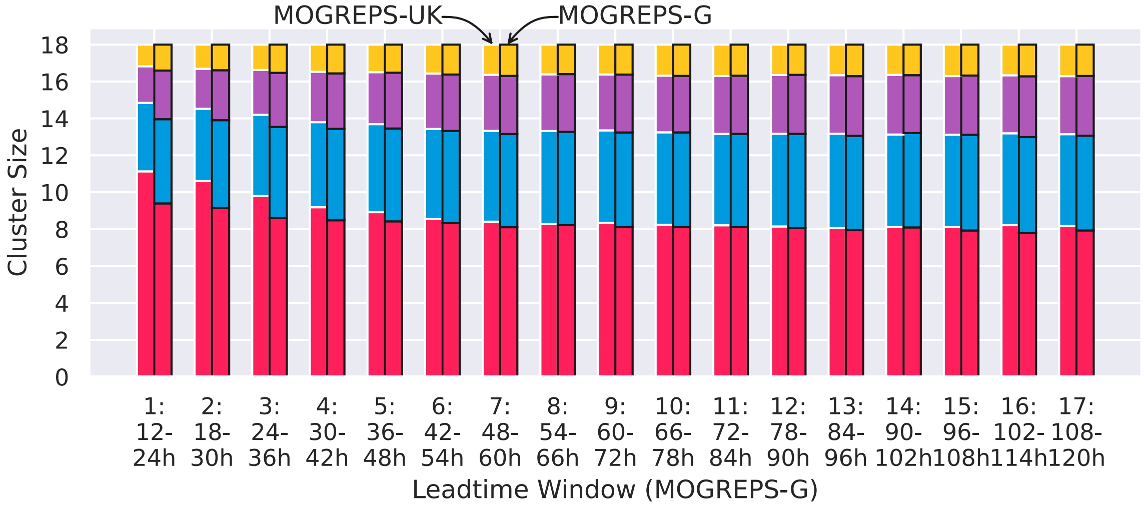


FIGURE 3 | Histogram of average MOGREPS-UK and MOGREPS-G cluster sizes for $k = 4$.

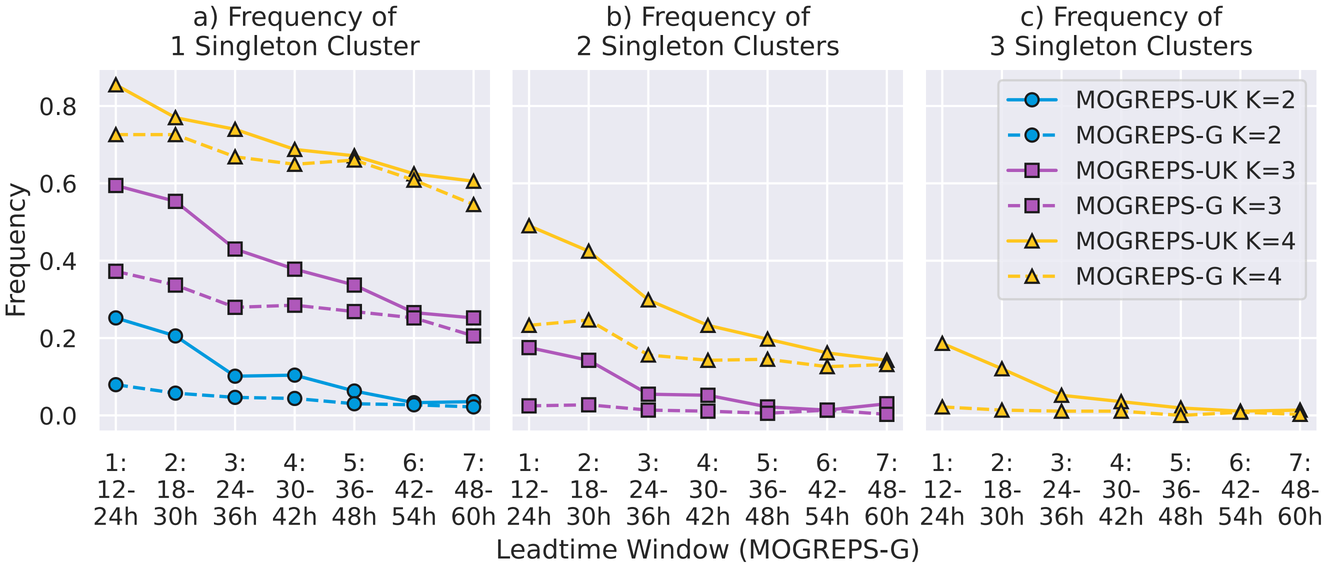


FIGURE 4 | Frequency with which (a) 1 singleton cluster, (b) 2 singleton clusters and (c) 3 singleton clusters occur in MOGREPS-UK (solid) and MOGREPS-G (dashed) cluster sets for the first seven leadtime windows.

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

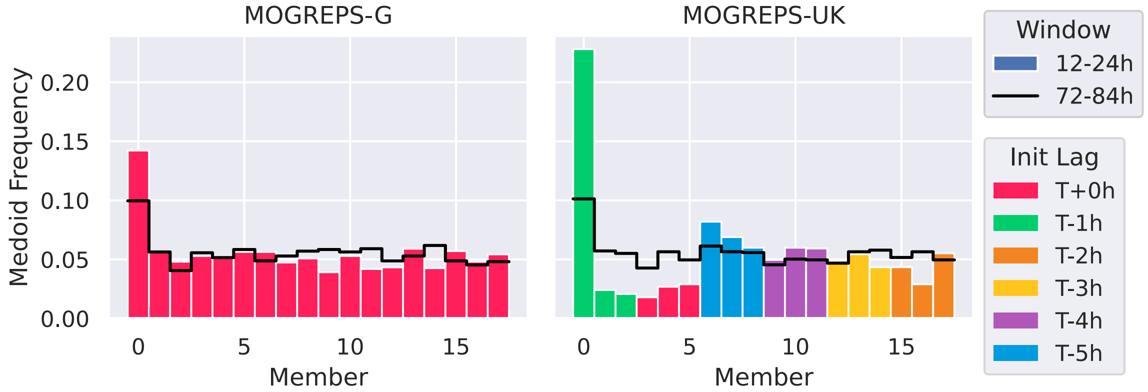


FIGURE 5 | Frequency with which each member in MOGREPS-UK and MOGREPS-G clusters is chosen as a $k = 4$ medoid, displayed by filled bars for the first leadtime window (T + 12–24 h in MOGREPS-G, T + 2–14 h in MOGREPS-UK) and by outline for the eleventh window (T + 72–84 h in MOGREPS-G, T + 62–74 h in MOGREPS-UK). MOGREPS-UK members are coloured by the time-lagged initialisation cycle, as explained further in text.

produced, while Figure 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 Figure 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 with only 38% of $k = 3$ MOGREPS-G clusters. In fact, MOGREPS-UK produces three singleton clusters 10 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 data sets. Consider an example of an ensemble pdf containing three distinct modes. When this data set is clustered with $k = 3$, the outputs will ideally reflect these distinct modes. If this data set 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.

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 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.

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,

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

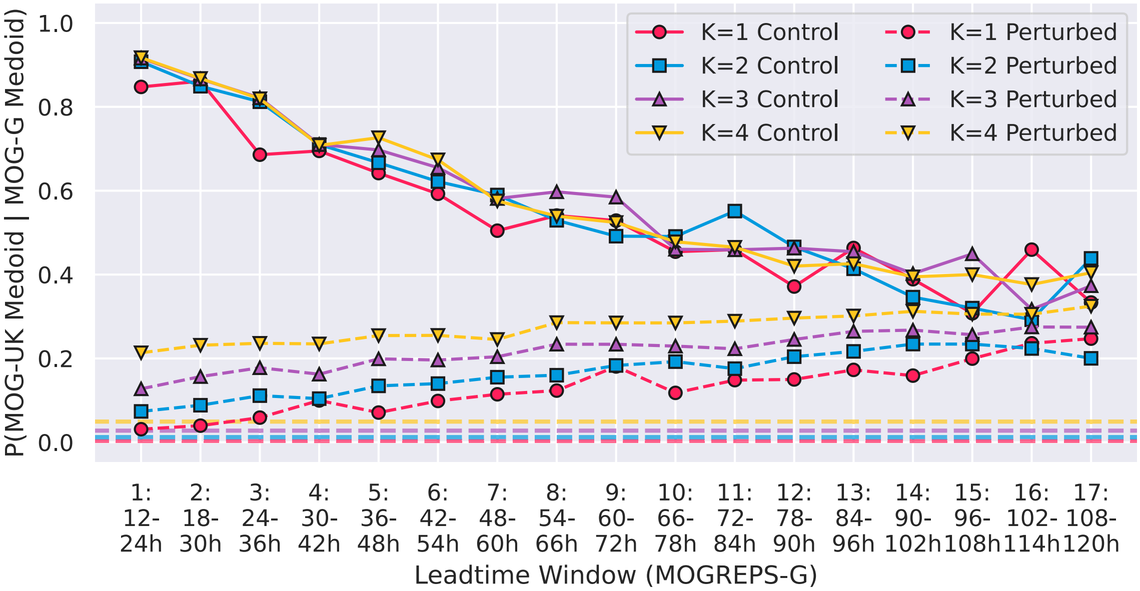


FIGURE 6 | Conditional probability of finding a given medoid in MOGREPS-UK given it is also a medoid in MOGREPS-G. Perturbed trends are average over probabilities for each perturbed member individually. Dashed lines without markers are the probabilities of finding the same medoid in each ensemble by chance (e.g., for $k = 1$ this is 1 / 18, for $k = 2$ this is $1 - (17 / 18^{*}16 / 17)$, etc.).

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 it 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, Figure 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 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.

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., members 1 and 5 are in clusters 1 and 2 in MOGREPS-G, but are both in cluster 1 in MOGREPS-UK).

The ARI is then calculated as:

$$\text{ARI} = \frac{2\left(N_{00}N_{11} - N_{01}N_{10}\right)}{\left(N_{00} + N_{01}\right)\left(N_{01} + N_{11}\right) + \left(N_{00} + N_{10}\right)\left(N_{10} + N_{11}\right)}, \tag{3}$$

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

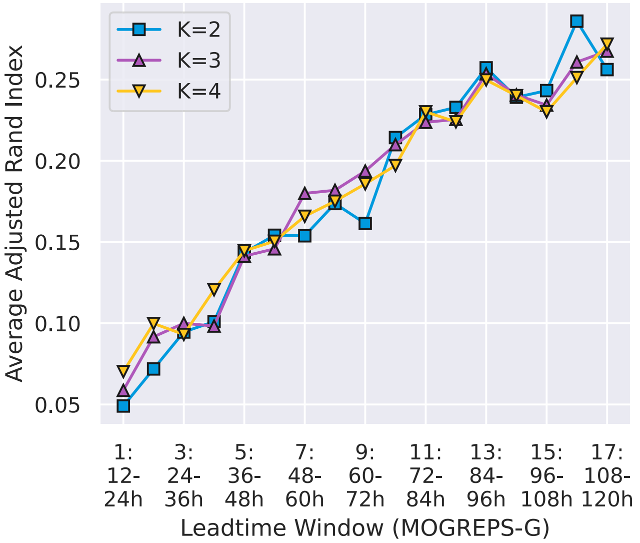


FIGURE 7 | Average Adjusted Rand Index between ensemble clusters evaluating similarity of ensemble membership.

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 7 shows the ARI calculated between the two ensembles for different k. As with the conditional medoid probabilities in Figure 6, 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.

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 is never complete agreement between the clusters, the clear trends with leadtime suggest 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.

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 Equations (1) and (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 8 shows the average PSD between the radar and the cluster medoids (using $k = 4$). For comparison, Figure 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 Supporting Information link this to the grid resolution. Therefore, Figure 8 should not be interpreted as evidence that MOGREPS-UK is drastically more skilful 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 (Figures 3 and 4a), as well as the associated singleton arguments discussed in Section 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, Figure 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 $k = 1$

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 skilful 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 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.

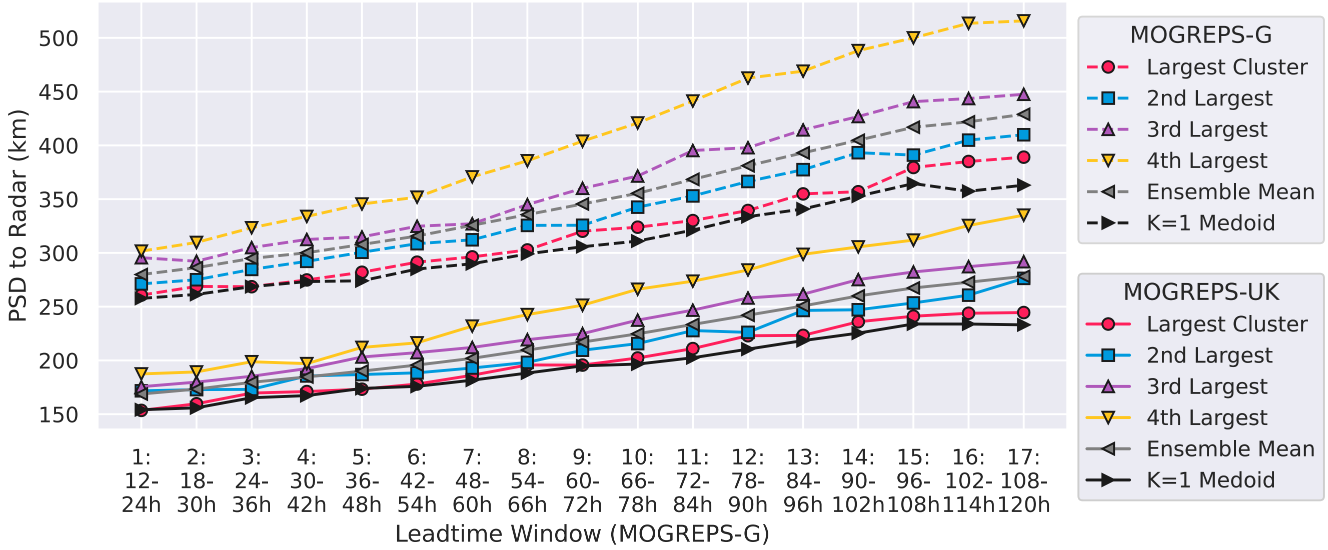


FIGURE 8 | Average PSD between radar and ensemble medoids for each $k = 4$ cluster. Ensemble mean represents the mean PSD between each ensemble member and the radar.

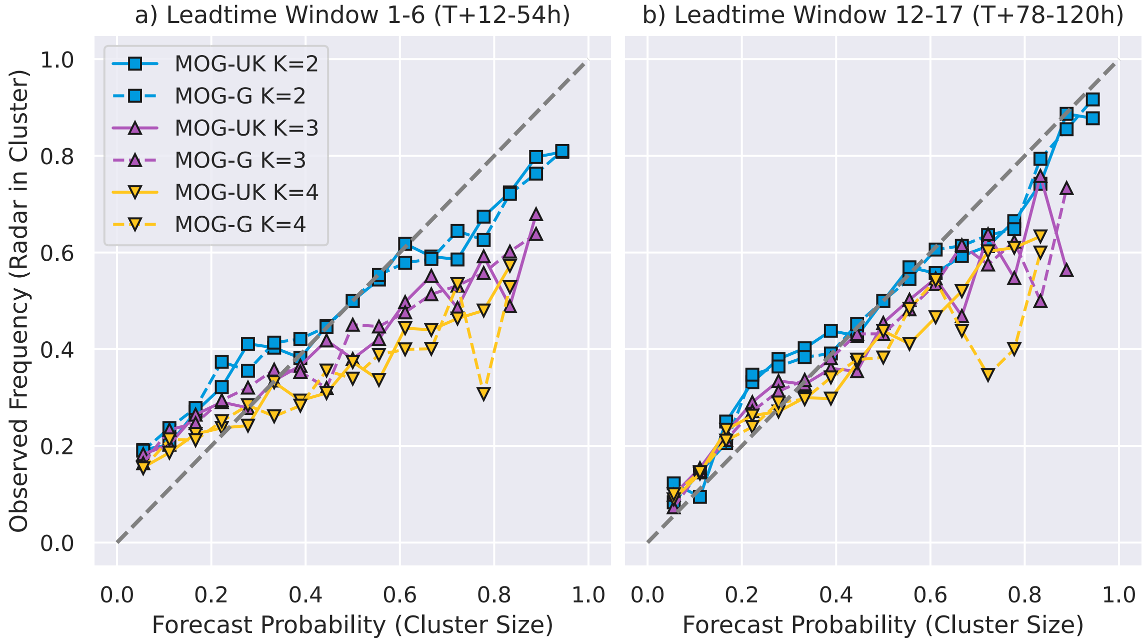


FIGURE 9 | Cluster reliability diagrams for (a) the first 6 leadtime windows and (b) the last 6 leadtime windows. Forecast probability is the cluster size, observed frequency is the frequency with which the radar is placed into a cluster of that size.

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

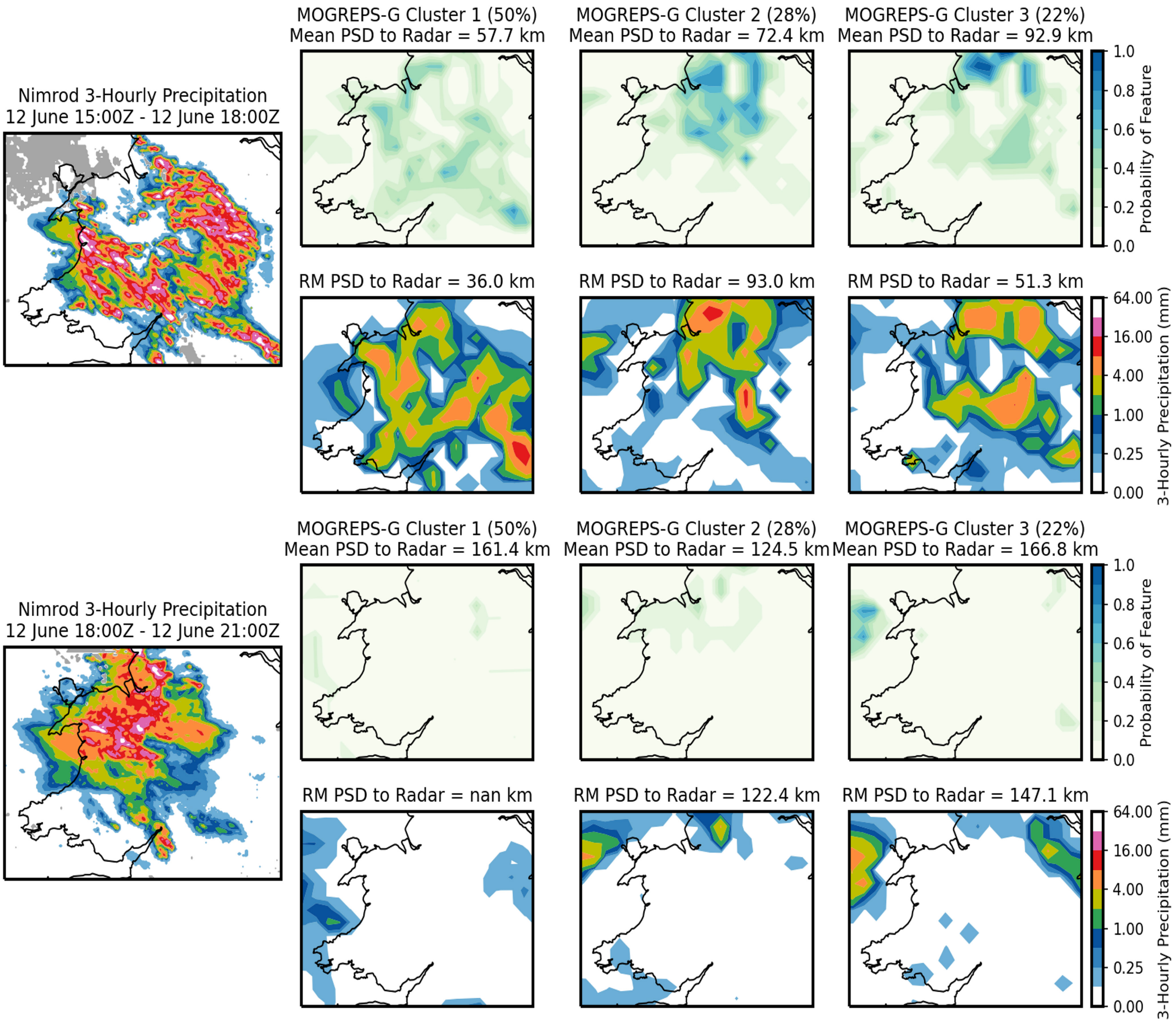

Nimrod 3-Hourly Precipitation
12 June 15:00Z - 12 June 18:00Z
MOGREPS-G Cluster 1 (50%)
Mean PSD to Radar = 57.7 km
MOGREPS-G Cluster 2 (28%)
Mean PSD to Radar = 72.4 km
MOGREPS-G Cluster 3 (22%)
Mean PSD to Radar = 92.9 km
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
RM PSD to Radar = 36.0 km
RM PSD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

FIGURE 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 min of missing data in a 1-h 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).

Figure 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 be 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 Figure 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

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 Supporting Information, 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 on 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 | 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 United Kingdom 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 affected regions. Impacts from surface-water flooding, hail, and thunderstorms were reported (UKMO 2023).

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 Figure 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 percentile, 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 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 11 and 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 northwest 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

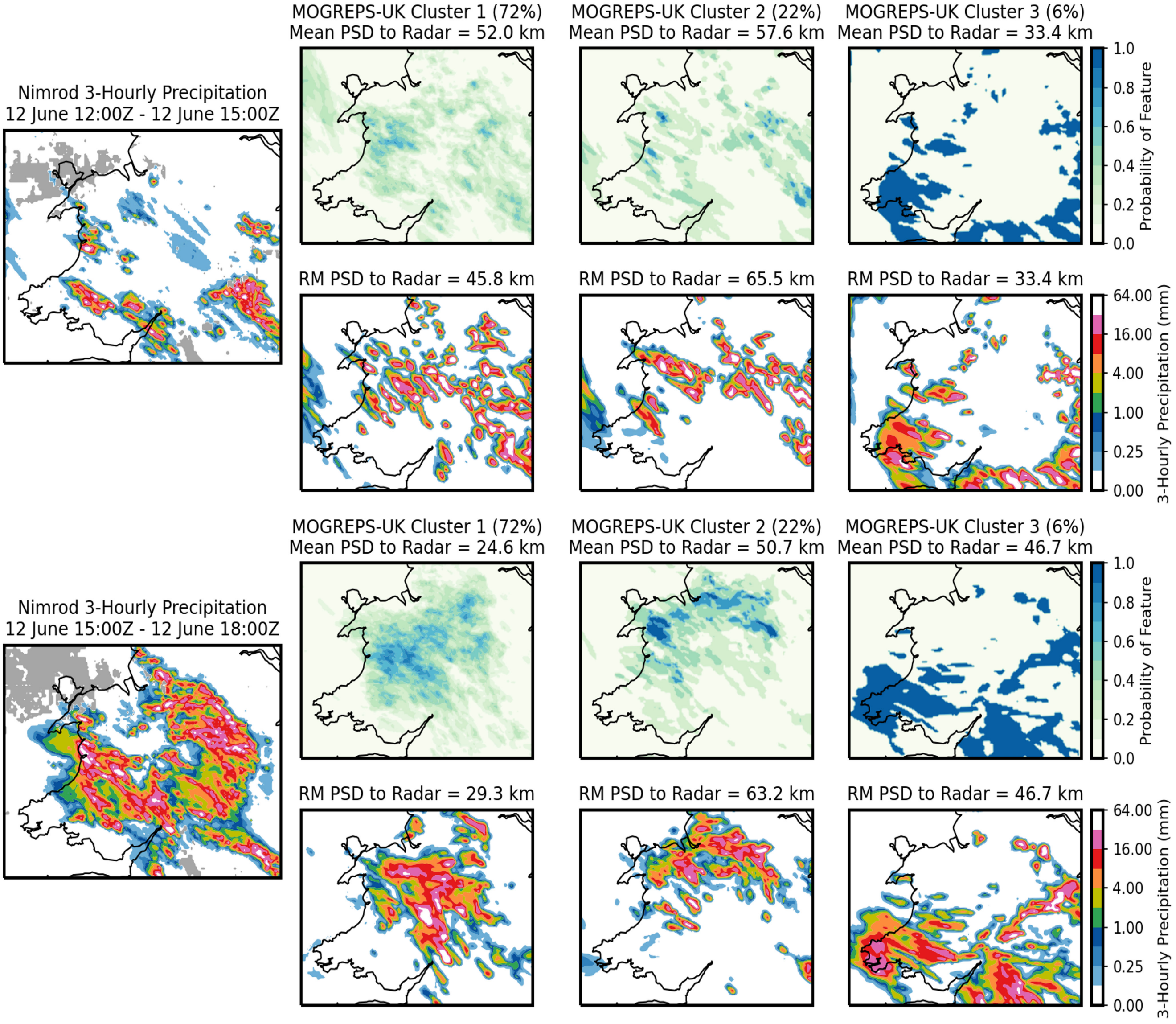
14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from this member, the largest cluster medoid provided the best forecast for this event.

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 data sets. 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 percentile 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 United Kingdom 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 indicate that clusters are most sensitive to large-scale features. A separate case study presented in Section 3 of the Supporting Information 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).



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
Probability of Feature
1.0
0.8
0.6
0.4
0.2
0.0
RM PSD to Radar = 45.8 km
RM PSD to Radar = 65.5 km
RM PSD to Radar = 33.4 km
3-Hourly Precipitation (mm)
64.00
16.00
4.00
1.00
0.25
0.00
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

FIGURE 11 | As with Figure 10 but for MOGREPS-UK clusters showing the first two case study accumulation periods.

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

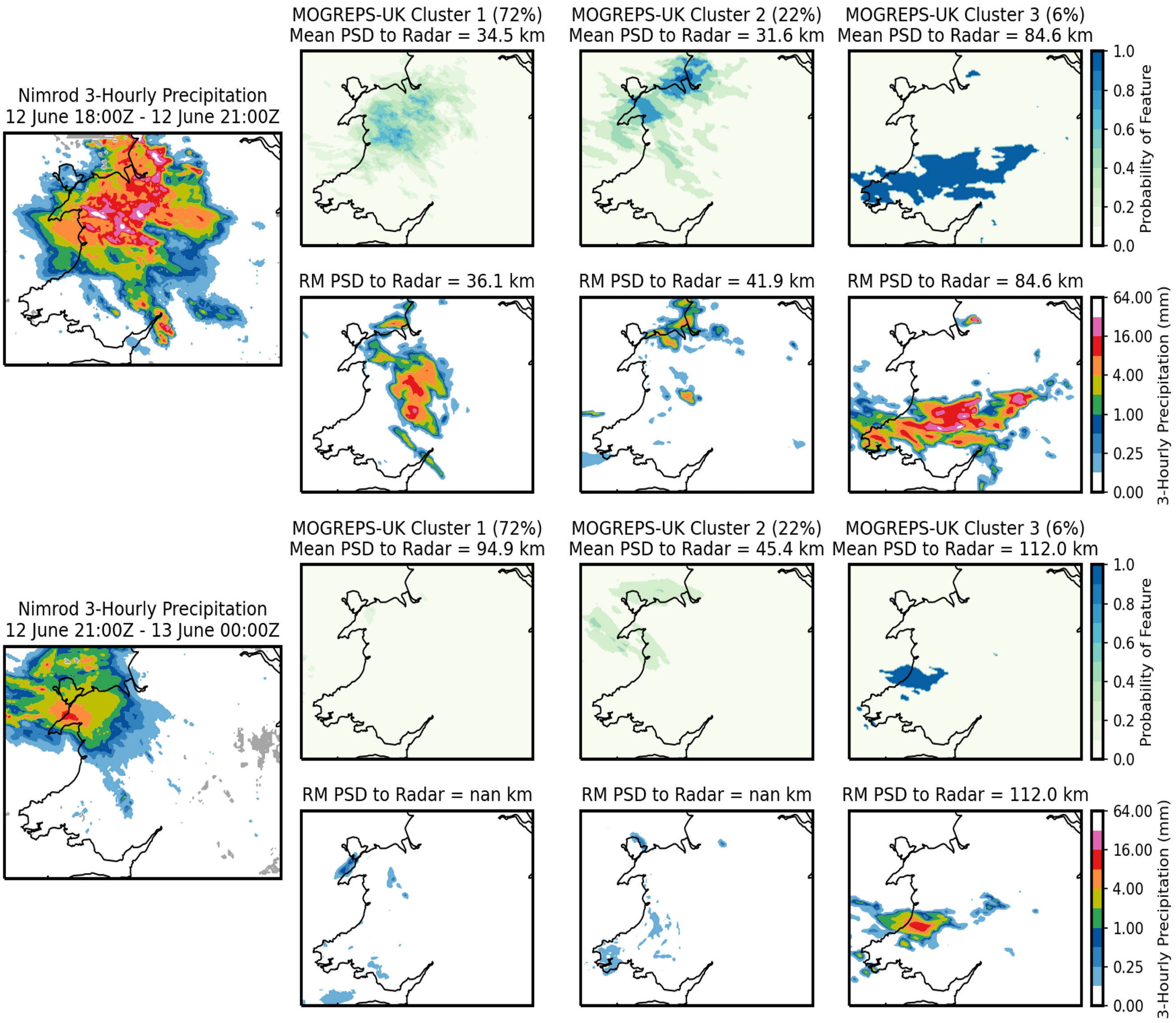

Nimrod 3-Hourly Precipitation
12 June 18:00Z - 12 June 21:00Z
MOGREPS-UK Cluster 1 (72%)
Mean PSD to Radar = 34.5 km
MOGREPS-UK Cluster 2 (22%)
Mean PSD to Radar = 31.6 km
MOGREPS-UK Cluster 3 (6%)
Mean PSD to Radar = 84.6 km
RM PSD to Radar = 36.1 km
RM PSD to Radar = 41.9 km
RM PSD to Radar = 84.6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

FIGURE 12 | As with Figure 11 but for the final two case study accumulation periods.

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 (e.g., 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 skilful than the ensemble average. Furthermore, the medoid representing the smallest cluster (when forced into four clusters) can be substantially less skilful 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,

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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 practice 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; Yussouf et al. 2004; Brill et al. 2015; Lamberson et al. 2023). 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.

Author Contributions

Adam Gainford: writing – original draft, methodology, software, formal analysis, data curation, investigation, visualization, conceptualization, validation. **Thomas H. A. Frame:** writing – review and editing, supervision, funding acquisition, methodology, conceptualization, project administration. **Suzanne L. Gray:** conceptualization, funding acquisition, methodology, writing – review and editing, project administration, supervision. **Robert Neal:** methodology, writing – review and editing, resources, software. **Aurore N. Porson:** conceptualization, funding acquisition, methodology, writing – review and editing, project administration, supervision, resources. **Marco Milan:** supervision, project administration, writing – review and editing, methodology, conceptualization, funding acquisition, resources.

Acknowledgments

The authors have nothing to report.

Funding

This work was supported by SCENARIO NERC Doctoral Training Partnership, University of Reading (NE/S007261/1), Met Office CASE Studentship, Met Office.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the Met Office. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available from the author(s) with the permission of the Met Office.

References

Alhamed, A., S. Lakshmivarahan, and D. J. Stensrud. 2002. "Cluster Analysis of Multimodel Ensemble Data From SAMEX." *Monthly Weather Review* 130: 226–256.

Atger, F. 1999. "Tubing: An Alternative to Clustering for the Classification of Ensemble Forecasts." *Weather and Forecasting* 14: 741–757.

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: 2381–2391.

Boykin, K. A. 2022. *Extracting Likely Scenarios From Ensemble Forecasts in Real-Time*. Ph.D. thesis.

Branković, e., B. Matjačić, S. Ivatek-Šahdan, and R. Buizza. 2008. "Downscaling of ECMWF Ensemble Forecasts for Cases of Severe Weather: Ensemble Statistics and Cluster Analysis." *Monthly Weather Review* 136: 3323–3342.

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: 873–891.

Cafaro, C., T. H. A. Frame, J. Methven, N. Roberts, and J. Bröcker. 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: 1780–1798.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez. 2016. "Convection-Permitting Models: A Step-Change in Rainfall Forecasting." *Meteorological Applications* 23: 165–181.

Craig, G. C., M. Puh, C. Keil, 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: 2325–2343.

Fereday, D. R., J. R. Knight, A. A. Scaife, C. K. Folland, and A. Philipp. 2008. "Cluster Analysis of North Atlantic–European Circulation Types and Links With Tropical Pacific Sea Surface Temperatures." *Journal of Climate* 21: 3687–3703.

Ferranti, L., and S. Corti. 2011. "New clustering products." https://www.ecmwf.int/node/17442.

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: 916–924.

Frogner, I.-L., U. Andrae, J. Bojarova, et al. 2019. "HarmonEPS—The HARMONIE Ensemble Prediction System." *Weather and Forecasting* 34: 1909–1937.

Gainford, A., S. L. Gray, T. H. A. Frame, A. N. Porson, and M. Milan. 2024. "Improvements in the Spread–Skill Relationship of Precipitation in a Convective-Scale Ensemble Through Blending." *Quarterly Journal of the Royal Meteorological Society* 150, no. 762: 3146–3166.

Gebhardt, C., S. Theis, M. Paulat, and Z. Ben Bouallègue. 2011. "Uncertainties in COSMO-DE Precipitation Forecasts Introduced by

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Model Perturbations and Variation of Lateral Boundaries." *Atmospheric Research* 100: 168–177. https://linkinghub.elsevier.com/retrieve/pii/S0169809510003455.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert. 2009. "Intercomparison of Spatial Forecast Verification Methods." *Weather and Forecasting* 24: 1416–1430.

Golding, B. W. 1998. "Nimrod: A System for Generating Automated Very Short Range Forecasts." *Meteorological Applications* 5: 1–16.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant. 2017. "The Met Office Convective-Scale Ensemble, MOGREPS-UK." *Quarterly Journal of the Royal Meteorological Society* 143: 2846–2861.

Hanley, K. E., D. J. Kirshbaum, S. E. Belcher, N. M. Roberts, and G. Leoncini. 2011. "Ensemble Predictability of an Isolated Mountain Thunderstorm in a High-Resolution Model." *Quarterly Journal of the Royal Meteorological Society* 137: 2124–2137.

Hohenegger, C., and C. Schar. 2007. "Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales." *Bulletin of the American Meteorological Society* 88: 1783–1794.

Inverarity, G. W., W. J. Tennant, L. Anton, et al. 2023. "Met Office MOGREPS-G Initialisation Using an Ensemble of Hybrid Four-Dimensional Ensemble Variational (En-4DEnVar) Data Assimilations." *Quarterly Journal of the Royal Meteorological Society* 149, no. 753: 1138–1164.

Johnson, A., X. Wang, F. Kong, and M. Xue. 2011. "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: 3673–3693.

Johnson, A., X. Wang, M. Xue, and F. Kong. 2011. "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: 3694–3710.

Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt. 2014. "The Impact of Downscaled Initial Condition Perturbations on Convective-Scale Ensemble Forecasts of Precipitation." *Quarterly Journal of the Royal Meteorological Society* 140: 1552–1562.

Lamberson, W. S., M. J. Bodner, J. A. Nelson, and S. A. Sienkiewicz. 2023. "The Use of Ensemble Clustering on a Multimodel Ensemble for Medium-Range Forecasting at the Weather Prediction Center." *Weather and Forecasting* 38: 539–554.

Lean, H. W., P. A. Clark, M. Dixon, 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: 3408–3424.

Lee, S. H., and G. Messori. 2024. "The Dynamical Footprint of Year-Round North American Weather Regimes." *Geophysical Research Letters* 51: e2023GL107161.

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: 7091–7108.

Marsigli, C., A. Montani, F. Nerozzi, 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: 2095–2115.

Mittermaier, M. P. 2021. "A "Meta" Analysis of the Fractions Skill Score: The Limiting Case and Implications for Aggregation." *Monthly Weather Review* 149: 3491–3504.

Molteni, F., R. Buizza, C. Marsigli, A. Montani, F. Nerozzi, and T. Paccagnella. 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: 2069–2094.

Montani, A., D. Cesari, C. Marsigli, and T. Paccagnella. 2011. "Seven Years of Activity in the Field of Mesoscale Ensemble Forecasting by the COSMO-LEPS System: Main Achievements and Open Challenges." *Tellus Series A: Dynamic Meteorology and Oceanography* 63, no. 3: 605. https://doi.org/10.1111/j.1600-0870.2010.00499.x.

Mounier, A., L. Raynaud, L. Rottner, M. Plu, and P. Arbogast. 2025. "Rainfall Classification of Kilometer-Scale Ensemble Forecasts Using Convolutional Neural Networks and SOMs." *Artificial Intelligence for the Earth Systems* 4, no. 4.

Neal, R., D. Fereday, R. Crocker, and R. E. Comer. 2016. "A Flexible Approach to Defining Weather Patterns and Their Application in Weather Forecasting Over Europe." *Meteorological Applications* 23: 389–400.

Neal, R., J. Robbins, R. Crocker, et al. 2024. "A Seamless Blended Multi-Model Ensemble Approach to Probabilistic Medium-Range Weather Pattern Forecasts Over the UK." *Meteorological Applications* 31: e2179.

Nuissier, O., B. Joly, B. Vié, and V. Ducrocq. 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: 2993–3011.

Pagano, T. C., B. Casati, S. Landman, et al. 2024. "Challenges of Operational Weather Forecast Verification and Evaluation." *Bulletin of the American Meteorological Society* 105: E789–E802.

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: 12–24.

Porson, A. N., J. M. Carr, S. Hagelin, et al. 2020. "Recent Upgrades to the Met Office Convective-Scale Ensemble: An Hourly Time-Lagged 5-Day Ensemble." *Quarterly Journal of the Royal Meteorological Society* 146: 3245–3265.

Rand, W. M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66: 846–850.

Roberts, N., B. Ayliffe, G. Evans, et al. 2023. "IMPROVER: The New Probabilistic Postprocessing System at the Met Office." *Bulletin of the American Meteorological Society* 104: E680–E697.

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: 78–97.

Serafin, S., L. Strauss, and M. Dorninger. 2019. "Ensemble Reduction Using Cluster Analysis." *Quarterly Journal of the Royal Meteorological Society* 145: 659–674.

Skok, G. 2022. "A New Spatial Distance Metric for Verification of Precipitation." *Applied Sciences* 12: 4048.

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: 414–425. https://doi.org/10.1002/qj.3212.

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: 677–702. https://doi.org/10.1002/qj.4410.

Tempest, K. I., G. C. Craig, M. Puh, and C. Keil. 2024. "Convergence of Ensemble Forecast Distributions in Weak and Strong Forcing Convective Weather Regimes." *Quarterly Journal of the Royal Meteorological Society* 150: 3220–3237. https://doi.org/10.1002/qj.4684.

UKMO. 2023. "Met Office Daily Weather Summary June 2023."

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: 2837–2854.

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Weidle, F., Y. Wang, W. Tian, and T. Wang. 2013. "Validation of Strategies Using Clustering Analysis of ECMWF EPS for Initial Perturbations in a Limited Area Model Ensemble Prediction System." *Atmosphere-Ocean* 51, no. 3: 284–295. https://doi.org/10.1080/07055900.2013.802217.

Woodhams, B. J., C. E. Birch, J. H. Marsham, C. L. Bain, N. M. Roberts, and D. F. A. Boyd. 2018. "What Is the Added Value of a Convection-Permitting Model for Forecasting Extreme Rainfall Over Tropical East Africa?" *Monthly Weather Review* 146: 2757–2780.

Young, M. V., and N. S. Grahame. 2024. "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: 72–80.

Yussouf, N., D. J. Stensrud, and S. Lakshmivarahan. 2004. "Cluster Analysis of Multimodel Ensemble Data Over New England." *Monthly Weather Review* 132: 2452–2462.

Zhang, L., J. Min, X. Zhuang, S. Wang, and X. Qiao. 2023. "The Lateral Boundary Perturbations Growth and Their Dependence on the Forcing Types of Severe Convection in Convection-Allowing Ensemble Forecasts." *Atmosphere* 14: 176.

Zhou, X., Y. Zhu, D. Hou, et al. 2022. "The Development of the NCEP Global Ensemble Forecast System Version 12." *Weather and Forecasting* 37: 1069–1084.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** met70139-sup-0001-Supinfo.pdf.

14698080, 2025, 6, Downloaded from https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70139 by Adam Gainford - <Shibboleth>-member@reading.ac.uk , Wiley Online Library on [12/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License