University of Reading

Extracting Likely Scenarios from Ensemble Forecasts in Real-time

PhD in Atmosphere, Oceans, and Climate

Department of Meteorology

Kristine Adelaide Boykin

September 2022

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

- Kristine Adelaide Boykin

Abstract

With the development of ensemble forecasting, operational meteorologists are faced with large amounts of constantly updating complex information which they must quickly interpret to issue forecasts and warnings. In this thesis a novel clustering technique is introduced that reduces ensemble forecasts to a few representative forecast trajectories. Clustering is performed using k-medoids with the distance metric defined by the Fractions Skill Score (FSS) of the gradient in 850hPa wet-bulb potential temperature to group ensemble members with similar frontal features. The number of clusters is selected using lead-time-coherence of the clusters over a window of interest when clustering is most distinct. Members nearest to the centre of each cluster during this window of interest are chosen as representative members to be viewed by forecasters. Clustering is found to be more coherent during low predictability events when ensemble spread is large. The clustering method was compared to an alternative that uses the FSS of large-scale rain rate and it was found that while similar, results are not interchangeable. The gradient of wet-bulb potential temperature had higher time-coherence and therefore was judged preferable. The method was evaluated during the Met Office winter testbed of 2021-22, and representative members found were found to correspond well to forecasters judgement of the distinct scenarios in the ensemble, hence providing a useful reduction in the data that needs to be considered in issuing forecasts. The method draws attention to low predictability events that appear across several forecasts. While this method has been created to fill a need with ensemble forecasting, it is anticipated that it can be used in many other areas of research such as identifying circulation patterns, seasonal and climate forecast trajectories, and exploring different meteorological phenomena by modifying variable choice and other parameters. The method is also planned for use at the Met Office.

Acknowledgements

I would like to express my deepest appreciation to my supervisors, John Methven, Tom Frame, Nigel Roberts, and Stephen Moseley, who's unwavering support and mentorship throughout my PhD saw me through both triumphs and pitfalls. I also could not have undertaken this challenge without the Met Office and NERC, who provided funding for my work. Additionally, I would like to express my gratitude to my monitoring committee, Sue Gray and Mike Lockwood, who's guidance was invaluable in helping me reach my goals. I would also like to thank Gregor Leckebusch, my external examiner, who I had an excellent discussion of my research with and provided valuable perspective and feedback on my work.

I am also grateful to my cohort, my office mates, and my fellow PhDs for their companionship and moral support. Thanks should also go to the Met Office Winter Testbed group and participants, who's contribution to my project was invaluable.

Finally, I would like to recognize my partner, Aiken Oliver, who was my constant cheerleader during my PhD. His never ending patience and encouragement kept me motivated. And for our cats, Ren and Rei, who were always there to offer love and cuddles while working from home during the pandemic.

Contents

List of Abbreviations

ECMWF	European Centre for Medium-Range Weather Forecasts
ENIAC	Electronic Numerical Integrator And Computer
EPS	Ensemble prediction system
FSS	Fractions Skill Score
MODE	Method for object-based diagnostic evaluation
MSD	Mean square difference
MSLP	Mean sea level pressure
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical weather prediction
RM	Representative member
RMSE	Root mean square error
SAL	Structure, amplitude, and location
SANDRA	Simulated annealing and diversified randomization
UK	United Kingdom
USA	United States of America
UTC	Coordinated universal time

Glossary

Centroid: The centre point of a K-means cluster, calculated as the mean of all cluster members.

Medoid: An ensemble forecast member identified as central to a cluster by the K-medoids algorithm.

Representative member: The ensemble forecast member selected by the algorithm that best represents the cluster over a designated window of interest within a forecast period. This member represents a potential forecast trajectory.

Traceability: The quality of how members remain in a cluster over a length of time within a forecast period.

Window of interest: A period of time (i.e. 48 hours) within a forecast period where the clustering stabilizes and representative members are chosen. The beginning of the window of interest is selected based on a reduction of the normalized sum over clusters of differences between members and their medoid.

Notation

D	Data set
D_{max}	Maximum of the data set
D_{min}	Minimum of the data set
D'	Normalized data set
D_{FSS}	Real distance in kilometers derived from the FSS
F	Forecast fractions
FSS_{MS}	FSS at a minimum size discrete neighbourhood length n that is closest to 0.5
I	The inertia, a measure of similarity in a K-means cluster
K	Kelvin
$\mathbf{K}$	Total number of clusters
k	The index used for clusters
km	Kilometer
m	Index labeling the medoid of a K-medoids cluster
MSD	Mean square difference
N	Domain length
$SICD$	The sum of intra-cluster distances
$SDist$	The sum distance of all intra-cluster sum distances ($SICD$)
x	Index used to label a member in a cluster
θ	Potential temperature
θ_e	Equivalent potential temperature
θ_w	Wet-bulb potential temperature
$\lvert\nabla\theta_w\rvert$	Magnitude of the gradient of the wet-bulb potential temperature
μ	Centroid of a K-means cluster

Chapter 1

Introduction

This chapter will briefly discuss the problem this study aims to address and a list of research questions it aims to satisfy.

1.1 The need for extracting forecast scenarios

Since the beginning of numerical weather prediction (NWP), when modeling the atmosphere became a reality, the need for better observational measurements, faster computation of the governing equations, higher resolution modeling, and deeper understanding of the atmosphere have been constant goals. In their paper, Bauer et al. (2015) discuss the revolution of NWP, its current state and future challenges. However, a major challenge of forecasting was that a single deterministic forecast was not able to account for the chaotic nature of the atmosphere. To overcome this, ensemble forecasts were developed, presenting observational meteorologists with a probabilistic view of the atmosphere with a myriad of different potential forecasts (Buizza, 2018). As computers have become more powerful, these forecasts have grown consistently in accuracy, resolution, duration of forecast, and complexity.

Although forecasts can be found on phone apps and websites, it takes an operational meteorologist to provide a narrative, offer advice, and spot errors. To do this, they must digest large amounts of complex data. Ensemble forecasts are a group of forecasts for the same period of time produced by a model. They can produce any number of forecasts at a time, for example the ECMWF ensemble forecast model produces 50 different forecasts per run. Each forecast contains an array of different atmospheric variables cal-

culated at different model levels (e.g. MOGREPS-UK has 70 different irregularly spaced levels (Hagelin et al., 2017)). Ensemble forecasts can be initialized multiple times a day, sometimes hourly, producing several different time-evolving 3-dimensional pictures of the atmosphere. They are used in risk-based decision making by informing operational meteorologists if there is a chance of a high-impact event even if the probability is relatively low. It is important to produce timely and accurate forecasts with the most up to date and comprehensive data available, especially as new data is produced. With the advent of hourly forecasts, this need becomes even more important to address. Therefore, it is imperative to find ways to improve the use of ensemble data so that an operational meteorologist can more easily digest and utilize all the relevant information while still maintaining the same level of accuracy and precision.

The purpose of this study is to design a method to reduce an ensemble down to the most distinctly different forecasts so that the message of the full ensemble and key signals can be quickly grasped and analysed. This will also provide a means for operational meteorologists to communicate different scenarios to users, particularly for impending high-impact events. By examining forecast members that are representative of the whole, they can avoid purely probabilistic forecasts and more easily see the connections between variables in each individual forecast and the progression of events leading to behaviours of the atmosphere. Ensemble forecasts are expensive to run in terms of computation power, time, and the necessary energy supply, so it is important to encourage their use by making the information contained within more easily accessible. Other studies have used clustering to achieve ensemble reduction, but this study presents a novel method which will be more effective as forecasts become more and more detailed. This method clusters a forecast at each lead time, traces the clusters through the forecast, determines a window of the forecast where clustering is strong, and then extracts representatives based on the clustering within the window. These representatives are then presented to the operational meteorologists as potential forecast scenarios.

1.2 Research questions

Within this study, the following scientific questions and key points about them will be explored:

- Can clustering be applied to an ensemble forecast and extract representative scenarios?

    1. What is the optimal number of clusters?
    2. Is there a preferred variable for clustering?
    3. Can the method be used on different meteorological fields?

- What are the effects of clustering at each lead time?

    1. Can the algorithm identify points in time within a forecast when the members become more diverged?
    2. Can clusters be traced through a forecast?
    3. Is there a relationship between the distinctness of the clusters and the coherence of the members staying in the same cluster over time?

- Can representative members be extracted from a forecast that represent distinct scenarios?

    1. Are the members that best represent the centre of each cluster (representative members) distinct from one another?
    2. Does the clustering algorithm produce representative members as scenarios that are useful to operational meteorologists?
    3. To what extent is there coherence in clusters and scenarios across different forecasts for the same end date and time (valid time) and do they connect to a particular weather event?
    4. Is it possible to quantify days when ensemble forecasts cluster better than others and identify why?

1.3 The significance and impact of the method

The novel method produced through this study aims to help improve ensemble use, decrease the amount of time an operational meteorologist must digest the data from an ensemble, and draw attention to low predictability situations in the atmosphere where

several distinct outcomes are possible. This would be a significant step for tools that will make forecasting easier, particularly for high-impact events.

1.4 Determining the scope of the study

There are a nearly limitless number of questions this project could attempt to address with the method presented and data it could be used on. Some examples include: different variables or combinations of variables could be studied, sensitivity analyses could be performed, climate and seasonal forecasts could be clustered and studied, using the method to cluster forecasts from high-resolution convection permitting ensembles, or tuning the method to better detect other specific weather patterns. However, certain limitations were put in place to keep the project manageable. The hope of this project is to be able to not only reduce an ensemble to its most salient information (according to operational meteorologists), but to be able to draw attention to potential high-impact weather events. The salient information will be presented as representative members from clusters that encompass the potential general progression of atmospheric motion. Whether or not these representatives result in a skillful forecast was not addressed in this work, however it can be a future study based on the probability of a representative member being the most likely forecast.

As much of the high-impact weather in the UK is due to heavy rainfall and high winds which are often associated with strong fronts, the gradient of the wet-bulb potential temperature, an indicator of frontal regions, is chosen as the variable this work will focus on. The method is developed using the global ensemble model MOGREPS-G as it is appropriate for examining frontal systems coming across the North Atlantic Ocean to impact the UK. It was tested on a three-month period from October to December 2018 to gain sufficiently robust results, i.e. the development of parameters used within the method that produce consistent outcomes. It is then implemented later in real time for assessment by operational meteorologists at the Met Office in a testbed.

1.5 Outline of thesis

The following chapters will discuss the relevant literature, the methodology of the study, the application of the method to the gradient of the wet-bulb potential temperature, a comparison of the gradient to large-scale rain, the application of the method during a Met Office testbed, and finally a conclusion that draws together the results of this work.

Chapter 2

Literature review

2.1 Introduction

Within this chapter, a brief discussion of numerical weather forecast development will be presented, including forecasting high-impact events and ensemble forecasts. Next, there will be a review of different methodologies for issuing advisories and warnings including the relatively recent push towards "impact based forecasting", followed by an analysis of clustering methods and meteorological applications of clustering. Finally, methods of forecast verification will be explored.

2.2 Numerical weather forecast development

Since ancient times, human beings have sought to understand and forecast the weather. Early methods frequently relied on current observations and folk traditions. During 1400-1900 AD, the science of studying the weather saw a surge with the invention of weather instruments and accurate measurements (Teague and Gallicchio, 2017). The modern era saw several key developments that changed how forecasting was done, such as the invention of the telegraph, radiosondes, and numerical weather prediction (NWP). Built on years of pioneering work, NWP steadily became a reality (Lynch, 2008). NWP calculations done by hand were prohibitively time consuming. The forecasts were completed hours to days after when they were relevant (Teague and Gallicchio, 2017). However, with the invention of computers, which could do the same calculations much faster than a person could, NWP became feasible. NWP progressed through several stages before

modern super computers, beginning with the first computer simulation of a single level barotropic model run on ENIAC (electronic numerical integrator and computer, (Charney et al., 1950)) in the USA, to later baroclinic and primitive equation models and beyond as computing power continually scaled upward allowing for these more complex models to be implemented (Shuman, 1989; Lynch, 2008). But the United States of America was not the only country interested in NWP. Many countries also began to develop their own NWP models with various successes and failures, detailed in the three articles by Persson (2005a). In his first paper, Persson (2005a) explores the development of NWP in Sweden. One of the most important developments for Swedish NWP was the return of Carl Rossby. As a prominent meteorologist at the cutting edge of scientific discovery, he changed the face of Swedish meteorology with his theories and international connections and developed a barotropic method for NWP. The success of ENIAC further drove Sweden to strive for their own NWP model, ultimately resulting in the first operational real-time forecast in Sweden in the autumn of 1954. In his second paper, Persson (2005b) discusses NWP development in twenty different countries across the world as they raced to join other prominent NWP groups and meteorological advancements. Finally, in his third paper, Persson (2005c) details the story of early British NWP, which chose a baroclinic methodology under the guidance of Reginald Sutcliffe. He argued that NWP should be primarily used for actual meteorological situations and operational activity, not just research. The advancement of NWP across several different countries provided competition and collaboration alike and today there are a large number of NWP models that are routinely run all over the world to forecast the weather and future climate. For example, the World Meteorological Organization runs the Global Data-Processing and Forecasting System program that brings together meteorological analyses and forecast products from 137 different centres and networks across every continent except Antarctica (WMO, 2019).

The importance and urgency of predicting high-impact weather has always been great, and as the study of meteorology progressed there was increasing focus on understanding and modeling these extreme events. However, there is still much ambiguity in the definition of extreme weather. Stephenson (2008) goes into depth about how extreme events might be labeled and diagnosed, pointing out that often the idea of an extreme event is relative and highly dependant on the situation. Stephenson notes that extreme events

are multi-dimensional, meaning they have a variety of attributes that must be taken into account, not simply one factor making it "extreme". One example he gives is a hurricane, which is often simply described by its maximum wind speed, which places it in a certain category of severity. However, hurricanes are large, slow moving events, that have high wind speeds and heavy rainfall, plus the potential to cause flooding due to storm surge. He also notes that due to the rarity of extreme events, predictions are prone to uncertainty to this day. Early numerical weather prediction via computer models provided broad scale forecasts only and high-impact event prediction was primarily up to local forecasters (Shuman, 1989), and still is today. Forecasting high-impact weather is so critical, it has led to collaborations across the world, such as the partnership between NOAA (the National Oceanic and Atmospheric Administration) in the USA and the Met Office in the UK, detailed by Kain et al. (2017). By combining their efforts in research and development, sharing post-processing strategies and tools, and undertaking experiments together like the Hazardous Weather Testbed, NOAA branches NSSL (National Severe Storms Laboratory, part of the Office of Oceanic and Atmospheric Research) and SPC (Storm Prediction Center, part of the National Weather Service) and the Met Office have all benefited. one such benefit was the improvement to predicting tornadoes, large hail, and damaging wind. Kain also notes that the prediction of high-impact events is very challenging, and collaborations between meteorological groups is the best way to address it.

2.2.1 The introduction of ensemble forecasting

A major shift in modeling occurred when the first ensemble forecasts were produced. With the chaotic nature of the atmosphere, any forecast made would eventually be wrong, no matter how well constructed (Buizza, 2018). Additionally, a single forecast could not adequately portray all the different possible outcomes for a forecast (Lynch, 2008; Buizza, 2018). There was also the matter of error propagation. Even beginning with two forecasts with nearly identical initial conditions, the further they are from the initial conditions the less likely one was to get the same forecast outcome, as even the smallest errors can quickly increase due the nature of error propagation (Lynch, 2008). Forecasting low predictability events was especially challenging as it was limited by single deterministic

forecasts, so to address this problem probabilistic forecasting by use of an ensemble was developed (Montani et al., 2011; Palmer, 2018). Producing more forecasts was more likely to lead to one or more of those forecasts being the most likely by probability (Lynch, 2008). Furthermore, by examining how much a set of forecasts diverges from one another operational meteorologists could evaluate the predictability of the weather. The first real time probabilistic monthly ensemble forecast was created in November of 1985 on the Met Office system (Folland and Woodcock, 1986; Murphy and Palmer, 1986; Palmer, 2018). Soon after, ECMWF (Buizza and Palmer, 1995; Molteni et al., 1996) and NCEP (Toth and Kalnay, 1993; Tracton and Kalnay, 1993) introduced the first operational ensemble forecasts that were probabilistic (Buizza, 2018).

2.2.2 Ensemble use around the world today

As ensemble models and the computers that ran them improved, ensembles became more and more commonplace. Now, they are a key part in modern forecasting and used across the world for short, medium, and long term forecasts and climate predictions. This wealth of data presented a new opportunity to examine and improve forecasting through collaboration. TIGGE, the THORPEX (The Observing System Research and predictability Experiment) Interactive Grand Global Ensemble, was founded to improve high-impact weather forecasting by bringing together forecast ensembles from across the globe (Richardson et al., 2005). The first TIGGE workshop set the expectations and rules of the collaboration, seeking to make the data collected by the project available to all researchers who sought to make use of it. Today, the TIGGE project is still being maintained online (Santoalla and Mladek, 2022), providing data from 13 different global NWP centre models: BoM, CMA, CPTEC, DWD, ECCC, ECMWF, IMD, JMA, KMA, Meteo-France, NCEP, NCMRWF, and UKMO. These ensembles are different from one another in many ways, such as the number of members (from as few as 12 (NCMRWF) to as many as 51 (ECMWF and JMA)), the resolution of the forecasts (from 7.5 km (Meteo-France) to 139 km (JMA)), and the length of the forecast (from 48 hours (Meteo-France) to 16 days (ECCC)). However, they must all provide previously agreed upon parameters to be included in the database: 5 pressure level parameters at 8 different pressure levels, 1 parameter at a potential temperature level, 3 potential vorticity level parameters at a

(a) (b)

Figure 2.1: Diagrams of a cold front (a) and warm front (b), as two-dimensional sections of the atmosphere when a front is passing in front of the observer's location. Cold air is depicted with blue arrows and warm air is depicted with orange and red arrows.

potential vorticity level, and 28 single level variables.

2.2.3 Challenges of ensemble forecasting

While the development of ensemble models has revolutionised forecasting, it has also presented a new problem. There is now so much data from ensemble forecasts, which are often run multiple times a day, that it is difficult for operational meteorologists to digest all the information before they must issue their own forecasts to the public. Finding a way to reduce the amount of data while still allowing forecasters to maintain the same level of accuracy in their forecasts is a pressing issue that this project addresses.

2.2.4 Diagnosis of fronts

In the UK, strong wind, heavy precipitation, and flooding are some of the primary high-impact events that occur. These events are often associated with frontal regions. Fronts are defined by Dunlop (2008) in A Dictionary of Weather as "the boundary or zone of transition between two air masses of different temperature or humidity, which thus differ in density." Characteristic features of fronts have been described in Meteorology Today (Ahrens and Henson, 2016) and consist of "1. sharp temperature changes over a relatively short distance, 2. changes in the air's moisture content, 3. shifts in wind direction, 4. pressure and pressure changes, 5. clouds and precipitation patterns." Figure 2.1 contains two diagrams of different front types. In figure 2.1a a cold front is depicted. During a cold front cold dry air is replacing moist warm air. The warm air rises above the cold air and often forms clouds and thunderstorms. A warm front is depicted in figure 2.1b, where warm air is overtaking retreating cold air. The warm air can still rise above the cold air as it moves forward, causing other types of cloud formations and precipitation (Ahrens and Henson, 2016). The close relationship of fronts to precipitation made them

a likely feature for this project to focus on.

However, fronts have been difficult to expressly define and plot on maps. The work of Renard and Clarke (1965) sought to determine if objective frontal analysis was possible. They listed six desirable aspects for what a robust numerical objective method should have in order to be useful for meteorologists. It should be able to "a. locate the warm-air boundary of each synoptic-scale baroclinic zone at one or more levels; b. attach a 'strength' label to every segment of a front; c. distinguish fronts according to movement: warm, cold, stationary; d. determine the frontolytical/frontogenetical character of the fronts; e. relate the frontal-zone slope and stage of development to vertical motion, clouds, precipitation, and development of pressure systems; and f. identify the air masses separated by the fronts." They also considered what variables would be most useful for determining fronts, with lesser importance being given to precipitation and wind, and more importance being given to thermal fields such as the wet-bulb potential temperature, the equivalent potential temperature, and the potential temperature. Their work concluded that using a constant pressure surface made the possibility of objectively identifying fronts a feasible goal, though there was still a great deal of work to do. Decades later, with frontal analysis still being dominated by subjective methods, Hewson (1998) strove to find a way to objectively classify fronts, using various quantities to determine where they were and plot them graphically. They listed many previously used diagnostics for objective frontal identification in their first table. In light of these previous attempts, they aimed to satisfy five key areas with the objective method they introduced: it should be simple, intelligible, accurate, tuneable, and portable. They combined a series of mathematical derivations and equations applied to thermal fields to determine frontal regions and neighboring baroclinc zones, then used further mathematical and graphical frontal analysis techniques to pinpoint and plot fronts by computer. They also determined that the wet-bulb potential temperature was the best parameter for frontal analysis. Berry et al. (2011) built off of this work to explore seasonal and annual frontal patterns, focusing on the 850 hPa pressure surface and once again using the wet-bulb potential temperature. They used the same method developed by Hewson (1998), except they apply numerical masking and an algorithm to aid in plotting. They explored fronts in the northern and southern hemispheres, finding there was a great deal of asymmetry between the two. Later, again using the same objective method, Catto et al. (2012) was able to link most storm track rainfall

to various front types. Their analysis showed that storm track precipitation over the ocean was most commonly associated with with cold fronts. Alternatively, warm fronts brought the most precipitation over continental land. Catto and Pfahl (2013) went on to complete a global study linking an overwhelming majority of extreme precipitation events to fronts, particularly in the midlatitudes, and associated with strong gradients of the wet-bulb potential temperature. A later study (Catto et al., 2015) was able to also associate warm conveyor belt fronts, associated with mid-latitude cyclones, with extreme precipitation. However, the Hewson and Berry method of objectively identifying fronts isn't the only method currently in use.

In their paper, Soster and Parfiti (2022) examine the sensitivity of two different objective frontal identification methods on reanalysis datasets: surface fronts identified by the Hewson (1998) method and by the Parfitt et al. (2017) method. This latter method uses the horizontal temperature gradient and the isobaric relative vorticity on a given pressure surface. These parameters were chosen due to the rapid temperature change across fronts and varying wind directions ahead of and behind the front. The method developed by Parfitt et al. (2017) is simpler than that of Hewson (1998) and had a high degree of agreement when applied to strong frontal events such as cyclones. However, when there wasn't as defined of a front, i.e. when a front would require more forecaster interpretation, they methods differed. Likewise, Soster and Parfiti (2022) came to a similar conclusion. Their analysis showed that there were large discrepancies between the two methods and between datasets. They cautioned that research being done with objective frontal analysis may be hindered by relying too much on a single dataset and method. This work emphasises how even today there is still not a single agreed upon method for objectively identifying fronts and how more work must be done in this field before a consensus can be reached.

However, this debate is less relevant for the work presented here. While the foundation of frontal analysis and how it relates to extreme precipitation has lead to the choice of variable used in my work (the gradient of the wet-bulb potential temperature), both the Hewson (1998) and Parfitt et al. (2017) method for identifying frontal regions are too complex for my purposes. This work focuses on the regions where fronts are likely, as opposed to the actual front itself, and is primarily concerned with the distances between these regions. Therefore, there is no need to identify a front exactly. The use of the gradient of the wet-bulb potential temperature is explained further in chapter 3).

2.3 The methodology of issuing advisories and warnings

When the forecasts indicate an extreme event is approaching, it is critically important how this information is conveyed to the end user. Issuing an advisory before the event is relatively certain to occur has the potential to risk damaging the trust of future forecasts. If the forecasted event and advisory had to later be reduced or removed, people may perceive the warnings as being untrustworthy and may therefore be less inclined to follow provided advice on how to stay safe (Losee and Joslyn, 2018). Similarly, if there isn't enough time to act when a warning is issued or the warning isn't strong enough to adequately describe the danger, then people will be unprepared and could potentially have their lives and livelihoods at risk. It is therefore a delicate balance of when and how to alert the public of dangerous weather. Recent esearch has been conducted on how best to achieve these goals via various methods. The 30^{th} volume of the International Journal of Disaster Risk Reduction, Communicating High Impact Weather: Improving warnings and decision making processes, has been dedicated to this issue. The work within has resulted in 5 common themes: "1) the move towards providing impact based weather warnings to better support decision making processes; 2) trust and its relationship with forecast uncertainty; 3) tailoring forecasts and warnings to meet the decision needs of different user groups; 4) the emerging role of social media in the dissemination and verification of weather warnings; and 5) the wider behavioural, social, cultural and political context in which weather warnings and forecast information are used in decision making" (Taylor et al., 2018). In their study Rodwell et al. (2020) presented a beach scenario and a camping scenario to participants with weather forecasts. They were able to associate forecast probabilities with user choices by evaluating how participants viewed the probability of a "bad weather" event occurring (i.e. the beach was cold and damp and the campsite experienced high winds). They also concluded that more guidance from forecasters might help users make better decisions when it comes to high-impact events such as the high wind during the camping scenario. This project aims to improve the ease of assessing the likelihood of high-impact events by quickly extracting scenarios with a novel clustering technique. Once operational meteorologists are able to review the scenarios, they will have

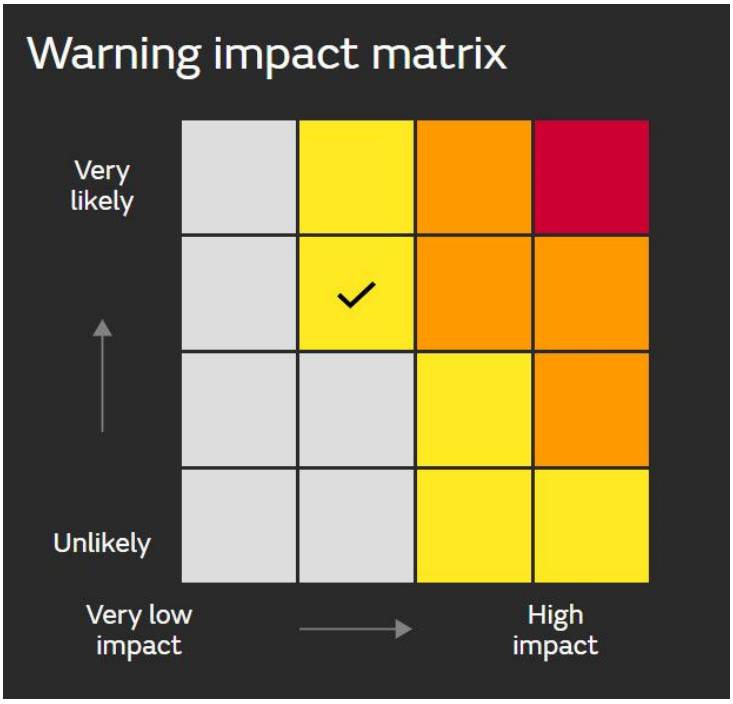


Figure 2.2: An example of the warning impact matrix used by Met Office during high-impact weather events. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.

a better understanding of the different potential forecasts and can use this information to evaluate the likelihood and severity of an event, thus informing the warning impact matrix (an example can be seen in figure 2.2) used to convey weather impacts to the end users. The matrix is based on the probabilities of an event occurring (unlikely to very likely) and how much it will impact the user (very low impact to high impact), which is based on ensemble forecast information (Met Office, 2021). The Met Office website explains the use of the warning impact matrix as such: the yellow warnings can indicate either a very likely weather event that will have minimal impacts on the populace or an unlikely event that could cause significant impacts, amber warnings mean there is an increased likelihood the severe weather will occur and will impact the populace, and red warnings are the most severe, indicating a high likelihood and severity of a weather event that may lead to loss of life.

The impact matrix is a way to reach end users and communicate potential hazards. This is an area of much research, where studies have been conducted on how best to convey severe weather risk to users. Potter et al. (2018) conducted a survey on impact-based warnings versus phenomena-based warnings. They concluded that impact-based warnings were more effective in how people perceived the risk of events, but it still was not clear that the participants would change their action based on the information. Another study

conducted by Mu et al. (2018) focused specifically how weather warnings were presented and set up an experiment to test how participants would react to these warnings if they were given costs associated with how they responded (e.g. if they would spend money to protect their assets or risk their costs in damages based on the information received in the weather warning). Kox et al. (2018) also found that how warnings were presented to specific users made a significant difference in how useful they were. For example, some groups required longer warning periods than others (such as road crews preparing for a winter storm versus fire and rescue crews responding to storm damage). They determined it was important for forecasters and end users in specific areas to work together to bring value, i.e. a high utility, to a forecast.

2.4 Clustering and its meteorological applications

Clustering is a machine learning technique that has long been used to reduce large data sets. A cluster can be defined as a group of similar members based on some metric, such as a distance (Omran et al., 2007). It can be used for a wide array of applications, including machine learning and image matching, and in an even wider array of fields, ranging from business analysis and customer care to the medical industry and environmental sciences (Wazarkar and Keshavamurthy, 2018). Within the environmental sciences, meteorology uses clustering quite often for several different applications (Wilks, 2019). Within this section, the primary applications of clustering within meteorology will be discussed, then there will be a brief discussion of the primary methods of clustering, ending with the method chosen for this project.

2.4.1 Clustering methods

The two primary clustering methods that are most often used are hierarchical clustering and partitional clustering (Omran et al., 2007). Each of these methods has a variety of different options and modifications that can be picked based on the need of a particular study. The effectiveness of the clustering method chosen can depend on a variety of factors, such as domain size, variable, and region. In his book, Wilks (2019) goes into depth about various statistical analyses, including clustering methods. It is very important to choose a measure of similarity between forecasts so that it is relevant to the problem ad-

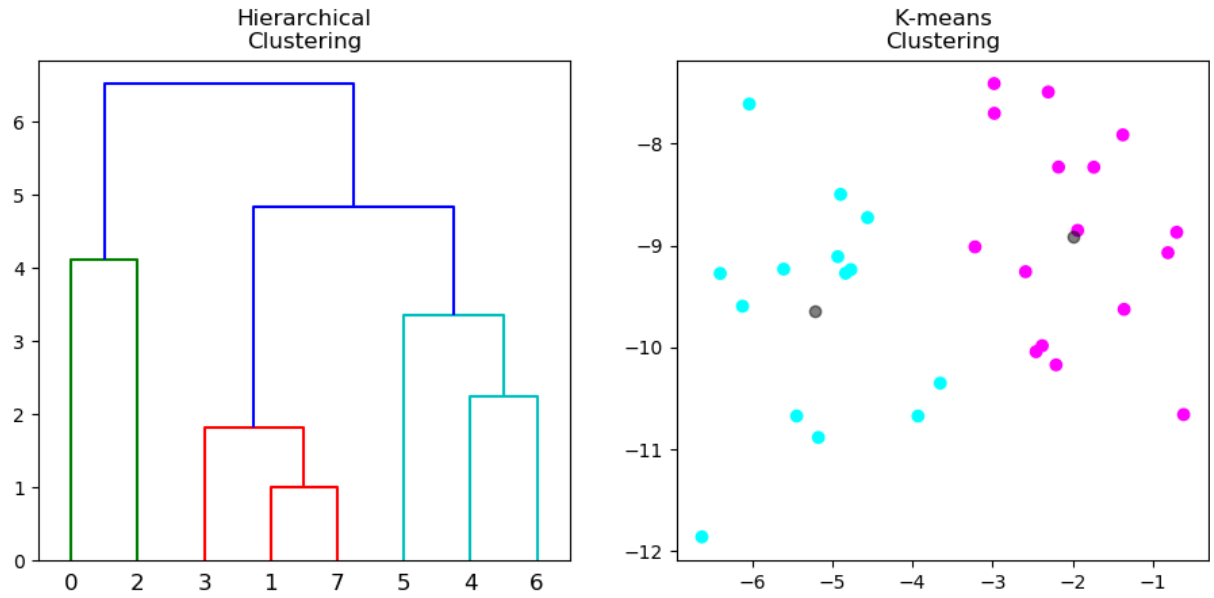


Figure 2.3: An example of a dendrogram plot used to visualize hierarchical clustering (left) and a scatter plot used to visualize k-means clustering by colouring the cluster members with the same colour and plotting the cluster centroids (mean of each clusters' members) in grey (right).

dressed. There are a variety of different approaches, though for the scope of this project, a simple clustering method was desired and a robust way to determine the optimal number of clusters based on ever evolving forecasts. The following sections briefly describe the two primary methods of clustering, their advantages, and their drawbacks.

2.4.1.1 Hierarchical clustering

Hierarchical clustering can either be agglomerative, where every member begins in a cluster of its own and then is joined into larger clusters until the desired number of clusters is reached, or divisive, where all members begin in a single cluster that is then split into more clusters until the desired number of clusters is reached (Omran et al., 2007). A visual example of hierarchical clustering can be seen in the left plot of figure 2.3, where a dendrogram tree shows how members (x-axis) are clustered based on the y-axis. There are several different distance metrics that can be used to compare the differences between members or combined members such as single, complete, centroid (Omran et al., 2007), average, and minimax linkage (Wilks, 2019). Clusters are first formed by joining the two nearest members. This process continues, with new clusters forming, members joining preexisting clusters, or two clusters joining together until the desired number of clusters is reached. As clusters grow in size, the average distance between all combinations of members between two clusters is calculated and used as the distance between two clusters.

While the hierarchical method is common, it does have a significant drawback. Once a member is assigned to a cluster, it is unable to switch to another cluster later in the process. This rigidity in membership may not beneficial for all applications, however, as multidimensional members may overlap with other clusters to some degree and may fit a different cluster better at a later stage than the one it was previously assigned to. This method can also lead to a snowballing effect, where there is one large cluster and one or more single member clusters. This drawback is important to consider because it can be difficult to distinguish if a singleton cluster is a genuine outlier forecast or it it appears to be an outlier due to the nature of the clustering method.

2.4.1.2 Partitional clustering

Partitional clustering, also known as non-hierarchical clustering, takes the number of clusters supplied by the user and attempts to determine the desired distribution of members by adjusting the centre points until the optimal distribution is found, instead of slowly adding or removing members until the desired number of clusters is reached. This begins with a number of points equal to the user specified number of clusters as the centre points. Then the members are sorted into clusters based on the centre they are closest to. The centre can then be recalculated any number of times until the optimal distribution of members to clusters is found. This can result in many different solutions depending on the distribution of the initial centre points. It can be a computationally expensive form of clustering, however just as with hierarchical clustering there are several varieties of partitional clustering and distance metrics that can be applied to achieve the desired results.

- K-means clustering

Omran et al. (2007) discusses several common varieties of partitional clustering, the most widely used being K-means clustering, often using the Euclidean distance between members. Like the hierarchical method, the distance metric used to calculate cluster membership can be adjusted. However, the K-means method requires that the number of clusters be chosen at the beginning of the process, and all members are considered and grouped into clusters at the same time. To begin, a number of initial centre points (centroids) equal to the number of suspected clusters are chosen within the data set.

These points are often seeded randomly. This can occasionally result in two or more different solutions if K-means is applied to a data set that has a high degree of variability between members and lacking a strong clustering signal. Members are assigned to a cluster based on which centre point they are closest to. The centre is then recalculated as a mean of all members within the cluster, and the assignment of members begins again. This iterative process can be run any number of times up to convergence at a specified tolerance. An example of K-means clustering can be seen in the right plot of figure 2.3, where two clusters of members are displayed as different colours and the centre points are in grey. K-means allows members to move between clusters during the next iteration if their current grouping isn't the best solution. This enforces a greater similarity between members in each cluster than the hierarchical method. This also reduces the snowball effect when there is sufficient variation among members, making singleton clusters rare but significant as they are genuinely separate from the rest of the clusters. A downside to K-means clustering, however, is that the centre point for each cluster is a mean of the members in that cluster. For smoothly varying fields such as pressure, this is not a significant issue. But when examining any field that contains binary or discrete features, such as fronts or precipitation patterns, the mean may no longer look like the original input fields. It may also obscure significant details by smearing or smoothing them out.

- K-medoids clustering

K-medoids clustering is similar to K-means. It has the same benefit as K-means where all members are clustered at the same time and members can be moved between clusters until a solution is found. In their work, Brusco et al. (2019) described four inherent advantages of K-medoids over K-means. The most significant advantage is that the centre point (medoid) is restricted to being a member of the cluster instead of a mean of members. This avoids losing any fine scale features or other significant details of the forecast as the mean would smooth them out and ensures the forecast that is the medoid is a solution to the atmospheric equations. As this project is designed to provide forecast scenarios, it is crucial that the representative forecasts chosen are actual forecasts. The second advantage Brusco et al. (2019) cited was how K-medoids will provide an exact solution for large data sets, as there are only so many combinations of members and medoids as clusters possible before finding the optimal distribution. Alternatively, K-

means may have difficultly converging on a single solution as the centroid is recalculated based on the membership of the clusters. Thirdly, they cite how K-medoids performs in terms of outlying members. As the centroid of K-means is not a stable point, genuine outliers may be grouped together with their nearest cluster. However, K-medoids is more likely to recognize an outlier as a singleton cluster if it is significantly different to other members. With respect to forecast scenarios, being able to extract true outlying forecasts is important as they may contain crucial information about the predictability of the atmosphere. In their final point, Brusco et al. (2019) describes how K-medoids has the ability to compare cluster members by any difference metric. This is a significant factor when comparing objects in a forecast where the typical difference metric of the Euclidean distance, used in K-means, is a less desirable metric for comparison. In this project, the differences between members is calculated by a verification technique (discussed further in section 2.5.1), therefore a clustering technique that accommodates this is key. All of these reasons lead to the decision to use K-medoids as the clustering algorithm for this project.

2.4.2 Applications of clustering

Clustering has been applied for a variety of different reasons to a variety of different meteorological phenomena. Often, clustering is used to develop a general picture of common patterns in the atmosphere. Some of the most common examples of clustering meteorological data are described in the following sections.

2.4.2.1 Circulation and synoptic classifications

One of the most common uses of clustering within atmospheric sciences is for classifying various synoptic weather patterns and atmospheric circulations, which are then used for weather prediction, climatology, or as a way to compare other variables associated with the patterns (Huth et al., 2008). The COST733 project (Huth et al., 2008; Philipp et al., 2010; Tveito et al., 2016) sought to create a catalog of circulation types and methods over Europe for various regions and domain sizes. Although various types of clustering made up many of the methods used to create the catalog, there were numerous other methods employed as well such as subjective methods, threshold based methods,

principal component analysis, and leader algorithms (Philipp et al., 2010). Many of the circulation classifications from COST733 have been used in further studies that compare the different methods, or compare a method to observations, or use the circulations from the methods as a way to investigate other variables. One such study by Beck and Philipp (2010) compared the classification methods and determined that non-hierarchical clustering methods, such as variations of k-means clustering, performed the best at classifying mean sea level pressure patterns. However, this did not translate to other variables, such as the temperature at 2 meters and total precipitation. In a study that sought to examine the relationship between precipitation and circulation patterns over Spain, Casado et al. (2010) used circulation catalogues from COST733 to determine which was best compared to observations. Overall, they found non-hierarchical clustering methods worked best for two of the three regions they designated. However, they also determined that in regards to Spanish precipitation, different classification use lead to different results and and conclusions. This implies that the the choice of the method can have an impact on the outcome of a study. In another study, Kassomenos (2010) used circulation patterns from the catalogue created by various methods to explore the relationship between the patterns and the occurrence of wild fires in Greece. Their study concluded that synoptic classification analysis for wild fire prediction was a feasible prospect, using both hierarchical and non-hierarchical classification methods, and a possible model could be developed to aid in forecasting.

In their paper, Neal et al. (2016) explored a new method for determining circulation types via K-means clustering. MSLP data was gathered from the UK and surrounding European area and used to create a set of circulation patterns that would then be used to make the first known weather regime forecasting tool, Decider, which is used in the Met Office. Furthering their work, Richardson et al. (2020) introduces Fluvial Decider, focusing on flood forecasting using the same weather patterns derived for use in Decider. In their paper, Richardson et al. (2020) explored how extreme precipitation could be related to circulation patterns and developed the Fluvial Decider tool to take advantage of these relationships. The tool can alert operational meteorologists to potential extreme precipitation events that might induce flooding. This tool was made operational in 2017 at the Met Office Flood Forecasting Centre. Ferranti and Corti (2011) used a modified K-means clustering technique created by Straus et al. (2007) on the 500 hPa geopotential

fields from ECMWF EPS forecasts and subsequently compared the clusters to pre-defined circulation patterns, summarizing ensemble information and providing additional climatological information for operational use. Kassomenos (2003a) and Kassomenos (2003b) use a combination of factor analysis to reduce the variable data into linear functions and k-means clustering to evaluate circulation types over southern Greece. Using this unique technique they were able to classify eight circulation patterns each for winter, spring, and autumn, and four for summer. The SANDRA method (simulated annealing and diversified randomization, (Philipp et al., 2007)) uses a modified k-means clustering algorithm to classify mean sea level pressure patterns before doing further analysis on long-term temperature variability. They found the conventional K-means clustering could not provide a stable result, but applying SANDRA did. Their work linked warming trends to the changes in circulation patterns over central Europe. Upper air circulation patterns derived from geopotential heights and thicknesses with K-means clustering by Enke and Spekat (1997) are used with downscaling to compare with observations of several variables. They found that their method provided a good agreement between their downscaling and observations, thereby providing a method that can reconstruct local weather variability. US east coast winter storm mean sea level pressure patterns were explored with a variation of partitional clustering by Zheng et al. (2017). Their method of clustering begins with an empirical orthogonal function analysis that extracts the leading principal components, then a fuzzy clustering technique (originally presented in Scott and Symons (1971)), which has some similarity to K-means, is applied. The results of their method allowed quick extraction of different forecast scenarios that could increase forecaster awareness of them.

2.4.2.2 Clustering trajectories

Another common use of clustering is with air mass and storm trajectories. Air mass trajectories describe the movements of air parcels over a given time. Air parcels arrive to any given destination from many different locations and pressure levels. The qualities of the parcels can greatly effect a region beyond just temperature and humidity; pollutants, various aerosols, and particulate matter (i.e. dust, ash, and pollen) are carried into an area from a neighbouring region, which directly affects the air quality. By calculating and clustering the parcel back trajectories, air quality in a region can be linked with synoptic patterns, improving the forecasting for these variables. A study by Delcloo

and Backer (2008) clustered 3 dimensional back trajectories of air parcels with concern to ozone concentration. They used a non-hierarchical clustering method and derived a way to determine the appropriate number of clusters. By using the root mean square deviation of a trajectory to its cluster centre, they were able to statistically derive the optimal number of clusters. Similarly, a study by Cape et al. (2000) was also concerned with the optimal number of clusters, albeit with a hierarchical clustering method applied to trace gas air parcel trajectories. They opted for a version of the RMS and R^2 values to help determine the right number of clusters. Using their method, they could detect different parcel trajectories for different measured ozone concentrations. Their method may have future applications in determining the chemical composition of the air at sites that currently lack this data but do have air parcel back-trajectories available. In the paper by Hart et al. (2015), they used a hierarchical agglomerative clustering technique to cluster air parcel trajectories in relation to extra-tropical cyclones (ETC), focusing their study on Cyclone Friedhelm, which impacted Scotland in 2011. They argued that applying a threshold to the airstreams would likely artificially restrict or inflate the number of trajectories and therefore couldn't be relied upon for clustering analysis. Instead, they chose parcel trajectories that were near either the warm conveyor belt or the cold conveyor belt of the ETC to cluster, resulting an a more objective method of identifying these airstreams.

In storm trajectory analyses, examining the clusters can lead to a deeper understanding of how the storm is interacting with the atmospheric flow and the likelihood it affected the storm's progression. The forecasted storm trajectories of hurricane Sandy, a tropical storm that interacted with a midtropospheric trough during its northward journey along the east coast of the USA that caused it to regain deadly strength before impacting New Jersey, were clustered and analysed via a regression mixture model by Kowaleski and Evans (2016). They examined the variations in the trough and storm interaction through the clustering results, reducing the data into a few distinct outcomes. Their work has the potential to be used in future tropical cyclone forecasting by providing probabilities of track occurrence.

2.4.2.3 Feature based / object-oriented clustering

Clustering circulation patterns and air mass trajectories can lead to a greater understanding of atmospheric motion and regional influences. When examining a smooth field,

such as MSLP, or a trajectory, comparing point-wise differences between members works well. However, this type of comparison is less desirable on fields that are not smooth, such as precipitation objects or fronts. When clustering regions of rainfall or related features such as fronts, forecasters are likely to be interested in spatial displacements between objects. This requires a different approach to clustering as a point-wise comparison will introduce a double penalty, where you get a large error from the displacement of a feature, both where it is forecast (and consequently not observed) and where it is observed but not forecast. To avoid this, regions of precipitation can be considered objects with various characteristics, i.e. those determined by MODE (the method for object-based diagnostic evaluation, further explained in section 2.5.1.1, (Davis et al., 2006a,b, 2009)), and then clustered. An example of this kind of clustering can be seen in the works by Johnson et al. (2011a) and Johnson et al. (2011b), who used data from convection-allowing ensemble forecasts and analysed it with an object oriented hierarchical clustering based on MODE to examine how perturbations affected the forecasts. They determined that ensemble design should depend on what it will be used for, i.e. there are different perturbations that would benefit near to surface variables more than upper-level variables, and vice versa.

2.4.2.4 Ensemble reduction

Many forecast models produce ensembles, some of which use a large number of forecast members. Ensembles provide many possibilities for how the atmosphere may evolve and large ensembles may use data reduction to extract the most salient information. This can result in clustering of the ensemble members to extract scenarios which can then be used for further analysis, such as with the works by Molteni et al. (2001), Marsigli et al. (2001), and Montani et al. (2011), who use hierarchical clustering on the wind vector, wind direction, or vorticity, in their method to extract representative members based on these fields or precipitation. These representatives are determined by the average distance between members in a given cluster and retain detailed features which is important as using the centroid of the cluster would result in smoothing the field. Once the representative is chosen, it can then be used for providing initial conditions for high-resolution model runs or as boundary conditions for nested forecasts. They found that by using a representative member to initialize a higher-resolution forecast they could obtain a far more detailed

forecast for local weather. This method is further used and studied in COSMO-LEPS (Montani et al., 2011), the COnsortium for Small-Scale MOdelling Limited-area Ensemble Prediction System. The system works by first joining two successive ensemble runs into a super-ensemble then grouping the members into five clusters (Montani et al., 2003, 2011). The variables of choice then go through a standardization process and then the distance between members is calculated. The five clusters are then created and a representative member is chosen from each. These members then provide initial and boundary conditions used to generate high-resolution limited area models. The cluster member that "minimises the ratio between its distance from the other members of its own cluster and its distance from the members of the other clusters" is chosen as the RM (representative member) (Montani et al., 2003). By defining the representative member in this way, COSMO-LEPS avoids the common issues when the centre does not accurately represent an atmospheric solution, such as a mean. This was also a crucial aspect of the method presented in this work, where the representative member must be an atmospheric solution and not a mean of a cluster. This drove the choice as to what clustering method would be used (see section 2.4.1.2, as well as how the representative member is chosen in section 3.3.4.2).

To evaluate the benefits of various clustering techniques, variable use, distance metrics, and resulting skill with regards to ensemble reduction being used to initialize limited area models, Serafin et al. (2019) found that the effectiveness of clustering is dependant on several factors. Variable choice and lead time of the forecast were very important aspects of getting good clustering results. However, they found that using clustering results for ensemble reduction are not necessarily any more skillful than random sampling when used as initial or boundary conditions for limited area ensembles, though it did improve with longer lead times. A key finding was that the clustering algorithms only had value when there were meaningful differences between members. Their work implies a careful choice must be made in the clustering variable and that clustering is most beneficial when members have had enough time to sufficiently perturb away from the control. In the method I developed, this was also a concern. Therefore, a new technique was developed (see section 3.3.4.1) to determine when clustering was becoming distinct and representative scenarios could potentially be extracted.

2.4.2.5 Temporal clustering

When temporal dimensions are considered, they are typically examined as a whole or over a window instead of at each iteration, i.e. if a data set included three days worth of data, each member to cluster would be three days worth of data or a segment would be extracted as a window of time, such as a single day's worth of data. Then three days (or a single day as a window of time) would be compared to each other at a time. An example of window clustering can be seen in the ECMWF EPS clustering system described in Ferranti and Corti (2011), where each member to be clustered contains data for a specified time window. They use K-means clustering at four different time windows to examine the evolution of 500 hPa geopotential synoptic development. The clusters are then categorized by pre-defined climatological regimes that affect the Euro-Atlantic region. This information is made available for forecasters to better inform them of potential atmospheric scenarios. In Leckebusch et al. (2008) 3 day episodes of 1000 hPa geopotential height or MSLP is clustered with K-means to examine the development of wind storms over Europe. They found using clustering to determine weather patterns associated with the wind storms over the 3 day window was very useful. However, this method is not as conducive to comparing real-time forecasts where time evolution of scenarios is critical. Clustering at each lead time allows members to move between scenarios if one particular scenario is a better fit than the previous. Therefore, this is the temporal method used within this project.

2.4.2.6 Clustering used in forecasting today and key points

Clustering can be a useful tool for examining meteorological phenomena and reducing large data sets. Some forecasting centres currently use clustering when analyzing their ensemble data and creating their forecasts. As noted in section 2.4.2.1, ECMWF uses clustering on the 500 hPa geopotential field to provide atmospheric evolution forecasting products over the North Atlantic and European region (Ferranti and Corti, 2011) and the Met Office uses Decider, which is a weather regime forecasting tool that utilizes clustering on MSLP (Neal et al., 2016; Richardson et al., 2020). NCEP (National Centers for Environmental Protection) and NOAA (National Oceanic and Atmospheric Administration) in the USA are currently prototyping a clustering tool for ensemble uncertainty, extremes,

and forecast scenarios (Rutz et al., 2022). Their method uses Empirical Orthogonal Functions on the 500 hPa height to create clusters. Deutscher Wetterdienst (DWD) provides a forecast product that clusters ensemble members based on the Grosswetterlagen (GWL) circulation patterns originally developed by Baur et al. (1944) (DWD, 2023). They produce a table that shows how many ensemble members match which GWL patterns over a 15 day period. Using 30 weather patterns developed by Neal et al. (2020) over the Indian subcontinent and source code provided by the Met Office, the India Meteorological Department creates similar weather pattern tables after clustering ensemble members to the most closely related pattern (Pattanaik, 2022). Météo-France uses cluster representatives from the PEARP ensemble to determine the lateral and upper boundary conditions for AROME-EPS, a convection-permitting ensemble (Bouttier et al., 2016).

Within atmospheric sciences, both hierarchical and partitional clustering are popular, and each method can have several variations applied to fit the study needs. Typically the goal of the project or application will determine what type of clustering is used, with results depending on the domain size, the complexity of the variable, whether the data is a smooth field or more similar to objects, and whether the data is from a single point in time or evolves. These reasons make clustering a useful tool for meteorological applications and should be considered when looking for patterns or the reduction of data sets.

2.5 Comparison methods

In clustering, how members and clusters are compared to one another is a key feature of any given technique. There are a variety of distance measures used, detailed at length in Wilks (2019), but the most common is the Euclidean distance, particularly for partitional algorithms such as K-means. Some other metrics Wilks (2019) notes are the Karl-Pearson distance, the cosine, and different types of correlations. Within hierarchical clustering, a second distance metric must be considered beyond how two members compare, which is the distance metric used to define how two clusters compare. These metrics include the single-linkage, complete-linkage, average-linkage, centroid, and minimum linkage. As this project uses K-medoids, it might be anticipated that the Euclidean distance would be used to compare members. However, this metric is not sufficient in this case. With K-medoids, the user can choose their own distance metric so trying unconventional options is

possible. As the project aims to extract a representative member for potential scenarios, it is important to consider how that scenario might be compared to observations at a later date. Therefore, this project uses a forecast verification method to compare members. The sections below will briefly discuss forecast verification methods and their varieties, and the method chosen for this project: the Fractions Skill Score.

2.5.1 Forecast verification methods

Forecast verification is an essential part of atmospheric sciences and seeks to establish how skillful a forecast is compared to observational data or other forecasts. In their book, Jolliffe and Stephenson (2012) discuss the various types of forecasts and methods of verification in depth and touch on the primary reasons for verifying forecasts. Traditional methods of forecast verification are based on a point-by-point comparison between a forecast an the observations (Gilleland et al., 2009). They might focus on the RMSE, mean error, or a hit or miss ratio. Non-traditional methods fall into four different categories of verification (neighborhood, scale separation, features/object based, and field deformation, (Gilleland et al., 2009; Jolliffe and Stephenson, 2012)), and it is important to match the forecast, verification data, and verification method to the needs of the user. Scale separation and field deformation methods tend to focus on the errors of a forecast, whereas neighborhood and feature based methods focus on the similarity between forecasts. As clustering is a method that is based on similarity between members, the latter verification methods are addressed further.

2.5.1.1 Feature based methods

Feature based methods, described at length by Jolliffe and Stephenson (2012) and Gilleland et al. (2009), compare features, otherwise known as objects, within a forecast. These objects can be maxima/minima within a field, such as high/low pressure centers, areas of rainfall, etc. A series of qualifiers are used to describe the objects, then the qualifiers can be used to compare objects between forecasts and observations, similar to how a forecaster might do so.

A well known feature based method is MODE, the method for object-based diagnostic evaluation, created and described by Davis et al. (2006a) and then further explored and

evaluated in their companion paper Davis et al. (2006b) and later in Davis et al. (2009). MODE begins by convolving the field of choice into pre-chosen shapes. This is done to smooth the field. A threshold is then applied to extract areas of interest. The boundaries of these areas can now be detected as objects. At this stage, these areas can now be related to simple object shapes for easier calculation. These final shapes are now given various qualifying data, such as the intensity of rainfall within the shape from the original field, the area the shape covers, the centre of mass, the major axis angle, the aspect ratio, and the curvature of the shape. These can then be used to match objects between forecasts and observations.

Another feature based method is SAL (structure, amplitude, and location) which was developed by Wernli et al. (2008). SAL uses three diagnostic quantities about an object for comparison. First, precipitation objects are identified by applying a threshold and selecting the contour around the maximum precipitation. Once the object has been identified, the diagnostic quantities can be assessed. The "structure" of an object relates to its volume (a function of precipitation), its "amplitude" is derived from domain-averaged precipitation, and its location is based on the distances between centres of mass, with a second component designed to account for when two different fields have the same centre of mass. The object can now be compared to objects derived from the observations. A major difference between SAL and other object oriented methods is that it doesn't require a similar object to be present in both fields in order to compare them.

These methods excel at categorizing precipitation objects. It likely would also work well for comparing the gradient of the wet-bulb potential temperature, which corresponds to regions likely to have fronts and will likely have many broken areas. However, as the method developed in this work is primarily concerned with simple displacement between frontal regions and less about the shape of the regions, there are other options for comparison that are more appealing for their simplicity at this time, such as neighbourhood verification methods.

2.5.1.2 Neighborhood methods

Feature based methods are useful for comparing objects that can be derived from forecast fields, but they are inherently limited to those objects. These methods are complex and depend on easily finding objects for comparison. Additionally, point-by-point meth-

ods are unable to detect closeness between objects. To alleviate both of these issues, a spatial neighbourhood comparison method can be used. Jolliffe and Stephenson (2012) found that using a spatial neighbourhood method to compare high-resolution fields, taking multiple potentially fine scale objects into account, is preferred. As detailed by Gilleland et al. (2009), these methods have several advantages over other techniques. Neighbourhoods, small grid boxes that can range in size and shape, are used to compare sections of a forecast to another forecast or observation using previously developed and tested verification scores as the metric. By comparing field to field in small areas, near-misses are no longer are subjected to a double penalty. As the neighbourhood size is variable, the forecast resolution can be adjusted by changing the size until there is a skillful match with the observations. These reasons make neighbourhood methods a good choice for comparing gradient fields.

- Fractions Skill Score

A notable neighbourhood method developed by Roberts and Lean (2008) is the Fractions Skill Score (FSS). Further analysis and additions to the method were completed in several later studies (Roberts, 2008; Skok, 2015, 2016; Skok and Roberts, 2016, 2018). Originally developed for use on precipitation forecasts, this method begins by applying a threshold to a precipitation field then converting the remaining data to binary. A neighbourhood, a sub-section of the domain size, is applied to a grid point, creating a sub-domain that is then used to calculate the fraction of hits (1s) within the area. The FSS neighbourhood is applied to all grid points in both the forecast and the observations. By comparing the fractions of each neighbourhood in a forecast to corresponding fractions from an observation field, it can be determined how similar the fields are. As the neighbourhood size can be adjusted, the appropriate forecast resolution to achieve the highest skill can also be determined. The highest skill will be given by the largest neighbourhood, so a balance must be achieved by finding the smallest neighbourhood size that results in a forecast that is more right than wrong. Formalized in Skok (2016); Skok and Roberts (2018), this balance was achieved when the FSS was equal to 0.5, which has a direct relationship to a measure of separation distance between features in a forecast. Gilleland et al. (2020) performed a study that tested several different distance metrics, one of which being the FSS distance, and found that as long as the frequency bias (the

difference in the number of non-zero points between the two fields) between members remained small, the separation distance between members derived from the FSS was good. The FSS can be applied to different types of features such as the idealized rainbands explored in Roberts and Lean (2008) and Skok (2015). Building on the foundational work by Roberts and Lean (2008) who first introduced the FSS used on an idealized rainband, Skok (2015) did an in-depth analysis of the FSS solution from idealized rainbands. While Skok's analytical solution of the FSS is restricted to this idealized case, it does pave the way for using the FSS on other rainband like objects, such as fronts. The ability to derive a displacement between forecasts is particularly useful, as when forecasts are viewed by operational meteorologists they can quickly see how well two fields agree, e.g. if one forecast has a front further west than another both the operational meteorologist and the FSS will recognize it. For these reasons, the FSS was chosen as the distance metric for the clustering algorithm (see section 3.3.1.1).

2.6 Conclusion

The development of ensemble numerical weather prediction has lead to a surge in better forecasting techniques and tools. By having multiple forecasts produced at a time, ensembles capture forecast uncertainty in a way a single deterministic forecast was not able to. This has lead to a massive increase in available data for operational meteorologists to use in creating their forecasts, particularly as resolution steadily increases. It is difficult to examine all the data to create a forecast, so it is important to find a way to extract potential scenarios to better understand the state of the atmosphere and convey that information to the user. This can be accomplished via clustering.

There are several varieties of clustering methods that have been used on meteorological applications, each with its own benefits and drawbacks. Picking the best clustering technique, variable, distance metric, and optimal number of clusters is of concern when choosing a method. Partitional clustering, such as K-means, is often used for grouping fields of data. However, the mean is not suitable for a scenario, which must be an actual forecast. Therefore, K-medoids was chosen as the clustering method for this project.

All clustering methods require a difference metric, or a way to describe a distance, between cluster members. One possibility of a metric is a verification score between

forecasts. Using the FSS provides a simple, direct measure of distance between members and avoids the pitfall of the double-penalty problem.

In the following chapter, the methodology of this project will be described. This will include a detailed discussions about the problem and the goals of the project, data and assumptions that have been made, the various design elements, and the computational algorithm.

Chapter 3

Clustering ensemble members to optimise consistency with lead time

3.1 Introduction

3.1.1 Outlining the Problem

Currently, forecasters have large amounts of complex high-resolution data available to them to review before issuing their forecasts. This includes multiple members within an ensemble and multiple fields of different variables. For high-impact weather, such as extreme precipitation, damaging wind storms, blizzards, flooding, and other events that pose a risk to the public, time is critical when delivering accurate and informative forecasts. Forecasters must be able to rapidly decide what information is most important then quickly and accurately create a forecast and any warnings required. To digest all the data necessary to issue a forecast in a time critical situation and to repeatedly review advisories as numerical forecasts are updated with later start times presents a difficult challenge. Forecasters would like to have all the complexity reduced to key messages as a starting point to know what to look for in more detail in the short time they have available, e.g. different scenarios that they can then ascribe probabilities to so they can construct a story.

3.1.2 Goals of the Methodology

The goal of this work is to help forecasters quickly identify different possible future scenarios from the large numbers of forecast products available. The information in the ensemble forecast is reduced to several scenarios (2-6 are considered) by clustering together the most similar ensemble members. These scenarios should be distinct from each other in terms of forecast impacts from the weather systems. To do this, a new clustering based approach has been developed, which is described in this chapter, where the similarity between forecasts, variable choice, and clustering technique are considered. The process was designed with the following in mind:

- the clusters should be coherent in time,
- clustering should be performed such that clusters are distinct with respect to features which forecasters routinely track to identify significant weather,
- clusters should be well represented by a single ensemble member (as opposed to an average),
- and the algorithm should be usable for different model resolutions and ensemble sizes.

The horizontal gradient of the wet-bulb potential temperature θ_w has been chosen as the meteorological parameter for this work as it is used for defining various different types of weather regimes, e.g. by defining air masses and precipitation (see section 3.2.2 for details).

3.1.3 Outline of chapter

This chapter will begin with a brief description of the data used to develop the algorithm and assumptions made during the process. Next, it will cover the design elements within the process which includes the clustering method, the distance metric used, variable choice, the technique used to match clusters between lead times, traceability (the quality of how members in a cluster at one time relate to members in a cluster at a different time) through lead time, and determining members that are most representative of each cluster

throughout the time window of interest. Finally, the computational algorithm and the processes will be discussed in depth.

3.2 Data and assumptions

3.2.1 Ensemble Forecast Data

To begin crafting the methodology, data were chosen that corresponded to a time period of significant high-impact weather, i.e. frontal regions that produced widespread heavy precipitation over the UK that caused flooding of homes and disruption to travel. The original chosen forecast began on 10/10/2018. This period of time saw storm Callum, a mid-latitude cyclone reaching a minimum pressure of 938 hPa, bring high winds and heavy rain to the UK on the 12^{th} and 13^{th} (Met Office, 2018), with 160 mm of rain falling in Libanus, Brecon over a 24 hour period (Prichard, 2018). The initial forecast data were taken from the operational Met Office MOGREPS-G ensemble (Bowler et al., 2008), which contains 18 members and runs for a total of 8.25 days at 3-hourly intervals, at the time this project began. The variable is the wet-bulb potential temperature θ_w in K and the gradient of the wet-bulb potential temperature $|\nabla\theta_w|$. The domain was chosen to encompass the UK and the surrounding area, particularly upstream into the North Atlantic, from 40°N to 70°N and from 45°W to 45°E. The MOGREPS-G data has a latitude resolution of 0.1875° and a longitude resolution of 0.28125°. As the algorithm developed, more data for the month of October 2018 was utilized to refine the processes.

3.2.2 Variable Choice

The wet-bulb potential temperature θ_w, which is a conserved property during both unsaturated and saturated reversible adiabatic processes, is a useful tracer of air parcels and is therefore a good way of identifying different air masses and how they move around (Dunlop, 2008). The gradient of θ_w is therefore an indicator of air mass boundaries and has been analysed in many studies on fronts. Renard and Clarke (1965) was the first study that endeavoured to design a method for computing fronts instead of relying on subjective determination. Although they intended to focus on θ_w, the equivalent potential temperature θ_e, and their derivatives, deficient hemispheric moisture fields forced them to

shift their focus to the potential temperature θ and its derivatives. Their work paved the way for many subsequent analyses of different variables which are reviewed extensively in Hewson (1998), which sought a simple and accurate way of objectively identifying fronts and concluded fronts obtained via analysis with θ_w were more useful than those obtained by other variable choices. Berry et al. (2011) used the methods created by Hewson (1998) on the ERA-40 reanalysis data (Uppala et al., 2005) to identify fronts and compile a global climatology. Catto et al. (2012) then used the method developed in Berry et al. (2011) to explore precipitation in relation to fronts. Extreme precipitation, a high-impact event that can threaten the UK, has been tied extensively with fronts, particularly within the mid-latitudes (Catto et al., 2012; Catto and Pfahl, 2013). Warm-conveyor frontal rain can cause severe flooding due to extended periods of light to moderate rainfall, especially with orographic enhancement over wind-facing hills. Severe rainfall can also be caused by deep convection. Although this type of rain is not necessarily associated with fronts, it still must occur in a favourable environment, which can be better identified by examining environments separated by the gradient of the wet-bulb potential temperature $|\nabla\theta_w|$. Although severe rainfall is the impact of interest, it is dependent on model resolution and the parametrization of convection, while the wet-bulb potential temperature is not. For these reasons, $|\nabla\theta_w|$ was chosen for developing the algorithm.

3.3 Design elements

3.3.1 Distance metrics

Clustering algorithms require a distance metric to compare members. Many algorithms have pre-set metrics to use, however, K-medoids (further described in section 3.3.2) allows for a variety of distance metrics to be employed, including user defined metrics. In the case of complex data sets that are not smoothly varying and have a high potential for the double penalty problem, such as gradient fields, algorithms for computing spatial distance can be used. Potential choices of distance metrics include various feature-based verification methods and neighbourhood verification methods, explored in chapter 2.

The Fractions Skill Score (FSS) (Roberts, 2008; Roberts and Lean, 2008; Skok, 2016; Skok and Roberts, 2016, 2018) is a notable neighbourhood analysis which has similarities

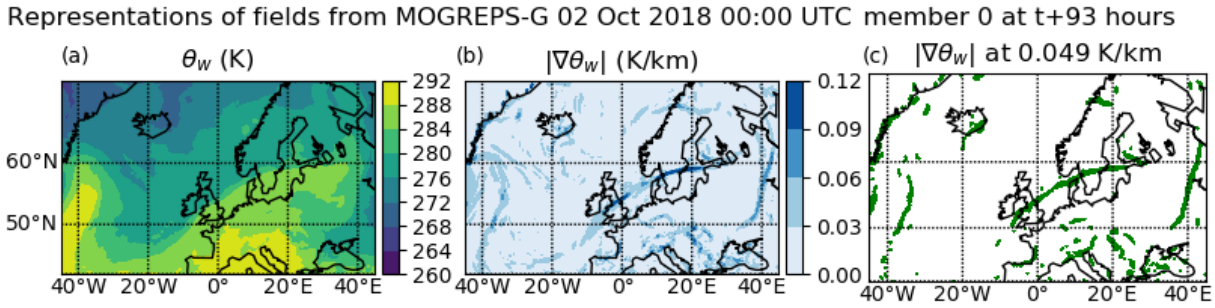


Figure 3.1: A comparison of θ_w at 850 hPa in K, $|\nabla\theta_w|$ at 850 hPa in K/km, and a binary field of $|\nabla\theta_w|$ at threshold 0.051 K/km from MOGREPS-G 02/10/2018 forecast at t+93 hours.

with the Brier Score (Brier, 1950). The FSS looks for agreement within a domain by using neighbourhoods instead of object to object agreement. Although many spatial-distance methods could be used in the algorithm, the FSS was chosen for developing the distance matrix for K-medoids clustering. This is due to how the FSS specifically avoids the double-penalty by focusing on domain wide agreement instead of object based agreement, and how it is convertible into a real distance between fields. This study is the first one where the clustering of ensemble members has been performed using FSS based distance measures.

3.3.1.1 Fractions Skill Score

The FSS is a neighbourhood analysis that first converts a field into a binary field on the model grid by setting all grid points with values above or equal to a given threshold to one and those below the threshold to zero. Figure 3.1 illustrates the conversion of θ_w at 850 hPa first to $|\nabla\theta_w|$, and then $|\nabla\theta_w|$ to a binary field to be used in the FSS, where values above/below the threshold of the 97^{th} percentile of $|\nabla\theta_w|$ values in figure 3.1b have been set to 1/0 in figure 3.1c. This process highlights frontal objects, locations where $|\nabla\theta_w|$ is large. This percentile was chosen through trial and error during the development phase of the project. Different thresholds were trialed until the visually desired level of frontal features were present and the resulting clustering appeared to draw out scenarios. Further refinement of the threshold and sensitivity testing is recommended for future work.

After the threshold is applied, the number of grid points exceeding the threshold is calculated over a square neighbourhood surrounding each grid point to give a fraction for the neighbourhood. Usually the neighbourhood is calculated for an n grid point by n grid point square around the point of interest. This is done for each point within the field and

is then converted to the FSS via equations 3.1 to 3.3,

$$FSS_{(n)} = 1 - \frac{MSD_{(n)}}{MSD_{(n)ref}} \tag{3.1}$$

$$MSD_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [F_{1(n)i,j} - F_{2(n)i,j}]^2 \tag{3.2}$$

$$MSD_{(n)ref} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F_{1(n)i,j}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F_{2(n)i,j}^2 \right] \tag{3.3}$$

where *MSD* is the mean square difference, F_1 and F_2 are two different member forecast fractions, n is the neighbourhood length, and N is the domain length (Roberts and Lean, 2008) and x and y are restricted to being the same size in this study. Following Skok and Roberts (2016), a neighbourhood size is chosen that gives, on average, a FSS approximately equal to 0.5. Then, the FSS can be converted to an approximate relative distance between fields with equation 3.4 (Skok and Roberts, 2018),

$$D_{FSS} = (1 - FSS_{MS})n * dn \tag{3.4}$$

where D_{FSS} is a distance in km and FSS_{MS} is the FSS at a minimum size discrete neighbourhood length n that is closest to 0.5 and dn is the average length of a grid square in the data.

3.3.2 Clustering of Ensemble Members

Clustering is a statistical technique that groups similar data together. It is used in this approach to reduce the dimensions of the data from 18 members to anywhere from 2-6 clusters of members (or forecasts). As discussed in section 2.4, common clustering methods used for meteorological applications include hierarchical clustering and partitional clustering, namely K-means and its variations. One such variation is K-medoids.

K-medoids clustering begins by finding the difference between members via a chosen distance metric and populating a table with the values. The distance table only needs to be calculated once for a set of data, reducing computation time. K-medoids chooses a member as the centre point of a cluster, known as a medoid. It proceeds through a series

of permutations of different combinations of k medoids and determines which group of members to medoids results in the least sum of squared distances seen in equation 3.6, where *SDist* is the sum distance of all intra-cluster sum distances *SICD*, where i is the index for the number of members to the total number of members n in the cluster, and $d_{x_i,m}$ is the distance between a member x_i and the medoid m.

$$SICD_m = \sum_{i=1}^{n}(d_{x_i,m}) \tag{3.5}$$

$$SDist = \sum_{i=1}^{k}(SICD_m) \tag{3.6}$$

Before deciding to use K-medoids, some basic tests were run to compare K-medoids and K-means to determine if they were comparable or one was clearly better at determining the optimal number of clusters than the other. K-means (analysed within this project using the *Scikit-learn* package in python by (Pedregosa et al., 2011)) aims to reduce a data set to a user specified number of clusters k and has a useful metric (inertia) by which to establish the optimal number of clusters. It begins with a set of k randomized data points (centroids) and the members of the cluster are grouped by similarity. The dispersion is typically measured by the mean-square difference from the centroid, which is the mean value of the members in a cluster. The sum over all clusters of the within-cluster dispersion is minimised for optimal clustering. To determine the smallest number of distinct clusters for a data set, the inertia is calculated for each possible number of clusters, summed, and normalized. The inertia is defined by:

$$I(N) = \sum_{j=1}^{N}\sum_{i\in j}(|\boldsymbol{x_i} - \boldsymbol{\mu_j}|^2) \tag{3.7}$$

where the summation runs over the cluster index, j, from 1 to N, summed over the number of members x (the field of data associated within the forecast ensemble) from i members within a cluster j. μ is the cluster centroid, e.g. the cluster mean of the j cluster members. The inertia is normalised by *I(N=1)* so that it is expected to decrease from one towards zero as the number of clusters increases from 1 to N. The total inertia is highest when there is only one cluster and lowest when each member is its own cluster. We expect the inertia to decrease monotonically as the number of clusters increases. For

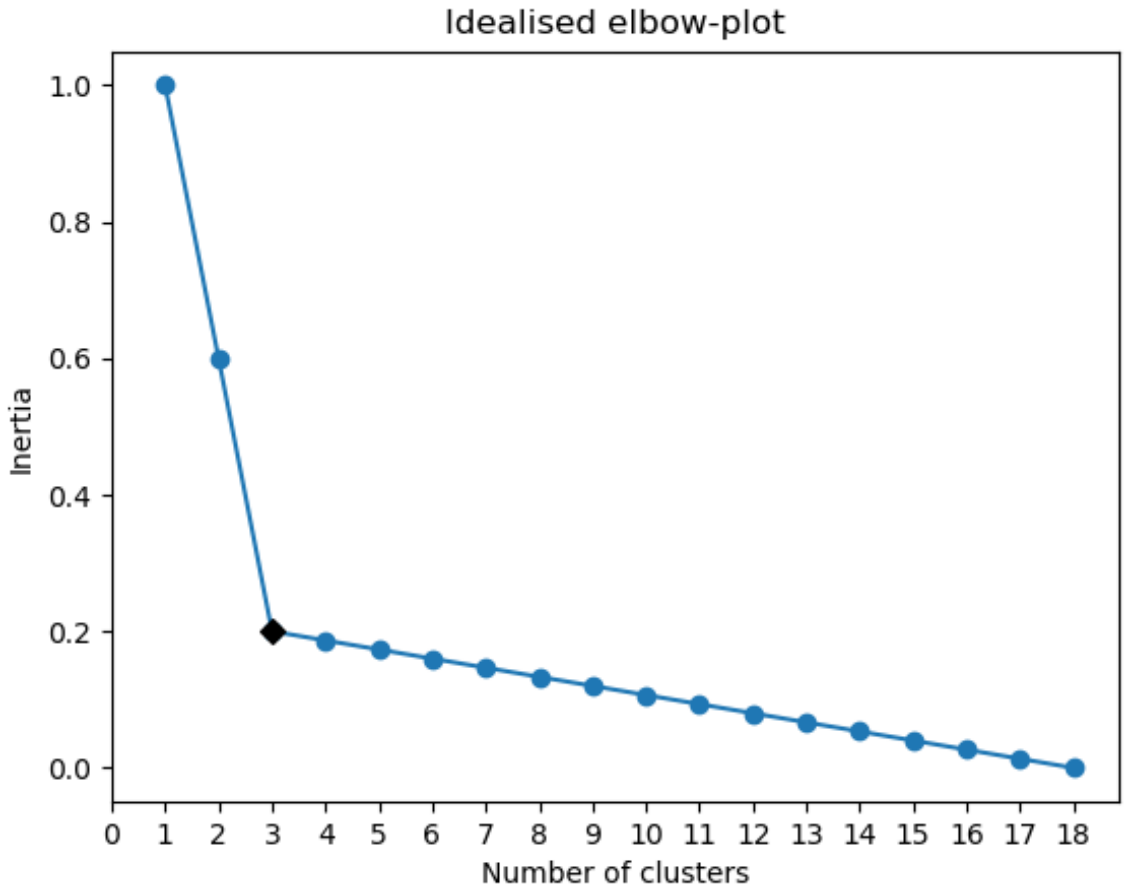


Figure 3.2: An idealized plot of normalized inertia versus the number of clusters, called an "elbow plot". The black diamond represents the bending point, indicating the ideal number of clusters for this data set is three.

any cluster with only one member, the distance between the member and the centroid is 0, therefore as we increase the number of clusters to equal the number of members, the inertia must tend towards 0. Determining the best number of clusters for the data set is then a trade-off between the reduction of data to fewer clusters and the variability between clusters. To visualize this problem, an idealized plot can be seen in figure 3.2. This plot is called an "elbow plot" and it can be used as a visual guide when choosing the correct number of clusters. It is generally accepted that the bend in the elbow, where the slope of the decrease in inertia has begun to level out or reduced greatly, is the best solution to the trade-off, where the fewest clusters can describe the most variability. In this example, the bend of the elbow is at 3 clusters, where the black diamond is located.

The elbow plot was used as a quick way to compare K-means and K-medoids to determine if there is a clear optimal number of clusters appearing from either method. The comparison can be seen in figure 3.3, where the θ_w and $|\nabla\theta_w|$ fields from the 02/10/2018 0000 UTC forecast (including all members) at time t+93 hours were clustered with both methods. As the sum distance and inertia are two different measures, they have each

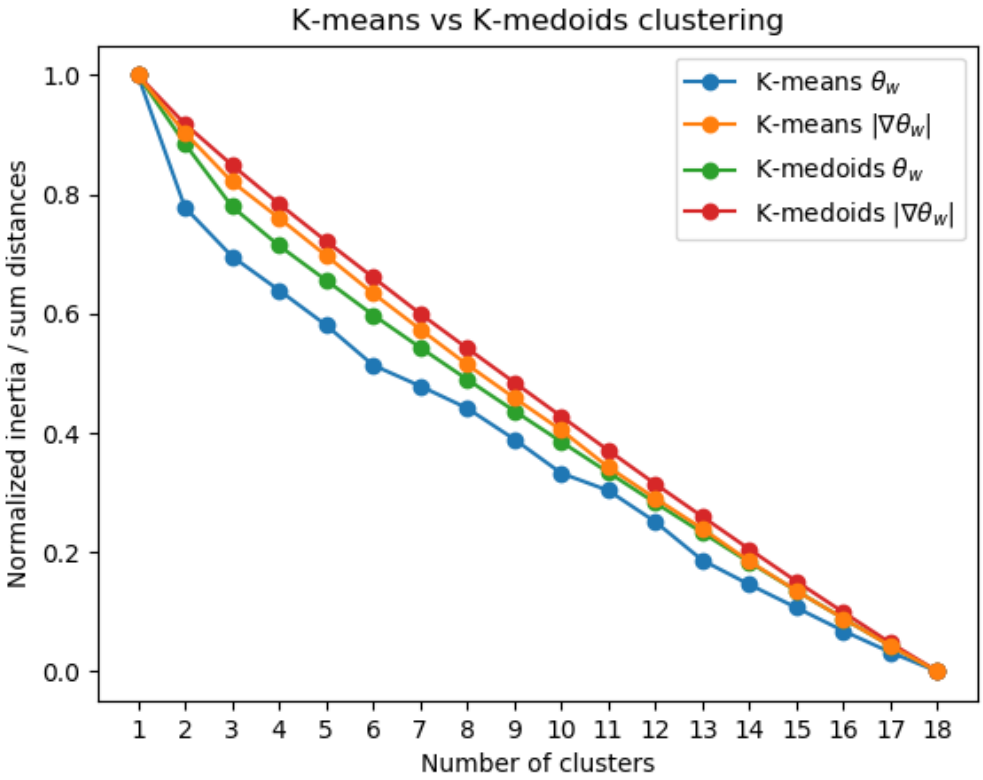


Figure 3.3: A comparison of inertia vs number of clusters (elbow plot) for K-means and K-medoids performed on θ_w and $|\nabla\theta_w|$, at t+93 hours from the 02/10/2018 0000 UTC forecast.

been normalized:

$$D' = \frac{D - D_{min}}{D_{max} - D_{min}} \tag{3.8}$$

where D' is the normalized data set, D_{min} is the minimum of the data set, and D_{max} is the maximum of the data set. At first glance, there appears to be little benefit for K-medoids over K-means as neither has an elbow indicating a clear optimal number of clusters, a common problem among large complex data sets. Some possible choices for optimal numbers of clusters are 2 clusters for K-means θ_w and 3 clusters for K-medoids θ_w, where there is a larger drop in inertia and the sum difference, respectively, with the increase in cluster number than any larger number of clusters and a slight but noticeable bend in the line. However, the locations of the bends in the K-means and K-medoids $|\nabla\theta_w|$ curves are not evident. Without a clearly better performing clustering algorithm with regards to the elbow plot, a different way to pick the preferred algorithm and to determine the optimal number of clusters must be chosen.

The elbow plot is only one tool to compare clustering methods and is not the only diagnostic used in clustering. Furthermore, it does not diminish the other benefits of using

K-medoids. Using K-means results in a mean of members as a representation of a cluster, however this is a limitation in its application to forecast problems. The ensemble mean field of a meteorological variable is not in itself a solution of the governing equations of the atmosphere, or by extension the forecast model. Using $|\nabla\theta_w|$, which is a rather noisy field but with small areas of sharp gradients, results in the mean being less useful. The mean smooths out the fields in question, losing the integrity of significant and fine scale features, thus causing issue with the relationships between variables. Therefore K-means is less useful for dealing with noisy fields and spatial differences, such as with small-scale spikes in a field that are not spatially predictable, where each forecast predicts a spike in a different place or does not include a spike at all.

Variable fields may also be better represented by non-traditional methods for comparison, such as the Mean Square distance. K-means requires calculation of the distance from the centroid, the mean of the cluster, which limits its use to distance metrics for which distance from the mean has a clear meaning. When two fields have similar fine scale features but each of these features is spatially displaced from each other by a distance greater than their width, both fields will have a high MS difference. These distances are then counted twice, which is the double penalty problem. It is therefore important to also consider using a distance metric other than the MS difference to account for such variability. K-means always uses the MS difference to determine how close members are, where flexibility in the distance metric is desired for this analysis. K-medoids, alternatively, does not require the mean or traditional distance metrics to be used in the clustering process.

As mentioned above, K-medoids uses one of the members of the cluster as the centre point, which removes the issues previously mentioned concerning K-means solutions and provides a single member solution per cluster, resulting in fields that keep their spatial distribution, fine-scale, and intensity intact. Therefore, K-medoids allows the user to input different distance metrics for clustering, such as a distance derived from the FSS, creating a distance matrix based on the metric of choice.

3.3.3 Analysing clusters and their evolution through time

Clustering of the data is the prime aim of the algorithm, therefore it is important to determine appropriate methods of cluster comparison. Given two sets of clusters, e.g.

clusters calculated at different lead times within the same forecast, there must be a simple yet informative measure of agreement between the assignment of members to clusters. Since the cluster labels themselves are arbitrary this must be based on maximising the agreement between the membership of clusters when all permutations of cluster labels are considered. To do this, the intersection between two sets of clusters is considered and a contingency table of matching members can be utilized. This, along with the Jaccard Index, which is a measure of similarity, is useful in exploring how cohesive clusters remain over time, i.e. their traceability.

Although it is possible to cluster the entire forecasts instead of per lead time, there is a distinct disadvantage to it that this method remedies. Clustering full forecasts reduces the variability inherent in the evolution of members through lead time, and it does not allow for narrowing the analysis of the clustering to any particular time period. By clustering at each time step, this method avoids over simplifying the data and can then use various techniques to determine representations of that data for scenarios.

3.3.3.1 Cluster comparison

During the clustering process, the label assigned to the cluster is arbitrary. To examine clusters at different lead times within a forecast, the clusters must be matched via membership comparisons. First, two sets of clusters are chosen for comparison, in this case two separate time steps (lead time t+93 in table 3.4a and t+111 in table 3.4b) from the same forecast start time (0000 UTC 02/10/2018). Each cluster at a given time step is assigned an arbitrary cluster label from 0 to 3. Then a 2D matrix (3.4c) is populated with the number of members that match between the clusters. For example, cluster 0 from t+93 has 6 members (where the first column in the table represents the number of members in the clusters from t+93) and cluster 0 from t+111 has 8 members (where the first row in the table represents the number of members in the clusters from t+111), and the number of members that match between them is 5. Next, the columns are held fixed and the rows of the table are reorganized by number of members in descending order, seen in table 3.4d, so that the largest (smallest) cluster is re-labeled 0 (3). Next the rows are held fixed and the columns are rearranged until the sum along the diagonal is the largest number possible, seen in table 3.4e, where the sum along the diagonal is 12. The final table (3.4f) is populated with the corresponding Jaccard Indices between the two clusters.

The Jaccard Index (Jaccard, 1912), or Jaccard Similarity Coefficient, is an equation in set theory that determines the similarity between two sets. In the following equation,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.9}$$

J is the Jaccard Index and A and B are two sets of numbers, in this case the members within two different clusters. This produces a ratio, where a perfect match would be 1 and a perfect mismatch would be 0. The Jaccard Index is a useful statistic to determine how well clusters match.

A visual representation of the contingency tables 3.4e and 3.4f, known as a cluster inter-comparison diagram, can be seen in figure 3.5, where the clusters at t+93 are compared to themselves in 3.5a and to t+111 in 3.5b. The x and y axes are are labeled with the cluster label and the number of members within the cluster inside the parentheses. The numbers within the boxes are the total intersecting members within the two clusters, and the colour bar is the Jaccard Index. The sum of intersecting members along the diagonal is written just outside the top left corner of the plot. If the sum is high, it indicates a strong correlation between the two time steps, indicating the clusters are remaining relatively intact through time. In figure 3.5a the sum is 18, which is expected as the clusters at t+93 are compared to themselves. In figure 3.5b, the sum is 12, where the clusters at t+93 are compared to the clusters at t+111. There is still a high degree of matching, as can be seen both by the colour of the squares (Jaccard Index) and the sum along the diagonal. Although some members of the clusters have moved to new clusters, the clusters from t+93 have remained relatively intact over time.

3.3.3.2 Traceability through time

The calculations in figure 3.4 and the corresponding diagrams in figure 3.5 demonstrate a quality called the traceability of a cluster. Traceability of clusters through time is critically important in determining respective scenarios within the ensemble. By choosing clusters at a single time step t_n then comparing previous t_{n-i} or future t_{n+i} clusters at different time steps to it, overall traceability can be determined. An example can be seen within figure 3.6. In the top plot, cluster members are coloured by the cluster they fall in most often (0: fuchsia, 1: gold, 2: chartreuse, 3: cyan, and 4: violet, 5: brown for 5 and 6

Labels	$Clusters_{t+93}$
0	0, 2, 7, 10, 11, 16
1	1, 5, 6
2	3, 4, 14
3	8, 9, 12, 13, 15, 17

(a)

Labels	$Clusters_{t+111}$
0	0, 5, 6, 7, 10, 11, 13, 16
1	1
2	4, 9, 14
3	2, 3, 8, 12, 15, 17

(b)

No. of mems	8	1	3	6
6	5	0	0	1
3	2	1	0	0
3	0	0	2	1
6	1	0	1	4

(c)

No. of mems	8	1	3	6
6	5	0	0	1
6	1	0	1	4
3	2	1	0	0
3	0	0	2	1

(d)

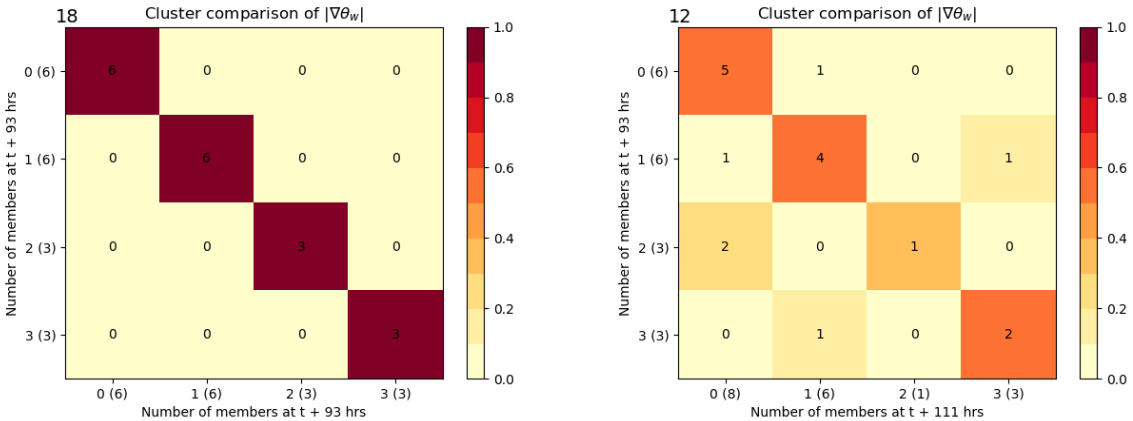
No. of mems	8	6	1	3
6	5	1	0	0
6	1	4	0	1
3	2	0	1	0
3	0	1	0	2

(e)

No. of mems	8	6	1	3
6	0.56	0.09	0	0
6	0.07	0.5	0	0.13
3	0.22	0	0.33	0
3	0	0.13	0	0.5

(f)

Figure 3.4: Tables and contingency tables of the cluster matching process where (a) and (b) contain the clusters in question from t+93 and t+111, respectively, (c) contains the number of matching members with within each set, (d) is the contingency table reordered by rows, (e) is the contingency table reordered by columns, and (f) contains the related Jaccard Indices used as a visual reference in figure 3.5.



(a) (b)

Figure 3.5: An example of a comparison between two time steps within the same forecast (t+93 and t+111 from 0000 UTC 02/10/2018). The x and y axes are the cluster number and the number of members within the cluster in parenthesis at the time steps *t*. The numbers within the squares are how many members between the x and y cluster match. The number at the bottom right is the sum along the diagonal. The colour bar is the Jaccard Index. Panel (b) is a visual representation of figure 3.4e where the color scale is based on the Jaccard Index in 3.4f.

clusters, respectively) over a time window of interest, which is marked by vertical dashed black lines and is further explained in section 3.3.4.1. The medoids of the cluster at any given time step are marked with black circles and the representative member is marked with a dotted blue line and is further explained in section 3.3.4.2. In the bottom plot, the normalized sum distance is in black and the ensemble spread is marked in a dashed blue line and calculated by:

$$Spread = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} DM_{D_{FSS}}}{n * (n-1)} \tag{3.10}$$

where *Spread* is the spread, $DM_{D_{FSS}}$ is the distance matrix of D_{FSS} values (calculated between members via equation 3.4) and n is the length of one side of the distance matrix.

Within the traceability plot, we can see how membership within clusters evolves over time. During early lead times, the majority of the members are within cluster 0, while every other cluster has a single member due to the number of clusters k being forced to remain the same throughout the forecast. The majority of members are in cluster 0 due to there being little difference between the various members and the control as they've all begun with similar and slightly perturbed initial conditions and there hasn't been enough time for error propagation to alter a forecast's evolution. These perturbations are described in Bowler et al. (2008) and Bishop et al. (2001) and involve the use of an ensemble transform Kalman filter, where by design the members will be distributed in a Gaussian way. In the early stages of a forecast, we expect most members to fall within the single Gaussian peak, or in this case, a single cluster. As time progresses and perturbations grow, members move further away from the control and naturally spread into other clusters. This typically takes two days or longer, and is typically when the window of interest begins. This results in cluster membership that appears predominantly stable for a period of time. When this happens, the process for selecting scenarios out of an ensemble of forecasts begins.

3.3.4 Selecting scenarios

As cluster members evolve over time and move further away from the initial conditions, different scenarios have the chance to form. Scenarios occur when clear clusters begin to appear. In figure 3.6 membership of cluster 0 begins to disperse around t+60. Around

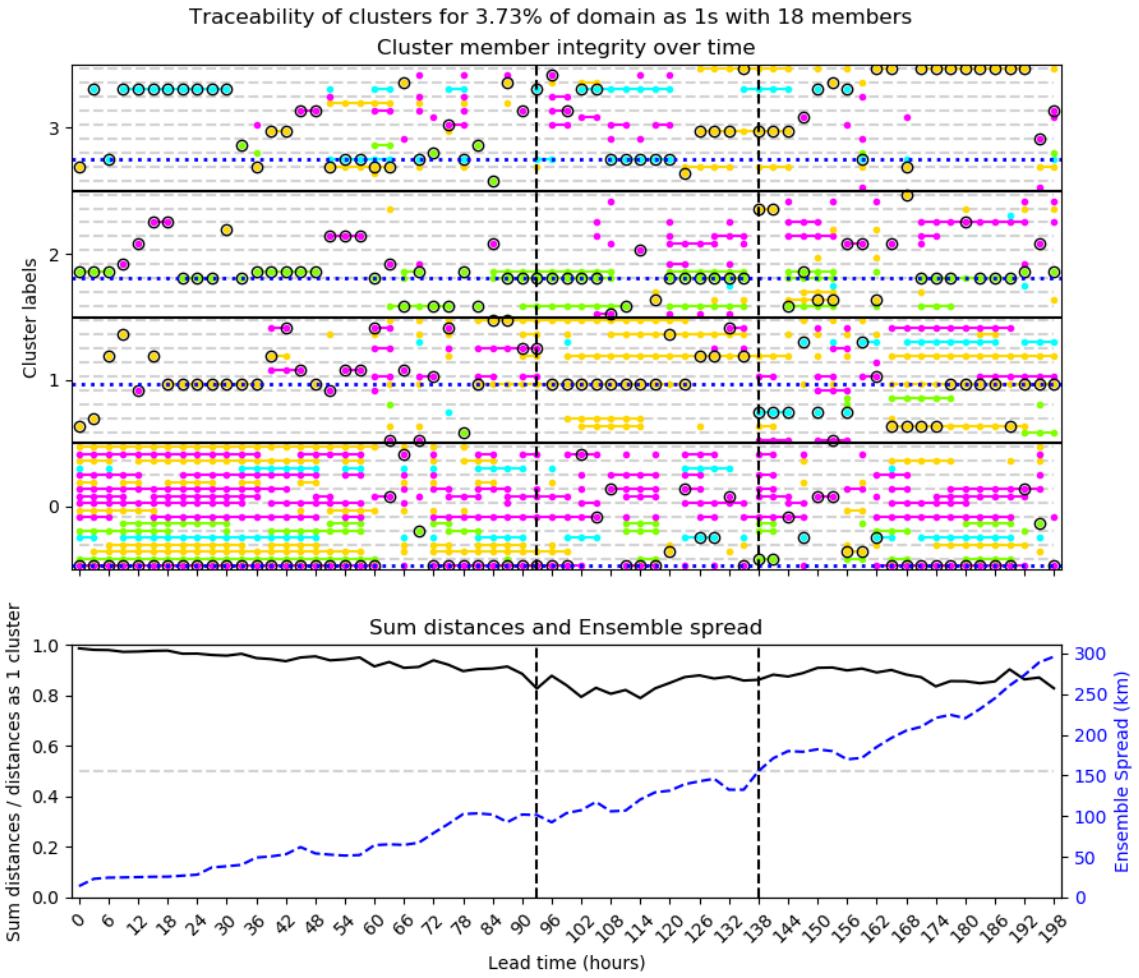


Figure 3.6: The traceability diagram of the clusters (top), and the sum distances and the ensemble spread (bottom) versus the lead time in hours from the forecast on 02/10/2018 0000 UTC. The colors in the top plot represent the cluster: fuchsia for 0, gold for 1, chartreuse for 2, cyan for 3, violet for 4. The vertical black dashed lines indicate the time window in which analysis is performed to produce a representative member, which are the blue dotted lines within each cluster in the top plot. The black circles represent the medoid of the cluster at a given lead time. In the bottom plot, the black line indicates the sum distances across all clusters, normalized by the sum distance of one cluster, and the blue dashed line is the ensemble spread.

t+93 it begins to stabilize with the same members staying in the cluster for several lead times, indicating distinct scenarios are beginning to form. To craft scenarios, the window of interest must first be determined and then a representative member extracted from the respective clusters.

3.3.4.1 Window of interest

The window of interest is a segment of time within the ensemble forecast where distinct clusters begin to emerge. The window is set to a 48 hour period and begins when the sum distance reduces to the 25^{th} percentile. When the sum distance has reached this point it indicates relatively distinct clusters are beginning to emerge as the members are closer to the medoid than they were previously. This also allows for a buffer of time around the most distinct clustering point so that evolution both into and out of this particular lead time can be observed. For the development of this project, the 25^{th} percentile was sufficient, however other percentiles were not tested at the time. This percentile may not always be beneficial, depending on when it occurs during a forecast. It is recommended that the sensitivity of this percentile is explored in the future and refined as the method is further utilised and explored. Once the significant window has been established, a representative member can be obtained.

3.3.4.2 Representative member

To reduce 18 forecasts to only a few scenarios, a representative member (RM) is chosen for each cluster. An RM is needed so the forecast can be seen evolving over time with physical consistency between variables, not a mean which would wash out information more and more as time progressed within the forecast. The RM should be similar to other forecasts within the cluster and either be, or be relatively close to, the centre of the cluster through time. This ensures the RM is a representative scenario. To determine the RM, a process of Least Sum is used.

The process of calculating the Least Sum begins with calculating the sum of distances between all members and the medoids during the window of interest, i.e. the distance of a member to the medoid at each lead time in the window is summed together to give a total distance during the window. If the member itself is the medoid at that time step, then the distance would be 0. The member that has the smallest distance, i.e. the Least

Sum, is then chosen as the RM of the cluster as it is the closest to the centre over the entire window. With this method, the RM may never be a medoid if the cluster medoid changes often during the window, but will still be a good representation of the cluster overall. Members may also move in and out of the cluster, but the RM will most likely remain within the cluster the longest. The longer the RM is part of a cluster, the more coherence the cluster has. Cases may occur where the member chosen as the RM appears infrequently within the cluster itself, though this is limited to a quarter of the window of interest except in extremely rare cases. The determination to allow the RM member to only appear a quarter of the time in the window of interest was made due to the likelihood restricting it further (e.g. where it must be in the cluster at least half of the window or longer) may unnecessarily remove potential scenarios from consideration.

Other possible processes of choosing the RM include simply choosing the ensemble member occurring most often as the cluster medoid or choosing the most commonly occurring member in the cluster. Choosing the member that appears most frequently as a medoid within a given cluster, either over the entire forecast or over a selected time window, as the RM is tempting as it is, by definition, the centre of the cluster. However, while it may be the centre of a cluster for so many points, it may not remain the centre or even close to the centre, as events evolve. Choosing a member, regardless if it is ever a medoid, that is within the cluster the longest for the RM is also tempting due to its high likelihood of similarity to other members in the cluster and low likelihood of switching clusters so it therefore has a high traceability. However, it may not be close enough to the centre to be representative of it. Therefore, the Least Sum is chosen as the RM to minimise the distance from the medoid over the window of interest, where by definition any other choice would on average be further away from the cluster center over the given window.

Within the traceability plot (figure 3.6 top), the RM, marked by a blue dotted line, is often a medoid within a cluster. For example, in all clusters, the RM is the most often occurring medoid within the cluster. Some RMs appear frequently in the cluster throughout the window, such as clusters 0, 1, and 2. In cluster 3 the RM predominantly appears in a single section of the window, but it still remains the member that was closest to the center throughout the window. Hence, by using the least sum, the RM is insured to be the member that's closet to the middle of the cluster, regardless of member longevity

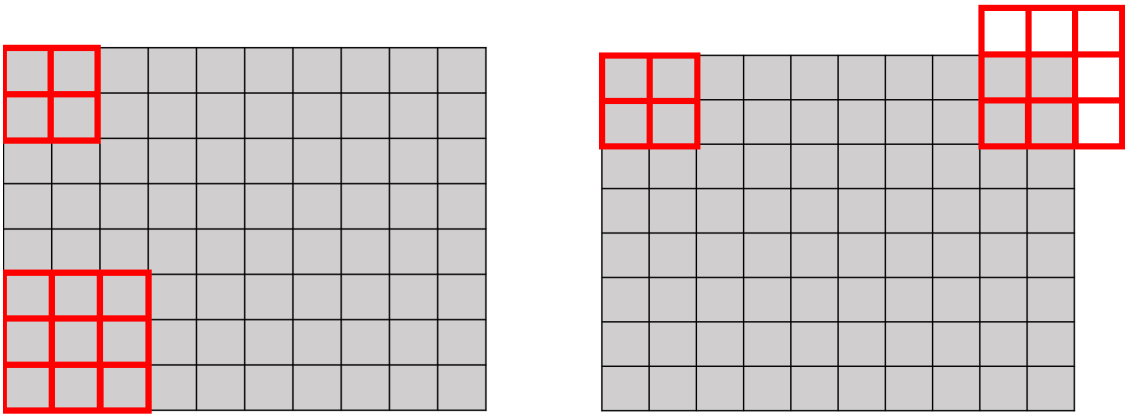
or medoid frequency.

3.4 Computational algorithm

The algorithm that performs the FSS calculations, the clustering, and traceability analysis is described in detail below in three stages, with corresponding flowcharts in figures 3.8, 3.9, and 3.11, respectively. For the purpose of this project, the algorithm was implemented in python.

3.4.1 Pre-processing

To begin, data first go through several pre-processing steps, seen in the beginning steps of the flowchart in figure 3.8. The θ_w field at 850 hPa is read from the forecast model data files and then a centred finite difference gradient is applied, followed by an application of a masking file to remove data below ground (i.e. mountainous regions above 850 hPa), which would otherwise introduce gradients that are not associated with frontal zones. To apply the FSS method, the data must first be converted into a binary field, which is done by applying a threshold. The threshold is calculated using the 96.27th percentile of data in the domain over the forecast, including all members within the ensemble throughout the forecast. Therefore, the 3.73% of data above or equal to the threshold will become 1 and the data below will become 0. This percentile was found via experimentation with the initial data (ensemble forecasts from 10/10/2018 0000 UTC). A threshold was needed that could extract the best frontal features in terms of producing distinct frontal objects and significantly reducing noise and fine scale features in the field. If the threshold is too low, the frontal features cannot be distinguished from the background gradient. If the threshold is too high, some of the frontal features will be missed. It was also necessary to determine a threshold that provided meaningful clustering results where visually similar large frontal objects were consistently grouped together in the same cluster. Using a percentile also means that on days when fronts are weaker they are still identifiable. However, this percentile has not been thoroughly tested for sensitivity or on warmer seasons. After the threshold has been applied, the data is now ready to calculate the FSS and begin clustering.



(a) (b)

Figure 3.7: A grid representation of the FSS neighbourhoods, where (a) represents how even n and odd n neighbourhoods (in red) compare, and (b) demonstrates how this method approaches the domain boundary (top left red grid) and how more traditional uses of the FSS approach the boundary (top right red grid).

3.4.2 Applying FSS to find a distance matrix

The details of creating the FSS matrix can be seen in the flowchart in figure 3.8. The field is converted to binary via the calculated threshold, so the data is converted to 1 at or above the threshold and 0 below. At t=0, the neighbourhood size n is initially set to 1, meaning it encompasses area of length and width both equal to 1, i.e. a single grid point. For this study, n is allowed to be both odd and even. For example, if n is equal to 2, the length and width of the neighbourhood are both equal to 2, encompassing 4 grid points total. If n is equal to 3, then the length and width of the neighbourhood is 3, encompassing a total of 9 grid points. This can be seen in figure 3.7 (a), where a neighbourhood of n=2 can be seen in the top left corner of the domain and a neighbourhood of n=3 can be seen in the bottom left corner. The FSS typically uses odd neighbourhood sizes, however this is done purely for the possibility of later calculating probabilities based on the center grid point. In this work, only the total neighbourhood size is relevant and not the centre point, as the fractions are calculated over the whole neighbourhood size and are only used for this purpose. Therefore, both even and odd neighbourhood sizes can be used. For faster calculation, simplicity in design, and due to the large size of the data (i.e. the domain of the forecast region), the neighborhoods are kept strictly within the boundary with no zero padding used. This can be seen in figure 3.7 (b), where the neighbourhood at the top left corresponds to what this study uses in terms of overlap with the domain boundary, and the neighbourhood at the top right is similar to what typical FSS uses,

where neighbourhood grid points that fall outside of the domain are padded with zeros. Regardless of the neighbourhood size, the neighbourhood is moved one grid point at a time and is recalculated to produce the fractions without crossing any boundaries. This reduces the array size of the calculated fractions by n on each side, i.e. if the size of the data is:

$$Data_{x,y} \tag{3.11}$$

where *Data* is an array of data and x and y are the lengths of the sides of the array, then the size of the resulting array of fraction scores is:

$$Fractions_{x-n,y-n} \tag{3.12}$$

where *Fractions* is an array of FSS fractions and n is the neighborhood size. This does mean that values along the domain boarder are used in the fractions calculations less than more central points in the domain, whereas the more common FSS fractions calculation would not have this issue. However, since the area of interest, i.e. the UK, is centered in the domain, this should not affect clustering of interest at this time. Although, a future study exploring which version of the FSS provides the best results or if any weighting should be added to correct it should be completed.

Next, the FSS is calculated between all members using equation 3.1 and a matrix is populated with the values (the FSS matrix). The average FSS is calculated from the matrix then tested to determine if it is approximately 0.5. If it is not equal to or above 0.5, the neighborhood size is increased until the average FSS is greater than 0.5. This FSS and the previous FSS with a smaller neighbourhood size are compared and the result closest to 0.5 is carried forward. The final neighbourhood size is then used as a first guess for the next time step for calculating the FSS. For this and all future lead times, the neighbourhood size is increased or decreased until the FSS is closest to 0.5, set within the range of n=1 and its maximum, the smallest maximum domain length of the data. Once the average FSS closest to 0.5 and the corresponding neighbourhood size are determined, a distance matrix is calculated as per equation 3.4, comparing each member to every other member, and can now be used for clustering.

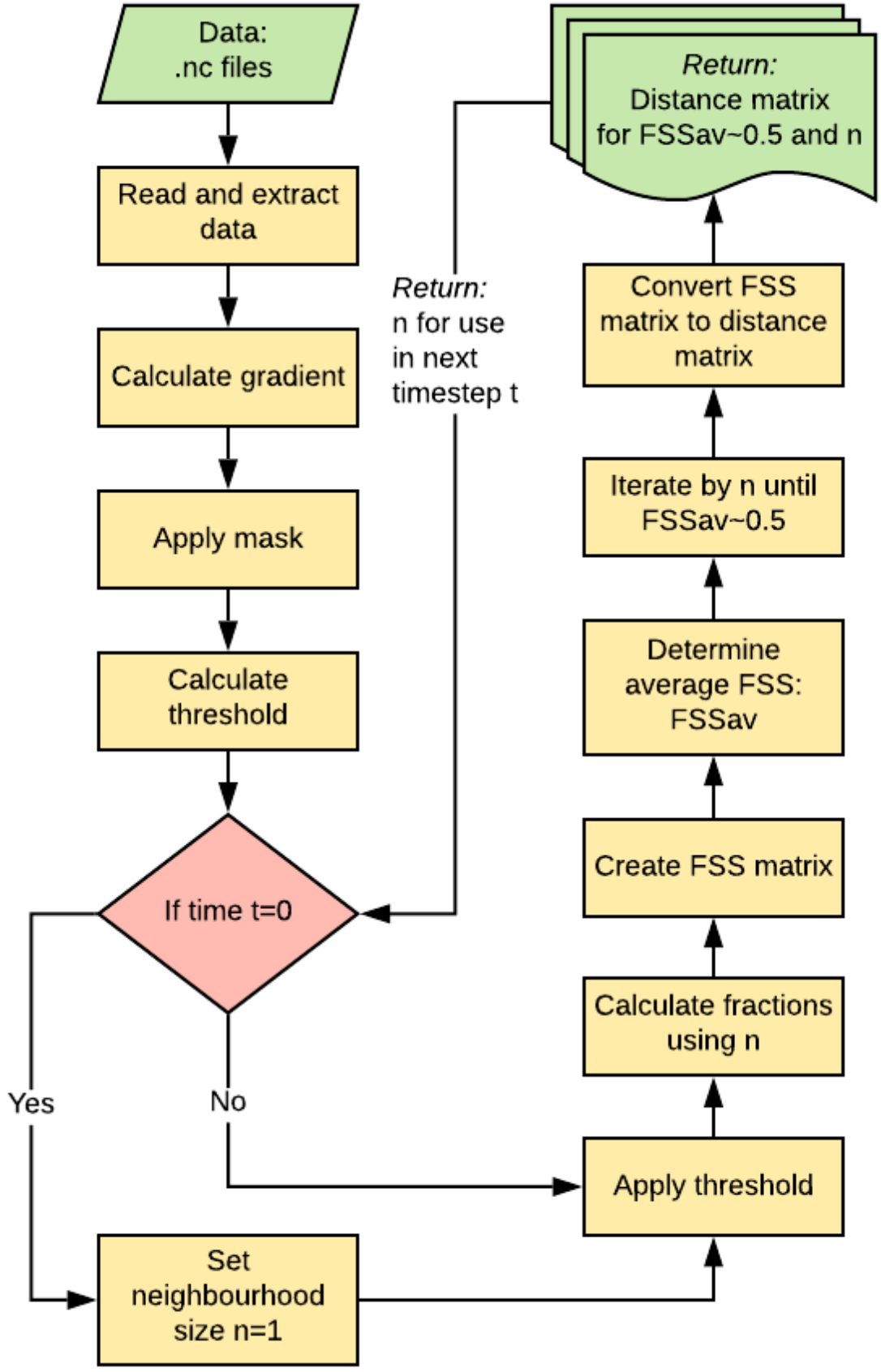


Figure 3.8: The pre-processing of the data and creating the FSS and distance matrices. (Created in Lucidchart, www.lucidchart.com)

3.4.3 Clustering of members

The distance matrices from the previous section are used to cluster the members into a set number of clusters k via K-medoids, seen in the flowchart in figure 3.9. The algorithm loops over the number of **K** clusters, from two to six, with the intention to determine the optimal number later. Within each k loop, there is a loop over time steps t, where each member is assigned a cluster label and corresponding cluster medoid. An example of the raw and resulting clustered distance matrices can be seen in figure 3.10. The original distance matrix (figure 3.10a) has the members along the x and y axes in numerical order from 0 to 17, with white squares along the diagonal where a member is compared with itself. After clustering results are reordered in a distance matrix as in figure 3.10b. The coloured boxes along the diagonal group clustered members with their medoids, the black circles, along the diagonal. In this instance of four clusters, cluster 0 is the fuchsia coloured box with member 0 as the medoid, cluster 1 is in gold with 13 as the medoid, 2 is chartreuse with member 5 as the medoid, and 3 is cyan with member 14 as the medoid. This plot provides a visual representation of the relative distances in km between members. For example, member 13, the medoid within the gold cluster, is closer to the members within its cluster than all other members. The cluster medoids and labels, along with the distance matrices are then fed into the next stage of the algorithm.

3.4.4 Determination of what number of clusters to use

Now that the data for all k numbers of clusters have been completed for all t time steps, the data must be examined to determine which set of **K** clusters provides the highest degree of traceability. This is broken down into three processes: membership, outliers, and unique RMs. These processes are described in detail below and can be seen in the flowchart in figure 3.11. The method begins with the smallest number of clusters (2) and increases the number when it passes a given criteria, or reduces it when it fails.

3.4.4.1 Membership criteria

The first step to determining the optimal number of clusters is to examine the membership of the cluster within the window of interest. The nature of clustering is to group similar members together, therefore, unless a member is an outlier (dealt with within the

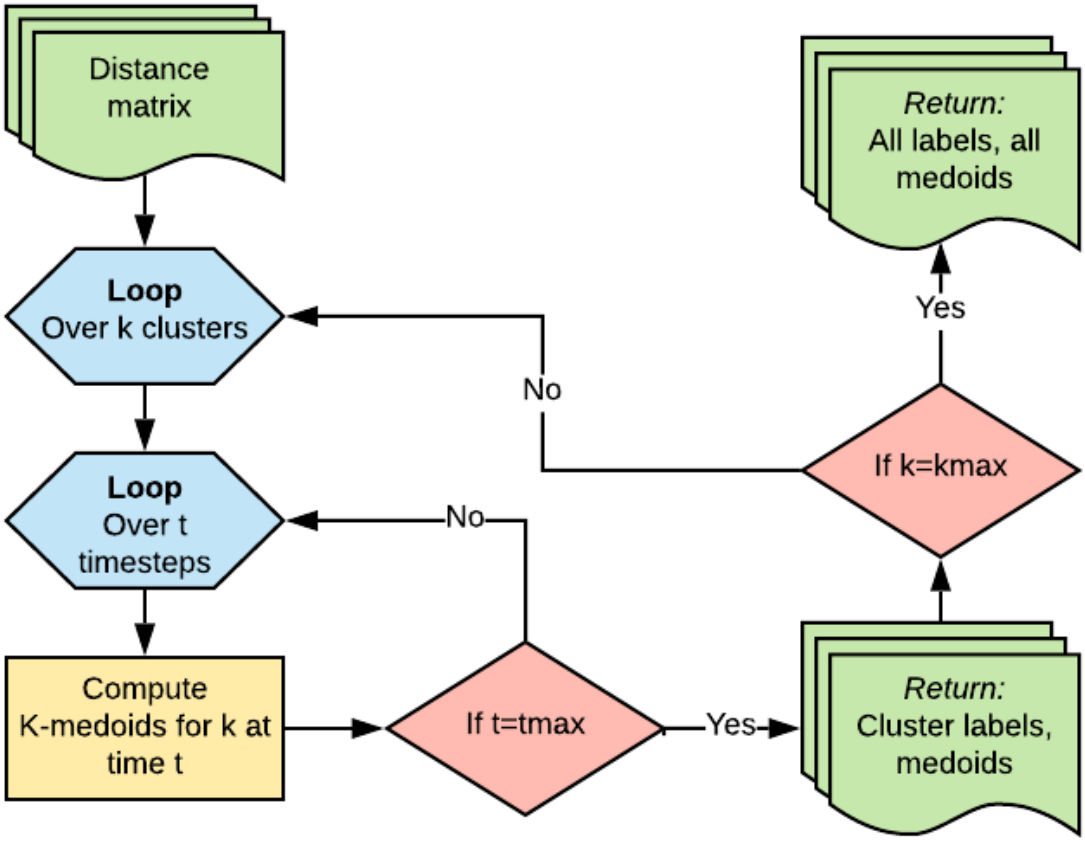
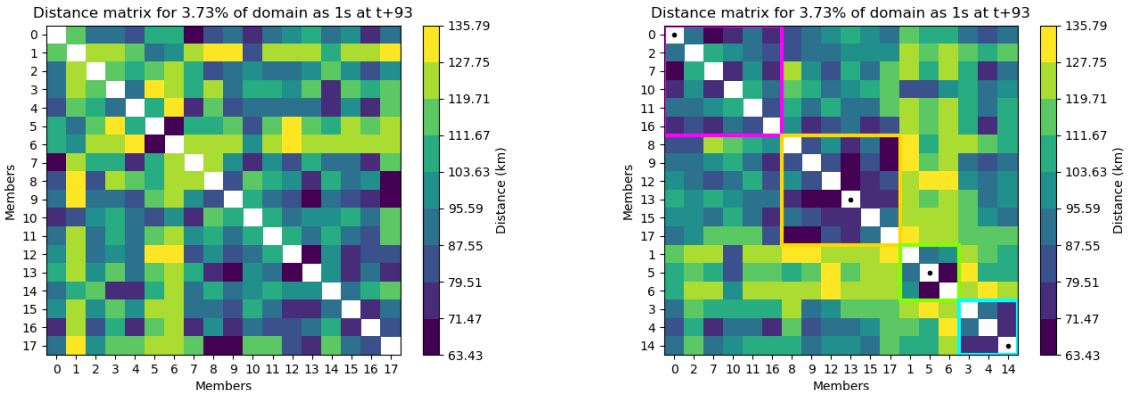


Figure 3.9: The clustering of the data via the distance matrix. (Created in Lucidchart, www.lucidchart.com)



(a) (b)

Figure 3.10: The raw distance matrix converted to km before clustering (a) and after clustering and re-ordering the members (b).

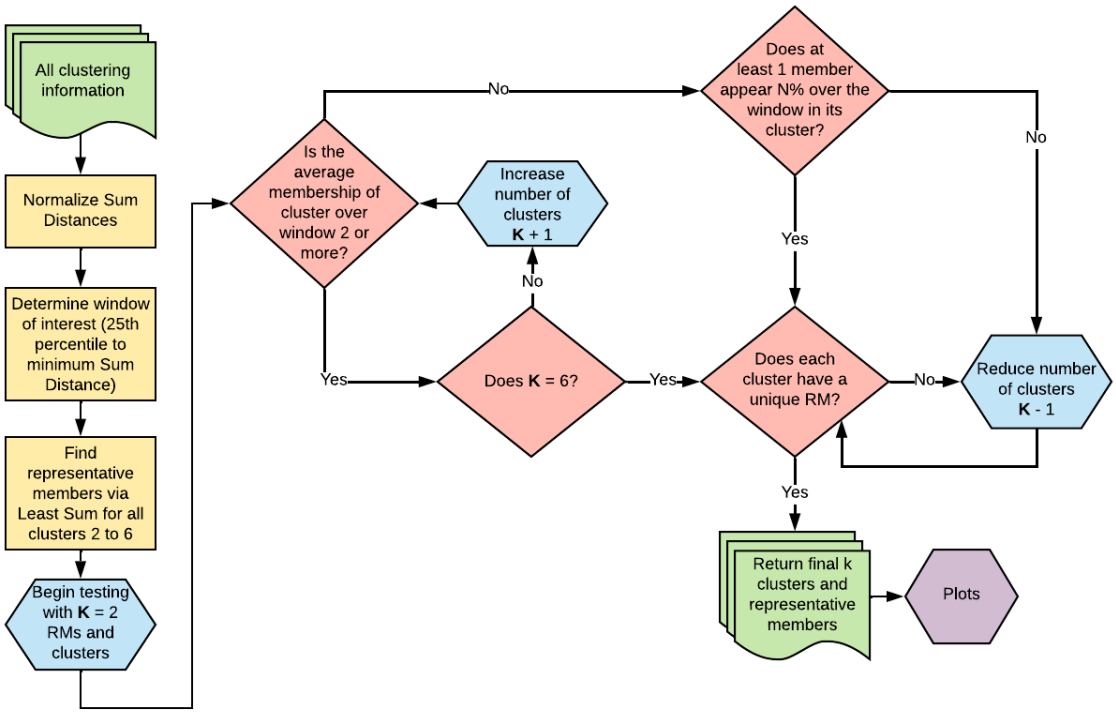


Figure 3.11: The analysis of clustered data to determine the representative member. (Created in Lucidchart, www.lucidchart.com)

next section), a cluster should have at least two members on average. The first step of the process is to examine each cluster within a set of **K** clusters to determine the average membership. This is considered a first sift of the data. If the membership is greater than or equal to two for every cluster with a set, the number of **K** clusters is increased. When a set of clusters fails this test, a second sift is applied to the clusters to check for outliers.

3.4.4.2 Outliers

When the membership of a cluster is below 2, there are two primary possibilities as to what is occurring. Either members are being forced into a unique cluster without being truly unique scenarios, which leads to low traceability as members jump in and out of other clusters, or there is a particularly strong outlying trajectory. To take outliers into account, any set of **K** clusters that produces an average membership over the time window of less than 2 is tested. If a member remains within the suspect cluster for 100% of the time within the window it will be considered a unique cluster and will move on to the final criteria check. If a single member does not remain with the suspect cluster for the desired time, the next fewer number of clusters will be passed to the final criteria.

3.4.4.3 Unique Representative Members

With the average membership and outliers analysed now the resulting RMs must be examined. Throughout the forecast, members can move between the clusters. It is therefore possible although unlikely that a member could move between two or more clusters and meet the criteria to be a representative member in more than one cluster. This would occur if the same member appeared within two clusters during the window of interest near 50% of the time and was frequently the medoid in both clusters. This indicates that these two clusters are much more likely to be similar than unique scenarios and **K** should therefore be reduced. Once every cluster has a unique RM, this solution is then fully processed and the resulting RMs are presented to the forecasters.

3.5 The Method Applied to a Forecast Encompassing Storm Callum

In the previous sections the method was described and portrayed by an ideal example forecast during the month of October. This was done to clearly show the different aspects of the method. In this section, a forecast that encompasses Storm Callum, the original event the method's creation began around, will be briefly discussed.

Figure 3.12 is the traceability plot for 1200 UTC on 07/10/2018. The window of interest begins at t+90 hours, which corresponds to 0600 UTC on 11/10/2018. During the window, the representative member in each cluster is the dominant medoid and remains in the cluster for most of the time period. This is significant in that it indicates these clusters may be well defined and more easily distinguished from one another. Figure 3.13 shows the representative members of the clusters at t+90 hours. These plots show the gradient of the wet-bulb potential temperature after it has had a threshold applied for use in the FSS. Each plot shows a very different position of the primary frontal object seen to the northwest of the UK. These different potential scenarios indicate there is significant uncertainty at this time when it comes to forecasting storm Callum. A further in-depth analysis of how the method performs and extract scenarios will be discussed in chapter 4.

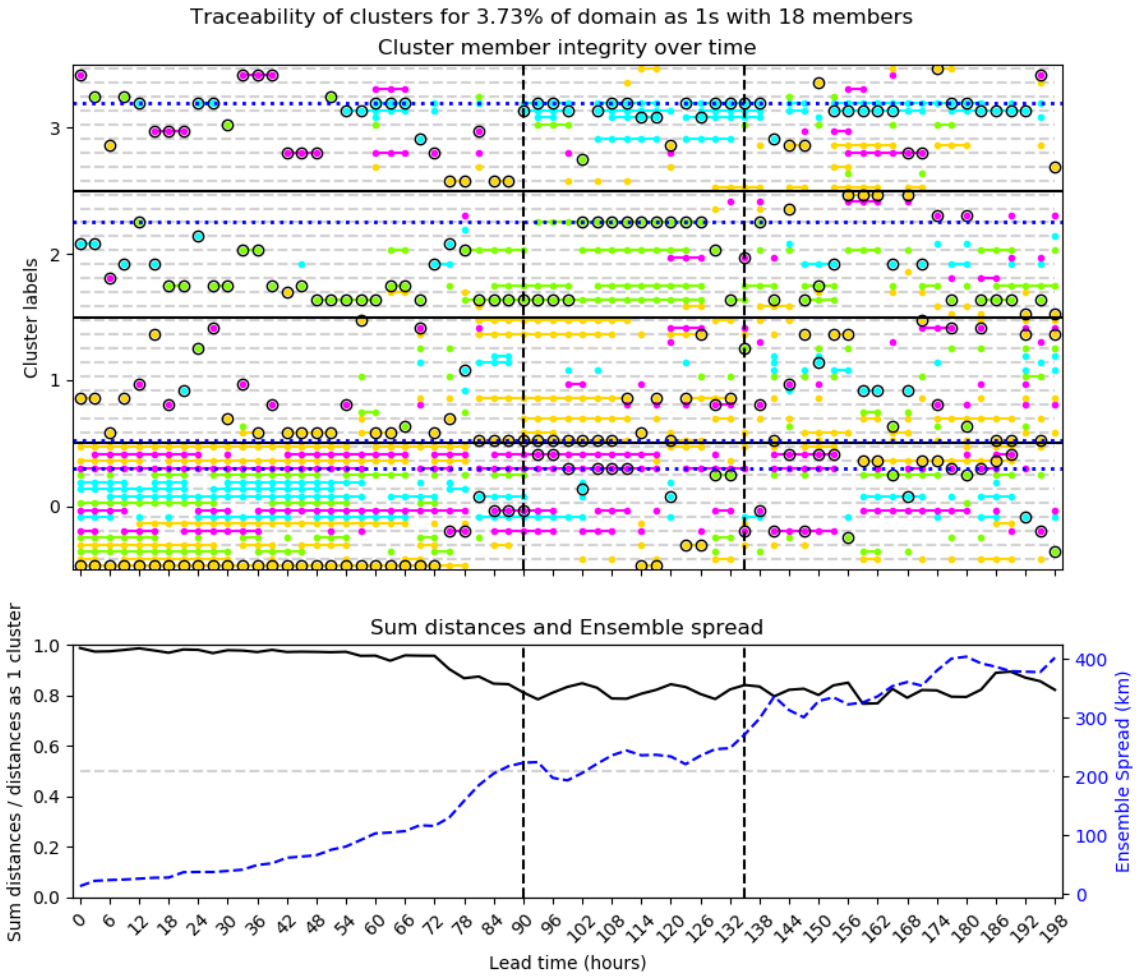


Figure 3.12: The traceability diagram of cluster membership (top), and the sum distances and the ensemble spread (bottom) versus the lead time in hours from the forecast on 07/10/2018 1200 UTC, over a domain of 40° to 70° north and 45° west to 45° east.

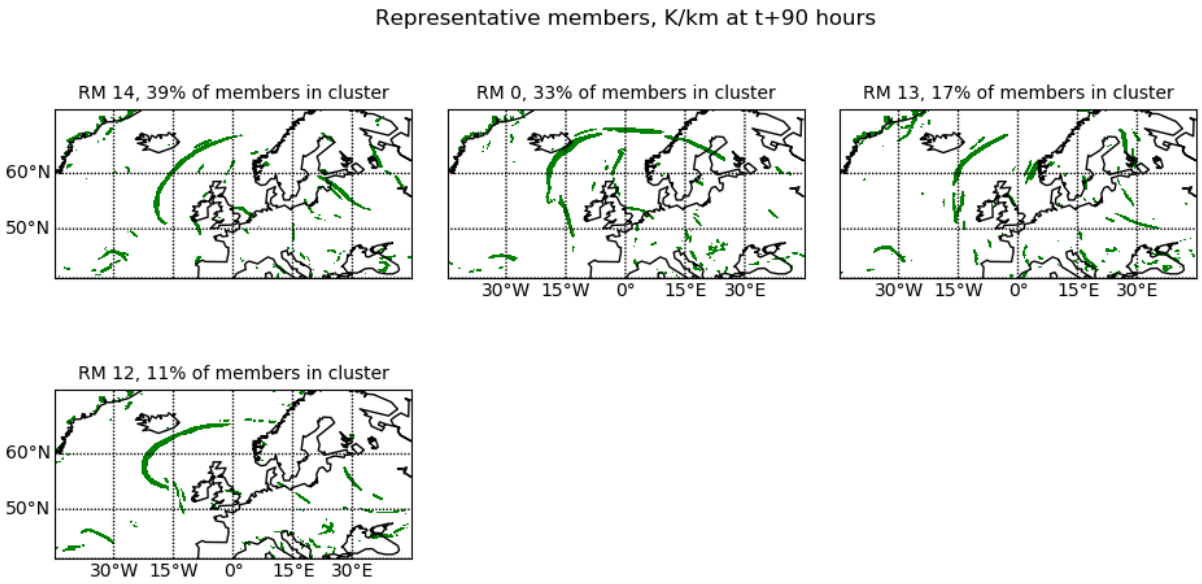


Figure 3.13: Representative member plots in θ_w at 850 hPa and frontal objects from $|\nabla\theta_w|$ at 850 hPa for the four cluster solution of the forecast on 07/10/2018 at 1200 UTC at a lead time of t+90 hours corresponding to the beginning of the time window of interest.

3.6 Conclusion

Within this chapter the novel clustering method was presented. It uses the K-medoids clustering algorithm with the FSS as the distance metric to compare ensemble clustering members. The variable used is the gradient of the wet-bulb potential temperature $|\nabla\theta_w|$. In the following chapters the method will be evaluated in different ways. In chapter 4, how the method works on $|\nabla\theta_w|$ for the months of October, November, and December of 2018 will be explored. Topics will include how the clustering compares to the ensemble spread, how the drop in sum distance relates to the beginning of the window of interest, visual methods of examining the clusters, how clustering compares across lead times as traceability, the variation of representative members, and finally extracting scenarios and predictability. In chapter 5, the dependence of the variable choice will be explored. It will include topics about how the clustering performs on a new variable, the large-scale rain rate, and how it compares to $|\nabla\theta_w|$, and if the membership, window of interest, and representative members are comparable across variables. In chapter 6 there will be a discussion of the method when it was used during the Met Office Winter Testbed and survey results from participants. Topics will include results from an in-depth case study, if the clusters represent distinct weather scenarios, how the products from the algorithm may influence forecasting communications during the testbed, if the method can detect high-impact scenarios, and how efficient the method is compared to crafting a forecast directly from the ensemble.

Chapter 4

Analysis of the method used for extracting scenarios from MOGREPS-G data over the Euro-Atlantic domain

4.1 Introduction

In this chapter, the performance of the ensemble clustering method introduced in chapter 3 is evaluated using several months of operational ensemble forecast data of the wet-bulb potential temperature θ_w at 850 hPa in a domain of 40° to 70° north and 45° west to 20° east. The gradient of the wet-bulb potential temperature is an excellent indicator of frontal regions (see section 3.2.2), which the method has been designed to use in order to determine different potential scenarios within an ensemble of forecasts. To support the use of this variable in the method, the clustering must be evaluated in depth. Within this chapter several topics will be considered. First, what the relationship is between the sum distance (equation 3.6) and the ensemble spread will be covered. Next, how the sum distance relates to the window of interest will be explored, followed by how members compare via distance measures and in terms of meteorology. How clusters are traced through a single forecast and how this traceability relates to the sum distance will then be covered. Next, the variation of representative members will be discussed. And finally, the scenario extraction and predictability will be explored.

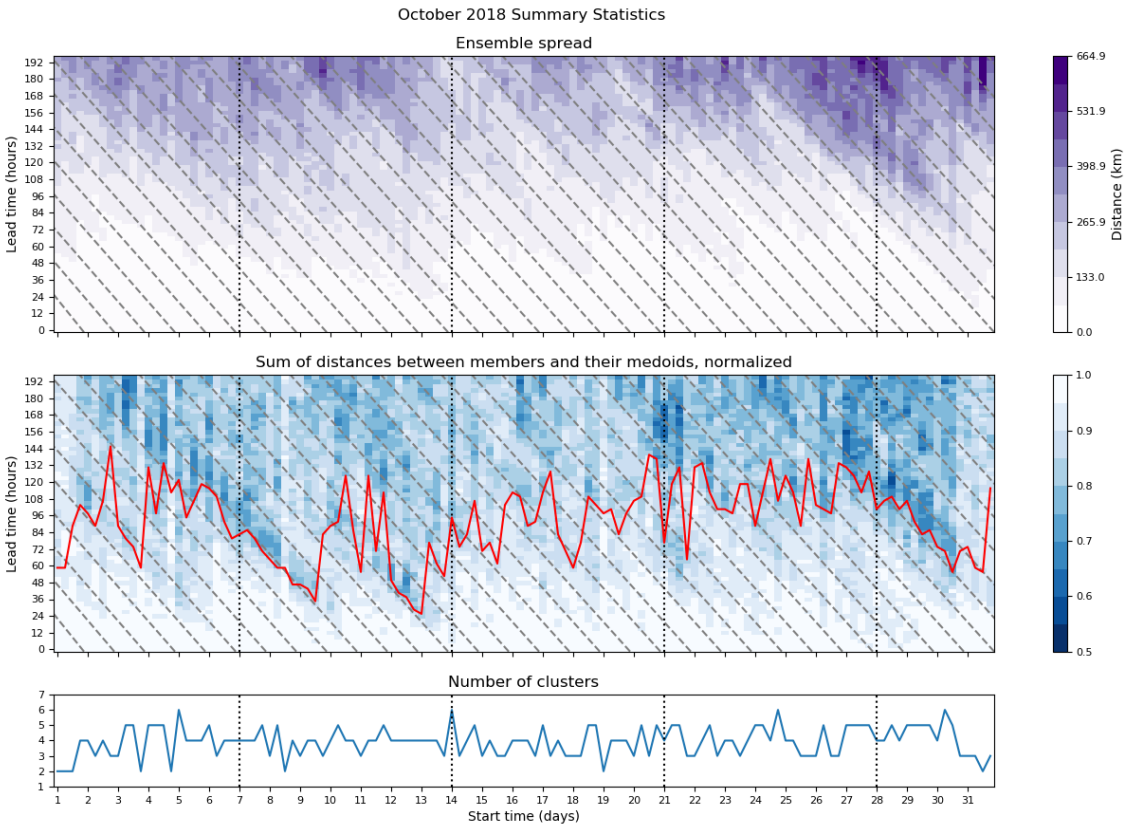


Figure 4.1: A "predictability plot" based on comparison of $|\nabla\theta_w|$ between ensemble members for the month of October 2018. The top plot is the spread. The middle plot is the sum of within cluster distances, normalized by the sum distances from the medoid of the whole ensemble. The red line marks where the window of interest (the point at which the sum distance has decreased to the 25$^{\text{th}}$ percentile) begins. The bottom plot is the optimum number of clusters chosen by the algorithm, denoted by the forecast start time on the x axis. The vertical black dotted lines mark every seven days. The diagonal dashed lines link the same verification times across forecasts.

4.2 Clustering of forecast data

To begin analysis of the clustering of $|\nabla\theta_w|$ it is important to look at trends of the sum distance, the primary means of comparing members and determining periods of time when clustering may become robust. Within this section how clustering relates to the ensemble spread and how the sum distance affects the window of interest will be discussed.

4.2.1 How clustering relates to the ensemble spread

Within an ensemble forecast each forecast member begins as nearly identical to the control member with nearly the same initial and boundary conditions and the ensemble spread is very low. As lead time progresses, forecast members begin to diverge from one another due to the chaotic nature of atmospheric dynamics, varying depending on the synoptic situation. This leads to a natural increase in spread over time, seen in figures 4.1

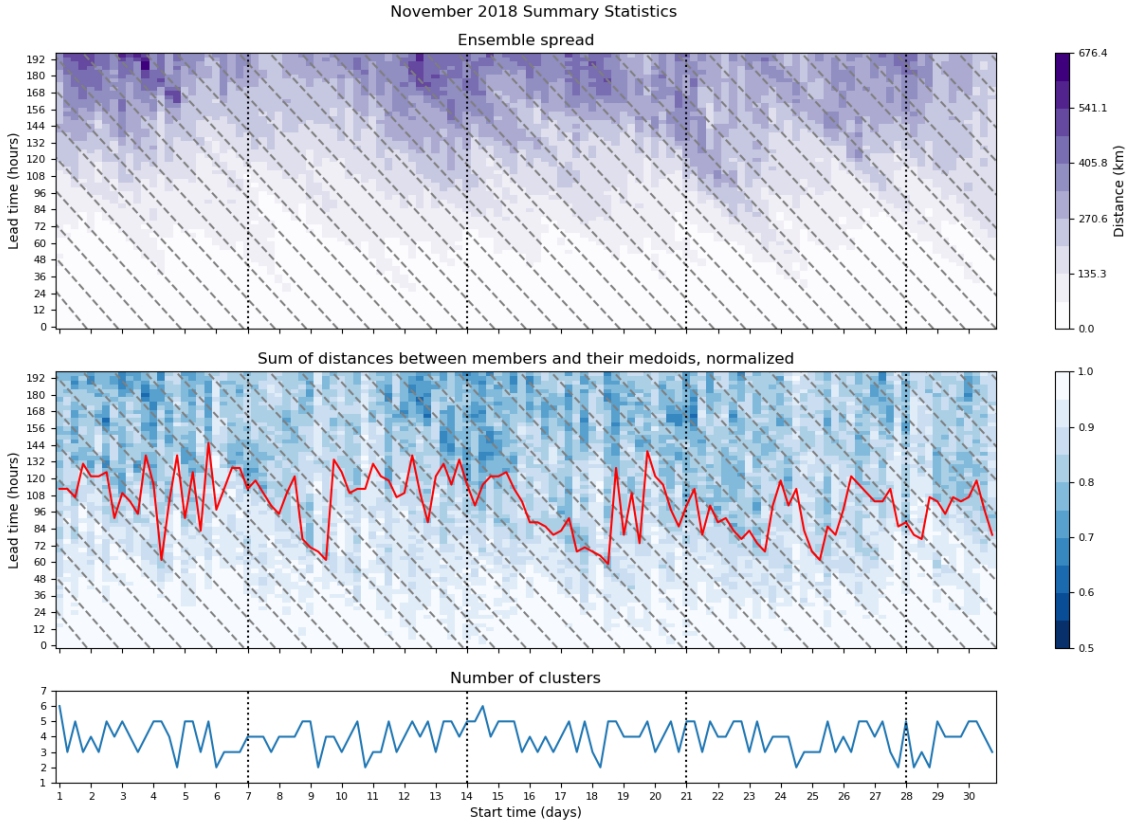


Figure 4.2: November 2018. Details can be found in figure 4.1.

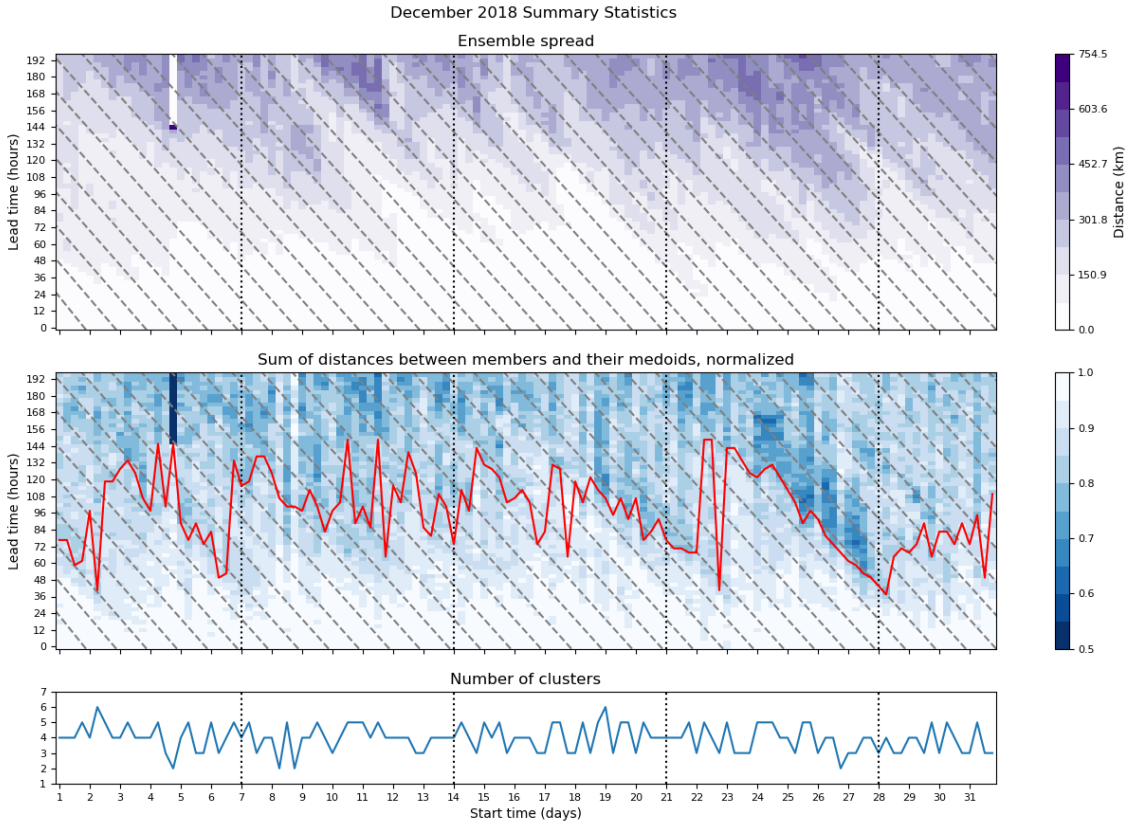


Figure 4.3: December 2018. Details can be found in figure 4.1.

to 4.3 in the top plot section. Each forecast increases in the lead time on the y-axis and the x-axis relates to different forecasts with a total of 4 per day at 0000, 0600, 1200, and 1800 UTC. Here, each ensemble forecast can be seen to start at a very low spread and increase as we get to the end of the forecast at t+198 hours. This is due the spatial separation between frontal regions on a synoptic time scale. However, there are some forecasts that increase in spread faster than others, notably seen towards the end of October (figure 4.1), between the 25$^{\text{th}}$ and 31$^{\text{st}}$, where a sudden increase in spread occurs towards the end of the forecasts. An increase in spread indicates an increase in uncertainty, and therefore a sharp increase may indicate the development of a particularly difficult to predict atmospheric situation, e.g. the intensity, location, and duration of frontal regions associated with a storm.

To better understand the variability and uncertainty in the atmosphere, statistical methods can be applied to determine if some members may follow similar trajectories and can therefore be grouped together. Applying the clustering process to the ensemble provides a useful metric to compare with the spread, as they are intrinsically linked via the distances between members. The sum distance (equation 3.6) is a calculation based on the distance between members and is a measure of how distinct the clusters are. In the October figure, the sharp increase in spread, associated with the average distance between ensemble members increasing corresponds with a sharp decrease in the sum distance, which is a measure normalised in such a way that a lower number indicates that the members can be grouped into more than one distinct cluster, seen in the middle plot. However, there are periods of time where the increase in spread is less sharp and more gradual but the quick decrease in sum distance is still apparent, such as between October 3$^{\text{rd}}$ and 8$^{\text{th}}$, and December 24$^{\text{th}}$ to 27$^{\text{th}}$. This may indicate some forecasts are more clustered than others. However, regardless of the intensity of the increase in spread or decrease in sum distance, the general pattern of these two metrics often co-vary, indicating they are related.

4.2.2 The sum distance and window of interest

An important question that must be considered is at what point in a forecast are clusters both sufficiently distinct from one another and they still maintain a strong simi-

larity amongst the cluster members. The red line across the sum distance plots indicate the beginning of the window of interest, defined in section 3.3.4.1. The method begins the window when there is a sufficient drop in sum distances, and although an increase in spread of the ensemble is often linked with a drop in sum distance, it is not always the case and cannot be relied on as a good indicator of when the window will begin. This is because the spread is a metric based just on the distances between members and the sum distance is related specifically to how distinct clusters are.

Figure 4.4 is a histogram of the sum distance values at the beginning of the windows of interest for the forecasts of October through December 2018. The window of interest begins when the sum distance has reached the 25^{th} percentile to its minimum value for the forecast. In the histogram it can be seen that the majority of windows begin with a sum distance of 0.75 to 0.87. The lower the sum distance the higher the chance for strong clustering, therefore significant drops in the sum distance, such as the 0.67 to 0.75 range, are more likely to signify strong clusters whereas small drops such as 0.87 to 0.91 are more likely to indicate weak clustering.

There is also the consideration of the relationship between the window start time, sum distance, and number of clusters. Figure 4.5 contains five box plots. Each box plot represents when the window of interested began most frequently for each number of clusters. It can be hypothesised that earlier lead times will tend towards smaller numbers of clusters due to forecasts converging towards a single solution. By examining these plots we can see this is indeed the case. There is a clear trend where the higher number of clusters have a window start time at later and later lead times. However, there are few occurrences of 2 and 6 clusters over the months of October to December, so to verify this trend is consistent throughout the number of clusters would require more data points for a more thorough analysis.

4.3 Visual representations of clustering

In the following subsections an in-depth examination of how the clustering performed at individual lead times will be discussed. This will include an overview of how the clustering can be analysed visually.

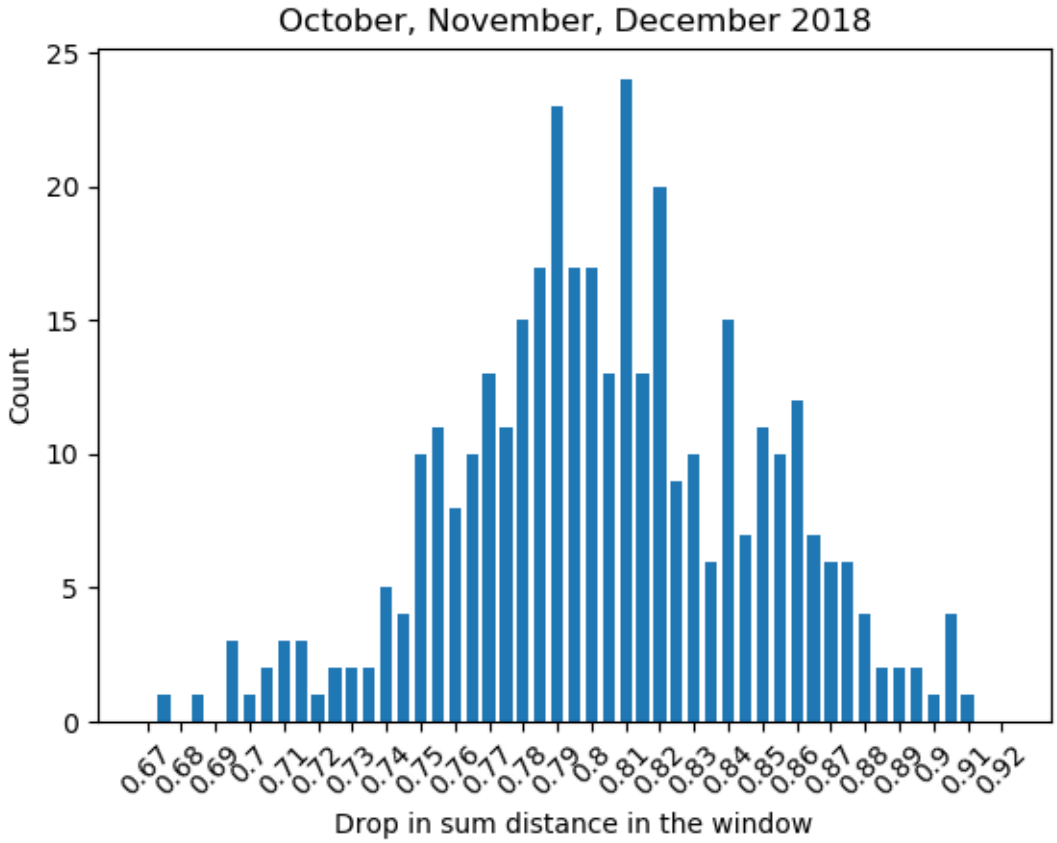


Figure 4.4: Histogram of the drop in sum distance at the window of interest. This value is the 25$^{\text{th}}$ percentile of the total drop in the sum distance, e.g. if the sum distance dropped from 1 to 0.8 the 25$^{\text{th}}$ percentile would be at 0.85. The x axis is the sum distance at the beginning of the window of interest and the y axis is how often this drop occurred.

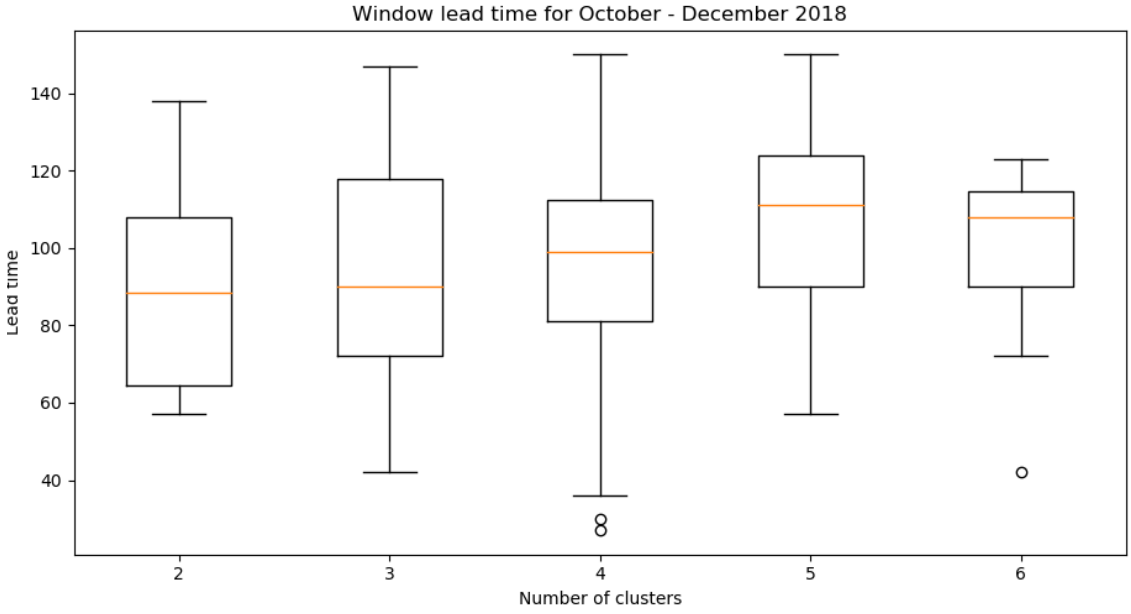
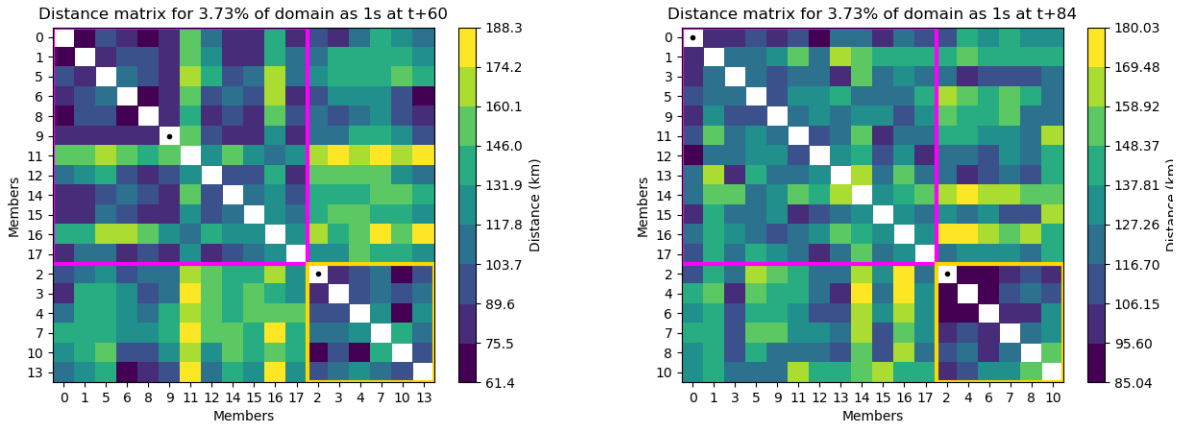


Figure 4.5: A series of box plots showing the distribution of how often a particular number of clusters is chosen compared to lead time when the window of interest begins.

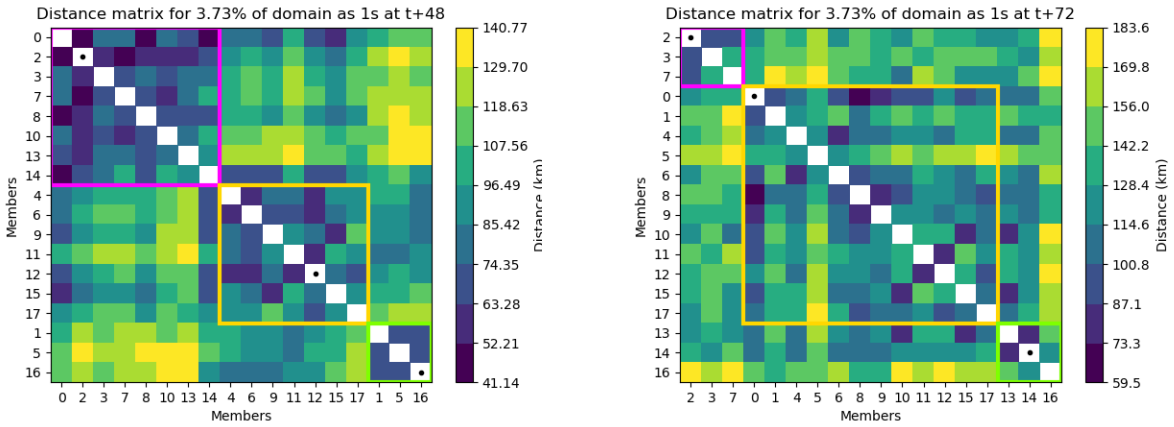


(a) (b)

Figure 4.6: Distance matrices for the forecast beginning on 08/10/2018 at 1200 UTC at t+60 hours (a) and t+84 hours (b). The members are listed by number along the x and y axes, sorted into their corresponding clusters. The clusters are designated by the coloured boxes along the diagonals (magenta for cluster 0, gold for cluster 1, and when applicable: chartreuse for cluster 2, cyan for cluster 3, indigo for cluster 4, and brown for cluster 5). The black dots along the diagonals indicate the medoid of that cluster.

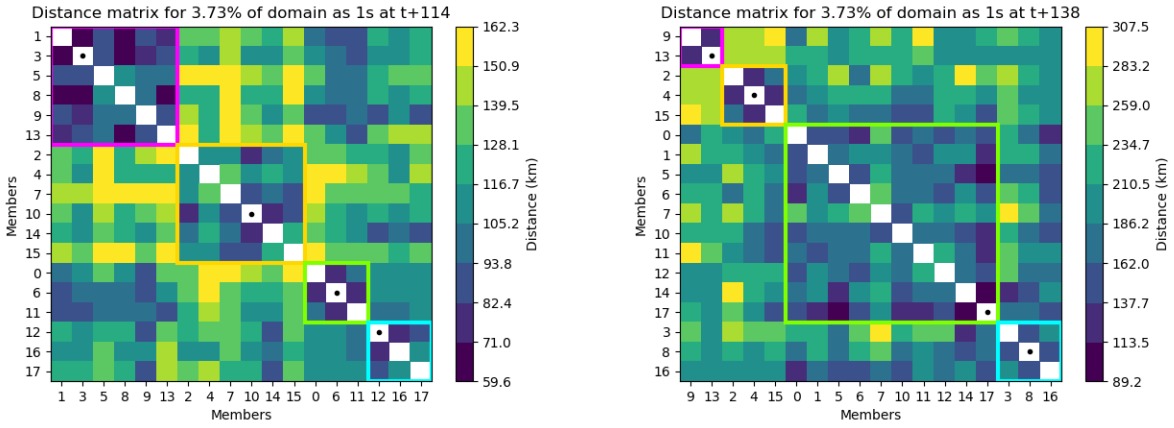
4.3.1 Examination of distances between members

When clustering simple data sets such as 2D data it is a relatively easy process to create a visualization of how the data clusters together. An example of this would be a scatter plot of the data with the different clusters plotted in different colours. However, visually representing the clustering of more complex data requires a different approach. Therefore, during the design of the algorithm a visualization of the distance matrix was created that displays the FSS distances calculated between members and rearranges them so they are grouped into their respective clusters. They provide a quick visual reference for inter and intra-cluster distance between members, cluster size, and the number of clusters. These plots also provide a visual representation of the sum distance and how it relates to clustering. Examples of these plots can be seen in figures 4.6 to 4.10, where each figure displays a different number of clusters for reference and comparison. Within these plots, the x and y axes are labeled with the member numbers from 0 to 18 and the colour bar is the distance between members calculated via the Fractions Skill Score. The members have been reorganized so that clusters can be represented via boxes drawn in outline colours representing their cluster. For example, in figure 4.6, representing two clusters, the magenta outline encloses all the members that make up cluster 0 and the gold outline encloses all the members that make up cluster 1. The black dots along the



(a) (b)

Figure 4.7: Distance matrices for the forecast beginning on 09/10/2018 at 0000 UTC at t+48 hours (a) and t+72 hours (b). Details available in figure 4.6

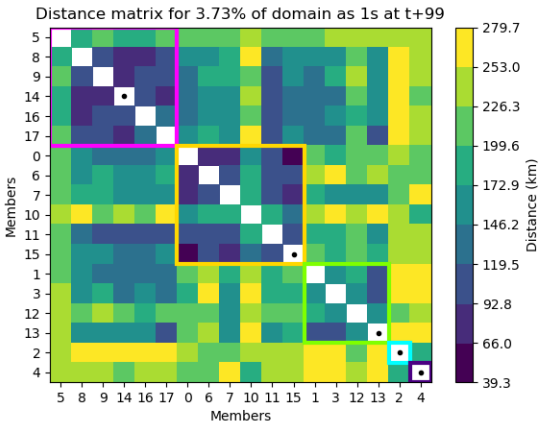


(a) (b)

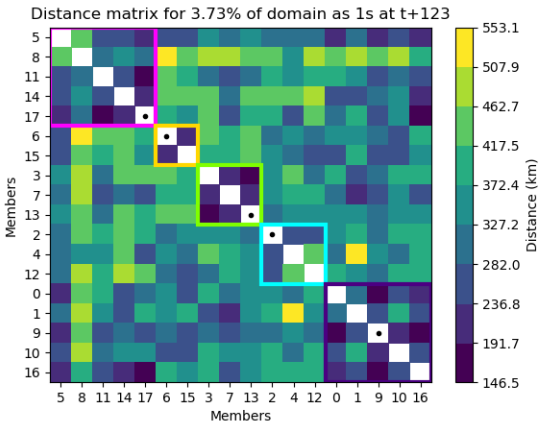
Figure 4.8: Distance matrices for the forecast beginning on 07/11/2018 at 0000 UTC at t+114 hours (a) and t+138 hours (b). Details available in figure 4.6.

diagonal represent the clusters' medoids at the given lead times, which are chosen at the beginning and halfway point of the window of interest. The k-medoids algorithm seeks to find the best member for use as a medoid and the best distribution of members to medoids to optimize the sum distance.

Within the distance matrix plots it can be seen that the members within a cluster are closer to their own cluster medoid than the any other medoid. However, in figure 4.7(a) member 14 appears close to both the cluster 0 medoid (member 2) and the cluster 1 medoid (member 12), but is ultimately closer to cluster 0. This is likely due to there being only small variations among clusters at such an early lead time. A clear example of members being closer to their medoid than any other medoid is plot (a) in figure 4.8. Here, looking at distances between members and medoids, it can be seen that members

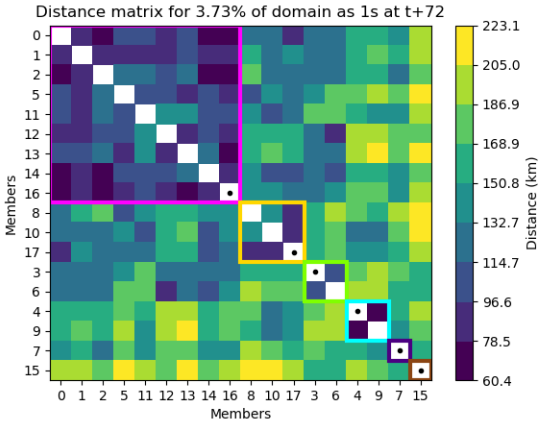


(a)

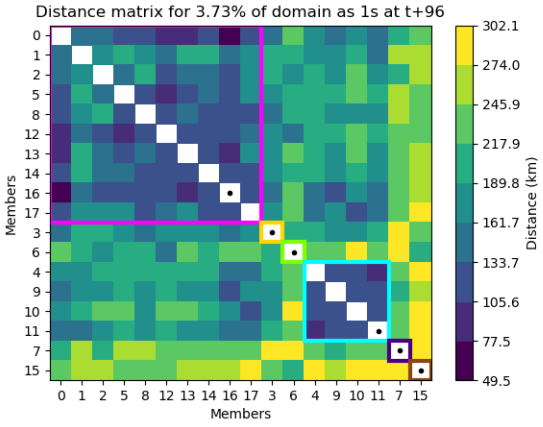


(b)

Figure 4.9: Distance matrices for the forecast beginning on 25/12/2018 at 1800 UTC at t+99 hours (a) and t+123 hours (b). Details available in figure 4.6.



(a)



(b)

Figure 4.10: Distance matrices for the forecast beginning on 30/10/2018 at 0600 UTC at t+72 hours (a) and t+96 hours (b). Details available in figure 4.6.

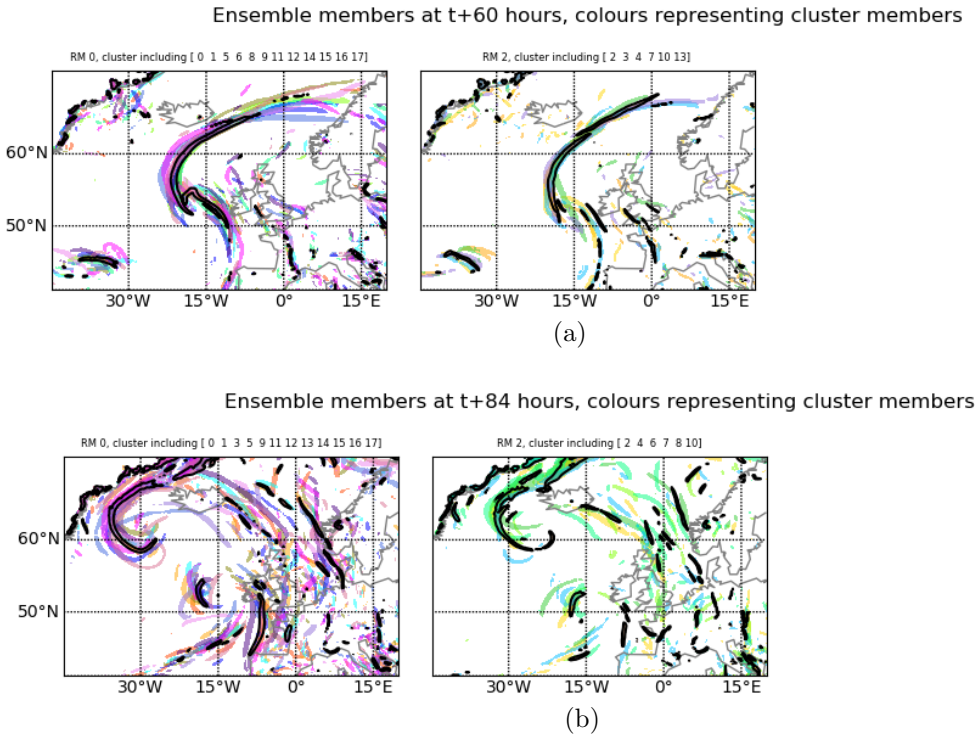


Figure 4.11: Paintball plots of $|\nabla\theta_w|$ at 850 hPa for the forecast beginning on 08/10/2018 at 1200 UTC with two clusters. The provided plots are for the lead times over the window of interest. Each member is portrayed by a unique colour and the representative member (when present within the cluster) is signified by a black outline.

within a cluster are typically very close to their medoid and other members within their cluster, but significantly further away from members and medoids outside their cluster. As cluster number is increased, membership will naturally tend to decrease as members are spread into more clusters, seen in figure 4.9, where the maximum cluster membership is 6 in plot (a). However, it is still possible for there to be one or more large clusters and several small clusters, such as in figure 4.10. Here there is one primary cluster and several small or single member clusters, indicating this forecast may contain several members that are very similar and a handful of members that show dramatically different scenarios as outliers. To fully explore this and ensure the clustering translates to meteorological features, members can be compared visually using paintball plots, named so due to the multi-coloured splotches representing different members.

4.3.2 Meteorological representation of clusters

Examining the paintball plots is beneficial in visually confirming cluster robustness. Figures 4.11 to 4.15 show the paintball plots associated with the previous distance plots (figures 4.6 to 4.10) at the beginning and halfway through the window of interest, respec-

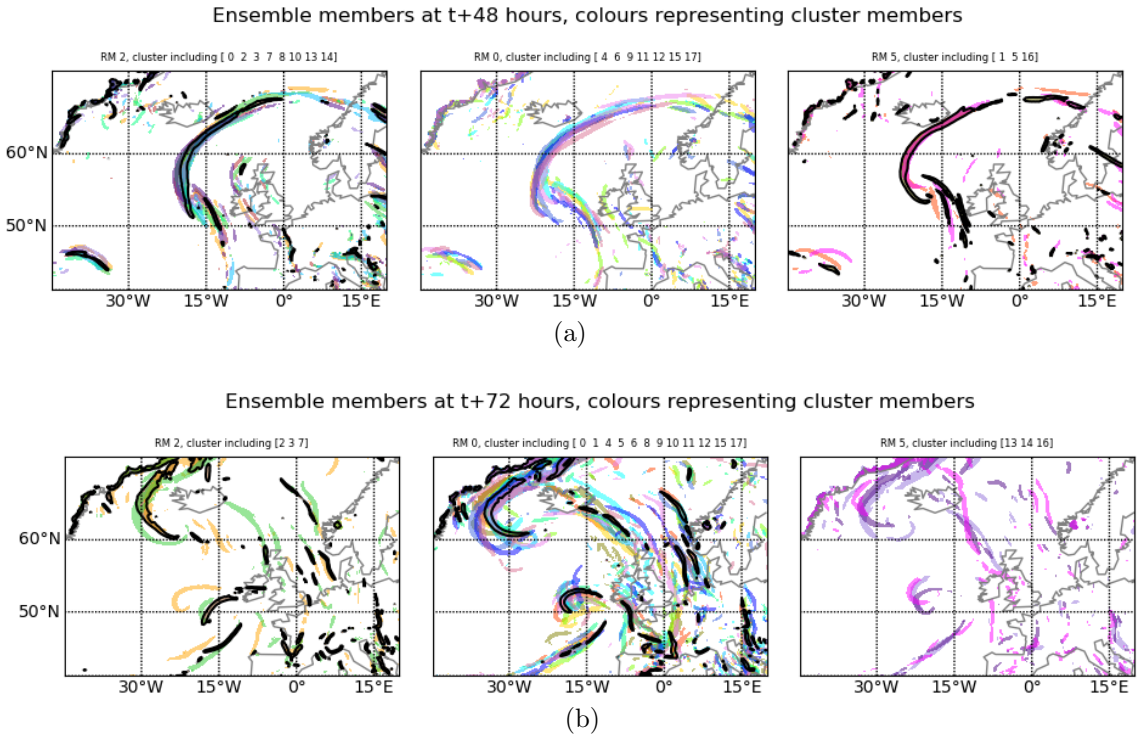


Figure 4.12: Paintball plots of $|\nabla\theta_w|$ at 850 hPa for the forecast beginning on 09/10/2018 at 0000 UTC with three clusters. Details can be found in figure 4.11.

tively. Similar to the previous section, these plots each have a different number of clusters for reference and comparison. The paintball plot displays the $|\nabla\theta_w|$ frontal objects for each member in a different colour, where members in the same cluster at the given lead time are plotted on the same map. The RM is outlined in a black contour when it is present in the cluster at the lead time given. Figure 4.11 is an example of two clusters at the beginning and half way through the window of interest. Plot (a) shows a similar frontal structure between the two plots displaying a mid-latitude cyclone, but there is a clear difference in displacement of the front between the clusters. There is a very strong visual match amongst members within each cluster. Figure 4.11b has a very similar situation, where the curvature and position of the primary frontal region is different amongst the members. This is a similar trend amongst the other paintball plots, concluding the robustness of cluster membership tracks through numerical and visual analysis.

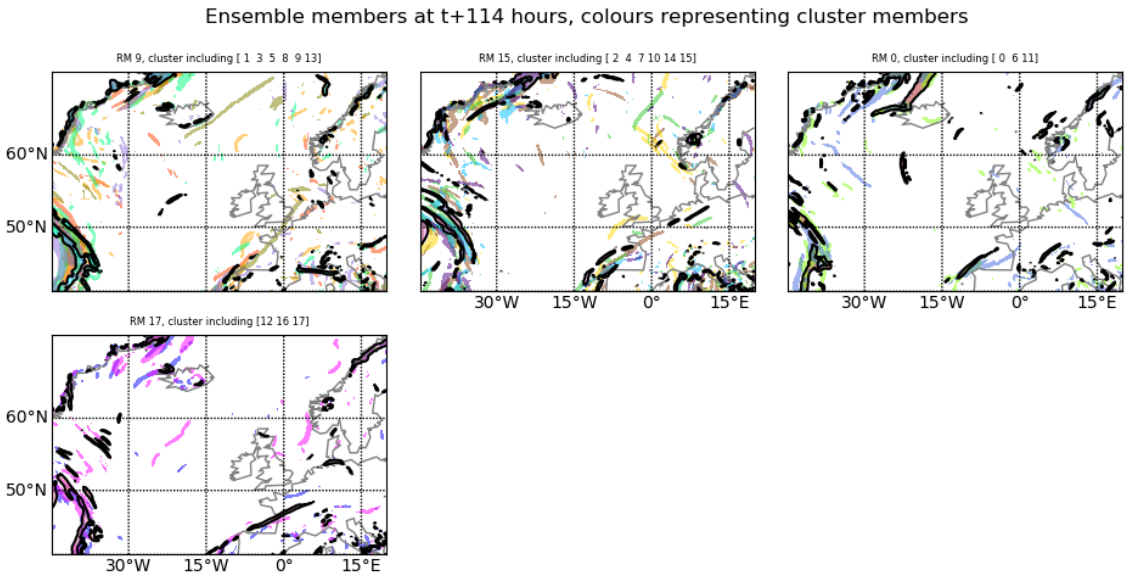
It is important to note that although Greenland itself was excluded in the data, there are frontal features appearing around the coastline in the majority of paintball plots that have been explored in this work. While an in-depth sensitivity analysis was not performed on whether or not these features affect the clustering, a visual analysis provided enough justification to assume their effect was generally minimal. As they are typically very small features and have a small displacement between members compared to the larger

frontal features the clustering focuses on, it can be assumed they are unlikely to affect the results of the FSS and by extension the clustering. However, a future sensitivity study is recommended as the method develops further for use.

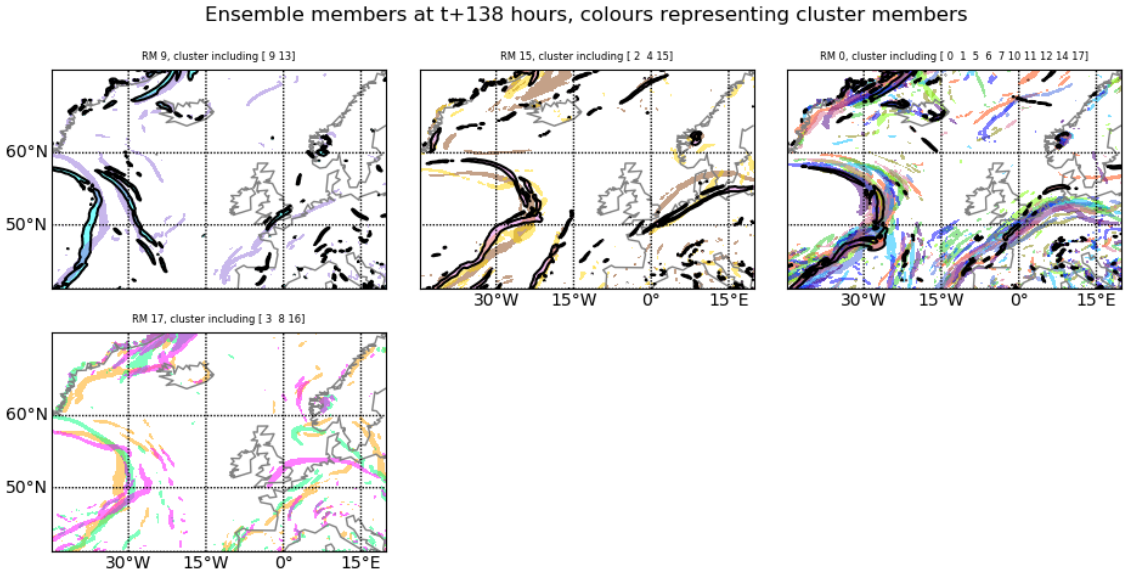
Similar to the two cluster example, the three cluster example in figure 4.12 shows a strong correlation of members in each cluster. It is important to note, however, that the clustering method is performed at each lead time, and therefore members may not always be grouped together. This is also true for representative members of a cluster. An example of this can be seen in the middle map in 4.12(a) and the right map in 4.12(b) where they are missing the representative member (the black contour seen in other plots). This is not necessarily an issue that must be resolved, as restricting the representative member too much (e.g. if the RM is not present in the window of interest for the full length of time then it is not a valid representative) will potentially force the optimal number of clusters into a much smaller number, likely missing some scenarios in the process. Another example of the movement of RMs is figure 4.14, where only two clusters in plot (b) contain their representative member at that lead time. In this instance, there was a strong relationship between members in plot (a), but later in the window the variation between members increased to the point that there was little cohesion. In this instance, the best period for clustering would be earlier in the window, and the frontal features in this forecast are too uncertain to provide significantly strong RMs.

Instead of strictly restricting the number of clusters based on how often the RM is present, it is more important to restrict the likelihood of outliers. Outlying representative members must be stable enough in their cluster to truly be considered an outlier and able to stand on its own without other members. Figure 4.15 is an excellent example of outlier RMs. In plot (a) at the beginning of the window of interest it can be clearly seen that the majority of members are in the top left cluster, following a well defined warm front over the North Atlantic and a cold front moving eastward. The second largest cluster, in the top middle map, has a similar warm frontal feature but is lacking a defined cold front. All of the remaining maps have distinctly different positions for a partial warm front. Progressing to figure 4.15b, we can see the majority of members are now part of the top left cluster, and the remaining clusters all have a very distinct and dissimilar evolution of the frontal system.

Figure 4.13 is an excellent example of a developing event entering the domain and



(a)



(b)

Figure 4.13: Paintball plots of $|\nabla\theta_w|$ at 850 hPa for the forecast beginning on 07/11/2018 at 0000 UTC with four clusters. The provided plots are for the lead times over the window of interest. Details can be found in figure 4.11.

the clustering method picking up on the uncertainty of its progression. In plot (a) there are several smaller frontal features in the domain but there is a potentially large feature moving in from the southwest. In plot (b) there is very strong clustering around the frontal region, with the top left map showing a nearly meridional frontal feature and the remaining maps displaying a long curved front, the primary difference between them being the northern displacement of the southeastern bend in the front.

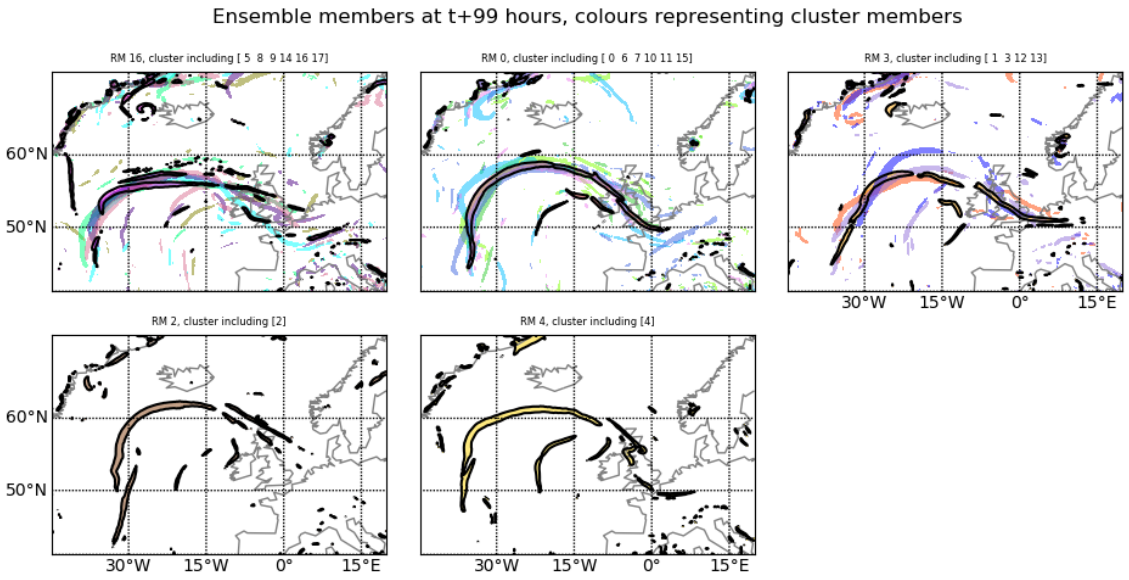
These paintball maps can be a valuable tool for operational meteorologists for quickly visually verifying cluster robustness, uniqueness, and longevity (i.e. when clustering is strong or beginning to break down), particularly within the window of interest. These maps also present a quick way to see outlying RMs, how similar members are to one another, and how big a cluster is.

4.4 Clustering across lead times

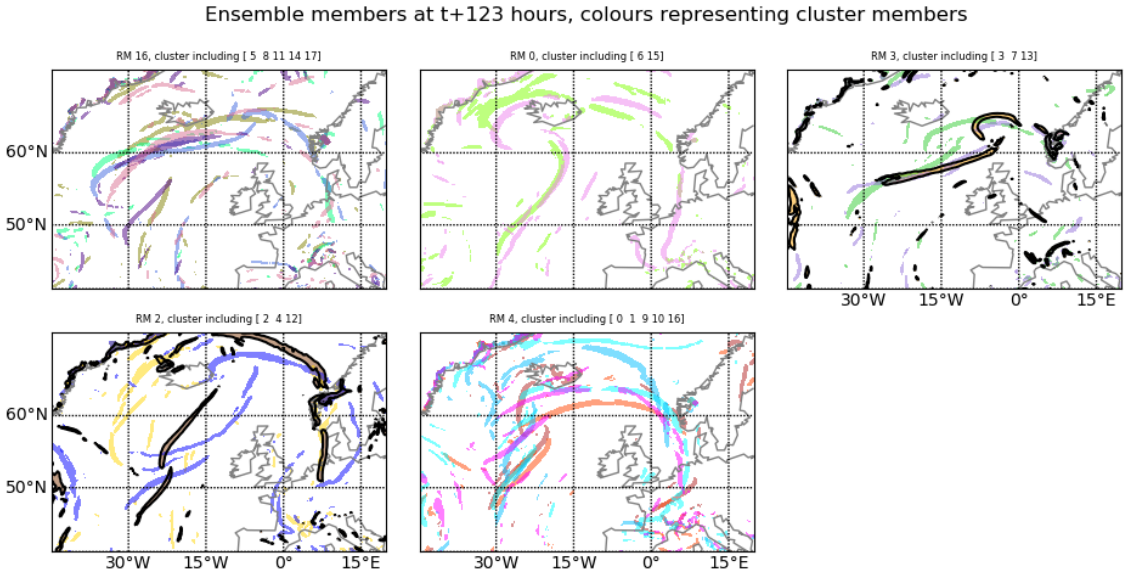
At individual lead times members are clustered effectively into groups. However, this means that each lead time is clustered independently of all other lead times, which presents a different issue: connecting clusters across lead times when the labeling is arbitrary. Therefore, a method to trace clusters across lead times must be employed, which must include re-labelling the clusters. The following sections will discuss tracing clusters across lead times, how the traceability of the clusters relates to the sum distance, and the variation in the representative members.

4.4.1 Tracing clusters by comparison

Cluster traceability can be defined by the degree to which cluster membership remains stable across lead times. By using the cluster inter-comparison diagrams (figure 4.16, described in 3.3.3) to compare cluster membership at different lead times to cluster membership at a set lead time, the cluster can essentially be traced across a forecast. The set lead time used for this process is the beginning of the window of interest. Due to the nature of an ensemble forecast, members are close to the control at early lead times and tend to group mostly in one cluster. As multiple clusters are imposed upon the data, at least one member must be in every cluster, but during the beginning of the forecast the clusters are typically not distinct or may contain a single member on the edge of

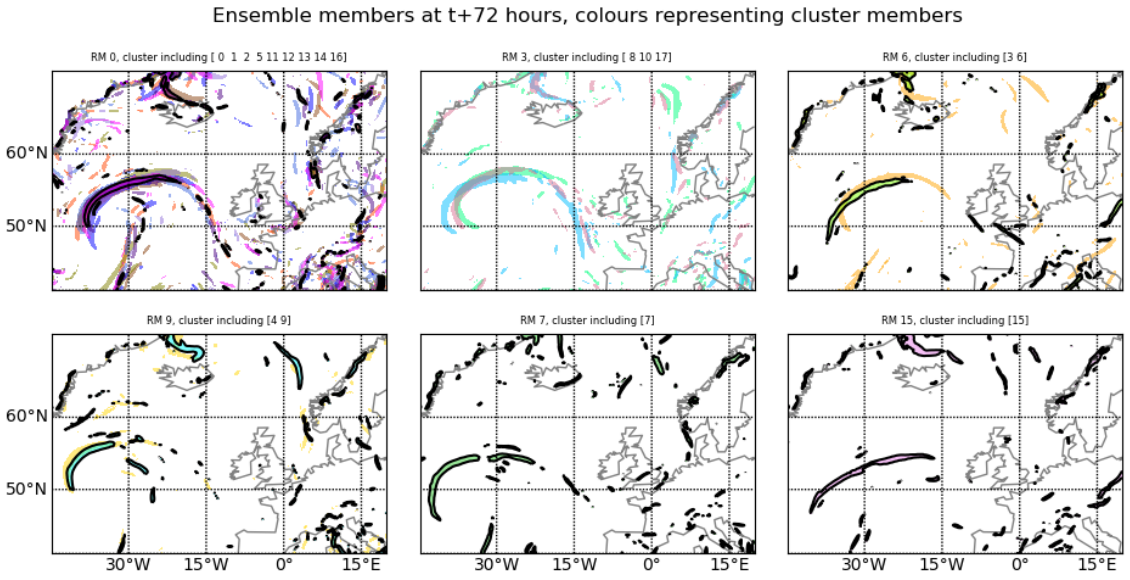


(a)

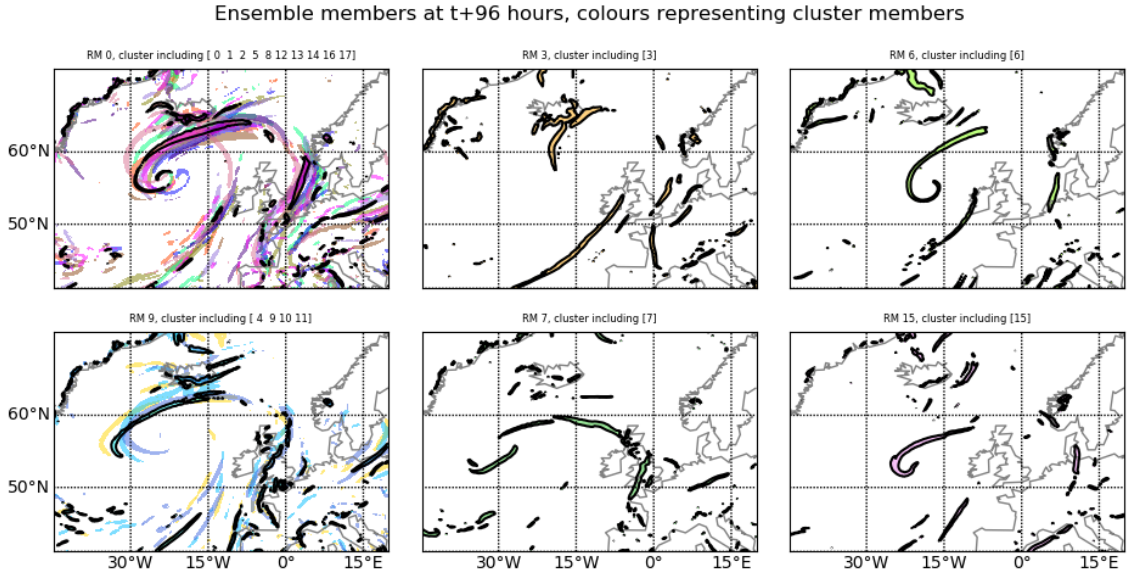


(b)

Figure 4.14: Paintball plots of $|\nabla\theta_w|$ at 850 hPa for the forecast beginning on 25/12/2018 at 1800 UTC with five clusters. The provided plots are for the lead times over the window of interest. Details can be found in figure 4.11.



(a)



(b)

Figure 4.15: Paintball plots of $|\nabla\theta_w|$ at 850 hPa for the forecast beginning on 30/10/2018 at 0600 UTC with six clusters. Details can be found in figure 4.11.

the distribution. This is expected due to the small amount of spread at this time. As the forecast progresses closer to the window of interest, the ensemble will naturally split around sensitive points in the atmospheric flow, resulting in members naturally grouping together more frequently. Then during the window, when clustering is at its strongest, membership tends to be more stable. After the window, membership tends to become more erratic as members drift further apart in similarity and there are not enough clusters to adequately represent all of the different outcomes as lead time increases. Therefore, it is expected that clustering into a few groups will be a transient behaviour that will occur between early and late lead times, leading to a window of interest.

To demonstrate how cluster membership changes across lead times when the number of clusters is specified, the set of inter-comparison diagrams in figure 4.16 are split between before the window of interest (t+0 to t+69) and during it (t+72 to t+117). At the beginning of the window of interest, clusters are relabeled in descending order of size with the largest cluster being labeled as 0. In this case, all lead times are compared to the clusters at t+72, and are relabeled to align with their closest match with the clustering labels assigned at t+72. As described in the previous paragraph, during the beginning of the forecast the majority of the cluster membership is within a single cluster. The membership of other clusters increases as time proceeds towards the window of interest. As the forecast approaches the window of interest a strong coherence between the cluster membership develops. This coherence increases during the beginning of the window then membership begins to change and traceability reduces towards the end of the window as members move further away from the control and each other in similarity. This period of coherence is what is important for extracting representative members.

4.4.2 Traceability and the sum distance

As previously mentioned, during the beginning of a forecast the ensemble is run with nearly the same set of initial conditions for each member leading to a tendency for members to group into a single cluster at early lead times. However, the nature of the method is such that the number of clusters must be defined before clustering. This often results in one large cluster and 1 or more single member clusters during the first few days of an ensemble forecast, indicating a fairly well predicted period. As lead time progresses, members begin

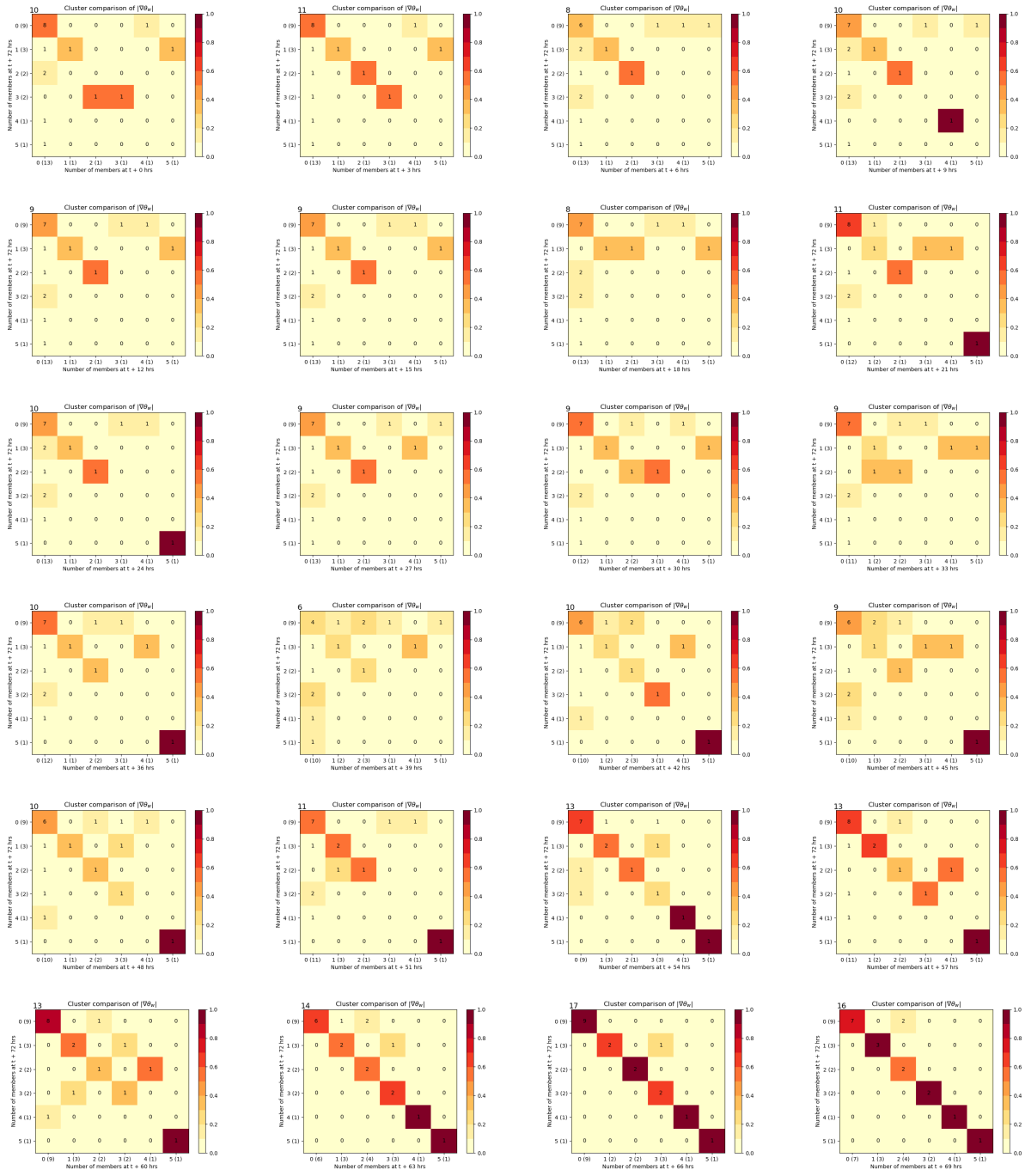


Figure 4.16: Cluster inter-comparison diagrams from 30/10/2018 at 0600 UTC with six clusters between t+0 and t+69 hours, before the window of interest, where the y-axis describes the clusters at the beginning of the window of interest (i.e. 0 is the cluster label and (6) is the number of members within that cluster), the x-axis describes the clusters at various lead times, the colour bar represents the Jaccard Index, and the number at the top left of the chart indicates the sum along the diagonal, where 18 indicates a perfect match.

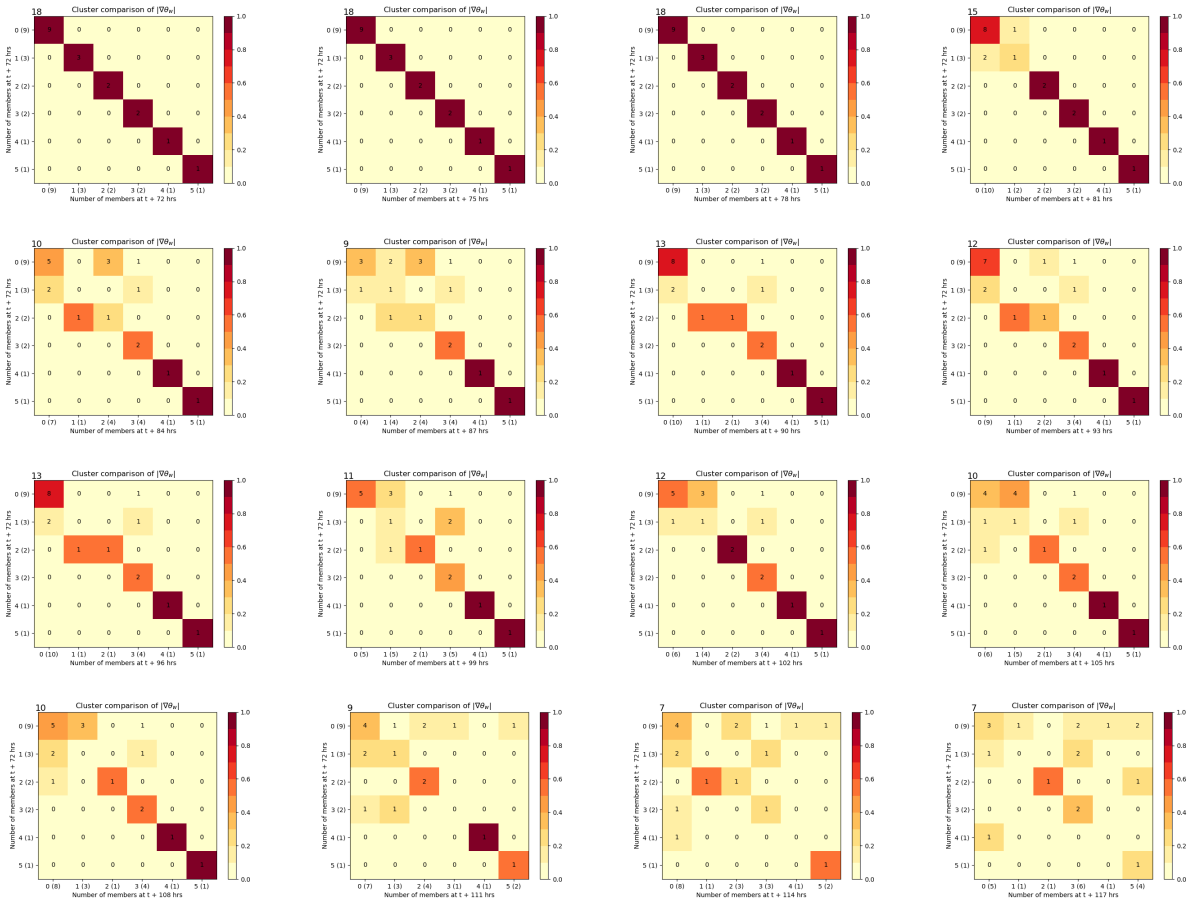


Figure 4.16: (Cont.) Cluster inter-comparison diagrams from 30/10/2018 at 0600 UTC with six clusters between t+72 and t+117 hours, during the window of interest.

to separate from one another, and clustering becomes more distinct. This can be seen in figure 4.17, which is a traceability plot for the 0600 UTC 30/10/2018 forecast with six clusters, which was the optimal solution for this case due to the strong outlying members in clusters 4 and 5 (described in section 3.3.3.2). Using the representative members derived from each set of clusters and windows, the number of optimal clusters can be determined (see section 3.4.4). The representative member (see section 3.3.4.2) is signified by a horizontal blue dotted line throughout the cluster. The bottom plot also contains the ensemble spread as a function of the FSS distance (equation 3.4) as a blue dashed line, which increases as forecast lead time increases. In figure 4.17, the membership of most of the clusters is erratic before the window of interest. When the window begins, membership of the clusters stabilizes, with clusters 4 and 5 showing strong outlying members, 0, 2 and 3 showing relatively stable membership, and clustering 1 showing a close relation with cluster 2, due to its RM being present often in both. After the window of interest, membership of the clusters becomes more evenly distributed but increasingly erratic in stability, although the RM in cluster 5 remains distinct through

to the end of the forecast. In this case, it might be expected that this member is a particularly distinct forecast scenario. How the clusters behave leads to the question of how robust the clusters are.

The sum distance (equation 3.6) is a measure of how distinct the clusters are. This is important as the clusters will be the most distinct from each other when the sum distance is the lowest. It is expected that the ensemble spread will increase during a forecast on average over many forecasts (see figure 4.17), but the sum distance is not expected to decrease continuously with lead time. It is also expected that the *SDist* begins near 1 at t+0 because the ensemble is well described by a single cluster. The *SDist* then dips to a minimum at intermediate times when **K** clusters are distinct and are a good description of the ensemble. After this time, the ensemble members will eventually diverge further from one another and *SDist* is expected to not be as low. This can be seen in figure 4.17, where there is a reduction in ensemble spread and increase in *SDist* almost halfway through the window of interest. One possible explanation of this is that the atmospheric feature that was the main focus of the clustering has left the domain or is otherwise no longer a feature.

4.4.3 Variation in representative members

The method extracts representative members (RMs, described in section 3.3.4.2) from the clusters based on what member is closest to the centre of the cluster during the entire window of interest and then presents them as potential scenarios for forecasters to review. This means the RM is often the medoid, i.e. the central member of the cluster, during the window, but it does not have to be the medoid at all. The RMs should also be unique from one another, regardless of the number of clusters present, which is enforced via the algorithm. A single member is chosen as the RM so that the scenario presented to operational meteorologists is a consistent solution of the atmospheric evolution through time and can be used to extract other variables in relation to that scenario for analysis.

The RMs are presented by their θ_w and $|\nabla\theta_w|$ fields in figures 4.18, 4.19, and 4.20, where the threshold used for creating a binary field for calculating the FSS has been applied to the $|\nabla\theta_w|$ field. The lead times presented are at the beginning of the window of interest (figure 4.18), the half way point of the window of interest (figure 4.19), and

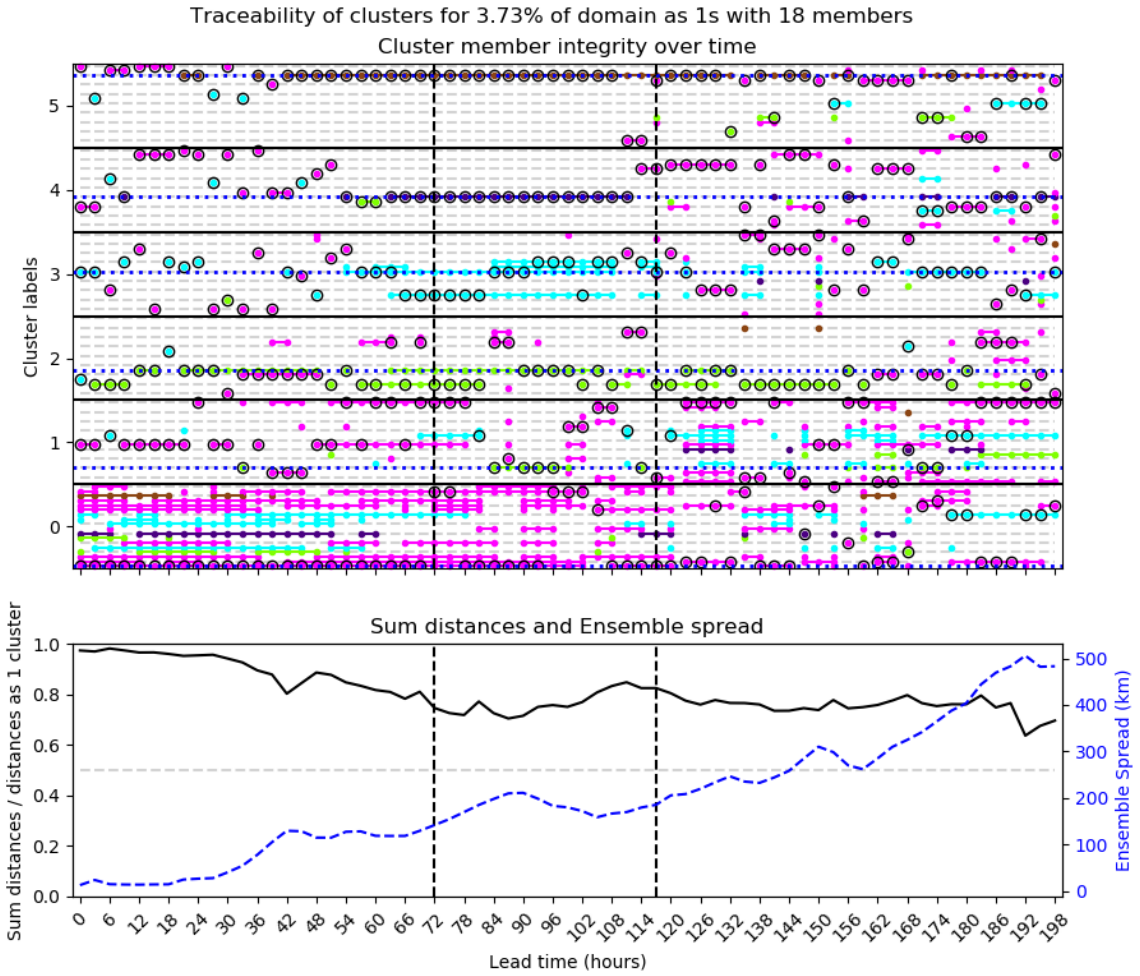
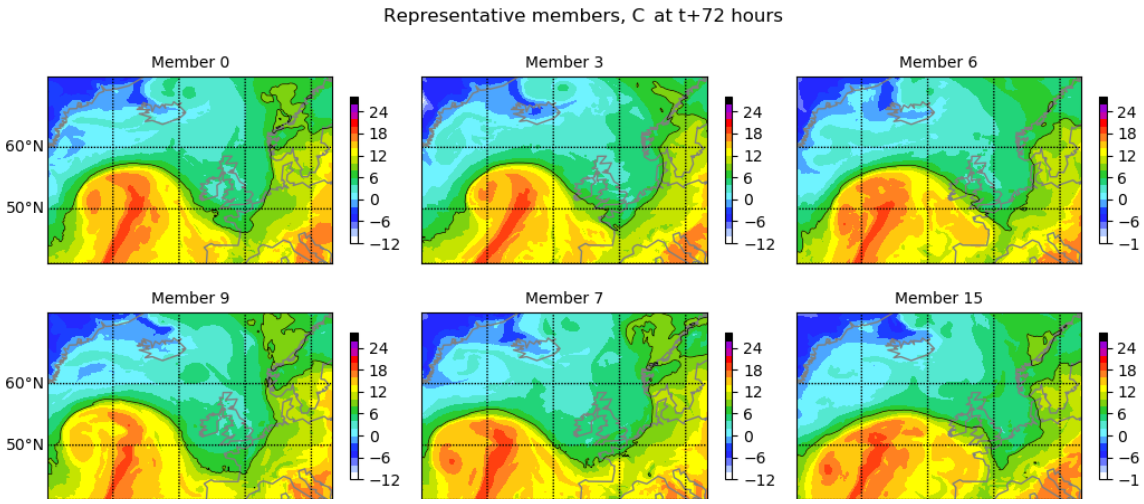
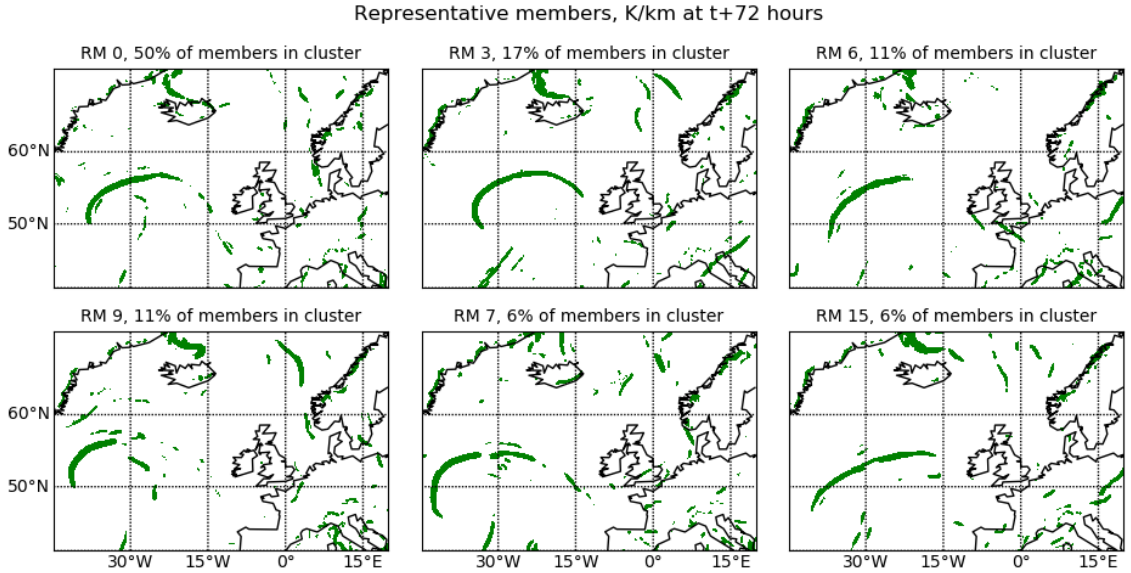


Figure 4.17: The traceability diagram of cluster membership (top), and the sum distances and the ensemble spread (bottom) versus the lead time in hours from the forecast on 30/10/2018 0600 UTC, over a domain of 40° to 70° north and 45° west to 20° east.

the end of the window of interest (figure 4.20). In figure 4.18 (a) there is a strong wave apparent in all the RMs, though its exact shape, position, and intensity varies significantly between members. Member 0 and member 3 are relatively similar, but when plot (b) is examined, there is a notable curve in the western section of the frontal region in member 3 that is not present in member 0, and a stronger southwesterly front near the domain boarder. Member 9 also has a similar shape of the wave, but the frontal region is less well defined in the gradient plot than in 0 and 3. The remaining 3 wet-bulb potential temperature plots (6, 7, and 15) all have varying differences from one another. Member 7 is further west, member 15 has the broadest range, and member 6 has the largest warm core. This results in vastly different frontal region plots, with a tight curve appearing in member 7, a partial frontal arc that has no southward curve to its eastern end in member 6, and a rather flat frontal feature in member 15. Moving to later in the window in figure 4.19, it becomes even clearer how these RMs differ. Member 0 has evolved into a tightly spiraled system, where the frontal region has continued northward towards Iceland. Member 3 still displays that strong southern front that is now impacting the UK. Members 9, 6, and 15 all show a frontal region similar to 0, however they all display a significant difference in position and curvature of the storm front. Member 7 shows the primary front of the cyclone breaking down but still impacting the UK. In figure 4.20, the wave has mostly moved into a trough, leaving scattered frontal regions impacting the British Isles and mainland Europe. Members 0 and 3 both have similar fronts affecting the southeastern English coast, however the distribution of fronts in the North Atlantic differs significantly. Members 6 and 7 also have fronts affecting the southwestern English coast, but the front in member 6 is less prominent and at a shallower angle than 7, and member 7 is both a stronger front and has a second front impacting Ireland. The front affecting England in member 9 is further inland, and member 15 displays no significant frontal features. Although it is clear all RMs are distinct from one another it is still up to an operational meteorologist to determine if they are in fact distinct weather scenarios, which is evaluated further in chapter 6.

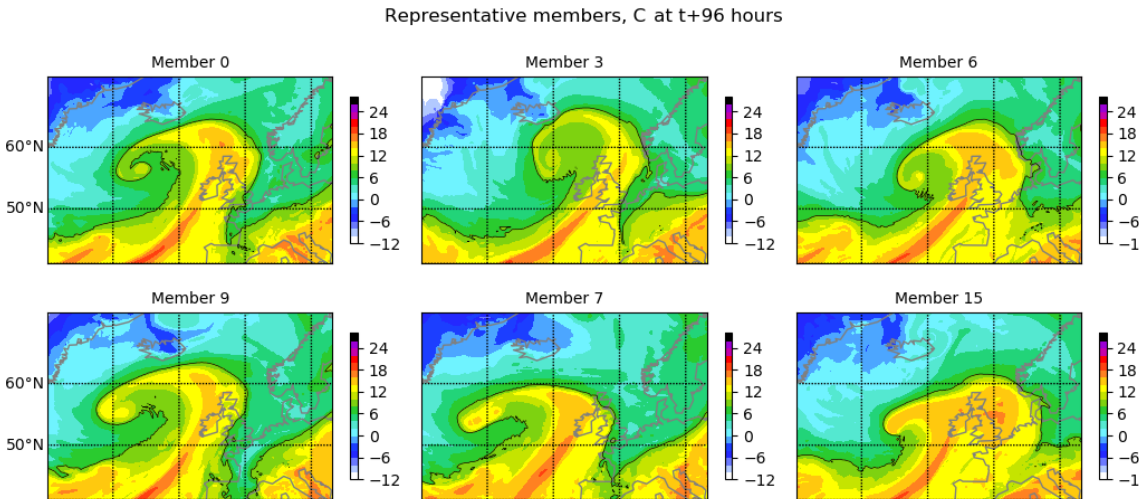


(a)

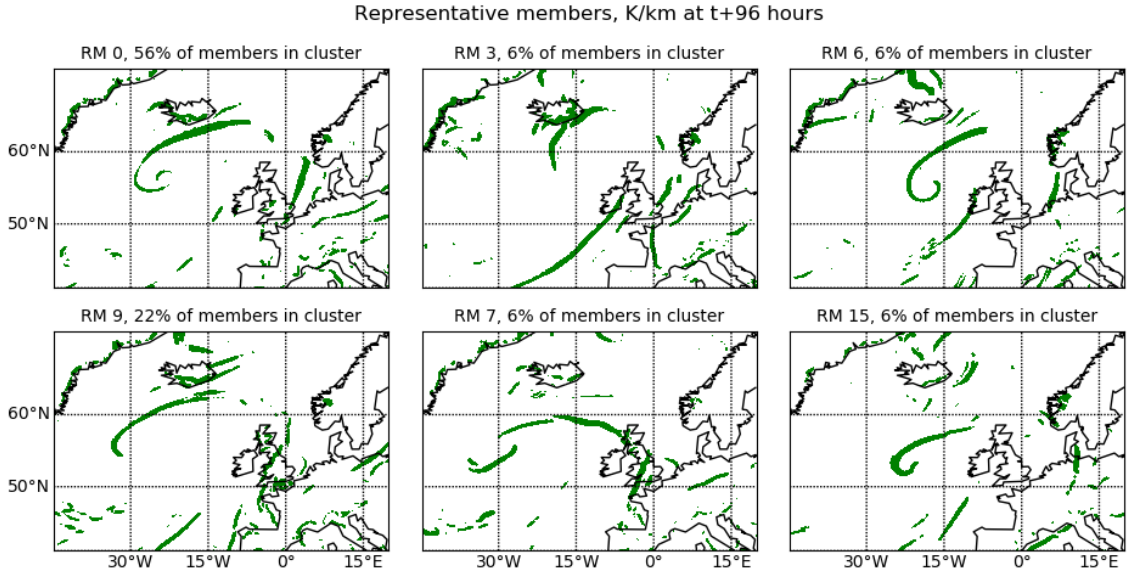


(b)

Figure 4.18: Representative member plots in θ_w at 850 hPa and frontal objects from $|\nabla\theta_w|$ at 850 hPa for the six cluster solution of the forecast on 30/10/2018 at 0600 UTC at a lead time of t+72 hours corresponding to the beginning of the window of interest.



(a)



(b)

Figure 4.19: Representative member plots in θ_w at 850 hPa and frontal objects from $|\nabla\theta_w|$ at 850 hPa for the six cluster solution of the forecast on 30/10/2018 at 0600 UTC at a lead time of t+96 hours corresponding to the centre of the time window of interest.

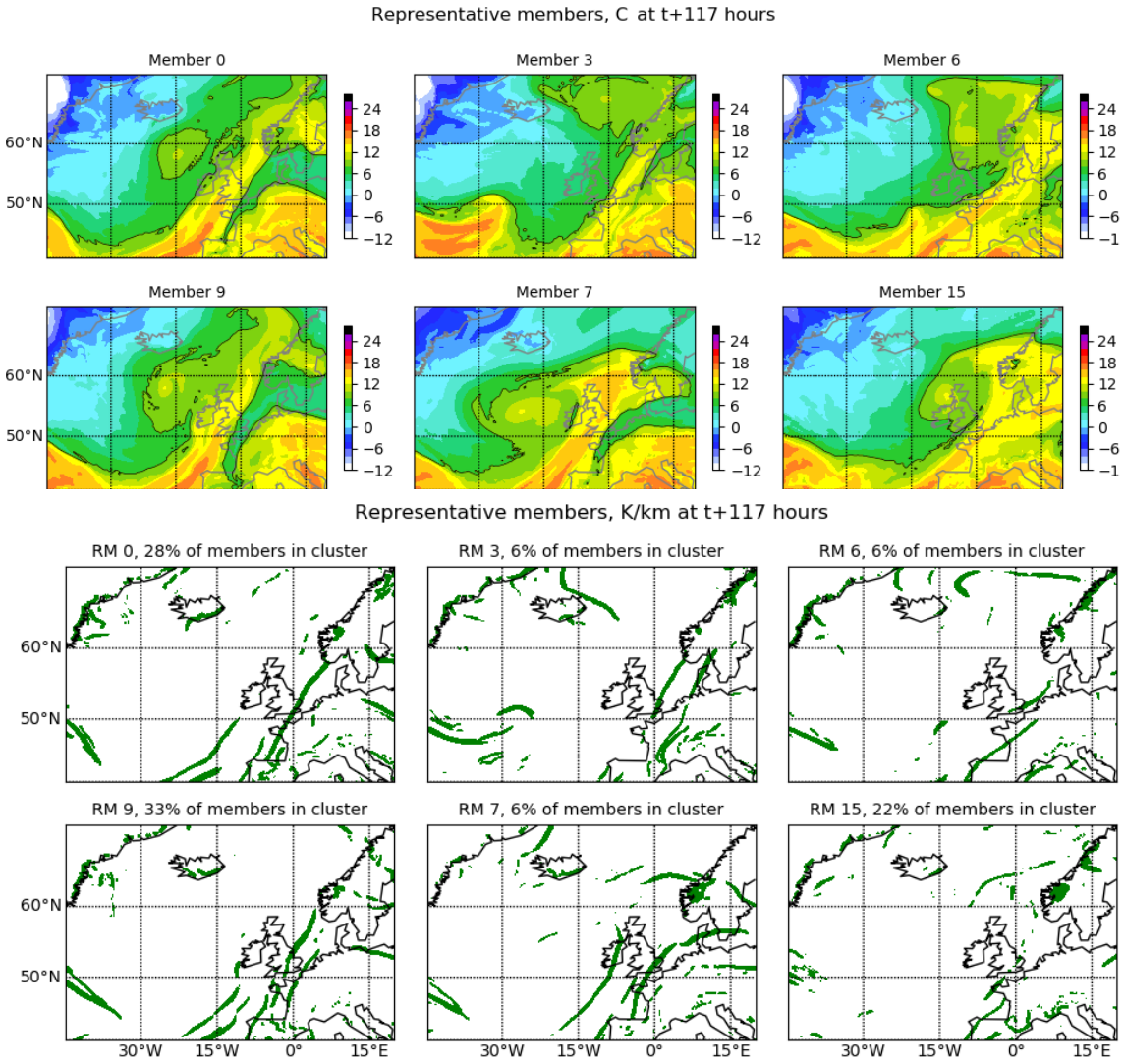


Figure 4.20: Representative member plots in θ_w at 850 hPa and frontal objects from $|\nabla\theta_w|$ at 850 hPa for the 0600 UTC 30/10/2018 forecast at t+117 hours, the end of the window of interest.

4.5 Scenarios and predictability

Extracting scenarios is a primary goal of the development of the method. However, during the development stage it became apparent that applying the algorithm over sequential forecasts and comparing the results had the potential to alert operational meteorologists to particularly uncertain events or atmospheric flows that did not quickly converge into a single (or several very similar) solution(s). Therefore, the clustering method may extend into determining the predictability of a system, which can be further explored. In the sections below, there will be a discussion of extracting scenarios, how scenarios appear across valid times, and the predictability of the scenarios in question.

4.5.1 Extracting potential scenarios

When crafting a method to reduce an ensemble to a few members, it was critical to consider how those members could be interpreted as potential weather scenarios. To increase the likelihood of extracting meaningful scenarios, the algorithm was finely tuned in determining the optimal number of clusters and how the representative members were chosen. This process is detailed in chapter 3, but examples of potential scenarios can be seen in figures 4.18 and 4.19. Here, each representative member for each cluster represents different scenarios. Section 4.4.3 has discussed the variation in the RMs presented, but the question remains what exactly makes an RM a "forecast scenario". Although the clustering is done at each lead time, the RMs are determined by seeking coherence in the clustering over the window of interest to encourage scenario extraction. An example of the atmospheric trajectories and their variations can be seen in figures 4.18 to 4.20. Within these plots six representative members are displayed as potential scenarios for the forecast evolution. While each RM has an overall similar pattern of a developing cyclone to the west of the UK and a warm plume stretching across the UK, there is significant variation in the spatial position and intensity of the features, especially in how the UK is impacted. With this example, it can be seen that the differences in the representative members increases across the window of interest, as expected. However, whether or not these trajectories can be interpreted as distinct scenarios and how they would potentially impact a forecast will be further explored in chapter 6.

4.5.2 Scenarios across valid times

A strong drop in sum distance at a fixed valid time can indicate a drop in predictability of the atmosphere. In the October summary plot (figure 4.1) there are two periods of the month where there appears to be a particularly uncertain event picked up by the method. The first valid time where this occurs is the 11th of October, in relation to the forecasts from October 6th to the 9th. The second is the valid time of November 2nd, which extends off the page, in relation to the forecasts from October 27th to the 30th. There are similar but less well defined valid time periods of the 12th, the 20th, and the 26th. However, in December, there is a very strong drop in sum distance with a well defined valid time correspondence for the 30th. When the same valid time is identified within the window of interest across forecasts it indicates the window in all of these forecasts is linked to the same event. This may alert operational meteorologists to pay careful attention to the state of the atmosphere around the valid time as it approaches. How the scenarios change between forecasts across valid times and whether or not the observed scenario can be traced across valid times from the beginning can be explored further with a case study.

Within figures 4.21 to 4.24 are the representative members for the forecasts from 28/12/2018 at 0600 UTC to 24/12/2018 at 0600 UTC, with the same valid time of 1200 UTC 30/12/2018 (referred to here as the observation) for each forecast. This group of forecasts were chosen as they all contained the same valid time within their window of interest, indicating the method was focusing on this particular period of uncertainty. The method was applied to each forecast and forced to four clusters for easier comparison. The binary field of each RM (calculated during the algorithm for use with the FSS, see section 3.3.1.1) is used to compare RMs across forecasts via the FSS. Each row in the figure includes the four resulting RMs from the respective forecast, identified in the boxes to the left. Each row is the preceding forecast for the previous row and has a lead time corresponding to the observation. The RMs have been linked between forecasts via arrows, which indicate the two RMs are the closest match between forecasts and include the FSS distance in km calculated between the RMs. Each row of maps is ordered from left to right by how close they are to the control member at t+0 (observation, first map in 4.21). The observation in figure 4.21 contains a meridional warm plume with the strongest values over the North Atlantic between 25° and 20° west and a hook-like

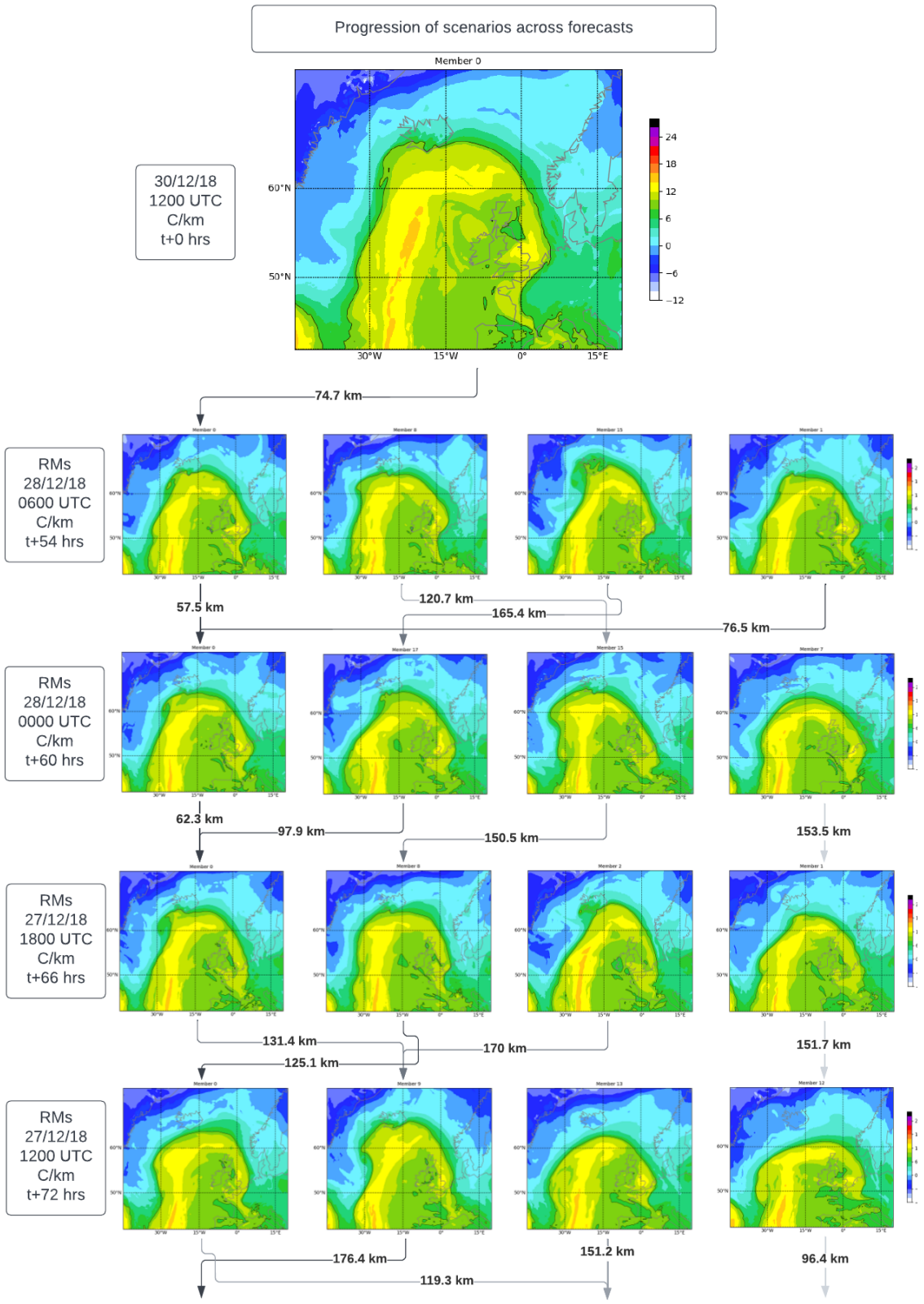


Figure 4.21: The progression of θ_w in C° at 850 hPa scenarios across forecasts with the valid time of 1200 UTC 30/12/2018 (presented as the first plot, obtained from the control member from the 1200 UTC 30/12/2018 forecast), where each consecutive row is a preceding forecast with increasing lead time. The forecasts begin with 0600 UTC 28/12/2018 and go to 1200 UTC 27/12/18 (labeled by the the boxes on the left). The numbers below each plot indicate the FSS distance in km between the two RMs connected by arrows, which are the closest matches between forecasts.

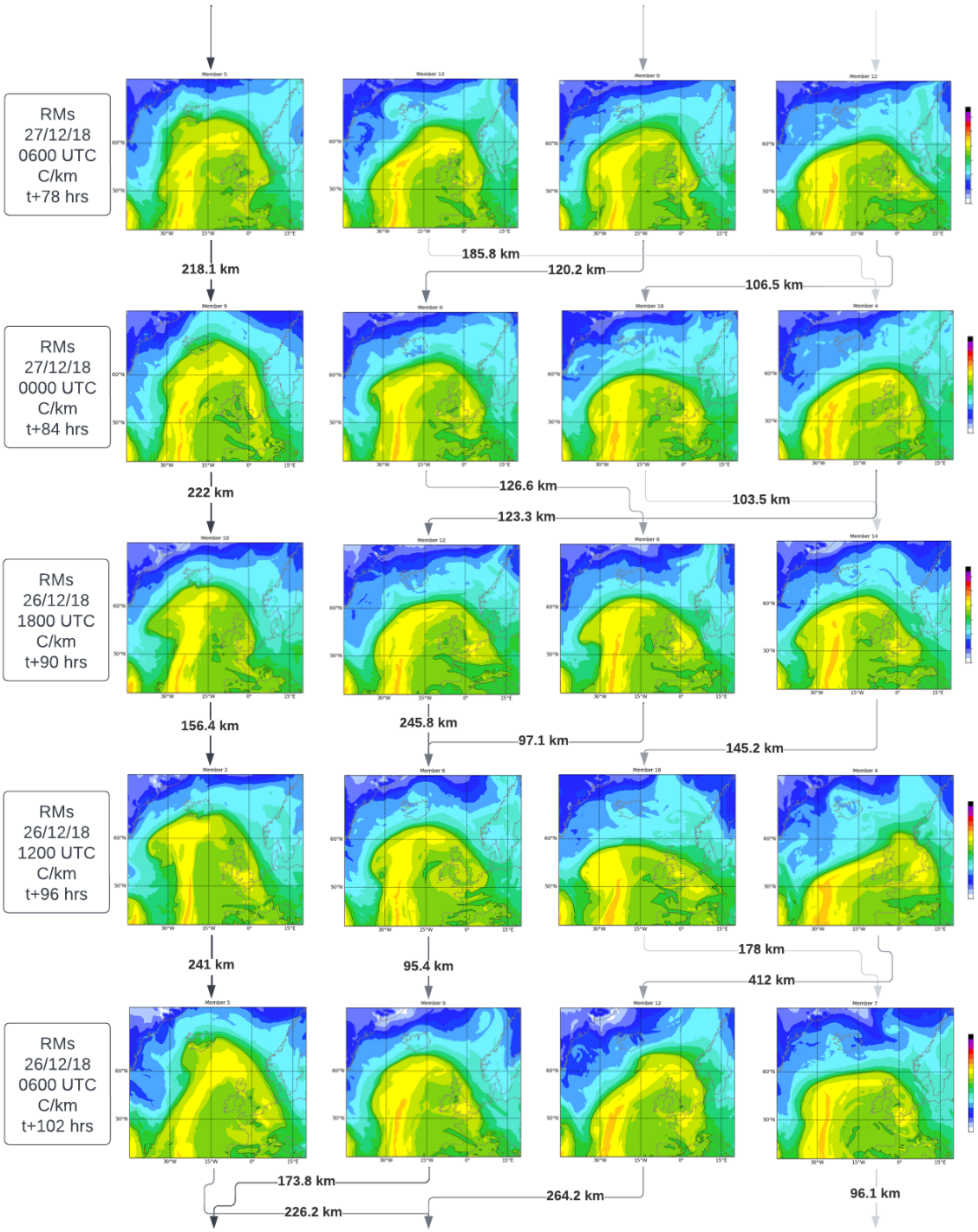


Figure 4.22: The progression of θ_w in C° at 850 hPa scenarios across forecasts with the valid time of 1200 UTC 30/12/2018, where each consecutive row is a preceding forecast with increasing lead time. The forecasts begin with 0600 UTC 27/12/2018 and go to 0600 UTC 26/12/18 (labeled by the the boxes on the left). The numbers below each plot indicate the FSS distance in km between the two RMs connected by arrows, which are the closest matches between forecasts

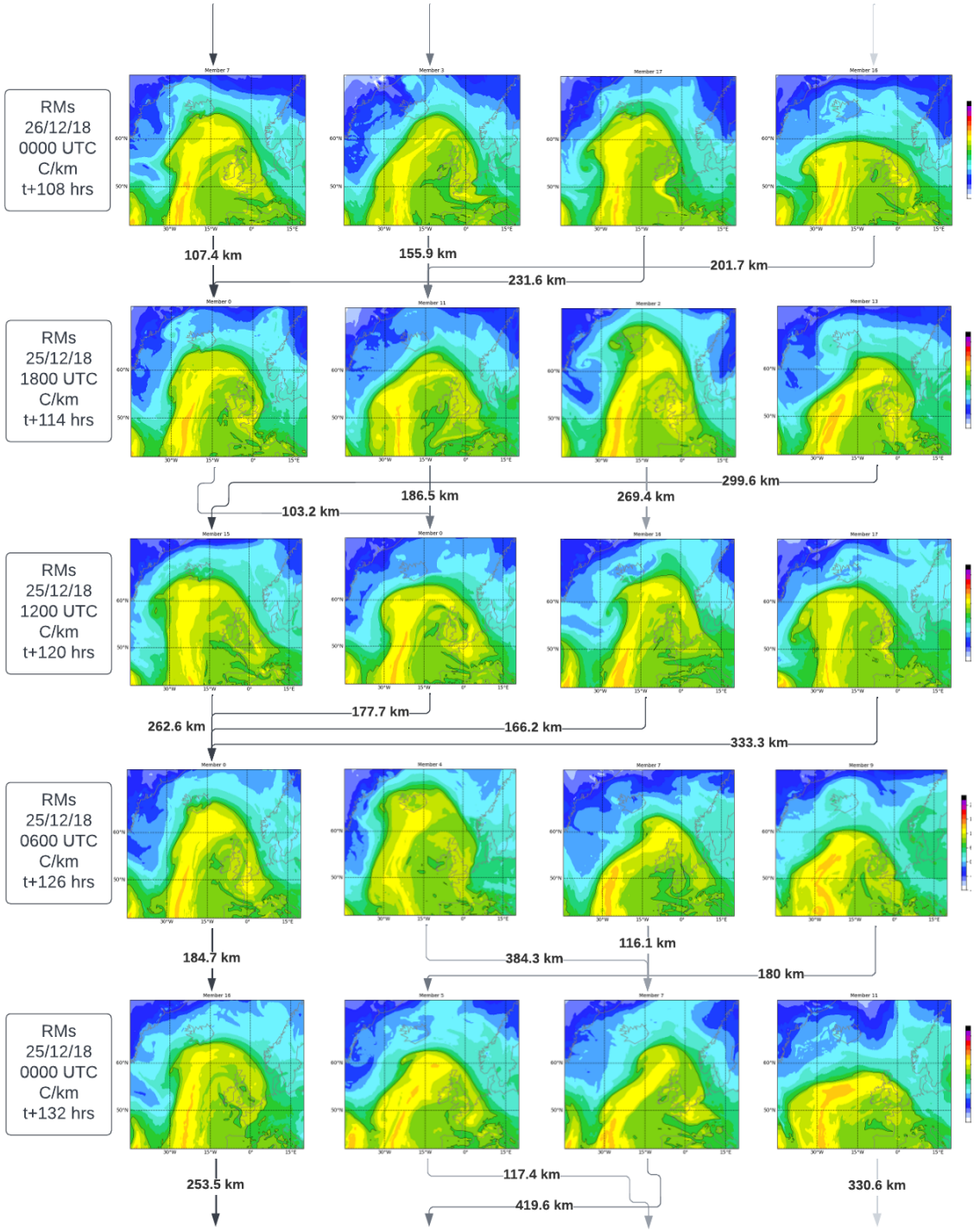


Figure 4.23: The progression of θ_w in C° at 850 hPa scenarios across forecasts with the valid time of 1200 UTC 30/12/2018, where each consecutive row is a preceding forecast with increasing lead time. The forecasts begin with 0000 UTC 26/12/2018 and go to 0000 UTC 25/12/18 (labeled by the the boxes on the left). The numbers below each plot indicate the FSS distance in km between the two RMs connected by arrows, which are the closest matches between forecasts

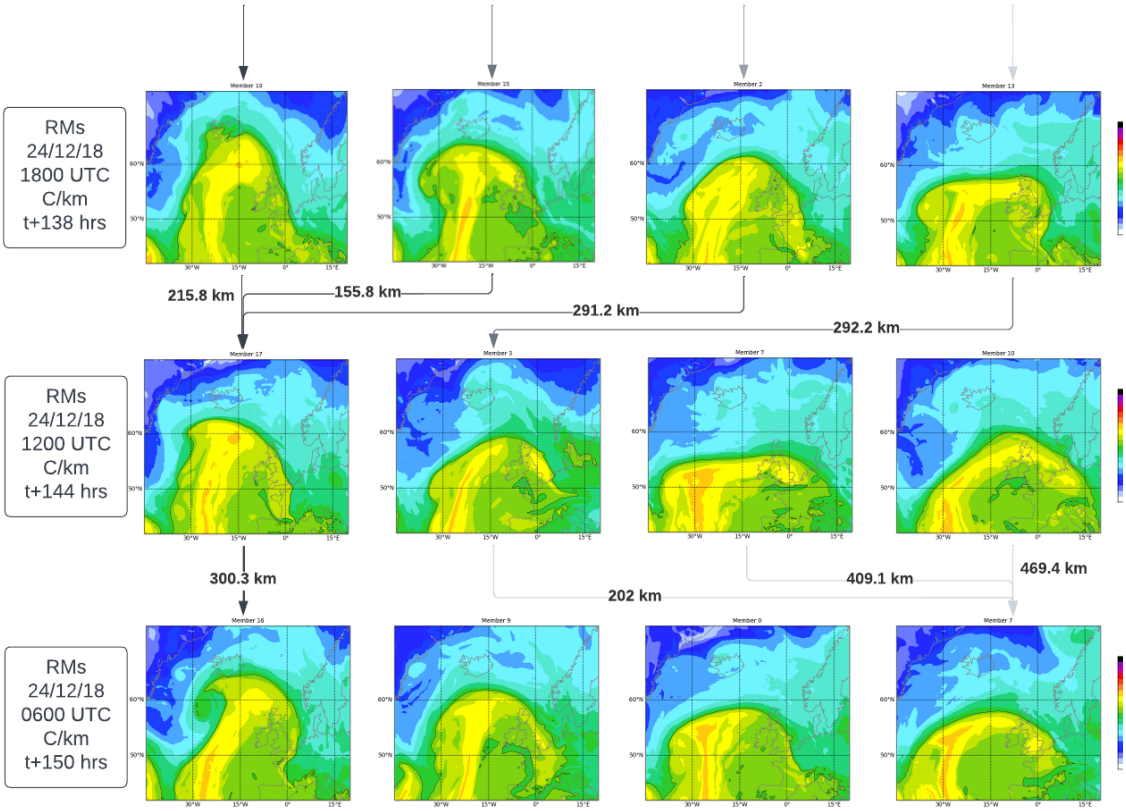


Figure 4.24: The progression of θ_w in C° at 850 hPa scenarios across forecasts with the valid time of 1200 UTC 30/12/2018, where each consecutive row is a preceding forecast with increasing lead time. The forecasts begin with 1800 UTC 24/12/2018 and go to 0600 UTC 24/12/18 (labeled by the the boxes on the left). The numbers below each plot indicate the FSS distance in km between the two RMs connected by arrows, which are the closest matches between forecasts

feature that curves around the southeast of the UK. Examining the forecasts at 0600 UTC 28/12/2018 (t+54) to 1200 UTC 27/12/2018 (t+72) it is clear that the closer the forecasts are to the observation the more similar the atmospheric patterns are to the observation, with a deeper meridional warm sector in the west of the plume, a domed northern edge, and the hook-like feature in the east. This is expected behaviour due to the reduction in spread leading to some scenarios dropping out as the event approaches. As lead time increases, spread will also increase and the older forecasts will exhibit multiple scenarios, which can be seen as later figures are explored. Further continuing to figure 4.22, with forecasts from 0600 UTC 27/12/2018 (t+78) to 0600 UTC 26/12/18 (t+102), shows that earlier and earlier forecasts had steadily more zonally structured and dome shaped plumes in their leading northern edge. At 1200 UTC 26/12/2018 (t+96) some RMS have plumes that are sharply slanted. In figure 4.23 are the RMs for the forecasts of 0000 UTC 26/12/2018 (t+108) to 0000 UTC 25/12/2018 (t+132). Progressing through these forecasts shows three common patterns: strongly meridional warm sector plumes, plumes with a strong zonal component, and plumes that have a more rounded northern shape. Finally, in figure 4.24, with forecasts from 1800 UTC 24/12/2018 (t+138) to 0600 UTC 24/12/2018 (t+150), the first forecasts that picked out this uncertain event are seen. These forecasts indicate there will be some kind of warm plume event occurring, though its structure is very different between RMs, ranging from a more meridional warm sector dominant in the western half of the plume to a very zonal warm sector that cuts across the UK. As expected, the oldest forecast RMs from the 0600 UTC 24/12/2018 (t+150) forecast look very different to the observation.

It is expected that the RM closest to the observations within each forecast at the valid time will be identified in all preceding forecasts and be traceable. This is indeed the case, where every RM in the left column has both at least one later forecast RM leading to it and one forecast RM it leads to. When a preceding forecast has no arrow connecting it to the later forecast (in the previous row) that scenario can be considered discontinued. An example is the last RM in forecast 0000 UTC 28/12/2018 (t+60) (figure 4.21), where no arrow connects to it from the later forecast 0600 UTC 28/12/2018 (t+54). However, this particular scenario can also be traced back through the forecasts to the very first forecast that picked up the uncertainty in this event (0600 UTC 24/12/2018, t+150, figure 4.24). This behaviour can be seen on all such RMs throughout the case, indicating scenarios can

indeed be traced through consecutive forecasts and determined when they are no longer plausible solutions. It is also expected that the RMs with the shortest distances between them will also be closest to the observation (i.e. in the left most column). This is definitely the case in the forecasts in figure 4.21, where all forecasts except 1800 UTC 27/12/2018 have RMs matching with the shortest distance to the preceding forecast that is closest to the observations. However, this trend does not carry through in earlier forecasts. Because this example has forced the number of clusters to four, it is likely that there will be some forecasts that have more scenarios than necessary (2 or 3 optimal clusters) and some forecasts that have fewer scenarios than the optimal number (5 or 6). However, the premise of scenarios appearing or disappearing as the valid time approaches will apply to both situations of forcing four clusters or allowing the algorithm to pick the optimal number of clusters.

4.5.3 Probability of Scenarios

The next step is to consider the probability of scenarios. Probability of a scenario may depend on many factors, but in terms of clustering it may be tied to clusters that have the most members. Table 4.1 contains columns of the lead times associated with the valid time, the RMs in order as they appear in figures 4.21 to 4.24 where the left most RM is associated with the shortest distance to the observation, the number of members associated with each RM, the distance of the RMs to the observation, and the distances of RMs between forecasts. Hypothetically, following the closest RMs across forecasts at the same valid time back from the observation will result in these RMs also being the closest to the observation and being the cluster with the most members. As seen in the previous section (4.5.2), the closest RMs to the observations were not always the closest RMs between forecasts. However, this did occur in later forecasts with shorter lead times. With regards to membership, a similar pattern can be seen. In table 4.1 the majority of lead times have the highest cluster membership count (or tied for the highest in some cases) associated with the RM closest to the observation. The exceptions are t+72, where the highest membership was associated with the RM that was third closest to the observations but was only one higher than the closest RM, and t+90 where the second and third closest RMs tied for the highest member count. In table 4.1 with much earlier

forecasts and later lead times, only two lead times have RMs that are both associated with the highest membership count and are the closest to the observations (t+126 and t+144). In total, 8 out of 17 forecasts had the highest member count associated with the closest RM to the observations, 5 forecasts had the highest member count associated with the RM that was second closest, and 4 forecasts had the highest member count associated with the third closest. No forecasts had the highest member count associated with the RM furthest from the observation.

The examination of the membership of the clusters provides mixed results, however higher membership closer to the observations does seem to indicate higher likelihood of the scenario resulting in the observations. This is expected due to the nature of reduced lead time indicating lower likelihood of divergence between members. A further point of importance is how often the control member, 0, is the closest to the observations. Not only does it appear as a RM 12 out of 17 lead times, it is the closest to the observations 6 of those times, most of which are shorter lead times. It is the second closest to the observations 3 times. The control member is expected to be the most likely member in an ensemble as it typically is not perturbed in any way and the appearance of the control member regularly as an RM and often the closest or near to the closest in observations supports this. There doesn't appear to be a relationship between the number of members and the shorter distances between RMs in consecutive forecasts. However, restricting the number of clusters to 4 may affect this outcome.

Further study of how scenarios can be traced across forecasts at the same valid time is needed, particularly to address the probability of scenarios. This can be achieved by comparing representative members produced by the clustering method in subsequent forecasts, such as in figures 4.21 to 4.24. As clustering can be performed with every new forecast, a series of statistics could be produced that compared membership between clusters at the same valid time at different forecasts, picking up trends that may point to a more probable scenario emerging.

4.6 Conclusion

In this chapter, experiments have been performed with operational global forecast data and determined a method and appropriate parameter ($|\nabla\theta_w|$) settings to identify

clustering behaviour and diagnose the beginning of the window of interest. If we pick a small number of clusters it is expected that there is a time window of interest when the ensemble is better described by clusters than a single distribution. After the window of interest, the ensemble members diverge further away from one another and no longer fit within a few clusters. This is related to the sum distance, which is not expected to decrease linearly over the forecast but to fluctuate after reaching its minimum value when clustering is most distinct. The algorithm allows for the freedom of members to move between clusters as necessary by clustering at individual lead times, which allows for more robust clusters. By comparing clusters across lead times, they can also be traced through a forecast. Once the clusters have been linked across lead time and when the clustering is near its most defined within a forecast, i.e. within the window of interest, the method produces distinct representative members that can then be presented to forecasters as potential scenarios. Highly unpredictable events can also be traced across forecasts with the same valid times, which may be beneficial in drawing operational meteorologist attention. Cluster membership has also been shown to potentially be tired to forecast probability, but requires more case studies to be verified.

Lead Time	*RM*	*Members*	*Distance Obs*	*Distance* RM_0	*Distance* RM_1	*Distance* RM_2	*Distance* RM_3
t+54	0	11	74.7				
	8	2	116.9				
	15	3	118.6				
	1	2	141.5				
t+60	0	10	101.1	57.5_{0}	124.5_{8}	181.9_{15}	76.5_{1}
	17	3	116.8	92.4_{0}	147.9_{8}	165.4_{15}	109.1_{1}
	15	1	146.8	137.4_{0}	120.7_{8}	195.3_{15}	183.5_{1}
	7	4	178.9	143.4_{0}	179.6_{8}	215.7_{15}	139.2_{1}
t+66	0	10	99.8	62.3_{0}	97.9_{17}	164.1_{15}	177.8_{7}
	8	4	112.4	150_{0}	178_{17}	150.5_{15}	204.2_{7}
	2	2	134.4	172.9_{0}	134.9_{17}	205.1_{15}	231.9_{7}
	1	2	175.4	158.4_{0}	99.5_{17}	184.4_{15}	153.5_{7}
t+72	0	6	132.4	140.8_{0}	125.1_{8}	220.7_{2}	208.2_{1}
	9	1	152.6	131.4_{0}	181.9_{8}	170_{2}	241.3_{1}
	13	7	251.2	233.8_{0}	227.8_{8}	287.7_{2}	173.9_{1}
	12	4	319.7	247.2_{0}	312.3_{8}	270.1_{2}	151.7_{1}
t+78	5	7	98.8	168.8_{0}	176.4_{9}	270_{13}	301.5_{12}
	13	3	204.3	260.5_{0}	262.9_{9}	266.9_{13}	212.7_{12}
	0	3	220	119.3_{0}	277.4_{9}	151.2_{13}	223.1_{12}
	12	5	337	299.1_{0}	368.8_{9}	195.3_{13}	96.4_{12}
t+84	9	6	172.5	218.1_{5}	261.8_{13}	255.9_{0}	401_{12}
	0	3	222.3	247.1_{5}	295.2_{13}	120.2_{0}	247.3_{12}
	16	6	346.4	370_{5}	278.6_{13}	235.2_{0}	106.5_{12}
	4	3	352.6	367.8_{5}	185.8_{13}	381.7_{0}	185.1_{12}
t+90	10	3	178.4	222_{9}	201_{0}	362.4_{16}	415.6_{4}
	12	7	259.4	348.8_{9}	252.7_{0}	117.7_{16}	123.3_{4}
	0	7	282.8	293.5_{9}	126.6_{0}	204.5_{16}	307.6_{4}
	14	1	315.3	369.2_{9}	185.2_{0}	103.5_{16}	248.5_{4}
t+96	2	6	245.2	156.4_{10}	415.5_{12}	321.9_{0}	366.4_{14}
	6	5	245.9	165_{10}	245.8_{12}	97.1_{0}	163.5_{14}
	16	6	457.3	378.3_{10}	260.7_{12}	226.5_{0}	145.2_{14}
	4	1	524.9	499.7_{10}	462.3_{12}	498.6_{0}	514.8_{14}

Table 4.1: Cluster membership during the 1200 UTC 30/12/2018 valid time case, where the Lead Time column is the matching forecast lead time for the valid time, the RM column lists the four representative members associated with that lead time and forecast, the Members column is how many members the associated cluster of the RM contained at that lead time, the Distance Obs column is the FSS distance in km of that RM compared to the control member of the 1200 UTC 30/12/2018 forecast at t+0, and the Distance RM columns are the FSS distances in km between the RMs of that forecast and the RMs of the RMs of the row above (a later forecast with a corresponding earlier lead time), denoted by subscript. The distances were calculated using the threshold associated with their respective forecast before the members were reduced to binary fields, i.e. RMs from lead time A will have the threshold from forecast A applied and RMs from lead time B will have the threshold from forecast B applied. As these threshold were calculated by the same percentage, they will result in a similar sized frontal regions present in the forecasts for better comparison.

Lead Time	*RM*	*Members*	*Distance Obs*	*Distance* RM_0	*Distance* RM_1	*Distance* RM_2	*Distance* RM_3
t+102	5	2	185.9	241_{2}	316_{6}	493.9_{16}	453.7_{4}
	0	5	220.8	255.8_{2}	95.4_{6}	271_{16}	596.8_{4}
	12	7	311.5	408.8_{2}	354.5_{6}	442.1_{16}	412_{4}
	7	4	401.5	476_{2}	266.5_{6}	178_{16}	540_{4}
t+108	7	2	148.5	231.6_{5}	173.8_{0}	352.4_{12}	344.6_{7}
	3	9	166.7	226.2_{5}	276.2_{0}	264.2_{12}	289.6_{7}
	17	4	197.1	274.5_{5}	244.9_{0}	440.5_{12}	446.8_{7}
	16	3	364.6	425.7_{5}	271.6_{0}	376.4_{12}	96.1_{7}
t+114	0	5	131.8	107.4_{7}	189.7_{3}	231.6_{17}	342.8_{16}
	11	7	264.4	193.4_{7}	155.9_{3}	404_{17}	201.7_{16}
	2	2	281.5	321.2_{7}	294_{3}	338.4_{17}	447.3_{16}
	13	4	343.7	325.1_{7}	266.6_{3}	418.7_{17}	321.8_{16}
t+120	15	4	174.7	204.6_{0}	336.4_{11}	318.2_{2}	299.6_{13}
	0	7	190.6	103.2_{0}	186.5_{11}	372.1_{2}	320.2_{13}
	16	4	198.2	249_{0}	221.1_{11}	269.4_{2}	393.1_{13}
	17	3	267.5	240.8_{0}	305.6_{11}	416.4_{2}	398.1_{13}
t+126	0	8	156.8	262.6_{15}	177.7_{0}	166.2_{16}	333.3_{17}
	4	3	284	270.5_{15}	437.4_{0}	388.5_{16}	465.8_{17}
	7	4	359.5	330_{15}	361.9_{0}	383.3_{16}	480.9_{17}
	9	3	364	425.8_{15}	347.8_{0}	407.4_{16}	398.1_{17}
t+132	16	6	191	184.7_{0}	429.9_{4}	329_{7}	390_{9}
	5	8	310.7	226.6_{0}	459.4_{4}	265.5_{7}	180_{9}
	7	2	358	342.1_{0}	384.3_{4}	116.1_{7}	368.4_{9}
	11	2	470	455.4_{0}	523.2_{4}	418_{7}	340.7_{9}
t+138	10	5	125.9	253.5_{16}	313.9_{5}	383.3_{7}	432.6_{11}
	15	4	267.2	295.4_{16}	380.5_{5}	419.6_{7}	444.4_{11}
	2	8	351.3	379.2_{16}	117.4_{5}	315.6_{7}	367.5_{11}
	13	1	386.4	480.1_{16}	430.6_{5}	423.6_{7}	330.6_{11}
t+144	17	5	161.3	215.8_{10}	155.8_{15}	291.2_{2}	361.9_{13}
	3	4	342.7	326_{10}	415_{15}	303.2_{2}	292.2_{13}
	7	4	614.6	537.8_{10}	529.5_{15}	544.5_{2}	416.3_{13}
	10	5	667.6	529.1_{10}	602.5_{15}	559.3_{2}	583.7_{13}
t+150	16	3	210.9	300.3_{17}	291.8_{3}	620.2_{7}	573.5_{10}
	9	4	352.1	352_{17}	447.8_{3}	545.6_{7}	512.8_{10}
	0	6	402.1	320.6_{17}	228.1_{3}	434.7_{7}	609_{10}
	7	5	427.8	493.2_{17}	202_{3}	409.1_{7}	469.4_{10}

Table 4.1: (Cont.) Details of the table can be found on the previous page.

Chapter 5

Dependence of clustering on variable used to compare members

5.1 Introduction

In this chapter, results obtained by clustering using distance measures based on the gradient of the wet-bulb potential temperature and large-scale rain are compared. This is to determine if there is a significant difference in the clustering of these two related variables and if so which is the better choice for this method. The analysis is performed for October 2018 over the domain of 40° to 70° north and 45° west to 20° east. First, the chapter will explore results of clustering on large-scale rain and will cover the cluster robustness, how traceable the clusters are, and how distinct the representative members are. Then, the results of this analysis are compared to the clustering performed on $|\nabla\theta_w|$. This comparison will first go into depth about whether or not the method extracts the same scenarios from a different variable field. Then, it will cover how the spread relates between the variables. After which it will cover when the window of interest begins, how comparable the cluster membership is, and if the same representative members are extracted from the forecasts.

5.2 How clustering performs where the distance is measured using the field of large-scale rain rate versus $|\nabla\theta_w|$

The wet-bulb potential temperature at 850 hPa and the large-scale rain rate are fundamentally different in nature. Instead of a continuous field of values, the rain rate is patchy, often with a great deal of fine scale features. Where, when, and how much rain falls can be greatly dependant upon variables other than the temperature and orography for example, vertical motion including convection. However, precipitation is likely at fronts, often appearing near sharp gradients in the wet-bulb potential temperature. Therefore, it is reasonable to consider if clustering on precipitation fields can be related to clustering performed on the gradient of the wet-bulb potential temperature, if it can be used instead of $|\nabla\theta_w|$ for clustering, if it is directly related to fronts or if it is isolated to non-frontal features, or if it provides any novel information for the forecast. For an in-depth evaluation of clustering, it is important to consider the robustness of the clusters, their traceability, and how distinct the representative members are and if they form sufficiently distinct potential scenarios. These topics will be covered in the following sections.

5.2.1 Robustness

A robust cluster is one that has stable membership and has smaller intra-cluster distances than inter-cluster distances (i.e. the members within a cluster are closer to their own medoid than they are to any other external cluster members) and the medoids are further apart from each other than they are from their respective members. As the medoid is actually a member, this nearly eliminates the likelihood of an unstable cluster sometimes found within the k-means method as the mean becomes the centroid. During the k-means clustering process, the centroid is first chosen at random and the members are clustered to their closest centroid. A mean of the members in the cluster is then calculated and used as the new centroid to re-evaluate cluster membership. This process can allow for multiple solutions, depending on where the initial centroids are placed and how many

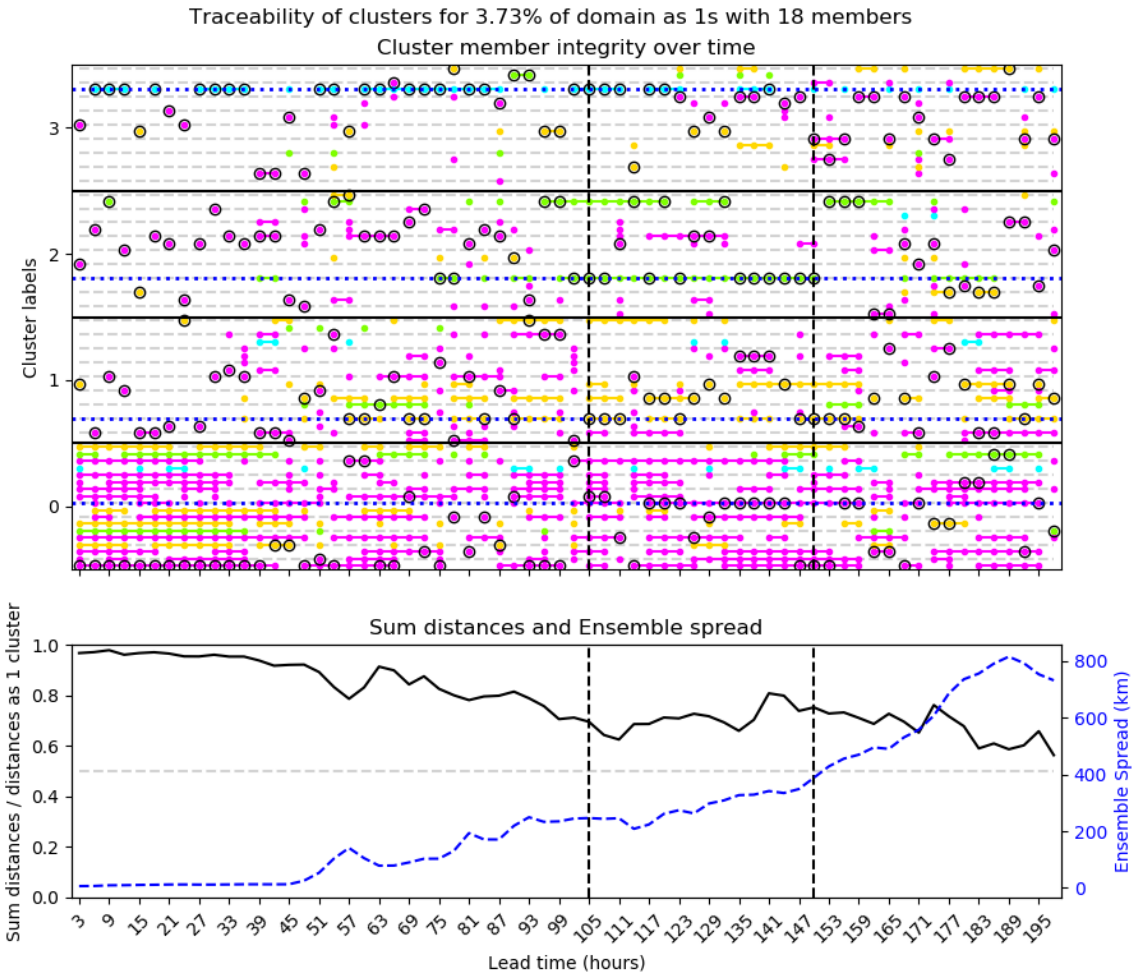


Figure 5.1: Clusters obtained using the large-scale rain rate for FSS distance (top), and the sum distances and the ensemble spread (bottom) versus the lead time in hours from the forecast on 06/10/2018 0000 UTC, over a domain of 40° to 70° north and 45° west to 20° east.

times the process of calculating a new centroid is repeated, if it isn't done to convergence.

Robustness can be examined by looking at how similar members are within a cluster at any given lead time via the sum distance, distance matrices, and paintball plots. The sum distance for the ensemble forecast run at 0000 UTC 06 October 2018 for the large-scale rain can be seen in figure 5.1, and for comparison the same plot can be seen for $|\nabla\theta_w|$ during the same forecast in figure 5.2. This forecast was chosen due to the mid-latitude cyclone it forecasts (seen in the later paintball and analysis figures 5.5 to 5.7), which has both significantly large frontal features and rain objects. For both variables, the lower the normalized sum distance, the stronger the variability in the clusters. When the sum distance decreases to a 75% reduction to its lowest value the window of interest (the vertical black dashed lines) begins. The window must be kept long enough to allow for meteorological events to progress but short enough to not lose cohesion of the clusters.

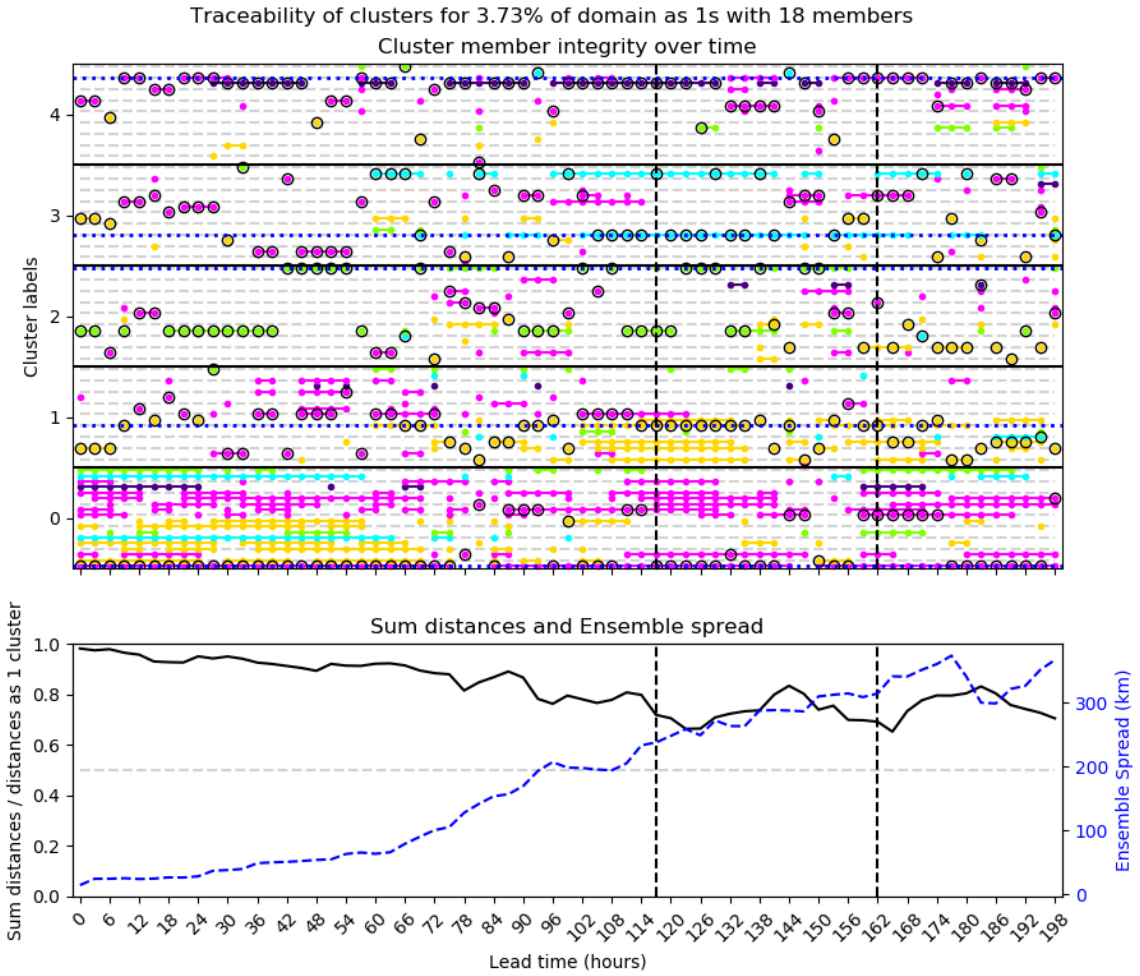
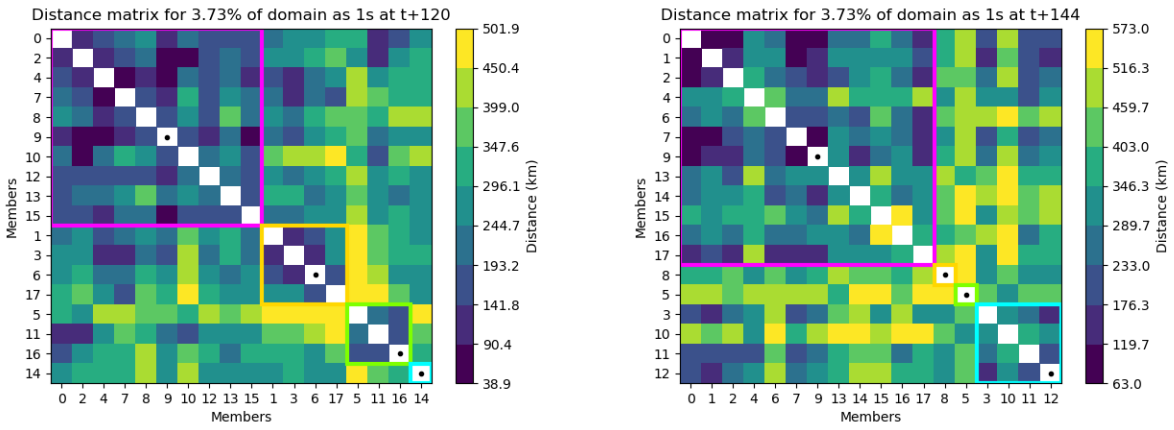


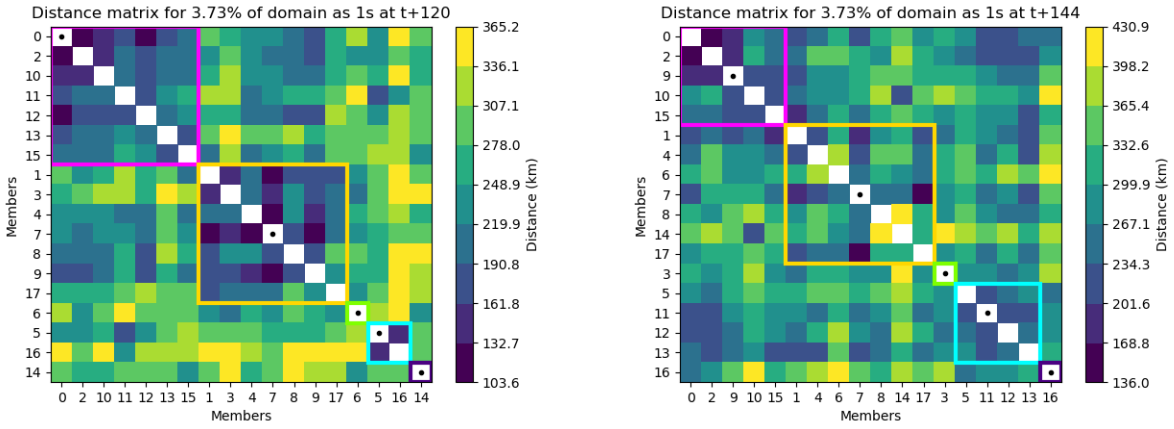
Figure 5.2: clusters obtained using $|\nabla\theta_w|$ for the FSS distance (top), and the sum distances and the ensemble spread (bottom) versus the lead time in hours from the forecast on 06/10/2018 0000 UTC, over a domain of 40° to 70° north and 45° west to 20° east.

This is where robustness is particularly strong, after members have evolved long enough to form one or more distinct scenarios. Figures 5.1 and 5.2 demonstrates this well, as cluster membership is stable within the window but fluctuates rapidly after it, indicating the the window is a particularly useful portion of the forecast for clustering.



(a) (b)

Figure 5.3: Distance matrices for the large-scale rain rate forecast beginning on 06/10/2018 at 0000 UTC at t+120 hours (a) and t+144 hours (b). The members are listed by number along the x and y axes, sorted into their corresponding clusters. The clusters are designated by the coloured boxes along the diagonals (magenta for cluster 0, gold for cluster 1, chartreuse for cluster 2, cyan for cluster 3). The black dots along the diagonals indicate the medoid of that cluster.



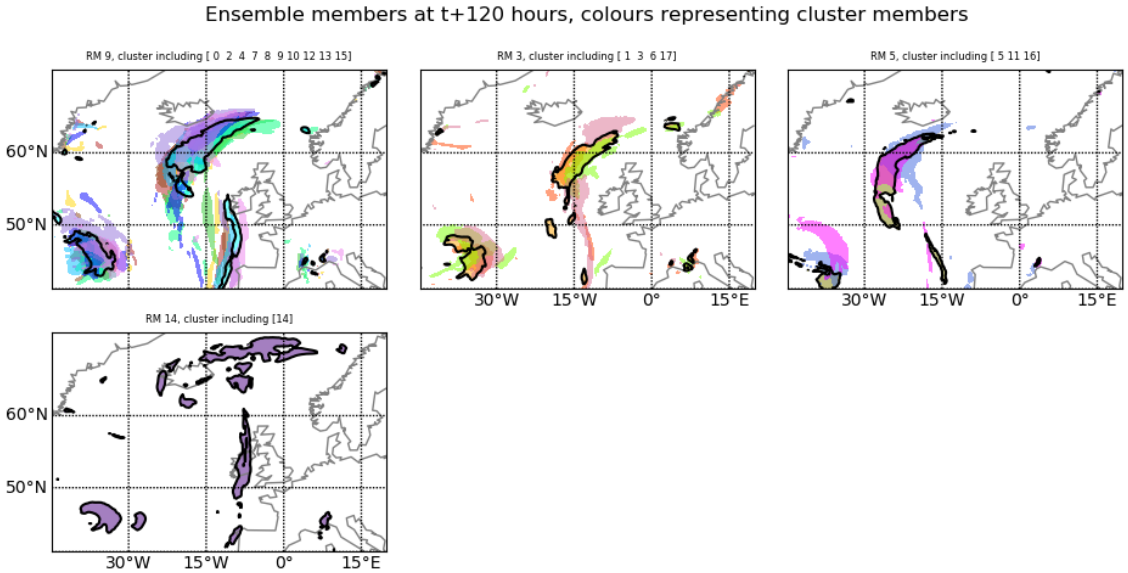
(a) (b)

Figure 5.4: Distance matrices for $|\nabla\theta_w|$ beginning on 06/10/2018 at 0000 UTC at t+120 hours (a) and t+144 hours (b). The members are listed by number along the x and y axes, sorted into their corresponding clusters. The clusters are designated by the coloured boxes along the diagonals (magenta for cluster 0, gold for cluster 1, chartreuse for cluster 2, cyan for cluster 3). The black dots along the diagonals indicate the medoid of that cluster.

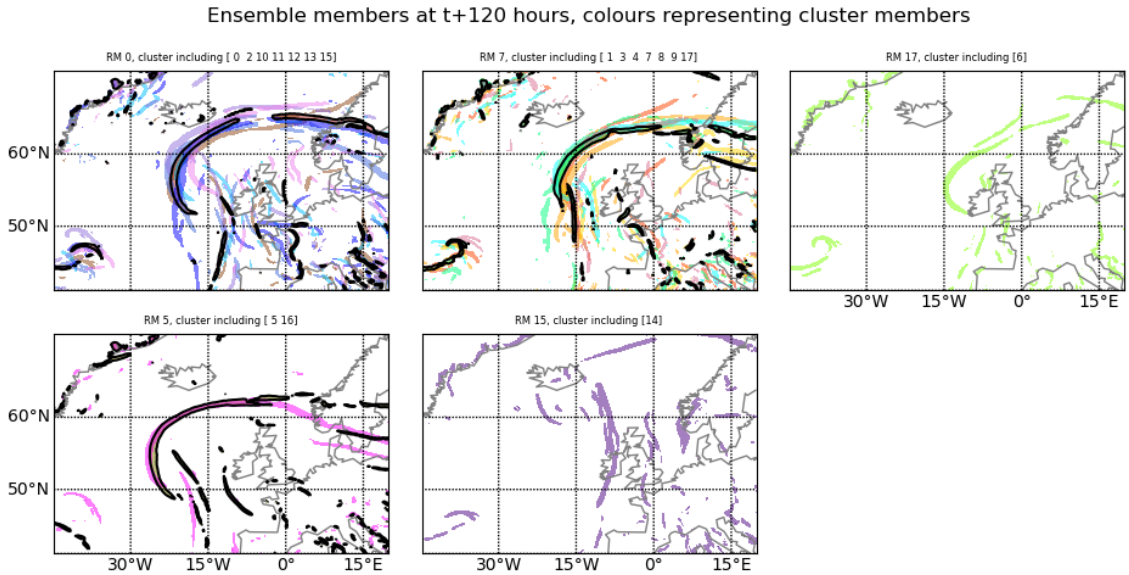
The distance matrices, seen in figures 5.3 and 5.4, show how close the members are to each other. The first plot (a) is within the window of interest for both $|\nabla\theta_w|$ and the

large-scale rain rate at t+120 hours and the second plot (b) is 24 hours later in the window at t+144 hours. An in-depth description of the distance matrix can be found in section 3.4.3. Both sets of distance matrices indicate the method is working as intended, with the distances between members and their medoid being closer than between members and other cluster medoids. However, it can be seen clearly in the distance matrix that cluster 0 is the primary cluster in the large-scale rain rate, whereas cluster 0 and 1 appear to be dominant clusters in $|\nabla\theta_w|$.

The paintball plots in figures 5.5 and 5.6 show maps of overlapping members of the rainfall rate and $|\nabla\theta_w|$ at t+120 and t+144 respectively. These were the optimal number of clusters chosen for each parameter for this particular forecast. Here can be seen what features primarily define the clusters and how well members actually match in terms of object shape, position, and size. Corresponding analysis charts for t+120 (0000 UTC 11/10/2018) and t+144 (0000 UTC 12/10/2018) are in figure 5.7. Beginning with the analysis chart corresponding to t+120 (figure 5.7a) shows a low pressure centre of a cyclone off the northwestern coast of Ireland with a cold front about to make landfall and the warm front off the southeastern Icelandic coast. In figure 5.5 the largest rain rate objects appear along the region of the warm front, with thinner rain rate bands appearing along the cold front. This relates well to the $|\nabla\theta_w|$ RMs in figure 5.5b where the warm front is the most prominent feature picked up by the method. Comparing the paintball plots of the two variables reveals that in general they cluster relatively similarly, even though the rain rate has four clusters and $|\nabla\theta_w|$ has five. The top three clusters of $|\nabla\theta_w|$ (clusters 0, 1, and 2) almost perfectly match the membership of the top left and center (0 and 1) clusters of the large-scale rain rate, with member 17 becoming a cluster to itself. The primary difference amongst the two variables is that the frontal objects in $|\nabla\theta_w|$ are grouped together based on the shape of the front itself, which is a long thin shape, whereas the rain rate objects are broad enough in shape that they overlap quite easily and therefore are not as clearly positioned as the front. However, for very strong differences in frontal position, there is a stronger distinction in how the rain rate clusters. Cluster 2 of the large-scale rain rate closely matches cluster 3 in $|\nabla\theta_w|$, with the exception of member 11 which is in cluster 0 in $|\nabla\theta_w|$. These clusters stand out because the frontal feature is much further west of the UK than any of the other clusters. Similar can be said of cluster 3 in large-scale rain rate and cluster 4 in $|\nabla\theta_w|$, where the cold front has already made

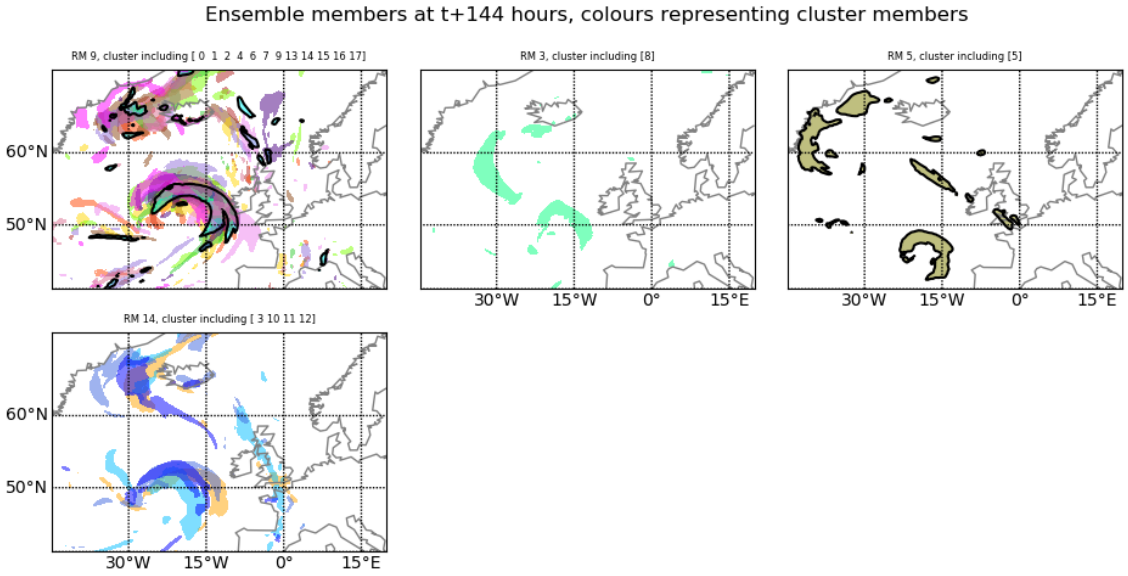


(a) Large-scale rain rate

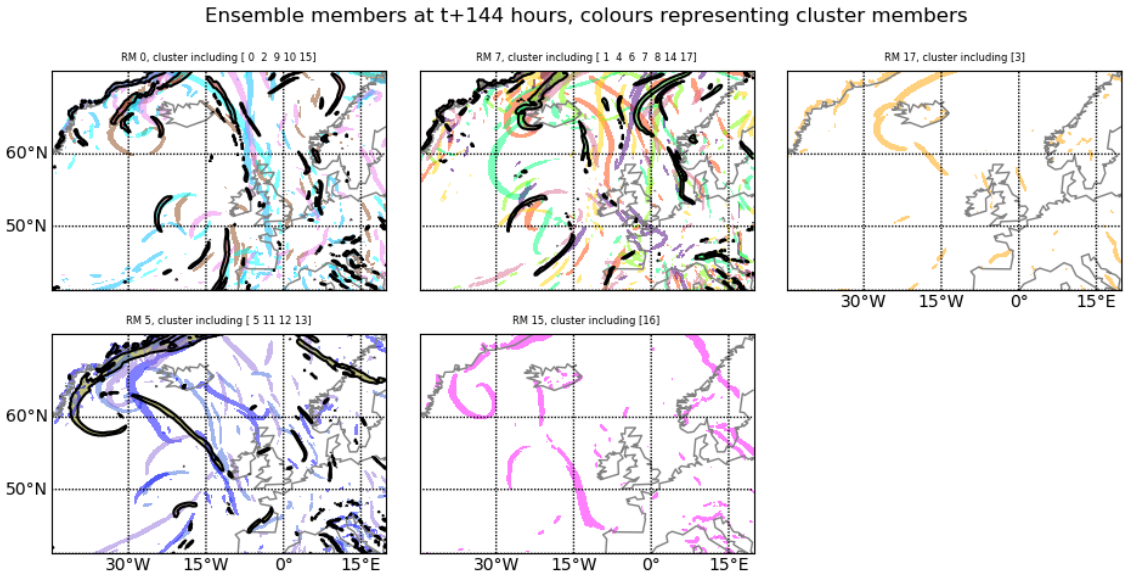


(b) $|\nabla\theta_w|$

Figure 5.5: Paintball plots for the forecast beginning on 06/10/2018 at 0000 UTC where (a) is from the large-scale rain and (b) is from $|\nabla\theta_w$ at 850 hPa. The provided plots are for the lead times over the window of interest. Each member is portrayed by a unique colour and the representative member (when present within the cluster) is signified by a black outline.

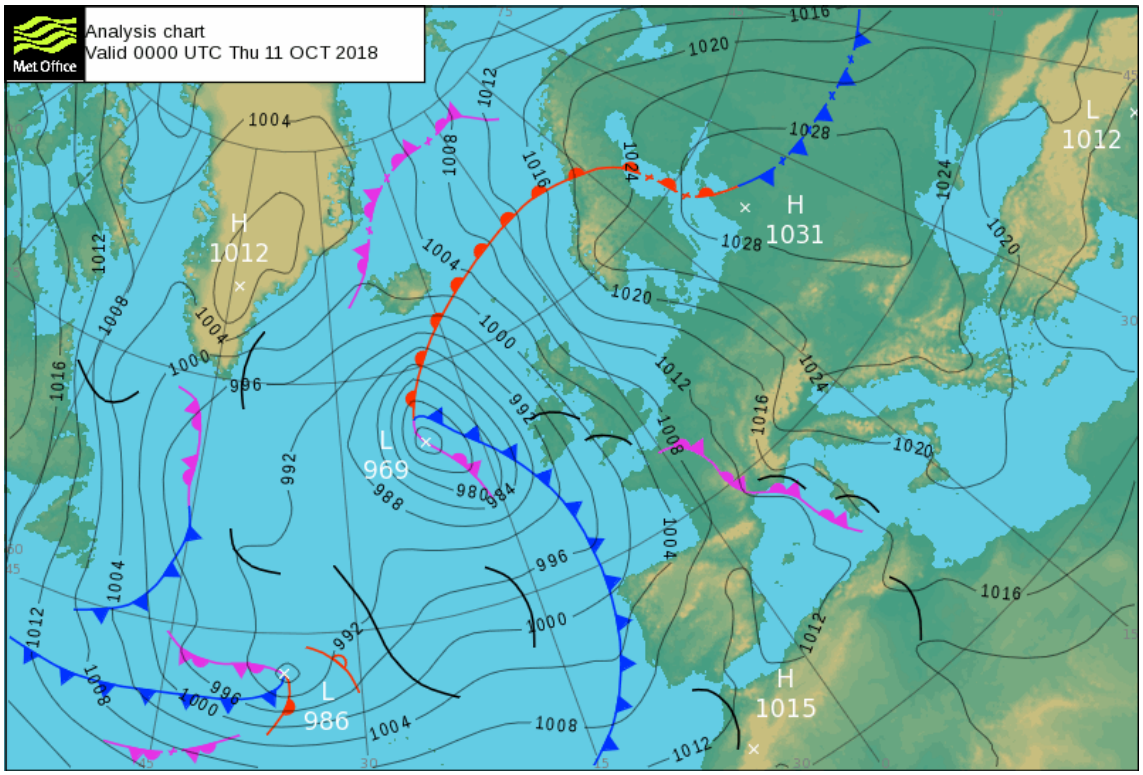


(a) Large-scale rain rate

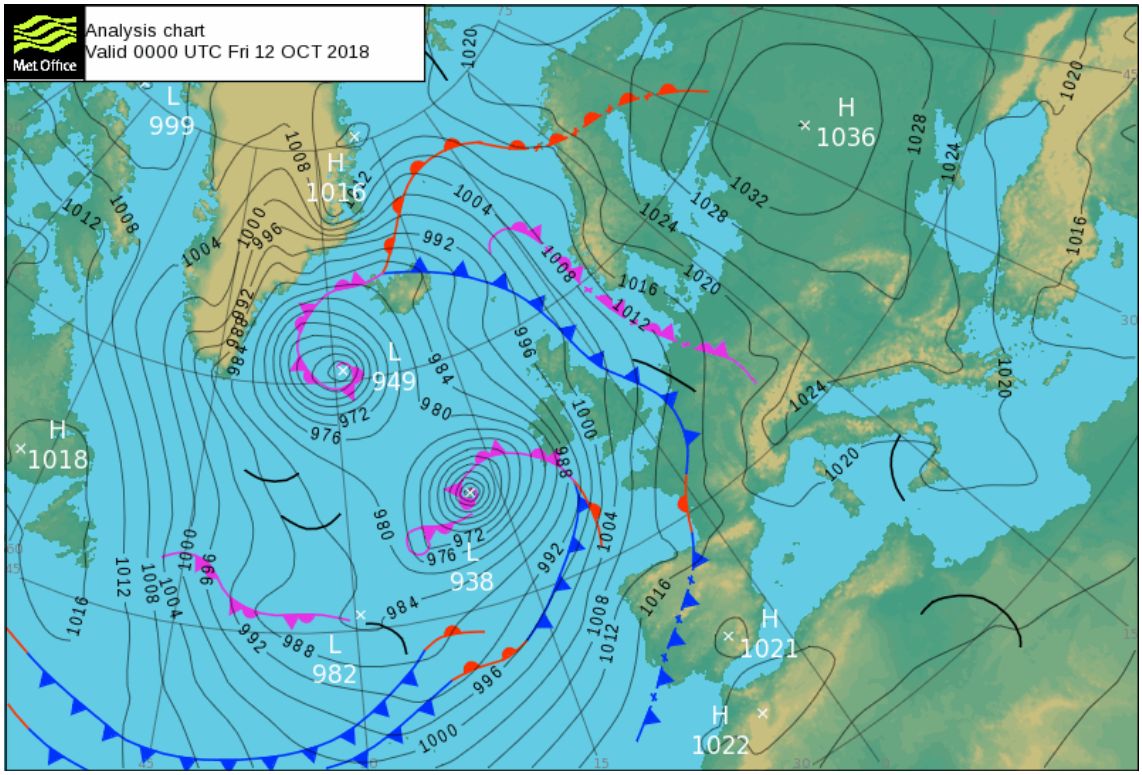


(b) $|\nabla\theta_w|$

Figure 5.6: Paintball plots for the forecast beginning on 06/10/2018 at 0000 UTC where (a) is from the large-scale rain and (b) is from $|\nabla\theta_w$ at 850 hPa. The provided plots are for the lead times over the window of interest. Each member is portrayed by a unique colour and the representative member (when present within the cluster) is signified by a black outline.



(a)



(b)

Figure 5.7: Analysis synoptic chart of 11/10/2018 at 0000 UTC and 12/10/2018 at 0000 UTC. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.

landfall in the UK and member 14 stands alone in both variables.

Later in the window of interest at t+144 the coherence of the clusters isn't as strong as members have moved further away from each other, which can be seen in figure 5.6, particularly in plot (b) where frontal objects don't line up as well as they did at t+120. The analysis chart (figure 5.7b) shows the decay of the storm in chart (a) and a secondary deeper low off the southwestern coast of Ireland. The rain rate objects associated with the original storm are predominantly still there at different distances between Greenland and Iceland, and the frontal objects are also still present. The reduced cohesion amongst the frontal objects is also reflected in the rain rate objects, where cluster members still overlap but are more spread apart from one another. The low off the southwestern Irish coast is not as prominent a feature amongst the frontal objects, but it does appear to be a significant feature amongst rain rate. Overall, there is still strong cohesion amongst the dominant features in both rain rate and $|\nabla\theta_w|$. By examining figures like 5.5 and 5.6, forecasters can see how similar members are within a cluster and how many members portray a particular feature. Seeing how the members overlap further supports the robustness of clustering, particularly within the window of interest.

5.2.2 Traceability

Similar to the $|\nabla\theta_w|$ clustering, large-scale rain rate cluster members are traceable across lead times, particularly within the window of interest where clustering is well defined. In figures 5.1 and 5.2 it is clear that most of the members remain close to the control early in the forecast, with the majority staying within cluster 0, even though the algorithm determined four clusters was the optimal choice for large-scale rain rate and five was the optimal choice for $|\nabla\theta_w|$. After 48 hours, cluster membership begins to be more evenly distributed, however there still isn't large variation between the clusters in the large-scale rain rate, evidenced by the still high sum distance value and the unstable cluster membership in figure 5.1. Alternatively, the $|\nabla\theta_w|$ in figure 5.2 has more stability in early lead times after members become more distributed. When clustering is examined within the window of interest, traceability is at its highest. Here, clustering is generally well defined and the representative members are, by design, chosen as the members that both remain in the cluster the longest and are the closest member to the centre of the

cluster throughout the window. It is only after the window when membership begins to appear more random as the members reach the end of their forecast lead time. However, this is expected due to the chaotic nature of the atmosphere. A notable difference in this case is that the spread of the $|\nabla\theta_w|$ members increases slower (reaching approximately 400 km) than the spread of the large-scale rain rate (reaching approximately 800 km). It's also important to note that the clusters appear more stable overall in $|\nabla\theta_w|$ than in large-scale rain rate and while the windows of interest do overlap for both variables, the window begins 12 hours later for $|\nabla\theta_w|$.

The traceability can be further seen and compared within figures 5.8 and 5.9. These figures contain the window of interest cluster inter-comparison diagrams from the ensemble forecast for 06 October 2018 at 0000 UTC with the large-scale rain rate (from t+105 to t+150) in figure 5.8 and $|\nabla\theta_w|$ (from t+117 to t+159) in figure 5.9. Details of these plots can be found in section 3.3.3. For both variables there are strong matches between clusters among the first half of the lead times within the window of interest. Towards the end of the window the high number of matches begins to tapper off, echoing the divergence of the state of the atmosphere across members in later lead times. Despite the difference of four versus five clusters for the respective variables, how the traceability behaves over the window of interest is very similar between them. Overall, these are another way to visualize traceability of clusters.

5.2.3 Variation in representative members

One of the key goals of this methodology is to provide forecast scenarios via representative members from the clusters. This means that representatives must be suitably distinct from one another. Examples of representative members from the large-scale rain rate and the wet-bulb potential temperature θ_w can be seen in figures 5.10 and 5.11 at t+120 and t+144, respectively.

In plot (a) of figure 5.10, each representative member is unique, showing four different examples of a large rain rate object near Iceland and a band of rain affecting the UK in at different times and intensities. Similarly, the wet-bulb potential temperature plots in (b) show a similarly shaped air mass but with significantly different positions of strong gradient regions. Advancing to t+144 in figure 5.11, all RMS are once again distinct, with

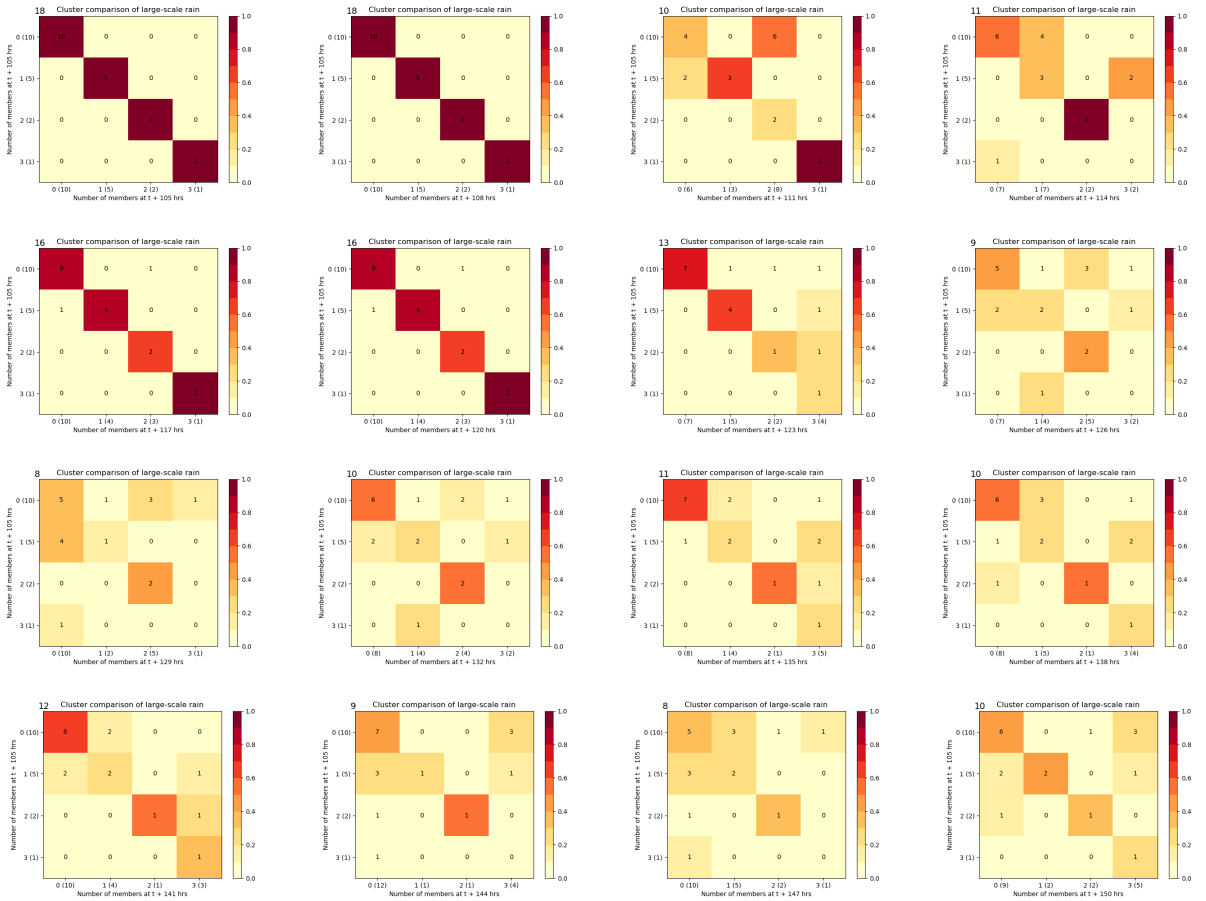


Figure 5.8: Cluster inter-comparison diagrams for the large-scale rain rate from 06/10/2018 at 0000 UTC between t+105 and t+150 hours (within the window of interest), where the y-axis describes the clusters at the beginning of the window of interest (i.e. 0 is the cluster label and (6) is the number of members within that cluster), the x-axis describes the clusters at various lead times, the colour bar represents the Jaccard Index, and the number at the top left of the chart indicates the sum along the diagonal, where 18 indicates a perfect match.

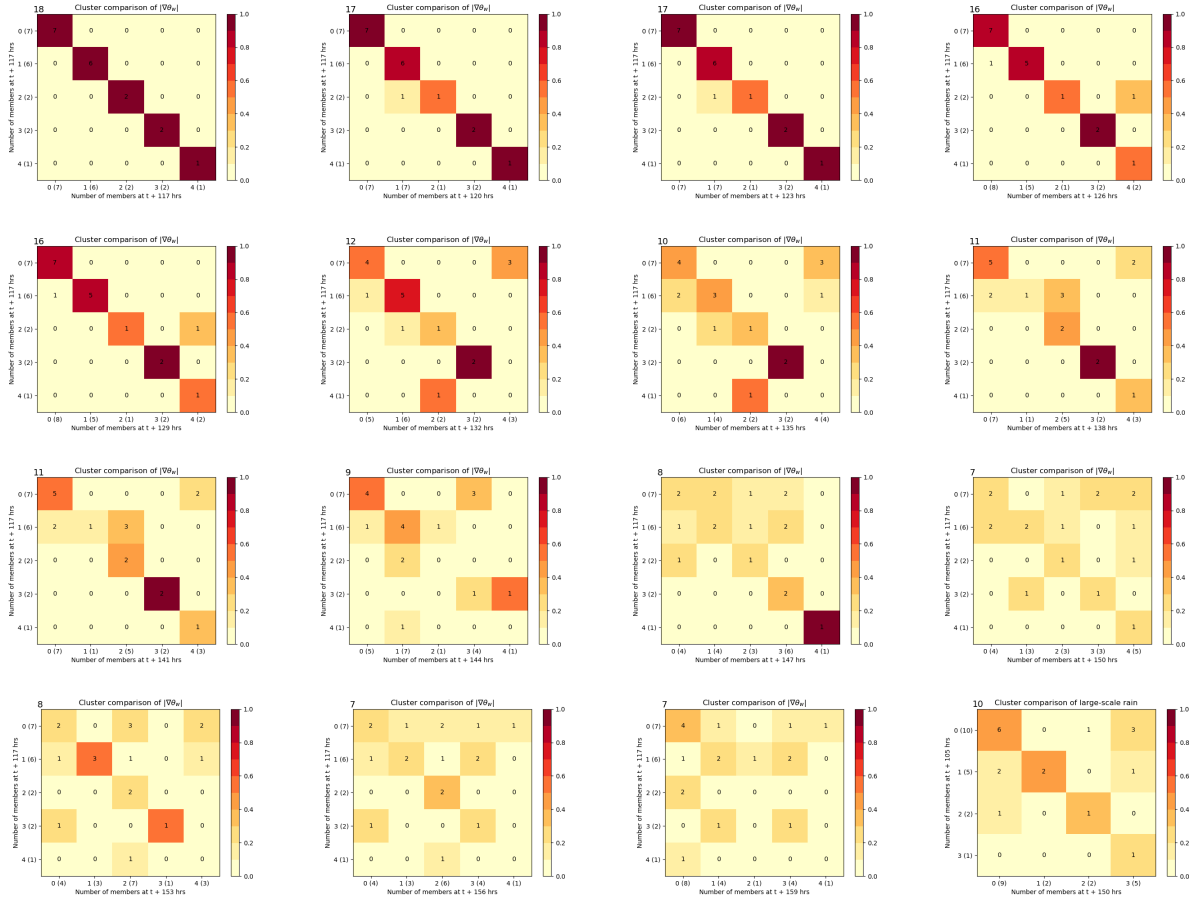
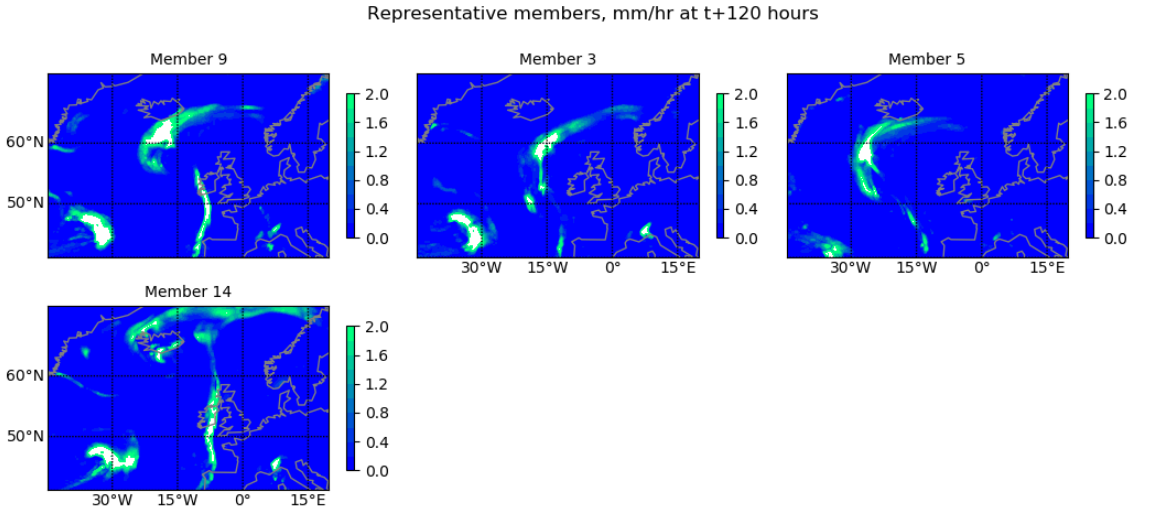


Figure 5.9: Cluster inter-comparison diagrams of $|\nabla\theta_w|$ from 06/10/2018 at 0000 UTC between t+117 and t+159 hours (within the window of interest), where the y-axis describes the clusters at the beginning of the window of interest (i.e. 0 is the cluster label and (6) is the number of members within that cluster), the x-axis describes the clusters at various lead times, the colour bar represents the Jaccard Index, and the number at the top left of the chart indicates the sum along the diagonal, where 18 indicates a perfect match.

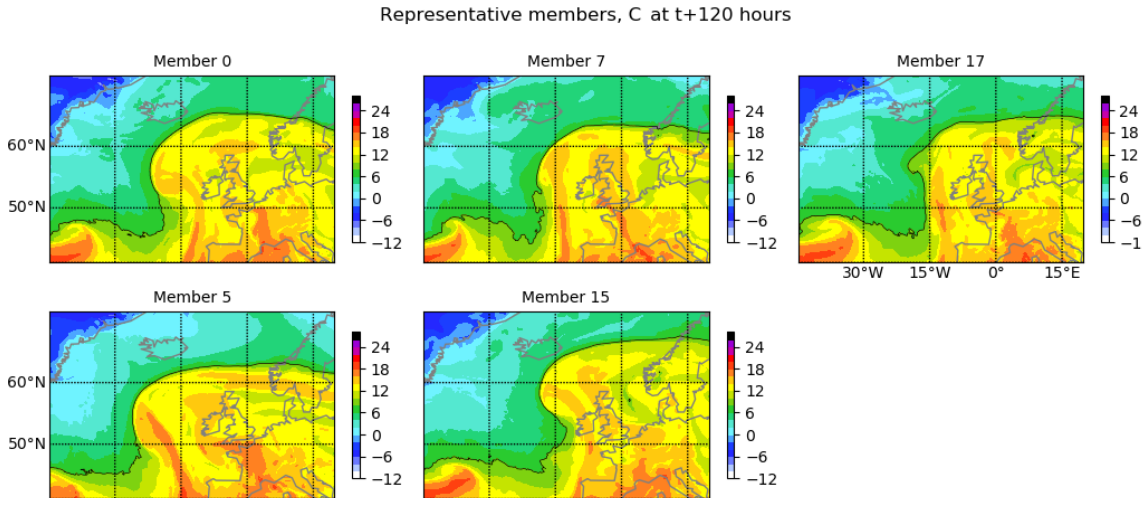
the rain rate reflecting the wave that appears in the θ_w field. Each rain RM contains a large rain object at different locations relative to the British Isles, some of which are already impacting Ireland. The wave depicted in θ_w RMs appear at different stages of progression, and dominate the field. It's clear that there is distinction between representative members with both variables. However, this is just a sampling of cases, and there likely will be forecasts in which there is very little variation amongst members to the point that there are only a small number of clusters and the variation between representative members does not require separate forecast scenarios even if the representative members are still distinct. This may happen in particular during highly predictable events such as slow moving high pressure systems or blocking events that may linger for several days, or when the window of interest begins very early in the forecast, where the uncertainty could be focused on finer details of a forecast. But how the algorithm is designed maximises the likelihood that this will be a rare occurrence. By specifically seeking out the lead times where there is strong clustering, by nature the algorithm is highly likely to provide suitably distinct representative members, as the example plots in figures 5.10 and 5.11 demonstrate.

5.3 How the large-scale rain rate and the gradient of the wet-bulb potential temperature forecasts compare

As previously stated, the wet-bulb potential temperature and the large-scale rain rate are two related but very different fields. Examining a single case can provide perspective as to how the two variables might relate to one another, but to provide a full understanding of each variables' strengths and weaknesses they must be examined over longer time periods and many cases. To compare them in depth, clustering has been restricted to 4 clusters. This restriction was chosen as the middle ground between 2 and 6 clusters and will result in less than optimal clustering of some forecasts, but is a necessary step to compare meteorological variables. The less than optimal clustering may result in too many clusters, which leads to lower traceability as two or more clusters may be very similar, or too few clusters, where one or more clusters contain more than one distinct scenario. However, as both variables are experiencing the same cluster restriction, it can

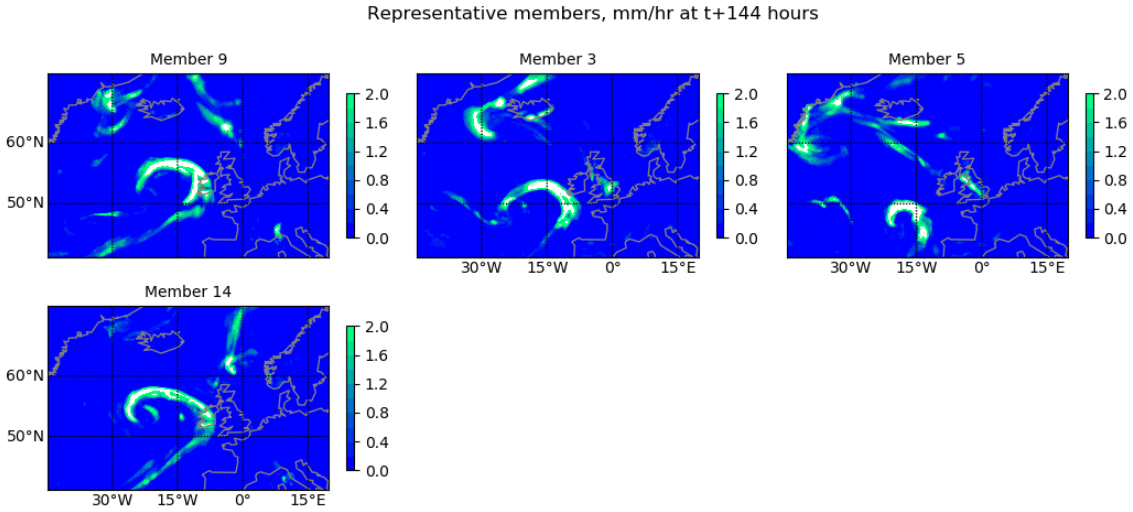


(a)

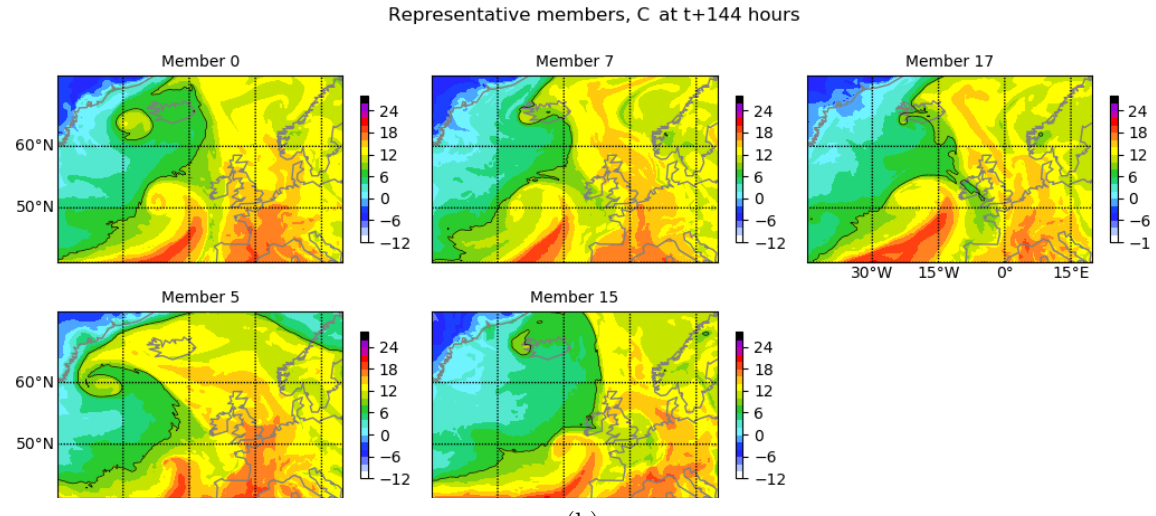


(b)

Figure 5.10: Representative members of (a) the large-scale rain rate and (b) θ_w at 850 hPa from the 0000 UTC 06 October forecast.



(a)



(b)

Figure 5.11: Representative members of (a) the large-scale rain rate and (b) θ_w at 850 hPa from the 0000 UTC 06 October forecast.

be anticipated that they will experience the same general results in clustering.

To begin the analysis, the summary plots for October (figure 5.12 for $|\nabla\theta_w|$ and 5.13 for the large-scale rain rate) can be compared for the two variables. The ensemble spread is higher for the distance measure based on the FSS comparing rain rate forecasts than for $|\nabla\theta_w|$, indicating a higher variation among members in general. There is also a steeper drop in the sum distance of the large-scale rain rate at the window of interest than there is in $|\nabla\theta_w|$ across lead times, noted by the lighter blues in the bottom plot of figure 5.12 and the deeper blues in the bottom plot of 5.13. This implies clustering of the large-scale rain rate is stronger than that of $|\nabla\theta_w|$. One feature of the $|\nabla\theta_w|$ summary plot mentioned in a previous chapter was how the start of the window of interest tends to start at the same valid time for multiple consecutive forecasts, indicating that the algorithm picked up on a single event dominating the uncertainty in forecasts (i.e. various storms in October 2018). In the rain rate summary plot, a similar pattern is seen, although the number of forecasts over which the valid times match is smaller. There is an approximately five day cyclic pattern of when the window of interest starts and some notable instances when the drop in sum distances occurs at the same valid time across many forecasts with different start dates. The nature of the FSS requires fields to be converted to binary, which results in the primary measure between members to be spatial. But in the process of applying a threshold to a field such as large-scale rain two similarly shaped rain objects could produce two very different binary fields if their intensity differed significantly. Some of this can be answered with further analysis of how the clustering matches up between variables. In the below sections, how the spread compares between the variables, if their cluster membership is similar, how the windows of interest start times line up between them, and how the representative members match between the variables will be examined.

5.3.1 Spread

Continuing from the analysis of the summary plots, the variation in the spread of the two variables can be seen in figure 5.14 where each dot represents the spread of the two variables at a matching forecast date and time in October 2018. In this plot, it's clear the spread, derived from the FSS distances in km, is greater for the rain rate than it is for $|\nabla\theta_w|$. They are however well correlated, which is expected since both variables are

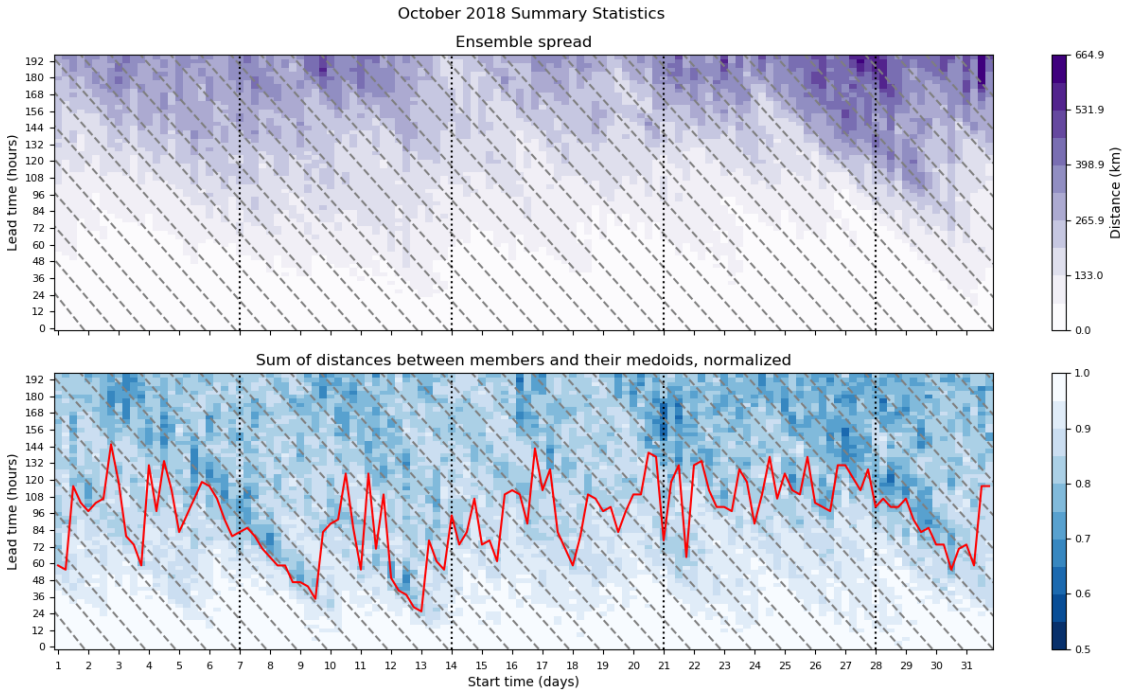


Figure 5.12: A "predictability plot" based on comparison of $|\nabla\theta_w|$ between ensemble members for the month of October 2018. The top plot is the ensemble spread calculated as the average FSS distance. The middle plot is the sum of within cluster distances, normalized by the sum distances from the medoid of the whole ensemble. The red line marks where the window of interest (the point at which the sum distance has decreased to the 25th percentile) begins. The forecast start time is denoted on the x axis. The vertical black dotted lines mark every seven days. The diagonal dashed lines link the same verification times across forecasts.

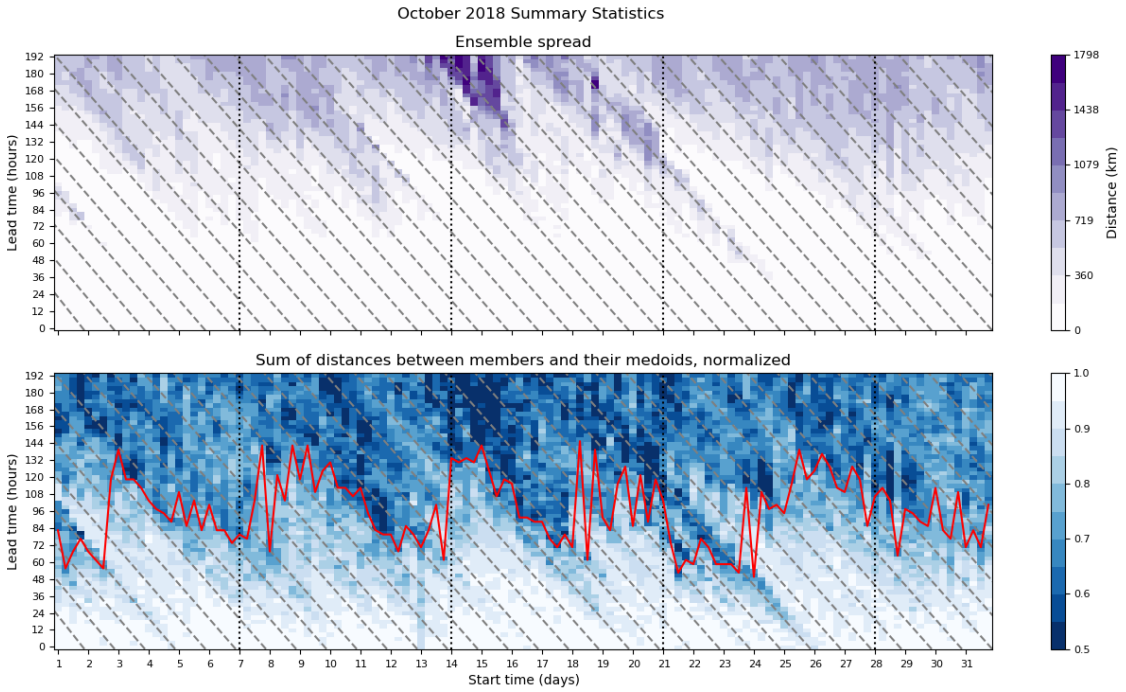


Figure 5.13: A "predictability plot" based on comparison of the large-scale rain rate between ensemble members for the month of October 2018. The top plot is the ensemble spread calculated as the average FSS distance. The middle plot is the sum of within cluster distances, normalized by the sum distances from the medoid of the whole ensemble. The red line marks where the window of interest (the point at which the sum distance has decreased to the 25th percentile) begins. The forecast start time is denoted on the x axis. The vertical black dotted lines mark every seven days. The diagonal dashed lines link the same verification times across forecasts.

associated with fronts. The larger spread of the rain rate compared to $|\nabla\theta_w|$ indicates a larger variation between members. This is likely to do with the more variable shape of the rain rate objects and their size (visible in the paintball plots of 5.5 and 5.6). As the spread is calculated from the FSS distances, it can be concluded that clusters are more likely to contain members that are similar in $|\nabla\theta_w|$ versus the rain rate, but $|\nabla\theta_w|$ clusters are also more likely to have similarities between them. This can also be seen in the paintball plots, where the positions of the rain rate objects tend to be much further apart between clusters than the frontal objects of $|\nabla\theta_w|$. In cases with greater spread, the sum distance is likely to decrease at a faster rate, evidenced in the predictability plots of figures 5.13 and 5.12, where the sum distance drops significantly faster in the large-scale rain rate than in $|\nabla\theta_w|$. The larger increase in spread of the FSS distance related to rain rate contributes to the faster drop of the sum distance as members more quickly perturb away from one another. This is likely to begin a window of interest for the rain rate before $|\nabla\theta_w|$, which is the likely explanation for the fairly regular five day window of interest across valid times pattern seen in the rain rate versus the fewer periods of window of interest across valid times in $|\nabla\theta_w|$.

5.3.2 Cluster membership

How similar clusters are between $|\nabla\theta_w|$ and the large-scale rain rate can be determined by using cluster inter-comparison diagrams, seen in previous chapters. Instead of comparing different lead times to the beginning of the window of interest to examine traceability of clusters, the cluster membership obtained using the two different variables as the input fields to the clustering algorithm based on the FSS distance can be compared to each other at each lead time via the cluster inter-comparison diagrams and the sum along the diagonal of the diagrams can be examined. The higher the sum, the stronger the matches between the variables, where 18 is a perfect match and 6 is the lowest possible score when clusters are rearranged for the largest matches to appear along the diagonal (i.e. if a 4 by 4 matrix had elements that summed up to 18 but were as evenly spread across the matrix as possible, two positions would still have 2 instead of 1, and when the largest possible digits are then moved to the diagonal, the diagonal sum can only be a minimum of 6). Figure 5.15 displays a histogram of diagonal sums from the cluster inter-comparison

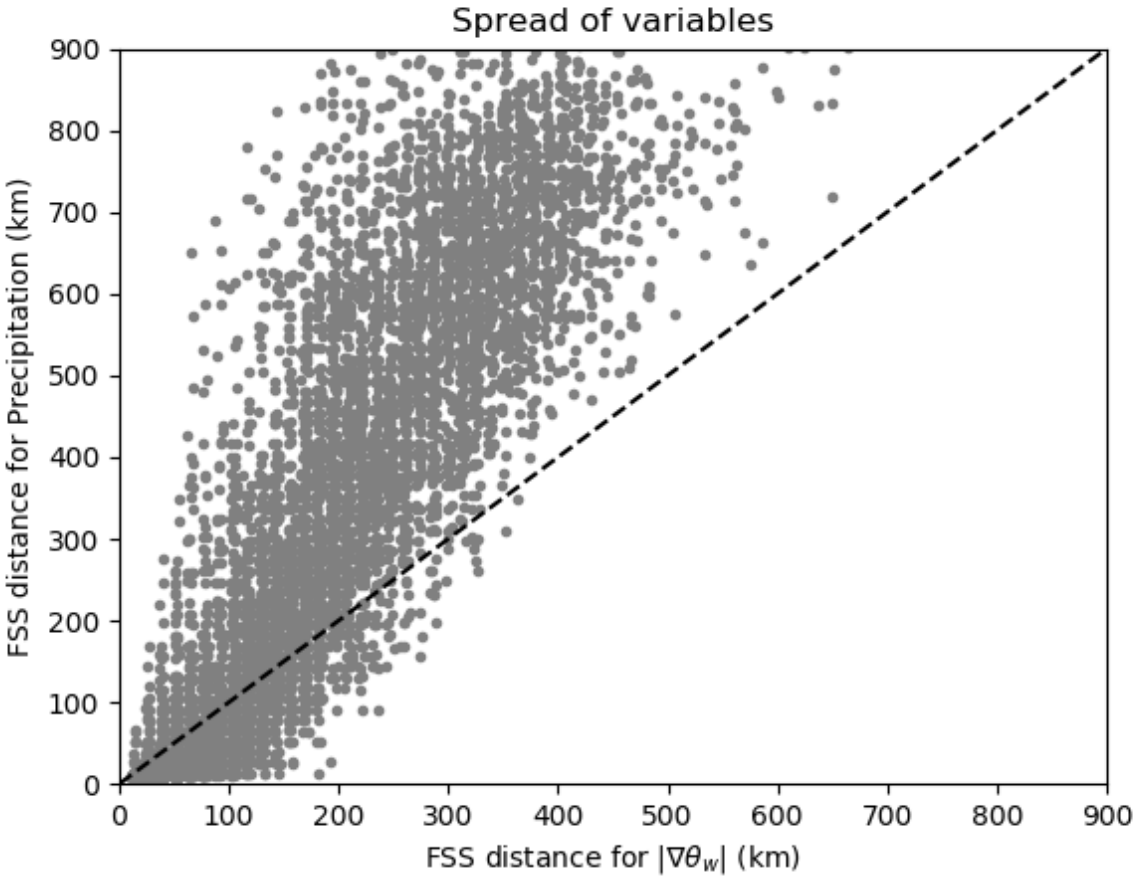


Figure 5.14: The spread of the FSS distance in km associated with large-scale rain rate compared to the spread of the FSS distance in km associated with $|\nabla\theta_w|$.

of two sets of four randomly generated clusters. The most common sum is 8, followed closely by 9. Therefore, anything above 9, where the cumulative distribution value is 0.473, is increasingly likely to not be a match purely by chance. A high match between clusters of $|\nabla\theta_w|$ and large-scale rain, i.e. a match of 9 or greater, will indicate a strong connection between the variables and increases the likelihood the representative members in one variable will reasonably carry over to another.

The diagonal sums of the two variables compared via cluster inter-comparison diagrams have been compiled for each lead time for the months of October 2018. In figure 5.16, these sums are displayed on a colour plot of lead time versus the number of matches between variables. A similar colour plot for the previous statistical figure 5.15 is included for comparison at the top. It is anticipated that there will likely be higher sums (i.e. higher likelihood of clusters matching) in the early lead times due to the likelihood of members predominantly being in a single clusters until there is enough variation among members to begin forming more strongly varied clusters. This is clearly the case in lead times t+3 to t+27 where the majority of sums equal 12 or 14. After this, the distributions

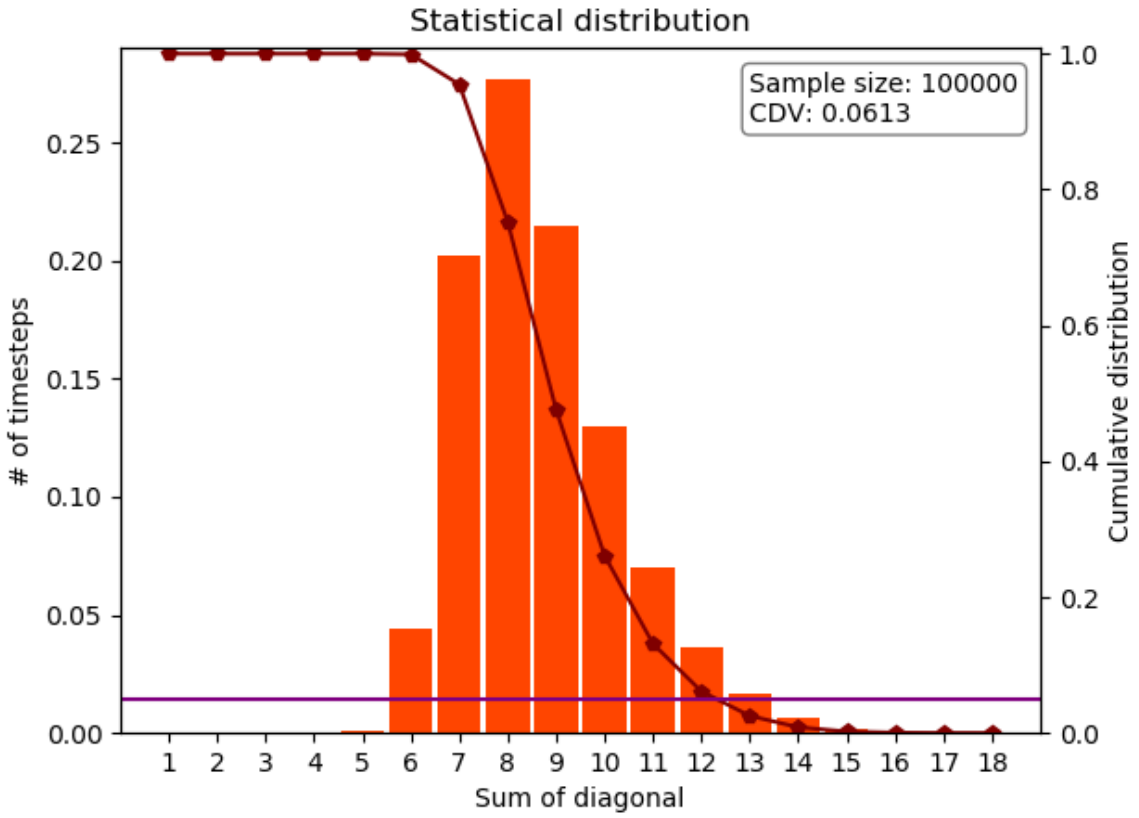


Figure 5.15: A histogram of diagonal sums obtained for 2 sets of 4 random clusters of 18 members, calculated from cluster inter-comparison diagrams, with a sample size of 100,000 and normalized so the sum of the bars is 1. The inverse cumulative distribution is calculated up to but not including the sum of values up to each point for each diagonal sum, i.e. the cumulative distribution plotted at 10 includes the values from 1 to 9 but not 10. The CDV is the cumulative distribution value closest at the 0.05 line.

of sums move into a more standard distribution. However, the cumulative distribution value closest to 5% is 0.06 at 12 matches, indicating a significant chance the matches between clusters were not just random at 12 or greater, although any matches of 9 and above, where the cumulative distribution is 50%, are sufficient to note. As can be seen in figure 5.16, the plot of the matches between the clusters associated with rain rate and $|\nabla\theta_w|$ is shifted to the right compared to the statistical distribution of the random cluster matches, indicating these two variables are related, as expected. However, after 48 hours the diagonal sums being at least 12 or more is less likely.

5.3.3 Window of interest

Another important consideration when examining how similar clustering is between two variables is examining when the window of interest begins. The beginning of the window of interest is when clustering is starting to become the most distinct and therefore indicates when there is notable uncertainty within the ensemble forecasts. As $|\nabla\theta_w|$ relates

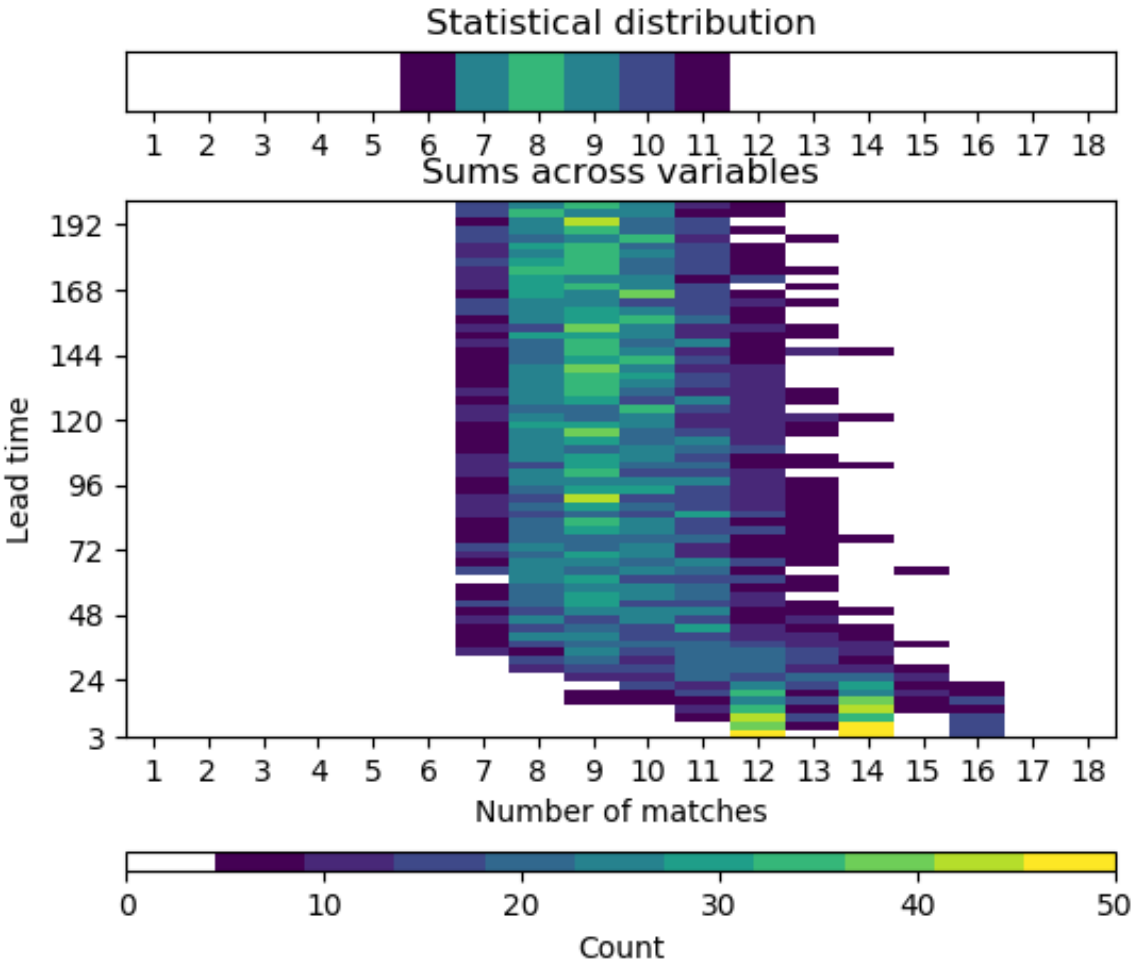


Figure 5.16: A colour plot of the statistical distribution of diagonal sums from cluster inter-comparison (see section 3.3.3.1) of two sets of random clusters of 18 members (top, see figure 5.15) and the diagonal sums from the cluster inter-comparison diagrams for the month of October 2018 between the clusters for the large-scale rain rate and the clusters for $|\nabla\theta_w|$ at individual lead times versus the number of matches (bottom).

to frontal positions, it can be anticipated that the window of interest will begin when there are still frontal pattern similarities among the members but significant enough variations between the clusters as to render their representative members (and therefore frontal patterns) distinct potential scenarios. With large-scale rain rate, the window of interest is also based on patterns amongst the members and variation in the clusters, but it is related to the size and shape of the area associated with the rain rate. Figure 5.17 shows the window of interest start time for the ensemble forecasts of October 2018 for $|\nabla\theta_w|$ (orange bars) and the large-scale rain rate (blue line) in the top plot as well as how many representative members match between variables (black) in the bottom plot. The Pearson's correlation coefficient comparing the window of interest start times between the variables is -0.0078, indicating there is almost no correlation between the two variables. This can be due to several reasons. When the large-scale rain rate window of interest begins earlier than the window for $|\nabla\theta_w|$ the positions of the fronts may be more certain than where the rain is located. There may also be periods of rainfall that aren't associated with fronts. Alternatively, if the window of interest for $|\nabla\theta_w|$ begins before the rain rate, this may indicate there is larger uncertainty in the development of frontal systems or the fronts are not yet well defined enough to be associated with heavy rain. However, it is important to note that there are periods of similarity that can be visually picked out. During the periods of October 4^{th} to the 7^{th} and the 25^{th} to the 31^{st} both window start times begin at very similar lead times for each variable, indicating there is an event that is dominating the uncertainty in the ensemble. The period of the 7^{th} to the 10^{th} picks out different valid times of interest for each variable. It is also important to note that the start of the window of interest is noisier on shorter timescales for the rain rate than for $|\nabla\theta_w|$. Regarding how many representative members match (out of 4) and how they compare to the window of interest start times, there is a relationship between the two variables as it is unlikely to see two more more matches between RMs if it were unrelated.

5.3.4 Representative members

How many representative members match between the variables can be further examined via figure 5.18. The figure displays a histogram of the number of matches between the representative members of $|\nabla\theta_w|$ and large-scale rain rate (where instances when forcing

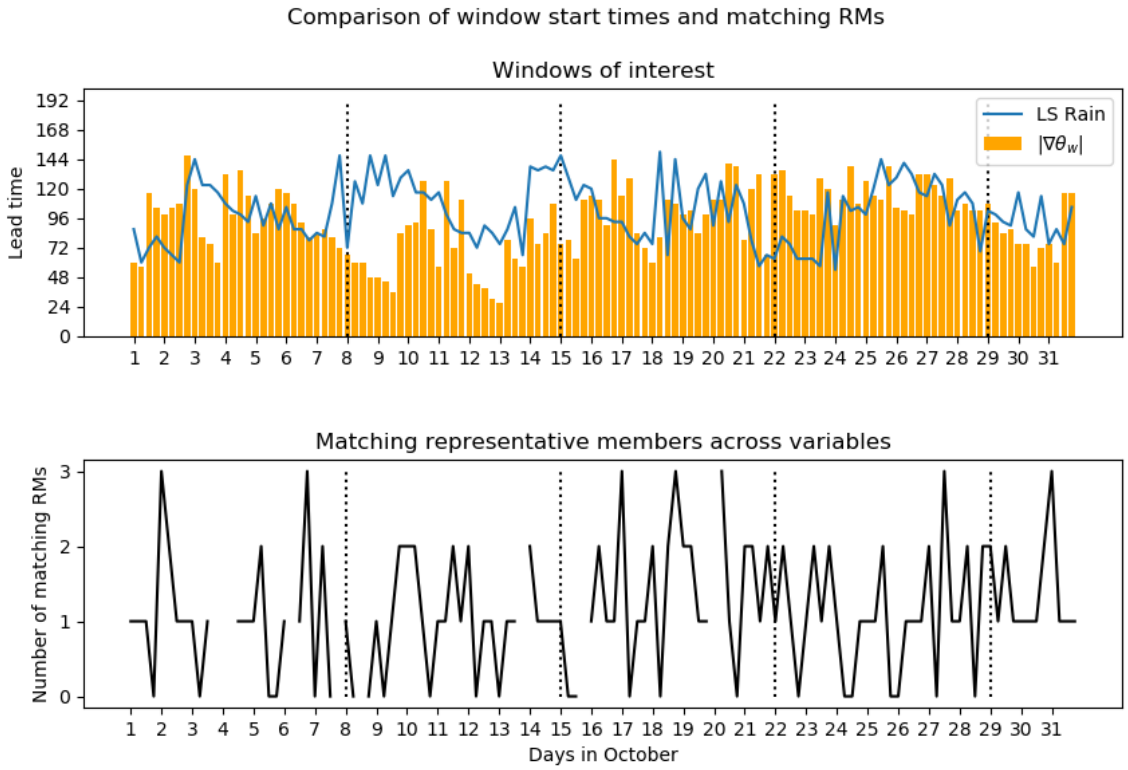


Figure 5.17: A plot of the window of interest start times for $|\nabla\theta_w|$ in orange bars and large-scale rain rate in a blue line plot. The number of matching representative members is in black in the bottom plot. Where the value is missing indicates where a representative member was repeated twice or more. This is due to forcing the clusters to 4.

the number of clusters to 4 resulted in duplicate RMs have been removed) over the month of October, which is displayed in blue. How often two sets of four random numbers from 0 to 17 have matches (sampled 10,000 times), where there are no repeating numbers within a set, is displayed in orange. Comparing these two histograms shows they are similar in some respects. The dominant number of random matches between two sets of four numbers (orange) is 1, closely followed by 0 matches, then 2, with a tiny fraction if any at all of 4 matching numbers. In the histogram for $|\nabla\theta_w|$ versus large-scale rain (blue), the dominant number of matches is again 1. There were fewer instances of 0 matches, but greater instances of 2 and 3 matches. However, throughout the month of October 2018 there were no sets of all four representative members matching across the variables. However, the p value between these two sets is 0.0053, indicating the probability of two or more RMs matching is significant and these two variables are clearly related.

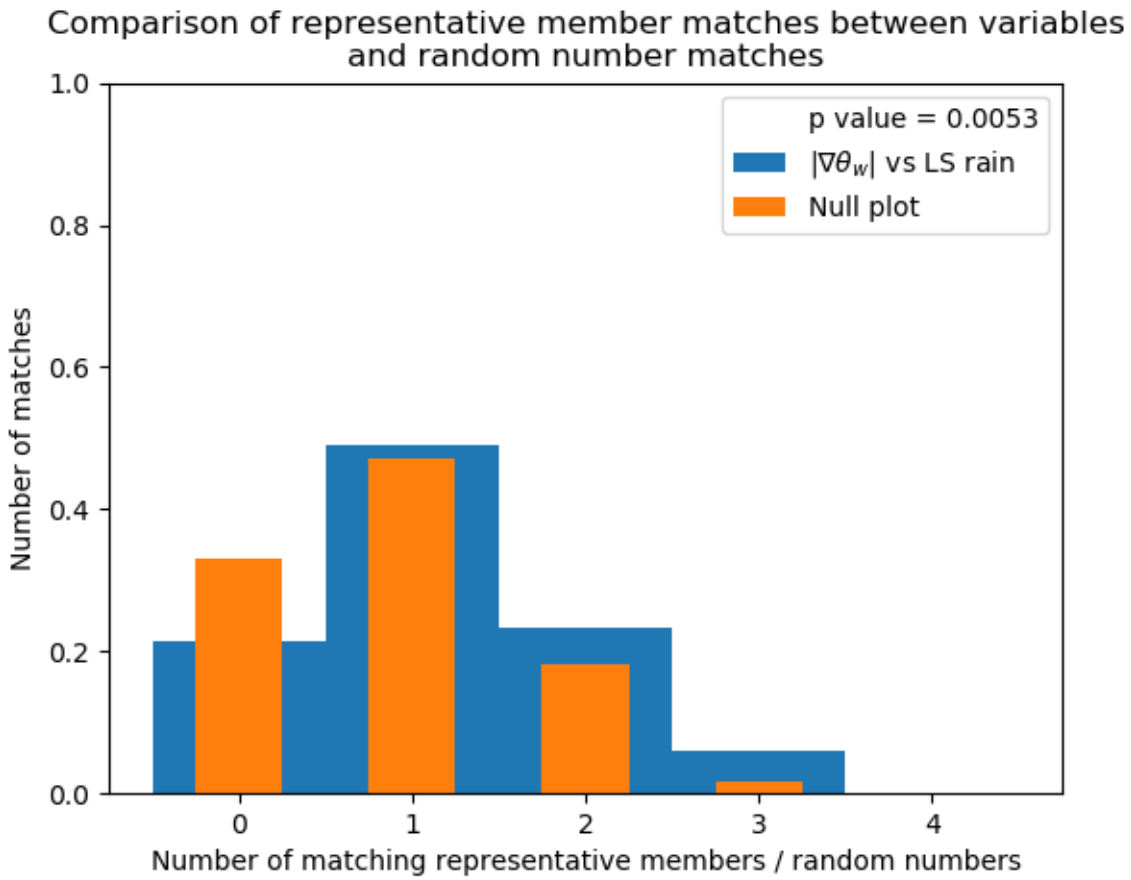


Figure 5.18: A histogram of matches between the representative members of $|\nabla\theta_w|$ and large-scale rain rate for the month of October 2018 (blue) and a histogram of matches between two sets of random numbers between 0 and 17 (orange). Both histograms are normalized so the sum is 1 and then a p value was calculated for them.

5.4 Conclusion

This chapter discussed how the large-scale rain rate compares to $|\nabla\theta_w|$ in terms of general clustering mechanics and statistically over the month of October. The clustering algorithm has been shown to work on both variables, however, the FSS distances between members are larger in the large-scale rain rate than $|\nabla\theta_w|$. There is similarity in dominant features when compared visually at the same lead time, indicating the large-scale rain rate often corresponds with frontal regions. Both variables produce clusters that have traceable membership across lead times and distinct RMs. Examining the variables over the month of October revealed a complex relationship. The spread is greater for large-scale rain rate but it positively correlates with $|\nabla\theta_w|$, which is likely to do with the more variable shape of the rain rate objects and their overall larger size than the thin frontal objects. The large-scale rain rate produced a cyclic pattern in the sum distances and by relation the beginning of the windows of interest. When comparing cluster membership between variables there was some similarity, which can be expected since the large-scale rain rate is often associated with fronts. There was no correlation between window of interest start times, however there are clearly periods of strong correlation mixed with periods when there are jumps in the sum distance and the different windows are picking up different events dominating the clusters. There is also a significant connection between RMs in both variables, which is surprising as the RMs are determined specifically by cluster membership within the window of interest.

There are many similarities between the variables indicating they are connected and clearly related. However, the relationship is not strong enough to say they are equivalent and interchangeable. The clustering associated with $|\nabla\theta_w|$ is better at picking up scenarios as it is more consistent across forecasts whereas the large-scale rain rate jumps more frequently between events. The rain rate can be intermittent and may not be related to $|\nabla\theta_w|$. With respect to the fundamental differences in the variables and the aim of the project, i.e. to extract high-impact scenarios from ensemble data, it can be concluded that using $|\nabla\theta_w|$ to find different frontal patterns is a better choice than using the large-scale rain rate to look at impact variability of regions of intense rain.

Chapter 6

Evaluation of the novel clustering method during an operational testbed

6.1 Introduction

In this chapter, the performance of the novel clustering method during the Met Office winter testbed will be discussed. This chapter aims to look in depth at the pros and cons of the method as evaluated by scientists and operational meteorologists during the testbed. The sections that follow include an explanation of the testbed setup, a brief summary of the state of the atmosphere and meteorological events during the testbed, and an in-depth analysis of a case taken during the testbed and related survey results from the participants.

6.2 Met Office winter testbed, January to February, 2022

The Met Office testbed ran daily from 09:15 to 1600 (approximately), for four weeks in January and February 2022 (Jan 10 - 14, Jan 24 - 28, Jan 31 - Feb 4, and Feb 7 - 11). It brought together both operational meteorologists and atmospheric scientists from UK universities with the goal of getting expert assessment of the projects for further development and refinement, bringing them closer to implementation.

The test bed comprised a set of daily activities for participants that repeated each day for its duration. Assessment of the clustering method I developed was one of the daily activities. Each week of the test bed began with a briefing for participants on products being tested. Each day comprised of several activities. Typically, the activities would begin with a weather briefing from an operational meteorologist, then the activity itself which included a survey. At the end of the activity there would be a discussion.

The method I developed for clustering ensemble forecasts in real-time for high impact scenarios was one of the daily sessions of the testbed. The daily forecast briefing focused on the current weather pattern over the North Atlantic and potential areas of uncertainty leading up to the window of interest in the latest forecasts where the clustering has begun to form sufficiently distinct clusters. The participants were then asked to answer a daily survey while they explored the clustering products for the current ensemble forecast. The survey included questions pertaining to how well the method clustered ensemble members, whether or not high-impact scenarios were present, if the scenarios present carried over multiple forecasts, and if the representative members (RMs) and related products influenced the participants' forecast message, either to the general public or special interest groups (i.e. aviation, emergency management, shipping, etc). The discussion after the clustering activity generally included how well the clustering performed overall and how distinct the RMs were, but they also frequently included questions or clarifications about the method, and ideas for further refinement and other applications.

6.3 Summary of the synoptic events during the winter testbed

During the months of January and February, several different weather types occurred: (i) a high pressure system that dominated for several days, (ii) a generally unsettled and variable period, (iii) a strong mid-latitude cyclone with a pressure centre that tracked up towards Greenland and caused fronts to affect the UK, and (iv) a period of several frontal systems moving across the UK (Met Office, 2022b,a). This meant that the method was trialed by the testbed participants for a variety of different weather patterns. This was advantageous as the behaviour of the clustering method during different regimes and

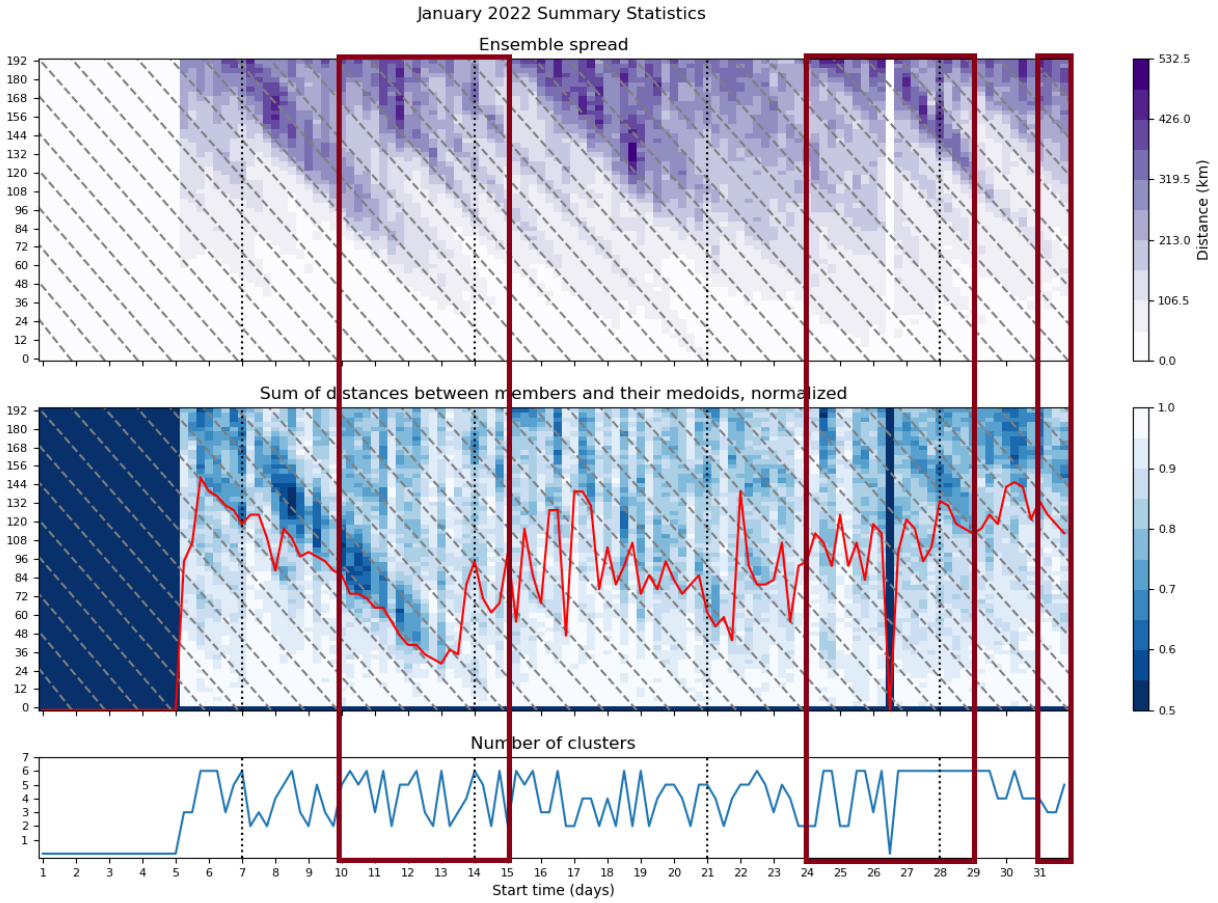


Figure 6.1: A "predictability plot" based on comparison of $|\nabla\theta_w|$ between ensemble members for the month of January 2022. The top plot is the spread. The middle plot is the sum of within cluster distances, normalized by the sum distances from the medoid of the whole ensemble. The red line marks the start of the window of interest (the point at which the sum distance has decreased to the 25th percentile). The bottom plot is the optimum number of clusters chosen by the algorithm, denoted by the forecast start time on the x axis. The vertical black dotted lines mark every seven days. The diagonal dashed lines link the same verification times across forecasts. The blocks of solid colour (white in the spread, dark blue in the sum distances) indicate missing data. The series of dates blocked by dark red rectangles are days in which the testbed was held.

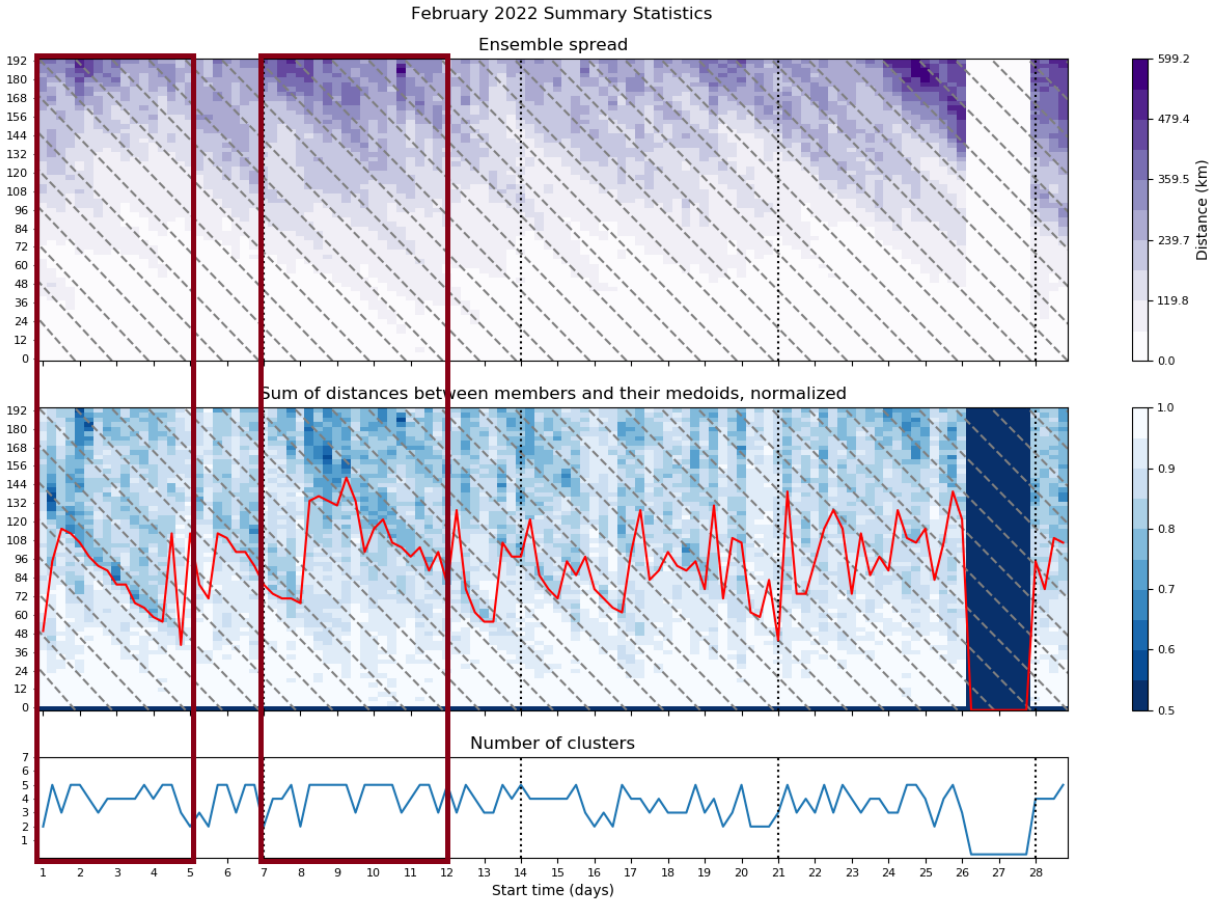


Figure 6.2: A "predictability plot" based on comparison of $|\nabla\theta_w|$ between ensemble members for the month of February 2022. Details as in figure 6.1.

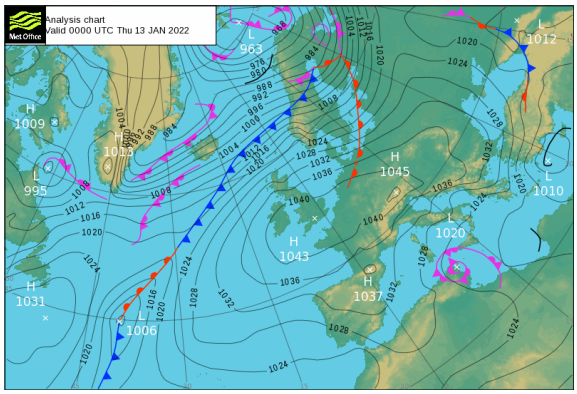
how this affected its perceived utility could be explored. Figures 6.1 and 6.2 show a brief summary of the spread, the sum distance, the number of clusters, and the window of interest start times. Each of the four weeks of the testbed are enclosed with a dark red rectangle and relate to the four weather systems mentioned previously (i through iv) during the window of interest, which was often (weeks 1, 3, and 4) strongly tied to specific valid times, implying that a particular event is determining the behaviour of the clustering. Week 1 was dominated by a high pressure system, week 2 was a generally variable period, week 3 followed a storm developing off the Canadian coast, and week 4 included both a storm in the North Atlantic and a series of fronts affecting the UK. In the following sections, the state of the atmosphere of each week of the testbed will be briefly described by its general characteristics before the windows of interest begin then what uncertainties and variations within these windows will be addressed. This section will then be followed by an in-depth look at a particular case study and the survey responses provided by the participants.

6.3.1 Week 1: January 10 - 14

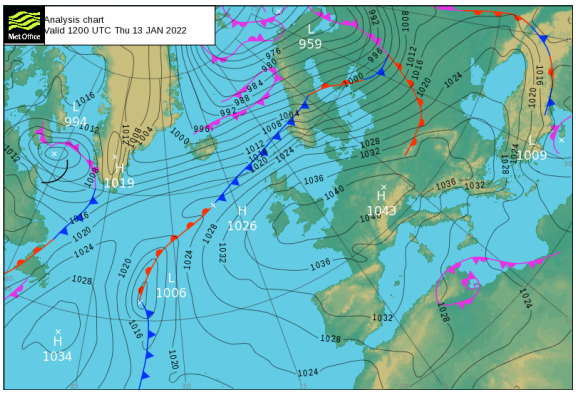
The first week, blocked in dark red in figure 6.1, had a significant event that was picked up by the method across several forecasts at the same valid times. Within the testbed week of the 10^{th} to the 14^{th} the method picked up on a significant drop in sum distance and increase in spread beginning with forecasts starting on the 8^{th} to the 13^{th}, indicating a high likelihood of strong clusters and low predictability of the event in question. Figure 6.3 is a series of analysis charts from Thursday the 13^{th} of January at 0000 UTC to Sunday the 16^{th} of January at 1200 UTC. The start of the event is identified with 0000 UTC on the 14^{th}. Plots (a) and (b) (0000 and 1200 UTC on the 13^{th}, respectively) show a fairly strong high persistent over the UK, western and central Europe and the beginning of a baroclinic wave. Plot (c) is the key time point associated with the start of the event captured by the window of interest and the baroclinic wave is now close to its stage of maximum growth rate forming a cyclone. In plot (d), the wave has a larger amplitude. The ensemble spread and drop in sum distances in figure 6.1 are clearly linked to this instability in the atmospheric flow. In plot (e) the trailing cold front south of the cyclone is unstable and frontal waves develop, one of which becomes a secondary cyclone more visible in plot (f). The primary cyclone begins to breakdown in plot (g) as the secondary cyclone continues to develop into plot (h). The baroclinic wave development and resulting cyclones were the primary focuses of the window of interest across forecasts at the same valid time, which can be seen in 6.4. These paintball plots are associated with the beginning (a) and end (b) of the window of interest for the 10^{th}. Here, it can be seen in plot a that the primary feature being clustered on is the frontal region to the west of the high. In plot b, there is very little cohesion amongst clusters, a possible indicator that how the high breaks down is where the primary uncertainty in this event lies.

6.3.2 Week 2: January 24 - 28

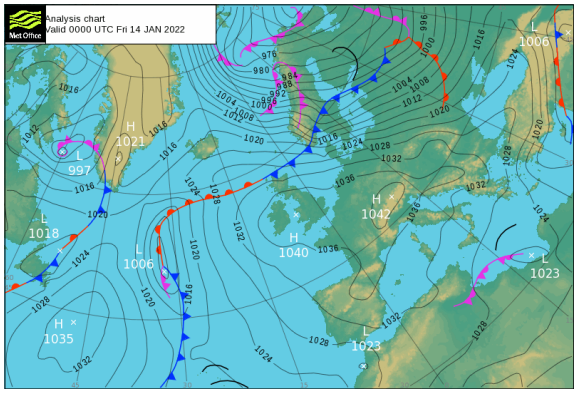
While the first week had a strong connection across forecasts at the same valid time, the second week did not. This can be seen in figure 6.1 in the second time period enclosed by a dark red rectangle. There was no strong connection between forecasts where the window of interest began, indicating that there wasn't a specific system that was strongly uncertain identified across a series of forecasts with different start dates. Figures 6.5 to



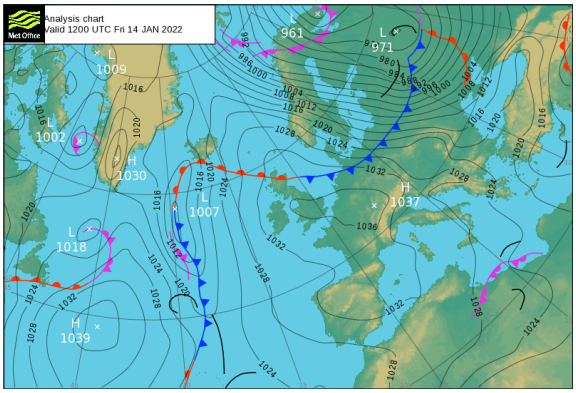
(a)



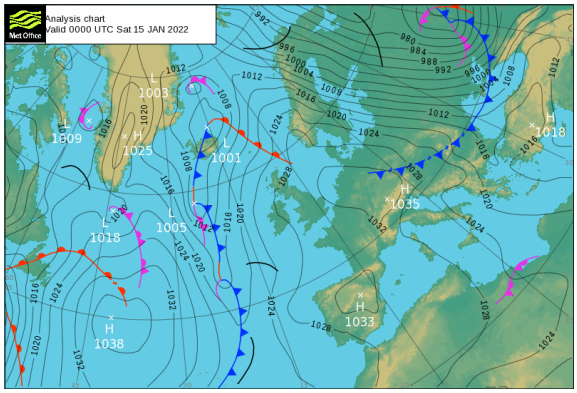
(b)



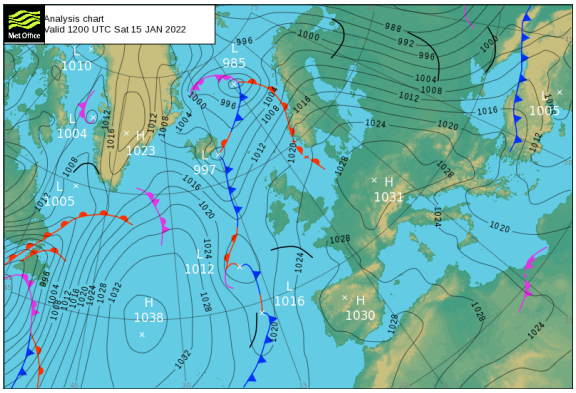
(c)



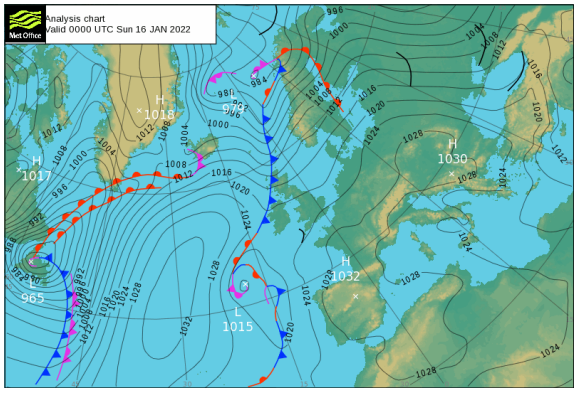
(d)



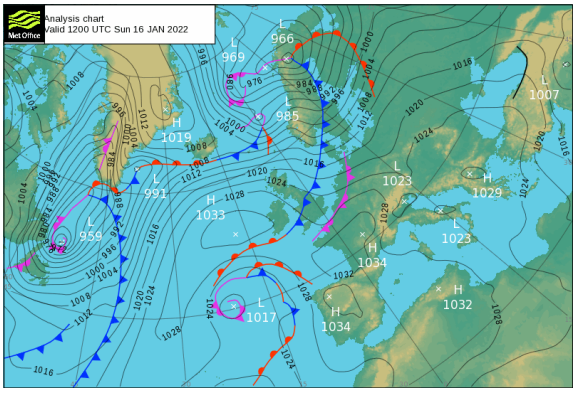
(e)



(f)

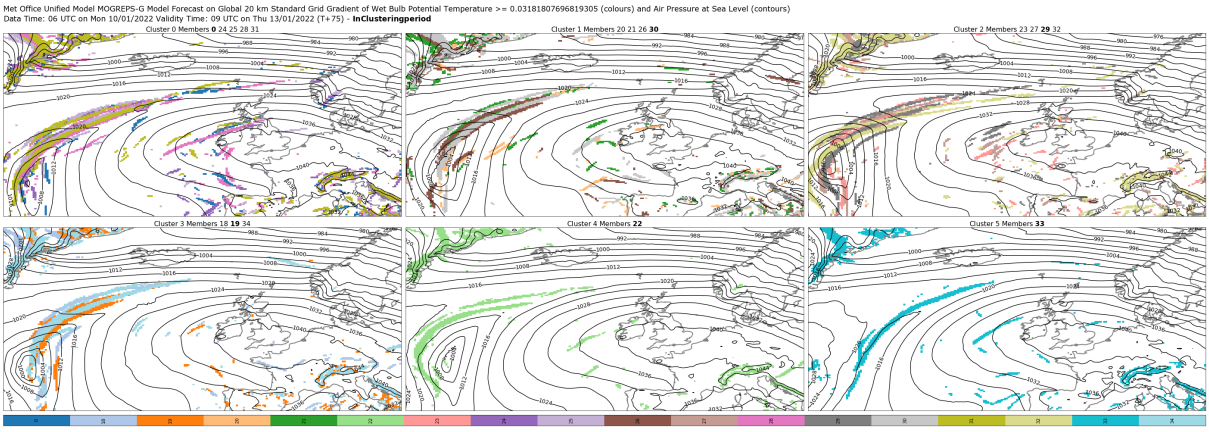


(g)

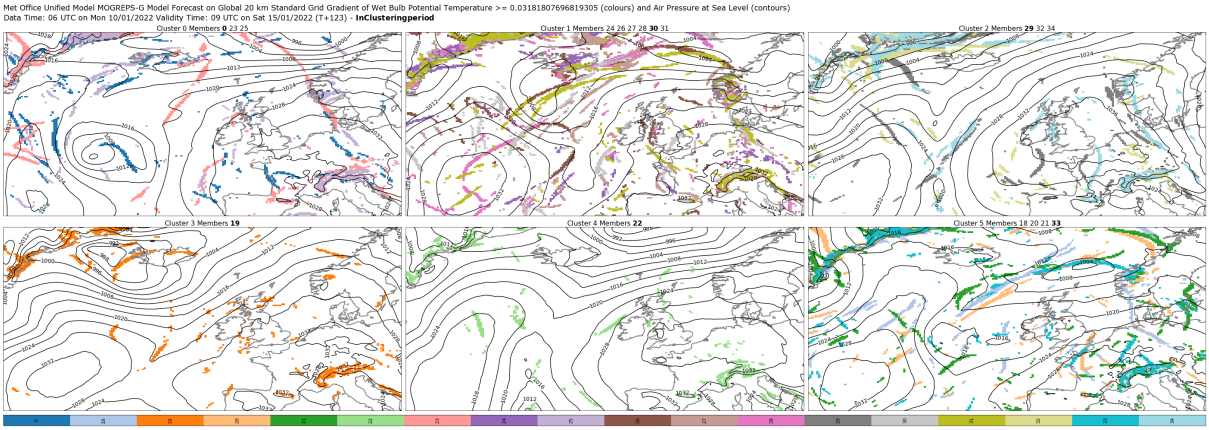


(h)

Figure 6.3: Analysis charts for 13-01-2022 to 16-01-2022, associated with forecasts from week 1 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.

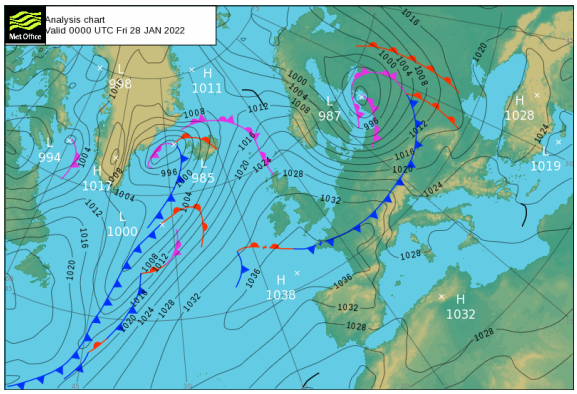


(a)

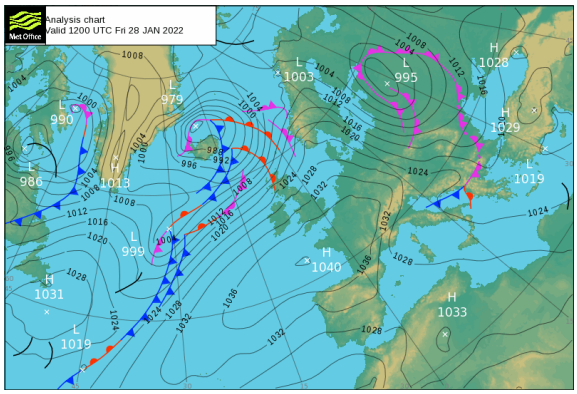


(b)

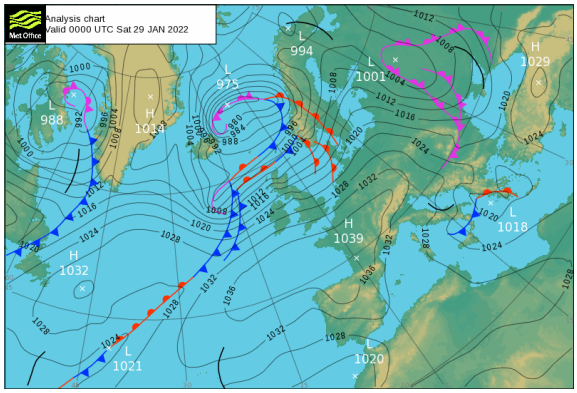
Figure 6.4: Paintball plots from week 1 of the testbed from $|\nabla\theta_w|$ at 850 hPa with the MSLP displayed of the representative member of each cluster from the 0600 UTC forecast on 10/01/2022 at lead time t+74, valid time 0900 13/01/2022, in plot (a), and at lead time t+123, valid time 0900 UTC 15/01/2022, in plot (b). The paintball plots represent the threshold applied to the $|\nabla\theta_w|$ fields, where each member is represented by its own colour.



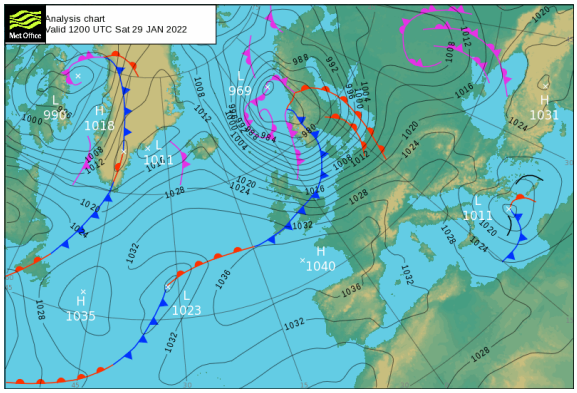
(a)



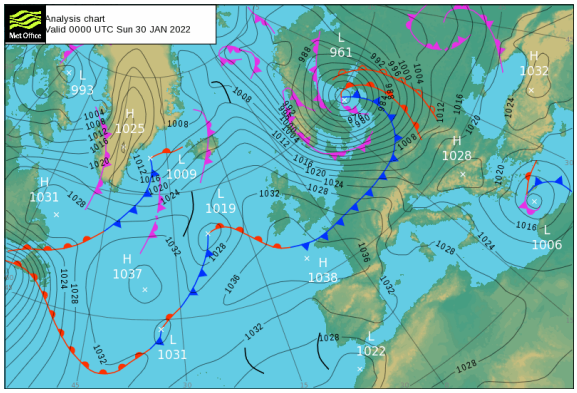
(b)



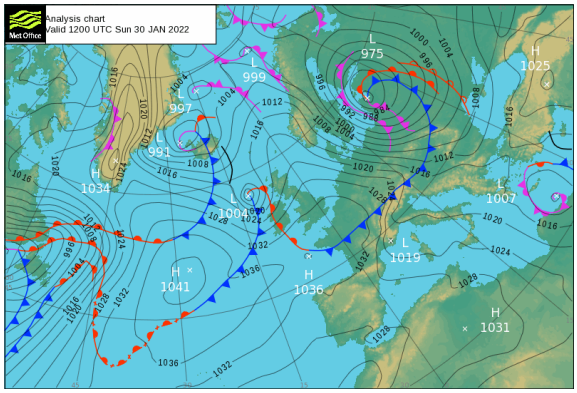
(c)



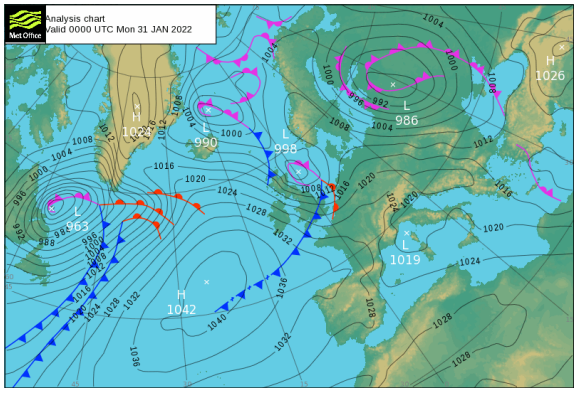
(d)



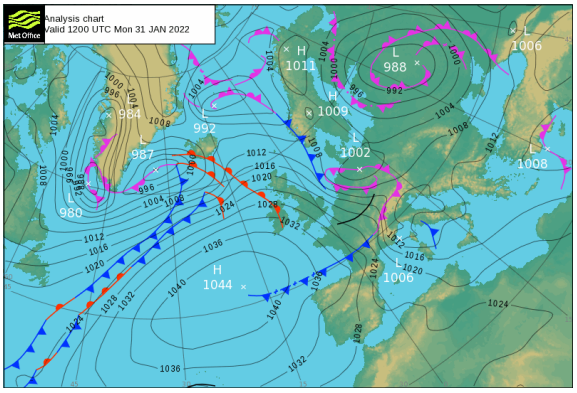
(e)



(f)

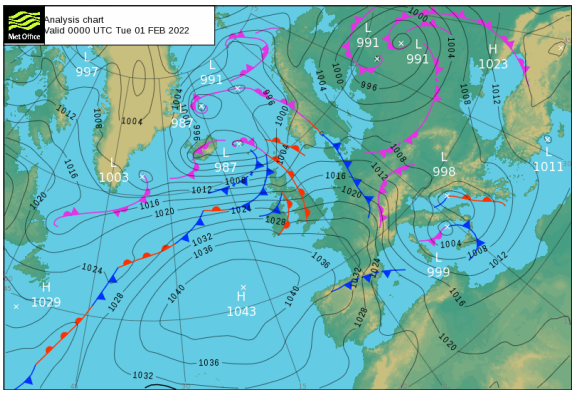


(g)

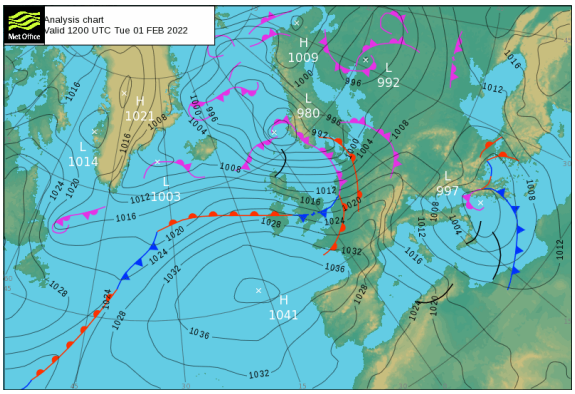


(h)

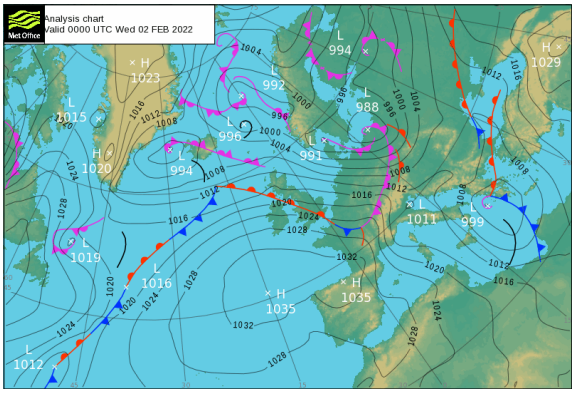
Figure 6.5: Analysis charts for 28-01-2022 to 31-01-2022, associated with forecasts from week 2 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.



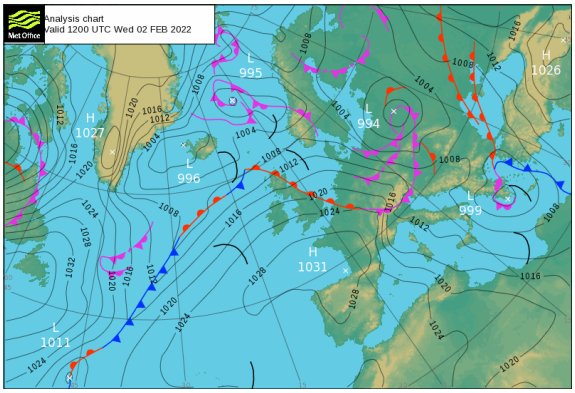
(a)



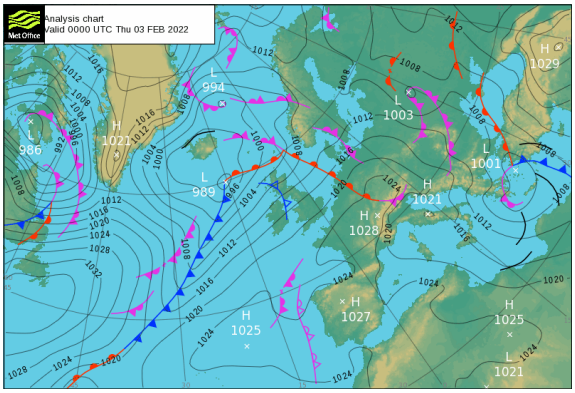
(b)



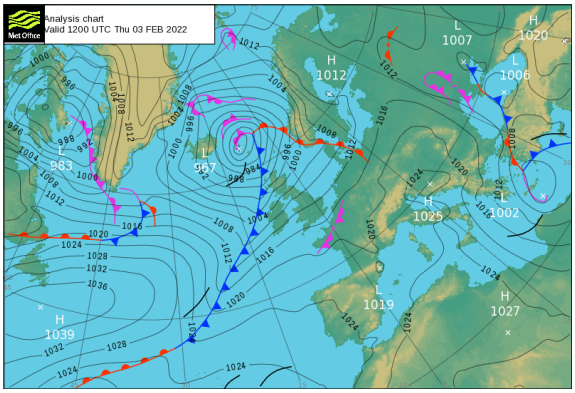
(c)



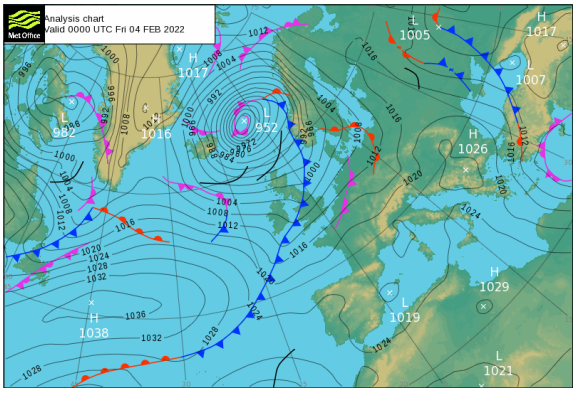
(d)



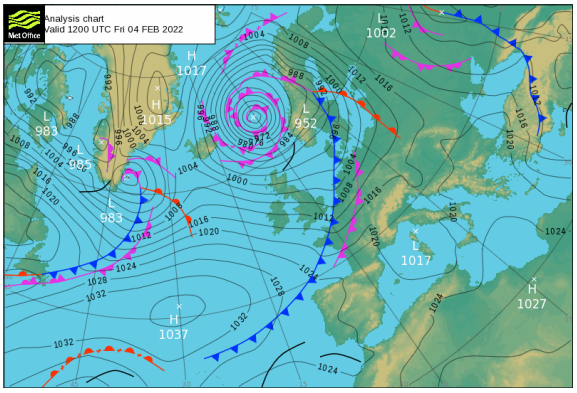
(e)



(f)

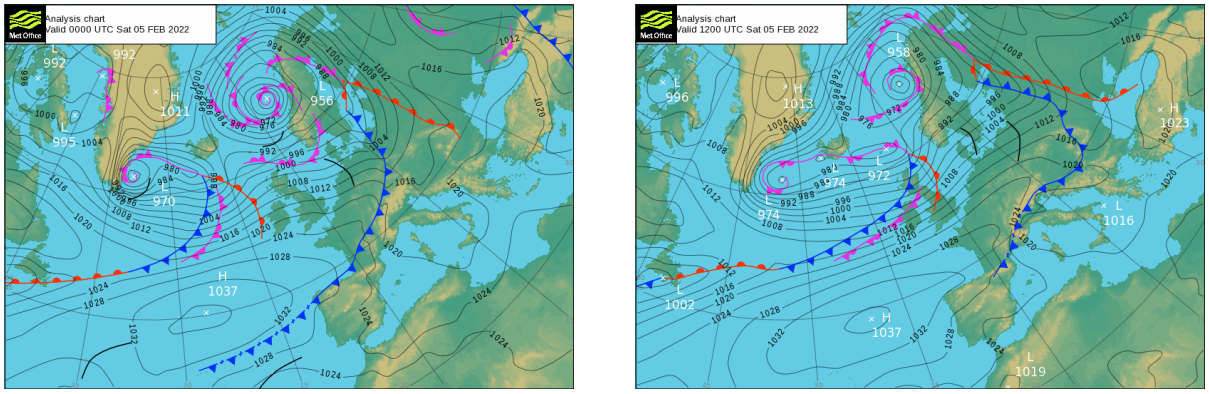


(g)



(h)

Figure 6.6: Analysis charts for 01-02-2022 to 04-02-2022, associated with forecasts from week 2 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.



(a) (b)

Figure 6.7: Analysis charts for 05-02-2022, associated with forecasts from week 2 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.

6.7 are the analysis charts for the approximate time enclosed by the various windows of interest during week 2. The week is dominated by a high pressure system that is quite stable as the rate of ensemble spread is relatively small and not dependent on any particular day. Figure 6.5 begins with the 0000 UTC analysis of the 28th of January, where a frontal system appears to the south of the UK and a series of fronts stretch out across the North Atlantic. A low deepens and moves past Iceland and towards Norway in plots (a) to (d), then progresses eastward in (e) and (f). A baroclinic wave is growing on the 30th, but doesn't appear to influence the ensemble spread. The low associated with the wave develops off the western coast of the UK in (d) to (f), and is pushed along by a high pressure system that moves into the North Atlantic. This high pressure system remains for some time, carrying over into the analysis plots in figure 6.6. This high pressure system lingers until plot (d), where it begins to move southward and a storm system begins to develop between Iceland and the UK in (e) and (f). This storm deepens rapidly and moves northward along the Norwegian coast in plots (g) and (h), with a secondary cyclone forming off the coast of Greenland. In figure 6.7 the secondary cyclone progresses across the North Atlantic and begins to impact the UK. This week, the clustering depended on multiple different frontal zones instead of a single event, resulting in the window of interest beginning at different lead times throughout the week.

6.3.3 Week 3: January 31 - February 4

This week saw the clustering primarily focusing on a single event on 1200 UTC 06/02/2022, seen in figures 6.1 and 6.2, where the rectangle encompassing the week spans

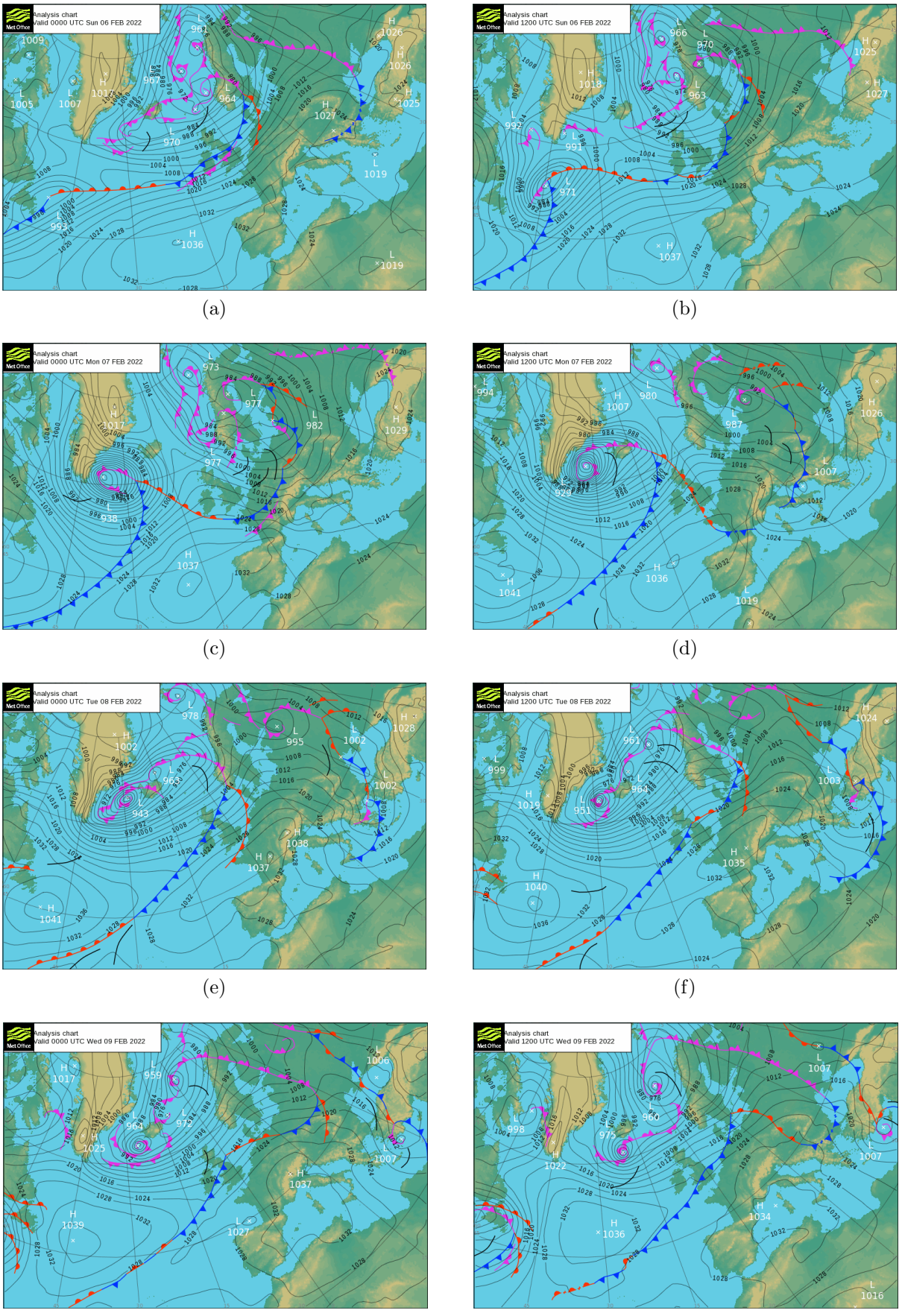
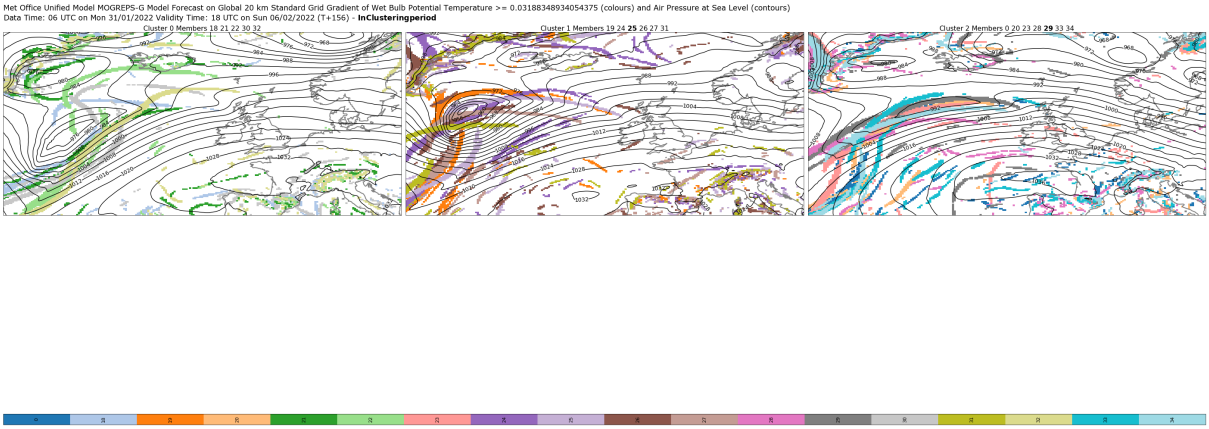
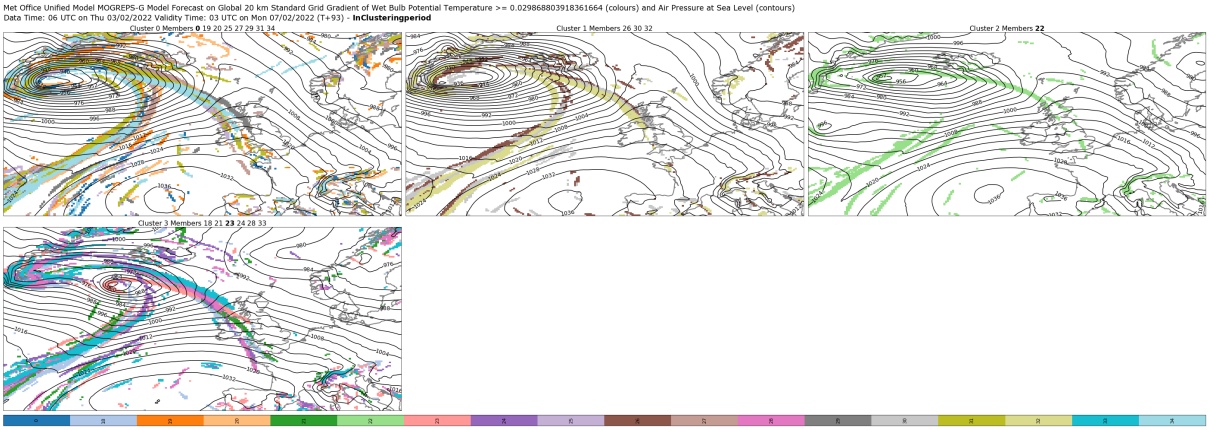


Figure 6.8: Analysis charts for 06-02-2022 to 09-02-2022, associated with forecasts from week 3 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.



(a)



(b)

Figure 6.9: Paintball plots from week 3 of the testbed from $|\nabla\theta_w|$ at 850 hPa with the MSLP displayed of the representative member of each cluster from the 0600 UTC forecast on 31/01/2022 at lead time t+156, valid time 1800 UTC 06/02/2022, in plot (a), and the 0600 UTC forecast on 03/02/2022 at lead time t+93, valid time 0300 UTC 07/02/2022, in plot (b). The paintball plots represent the threshold applied to the $|\nabla\theta_w|$ fields, where each member is represented by its own colour.

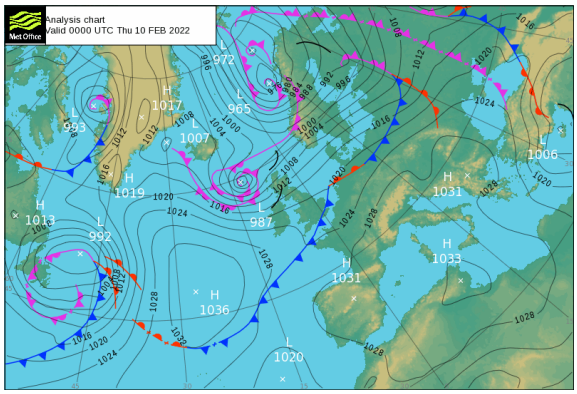
across the plots. The analysis plots can be seen in figure 6.8. Here, the primary feature of uncertainty is the storm developing off the Canadian coast and its associated fronts. In plots (a) through (d) the storm center deepens and moves towards the coast of Greenland, where it remains for the remainder of the window, causing frontal zones to cross the UK. The low predictability of this system and the temporal and spatial qualities of the associated frontal zones can be seen in figure 6.9. Initially, there was uncertainty as to when the storm would develop, seen in plot (a), but as the week progressed the uncertainty became primarily associated with the fronts, seen in plot (b). It is clear in this figure that the ensemble spread is associated with the fronts embedded within the rapidly developing cyclone. This week will be discussed in depth within the case study.

6.3.4 Week 4: February 7 - 11

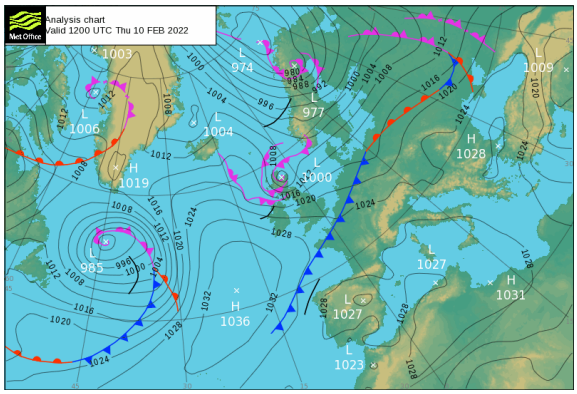
The final week, seen as the last dark red rectangle in figure 6.2, followed two separate events across forecasts over two different valid times. The first event was a storm moving across the North Atlantic with a strong zonal flow (figure 6.10, plots (a) to (d)). The clustering focused on the frontal region moving along with the storm on Thursday and Friday (figure 6.12). The second and main event was a series of fronts moving across the UK (figure 6.11) beginning on the 15$^{\text{th}}$. In figure 6.13, the same valid time of 0300 UTC 16/02/2022 is presented for three different forecasts (0600 UTC 09/02/2022, 0000 UTC 10/02/2022, and 0600 UTC 11/02/2022). Here the series of fronts can be seen over the course of several forecasts, illustrating how the structure of the fronts becomes more defined as lead time reduces.

6.4 Case study

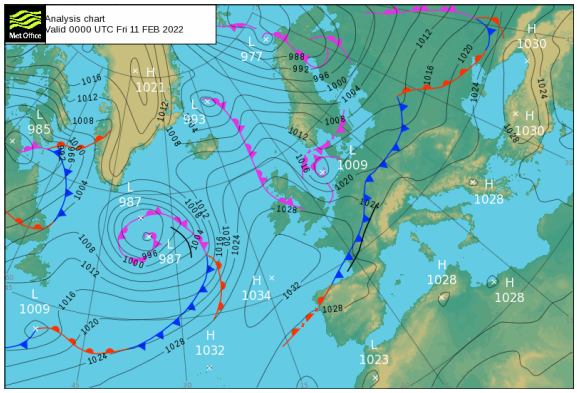
Within this section the third week of the testbed, January 31$^{\text{st}}$ to February 4$^{\text{th}}$, will be expanded upon with a deeper analysis of the clustering and an in-depth discussion of the survey results from the testbed during the week. This week was chosen because the event the windows of interest (beginning around 1200 UTC 06/02/2022) focused on carried across valid times throughout the week. This allowed the participants to see how the clustering differed each day and how the RMs and their potential scenarios evolved and carried across valid times.



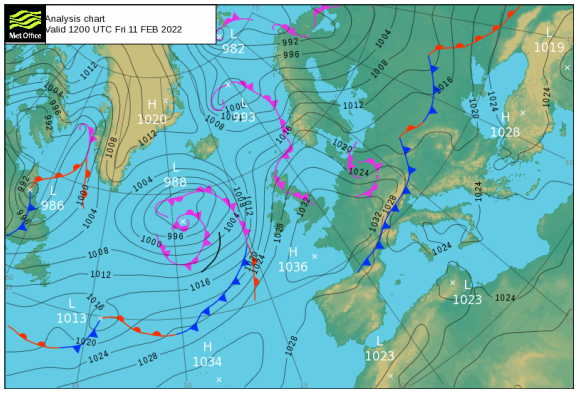
(a)



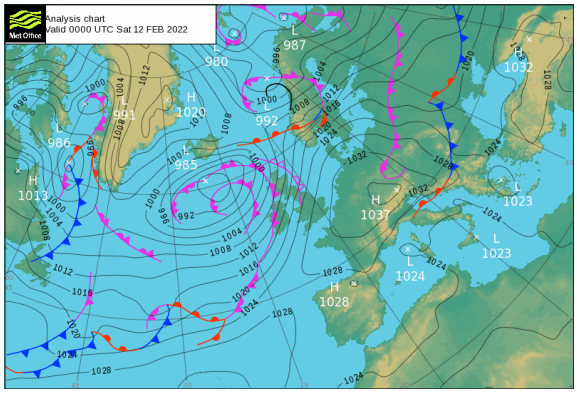
(b)



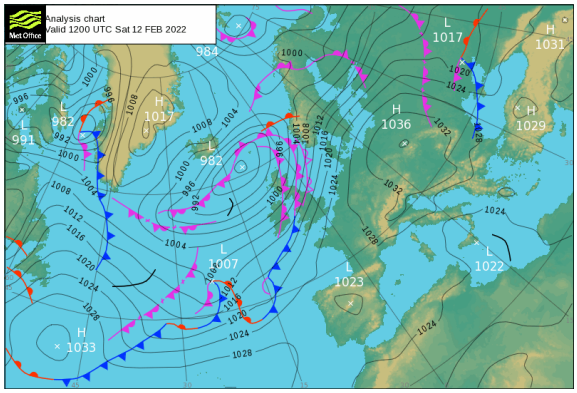
(c)



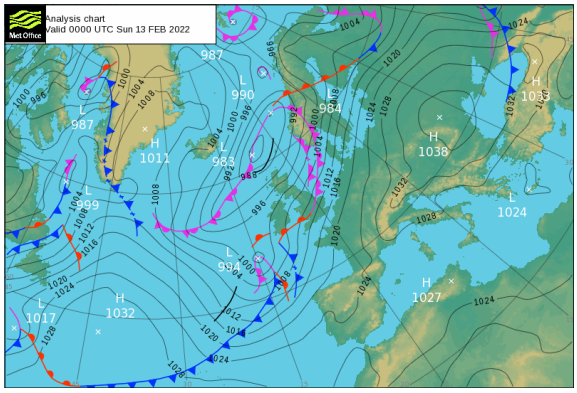
(d)



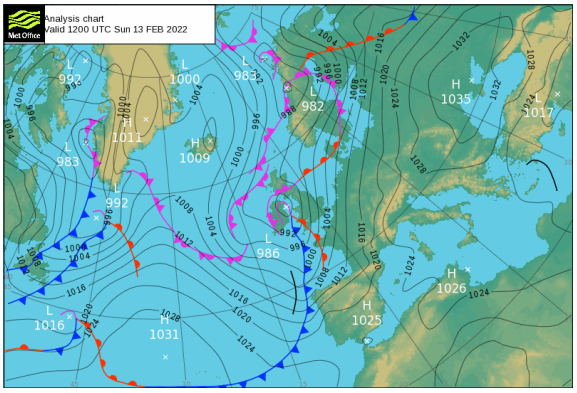
(e)



(f)

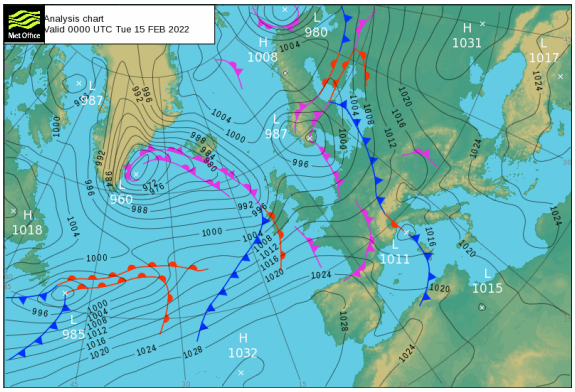


(g)

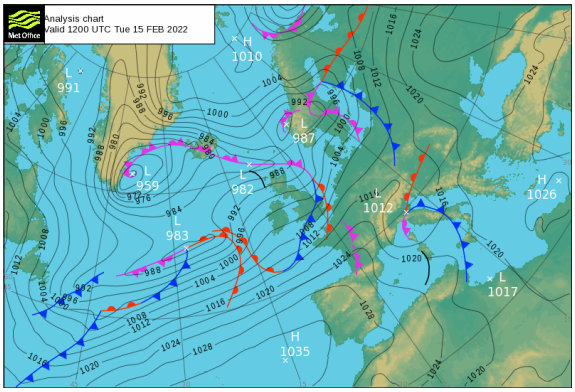


(h)

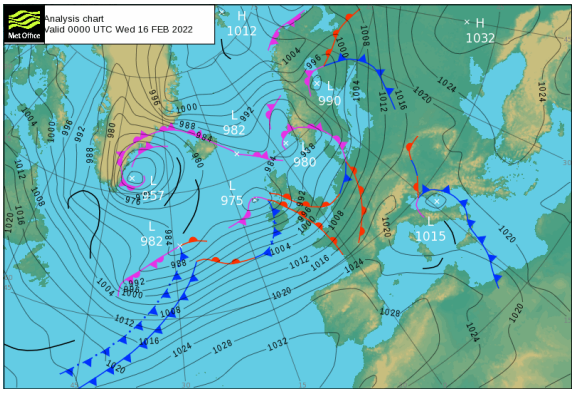
Figure 6.10: Analysis charts for 10-02-2022 to 13-02-2022, associated with forecasts from week 4 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.



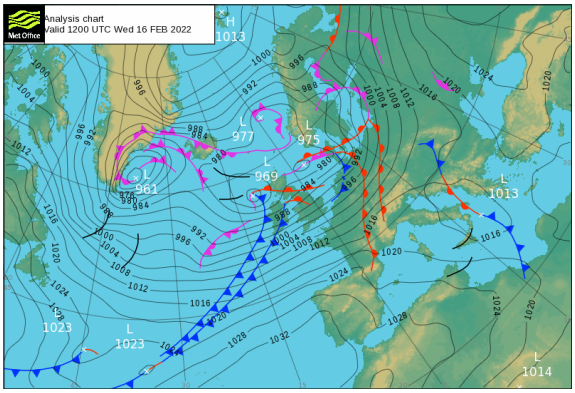
(a)



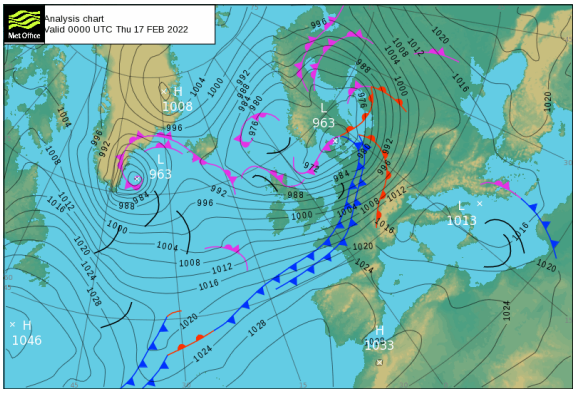
(b)



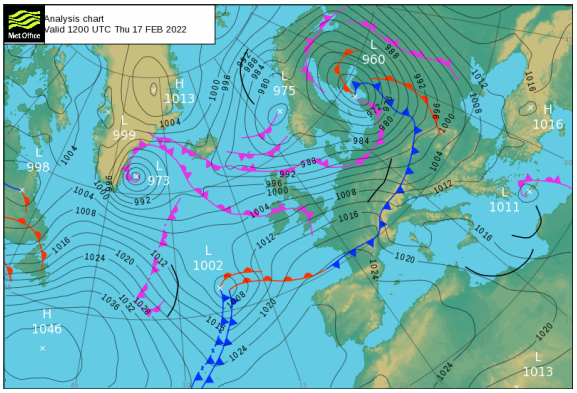
(c)



(d)



(e)



(f)

Figure 6.11: Analysis charts for 15-02-2022 to 17-02-2022, associated with forecasts from week 4 of the testbed. Contains public sector information licensed under the Open Government Licence v3.0, ©Crown copyright, Met Office.

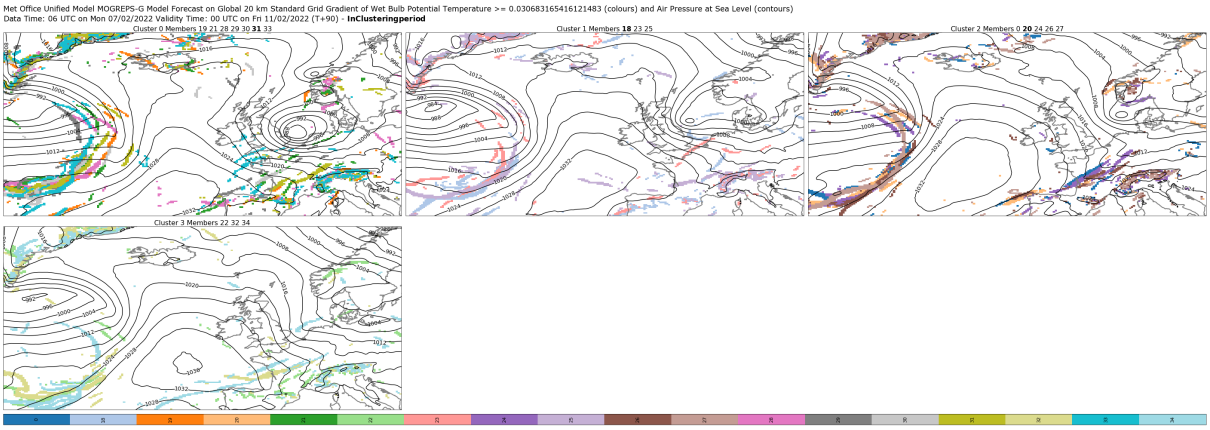
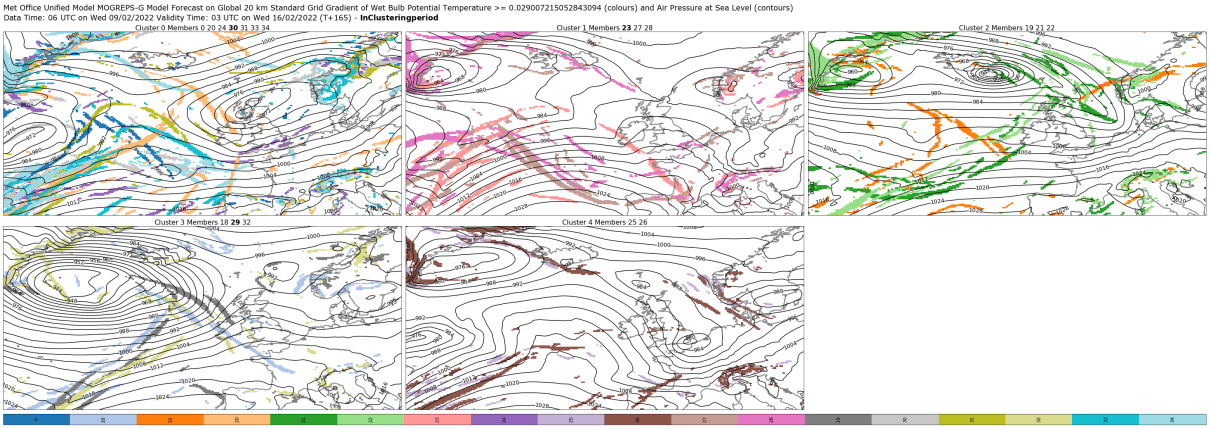


Figure 6.12: A paintball plot from week 4 of the testbed from $|\nabla\theta_w|$ at 850 hPa with the MSLP displayed of the representative member of each cluster from the 0600 UTC forecast on 07/02/2022 at lead time t+90, valid time 0000 UTC 11/02/2022, halfway through the window of interest. The paintball plots represent the threshold applied to the $|\nabla\theta_w|$ fields, where each member is represented by its own colour.

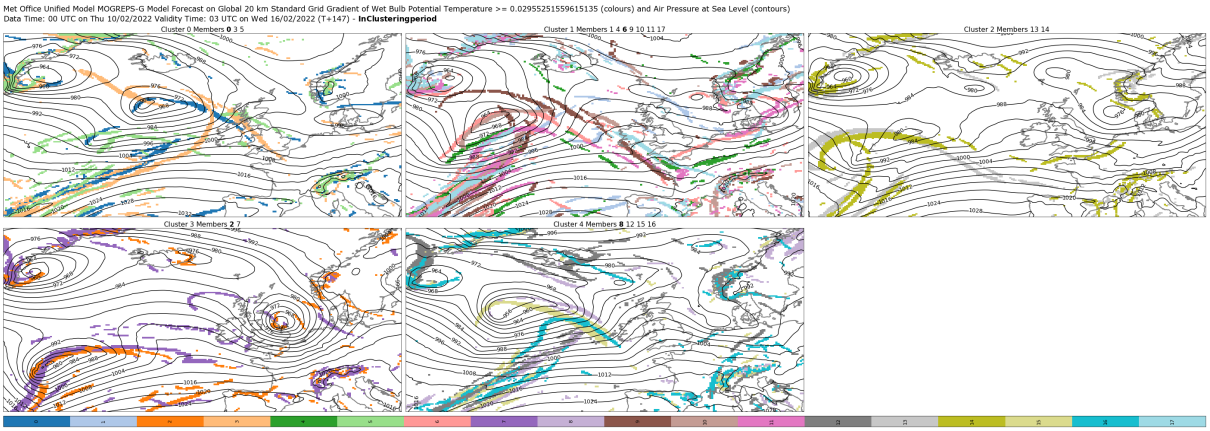
There is a discrepancy between the clustering participants observed from the 0600 UTC forecast on Monday January 31st and the clustering summary on figure 6.1. This was due to a bug fix, but the implementation was not yet available to participants until the next day. The bug caused a discrepancy in the optimal number of clusters chosen. For this reason, the 31st has been removed from the survey results of the case study, though the clustering products will still be presented for context.

6.4.1 Prevailing weather pattern leading up to the window of interest

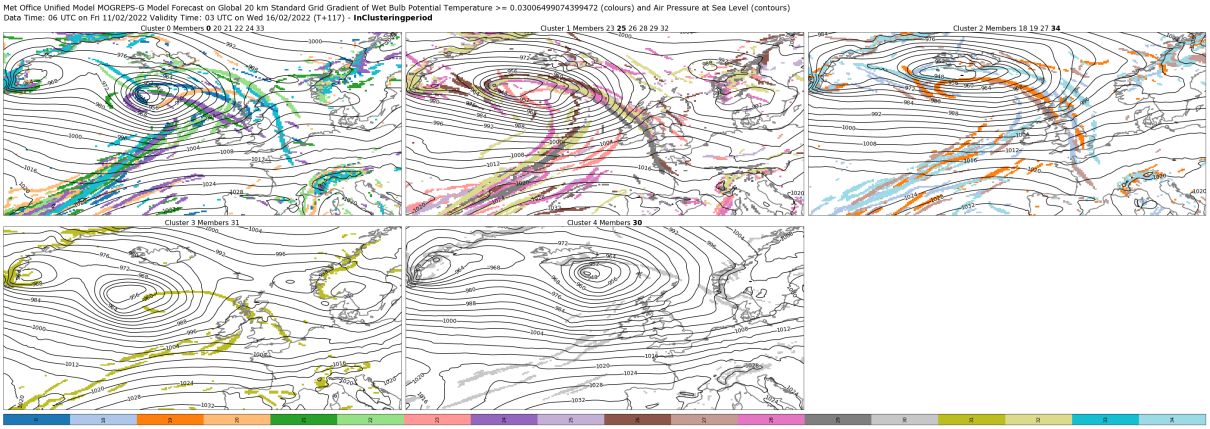
The prevailing weather patterns are key to understanding why an event has such a high level of uncertainty. Therefore, as week 3 contained a particularly uncertain event, it is necessary to evaluate these patterns before examining the clustering in depth. For this analysis, we'll examine the prevailing pattern from the first forecast to the approximate window of interest start time for all forecasts picking up on the same event. There is some variation in the window start time progressing across forecasts, but the valid time remains relatively close, varying from 0600 UTC on the 5th to 1500 UTC on the 6th. As the window lasts for 48 hours, there is enough overlap across forecasts to consider the clustering is focusing on a single event across valid times. Therefore, the prevailing weather patterns beginning on the 31st of January and ending just before the on the 6th of February will be considered.



(a)



(b)



(c)

Figure 6.13: Paintball plots from week 4 of the testbed from $|\nabla\theta_w|$ at 850 hPa with the MSLP displayed of the representative member of each cluster from the 0600 UTC forecast on 09/02/2022 at lead time t+165, valid time 0300 UTC 16/02/2022, in plot (a), the 0000 UTC forecast on 10/02/2022 at lead time t+147, valid time 0300 UTC 16/02/2022, in plot (b), and the 0600 UTC forecast on 11/02/2022 at lead time t+117, valid time 0300 UTC 16/02/2022, in plot (c). The paintball plots represent the threshold applied to the $|\nabla\theta_w|$ fields, where each member is represented by its own colour.

The week begins on the 31st of January with a high pressure system over the North Atlantic that remains to the southwest of the UK before beginning to break down on Wednesday the 2nd (figure 6.5, plots (g) and (h), figure 6.6, plots (a) to (d)). A low deepens to the east of Iceland on Thursday and a trough from the west brings fronts across the North Atlantic (figure 6.6, plots (e) and (f)). The low moves further to the north of Scotland between Iceland and Norway, and persists through Friday, while another low develops off the coast of Greenland (figure 6.6, plots (g) and (h)). The low moves further to the east and the flow becomes more zonal as the forecast moves into Saturday and Sunday, where the windows of interest begin (figure 6.7).

6.4.2 Progression of uncertainty within the window of interest

Figure 6.14 shows a 12 hour progression through the window of interest from the forecast on Monday 31/01/2022 at 0600. It begins with a valid time of 0600 06/02/2022, with a relatively zonal flow over the UK and a trough moving in, carrying a front (a, b). As the forecast progresses, a low begins to develop towards the west southwest of Iceland. This is an area of a lot of uncertainty, with the position of the associated fronts and the timing of the formation of the cyclone varying across the clusters (c, d). The variation in the clusters becomes stronger as some clusters show the cyclone dissipating and some show it deepening (e, f), which continues through the rest of the window (g to j).

On Tuesday (figure 6.15), from the 0600 UTC 01/02/2022 forecast, the window begins 24 hours earlier at a valid time of 0600 UTC 05/02/2022, so the initial uncertainty is in some western frontal regions that develop into a wave and the low seen in figure 6.14 (a to f). The development of the cyclone appears more certain as all five clusters show the storm in the later half of the window (g to j).

The window of interest again begins on Sunday with a valid time of 1500 UTC 06/02/2022 for the Wednesday forecast from 0600 UTC 02/02/2022 and opens with uncertainty on where the center of the storm is (figure 6.16) and the progression of the fronts. Some clusters have the fronts crossing the the UK earlier than others. The forecasts on Thursday (0600 UTC 03/02/2022) and Friday (0600 UTC 04/02/2022) both have windows that begin just slightly later at valid time 1500 UTC 06/02/2022, covering the same storm progression (figures 6.17 and 6.18). Again, the main uncertainty is regarding the

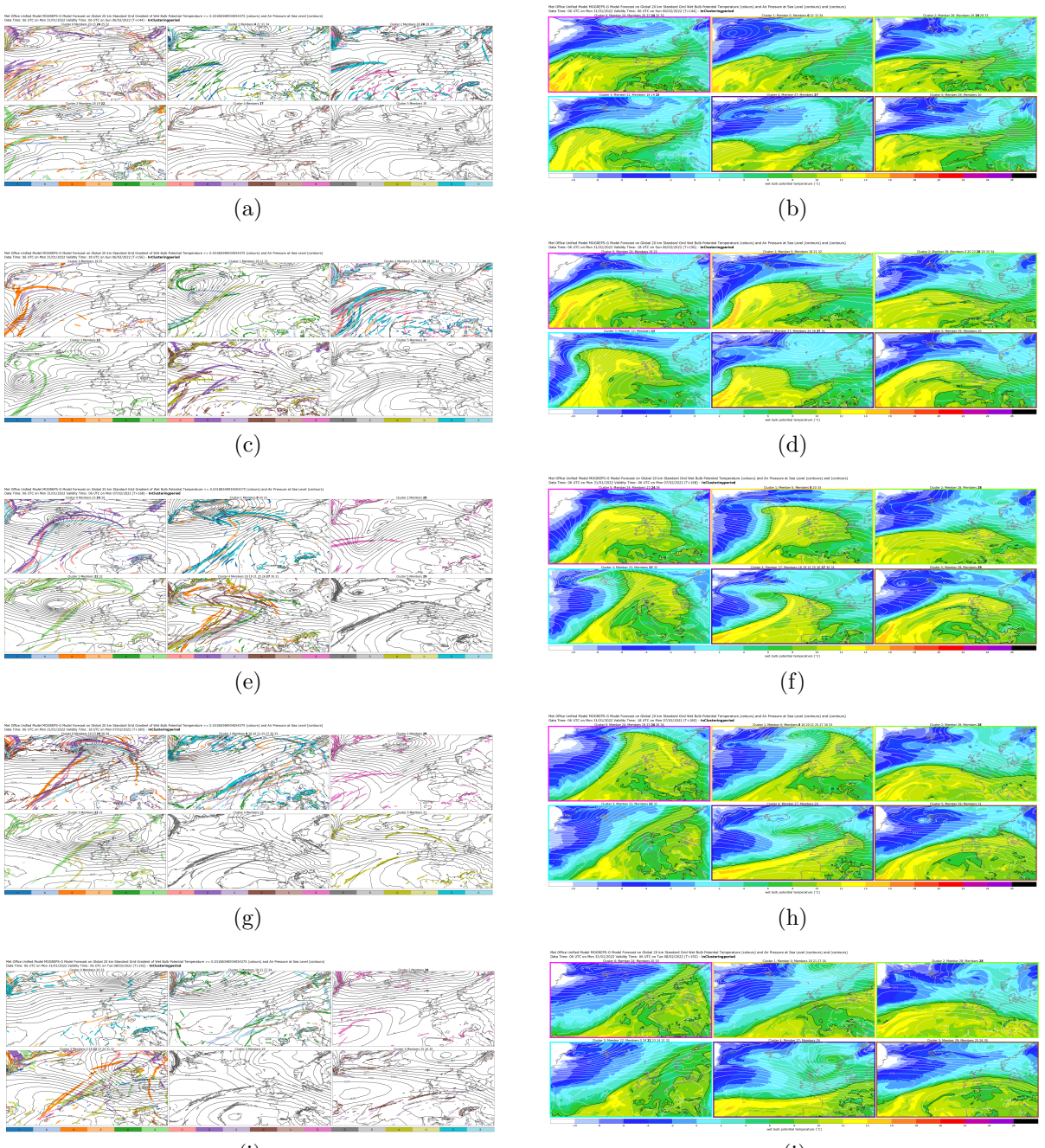


Figure 6.14: A comparison of paintball plots to representative member plots for the forecast at 0600 on Monday, 31/01/2022. The paintball plots (a, c, e, g, and i) represent the threshold applied to the $|\nabla\theta_w|$ at 850 hPa fields, where each member is represented by its own colour. The representative member plots (b, d, f, h, and j) are presented in θ_w at 850 hPa. Plots (a) and (b) are at valid time 0600 UTC 06/02/2022, plots (c) and (d) are at valid time 1800 UTC 06/02/2022, plots (e) and (f) are at valid time 0600 UTC on 07/02/2022, plots (g) and (h) are at valid time 1800 UTC 07/02/2022, and plots (i) and (j) are at valid time 0600 UTC on 08/02/2022.

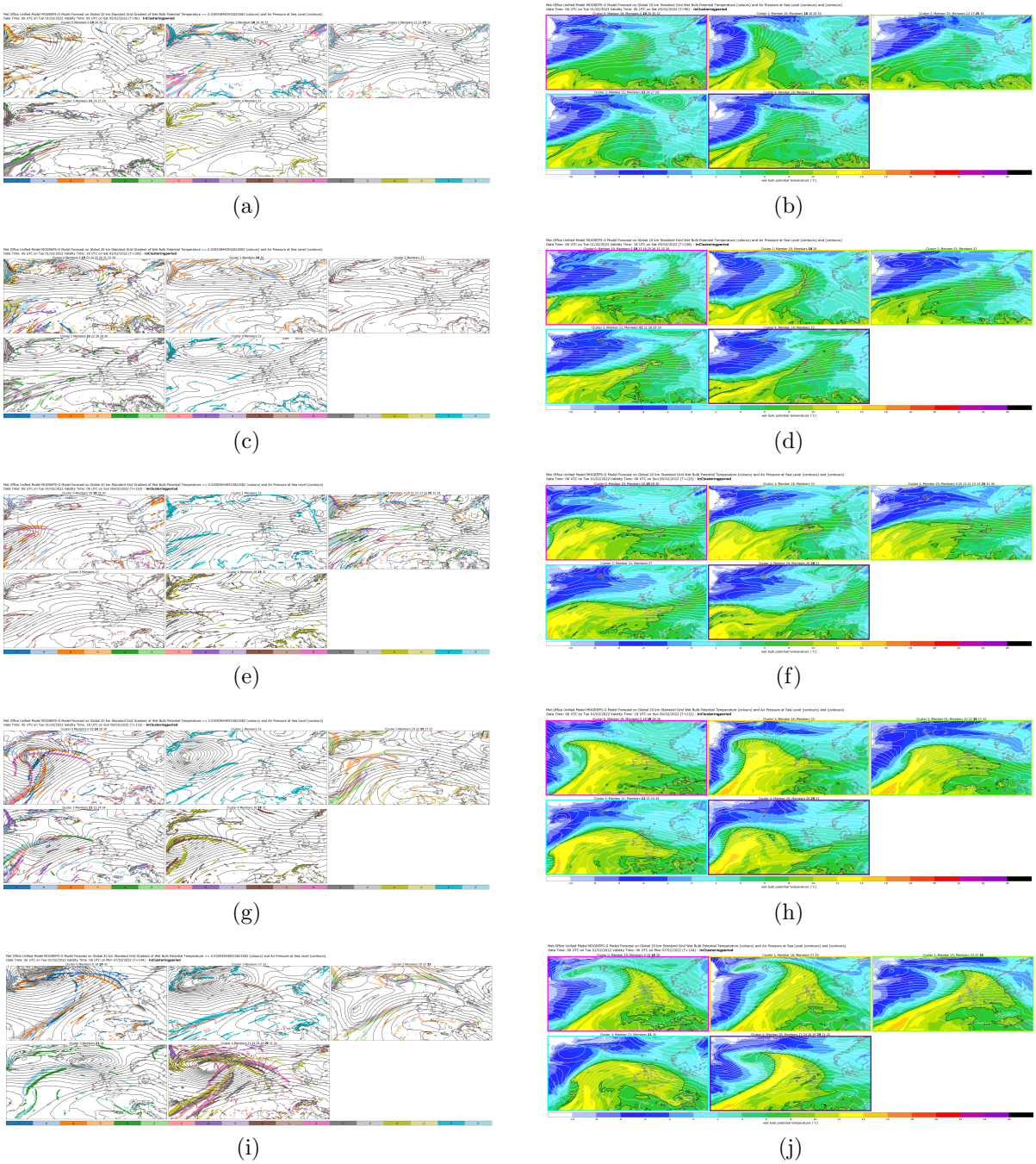


Figure 6.15: A comparison of paintball plots to representative member plots for the forecast at 0600 on Tuesday, 01/02/2022. The paintball plots (a, c, e, g, and i) represent the threshold applied to the $|\nabla\theta_w|$ at 850 hPa fields, where each member is represented by its own colour. The representative member plots (b, d, f, h, and j) are presented in θ_w at 850 hPa. Plots (a) and (b) are at valid time 0600 UTC 05/02/2022, plots (c) and (d) are at valid time 1800 UTC 05/02/2022, plots (e) and (f) are at valid time 0600 UTC on 06/02/2022, plots (g) and (h) are at valid time 1800 UTC 06/02/2022, and plots (i) and (j) are at valid time 0600 UTC on 07/02/2022.

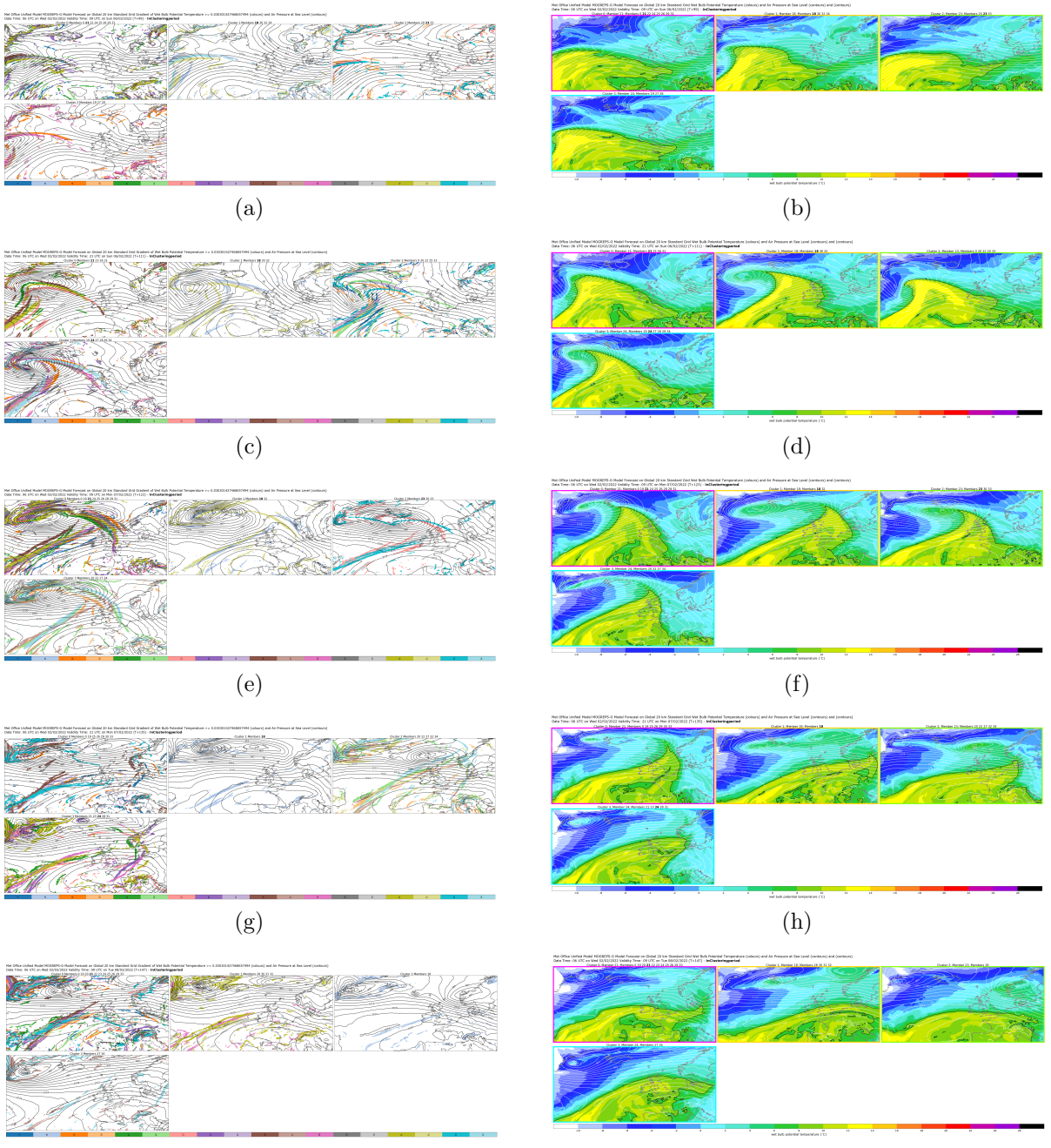


Figure 6.16: A comparison of paintball plots to representative member plots for the forecast at 0600 on Wednesday, 02/02/2022. The paintball plots (a, c, e, g, and i) represent the threshold applied to the $|\nabla\theta_w|$ at 850 hPa fields, where each member is represented by its own colour. The representative member plots (b, d, f, h, and j) are presented in θ_w at 850 hPa. Plots (a) and (b) are at valid time 0900 UTC 06/02/2022, plots (c) and (d) are at valid time 2100 UTC 06/02/2022, plots (e) and (f) are at valid time 0900 UTC on 07/02/2022, plots (g) and (h) are at valid time 2100 UTC 07/02/2022, and plots (i) and (j) are at valid time 0900 UTC on 08/02/2022.

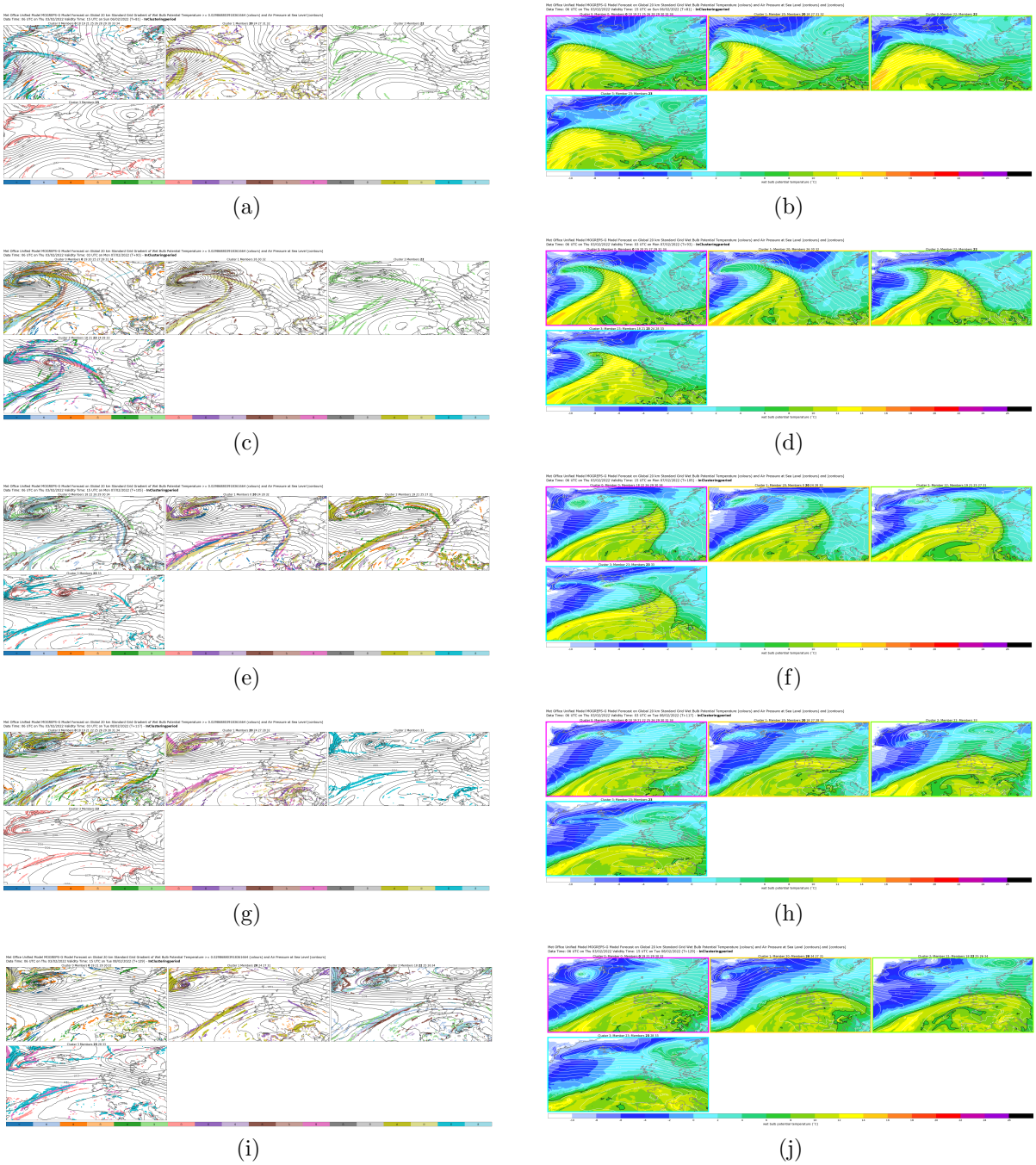


Figure 6.17: A comparison of paintball plots to representative member plots for the forecast at 0600 on Thursday, 03/02/2022. The paintball plots (a, c, e, g, and i) represent the threshold applied to the $|\nabla\theta_w|$ at 850 hPa fields, where each member is represented by its own colour. The representative member plots (b, d, f, h, and j) are presented in θ_w at 850 hPa. Plots (a) and (b) are at valid time 1500 UTC 06/02/2022, plots (c) and (d) are at valid time 0300 UTC 07/02/2022, plots (e) and (f) are at valid time 1500 UTC on 07/02/2022, plots (g) and (h) are at valid time 0300 UTC 08/02/2022, and plots (i) and (j) are at valid time 1500 UTC on 08/02/2022.

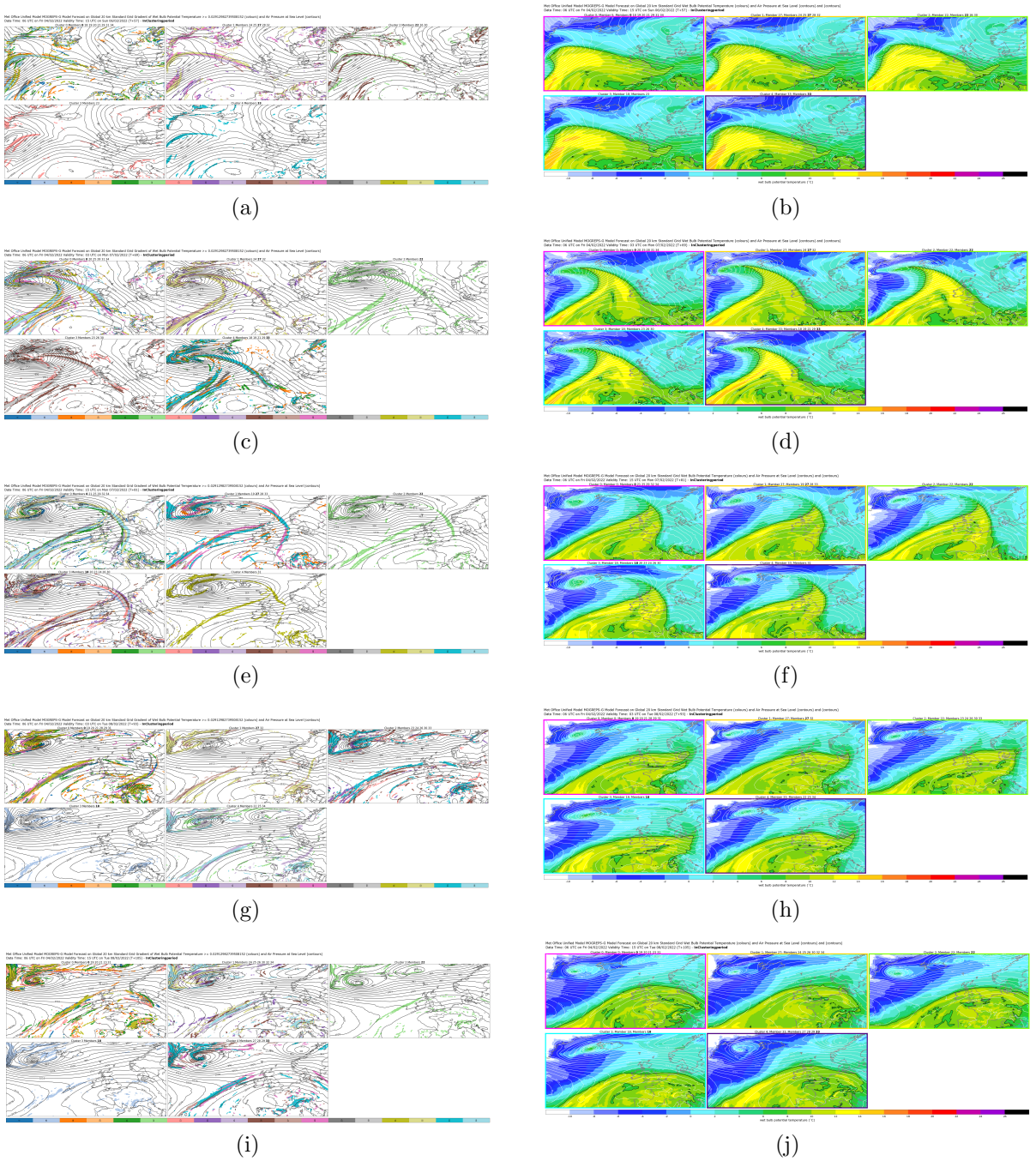


Figure 6.18: A comparison of paintball plots to representative member plots for the forecast at 0600 on Friday, 04/02/2022. The paintball plots (a, c, e, g, and i) represent the threshold applied to the $|\nabla\theta_w|$ at 850 hPa fields, where each member is represented by its own colour. The representative member plots (b, d, f, h, and j) are presented in θ_w at 850 hPa. Plots (a) and (b) are at valid time 1500 UTC 06/02/2022, plots (c) and (d) are at valid time 0300 UTC 07/02/2022, plots (e) and (f) are at valid time 1500 UTC on 07/02/2022, plots (g) and (h) are at valid time 0300 UTC 08/02/2022, and plots (i) and (j) are at valid time 1500 UTC on 08/02/2022.

fronts moving towards the UK and when and where they will impact. As the beginning of the window gets closer to the forecast initialization day and time, the variability in the RMs gets smaller.

6.4.3 Survey results

The survey provided to testbed participants covered several important themes with regards to the method and its products. The first set of questions (see appendix A) sought to quantify the extent to which the clustering method matched participants' judgements about what is considered a distinct weather scenario. The second set of questions was aimed at determining how much the clustering results might impact the participants' forecast. The next section questioned the extent to which the clusters pointed to potential high impact weather events. Lastly, the participants were asked about how efficient using the method was compared to looking through the whole ensemble. Within the following sections the survey questions and their results will be explored in detail.

6.4.3.1 Do clusters represent what constitutes a distinct weather scenario?

Whether or not clusters represent distinct weather scenarios is a theme in the survey and it can be explored by asking several questions, which are listed as follows:

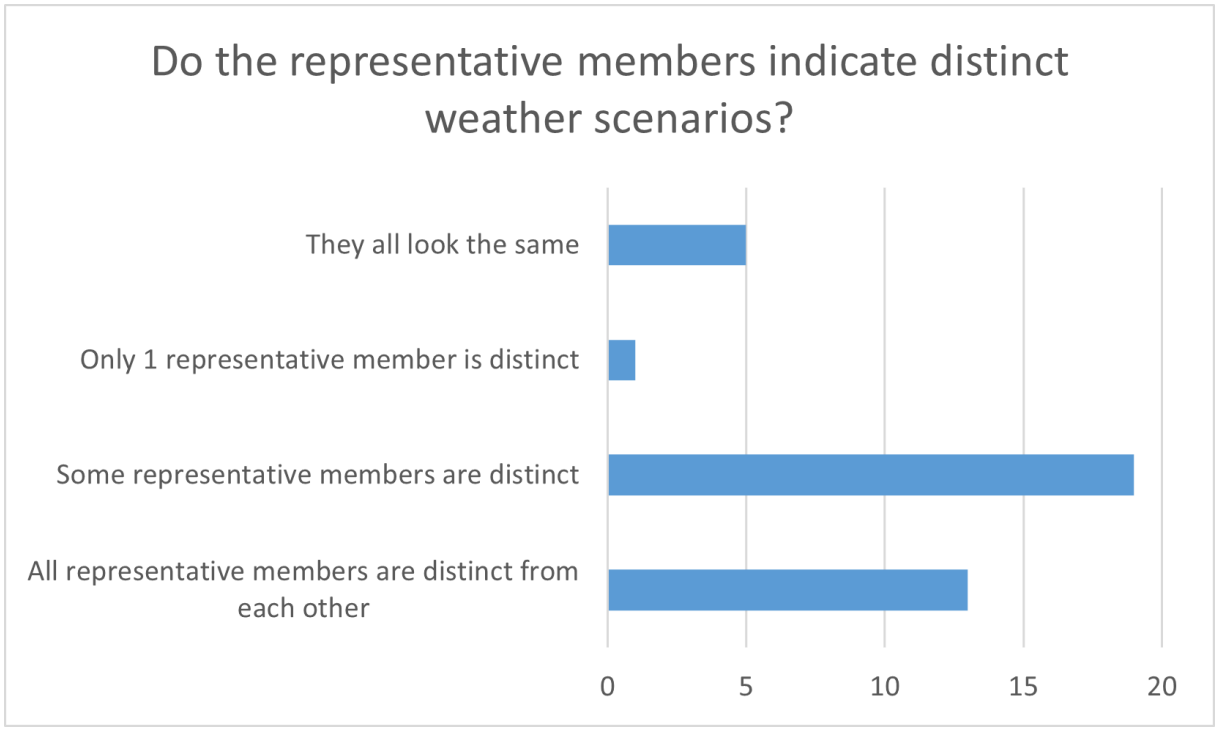
1. Do the representative members indicate distinct weather scenarios?
2. Explain why they are distinct or not.
3. Is there an important meteorological event in the full ensemble that does not have a close representative member?
4. If yes, which member(s) and which representative member(s) are they different from?
5. Would you cluster the members similarly to how they are being clustered (using only the lead time at the beginning of the window of interest)?
6. If you answered "no" to the previous question, please elaborate on how you would cluster members differently.
7. Are there too many or too few clusters?

The goal of the clustering method is to group members together in a way that provides an optimal number of clusters that produce distinct representative members. Therefore, these questions were chosen specifically to measure how effective the clustering method was compared to a forecaster's judgement of an ensemble forecast. To answer these questions, participants were asked to view the 48 hour window of interest and assess whether the RMs represented distinct weather scenarios. The representative members and the ensemble postage stamp plots were provided in the wet-bulb potential temperature, the gradient of the wet-bulb potential temperature, the rainfall accumulation, the snowfall accumulation, the precipitation rate, and the maximum wind gust. The following bullet points and discussion will address each question in turn, though related questions (such as items 1 and 2) will be joined together in a single section.

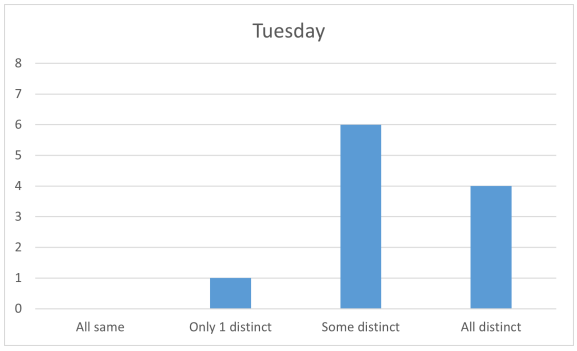
Q1. Do the representative members indicate distinct weather scenarios?

Q2. Explain why they are distinct or not.

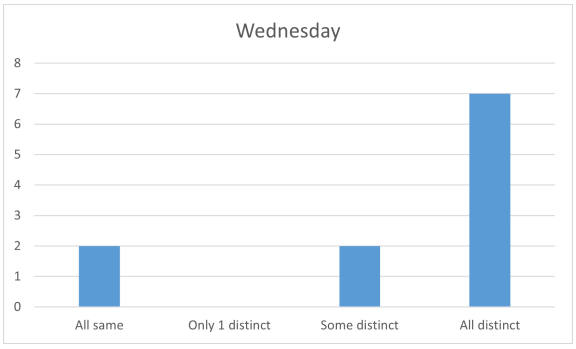
This question, and the follow up explanation, is to test of how unique the representative members are from one another and if there's enough distinction between them to warrant separate potential scenarios for review. Figure 6.19 contains bar charts of the answers provided in the survey, with (a) representing the total for the week and (b) through (e) representing each individual day during the week, respectively. 86% of responses through the week indicated that some or all of the clusters represented different scenarios. Some examples of the explanations for responses that indicated they all looked the same or only one scenario was distinct included: "at T+96 (start of period) only cluster 1 is noticeably different - the others all have westerly flow over the UK with precip[itation] over Scotland. By T+120 they're all pretty similar, with westerly flow and a front oriented E-W across the middle of the UK around T+120, with only small differences in the precise location and orientation of this front" on Tuesday, "all show some form of deep Atlantic low pressure with cold front crossing the British Isles late on Monday" on Wednesday, "evolution of the developing lows to north and progression of fronts very similar in all clusters," on Friday. These responses appear to indicate the participants were expecting radically different scenarios instead of scenarios that are roughly similar but differ in key areas like timing, position, and intensity. This is supported by the following examples of the explanations of the participants who responded some or all of the RMs are distinct



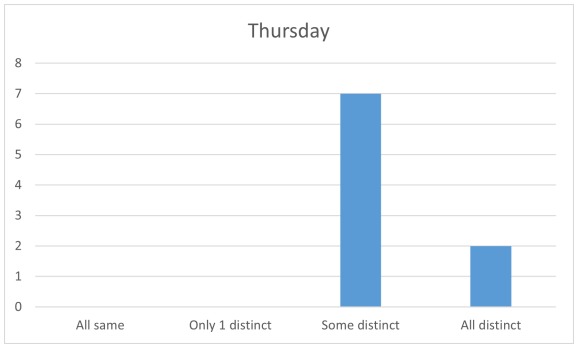
(a)



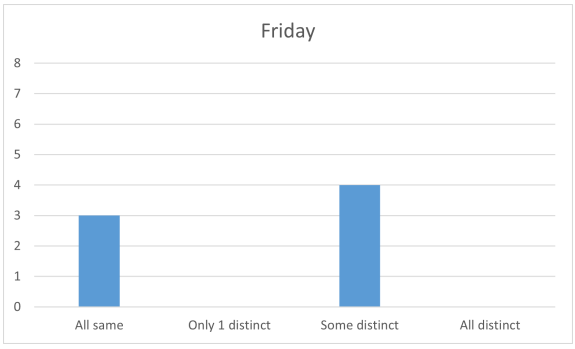
(b)



(c)



(d)



(e)

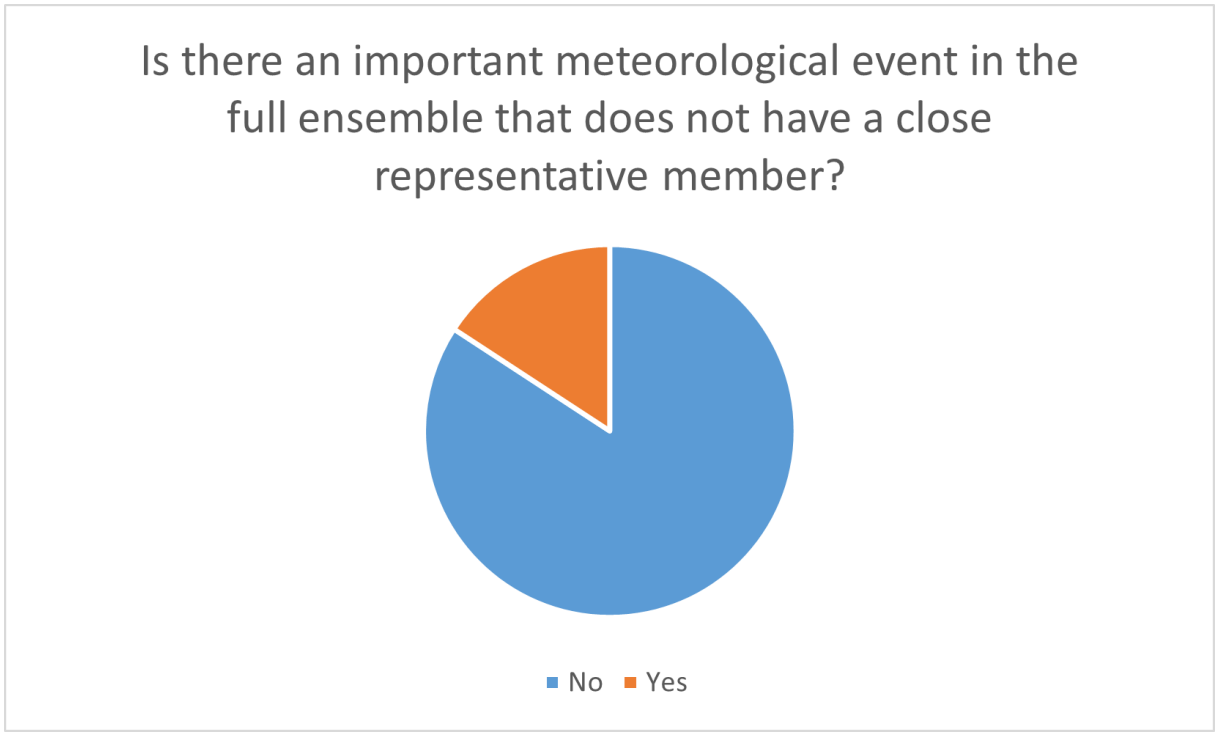
Figure 6.19: Bar plots on if the representative members indicate distinct weather scenarios. The results for the entire week (excluding Monday) are in (a), with 5 responses indicating the all look the same, 1 response indicating only one representative member is distinct, 19 indicate some are distinct, and 13 indicate all are distinct. Bar plots for Tuesday through Friday are in (b) to (e).

from one another: "each has a slightly different position of frontal systems" on Tuesday, "all representative members develop the deep low in the NW Atlantic in different ways and thus bring time differences to the frontal passage across the UK. Also the shape (level of elongation and multi-centres, or not) of the low to the north of the UK are different across the representative members" on Wednesday, "they all tell broadly the same story, differences mainly relating to uncertainties in timing of the frontal system. Even then, I'd say clusters 0 and 1 had no significant differences between each other. Cluster 2 is noticeably quicker with the progression of the frontal system. Cluster 3 is quicker as well, but also takes the cold front much further south by the end of the period, and has a unique evolution in terms of the low pressure systems in the Atlantic" on Thursday, and "distinct only on the position of the front over the UK, and perhaps how the 1st low occludes when moving to the north of the UK" on Friday. This indicates that what participants tended to choose depended greatly on how they viewed the forecast in general and what they were looking for in terms of features that would make them notably distinct. Some participants considered timing, position, or intensity to be enough to conclude the RMs were distinct while others looked at the general atmospheric pattern and considered them all the same unless there was a notable deviation.

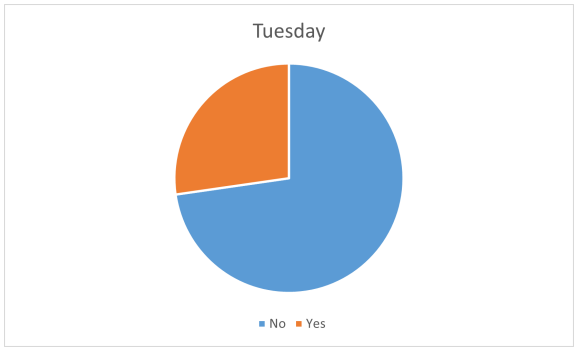
Q3. Is there an important meteorological event in the full ensemble that does not have a close representative member?

Q4. If yes, which member(s) and which representative member(s) are they different from?

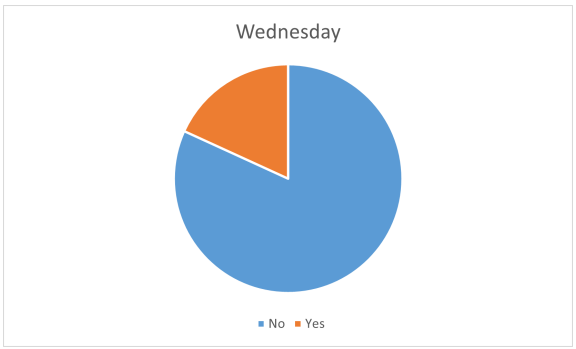
It is important that the RMs are both distinct and represent the different potential scenarios adequately. Therefore, it is vital to know if the RMs miss a potential scenario or meteorological event and why it was missed. If a single member is quite different than the rest of the ensemble, it is likely to be considered an outlier and grouped with other members in the closest cluster. This would result in that member likely not having a strong similarity with an RM. However, if this outlying member is distinct enough throughout the window, it will likely be the RM of a cluster and predominantly on its own (see section 3.4.4.2). During the week (figure 6.20), 84% of responses indicated they were confident no important meteorological events were missed or were too dissimilar from their representative member. However, during this week participants picked up on a few members that stood out within the forecasts. How significant these stand out members



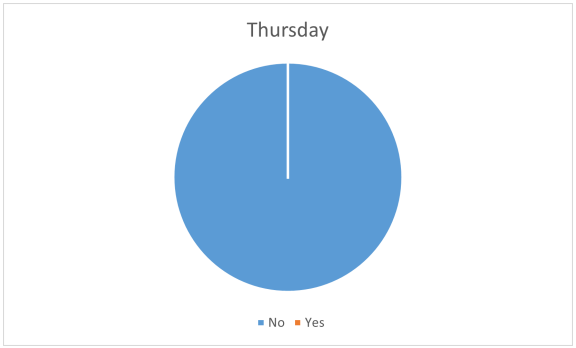
(a)



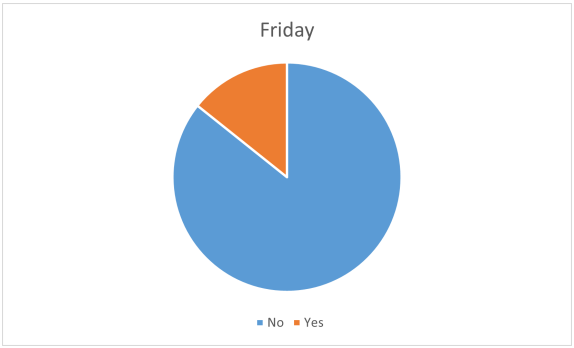
(b)



(c)



(d)



(e)

Figure 6.20: Pie charts on whether or not the representative members do not adequately represent an important meteorological event in the full ensemble. The results for the entire week (excluding Monday) are in (a), and Tuesday through Friday are in (b) to (e).

were and if they should have been their own cluster would be an area of future study and refinement of the method, if necessary.

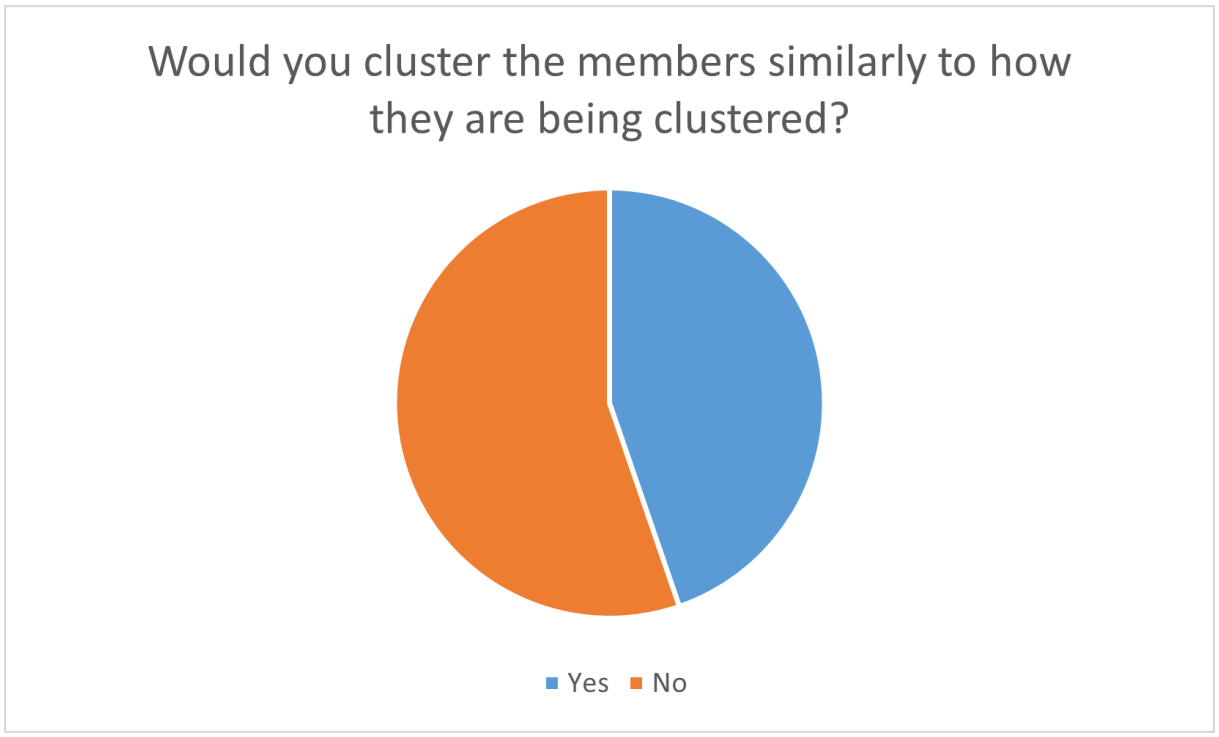
Q5. Would you cluster the members similarly to how they are being clustered?

Q6. If you answered "no" to the previous question, please elaborate on how you would cluster members differently.

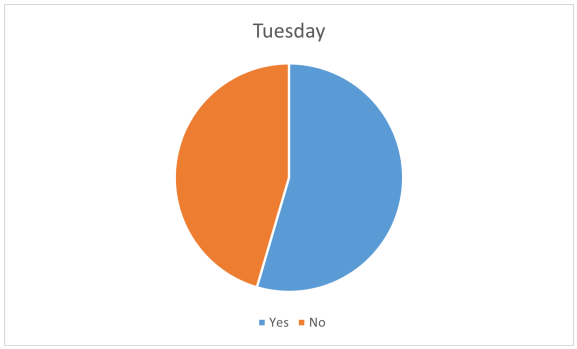
These questions had the lead time restricted to only the first lead time in the window of interest, which likely effected the results. The start of the window of interest is when clusters are beginning to become distinct, when the sum distance has dropped to the 25^{th} percentile, but is not yet at its lowest point. Therefore, clustering at the beginning of the window will often be less distinct than when viewed later in the window. This question would have been better presented at the point when the sum distance had dropped to its lowest point within the window, where clusters would be the most distinct. 55% of responses indicated that participants would not cluster this particular lead time the same way (figure 6.21) and would choose fewer clusters. Over the course of the testbed, this has led to a potential idea for future development and refinement of the method, where the number of clusters was not fixed throughout the forecast but was allowed to change to best fit the data at each lead time. However, this would be a significantly challenging task as allowing the optimal number of clusters to fluctuate at each lead time would make it very difficult to establish traceability between lead time clusters.

Q7. Are there too many or too few clusters?

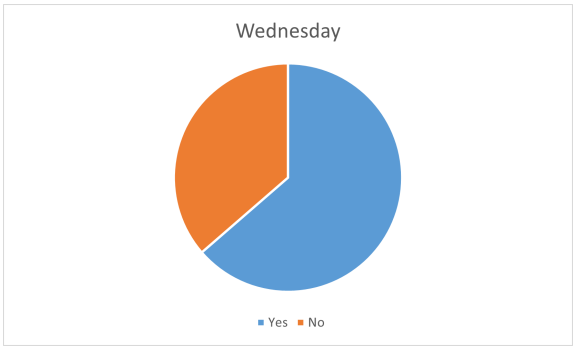
Closely related to the previous question, but with the freedom to examine the entire window of interest, the question of how many clusters is the correct number can depend greatly on what the participants want to see within the forecast, i.e. if temporal or spatial displacements are distinct scenarios, and a potential bias as to what number people might prefer. Both Tuesday and Friday had 5 clusters and Wednesday and Thursday had 4. The responses in figure 6.22 don't appear to show a strong bias for either 4 or 5 clusters, but more data would be beneficial. Figure 6.22 also echos the results from the previous question, with 55% of responses indicating there are too many clusters over the week. The majority of comments from participants who said they would not cluster the members the same way and that there were too many clusters said they would combine some of



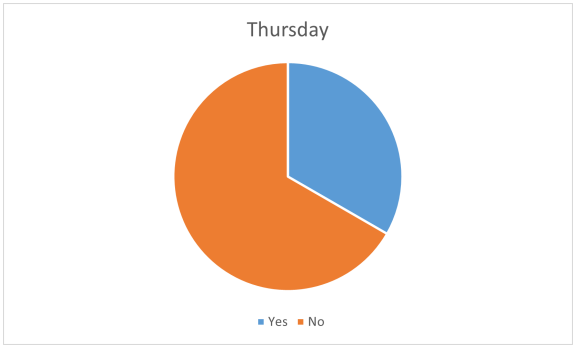
(a)



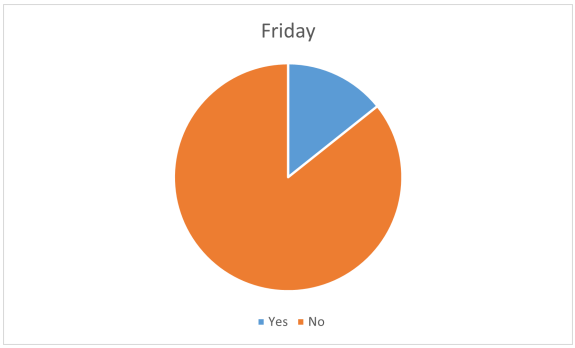
(b)



(c)

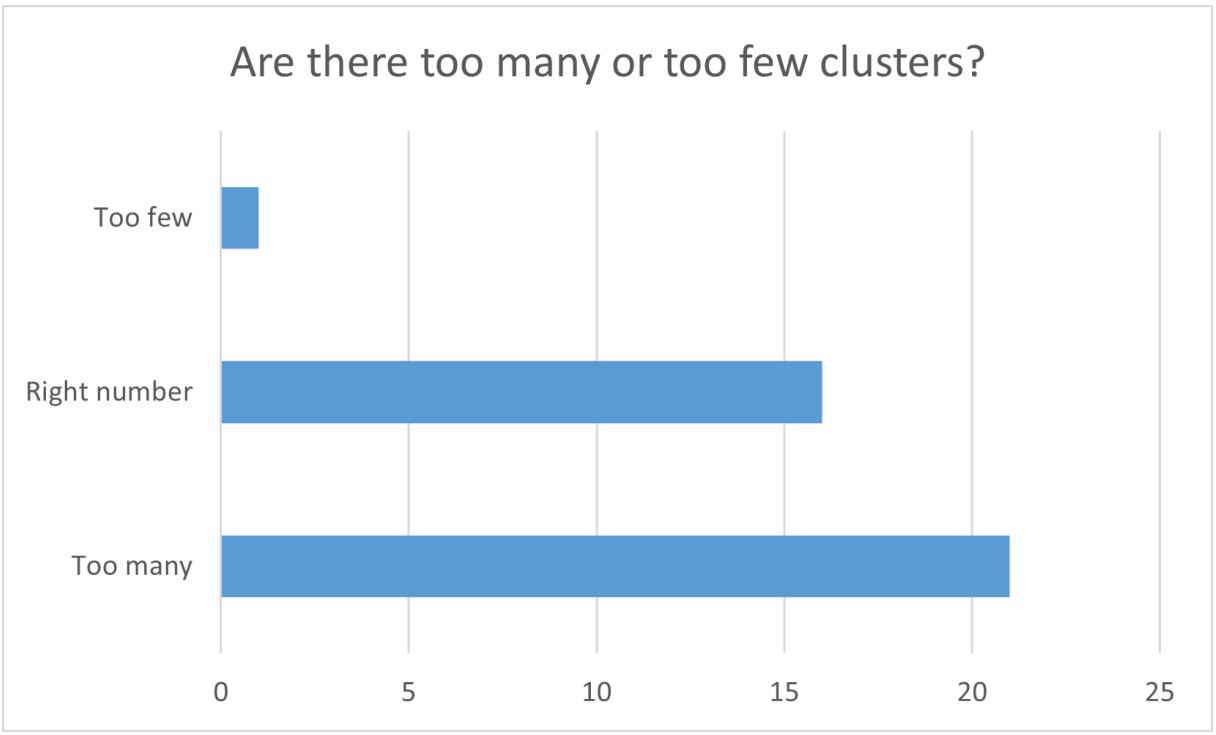


(d)

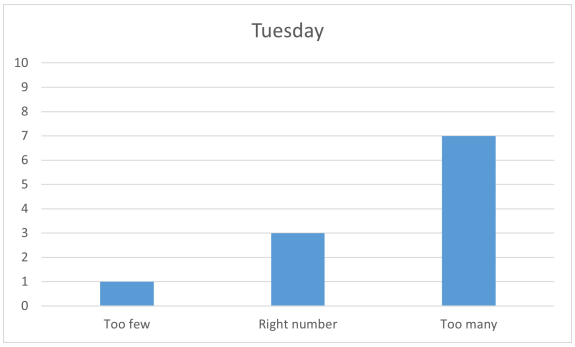


(e)

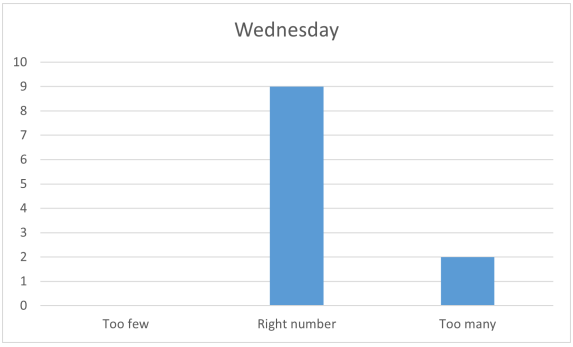
Figure 6.21: Pie charts on whether or not participants would cluster the members similarly. This was restricted to the beginning of the window of interest. The results for the entire week (excluding Monday) are in (a), and Tuesday through Friday are in (b) to (e).



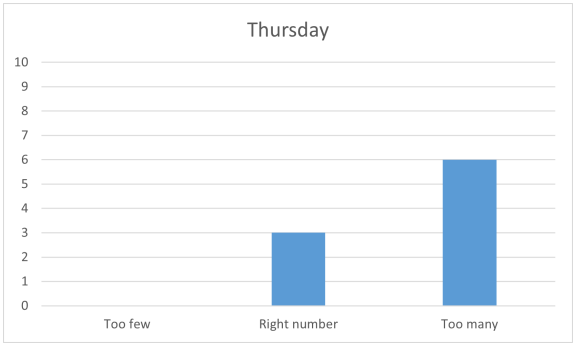
(a)



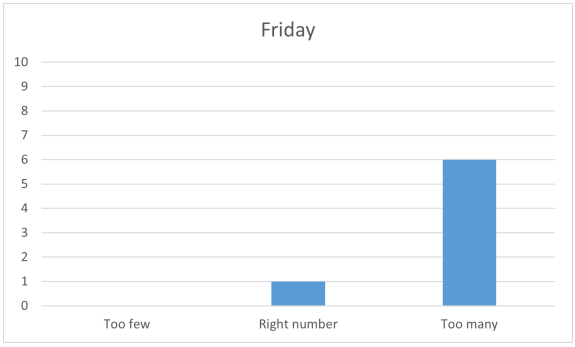
(b)



(c)



(d)



(e)

Figure 6.22: Bar plots on if the number of clusters is an accurate representation of scenarios operational meteorologists see within the full ensemble. The results for the entire week (excluding Monday) are in (a), with 1 response indicating too few clusters, 16 responses indicating the right number of clusters, and 21 responses indicating too many clusters. Bar plots for Tuesday through Friday are in (b) to (e).

the clusters as there wasn't enough distinction between them. This could be addressed by choosing a different threshold when comparing members, adjusting the requirements for how the optimal number of clusters are chosen, or by reducing the domain size to encompass a noteworthy event for clustering instead of looking at a larger area.

6.4.3.2 What impact does the clustering algorithm have on forecasting and communication with end users?

The following questions were used for this section:

1. Would the cluster information (representative members) for this forecast be useful in creating a forecast communication to the general public (e.g., informing the warning impact matrix)?

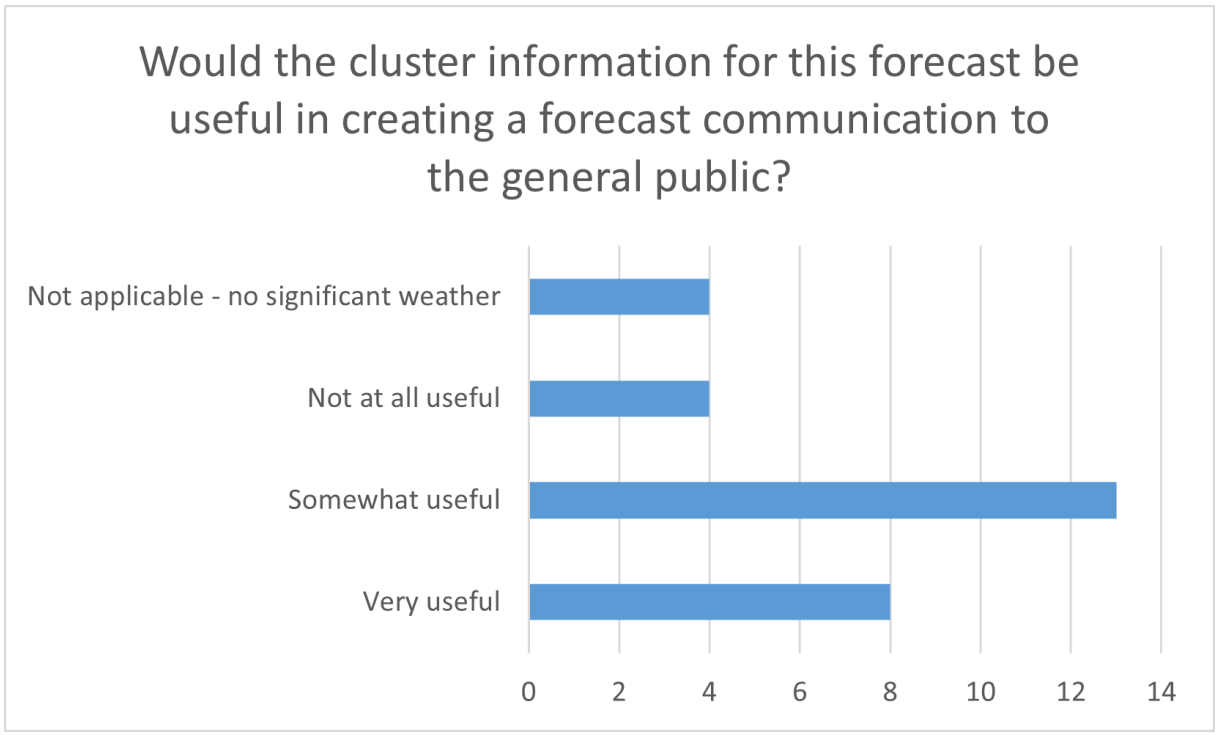
2. Is the cluster information useful in creating a forecast message to specific users?

3. If you answered "very useful" or "somewhat useful" to the previous questions, what areas of interest would it be for (i.e., emergency response, local authorities, aviation)?

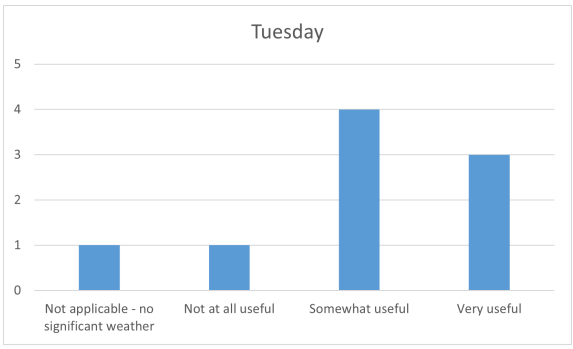
These questions were chosen to gauge how useful the products (i.e. the representative members) were for the participants to shape their forecasts and how the information might be used in communicating with various industries and the public. The method was designed to reduce the amount of time forecasters needed to spend digesting ensemble data, while still retaining the accuracy of the forecasts. It is therefore important to determine if participants found the method adequate for this purpose, which will be explored in the following bullet points.

Q1. Would the cluster information (representative members) for this forecast be useful in creating a forecast communication to the public (e.g., informing the warning impact matrix)?

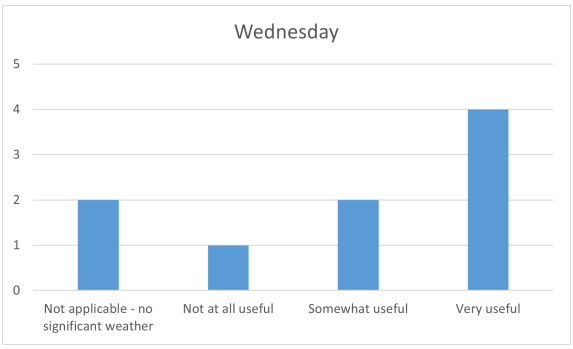
One of the goals of this method is to provide useful information from the ensemble for operational meteorologists to use in creating their forecasts. This question is key for gauging how much an operational meteorologist might use or rely on the method products to influence their forecasts. First, we ask if the information is useful for informing



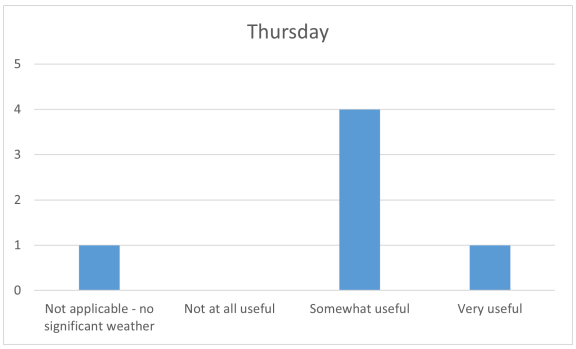
(a)



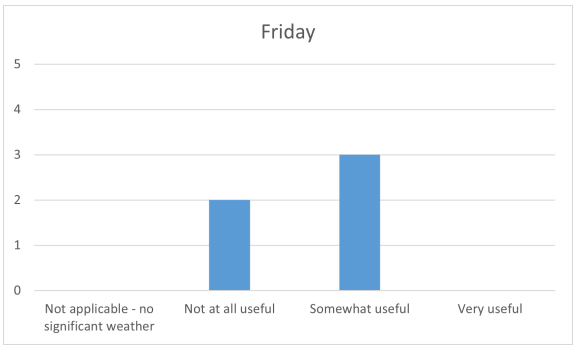
(b)



(c)



(d)



(e)

Figure 6.23: Bar plots on if the cluster information is useful in creating a forecast communication to the general public. The results for the entire week (excluding Monday) are in (a), with 4 responses indicating this question is not applicable, 4 responses indicating it is not at all useful, 13 responses indicating it is somewhat useful, and 8 responses indicating it is very useful. Bar plots for Tuesday through Friday are in (b) to (e).

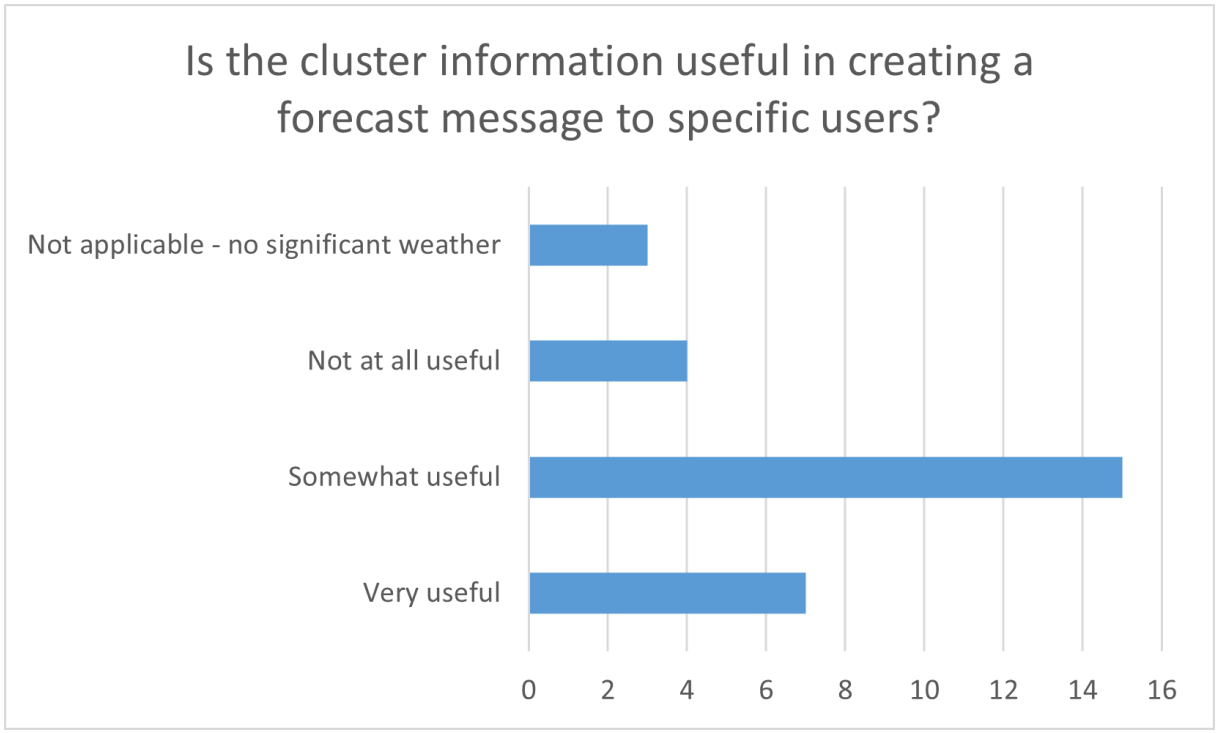
the warning impact matrix for the general public. In figure 6.23 we can see that the majority of participants considered the current products somewhat or very useful for this purpose at 72% of responses. This trend was evident every day for this particular weather scenario (weather event (iii) in section 6.3). This is also important in light of responses to previous questions suggesting there may be too many clusters for this event and not enough variation between them. Even though refinement of the method will be beneficial, the current products are already providing an important service for operational meteorologists.

Q2. Is the cluster information useful in creating a forecast message to specific users?

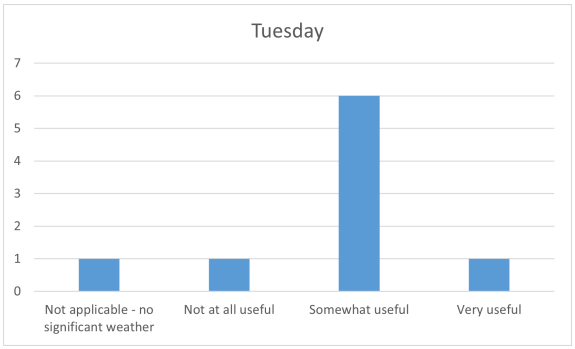
Similar to the previous question, this question focused on specific users and resulted in nearly the same answer at 75% of responses saying some or very useful. The majority of participants found this was the case throughout the week for specific users. A common theme of the comments related to how the variation between RMs could be useful, or for general longer lead time guidance or early warnings.

Q3. If you answered "very useful" or "somewhat useful" to the previous questions, what areas of interest would it be for (i.e., emergency response, local authorities, aviation)?

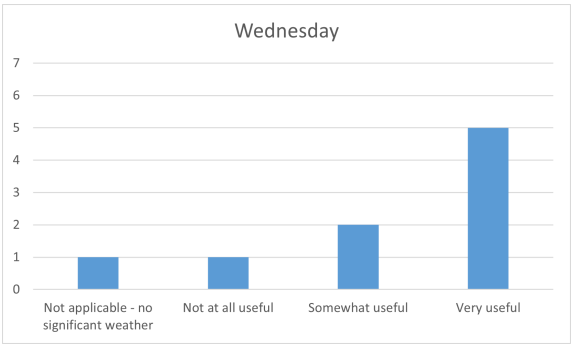
The responses to this question varied. Some responses noted specific sectors, such as aviation, emergency response, and marine forecasts. Others commented that the products were useful in a more general aspect. Some notable comments included: "in this sort of situation there's enough consistency in the ensemble to get a picture of the main themes in the weather story, but enough differences to appreciate some uncertainty in smaller scale details such as the timings of frontal systems and the track of the low in the Atlantic near the end of the period," "as the position of the front (and associated rainfall) is different in each cluster this would be useful for showing the range of possible outcomes in a more in-depth TV/video forecast," "as always, useful for informing overarching guidance products," and "as often the case, very useful for a general heads up as to the main themes of the forecast and therefore informing the potential for issuing of warnings etc. Because the representative members are broadly similar across the UK, specific users, e.g. those with site specific requirements, may also benefit from this output." These responses



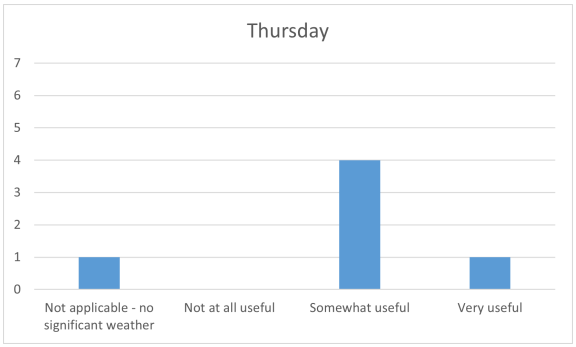
(a)



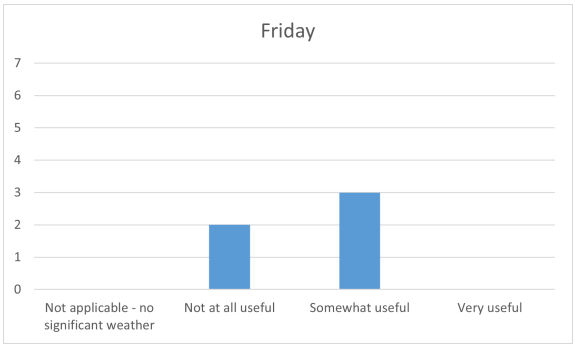
(b)



(c)



(d)



(e)

Figure 6.24: Bar plots on if the cluster information is useful in creating a forecast message to specific users. The results for the entire week (excluding Monday) are in (a), where 3 responses indicate this question is not applicable, 4 responses indicating it is not at all useful, 15 responses indicating it is somewhat useful, and 7 responses indicating it is very useful. Bar plots for Tuesday through Friday are in (b) to (e).

indicate that the representative members can provide a good indication of the variability in the atmosphere without having to examine the entire ensemble, potentially allowing forecasters to communicate the variability more effectively with end users.

6.4.3.3 Does the method detect high-impact scenarios? Are scenarios that have a low predictability detected across forecasts at the same valid time?

The method uses the gradient of the wet-bulb potential temperature, which is associated with frontal zones and therefore the potential for high-impact scenarios. To determine if the clustering algorithm picked up on these events, participants were asked questions related to high-impact weather within the RMs. There is also the possibility that some events that have a low predictability are picked up by the method across valid times over several forecasts until the trajectory of the scenarios converge into a single solution. During the winter, the likelihood of such events occurring is high, so a series of questions were also prepared to get participants feedback if and when they occurred.

1. Are any of the representative members at the lead time provided displaying a possible high impact (i.e., is there a possibility this member would require issuing a warning for rain, wind, snow etc.) scenario?

2. What type of high impact weather is associated with the scenario?

3. How many clusters contain a high-impact scenario?

4. Is the high-impact scenario you previously noticed present across multiple initializations?

5. Is the presence of this high-impact event as a potential scenario likely to impact your forecast message?

6. Why or why not?

7. Do the scenarios appear across multiple initializations?

If the answer to question 1 was 'no' participants were directed to question 7. This allowed for analysis across valid times even when high-impact scenarios weren't present.

This particular case included an event that had the potential for high-impact weather, so most participants answered questions 2 to 6. The results to all the questions will be explored in the following bullet points.

Q1. Are any of the representative members at the lead time provided displaying a possible high impact (i.e., is there a possibility this member would require issuing a warning for rain, wind, snow, etc.) scenario?

The responses to this question were divided (figure 6.25), though 69% indicated they saw potential high-impact scenarios within the window of interest of the RMs over the week. Mid-latitude storms, which the clustering window predominantly focused on, tend to be associated with strong winds and heavy rain or snow. The next questions explores what types of high-impact weather participants expected.

Q2. What type of high impact weather is associated with the scenario?

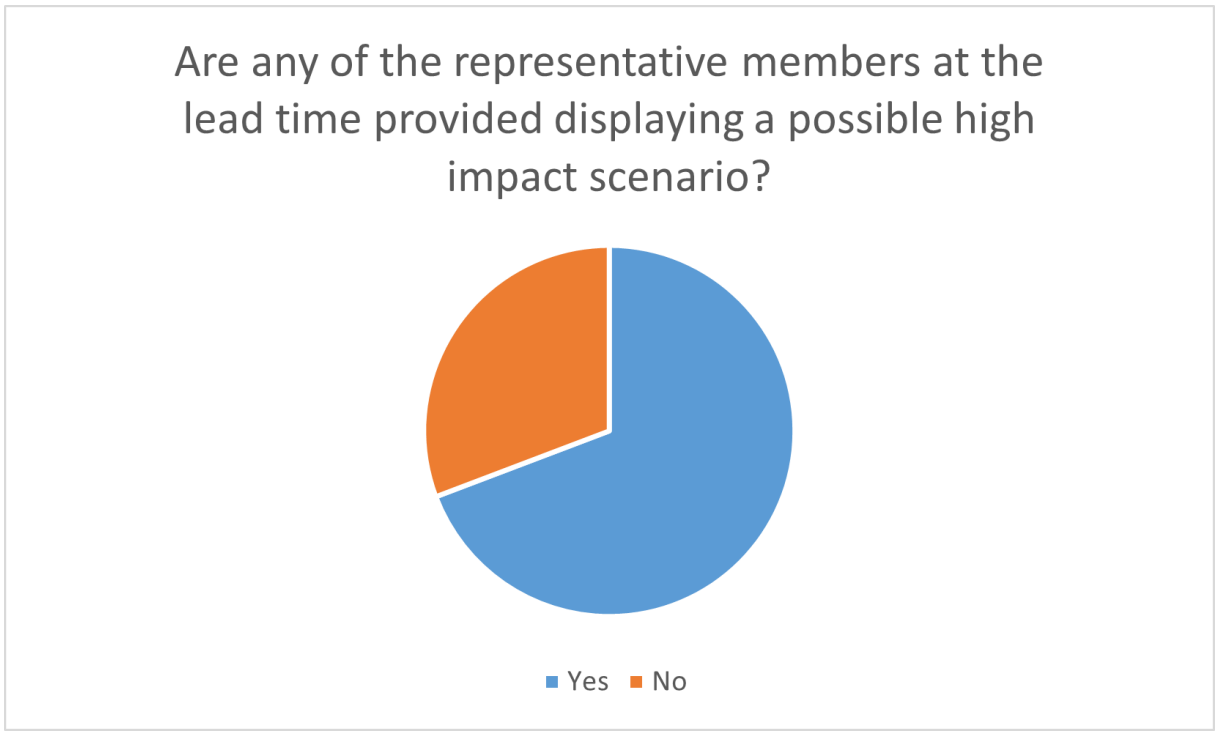
We can see from the previous question that the majority of participants did see some high-impact potential during the window of interest. As the domain size is rather large, and the window is over a 48 hour period, their focus may or may not have been specifically directed towards high-impact weather over the UK (though it is fairly likely) and it may not have been related specifically to the storm. However, the majority of responses (figure 6.26) are associated with high wind (89% of 18 responses), snow (50%), and heavy rain (33%), likely all related to when the frontal regions of the storm cross the UK.

Q3. How many clusters contain a high-impact scenario?

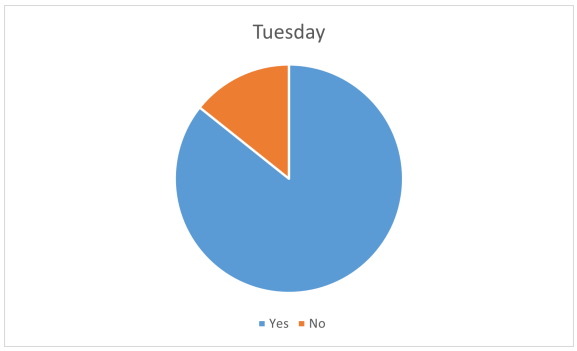
This question sought to establish whether or not a high-impact scenario was present in 1 or more clusters and only appeared if participants indicated they saw a potential high-impact scenario in the previous questions. Responses varied in the beginning of the week (figure 6.27), however as the event neared, the potential high-impact scenario appeared in more clusters, indicating the uncertainty of the event happening was reducing.

Q4. Is the high-impact scenario present across multiple initializations?

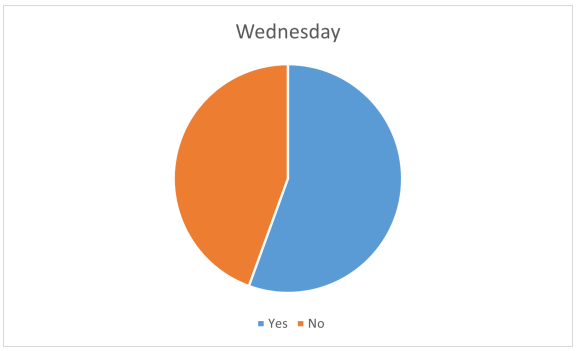
During the survey, participants were asked to look back through older forecasts, up to 24 hours (four initializations), to determine if the high-impact scenario they previously



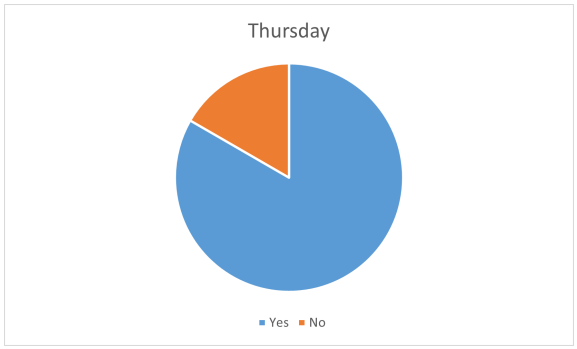
(a)



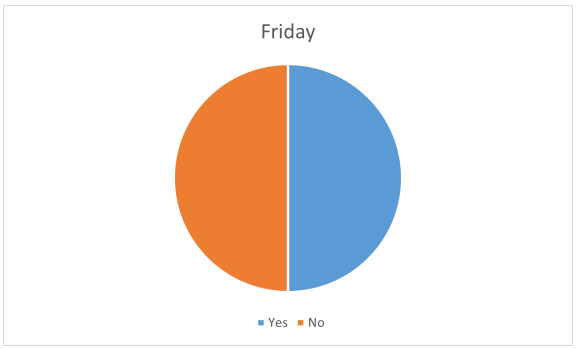
(b)



(c)

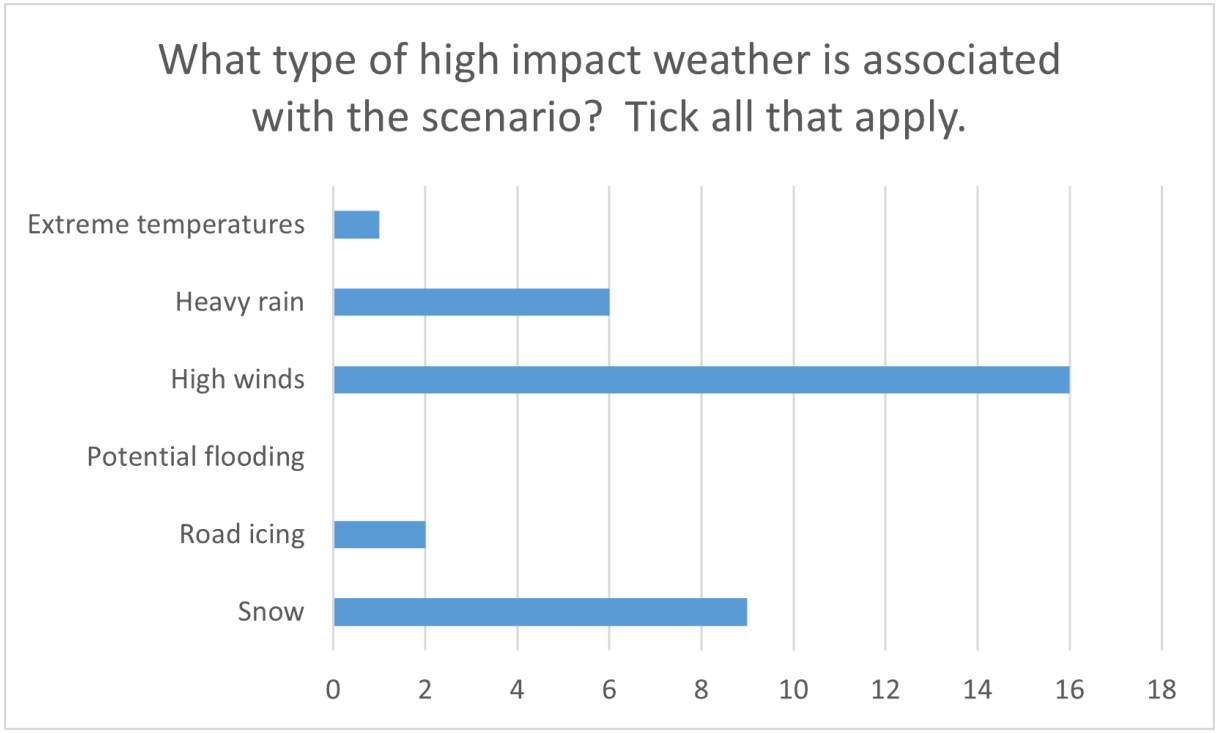


(d)

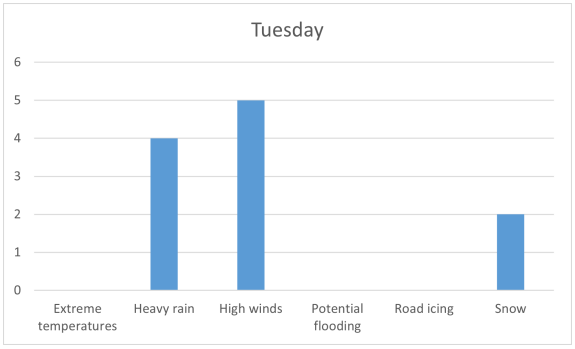


(e)

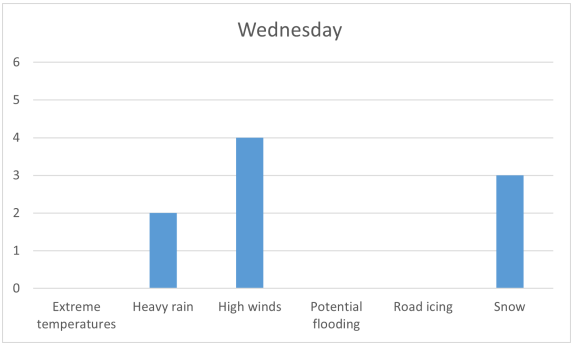
Figure 6.25: Pie charts on whether or not the representative members, at the beginning of the window of interest (i.e. the lead time provided) contain a possible high impact scenario. The results for the entire week (excluding Monday) are in (a), and Tuesday through Friday are in (b) to (e).



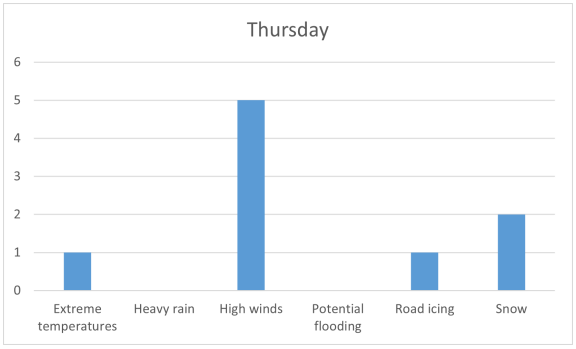
(a)



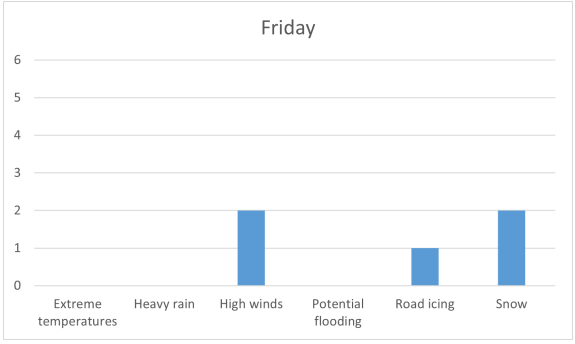
(b)



(c)

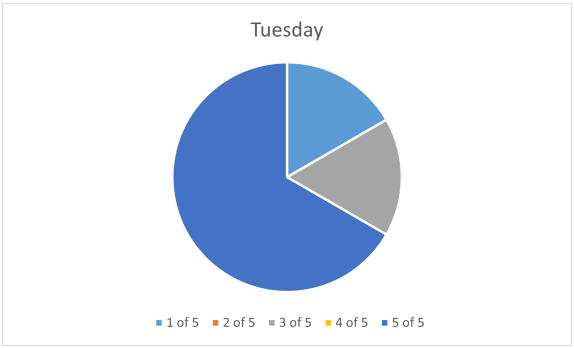


(d)

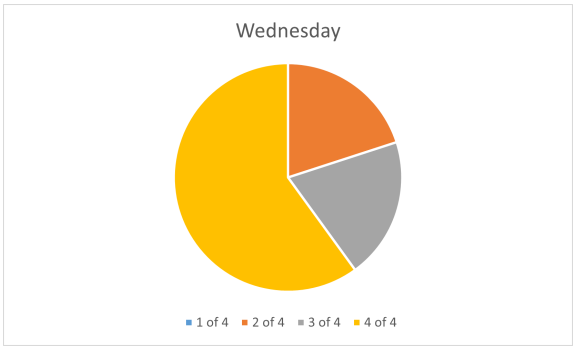


(e)

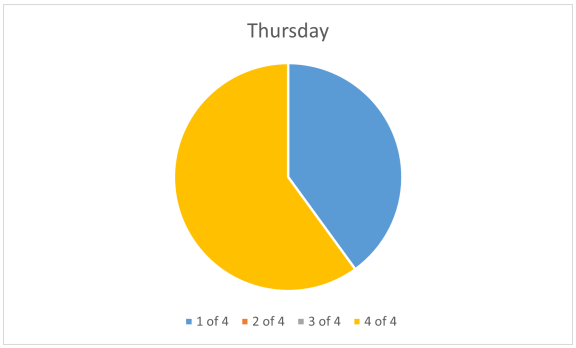
Figure 6.26: Bar plots of what types of high-impact weather were apparent in the scenario. The results for the entire week (excluding Monday) are in (a), where 1 response indicates extreme temperatures, 6 responses indicate heavy rain, 16 responses indicate high winds, 0 responses indicate potential flooding, 2 responses indicate road icing, and 9 responses indicate snow. Bar plots for Tuesday through Friday are in (b) to (e).



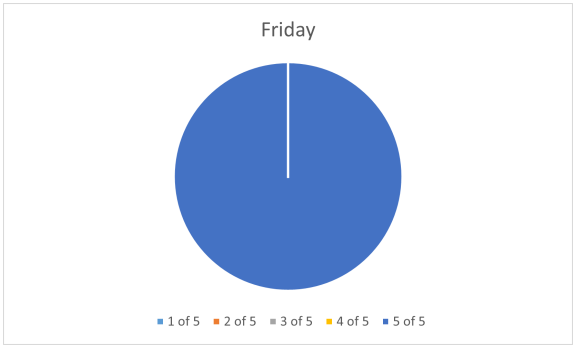
(a)



(b)



(c)



(d)

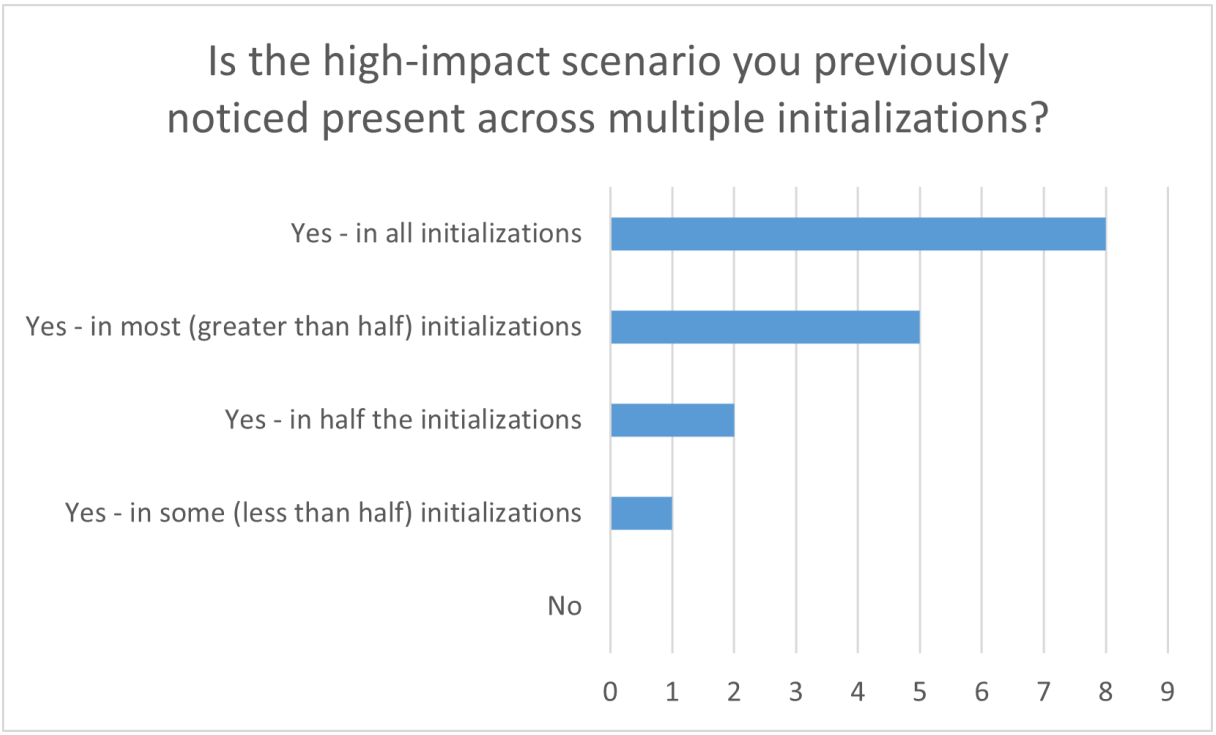
Figure 6.27: Pie charts of how many clusters in a forecast contain a high-impact scenario, where Tuesday (a) and Friday (d) had 5 clusters, and Wednesday (b) and Thursday (c) had 4 clusters.

noticed was present. This can be a useful determination to see if a particular member/forecast is just an outlier, if the formation of the event itself is uncertain, or if the occurrence of the event is reasonably certain but there is uncertainty about it's timing, position, or intensity. Figure 6.28 shows that the high-impact event participants noticed appeared consistently across initializations throughout the entire week. This supports the summary plot, figure 6.2, where the beginning of the window of interest followed the same valid time across multiple forecasts in week 3.

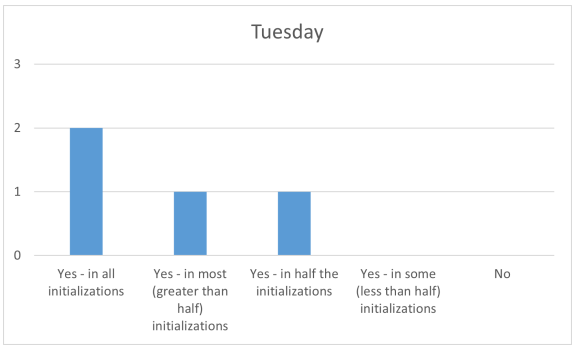
Q5. Is the presence of this high-impact event as a potential scenario likely to impact your forecast message?

Q6. Why or why not?

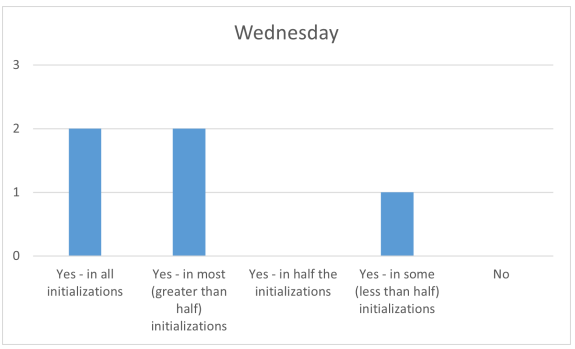
For the group of participants that said there was a high-impact scenario present, almost all of them every day said that the presence of this event would impact their forecast message (figure 6.29). Comments over the week mostly included mentions of high wind, such as "message can focus on Atlantic frontal systems bringing strong winds and snow. Mostly over Scotland but potential for these to sink further S and impact



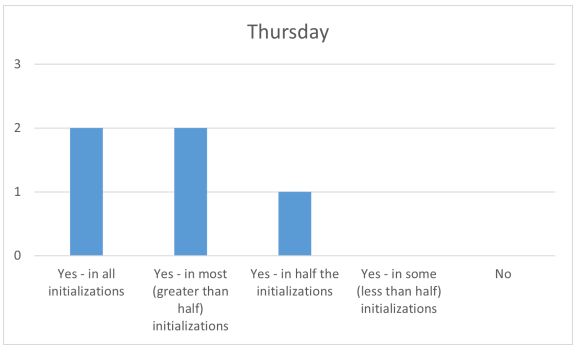
(a)



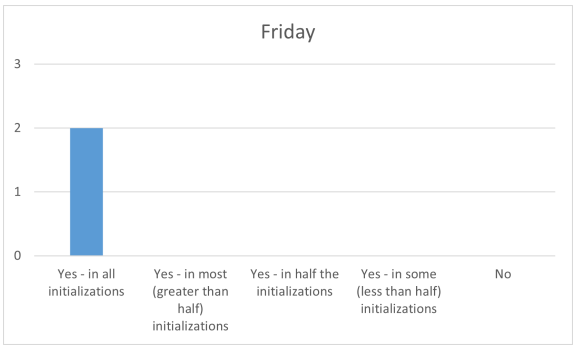
(b)



(c)

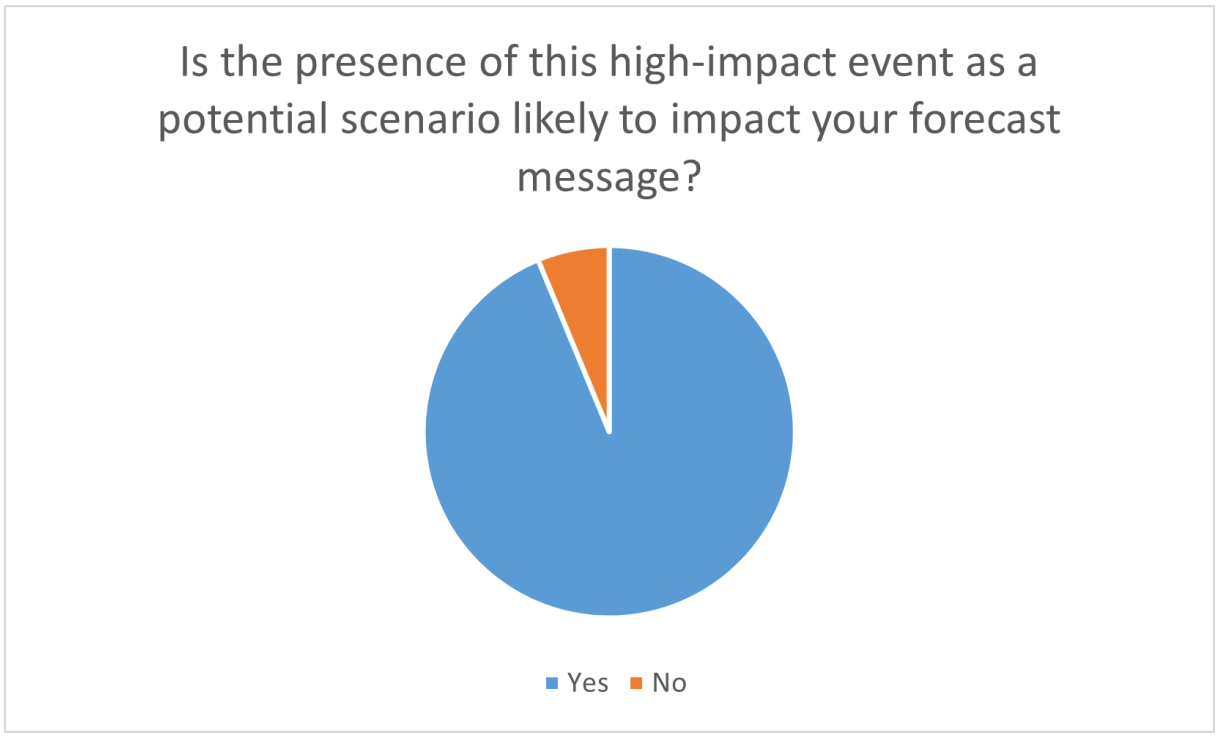


(d)

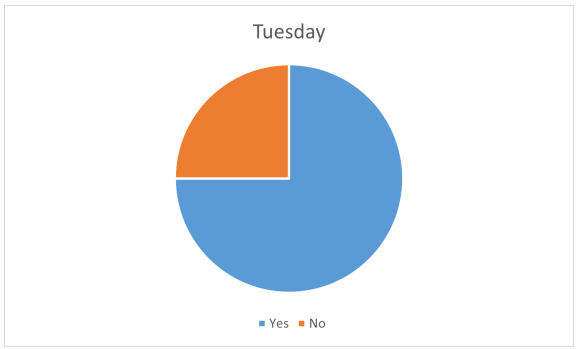


(e)

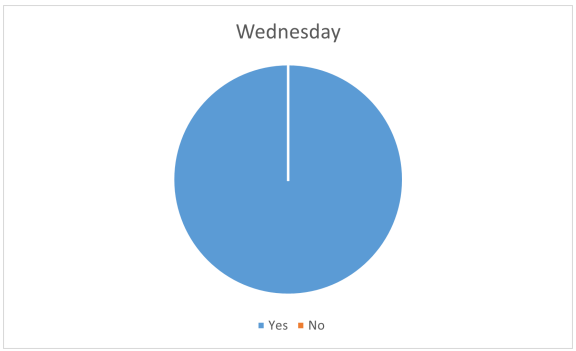
Figure 6.28: Bar plots of if the high-impact scenarios presented by the representative members appear across multiple initializations. The results for the entire week (excluding Monday) are in (a), where 8 responses indicate they were present in all initializations, 5 responses indicate they were in greater than half of the initializations, 2 responses indicate they were in half the initializations, 1 response indicates they are in less than half the initializations, and 0 responses indicating they were in no previous initialization. Bar plots for Tuesday through Friday are in (b) to (e).



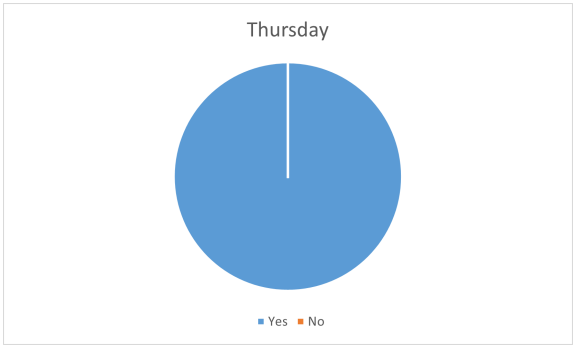
(a)



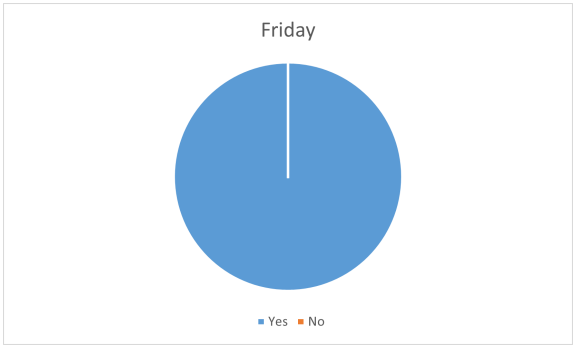
(b)



(c)



(d)



(e)

Figure 6.29: Pie charts of if the high-impact event was likely to affect the participant's forecast message. The results for the entire week (excluding Monday) are in (a), and Tuesday through Friday are in (b) to (e).

N England as well," and "high winds across Scotland for a time in almost all the runs, but with a variety of intensity and extent. For example cluster 3 in the latest run is the most impactful-looking." This is important to note because if the method can draw the attention of operational meteorologists to a potential high-impact event quicker than going through an ensemble's worth of data it will save them valuable time and resources.

Q7. Do the scenarios appear across multiple initializations?

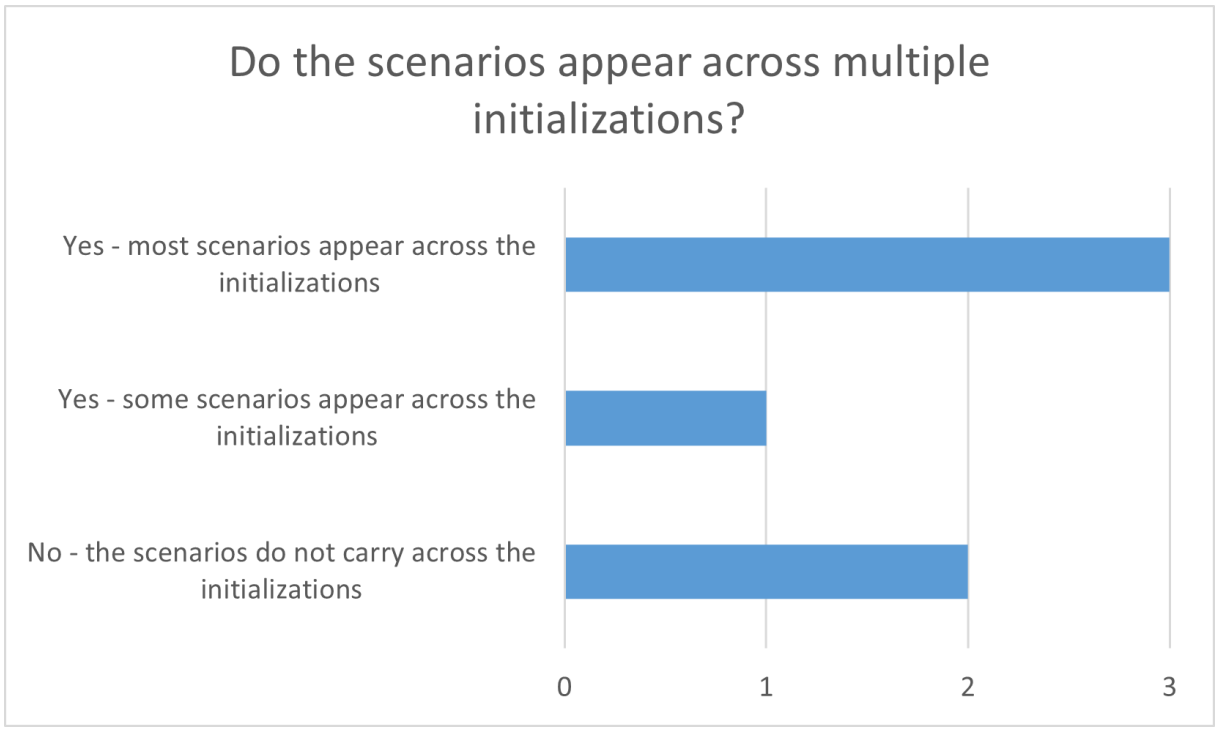
This question was specifically for participants who said there was not a high-impact scenario present, which only resulted in 1 or 2 responses each day (figure 6.30). There is less of a clear pattern here, as there are very few responses and they fall into all categories. A possibility for these answers is that participants may have been focusing on very specific aspects of a scenario that they expected to translate across valid times, or they were expecting the RMs to directly translate from one valid time to another. An example of how scenarios may appear and change over valid times can be seen in figures 4.21 to 4.24. Between any two sets of RMs, there is a range of differences from minor changes in frontal position to major changes in frontal shape, and the associated table 4.1 indicates there can be very small (57.5 km) to very large (529.1 km) differences in FSS distances between the closest RMs across valid times. This supports the hypothesis that tracing scenarios across valid time by just visual interpretation is likely subject to what the participant is looking for.

6.4.3.4 Efficiency of the clustering algorithm versus evaluating the ensemble as a whole

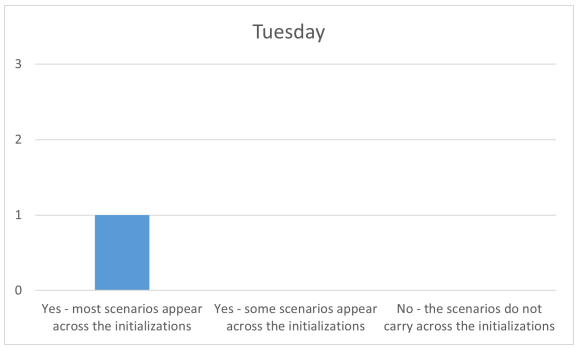
Now that participants had time to explore the method's products, evaluate the weather scenarios, and answer directed survey questions, there was one remaining question to ask:

1. Considering today's forecast, to what extent do you find the clustering more efficient than looking though the full ensemble?

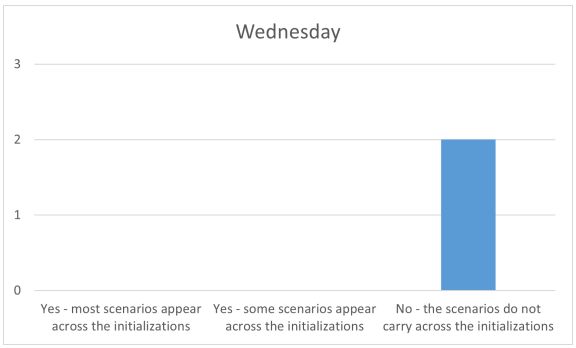
This question is the sum of their experience with the activity, and the end goal of designing the novel clustering method. The clusters and representative members must make using the ensemble easier and faster, so these responses are particularly important to consider for the project.



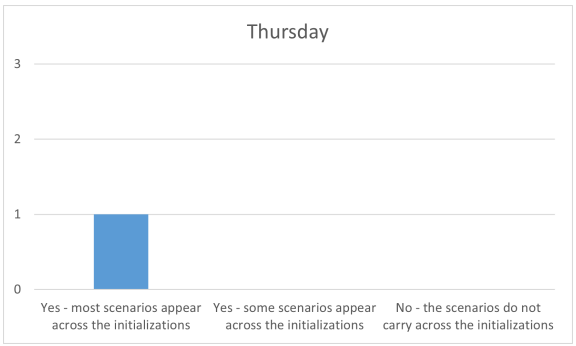
(a)



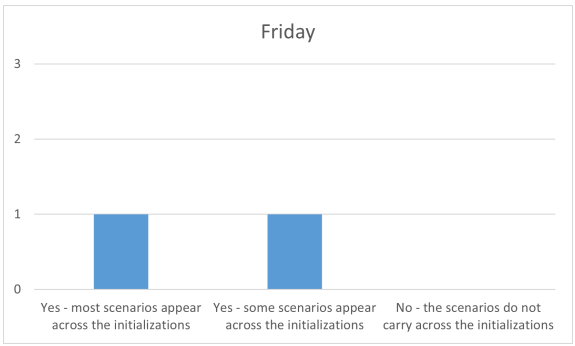
(b)



(c)



(d)



(e)

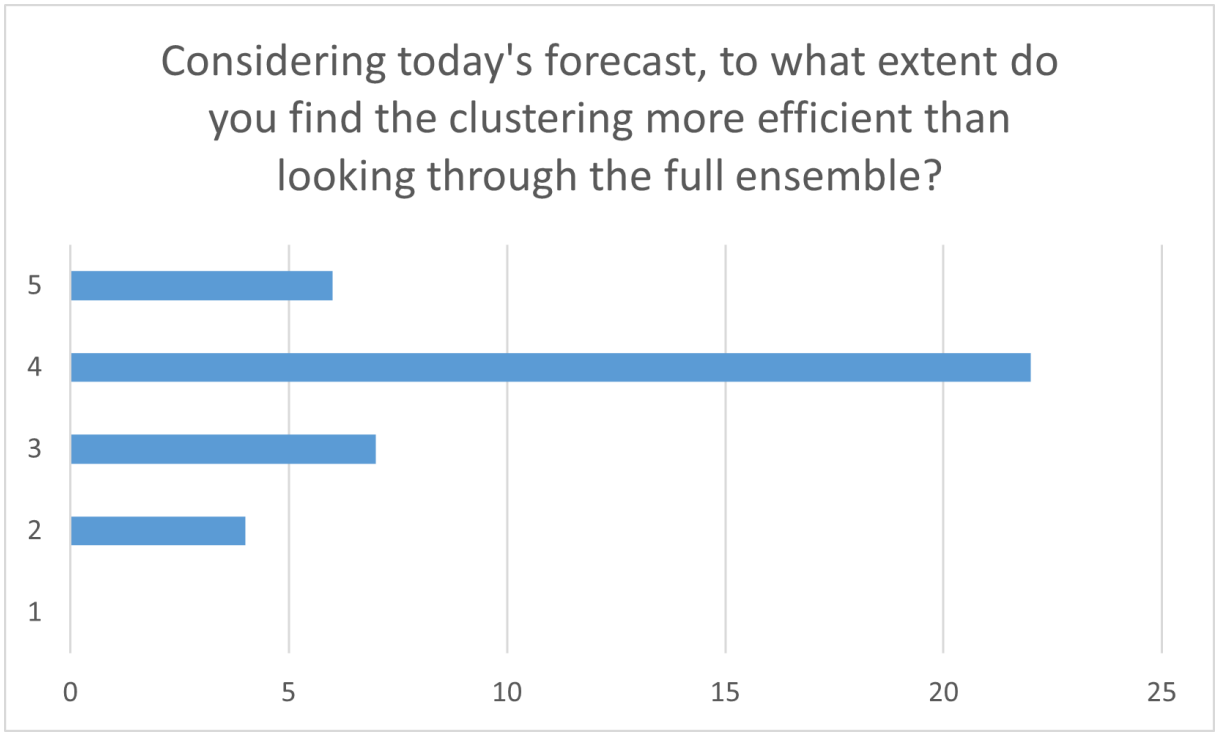
Figure 6.30: Bar plots of if the scenarios presented by the representative members appear across multiple initializations. This question appeared if participants declared there were no high-impact scenarios amongst the representative members. The results for the entire week (excluding Monday) are in (a), where 3 responses indicated most scenarios appeared across the initializations, 1 response indicated some appeared across initializations, and 2 responses indicated no scenarios appeared across the initializations. Bar plots for Tuesday through Friday are in (b) to (e).

Q1. Considering today's forecast, to what extent do you find the clustering more efficient than looking through the full ensemble?

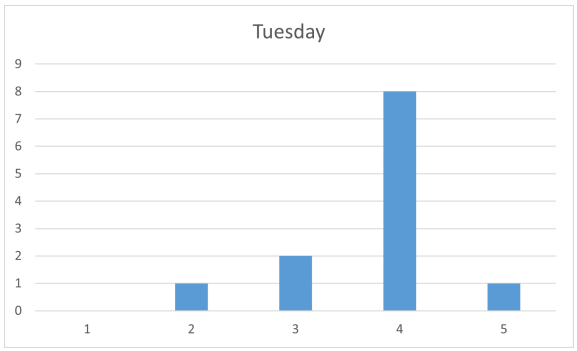
In figure 6.31, the scale is from 1 to 5, where 1 is not at all efficient and 5 is very efficient. Overwhelmingly (15% very efficient, 56% efficient), participants found using the clustering method and RMs to be more efficient than using the full ensemble, which is one of the desired goals. There were only two days in which participants found the clustering less efficient than looking at the ensemble, one of those days being Friday, where the window of interest was two days out from the forecast initialization. This could indicate a couple of possibilities: 1) the beginning of the window was close enough to the forecast initialization that there was not enough variation between the clusters to make them visually distinct, thereby drawing forecaster attention unnecessarily when a glance over the ensemble could have more quickly provided the same determination, or 2) the variables necessary to see the distinction between the clusters were not provided, thereby requiring access to the full ensemble regardless of the clustering. However, these are issues that could be addressed by further refinement of the method.

6.5 Conclusion

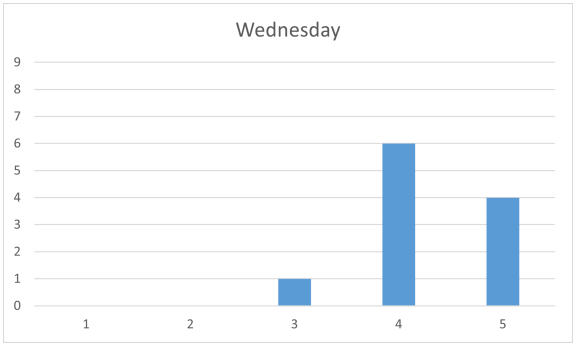
The testbed provided valuable feedback from operational meteorologists, the intended users of the method and its products, on how well the method performed and how useful and efficient it is with regards to looking at the full ensemble instead. The results of the survey indicate the method is somewhat subjective, as it is often focusing on a medium-range forecast due to how the window of interest is determined, and how distinct the representative members appear to the operational meteorologist can depend greatly on how they tend to view medium-range forecasts as a whole. Some participants focused on the general pattern of the atmosphere, others were more concerned with the spatial, temporal, and intensity variation of the frontal systems the clustering focused on. This lead to the conclusion that the method can produce unique and important scenarios similar to what an operational meteorologist would pick, though it may be more beneficial for specific types of forecasting, i.e. medium range forecasting. The clustering method is versatile, and applying it to different domain sizes, variables, and time scales may produce results more applicable to short range forecasting or local forecasting.



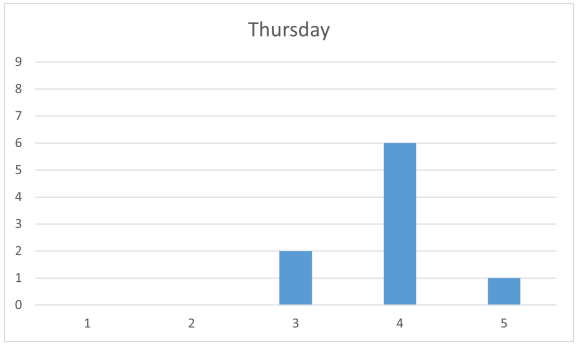
(a)



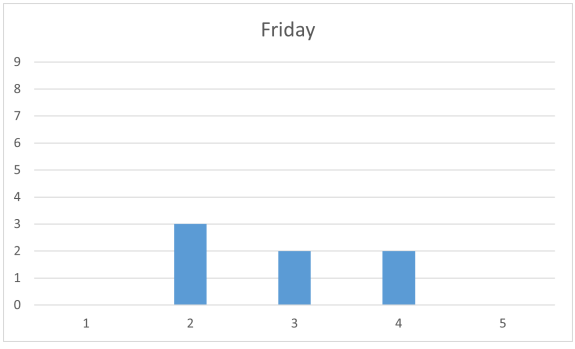
(b)



(c)



(d)



(e)

Figure 6.31: Bar plots of how efficient the clustering was versus looking at the full ensemble, where 1 is not at all efficient and 5 is very efficient. The results for the entire week (excluding Monday) are in (a), and Tuesday through Friday are in (b) to (e).

The clustering method did provide valuable information to operational meteorologists, making it a useful tool that extracts potential scenarios from an ensemble for further review. The majority of participants found the clustering method and its products to be somewhat to very useful in informing their forecasting decisions for both general and specific users for high-impact events. Overwhelmingly, participants found the method was generally more efficient than using the full ensemble to inform their forecast. Participants were also largely able to trace a scenario back through forecast initializations, echoing the case study in section 4.5.2 but by purely visual inspection. Being able to follow scenarios back across forecast initializations is a key finding and an important feature of the clustering method that will allow operational meteorologists to see how potential scenarios are evolving, when some scenarios are no longer forecast, and when new scenarios appear. These results support the further refinement and implementation of the method as another tool for operational meteorologists to reduce the amount of time required to absorb relevant information from an ensemble forecast before issuing a forecast.

Chapter 7

Conclusions and discussion

7.1 Introduction

This thesis has sought to provide a solution for the abundance of complex data provided by ensemble forecasts and the amount of time operational meteorologists must take to digest this information. A novel clustering technique was developed to extract distinct forecast scenarios from ensemble forecasts, thus extracting the most salient information for operational meteorologists to use when creating their forecast message. Using the novel choices of the gradient of the wet-bulb potential temperature, the Fractions Skill Score (FSS), and K-medoids clustering, this new method has been evaluated using 3 months of operational ensemble forecast data. The gradient of the wet-bulb potential temperature was chosen due to its relationship with air mass boundaries and frontal regions, which are often associated with high wind and heavy precipitation. The FSS was chosen as the distance metric for comparing ensemble members due to its unique neighborhood comparison and the ability to avoid the double-penalty problem. K-medoids was chosen as the clustering algorithm for the method for two primary reasons. It both has the ability to use non-standard distance metrics and it uses a member as the centre of a cluster instead of a mean. Clustering around a member means that the centre of a cluster is a forecast solution of the ensemble, which aligns with the goal of presenting representative members (RMs) from the ensemble as possible scenarios. Different meteorological variables have been compared over the month of October in the measure of distance between forecasts. Finally, the method was reviewed by expert elicitation of opinions in the relation between the objective clustering technique and relevant scenarios in forecast messages. Within this

concluding chapter there is a discussion of the results of the study, a discussion of how this work contributes to the field of meteorology and clustering applications, limitations of the work, and recommendations for future work and analysis.

7.2 Study results

The gradient of the wet-bulb potential temperature was the first variable to which the new method was applied due to its relationship with high impact events (high wind, heavy precipitation) in the UK. This led to further analysis using the large-scale rain rate to compare to $|\nabla\theta_w|$ to determine if the variables were linked, if they could be interchanged, or if they were better used for specific purposes. Then, the method was analysed in a testbed setting, using $|\nabla\theta_w|$, where operational meteorologists provided their feedback on its products, use, and benefits.

The method was able to consistently extract distinct weather scenarios from the ensemble, using either $|\nabla\theta_w|$ (chapter 4) or the large-scale rain rate (chapter 5). However, the level of distinction may depend upon chosen parameters, such as the threshold applied to the fields of data to create binary fields before the FSS was applied. This threshold determines how fine a scale of features is permitted and affects the FSS distance, and in turn the clustering, making it a key parameter. A limitation of this thesis was that the sensitivity of the threshold was not tested. The level of distinction between RMs associated with $|\nabla\theta_w|$ was also subject to operational meteorologist preference, as noted in chapter 6. Some operational meteorologists found ample distinction between RMs, each warranting a different forecast. Others said there was not enough distinction at such a domain size. This may in part be due to what their areas of interest were (i.e. short or medium range forecasting) or their expectations on how different the RMs would be on average. The distinction also often relied on when the window of interest began. As discussed in section 4.4.2, there exists a period of time at the beginning of a forecast when there isn't much distinction between clusters, then a period when clustering becomes the most distinct, and the finally a period when members have diverged far enough away from one another they could be considered individual clusters. This is related to the ensemble spread, which is expected to increase throughout the forecast. However, the sum distance is not expected to decrease indefinitely, which makes it ideal for use in determining

this window of interest within a forecast. However, operational meteorologists during the testbed mentioned that if the window began too soon, e.g. around t+48, there may not have been enough distinction between RMs to justify different forecasts. If it began too late, e.g. around t+144, it may be too uncertain and too far out in lead time to be relevant to the forecast. If the start time of the window is perceived to be too early or vary too much, the parameters of the method could be updated to adjust how the start point is identified.

A key feature of the novel method is clustering at each lead time, which allows members to move between clusters. This provides a unique opportunity to see when membership begins to stabilize by examining a traceability diagram (e.g. figure 3.6). Cluster stability is linked to a drop in the sum distance, indicating clusters have become distinct from one another and members have diverged enough from the control as to have developed different scenarios. Towards the end of the window of interest, the cluster membership begins to scatter as the ensemble continues to spread and members become decorrelated from each other, as expected in a chaotic system. It is therefore important to examine the results within the window for the best clustering. However, clustering is stronger on some days than others. Some forecasts may contain an event that is fairly stagnant, such as a block, or simply have a relatively inactive atmosphere in terms of frontal development. This often results in clusters that are predominantly very similar in nature, or fewer clusters due to the lack of variation in potential scenarios. Although it is expected that the optimal number of clusters are larger for later window of interest start times and smaller for earlier windows, that does not appear to be the case. However, this point could use further analysis with more case studies.

There is evidence that scenarios do connect across different forecasts for the same valid time (section 4.5.2). Cluster membership also appeared connected to the probability of a scenario occurring, although the FSS distances between RMs from consecutive forecasts did not strongly indicate a link between smaller distances and the observed forecast until the event was closer. During the testbed, operational meteorologists also determined that some events were picked up across forecasts at the same valid time for several forecasts (see section 6.4.3.3), where the window of interest focused primarily on a single event. Notable instances of this can be seen on the various summary plots (e.g. figure 4.1) where the window of interest follows a valid time across several forecasts. However, more

work must be done and case studies evaluated to establish what methods may be used by operational meteorologists in real time to determine which scenarios might be most likely.

With regards to the use of other variables, the method can be used on any variable available, however the usefulness of that variable depends on the user's end goals. The gradient of the wet-bulb potential temperature can be used as a variable for clustering around air mass boundaries and potential frontal regions. With an appropriate domain size such as that chosen for this study, the field captures synoptic-scale atmospheric motion such as mid-latitude cyclone development, seen in many examples within this work (see figures 4.11, 4.19, 5.5, and 6.9). The large-scale rain rate performed similarly to $|\nabla\theta_w|$, which is expected. The two variables are clearly related, as evidenced by some similarity in cluster membership and RMs. While the large-scale rain rate typically corresponded with frontal objects found with $|\nabla\theta_w|$ they often were much larger objects, making the clusters less distinct. It is also highly variable and may not accurately indicate the likelihood of high-impact weather, whereas $|\nabla\theta_w|$ indicates frontal zones which are more closely linked with high-impact weather. For these reasons, $|\nabla\theta_w|$ is the recommended variable of choice for this application. However, clustering on the large-scale rain rate may be beneficial when examining a smaller UK sized domain for a shorter time span, e.g. clustering rainfall associated with storm landfall to better estimate impact for specific areas.

In the final research chapter (6) the results of the method being evaluated in a testbed setting were explored. Overall, the results of the testbed were positive, with many operational meteorologists finding the representative members useful for their forecasts and impact analysis. While more work should be done to finely tune the method to the needs of the operational meteorologists, it can be concluded that the aim of the project was met: reducing the amount of ensemble data an operational meteorologist must digest before issuing their forecast. Additionally, the method highlights potentially impactful weather with low predictability, drawing operational meteorologist attention to the variability in the atmosphere when different distinct scenarios begin to form. The testbed was run in real-time, and there were a few distinct events during it that dominated the predictability and the occurrence of clustering. The method clearly picked out these events, including their start time via the window of interest, and presented them to the meteorologists.

7.3 Discussion and contribution of the method

This work adds to the previous clustering work done in atmospheric science, but it also brings a new perspective and a novel methodology. Firstly, the design of this method uses K-medoids as the clustering algorithm. Previous methods of clustering used hierarchical (Cape et al., 2000; Hart et al., 2015; Johnson et al., 2011a,b; Molteni et al., 1996; Marsigli et al., 2001; Montani et al., 2011) or partitional clustering typically based on K-means (Neal et al., 2016; Richardson et al., 2020; Ferranti and Corti, 2011; Kassomenos, 2003a,b; Philipp et al., 2007; Enke and Spekat, 1997; Zheng et al., 2017; Delcloo and Backer, 2008; Leckebusch et al., 2008). K-medoids is closely related to K-means, but has the advantage of using a member as the centre point which retains the forecast information as opposed to smoothing out sharp features such as with a mean. By being a partitional method, it also has the advantage over hierarchical methods by allowing cluster membership to vary as the number of clusters changes. Clustering has been done on ensembles before for several reasons: more information to provide forecasters concerning circulation patterns (Ferranti and Corti, 2011), analysis of perturbation effects (Johnson et al., 2011a,b), and ensemble reduction (Molteni et al., 2001; Marsigli et al., 2001; Montani et al., 2011, 2003). The clustering performed here is similar to ensemble reduction in that an ensemble of forecasts is reduced to a few representative members. Montani et al. (2003) and Montani et al. (2011) introduced the ensemble reduction method developed for COSMO-LEPS, which results in a representative member that is then used as initial and boundary conditions for a high resolution forecast. Similar to the method I used, their RM is a member, not a mean. However, the method they take to get to this stage is complex and time consuming, using 153 members and a combination of several variables at various pressure levels for clustering (Marsigli et al., 2001). In contrast, the method presented in my work is intended for a small ensemble and is designed to be run quickly, providing operational meteorologists with RMs soon after the ensemble forecast is produced. K-medoids works well for small ensemble sizes, but a drawback of this method is that computation time increases significantly as the ensemble size increases.

As implied by Serafin et al. (2019), careful considerations must be made of the clustering algorithm and variable choice in regards to getting good clustering results for use in limited area models. This also extends to my study, although at this time the RMs

are not used to initialize a limited area model. However, the variable choice of $|\nabla\theta_w|$ was made to have the largest impact in the simplest way possible, i.e. extracting frontal regions that could be associated with extreme precipitation. This variable does provide a synoptic view of the atmosphere, albeit a focused one. Previous studies that were focused on atmospheric patterns tended to use pressure or geopotential height variables (Huth et al., 2008; Philipp et al., 2010; Beck and Philipp, 2010; Casado et al., 2010; Neal et al., 2016; Richardson et al., 2020; Ferranti and Corti, 2011; Philipp et al., 2007; Enke and Spekat, 1997). Other studies such as Kassomenos (2003a,b) used several different variables to look at seasonal circulation patterns and air masses. However, my method is not focused on circulation pattern categorization but scenario extraction. Though they are similar, as scenarios are representations of potential circulation patterns, they are focused on potential high-impact weather by using $|\nabla\theta_w|$ as the variable choice. This gains the benefit of distinguishing air mass boundaries at a glance and alerting forecasters to potentially strong frontal regions. A drawback of this method, however, is that it is limited in the amount of information it provides. The RM extracted from each cluster can also be shown with other variables, such as precipitation or wind, but because those variables are not integrated into the method and therefore are not part of the clustering, there is the potential that the members of the cluster may vary significantly in relation to them. E.g., while the frontal region is very similar in shape, therefore clustering two members together, one forecast may have more intense rain and wind associated with it than the other. Which raises the question, how representative are the RMs? This is an area of future work that should be explored.

A novel distance metric for the clustering process was also introduced: the Fractions Skill Score. K-means methods are restricted to Euclidean distances between members and hierarchical methods are also generally restricted to a few variations of distance calculations, such as average-linkage and Ward's method. Therefore, some previous clustering studies focused their attention on unique ways in which to treat the data before clustering, such as reducing it by empirical orthogonal functions (Ferranti and Corti, 2011; Zheng et al., 2017), factor analysis (Kassomenos, 2003a,b), scaling (Hart et al., 2015), or standardization (Montani et al., 2003). However, by using K-medoids the method I developed can use the FSS distance, allowing a new way to cluster members. The only pre-processing required of the wet-bulb potential temperature data before clustering is calculating the

gradient then applying a threshold to create a binary field that is used to calculate the FSS distance between the members. The distances are then used directly in K-medoids clustering. There are many advantages to using the FSS distance as a distance metric over the Euclidean distance and other standard distance metrics. It allows two fields that contain objects (in this case, the frontal regions identified by the threshold process) to be compared to one another while avoiding the double penalty problem. It is also a fairly simple and straightforward method, as opposed to MODE (Davis et al., 2006a,b, 2009) and SAL (Wernli et al., 2008) that define object characteristics to compare. Johnson et al. (2011a,b) clustered precipitation objects using MODE, but they were still limited to using Ward's method for their hierarchical clustering method. Introducing K-medoids with the FSS distance creates a whole new opportunity for researchers to explore in terms of clustering. However, there are some uncertainties with this method. While the fields compared in this study had regions masked out to remove erroneous values, such as values over Greenland, there is still likely to be many frontal objects appearing of various sizes in the field. By current experimentation and subjective opinion, the algorithm appears to cluster members based on long well defined fronts first while smaller broken frontal objects are less likely to impact the results. But this has not been extensively evaluated and should be considered in future applications.

Also introduced was a new way to temporally cluster a forecast, i.e. clustering at individual lead times and tracing clusters across the forecast. Circulation studies tend to be focused on the fields, such as the studies by Neal et al. (2016) and Philipp et al. (2007), that clustered daily mean sea-level pressure (MSLP) fields, or the study by Richardson et al. (2020) that clustered daily MSLP anomalies and the study by Enke and Spekat (1997) that clustered daily geopotential heights and thicknesses. Zheng et al. (2017) did cluster MSLP and 500 hPa geopotential height at individual lead times within a forecast, but the clusters are not linked across lead times. Several studies cluster over windows of time, such as Ferranti and Corti (2011) who applied clustering to the 500 hPa geopotential height at four different time windows to maintain synoptic consistency, Johnson et al. (2011a) and Johnson et al. (2011b) who clustered 24-hour forecasts of 1-hour precipitation accumulation, Molteni et al. (2001) and Marsigli et al. (2001) who clustered day-5 forecast fields of geopotential height, and Leckebusch et al. (2008) who clusters 3-day forecast periods of 1000 hPa geopotential height or MSLP to analyse cyclones. Storm tracks and air

parcel trajectories similarly cluster the paths calculated over several hours or days, such as the studies by Delcloo and Backer (2008) and Cape et al. (2000) who clustered 5-day back trajectories, the study by Hart et al. (2015) that clustered conveyor belt airstreams and mesoscale jet structure over a 7-hour period, and the study by Kowaleski and Evans (2016) that clustered storm tracks and cyclone phase space for hurricane Sandy over different time segments. Kassomenos (2003a) and Kassomenos (2003b) cluster daily variables and then perform a temporal analysis of the results by examining inter-annual variation in cluster frequency of occurrence and the frequency of event occurrence over the next two days after it first appears. The temporal aspects of a study can clearly be important. The method developed for this study is key in accounting for the variability within an ensemble forecast, allowing for flexibility in extracting scenarios. Ensemble forecast trajectory behaviour is complex and allowing cluster membership to change throughout a forecast takes this into account. Clusters must then be linked through lead times, which can then be used to determine representative members. However, clustering at each lead time and then linking the clusters through a forecast can be time intensive, and if there are not clearly different scenarios appearing within the ensemble clustering in this fashion becomes less clear. E.g., members may move so often between similar clusters that there is no clear distinction between clusters.

Finally, this method introduces a different way to extract a representative member from ensemble forecast clusters. In Ferranti and Corti (2011), the representative member is chosen as the member closest to the centre of the K-means cluster. Molteni et al. (1996); Marsigli et al. (2001); Montani et al. (2011, 2003) determines the RM by selecting the member with the smallest ratio between the average distances of members in the cluster and to members in other clusters. These methods were chosen because they were suitable for the study. However, with the new method presented here, a different way of choosing a representative member was required. As clusters were calculated at each lead time then linked throughout a forecast, a window of interest was required to determine when clustering was at its peak and an RM would be most likely to represent the scenarios extracted from the ensemble. The RM selection method used here has the disadvantage that it is somewhat complex. However, the ability to tune it may be an advantage if this technique is used on other variables or combinations of variables in the future.

7.4 Limitations

The primary limitation of this study was that a limited number of forecasts were investigated and it has only been tuned using the MOGREPS-G ensemble forecast. The method was originally intended to be used on convection-permitting ensemble data, such as MOGREPS-UK, on precipitation data. However, during early development the choice was made to first build the method on global data (i.e. MOGREPS-G) as both a proof of concept and to explore contributing factors to precipitation. As work began, it was clear that developing the method on MOGREPS-G with $|\nabla\theta_w|$ would have a great deal of utility, and so the main analysis switched to global data. If there was time permitting, the method would have also been tuned for precipitation in MOGREPS-UK. The method can be used on any ensemble forecast with any variable, but for this study it was restricted to being evaluated with $|\nabla\theta_w|$ and the large-scale rain rate. A further limitation was that the sensitivity of the clustering behaviour to the choice of preset values, such as the choice of threshold, was not evaluated. However, the results show significant promise in the utility of the algorithm but it requires more study and data to solidly confirm these results.

7.5 Recommendations for future work

While the method can be used immediately by the Met Office, who are conducting further research on it and integrating it into operational use, there are several areas for future work and studies that could be conducted. The following sections will go over several potential areas of research.

- What are the optimal values for the thresholds that are used and how dependant are they on the domain size, weather regimes, and seasons?

An in-depth sensitivity analysis should be performed on the threshold applied to $|\nabla\theta_w|$, the value of sum distance the window of interest begins, and the FSS neighborhood sizes and boundaries. The threshold was chosen subjectively as a percentile of $|\nabla\theta_w|$ values to extract the largest gradients. But the question of how well this threshold works on different domain sizes, during different weather patterns, and during different seasons,

remains. Therefore, it is recommended that a series of studies be conducted that varies the threshold value and compares clustering results and how distinct scenarios are from one another. It is anticipated that depending on the variable of choice and the domain size the threshold may need to be adjusted to get the desired results. Feedback on the window of interest start time indicated that when the window began too early there wasn't enough distinction between scenarios. The drop in sum distance begins the window of interest, therefore an analysis should be done on what the optimal drop should be. This could be achieved by varying the percentage drop required in the sum distance for the window to begin or restricting how early the window can begin. This may also be dependant on the atmospheric situation, e.g. a more active or unpredictable atmosphere is more likely to have strong frontal regions and distinct scenarios, whereas a calm or predictable atmosphere has very little variation amongst members and therefore less distinct clustering and scenarios. The FSS calculations used for this method did not include any domain boundary padding, i.e. when the center of the neighborhood is at the edge of the domain and the values of the neighborhood that fall outside of the domain are filled with zeros. This may be a limitation to the method, or it may not impact it at all as the current domain was chosen so that events of interest that would impact the UK would tend to more through the centre. However, this should be thoroughly tested by varying the boundary conditions used in the FSS calculations and evaluating the similarity of the clustering results.

- Can the method be tuned for different seasons, different variables, and convection-permitting models?

The current method was designed specifically for MOGREPS-G and used during the autumn and winter months, when atmospheric patterns often result in unstable weather and storms and are associated with sharp gradients between air masses. While the method still produced clusters and scenarios during calmer periods, it remains to be seen if it should be tuned for different atmospheric patterns and seasons. To begin this work would require at least a year's worth of data and clustering analysis. Examining the clustering results could lead to adjusting the various preset values, such as the threshold and sum distance mentioned previously, or choosing different variables for different seasons, such as maximum temperature, cloud cover, pressure gradients, or precipitation. Depending

on the ensemble model, such as a convection-permitting model like MOGREPS-UK, the method could be tuned specifically for precipitation events. How versatile the method is should be fully explored by applying it to several different ensemble models with the same parameters and variable choice, with the clustering results explored in detail to determine how robust the results are and if the extracted scenarios match an operational meteorologist's analysis of the ensemble.

- Can a technique be developed that allows scenarios to be linked reliably across forecasts and can it be used to indicate the probability of a scenario occurring?

Section 4.5.2 introduced a way to potentially link scenarios across forecasts along the same valid time. If a reliable technique can be developed that can achieve these connections with relative accuracy, it could dramatically improve the output and value of the method. This would require several case studies that have strong clustering, such as forecasts that include a storm with uncertainty in its position that the method picks out within the window of interest over several forecasts. Instead of restricting the clustering to four clusters, it would be better to allow the method to choose the optimal number and then compare the clusters across forecasts either cluster to cluster or RM to RM. A verification analysis of the scenario against the observations could then be performed and used to further investigate how linking scenarios across forecasts may lead to more probable outcomes. This can also be coupled with an analysis on the probability of a scenario occurring and the number of members in the associated cluster.

- How closely does the method mimic the way an operational meteorologists would cluster the ensemble and does it provide all the information they need from the ensemble to inform their forecasts?

Finally, it is recommended that a longer and more comprehensive survey of operational meteorologists be performed to further refine and tune the method to their needs. The results presented in chapter 6 are limited to only four weeks of data and survey results, thus limiting the conclusions drawn to being more subjective in nature. However, a more detailed survey conducted over several months or seasons could be used to gather both more data and more forecaster analysis. These results could then be compared and verified with observational data and more objective conclusions could be drawn about how the

method performs on high-impact scenarios and probabilistic ensemble forecasting as a whole. As the method continues to be studied and is being implemented at the Met Office for operational use, it is likely another survey will be conducted in the future.

Bibliography

Ahrens, C. and Henson, R. (2016). *Meteorology today: an introduction to weather, climate, and the environment. Eleventh edition.* CENGAGE Learning, 20 Channel Center Street, Boston, MA 02210, USA.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525:47–55.

Baur, F., Hess, P., and Nagel, H. (1944). Kalendar der Grosswetterlagen Europas 1881-1939. *Nature*.

Beck, C. and Philipp, A. (2010). Evaluation and comparison of circulation type classifications for the European domain. *Phys. Chem. Earth*, 35:374–387.

Berry, G., Reeder, M., and Jakob, C. (2011). A global climatology of atmospheric fronts. *Geophys. Res. Lett.*, 38(L04809):1–5.

Bishop, C., Etherton, B., and Majumdar, S. (2001). Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Weather Rev.*, 129:420–436.

Bouttier, F., Raynaud, L., Nuissier, O., and Ménétrier, B. (2016). Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Q. J. R. Meteorol. Soc.*, 142 (Suppl 1):390–403.

Bowler, N., Arribas, A., Mylne, K., Robertson, K., and Beare, S. (2008). The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, 134:703–722.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78:1–3.

Brusco, M., Steinley, D., and Stevens, J. (2019). K-medoids inverse regression. *Commun. Stat. Theory*, 48:4999–5011.

Buizza, R. (2018). Introduction to the special issue on "25 years of ensemble forecasting". *Q. J. R. Meteorol. Soc.*, 145 (Suppl. 1):1–11.

Buizza, R. and Palmer, T. (1995). The Singular-Vector Structure of the Atmospheric Global Circulation. *J. Atmos. Sci.*, 52(9):1434–1456.

Cape, J., Methven, J., and Hudson, L. (2000). The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland. *Atmos. Environ.*, 34:3651–3663.

Casado, M., Pastor, M., and Doblas-Reyes, F. (2010). Links between circulation types and precipitation over Spain. *Phys. Chem. Earth*, 35:437–447.

Catto, J., Jakob, C., Berry, G., and Nicholls, N. (2012). Relating global precipitation to atmospheric fronts. *Geophys. Res. Lett.*, 39(L10805):1–6.

Catto, J., Madonna, E., Joos, H., Rudeva, I., and Simmonds, I. (2015). Global Relationship between Fronts and Warm Converyor Belts and the Impact on Extreme Precipitation. *J. Climate*, 28:8411–8429.

Catto, J. and Pfahl, S. (2013). The importance of fronts for extreme precipitation. *J. Geophys. Res. Atmos.*, 118:10,791–10,801.

Charney, J., FjÖrtoft, R., and von Neumann, J. (1950). Numerical Integration of the Barotropic Vorticity Equation. *Tellus*, 2:4:237–254.

Davis, C., Brown, B., and Bullock, R. (2006a). Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Weather Rev.*, 134:1772–1784.

Davis, C., Brown, B., and Bullock, R. (2006b). Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Weather Rev.*, 134:1785–1795.

Davis, C., Brown, B., Bullock, R., and Halley-Gotway, J. (2009). The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Weather and Forecasting*, 24:1252–1267.

Delcloo, A. and Backer, H. (2008). Five day 3D back trajectory clusters and trends analysis of the Uccle ozone sounding time series in the lower troposphere (1969-2001). *Atmos. Environ.*, 42:4419–4432.

Dunlop, S. (2008). *A Dictionary of Weather (2 ed.)*. Oxford University Press, Oxford, United Kingdom.

DWD (2023). Grosswetterlagen Forecast (GWL). `https://www.dwd.de/EN/research/weatherforecasting/met_applications/nwp_applications/grosswetterlagen_forecast.html` [Accessed: Feb 2023].

Enke, W. and Spekat, A. (1997). Downscaling climate model outputs into local and regional weather elements by classification and regression. *Clim. Res.*, 8:195–207.

Ferranti, L. and Corti, S. (2011). New clustering products. *ECMWF Newsletter*, 127:6–11. `https://www.ecmwf.int/node/17442`.

Folland, C. and Woodcock, A. (1986). Experimental montly long-range forecasts for the united kingdom. part i. description of the forecasting system. *Meteorol. Mag.*, 115:301–318.

Gilleland, E., Ahijevych, D., Brown, B., Casati, B., and Ebert, E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24:1416–1430.

Gilleland, E., Skok, G., Brown, B., Casati, B., Dorninger, M., Mittermaier, M., Roberts, N., and Wilson, L. (2020). A novel set of geometric verification test fields with application to distance measures. *Mon. Weather Rev.*, 148:1653–1673.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W. (2017). The Met Office convective-scale ensemble, MOGREPS-UK. *Q. J. R. Meteorol. Soc.*, 143:2846–2861.

Hart, N., Gray, S., and Clarck, P. (2015). Detection of Coherent Airstreams Using Cluster Analysis: Application to an Extrotropical Cyclone. *Mon. Weather Rev.*, 143:3518–2531.

Hewson, T. (1998). Objective fronts. *Meteorol. Appl.*, 5:37–65.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J., and Tveito, O. (2008). Classification of Atmospheric Circulation Patterns - Recent Advances and Applications. *Ann. N.Y. Acad. Sci.*, 1146:105–152.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytol.*, 11:37–50.

Johnson, A., Wang, X., Kong, F., and Xue, M. (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. *Mon. Weather Rev.*, 139:3673–3693.

Johnson, A., Wang, X., Ming, X., and Kong, F. (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. *Mon. Weather Rev.*, 139:3694–3710.

Jolliffe, I. and Stephenson, D. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science (2 ed.)*. John Wiley & Sons, Ltd., Chichester, United Kingdom.

Kain, J., Willington, S., Clark, A., Weiss, S., Weeks, M., Jirak, I., Coniglio, M., Roberts, N., Karstens, C., Wilkinson, J., Knopfmeier, K., Lean, H., Ellam, L., Hanley, K., North, R., and Suri, D. (2017). Collaborative efforts between the United States and United Kingdom to advance prediction of high-impact weather. *Bull. Am. Meteorol. Soc.*, 98:937–948.

Kassomenos, P. (2003a). Anatomy of the synoptic conditions occurring over southern Greece during the second half of the 20^{th} century. Part I. Winter and summer. *Theor. Appl. Climatol.*, 75:65–77.

Kassomenos, P. (2003b). Anatomy of the synoptic conditions occurring over southern Greece during the second half of the 20^{th} century. Part II. Autumn and spring. *Theor. Appl. Climatol.*, 75:79–92.

Kassomenos, P. (2010). Synoptic circulation control on wild fire occurrence. *Phys. Chem. Earth*, 35:544–552.

Kowaleski, A. and Evans, J. (2016). Regression Mixture Model Clustering of Multimodel Ensemble Forecasts of Hurricane Sandy: Partition Characteristics. *Mon. Weather Rev.*, 144:3825–3846.

Kox, T., Kempf, H., Lüder, C., Hagedorn, R., and Gerhold, L. (2018). Towards user-oriented weather warnings. *Int. J. Disast. Risk Re.*, 30:74–80.

Leckebusch, G., Weimer, A., Pinto, J., Reyers, M., and Speth, P. (2008). Extreme wind storms over Europe in present and future climate: a cluster analysis approach. *Meteorol. Z.*, 17:67–82.

Losee, J. and Joslyn, S. (2018). The need to trust: How features of the forecasted weather influence forecast trust. *Int. J. Disast. Risk Re.*, 30:95–104.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *J. Comput. Phys.*, 227:3431–3444.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F., and Buizza, R. (2001). A strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.*, 127:2095–2115.

Met Office (2018). Storm Callum. `https://www.metoffice.gov.uk/weather/warnings-and-advice/uk-storm-centre/storm-callum`.

Met Office (2021). *Weather warnings guide*. `https://www.metoffice.gov.uk/weather/guides/warnings` [Accessed: Dec 2022].

Met Office (2022a). Daily weather summary | February 2022. `https://digital.nmla.metoffice.gov.uk/IO\_7da1a384-a514-429f-ba33-61221a81e67e/`.

Met Office (2022b). Daily weather summary | January 2022. `https://digital.nmla.metoffice.gov.uk/IO\_67cbebee-317f-49a6-a08d-b9a2430578fb/`.

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F., and Paccagnella, T. (2001). A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments. *Q. J. R. Meteorol. Soc.*, 127:2069–2094.

Molteni, F., Buizza, R., Palmer, T., and Petroliagis, T. (1996). The ECMWF Ensemble Prediction System: Methodology and validation. *Q. J. R. Meteorol. Soc.*, 122:73–119.

Montani, A., Cesari, D., Marsigli, C., and Paccagnella, T. (2011). Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A*, 63A:605–624.

Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T., Tibaldi, S., and Buizza, R. (2003). The Soverato flood in Southern Italy: performance of global and limited-area ensemble forecasts. *Nonlinear Proc. Geoph.*, 10:261–274.

Mu, D., Kaplan, T., and Dankers, R. (2018). Decision making with risk-based weather warnings. *Int. J. Disast. Risk Re.*, 30:59–73.

Murphy, J. and Palmer, T. (1986). Experimental monthly long-range forecasts for the united kingdom, part ii. a real-time long-range forecast by an ensemble of numerical integrations. *Meteorol. Mag.*, 115:337–349.

Neal, R., Fereday, D., Crocker, R., and Comer, R. (2016). A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorol. Appl.*, 23:389–400.

Neal, R., Robbins, J., Dankers, R., Mitra, A., Jayakumar, A., Rajagopal, E., and Adamson, G. (2020). Deriving optimal weather pattern definitions for the representation of precipitation variability over India. *Int. J. Climatol.*, 40:342–360.

Omran, M., Engelbrecht, A., and Salman, A. (2007). An overview of clustering methods. *Intell. Data Anal.*, 11:583–605.

Palmer, T. (2018). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Q. J. R. Meteorol. Soc.*, 145 (Suppl. 1):12–24.

Parfitt, R., Czaja, A., and Seo, H. (2017). A simple diagnostic for the detection of atmospheric fronts. *Geophys. Res. Lett.*, 44:4351–4358.

Pattanaik, R. (2022). A Report on Numerical Weather Prediction Products For Sectoral Applications. `https://nwp.imd.gov.in/NWP_REPORT_2022.pdf` [Accessed: Feb 2023].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau,

D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830.

Persson, A. (2005a). Early operational Numerical Weather Prediction outside the USA: an historical introduction Part I: Internationalism and engineering NWP in Sweden, 1952-69. *Meteorol. Appl.*, 12:135–159.

Persson, A. (2005b). Early operational Numerical Weather Prediction outside the USA: an historical introduction Part II: Twenty countries around the world. *Meteorol. Appl.*, 12:269–289.

Persson, A. (2005c). Early operational Numerical Weather Prediction outside the USA: an historical introduction Part III: Endurance and mathematics - British NWP, 1948-1965. *Meteorol. Appl.*, 12:381–413.

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., Huth, R., James, P., Jourdain, S., Kreienkamp, F., Krennert, T., Lykoudis, S., Michalides, S., Pianko-Kluczynska, K., Post, P., Álvarez, D., Schiemann, R., Spekat, A., and Tymvios, F. (2010). Cost733cat - A database of weather and circulation type classifications. *Phys. Chem. Earth*, 35:360–373.

Philipp, A., Della-Marta, P., Jacobeit, J., Fereday, D., Jones, P., Moberg, A., and Wanner, H. (2007). Long-Term Variability of Daily North Atlantic-European Pressure Patterns since 1850 Classified by Simulated Annealing Clustering. *J. Climate*, 20:4065–4095.

Potter, S., Kreft, P., Milojev, P., Nobel, C., Montz, B., Dhellemmes, A., Woods, R., and Gauden-Ing, S. (2018). The influence of impact-based severe weather warnings on risk perceptions and intended protective actions. *Int. J. Disast. Risk Re.*, 30:34–43.

Prichard, B. (2018). Weather log. *Weather*, 73(12):i–iv.

Renard, R. and Clarke, L. (1965). Experiments in numerical objective frontal analysis. *Mon. Weather Rev.*, 93:547–556.

Richardson, D., Buizza, R., and Hagedorn, R. (2005). First workshop on the thorpex interactive grand global ensemble (tigge), final report. *World Meteorological Organization Report*, WMO/TD-No. 1273:1–39.

Richardson, D., Neal, R., Dankers, R., Mylne, K., Cowling, R., Clements, H., and Millard, J. (2020). Linking weather patterns to regional extreme precipitation for highlighting potential flood events in medium- to long-range forecasts. *Meteorol. Appl.*, 27(4).

Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorol. Appl.*, 15:163–169.

Roberts, N. and Lean, H. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Mon. Weather Rev.*, 136:78–97.

Rodwell, M., Hammond, J., Thornton, S., and Richardson, D. (2020). User decisions, and how these could guide developments in probabilistic forecasting. *Q. J. R. Meteorol. Soc.*, 146:3266–3284.

Rutz, J., Staudenmaier, M., Jeglum, M., and Lamberson, B. (2022). Understanding the WPC Cluster Analysis Tools. `https://origin.wpc.ncep.noaa.gov/wpc_ensemble_clusters/cluster_analysis.pdf` [Accessed: Feb 2023].

Santoalla, D. and Mladek, R. (2022). *TIGGE archive*. `https://confluence.ecmwf.int/display/TIGGE` [Accessed: Dec 2022].

Scott, A. and Symons, M. (1971). Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27.

Serafin, S., Strauss, L., and Dorninger, M. (2019). Ensemble reduction using cluster analysis. *Q. J. R. Meteorol. Soc.*, 145:659–674.

Shuman, F. (1989). History of Numerical Weather Prediction at the National Meteorological Center. *Weather and Forecasting*, 4:286–296.

Skok, G. (2015). Analysis of Fraction Skill Score properties for a displaced rainband in a rectangular domain. *Meteorol. Appl.*, 22:477–484.

Skok, G. (2016). Analysis of fraction skill score properties for a displaced rainy grid point. *Atmos. Res.*, 169:556–565.

Skok, G. and Roberts, N. (2016). Analysis of fractions skill score properties for random precipitation fields and ECMWF forecasts. *Q. J. R. Meteorol. Soc.*, 142:2599–2610.

Skok, G. and Roberts, N. (2018). Estimating the displacement in precipitation forecasts using the fractions skill score. *Q. J. R. Meteorol. Soc.*, 144:414–425.

Soster, F. and Parfiti, R. (2022). On Objective Identification of Atmospheric Fronts and Frontal Precipitation in Reanalysis Datasets. *J. Climate*, 35:4513–4534.

Stephenson, D. B. (2008). *Definition, diagnosis, and origin of extreme weather and climate events*, page 11–23. Cambridge University Press.

Straus, D., Corti, S., and Molteni, F. (2007). Circulation Regimes: Chaotic Variability versus SST-Forced Predictability. *J. Climate*, 20:2251–2272.

Taylor, A., Kox, T., and Johnston, D. (2018). Communicating high impact weather: Improving warnings and decision making processes. *Int. J. Disast. Risk Re.*, 30:1–4.

Teague, K. and Gallicchio, N. (2017). *The Evolution of Meteorology*. John Wiley & Sons, Ltd., Chichester, United Kingdom.

Toth, Z. and Kalnay, E. (1993). Ensemble Forecasting at NMC: The Generation of Perturbations. *Bull. Am. Meteorol. Soc.*, 74(12):2317–2330.

Tracton, M. and Kalnay, E. (1993). Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects. *Weather and Forecasting*, 8:379–398.

Tveito, O., Huth, R., Philipp, A., Post, P., Pasqui, M., Esteban, P., Beck, C., Demuzere, M., and Prudhomme, C. (2016). *COST Action 733: Harmonization and Application of Weather Type Classifications for European Regions*.

Uppala, S., Kållberg, P., Simmons, A., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J., Haseler, J., Hernandez, A., Kelly, G., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R., Andersson, E., Arpe, K., Balmaseda, M., Beljaars, A., Vande Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B., Iskasen, L., Janssen, P., Jenne, R., McNally, A., Mahfouf, J., Morcrette, J., Rayner, N., Saunders, R., Simon, P., Sterl, A., Trenberth, K., Untch, A., Vasiljevic, D., Viterbo, P., and Wollen, J. (2005). The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.*, 131:2961–3012.

Wazarkar, S. and Keshavamurthy, B. (2018). A survey on image data analysis through clustering techniques for real world applications. *J. Vis. Commun. Image Represent.*, 55:596–626.

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL-A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Weather Rev.*, 136:4470–4487.

Wilks, D. (2019). *Statistical Methods in the Atmospheric Sciences (4 ed.)*. Elsevier, Amsterdam, Netherlands.

WMO (2019). Manual on the Global Data-processing and Forecasting system, Annex IV to the WMO Technical Regulations. `https://library.wmo.int/?lvl=notice_display&id=12793#.Y6be6Xany3A` [Accessed: Dec 2022].

Zheng, M., Chang, E., Colle, B., Luo, Y., and Zhu, Y. (2017). Applying Fuzzy Clustering to a Multimodel Ensemble for U.S. East Coast Winter Storms: Scenario Identification and Forecast Verification. *Weather and Forecasting*, 32:881–903.

Appendix A

Testbed Survey Questions

The method presented in this work was evaluated by participants of the Met Office winter testbed 2021-22 (see chapter 6). The full list of questions in the survey are presented below for reference, separated by sections.

Survey questions

1. Please input today's date.

Ensemble initialisation and lead time information

2. Please choose the initialization time of today's ensemble run you are viewing.
    - 0000 UTC
    - 0600 UTC
    - 1200 UTC
    - 1800 UTC
3. How many clusters are present?
    - 2
    - 3
    - 4
    - 5
    - 6
4. What is the window of interest period?

5. To answer the following sections you will examine the plots and maps using a designated series of lead times provided during the brief. Please indicate what lead times you are using in hours.

6. Do you have time to answer questions about the clustering?

    - Yes - continue to the next section
    - No - skip to closing comments and the end of the survey

Clustering questions

7. Do the representative members indicate distinct weather scenarios?

    - All representative members are distinct from each other
    - Some representative members are distinct
    - Only 1 representative member is distinct
    - They all look the same

8. Explain why they are distinct or not.

9. Is there an important meteorological event in the full ensemble that does not have a close representative member?

    - Yes
    - No

10. If yes, which member(s) and which representative member(s) are they different from?

11. Would you cluster the members similarly to how they are being clustered (using only the lead time at the beginning of the lead time series provided during the brief)?

    - Yes
    - No

12. If you answered "no" to the previous question, please elaborate on how you would cluster members differently.

13. Are there too many or too few clusters?

- Too many
- Right number
- Too few

14. Do you have time to answer questions about forecasting and communication?

    - Yes - continue to the next section
    - No - skip to closing comments and the end of the survey

Forecasting and communication

15. Would the cluster information (representative members) for this forecast be useful in creating a forecast communication to the general public (e.g., informing the warning impact matrix)?

    - Very useful
    - Somewhat useful
    - Not at all useful
    - Not applicable - no significant weather

16. Is the cluster information useful in creating a forecast message to specific users?

    - Very useful
    - Somewhat useful
    - Not at all useful
    - Not applicable - no significant weather

17. If you answered "very useful" or "somewhat useful" to the previous questions, what areas of interest would it be for (i.e., emergency response, local authorities, aviation)?

18. Do you have time to answer questions about high-impact scenarios and scenarios across valid times?

    - Yes
    - No

High-impact scenarios and scenarios across valid times

19. Are any of the representative members at the lead time provided displaying a possible high impact (i.e., is there a possibility this member would require issuing a warning for rain, wind, snow etc.) scenario?

    - Yes
    - No

20. What type of high impact weather is associated with the scenario? Tick all that apply.

    - High winds
    - Heavy rain
    - Snow
    - Potential flooding
    - Extreme temperatures
    - Road icing
    - Fog or low visibility
    - Other

21. How many clusters contain a high-impact scenario?

    - 1
    - 2
    - 3
    - 4
    - 5
    - 6

22. Is there a high-impact scenario you previously noticed present across multiple initializations?

    - Yes - in all initializations

- Yes - in most (greater than half) initializations
- Yes - in half the initializations
- Yes - in some (less than half) initializations
- No

23. Is the presence of this high-impact event as a potential scenario likely to impact your forecast message?
    - Yes
    - No
24. Why or why not?
25. Do the scenarios appear across multiple initializations? (Available if "No" was chosen for Q19 and instead of Q20 - Q24.)

Closing comments

26. If you have any further comments about today's clustering, please add them here.
27. Considering today's forecast, to what extent do you find the clustering more efficient than looking through the full ensemble? (From 1 to 5 stars where 1 is not at all efficient and 5 is very efficient).
28. Please provide your name.