MASTERARBEIT

Titel der Masterarbeit

Clustering ECMWF ENS ensemble predictions to optimise FLEXPART plume dispersion ensembles

verfasst von

Robert Klonner, B.Sc.

angestrebter akademischer Grad

Master of Science
(M.Sc.)

Wien, 2013

Studienkennzahl:	A 066 614
Studienrichtung:	Masterstudium Meteorologie
Betreuer:	Ao. Univ.-Prof. Mag. Dr. Leopold Haimberger

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment. [...] The never-satisfied man is so strange; if he has completed a structure, then it is not in order to dwell in it peacefully, but in order to begin another.

Wahrlich es ist nicht das Wissen, sondern das Lernen, nicht das Besitzen sondern das Erwerben, nicht das Da-Seyn, sondern das Hinkommen, was den grössten Genuss gewährt. [...] so sonderbar ist der nimmersatte Mensch, hat er ein Gebäude vollendet so ist es nicht um nun ruhig darin zu wohnen, sondern um ein andres anzufangen.

Carl Friedrich Gauss (1777-1855)

Abstract

It is commonly known today that ensemble forecasting can yield very valuable information about the uncertainty of a forecast. However, implementing ensemble forecasts in an operational environment is difficult due to the huge amounts of data that have to be retrieved and preprocessed and the computational demands for simulating several dispersion forecasts. Therefore, this thesis deals with the construction of reduced ensembles and the designed methodology is used to forecast the dispersion of volcanic ash.

A method is developed to reduce the number of ensemble members to a reasonable amount according to a physical basis. The idea is to find a linkage between the horizontal wind field in the meteorological input data and corresponding ash concentration fields in the output of the dispersion model FLEXPART with the help of cluster analysis (CA). Results of three clustering algorithms are investigated, namely Ward's method, K-means and Partitioning Around Medoids (PAM). The clustering is also refined by adapting the area of clustering to the domain that is affected by the ash plume. It is shown that the main patterns can be reproduced only with information about the wind field. This allows the construction of a reduced ensemble while keeping retrieval, preprocessing and simulation times short enough for operational purposes.

To prove the connection between wind and ash concentration fields three validation methods have been developed. The essential part for validation is the definition of a reference cluster solution to compare the findings. This is solved by using the cluster solution of an ensemble of concentration forecasts simulated with all members of the meteorological input. A qualitative approach is based on the visualisation of a classification by using multidimensional scaling (MDS) to analyse the reproduced groups. Furthermore, the cluster comparison measures Rand Index (RI) and Variation of Information (VI) are used. As a main target, the reduced ensemble should have the same ensemble spread as the original one. For measuring the reproduced spread the variance against the ensemble mean for both ensembles is calculated and compared.

To test the method for practical applications it is applied to hypothetical volcanic events. It is found that for synoptic situations where the forcing through the horizontal wind field is strong enough, the original ensemble spread is sufficiently reproduced with a set of 5-7 representative members.

Zusammenfassung

Es ist allgemein bekannt, dass Ensemblevorhersagen wertvolle Informationen über die Unsicherheit von Vorhersagen geben können. Die Implementierung solcher Systeme für operationelle Anwendungen wird jedoch durch die großen Datenmengen die bearbeitet werden müssen und durch die rechenaufwendigen Ausbreitungssimulationen erschwert. Deshalb beschäftigt sich diese Masterarbeit mit der Konstruktion reduzierter Ensembles und der Anwendung dieser Methode für die Vorhersage von Vulkanascheausbreitung.

Es wird ein Verfahren entwickelt um die Anzahl der Ensemblemitglieder aufgrund physikalischer Zusammenhänge zu verringern. Die Idee ist eine Verbindung zwischen dem horizontalen Windfeld in den meteorologischen Eingangsdaten und den zugehörigen Aschekonzentrationsfeldern vom Ausbreitungsmodel FLEXPART mittels Clusteranalyse (CA) zu finden. Die Ergebnisse von drei Clusteralgorithmen Ward's method, K-means und Partitioning Around Medoids (PAM) wurden untersucht. Es zeigt sich, dass ein Großteil der Muster nur anhand von Informationen über das Windfeld reproduziert werden kann. Das erlaubt die Zusammenstellung eines reduzierten Ensembles während die Bearbeitungs- und Simulationszeit auf ein Minimum beschränkt wird.

Um die Verbindung zwischen den beiden Feldern zu validieren wurden drei Verfahren entwickelt. Die Hauptaufgabe dabei ist die Definition einer Referenzlösung um die Ergebnisse zu kontrollieren. Das Problem wird durch Heranziehen einer Clusterlösung von Konzentrationsensemblevorhersagen gelöst. Ein qualitativer Ansatz basiert auf der Visualisierung von Klassifikationen mithilfe einer multidimensionalen Skalenanalyse (MDS) um die reproduzierten Gruppen zu analysieren. Weiters werden die Indizes Rand Index (RI) und Variation of Information (VI), die zum Vergleich von Clusterlösungen dienen, herangezogen. Ein Hauptziel ist es, den Spread des originalen Ensembles zu erhalten. Als Maß für den Spread wurde die Varianz gegen das Ensemblemittel für beide Ensembles berechnet und verglichen. Ein weiteres Konzept wird entwickelt um die Fläche für das Clustering der von der Aschewolke betroffenen Region anzupassen.

Um die Methode an einer praktischen Anwendung zu demonstrieren, wird diese an theoretischen Vulkanausbrüchen getestet. Bei synoptischen Situationen mit ausreichend Einfluss des horizontalen Windfeldes kann der originale Spread des Ensembles mit 5-7 repräsentativen Mitgliedern ausreichend abgebildet werden.

Contents

List of Figures

List of Tables

List of Acronyms

a.s.l. above sea level

CA cluster analysis

CP convective precipitation

cRI corrected Rand Index

ECMWF European Center for Medium-range Weather Forecasts

EDA Ensemble of Data Assimilations

EM ensemble mean

ENS ensemble of lower-resolution forecasts

EPS Ensemble Prediction System

EWSS east-west surface stress

GFS Global Forecast System model

HRES high-resolution forecast

HYSPLIT Hybrid Single Particle Lagrangian Integrated Trajectory Model

IFS Integrated Forecasting System

LL latitude/longitude grid

LSP large scale precipitation

MARS Meteorological Archival and Retrieval System

MDS multidimensional scaling

NEMO Nucleus for European Modelling of the Ocean

NEMOVAR variational data assimilation system for the ocean model NEMO

NSSS north-south surface stress

nVI normalised Variation of Information

PAM Partitioning Around Medoids

PCoA principal coordinate analysis

RGG reduced Gaussian grid

RI Rand Index

RMSE root mean square error

SD standard deviation

SH spherical harmonics

SKEB stochastic back-scatter scheme

SPPT stochastically parametrised tendency scheme

SSHF surface sensible heat flux

SSR surface solar radiation

VAR variance

VAST Volcanic Ash Strategic-initiative Team

VI Variation of Information

WAM Wave Model

WRF Weather Research and Forecasting model

ZAMG Zentralanstalt für Meteorologie und Geodynamik

1 Introduction

1.1 Motivation

In April and May 2010, the eruption of the volcano Eyjafjallajökull in Iceland brought huge amounts of volcanic ash into the atmosphere. As a result of high ash concentrations the European airspace was shut down for several days. The financial damage caused by this event has been estimated at €5 billion (Oxford Economics, 2010). The significant economic impact and public awareness of these air traffic restrictions were the reason for the initiation of several projects to minimise the impact of volcanic eruptions on civil aviation (VAST technical proposal, 2012).

One of these projects is the Volcanic Ash Strategic-initiative Team (VAST) project and its targets are described in the VAST technical proposal (2012). An aim of VAST is the demonstration of an operational volcanic ash monitoring and forecasting service. To estimate the uncertainty of the forecasts ensemble techniques will be used. Within the VAST project two methods shall be explored, namely the use of multiple-model multiple-input ensembles (e.g. JRC ENSEMBLE platform) and single-model multiple-input ensembles. This master thesis deals with the latter approach. The dispersion model FLEXPART (3.1) is used as single-model and the multiple meteorological input comes from the European Center for Medium-range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) (2.3.3). The ECMWF ensemble of lower-resolution forecasts (ENS) currently consists of 51 ensemble members and since it is not feasible to run 51 dispersion simulations with FLEXPART operationally, the number of ensemble members needs to be reduced. A method to accomplish this task is to apply cluster analysis (CA) techniques (2.1.1) to the ensemble.

1.2 State of research

The benefits of ensemble forecasting are commonly known today and therefore also used in the field of dispersion forecasting. Galmarini et al. (2004) investigated the multi-model ensemble approach because the results of different dispersion models do not yet converge in every case and so additional information about forecast uncertainty is essential. With a single-model multiple input approach, Challa et al. (2008) made ensemble forecasts for SO_2 with the dispersion model HYSPLIT. They ran the Weather Research and Forecasting model (WRF) model with different physical parametrisation schemes to gain an ensemble of meteorological input.

Applying CA to ensembles can be a helpful tool to summarise the large amount of information they provide. For example, ECMWF implemented a modified K-means algorithm to cluster 500 hPa geopotential fields of their ENS product over the North Atlantic and Europe for several time windows (Ferranti and Corti, 2011). They try to identify different synoptic flow patterns and represent each cluster by its central member. The number of clusters varies between 1-6 depending on the ensemble spread of the current forecast.

Another example for using CA is clustering trajectories to identify synoptic patterns which are responsible for heavy precipitation (Seibert et al., 2007). They combined a hierarchical centroid linkage algorithm with a rearrangement procedure after each iteration as it is often used in nonhierarchical algorithms. Marzban and Sandgathe (2005) clustered forecast and observation precipitation fields to identify given features and compared the results for verification purposes. Beaver and Palazoglu (2006) used CA to investigate hourly wind measurements to determine synoptic regimes that can affect air quality.

As mentioned by Lee et al. (2012) it is desirable to reduce the number of a given forecast ensemble. Computational resources often limit the size of ensembles and so choices must be made about which members to include in the forecast. They use principal coordinate analysis (PCoA) and the clustering algorithm K-means to reduce the ensemble size.

1.3 Goals and outline

The most essential factor for operational forecasts after an event like a volcano eruption is time. Decision makers need as much information as possible in the shortest achievable time range. On the side of the operational system we see computational limitations and a huge amount of data that need to be retrieved and preprocessed. For these reasons the main goal of the investigation is to analyse to which amount the ensemble can be reduced compared to the full ensemble while keeping the necessary information content and satisfying operational demands.

In order to reduce the ensemble it seems reasonable to use CA. For this method it has to be examined which meteorological parameters should be chosen for clustering and which clustering algorithm will be applied. The parameters which have the strongest linkage to the FLEXPART concentration fields should be chosen. The choice of spatial resolution, temporal resolution and a selection of vertical levels for clustering are also crucial questions. It would be a big advantage if one does not have to retrieve the whole ensemble for the clustering input in order to find representative members for the forecasts. The developed method should then be applied to some test scenarios to produce dispersion forecasts for volcano eruptions from a reduced ensemble. Furthermore, validation measures must be found to check the quality of the clustering.

Chapter 2 of this thesis describes the theoretical background for CA and ensemble forecasting. The dispersion model FLEXPART and properties of provided meteorological data are presented in chapter 3. The developed concepts and their implementation are described in chapter 4. Chapter 5 shows the results of the work. Chapter 6 contains a discussion of the findings and in the last chapter conclusions are drawn and suggestions for future studies are provided.

2 Theoretical background

2.1 Cluster analysis

2.1.1 Introduction

This introduction follows the presentation of Wilks (2006) unless denoted otherwise. Cluster analysis (CA) is used to find groups in data. At the beginning neither the identities of the groups nor the correct number of groups is known. The data consist of individual observations (objects) and the target is to assign a group membership for every observation so that similar observations belong to the same group. To achieve this assignment, a measure of the attribute similarity has to be defined. This is done by using the idea of distance between objects. Data, on which the distance measure are applied, are arranged as a $(n \times P)$ data matrix where everyone of the n rows is a data vector in a P-dimensional space. A data vector can be a single value but it can also consist of different variables at multiple times. The most commonly used measure between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is the Euclidean distance (2.1.1). For this measure, the observations in a cluster should be separated by small distances whereas the distances to the members of the other clusters should be relatively high. There are many other possible choices for distance measures, for example, the angle between the pairs of vectors or the correlation between them.

$$d_{i,j} = \left[\sum_{k=1}^{P}(x_{i,p} - x_{j,p})^2\right]^{\frac{1}{2}} \tag{2.1.1}$$

After choosing a distance measure, all distances between the n(n-1)/2 possible different combinations are calculated and the resulting values can be arranged in a $(n \times n)$ matrix called distance matrix. This matrix is symmetric and the values of the main diagonal are zero (zero distance between $\mathbf{x}$ and itself). The indices i, j in equation (2.1.1) indicate the entries of this matrix. Selecting the distance measure is the basis for clustering and it has to be carefully chosen according to the data. For example, if the data consists of variables with different measurement units, the result will be strongly influenced by the variable with higher values when using the Euclidean distance. To overcome this problem data can be standardised by the variance of the observations. If the observations contain outliers, standardising by the ranges of the variables would be a better choice.

In CA two main categories are distinguished, namely hierarchical and partitioning (nonhierarchical) methods. The first category can be further divided into agglomerative and divisive methods. The most commonly used methods are hierarchical and agglomerative. They build a hierarchy of sets of groups which is formed by merging previously defined groups in several steps. At the beginning, there is no group structure; this means every of the n observations is its own group. The first step is to find the two most similar observations in their P-dimensional space and merge them to a new group. There are many possible methods for calculating the representative distance for this new group (e.g. average-linkage). Once an observation is assigned it stays in that group. It can only change its assignment if its group gets merged with another group. After this step n-1 groups exist; one of them with two members. The process ends after n-1 steps when all members are assigned to one group. Here, neither the grouping at the beginning nor at the end is helpful but hopefully a useful clustering emerges at an intermediate stage that reflects similar data generating processes. This stage minimises the distances within a cluster and maximises the distances among the clusters. Divisive methods start from the other side of the hierarchy. At the beginning, all observations are in one group and this group is split stepwise into more similar groups. A weakness of hierarchical methods is that a member cannot change its group assignment during the clustering process. This can be a disadvantage if an observation is misclassified at the beginning and would better fit to another group at the end. Methods where reassignment is possible during the analysis are called nonhierarchical. These clustering algorithms also use a distance measure to group observations.

An important practical problem in CA is to determine the number of clusters (groups) which should be used. As mentioned above, hierarchical methods provide grouping solutions for every intermediate step. A popular approach to find the most suitable clustering stage is to analyse a plot where the ordinate shows the distance between merged clusters as a function of the analysis stage. This distance will be small at early stages where similar clusters get merged. When clusters are merged whose inner cluster distances are small, the distances between these groups are large. If a distinct increase of distance is found in this plot, the process can be stopped right before this jump. For clearly separated groups, the best number of clusters might be seen directly in the dendrogram (tree diagram). In case of a commonly used nonhierarchical algorithm, K-means, this approach is not possible because the number of clusters has to be fixed at the beginning. As mentioned by Kaufman and Rousseeuw (2005), one

possibility is to run the algorithm several times for different numbers of clusters and to analyse the results with certain graphics or characteristics. A useful graphical tool to analyse and compare the quality of such clustering solutions is the silhouette plot introduced by Rousseeuw (1987).

As CA is, for the main part, an exploratory data analysis tool, it can find groups that may have been overlooked with more inferential tools. In the best case it leads to an empirically useful separation of the data or it possibly suggests physical bases that lead to the structure of the underlying data. For example, Cheng and Wallace (1993) applied CA to data of geopotential height to identify atmospheric flow regimes. CA can also be used to classify solutions that are developing in an ensemble forecast due to different possible weather regimes as it is illustrated in figure (2.3).

2.1.2 Wards's method

The presentation of Ward's method follows Kaufman and Rousseeuw (2005) unless denoted otherwise. Ward's method or Ward's minimum variance method belongs to the category of hierarchical CA methods and was introduced by Ward (1963) who suggested an agglomerative approach. It is designed to be used for data consisting of interval-scaled measurements (Kaufman and Rousseeuw, 2005) and it does not operate on the distance matrix (Wilks, 2006). To understand the distance measure one needs the definition of a centroid, which is the average of all observations in a cluster. We denote a cluster R that consists of $n = (1, ..., f)$ observations, where x_{if} is the fth measurement ($f = (1, ...p)$) of the ith observation. For the centroid of R we get the point $\bar{x}(R) = (\overline{x_1}(R), \overline{x_2}(R), ..., \overline{x_p}(R))$ and its fth coordinate is defined by

$$\overline{x_f}(R) = \frac{1}{|R|} \sum_{i \in R} x_{if} \tag{2.1.2}$$

If all points of R have a physical mass, the centroid $\bar{x}(R)$ would be the center of gravity of R. After calculating the centroid for each cluster, the error sum of squares (ESS) is determined. This is done by taking the sum of squared Euclidean distances between the observations of the cluster C and its centroid

$$ESS(C) = \sum_{i \in C} ||x_i - \bar{x}(C)||^2 \tag{2.1.3}$$

which is a measure of the tightness of a cluster. To know the total ESS at each step, the sum of ESS(C) over all clusters has to be computed. For the next stage two clusters A and B are merged to form a new cluster R. We are interested in how much the total ESS would increase through this particular fusion. The ESS of A and B combined can only increase or stay the same (in rare cases) and the ESS of the other clusters remain the same. So the increase of total ESS is

$$\Delta ESS = ESS(R) - ESS(A) - ESS(B) = \tfrac{1}{2} d^2(A, B) \tag{2.1.4}$$

Wishart (1969) had the idea to use the dissimilarity measure

$$d^2(A, B) = \frac{2|A||B|}{|A| + |B|} ||\overline{A} - \overline{B}||^2 \tag{2.1.5}$$

to establish (2.1.4).

In order to choose the best pair to merge from C+1 clusters to get C clusters, ΔESS has to be calculated for all of the C(C+1)/2 possible pairs (Wilks, 2006). The pair with the smallest ΔESS gets fused at each step (Ward, 1963).

Summarised we can say that Ward's method minimises the sum of squared distances between observations and the centroids of their respective groups, summed over all clusters. Or in simpler words, it minimises the sum of within-groups variances.

2.1.3 K-means

This section follows the presentation of Morissette and Chartier (2013) unless denoted otherwise. K-means is the most commonly used nonhierarchical method and has existed for more than 50 years now (Steinhaus, 1956). Although there are many other modern algorithms available, it is still popular because it is easy to implement and computationally efficient. K stands for the number of groups into which the data gets clustered. In contrast to hierarchical algorithms K must be fixed at the beginning of the clustering process. Similar to Ward's method (2.1.2) it is a variance minimisation technique that is based on construction of central points, namely centroids as defined in (2.1.2) (Kaufman and Rousseeuw, 2005). There exist several different adaptations, of

which the most widely used are the algorithms of Hartigan and Wong (1979), MacQueen (1967), Forgy (1965) and Lloyd (1957, 1982). In the following, the method described by MacQueen (1967) will be presented.

As a first step, the initial centroids for the K clusters have to be chosen. There are several possible ways of doing that. For example, one can use previous empirical knowledge of the data if available, pick K random observations from the data set or just set them to random values within the data space of the observations. After that, each observation of the data gets assigned to its nearest centroid according to a distance metric, e.g. Euclidean distance mentioned in (2.1.1). Now the initial centroids can be calculated by averaging the observations within every cluster. Finally, an iterative process is started to check whether there exists a centroid for a case i that it is closer to. If yes, it gets reassigned and the centroids for the two affected clusters are recalculated. The process proceeds until no reassignment for an observation has taken place in a cycle. In the following pseudocode the steps are summarised:

1. Choose the number of clusters
2. Choose the distance measure
3. Choose a way to pick initial centroids
4. Assign initial centroids to all observations
5. While there occur reassignments in a cycle
   a. For $i \leq$ number of observations
      i. Assign observation i to its nearest centroid according to metric
      ii. Recompute centroids for the two involved clusters
   b. Recompute centroids

The cluster centres are initialised in the same way for all four implementations mentioned. The main difference between MacQueen (1967) to the Lloyd (1957, 1982) and Forgy (1965) algorithms is that there is neither a recalculation of the centroid after an observation is reassigned nor after each cycle. In contrast to the other algorithms, the version of Hartigan and Wong (1979) searches for the solution of locally optimal within-cluster sum of squares of errors. That means an observation is possibly assigned to another cluster although it currently belongs to the cluster with the closest centroid. By doing this, the method minimises the total within-cluster sum of squares.

If the determination of the initial centroids is found by a random selection, an

important characteristic of the algorithm is that it must not converge to the same clustering solution (global minimum) for several runs on the same data set. This behaviour can often be seen for data sets containing very similar observations where multiple local minima can exist.

2.1.4 Partitioning Around Medoids

The presentation of this section follows Kaufman and Rousseeuw (2005) unless denoted otherwise. Partitioning Around Medoids (PAM) is a nonhierarchical method introduced by Kaufman and Rousseeuw (1987). It is related to the K-means algorithm (2.1.3) in a way that the number of clusters have to be defined at the beginning of the clustering. Also, both try to minimise the distance between observations in a cluster and an element which is designated as the centre of the cluster. While for K-means this central element is the centroid of a cluster it is the medoid for PAM. A medoid has the following characteristics: it must be a member (observation) of the data set and its average dissimilarity to all other observations is minimal. PAM aims to minimise the sum of the average dissimilarities of observations to their closest representative observation, the medoid.

The algorithm consists of two phases. At first, an initial clustering is set up by the BUILD phase. This is done by the stepwise selection of representative observations until K representative objects (medoids) have been found where K is the requested number of clusters into which the data should be separated. The process starts by determining the most centrally located observation in the data (minimal average dissimilarity to all other objects). For each iteration that comes next another medoid is chosen by carrying out the procedure which is explained by the following pseudocode:

1. For all objects i which are not yet a medoid
   a. For all other nonselected objects j
      i. Compute the difference between j and its distance D_j to the nearest previously selected medoid an its distance $d(j, i)$ to object i
      ii. Only a positive distance should result in a contribution to select object i as a medoid, thus we compute: $C_{ji} = max(D_j - d(j, i), 0)$
   b. Compute the total gain for the hypothetical medoid i: $TG_i = \sum_j C_{ji}$
2. Set that i as next medoid which maximises TG_i

Phase two of the algorithm is called SWAP phase and it should improve the set of selected medoids and therefore improve the clustering emerging from this set. To do this, all pairs of objects (i, h), where i has been selected as medoid and h has not, undergo a process. It is checked how the clustering is effected by swapping i and h, i.e. setting h as a representative object and not i. To calculate this effect a third object j is taken into account and for the constellation of those three the contribution to the swap is determined. This is illustrated by following pseudocode:

1. Choose a nonselected object j and compute its contribution C_{jih} to the swap by considering the following three differentiations:
   a. If j is farther away from i and from j than from one of the other medoids:
      $C_{jih} = 0$
   b. If j is not more distant from i than from any other medoid ($d(j, i) = D_j$, see BUILD phase), there are two possible cases. By denoting E_j to be the dissimilarity between j and the second most similar medoid, we can define the contribution for the two cases:
      i. j is nearer to h than to the second closest medoid ($d(j, h) < E_j$):
         $C_{jih} = d(j, h) - d(j, i)$
      ii. j is at least as far away from h than from the second closest medoid ($d(j, h) \geq E_j$):
         $C_{jih} = E_j - D_j$
   c. If j is further from i than from at least one of the other medoids but closer to h than to any medoid:
      $C_{jih} = d(j, h) - D_j$
2. Compute the total contribution to the swap: $T_{ih} = \sum_j C_{jih}$
3. Choose the pair (i, h) that minimises T_{ih}
4. If the minimum $T_{ih} < 0$: carry out the swap,
   if the minimum $T_{ih} \geq 0$: do not swap i and h

In contrast to the K-means method the initial determination of the central objects does not depend on a random process. Therefore, the algorithm yields the same clustering solution if it is applied multiple times on the same data set.

2.2 Cluster visualisation and validation

2.2.1 Introduction

To investigate the result of a CA method it can be very helpful to use graphical tools for visualisation. An often used graphical representation of a cluster solution is the dendrogram. It has a tree like structure and it shows the relationship between observations at different levels of aggregation. However, this tool can only be used for hierarchical clustering algorithms (Wilks, 2006). A method that can be applied to both, hierarchical and partitioning methods, is the silhouette plot mentioned in section (2.1.1) where one can see how well the observations lie within their cluster. To visualise how the clusters are related to each other and to identify similar groups another method might be more intuitive to interpret. It is called multidimensional scaling and is presented in section (2.2.2).

There exist several cluster validation tests and they can be classified into three different types after Jain and Dubes (1988): *External* tests that compare a classification (cluster solution) with information that was not used for constructing the classification, *internal* tests that compare parts of the solution with the initial data set and *relative* tests that compare several different results of classifications gained by the same data set. For the first category two validation methods are presented, namely the Rand Index (2.2.3) and Variation of Information (2.2.4).

2.2.2 Multidimensional scaling

Referring to Marida et al. (1997) and Gordon (1999), Multidimensional scaling (MDS), also referred to as principal coordinate analysis (PCoA) or ordination in ecology, deals with the representation of objects as points in a low dimensional space (usually Euclidean space) by using the distances between the objects. The interpoint distances should represent the distances which the objects have in higher space as good as possible but it might not be feasible to capture all the variability that is present in the data set. With this method one can visualise the relationship between the objects in a two- or three-dimensional plot that hopefully provides a meaningful interpretation of the data set. In this plot it can be observed which objects are similar to each other (small interpoint distances) and which are not. Therefore, this technique can be a helpful tool in CA to visualise the cluster solution and to check if the clusters contain similar

objects or not.

The following methodology is sometimes called *classical scaling* and was developed by Schoenberg (1935), Young and Householder (1938) and Torgerson (1952). Gower (1966) popularised this method under the title *principal coordinate analysis*.

The presentation in the next section completely follows Gordon (1999) unless denoted otherwise. The main target can be formulated as a converse problem: given a $n \times n$ distance matrix Δ which represents the squared distances within a set of n points in some Euclidean space, gain the coordinates of these points. For a set of n points in p-dimensional Euclidean space, $\{P_i(i = 1, ...n)\}$ with coordinates $\{x_{ik}(i = 1, ...n; k = 1, ...p)\}$, the squared Euclidean distance Δ_{ij} between any pair of points (P_i and P_j) can be calculated by

$$\Delta_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \tag{2.2.1}$$

For a configuration of points we require that the centroid of the set of points lies at the origin of the coordinates by setting

$$\sum_{i=1}^{n} x_{ik} = 0 \ (k = 1, 2, ...) \tag{2.2.2}$$

Obtaining the values $\{x_{ik}\}$ is a two-stage process where as a first step the matrix Δ_{ij} is transformed into an inner-product matrix $\mathbf{B} \equiv (b_{ij})$, where

$$b_{ij} \equiv \sum_{k} x_{ik} x_{jk} \ (i, j = 1, ..., n) \tag{2.2.3}$$

and from equation (2.2.2)

$$\sum_{j=1}^{n} b_{ij} = 0 = \sum_{i=1}^{n} b_{ij} \tag{2.2.4}$$

It can be shown that

$$b_{ij} = -\tfrac{1}{2}(\Delta_{ij} - \Delta_{i.} - \Delta_{.j} + \Delta_{..}) \tag{2.2.5}$$

where

$$\begin{aligned} \Delta_{i.} = \Delta_{.i} &\equiv \sum_{j=1}^{n} \Delta_{ij}/n \ (i = 1, ..., n) \\ \Delta_{..} &\equiv \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij}/n^2 \end{aligned}$$

The matrix $\mathbf{B}$ can be diagonalised because it is a real symmetric matrix:

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' \tag{2.2.6}$$

where $\mathbf{V} \equiv (\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n})$ is an orthogonal matrix and its columns contain the eigenvectors which correspond to the eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_n)$ that appear in the $(n \times n)$ diagonal matrix $\mathbf{\Lambda}$. The eigenvectors $\mathbf{v_i'v_i}$ have unit length and the eigenvalues λ_i are supposed to be in decreasing order of magnitude $(\lambda_1 \geq \lambda_2 \geq ...\lambda_n)$.
If $\mathbf{X} \equiv (x_{ik})$ we can equivalently express equation (2.2.3) as

$$\mathbf{B} = \mathbf{X}\mathbf{X}' \tag{2.2.7}$$

We can gain the unknown matrix $\mathbf{X}$ by comparing equation (2.2.6) and (2.2.7):

$$\mathbf{X} = V\Lambda^{\frac{1}{2}} = (\sqrt{\lambda_1}\mathbf{v_1}, \sqrt{\lambda_2}\mathbf{v_2}, ..., \sqrt{\lambda_n}\mathbf{v_n}) \tag{2.2.8}$$

The set of interpoint distances can be recovered from a s-dimensional configuration (where $s < n$) by setting the eigenvalues $\lambda_j = 0$ for $(j = s + 1, ..., n)$.
Every row of $\mathbf{X}$ contains the coordinates for the point of the corresponding object in the data. The coordinates are determined by the first s eigenvectors of $\mathbf{B}$. A good approximation for Δ in the subspace R^s is obtained if the first s eigenvalues are large and positive and the others are near 0 (positive or negative). (Marida et al., 1997)

To evaluate the quality of the representation of Δ in R^s through a classical MDS Marida et al. (1997) suggests two *agreement measures*:

$$A_1(s) = \left(\sum_{i=i}^{s} \lambda_i \Big/ \sum_{j=1}^{n} |\lambda_j| \right) \times 100\% \tag{2.2.9}$$

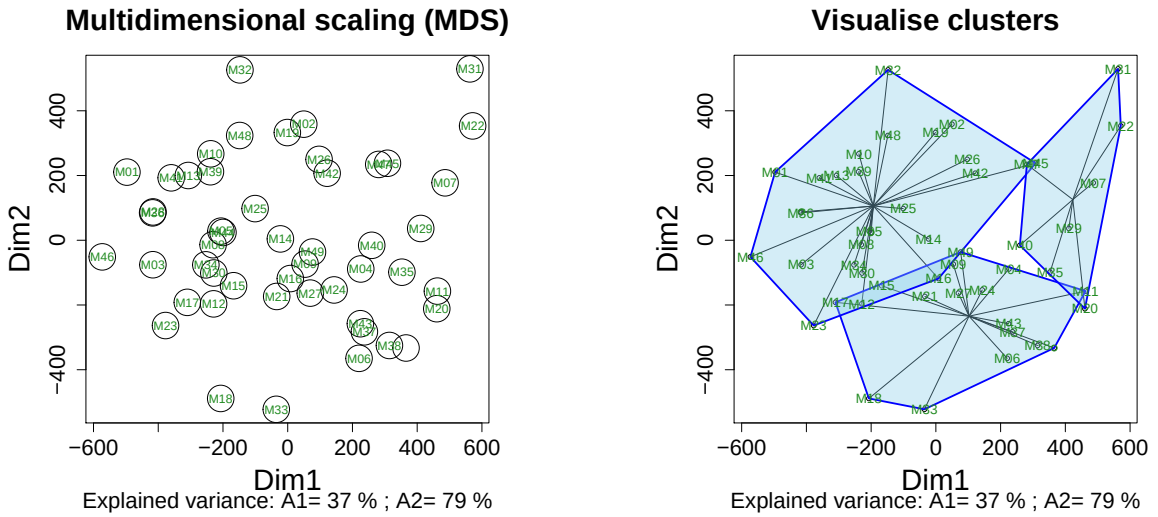
and

$$A_2(s) = \left(\sum_{i=i}^{s} \lambda_i^2 \Big/ \sum_{j=1}^{n} \lambda_j^2 \right) \times 100\% \tag{2.2.10}$$

Absolute values are taken in (2.2.9) because some of the smaller eigenvalues might be negative (Marida et al., 1997). However, the result should be treated carefully if any large negative eigenvalues arise (Gordon, 1999).

An example how to use MDS to visualise a clustering of an ensemble forecast is given in figure (2.1). At first the distances between the members in their higher dimensional space are tried to be represented in two dimensions (figure 2.1a). The amount of information kept can be measured with (2.2.9) and (2.2.10) and are $A_1 = 37$ % and $A_2 = 79$ % for this example. In a second step, one can surround the members for every cluster with a line to denote their group membership. This is shown for a classification into three clusters in figure (2.1b). A clear separation of the groups can indicate a good clustering. Nevertheless, it has to be kept in mind that only a part of the variance can be explained when applying this method to complex high dimensional data. Therefore, the areas of different clusters can overlap in such plots although the clustering is clear in higher dimensions.

As another example for the application of MDS given by Marida et al. (1997) one could consider a distance matrix which holds the road distances (not the direct distances) between cities. If the analysis is carried out properly to represent the information in a two-dimensional space ($s = 2$) the resulting plot should approximately reproduce a geographical map of these cities.



(a) Visualising the distances between ensemble members with inter-point distances in two dimensions gained from MDS.

(b) Surrounding members of the same cluster (dark blue line, light blue area) to indicate their group membership. Black lines point from each member to the centre of the cluster.

Figure 2.1: Example for usage of MDS to visualise a cluster solution with three groups for a 72 h ECMWF ENS forecast 10 km.

2.2.3 Rand Index

The presentation in the next section follows Gordon (1999) unless denoted otherwise. With the Rand Index (RI) one can hopefully answer the question which cluster solution agrees best with an externally-specified partition. A study where five indices were compared by Milligan and Cooper (1986) concluded that the corrected Rand Index (cRI) (also adjusted Rand Index) was best suited for this investigation.

One can consider two partitions of the same set of n objects and denote them by

$$P_1 \equiv \{C_{1k}(k = 1, ..., c_1)\} \;\; and \;\; P_2 \equiv \{C_{2k}(k = 1, ..., c_2)\} \tag{2.2.11}$$

The similarity between these partitions can be determined by using information which is held by the $c_1 \times c_2$ cross-classification table (also contingency table) (n_{ij}), where n_{ij} is defined as the number of objects which are assigned to both partitions $C_{1i}(i = 1, ..., c_1)$ and $C_{2j}(j = 1, ..., c_2)$. One can categorise the $\binom{n}{2}$ possible pairs of observations into

three different types:

- pairs that are part of the same class in P_1 and P_2
- pairs that are part of a different class in P_1 and P_2
- pairs that belong to the same class in one of the partitions and to a different class in the other partion

We summarise the total number of pairs of the first two categories as A ('agreement') and define the number of pairs of the third category as D ('disagreement'). If we denote

$$n_{i.} \equiv \sum_{j=1}^{c_2} n_{ij} \quad \text{and} \quad n_{.j} \equiv \sum_{i=1}^{c_1} n_{ij} \tag{2.2.12}$$

it can be demonstrated that

$$A = \binom{n}{2} + 2\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i.}}{2} + \sum_{j=1}^{c_2}\binom{n_{.j}}{2}\right] \tag{2.2.13}$$

With A and D several measurements for similarity can be defined. The Rand Index (RI) introduced by Rand (1971) has the form

$$RI \equiv \frac{2A}{(n(n-1))} \tag{2.2.14}$$

Hubert and Arabie (1985) proposed an adjustment to the original RI. They argued that RI should be corrected to account for agreement by chance:

$$\begin{aligned} cRI &= \frac{RI - expected\ RI}{1 - expected\ RI} \\ &= \frac{\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}/\binom{n}{2}}{\left[\sum_{i=1}^{c_1}\binom{n_{i.}}{2} + \sum_{j=1}^{c_2}\binom{n_{.j}}{2}\right]/2 - \sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}/\binom{n}{2}} \end{aligned} \tag{2.2.15}$$

The range of cRI is $[0, 1]$ where 1 indicates identical cluster solutions. For very low agreement the value of cRI can also get slightly negative.

2.2.4 Variation of Information

The presentation of the measure Variation of Information (VI) in the next section follows Meila (2007) unless denoted otherwise. This relatively new cluster validation method was introduced by her. The criterion does neither fall into the category *comparing clusterings by counting pairs* (e.g. Rand Index (2.2.3)) nor *comparing clusterings by set matching*. It measures the amount of information lost and gained in changing from one clustering C to another clustering C'.

If we have a data set D with n objects, cluster it into K groups and pick an observation, the probability for this observation to be in the cluster C_k is

$$P(k) = \frac{n_k}{n} \tag{2.2.16}$$

For this discrete random variable one can calculate the entropy

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k) \tag{2.2.17}$$

that is associated with the clustering C and equal to the uncertainty for the consideration above. Entropy is measured in *bits* and is always non-negative.

We further can define the information that one clustering has about the other, called mutual information. The random variables associated with the clusterings C and C' are denoted by $P(k)$ $(k = 1, ..., K)$ and $P'(k')$ $(k' = 1, ..., K')$. The joint distribution $P(k, k')$, which is

$$P(k, k') = \frac{|C_k \cap C'_{k'}|}{n} \tag{2.2.18}$$

represents the probability that a point is assigned to C_k in clustering C and to $C'_{k'}$ in C'. Now we denote the mutual information between the associated random variables (P_k, $P'_{k'}$) to be equal to the mutual information $I(C, C')$ between the investigated clusterings C and C':

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P'(k')} \tag{2.2.19}$$

To better understand $I(C, C')$ we can think of it the following way: Having in mind an observation in D we can measure the uncertainty it has about its cluster in C' with $H(C')$. If we then know to which cluster it belongs in C the uncertainty about C' is reduced, namely by $I(C, C')$. So $I(C, C')$ is equal to the reduction of uncertainty averaged over all observations n. Its properties are that it is always non-negative and symmetric ($I(C, C') = I(C', C) \geq 0$). Furthermore, it can never exceed the total uncertainty in a clustering ($I(C, C') \leq \min(H(C), H(C'))$), where equality occurs for identical clusterings C and C'.

Variation of Information is proposed as the quantity

$$\begin{aligned} VI(C, C') &= H(C) + H(C') - 2I(C, C') \\ &= |H(C) - I(C, C')| + |H(C') - I(C', C)| \end{aligned} \tag{2.2.20}$$

If the two terms in (2.2.20) are expressed as conditional entropies, the equation reduces to

$$VI(C, C') = H(C|C') + H(C'|C) \tag{2.2.21}$$

For better understanding and imagination, equation (2.2.21) is visualised in figure (2.2). The value of VI does not directly depend on n but on the relative proportions of the clusters. However, the upper bound of VI does depend on n ($VI(C, C') \leq \log n$). Therefore VI ranges from $[0, \log n]$ where 0 indicates identical clusters. It is easy to normalise by $\log n$ to get a range between 0 and 1:

$$nVI(C, C') = \frac{1}{\log n} VI(C, C') \tag{2.2.22}$$

This can be helpful if classifications on the same data set are compared (or two different

data sets have the same number of observations n) because the measure is then easier to interpret. Normalising is not recommended if distances obtained by different data sets should be analysed because then VI loses a valuable property: comparability across many different conditions.

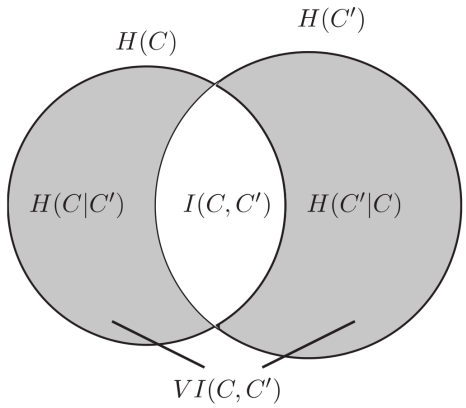


Figure 2.2: Variation of Information (sum of coloured areas) and related quantities (Meila, 2007)

2.3 Probabilistic weather forecasting

2.3.1 Introduction

In the past two decades, the fundamental insight that the atmosphere is a chaotic system and therefore the deterministic weather prediction has intrinsic limitations led to the development of ensemble prediction systems (Wernli, 2012). Ensemble prediction systems provide an approach to account for uncertainties in both initial conditions and model formulation that also depend on the flow (Ehrendorfer, 1997; Palmer, 2000).

Further details follow Wilks (2006). Initial uncertainty is due to the fact that the atmosphere cannot be completely observed in terms of spatial coverage and measurement accuracy. Model uncertainty emerges from parameterisations and numerical approximations of the governing equations. Knowing this, one sees that one conventional deterministic forecast that describes the evolution of a single initial state, which is assumed to be the true initial state, is not capable to account for this given uncertainty. The idea of probability forecasts is to let the deterministic governing equations operate on the probability distribution which describes the initial uncertainty. This distribution then evolves in phase space with time and describes a forecast probability distribution. A phase space is a geometrical illustration of all hypothetically possible states of the underlying dynamical system. The evolution of the probability distribution can be described by introducing a continuity equation for probability (Ehrendorfer, 1994). As a consequence of the high dimensionality of the phase space of an atmospheric model this equation needs to be simplified, which is possible with ensemble forecasting.

The ensemble forecast process starts by generating a finite sample of perturbed initial states from the initial probability distribution. The generation of these perturbations is non-trivial, as only linear combinations of the fastest governing perturbations lead to a realistic ensemble spread (Buizza, 1997; Ehrendorfer, 1997). This ensemble of initial conditions should represent a plausible representation of the initial state uncertainty. Each member of this ensemble is used as initial condition for a separate model run. With the help of these forecasts it is possible to approximate what the future probability distribution will look like after the initial distribution has been transformed by the physical laws of the model with time. Figure (2.3) illustrates this process. It visualises how the trajectories of the single runs evolve in time. At the intermediate stage of the forecast all members show a similar state of the atmosphere, whereas two different solutions, indicated through the accumulation of two groups of points that are

close to each other, are seen in the final forecast projection. We notice that the result of the single best deterministic run (heavy line) would be the upper solution but the ensemble forecast shows that there is a second probable solution indicated by six of nine ensemble members. This is very important information for forecasters that would not be available with the purely deterministic approach.

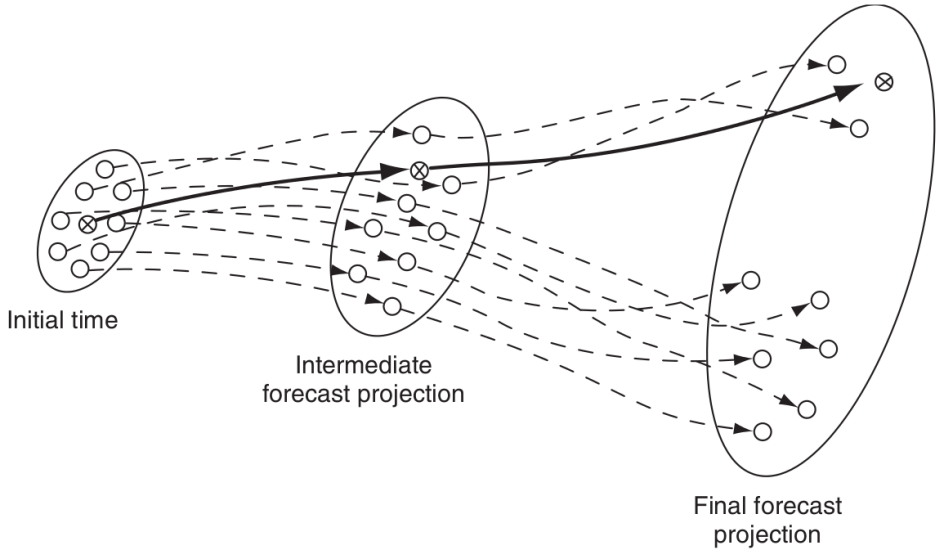


Figure 2.3: Illustration of a schematic ensemble forecast in an idealised two-dimensional phase space by Wilks (2006). The heavy line represents the single best deterministic forecast and dashed lines correspond to the individual ensemble members. Ellipses at each state represent the phase space. At the intermediate state all members are still close together whereas in the final forecast projection two possible solutions have developed.

2.3.2 Ensemble spread

As mentioned by Grimit and Mass (2007) the raw output of a well-designed ensemble forecast already provides useful information by analysing the relative frequencies of the ensemble members. To gain more information, statistical postprocessing of the output is required.

According to Wilks (2006) one simple postprocessing step is to average over the members of the ensemble to obtain a single forecast. This corresponds to the centre of the ensemble and therefore approximates the centre of the probability distribution. The ensemble mean is, in general, more skillful than a single deterministic forecast as elements of disagreement are averaged out and similar characteristics shared by the big-

ger part of the ensemble are emphasised. This is true for the beginning of the forecast but as forecast time proceeds, a regime change or a change in the long-wave pattern may occur (Palmer, 1993). The resulting problem can be observed in figure (2.3). If the ensemble members are averaged, the ensemble mean would indicate a very improbable state in between the two main solutions. Hence, an additional important aspect of ensemble forecasting is to characterise the magnitude and nature of the uncertainty in a forecast that is state-dependent (i.e. forecast uncertainty is different for different occasions and weather regimes).

The upcoming paragraphs follow the presentation of Grimit and Mass (2007) unless stated otherwise. One measure to quantify the forecast uncertainty is the ensemble spread (ensemble dispersion). Ensemble spread is an ambiguous term because there exist several definitions which are summarised in table (2.1).

Spread metric	Acronym	Mathematical expression	Type
Std dev	SD	$\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(x_m-\bar{x})^2}$	Continuous
Variance	VAR	$\frac{1}{M-1}\sum_{m=1}^{M}(x_m-\bar{x})^2$	Continuous
Mode frequency	MOD	$\frac{1}{M}\max(M_i)$	Categorical
Statistical entropy	ENT	$-\sum_{i=1}^{\mathrm{nbin}}\frac{M_i}{M}\log\left(\frac{M_i}{M}\right)$	Categorical

Table 2.1: Several definitions of ensemble spread summarised by Grimit and Mass (2007). Ensemble size is given by M, x_m corresponds to forecast ensemble member m and $\bar{x}$ indicates the ensemble mean. *nbin* stands for the number of categories defined from climatology and M_i refers to the population of forecast ensemble members within a climatological bin i.

These measures can be categorised according to their type into continuous and categorical groups. The latter group is for end users that have a categorical verification

strategy. Here, the ensemble spread is measured by the degree to which the members are distributed among climatological bins. Certainly, there are end users, for example utility companies, who need power demand projections, that define their cost functions in a continuous way. The used measures for the continuous type, standard deviation (SD) and variance (VAR), can be calculated with different reference points. One can use the ensemble control forecast (i.e. operational run with lower resolution) as mentioned by Buizza and Palmer (1998). Another possibility is to use the ensemble mean

$$\bar{x} = \frac{1}{M} \sum_{m=1}^{M} x_m \tag{2.3.1}$$

as it is defined in table (2.1).

2.3.3 ECMWF operational ensemble forecasts

This section follows the presentation of Barkmeijer et al. (2012) unless denoted otherwise. In November 1992 the first ensemble forecasts were produced at the European Center for Medium-range Weather Forecasts (ECMWF). For the last twenty years the system has undergone many developments that improved forecast quality. At the beginning, the ensemble system was based on the idea of Farrell (1982), who suggested to use singular vectors to efficiently represent the initial state uncertainty. Further developments involved the representation of model uncertainty by using stochastic physics tendencies, increase of resolution and several model upgrades. To better account for the initial uncertainty, singular vectors are combined with perturbations derived from the Ensemble of Data Assimilations (EDA) in recent years.

Using the new terminology for the ECMWF products introduced by the ECMWF Directorate (2012), the ECMWF operational ensemble forecasts consist of a high-resolution forecast (HRES) and an ensemble of lower-resolution forecasts (ENS) which are provided by the ECMWF forecasting system called Integrated Forecasting System (IFS). Former terminologies were operational or deterministic run for HRES and Ensemble Prediction System (EPS) for ENS. The current configuration of the ECMWF operational ensemble forecasts has the following properties.

ENS consists of 51 members, namely one control forecast (HRES at ensemble resolution) and 50 forecasts with initial perturbations. Twice a day at 00 and 12 UTC

a run is started with a 15 day forecast range. Additionally, twice a week on Monday and Thursday 00 UTC, the forecast range is extended to 32 days. To describe the atmosphere, on a spatial linear grid, a horizontal resolution of a spectral triangular truncation T639 (around 32 km) up forecast day 10 is used. Afterwards, the resolution is decreased to T319 (around 65 km). The vertical resolution is given by 62 levels up to 5 hPa. The HRES forecast, which is suggested to be used together with ECMWF ENS, has a horizontal resolution of T1279, the model top is at 0.01 hPa and the vertical is divided into 91 levels.

In this forecast system the atmospheric model is combined with information about the ocean. ENS is coupled to the WAM wave model with 55 km resolution. It accounts for 24 directions and 30 frequencies up to forecast day 10, afterwards for 12 directions and 25 frequencies. Ocean currents are modelled on the ORCA100z42 grid which corresponds to 1-degree horizontal resolution and 42 vertical levels. Up to day 10, ENS runs with constant sea surface temperature anomalies. For the rest of the forecast time the system is coupled to the NEMO ocean model. Initial ocean current uncertainties are simulated by making use of the NEMOVAR ensemble of ocean analysis.

ENS is constructed to account for both initial atmospheric and model uncertainties. The basis for the initial uncertainties is the HRES-4DVAR analysis with a horizontal resolution of T1279 (around 16 km) and 91 layers. Initial uncertainty is modelled by adding two sets of perturbations. On the one hand, 50 singular vectors are determined over different areas, namely the northern and southern hemisphere and the tropics. Those 50 vectors whose total-energy grows fastest over 48 hours are chosen. On the other hand, an additional perturbation set is computed from 6 hour forecasts started from 11 perturbed members of EDA with a resolution of T399. This is due to the fact that singular vectors have two major weaknesses: they are marginally sensitive to observation characteristics (general error, coverage and representativeness) and they poorly sample some areas of the world, like the tropics.

To account for model uncertainty, which is done only for the atmosphere, two stochastic schemes are applied. One is the stochastically parametrised tendency scheme (SPPT) which simulates random model errors due to parameterised physical processes. The other is the stochastic back-scatter scheme (SKEB) designed for modelling the upscale energy transfer. This energy originates from unresolved scales and can act as an energy source for developing perturbations on resolved scales (Shutts, 2005).

3 Dispersion model and input data

3.1 FLEXPART

The presentation of FLEXPART follows Stohl et al. (2010) unless denoted otherwise. FLEXPART is a Lagrangian particle dispersion model. Its first release was in 1998 and it was designed to simulate the long-range dispersion of air pollutants like after a nuclear power plant accident. Over the years it has evolved into a comprehensive model for atmospheric transport simulations. Other examples for applications besides nuclear accidents are dispersion simulations of volcanic ash and emissions of forest fires. FLEXPART is used for both research and operational purposes by many institutes.

The main advantage of the Lagrangian approach compared to the Eulerian one is that no numerical diffusion occurs. Furthermore, tracers released from a source are not instantaneously mixed into a grid cell as Lagrangian models are independent of a computational grid. Of course, this approach is not free of inaccuracy, for example, interpolation errors occur when particle positions are interpolated to an output grid (Arnold, 2013). The computational particles that are used for the simulation can be thought of as infinitesimally small air parcels. One can assign properties of real particles to them according to the substance which is under investigation. A source can be defined as point, line, area or volume source.

FLEXPART simulates long-range and mesoscale transport, diffusion, dry and wet deposition, convection and radioactive decay of tracers. The model is self-adjoint which means that it can be run forward and backward in time (Arnold, 2013). Therefore, it is not only possible to simulate the dispersion of particles from a source but also to estimate the source from given measurements (e.g. station measurements). The following summary of Arnold (2013) describes the function of FLEXPART in a simple way.

FLEXPART in a nutshell

1. track particles in a given velocity field
2. model sub-grid scale processes (boundary layer turbulence, mesoscale turbulence and cumulus convection) that effect the particle dispersion
3. modify particle properties according to locally acting processes, e.g. radioactive decay

4. count particles in a volume and calculate the concentration value

FLEXPART is an offline model; that means one has to provide meteorological input data to run the model. It has no internal preprocessor but there are several adapted versions for data of important meteorological forecast models such as the ECMWF IFS, GFS and WRF. As ECMWF data is used for this thesis, two main preprocessing steps for FLEXPART should be mentioned:

i Since the fluxes provided by ECMWF are accumulated and because FLEXPART needs de-accumulated fluxes, a de-accumulation process must be applied. A method is described in (3.3.2).

ii Due to the fact that parameterised random velocities for the boundary layer of the atmosphere are treated in [m/s] within the dispersion model, all three-dimensional variables of the meteorological input data from ECMWF, that are on model levels, are transformed to terrain-following Cartesian coordinates $\tilde{z} = z - z_t$. Here z_t is the height of the topography. To convert the vertical wind speeds from the eta coordinate system to terrain-following coordinates the formula

$$\tilde{w} = \dot{\tilde{z}} = \dot{\tilde{\eta}} \left(\frac{\partial p}{\partial \eta} \right)^{-1} + \left. \frac{\partial \tilde{z}}{\partial t} \right|_{\eta} + \mathbf{v}_h \cdot \nabla_\eta \tilde{z} \tag{3.1.1}$$

is used, where $\dot{\tilde{\eta}} = \dot{\eta} \, (\partial p / \partial \eta)$. The second term on the right hand side is not implemented in the transformation because it has been shown that it is negligible compared to the other terms. As $\dot{\eta}$ is not available for the ECMWF ENS product, an alternative way has to be found to derive $\dot{\eta} \, (\partial p / \partial \eta)$ which is described in (3.3.3).

Since the output of FLEXPART usually contains many grid cells with zero concentration, a special output format was designed to compress file size. To provide assistance in reading the non-standard output format and plotting FLEXPART fields the Python routine *pflexible* was developed. This module is freely available and its main developer is John Burkhart (Burkhart, 2011).

3.2 Meteorological input data

For the thesis, two global ECMWF operational ensemble forecast data sets (2.3.3) were retrieved from the ECMWF Meteorological Archival and Retrieval System (MARS) and provided by ZAMG. One part of the ensemble is the HRES forecast which gets already preprocessed for FLEXPART at ZAMG and is given on a 1° LL with 91 vertical levels. The other part is the ensemble of 50 perturbed forecasts which was provided in two different ways. As input for the dispersion simulations the full ensemble was retrieved (all 62 model levels, surface fields and humidity on a N160 reduced Gaussian grid (RGG) and vorticity, divergence and temperature in SH with a triangular truncation of T319). Table (3.1) lists the characteristics of the two meteorological data sets. For the clustering part, the meteorological parameters on model levels were interpolated from the dispersion simulation input ensemble from T319 to 1° LL. The data on pressure levels was downloaded on a LL grid with 0.5° resolution only at desired heights and interpolated to the same output grid as configured for FLEXPART output (1° LL).

A problem that occurred is linked to the de-accumulation scheme described in (3.3.2). The scheme averages the fluxes over five time steps and this value is valid for the central time step. As a consequence, the scheme cannot provide de-accumulated fluxes for the first two time steps of the forecast. For the first data set this was solved by setting the fluxes for these two times equal to the HRES fluxes for all 50 ensemble members. A better way to get the initial fluxes is to additionally retrieve the t+21 h forecast for the six fluxes, initialised at 12 UTC the day before, de-accumulate these fluxes and use the two that correspond to the missing time steps of the main forecast. This was done for the second data set.

The used ensemble consists only of 50 members because the control forecast was not used. This is due to the fact that this member is equal to the HRES run but with lower resolution. The ensemble is centred around the HRES run and as a result, the variance of the control forecast is very low compared to the ensemble mean. Furthermore, the dispersion run for the HRES input is always simulated in the implemented process and the control forecast would not yield much more information.

Data set 1

- Clustering input
  - Data: ECMWF ENS
  - Date: 14.03.2013 00 UTC +72 h
  - Temporal resolution: 3 h
  - Area: global
  - Surface parameter: precipitation
  - 3D parameters: u/v wind speed
  - Members: 50
  - Resolution: 1° LL
  - Model levels: 31, 27, 23

- Model input
  - Data: ECMWF HRES (ZAMG)
  - Date: 14.03.2013 00 UTC t+78 h
  - Temporal resolution: 3 h
  - Area: global
  - Parameters: all
  - Resolution: 1° LL
  - Model level: all 91

  - Data: ECMWF ENS
  - Date: 14.03.2013 00 UTC t+78 h
  - Temporal resolution: 3 h
  - Area: global
  - Members: 50
  - Surface parameters: all
  - 3D parameters: voriticity, divergence, temperature, humidity
  - Resolution: RGG N160, SH T319
  - Model level: all 62

Data set 2

- Clustering input
  - Data: ECMWF ENS
  - Date: 21.03.2013 00 UTC t+72 h
  - Temporal resolution: 3 h
  - Area: global
  - Surface parameter: precipitation
  - 3D parameters: u/v wind speed
  - Members: 50
  - Resolution: 1° LL
  - Model levels: 31, 27, 23
  - Pressure levels: 400, 250

- Model input
  - Data: ECMWF HRES (ZAMG)
  - Date: 21.03.2013 00 UTC t+78 h
  - Temporal resolution: 3 h
  - Area: global
  - Parameters: all
  - Resolution: 1° LL
  - Model level: all 91

  - Data: ECMWF ENS
  - Date: 21.03.2013 00 UTC t+78 h
  - Temporal resolution: 3 h
  - Area: global
  - Members: 50
  - Surface parameters: all
  - 3D parameters: voriticity, divergence, temperature, humidity
  - Resolution: RGG N160, SH T319
  - Model level: all 62

  - Data: ECMWF ENS
  - Date: 20.03.2013 12 UTC t+21 h
  - Temporal resolution: 3 h
  - Area: global
  - Members: 50
  - Surface parameters: all 6 fluxes
  - Resolution: 1° LL
  - Model level: Surface

Table 3.1: Characteristics of the provided meteorological data sets

3.3 Preprocessing for FLEXPART

3.3.1 Interpolation - grid, resolution and horizontal wind field

Staring from the raw data sets described in table (3.1) all needed input fields for FLEXPART can be derived. An essential tool to handle ECMWF fields is the Fortran library EMOSLIB provided by ECMWF. The package is used by ECMWF MARS and is also available for users to preprocess data to their needs. Its features are shown in figure (3.1).

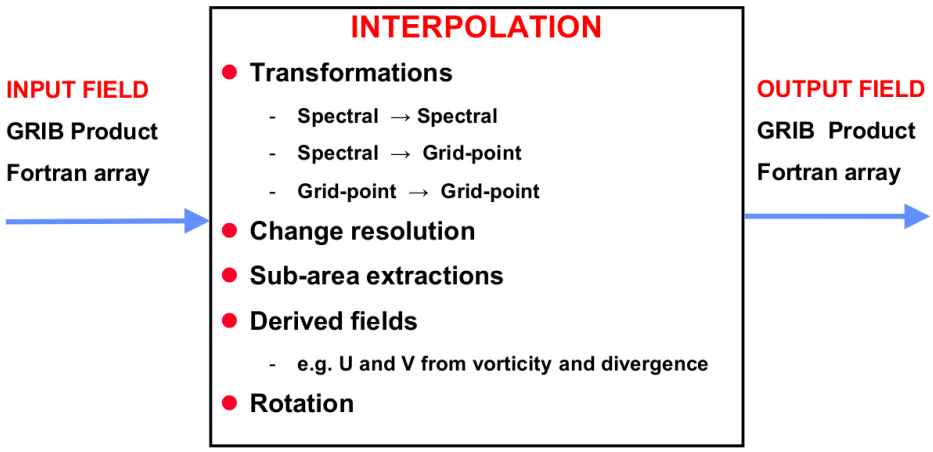


Figure 3.1: Features of the ECMWF Fortran library EMOSLIB (Dando, 2013)

As a first step, all fields are interpolated to a 1° LL grid. From vorticity and divergence the horizontal wind field is derived and interpolated to a 1° LL grid for the FLEXPART input and to have the fluxes available on this grid for deaccumulation (3.3.2). Additionally, the horizontal wind field gets interpolated to SH T319 for calculating the verical velocity $\dot{\eta}\,(\partial p/\partial \eta)$ (3.3.3). This is one of the reasons why the ensemble data for the meteorological input (3.1) wasn't retrieved already on a LL grid.

3.3.2 De-accumulation of surface fluxes

Fluxes are given as accumulated fields in ECMWF data. FLEXPART needs de-accumulated fluxes as part of the meteorological input data. Therefore, the six fluxes, large scale precipitation (LSP), convective precipitation (CP), surface sensible heat flux (SSHF), surface solar radiation (SSR), north-south surface stress (NSSS) and east-west surface stress (EWSS) must be de-accumulated. This is done by using the interpolation methods documented by James (2000). He described two different interpolations which

are presented in the next paragraphs, one for the rainfall fluxes and one for the other four fluxes.

Converting units

Before the de-accumulation methods are applied, the accumulated fluxes have to be converted into equivalent rates. This is done according to table (3.2).

Quantity	Acronym	$\text{Unit}_{\text{ECMWF}}$	Factor	$\text{Unit}_{\text{FLEXPART}}$
large scale precipitation	LSP	m	1000/3	mm/h
convective precipitation	CP	m	1000/3	mm/h
surface sensible heat flux	SSHF	J/m^2	1/(3600·3)	W/m^2
surface solar radiation	SSR	J/m^2	1/(3600·3)	W/m^2
north-south surface stress	NSSS	N/m^2s	1/(3600·3)	N/m^2
east-west surface stress	EWSS	N/m^2s	1/(3600·3)	N/m^2

Table 3.2: Conversion of accumulated fluxes into equivalent rates

De-accumulation method for rainfall

This method can be applied to the rainfall fluxes LSP and CP. To explain the method we denote four consecutive accumulated rainfall values, A, B, C, D with unit [mm/h]. These values have been accumulated between the time blocks 0-1, 1-2, 2-3, 3-4, respectively. We try to estimate the value of the de-accumulated variable S at the center of these four time ranges which corresponds to time step 2.

The characteristics of rainfall are that it is a peaky quantity and it cannot be negative. For this reason a polynomial solution is not useful. Further, it would be desirable to account for situations, for example, if B=5 mm/h and C=0 mm/h. Here, it is impossible that $S > 0$ in reality because rainfall must have ended before t=2 (or exactly at t=2 which is very unlikely). So we need a method that conserves total rainfall, represents the peaky nature and reduces rainfall in the mentioned example as close as possible to zero.

To accomplish these properties he suggests an adjusted version of the simple linear interpolation

$$S = (B + C)/2 \tag{3.3.1}$$

The method changes fractions of each accumulated total by looking forward and backward in time to see how the totals are currently changing relative to each other. By looking at the ratios A or C to $(A + C)$ we can divide B into B_b, by going one time step backwards, and B_f by going one time step forward. These fractions are defined by

$$B_b = B \cdot C/(A + C) \text{ and } B_f = B \cdot A/(A + C) \tag{3.3.2}$$

We can also calculate the fractions for C to get C_b and C_f. The accumulated value for S would then be

$$\begin{aligned} S &= B_f + C_b \\ &= B \cdot C/(A + C) + B \cdot C/(B + C) \\ &= B \cdot C \cdot (A + B + C + D)/((A + C) \cdot (B + D)) \end{aligned} \tag{3.3.3}$$

If the denominator is zero, the fraction is assumed to be $0.5B$. Further, S is constrained such that $S \geq 0$ and $S \leq (B + C)$.

De-accumulation method for other fluxes

For the purpose of de-accumulating SSHF, SSR, NSSS and EWSS a cubic-polynomial interpolation method is presented. The full unconstrained polynomial solution can only be applied to SSHF. Since the other three fluxes are positive quantities, the value given by the polynomial is set to zero if slightly negative values occur.

As for the rainfall interpolation we denote four consecutive accumulated values A, B, C, D. These values are accumulated over the time blocks 0-1, 1-2, 2-3, 3-4, respectively, and already have the unit [W/m^2] or [N/m^2]. We again try to estimate the de-accumulated variable S for the central time step at t=2. This method can be applied to every regular time interval, so the temporal resolution of the data set does not matter.

The integrals of the polynomial function $f(t)$ must be identical to the respective original accumulated values, i.e.

$$\int_{t=0}^{1} f(t)dt = A \ , \ \int_{t=1}^{2} f(t)dt = B \ , \ \int_{t=2}^{3} f(t)dt = C \ , \ \int_{t=3}^{4} f(t)dt = D \tag{3.3.4}$$

These four constraints require a cubic polynomial. A unique ideal cubic-polynomial solution is given by

$$f(t) = a \cdot t^3 + b \cdot t^2 + c \cdot t + d \tag{3.3.5}$$

and therefore the accumulated value S at time step t=2 can be computed by

$$S = 8a + 4b + 2c + d \tag{3.3.6}$$

where

$$\begin{aligned} a &= (D - A + 3(B - C))/6 \\ b &= (C + A)/2 - B - 9a/2 \\ c &= B - A - 7a/2 - 2b \\ d &= A - a/4 - b/3 - c/2 \end{aligned}$$

For each de-accumulation step the values A-D are set to the next four values by moving one time step forward and then a new polynomial is calculated.

3.3.3 Calculating the vertical velocity $\dot{\eta}\,(\partial p/\partial \eta)$

As mentioned in (3.1), FLEXPART needs the vertical velocity $\dot{\eta}\,(\partial p/\partial \eta)$ on $half\ levels$. The parameter $\dot{\eta}$ that is needed for this calculation is provided for the ECMWF HRES forecast but not for the ENS product. A method to derive $\dot{\eta}\,(\partial p/\partial \eta)$ out of available parameters is described by Haimberger (2009) and the following paragraphs follow his presentation.

The method is based on the the continuity equation in a pressure based hybrid coordinate system

$$\frac{\partial}{\partial t}\left(\frac{\partial p}{\partial \eta}\right)+\nabla\cdot\left(\mathbf{v}\frac{\partial p}{\partial \eta}\right)+\frac{\partial}{\partial \eta}\left(\dot{\eta}\frac{\partial p}{\partial \eta}\right)=0 \tag{3.3.7}$$

as used by ECMWF. By rearranging terms and integrating over η we yield explicit expressions for the rate of change of surface pressure p_s (with boundary conditions $\dot{\eta}=0$ at $\eta=0$ and $\eta=1$) and for $\dot{\eta}$:

$$\frac{\partial p_s}{\partial t}=-\int_0^1 \nabla\cdot\left(\mathbf{v}\frac{\partial p}{\partial \eta}\right)d\eta \tag{3.3.8}$$

$$\dot{\eta}\left(\frac{\partial p}{\partial \eta}\right)=-\frac{\partial p}{\partial t}-\int_0^\eta \nabla\cdot\left(\mathbf{v}\frac{\partial p}{\partial \eta}\right)d\eta \tag{3.3.9}$$

For the distribution of the levels we define that prognostic variables, such as the horizontal wind field are on $full\ levels\ k=1,...,NLEV$. Values for the vertical coordinate, and thus pressure, are defined at intermediate levels, $\eta_{k+\frac{1}{2}}$ and $p_{k+\frac{1}{2}}$, respectively. In its discretised form, after Simmons and Burridge (1981), surface pressure tendency is given by

$$\frac{\partial p_s}{\partial t}=-\sum_{k=1}^{NLEV}\left[\Delta p_k\nabla\cdot\mathbf{v}_k+\left(\mathbf{v}_k\cdot\nabla p_s\right)\Delta B_k\right] \tag{3.3.10}$$

and the desired vertical velocity on $half\ levels$ is written as

$$\left(\dot{\eta}\frac{\partial p}{\partial \eta}\right)_{k+\frac{1}{2}}=-B_{k+\frac{1}{2}}\frac{\partial p_s}{\partial t}+\sum_{j=1}^{k}\left[\Delta p_j\nabla\cdot\mathbf{v}_j+\left(\mathbf{v}_j\cdot\nabla p_s\right)\Delta B_j\right] \tag{3.3.11}$$

where $\Delta p_k=p_{k+\frac{1}{2}}-p_{k-\frac{1}{2}}$ is the pressure difference and $\Delta B_k=B_{k+\frac{1}{2}}-B_{k-\frac{1}{2}}$ is one of two parameters which specify the hybrid vertical coordinates.

For the case $\dot{\eta}$ is given on $full\ levels$ ($\dot{\eta}_k$), as it is in the ECMWF HRES forecast, $\dot{\eta}\,(\partial p/\partial\eta)$ on $half\ levels$ can be calculated by using

$$\left(\dot{\eta}\frac{\partial p}{\partial \eta}\right)_{k+\frac{1}{2}} = 2\dot{\eta}_k \left(\frac{\partial p}{\partial \eta}\right)_k - \left(\dot{\eta}\frac{\partial p}{\partial \eta}\right)_{k-\frac{1}{2}} \tag{3.3.12}$$

where the factor $(\partial p/\partial \eta)_k$ is given by

$$\left(\frac{\partial p}{\partial \eta}\right)_k = p_s \frac{\Delta A_k/p_s + \Delta B_k}{\Delta A_k/p_0 + \Delta B_k} \tag{3.3.13}$$

The coefficients $\Delta A_k = A_{k+\frac{1}{2}} - A_{k-\frac{1}{2}}$ and $\Delta B_k = B_{k+\frac{1}{2}} - B_{k-\frac{1}{2}}$ describe the hybrid vertical coordinate system and p_s=1013.25 hPa is the reference pressure.

To calculate $\dot{\eta}\,(\partial p/\partial \eta)$, one needs the horizontal wind field $\mathbf{v}$, divergence and $\log p_s$ given in SH as input. If $\dot{\eta}\,(\partial p/\partial \eta)$ has to be calculated for the HRES forecast, one needs to provide $\dot{\eta}$ and $\log p_s$. Table (3.3) summarises the steps performed to yield $\dot{\eta}\,(\partial p/\partial \eta)$ on a LL grid.

	SH	RGG	SH	LL
ENS	Provide input $\mathbf{v}$, $\nabla \cdot \mathbf{v}$ and $\log p_s$, calculate $\nabla \log p_s$	Compute $\dot{\eta}\,(\partial p/\partial \eta)$ by using (3.3.11)	Truncate the following fields to a spectral resolution which corresponds to the LL output grid: $\dot{\eta}\,(\partial p/\partial \eta)$, $\mathbf{v}$	Interpolate $\dot{\eta}\,(\partial p/\partial \eta)$ to a LL grid
HRES	Provide $\dot{\eta}$ and $\log p_s$			Compute $\dot{\eta}\,(\partial p/\partial \eta)$ by using $\dot{\eta}$ and equation (3.3.12)

Table 3.3: Overview of steps performed to yield $\dot{\eta}\,(\partial p/\partial \eta)$ (Haimberger, 2009)

In order to estimate the error which results through calculation and interpolation it is suggested to compare the root mean square error (RMSE) difference of given fields. Since $\dot{\eta}$ cannot be retrieved, ω is a good alternative. It can be downloaded for ENS data and it is also calculated during the procedure of getting $\dot{\eta}\,(\partial p/\partial \eta)$. An error lower than 2% is documented for several experiments if the input fields are provided at a resolution of T799 and the output grid is 0.2° LL. Unfortunately, this is only true for

lower levels because errors can increase up to 55% for individual levels. Furthermore, it has to be noted that accuracy strongly decreases when the resolution of the input data and output grid is reduced. This measure was applied to a test data set provided for the thesis and with the resolution of T319 for the input and 1° for the output grid an error of 31-44% was estimated for middle model levels where usually the largest errors occur.

4 Proposed concept and practical considerations

4.1 Proposed concept

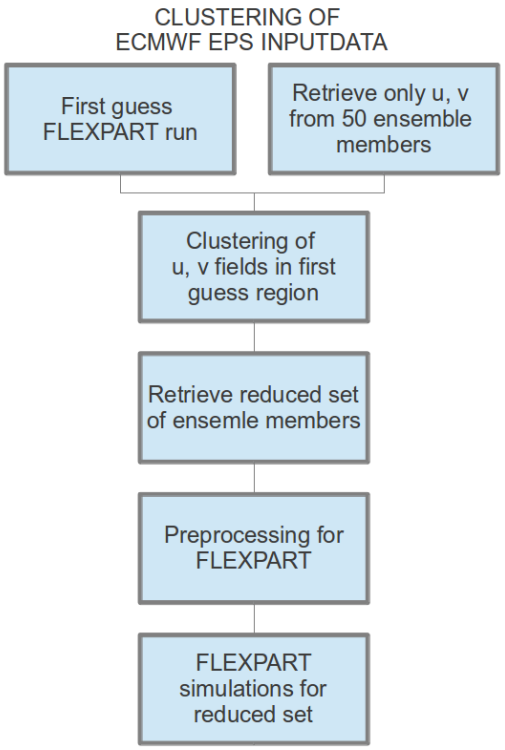
Concerning the meteorological input data it was clear from the beginning that ensemble members from the ECMWF ENS product will be used. However, it was not clear from the outset which meteorological parameters should be used for clustering or if it would be better to cluster concentration fields calculated by FLEXPART with different ensemble members using small particle samples. Therefore two fundamentally different clustering concepts have been implemented which are referred to as *proposed* (or *operational*) *concept* and $reference\ concept$ as outlined in figure (4.1).

Reference concept

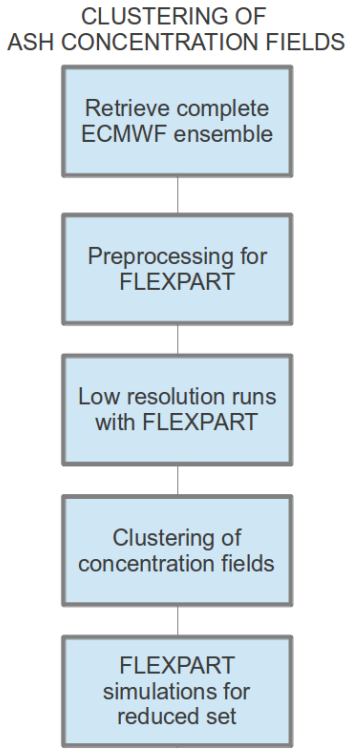
During a discussion, Prof. Petra Seibert suggested to make dispersion simulations for every member of the ECMWF ENS with reduced resolution and fewer computational particles and to cluster the concentration fields directly. This is a very good basis for finding representative members of the dispersion ensemble forecast since all parameters and processes that describe the transport would be included. It would have been the concept of choice if it had not exceeded the operational time constraints for retrieval, preprocessing and simulation. Additionally, it implicates a sampling problem which arises when too few particles are used for the dispersion simulation. Nevertheless, this concept is perfect as a reference for validating the cluster solutions found by clustering the meteorological data as described by the operational concept. Usually the validation should be carried out with real observations but since those are not available the reference concept is applied to case studies (Table 5.1) instead. A flow chart of this concept is shown in figure (4.1b). In contrast to the concept (4.1b), the resolution of both the meteorological input data and the FLEXPART simulations are kept at T319 and 1° LL, respectively, to have a better reference.

Operational concept

Figure (4.1a) shows the operational concept which has the capabilities to account for the operational requirement to gain quickly available forecasts with limited computational resources. It is based on clustering meteorological input parameters that have a



(a) Proposed concept: yields a quickly available ensemble forecast but less information is used for clustering



(b) Reference concept: forecasting time exceeds operational demands, would have a better information basis for clustering

Figure 4.1: Derived concepts for dispersion ensemble forecasts

strong linkage to the modelled concentration fields. Further the concept is designed to optimise the area for clustering the parameters regarding to the area that is affected by the ash plume. How the meteorological parameters are chosen and the clustering area is defined is described in sections (4.3.1) and (4.3.2), respectively. To get straight to the point, this approach is designed to daily retrieve the global u/v wind component of a specific height for the complete ensemble. In case of an event, a first guess run (i.e. one dispersion simulation) will be made with the ECMWF HRES forecast, which is already available and preprocessed for FLEXPART at ZAMG. Now it is possible to estimate the area of interest, cluster the meteorological parameters there and determine the new ensemble which consists of representative members that are found in every cluster. After this, the full forecast data for every member of the reduced ensemble will be retrieved and preprocessed for FLEXPART (3.3). As a final step, the dispersion simulations are carried out to gain the ensemble forecast for ash plume concentrations.

4.2 Model configuration

To simulate the dispersion of volcanic ash the dispersion model FLEXPART (3.1) is used. For configuration three important files must be edited to describe general settings, the output grid and the source. Examples of these configuration files are shown in (A) and the most relevant settings are mentioned below.

Important general settings are that FLEXPART is run in forward mode for 72 h, model output is averaged over 3 h and written every 3 h as concentration fields. The adaption of the computation time step to the Lagrangian time scales is switched off ($ctl < 0$) and a constant time step of one synchronisation time (900 s) is used (Stohl et al., 2010). With this option turbulence is not well described but for large scale application it works very well (Stohl et al., 1998). Furthermore, moist convection is switched on. For details see (A.1).

The output grid is defined by setting the lower left corner of the grid. From this point, the number of grid cells in longitudinal and latitudinal direction must be specified. The domain is adapted for every test scenario and for every case a resolution of 1° LL is used. To define the vertical resolution one has to specify layers by setting their upper boundaries. For this thesis, continuous layers of 500 m thickness between 5-14 km above sea level (a.s.l.) are assigned. For details see (A.2).

Volcanic ash is modelled with sulphate aerosol (SO_4-aero) as tracer. The location of the source is defined by the coordinates of the volcano and the tracer is released in a column over this point. To approximate the released mass and the number of particles to simulate the events, estimations for Eyjafjallajökull from ZAMG are used for a release height up to 14 km a.s.l. and a release time for 72 h. Within this release time, a total mass of $6.16 \cdot 10^{10}$ kg is released between 6-14 km a.s.l. and modelled with 1440000 particles. These values are used for each test scenario. Of course, this does not account for a real source but for this study it is enough to investigate the relative differences between the ensemble members. For details see (A.3).

4.3 Considerations about clustering

4.3.1 Choosing meteorological variables for clustering

The meteorological parameters should be chosen according to the problem which is given. Since this thesis is concerned with dispersion of volcanic ash which can affect air traffic we note the following details.

For dispersion the most important parameters are the three-dimensional wind field, which strongly influences the ash transport and also precipitation, which is responsible for wet deposition. The vertical velocity is not available on model levels for the ECMWF ENS product in [m/s]. It also is not advisable to use the calculated vertical velocity due to non-negligible errors which occur during the interpolation when a relatively coarse output grid with 1° is used (section 3.3.3). Therefore, only the horizontal wind field is utilised. The forecast should be valid for that level at which a plane stays the longest time. That is the cruising altitude which can lie roughly between 8-12 km. As precipitation only plays a major role below this layer it is not expected to contribute too much useful information. It was tried to combine the horizontal wind field and total precipitation (LSP + CP) in the clustering distance measure but this led to fuzzier clusters than without precipitation.

The relevant forecast time for the flight economy is about 24 h up to 72 h because it is most important to inform currently flying planes and to schedule the flight plans for the upcoming hours. Thus, 3 day horizontal wind forecasts were used with a temporal resolution of 3 h.

To determine a level for clustering, three levels of the cruising altitude mentioned above were examined individually, namely 8, 10, and 12 km. For assigning each of these levels a pressure level, and therefore a model level of the meteorological data, the assumption of a standard atmosphere was made. After Holton (2004) we can, by using the hydrostatic equation

$$\frac{dp}{dz} = -\rho g \tag{4.3.1}$$

derive the barometric formula

$$p(Z) = p(0) \exp^{-Z/H} \tag{4.3.2}$$

where $H = R\overline{T_v}/g_0$ is the scale height defined for an isothermal atmosphere. If we insert values for the specific gas constant for dry air $R = 287.058$ J/kgK, an assumed mean virtual temperature $\overline{T_v} = 288$ K and the global average of gravity at mean sea level $g_0 = 9.80665$ m/s^2 a scale height of $H = 8.430$ km is found. By using the three heights and a reference pressure $p_0 = 1013.25$ hPa for equation (4.3.2) we can determine the pressure levels. For the clustering input on pressure levels (see data set 2 (table 3.1)) only two levels are available that make sense to compare, namely 400 hPa and 250 hPa. To compare the clustering of the horizontal wind field with the clustering of a concentration field, a corresponding output layer of the FLEXPART grid has to be assigned. The defined FLEXPART output grid is described in section (4.2). Table (4.1) summarises the layers which are examined and the corresponding levels for the meteorological data.

cruising altitude [km]	pressure (4.3.2) [hPa]	ECMWF ENS model level [1]	ECMWF ENS pressure level [hPa]	FLEXPART output layer [km]
8	392.26	31	400	7.5 - 8
10	309.41	27	-	9.5 - 10
12	244.06	23	250	11.5 - 12

Table 4.1: Overview of investigated height levels and their corresponding model levels

4.3.2 Choosing the area for clustering

Concerning the domain for clustering, it first was planned to estimate the area by the range a particle would be transported under maximum wind speeds and use this measure to define a rectangular domain of interest centred over the volcano. During a discussion, the legitimate argument was brought up by Prof. Petra Seibert that this is too rough an assumption. In the beginning, the area with non-zero ash concentrations is very small. If we consider a volcanic eruption in Italy and a south-westerly wind direction in that region, airspace in eastern Europe would be affected. If further the domain of clustering would be the whole of Europe it is not improbable that patterns, which imply a big amount of uncertainty (e.g. a low pressure system), are located elsewhere in Europe. As a consequence, the reduced ensemble would be optimised to

represent the spread for all Europe, and not as desired, in the area of the ash plume. In other words, the area where the horizontal wind field gets clustered should correspond to the area of the ash plume as well as possible so that the distance measure for clustering acts in the same area. The idea is to make a first guess FLEXPART simulation with the ECMWF HRES forecast which is already available at ZAMG every day. This provides a first estimate of where the ash will be transported in the investigated level and allows to choose a rectangular domain that encloses the plume as shown in figure (4.2).

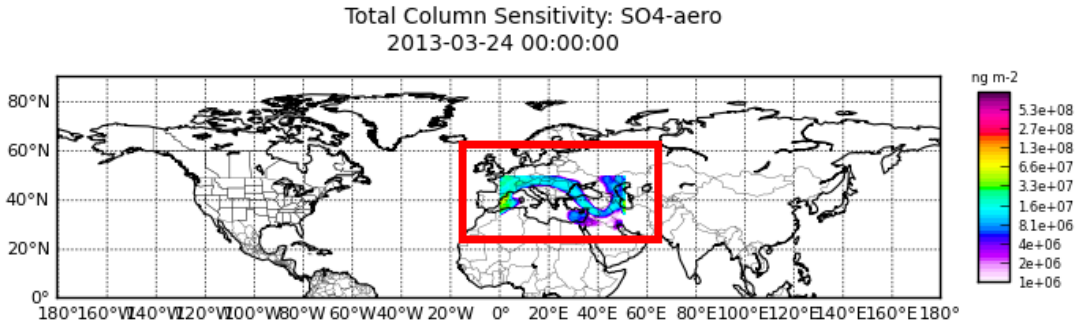


Figure 4.2: Determining the area of interest for clustering. Forecast for a theoretical eruption of the Italian volcano Etna after 72 h.

Moreover, one can check where grid cells have zero concentration values in this domain to further exclude parts of the area with unimportant information. One can do this by creating a mask according to the grid cells with non-zero values of the concentration field and applying it to the grid of the horizontal wind field when reading out the values for clustering. For the practical implementation this means it is checked whether a grid cell exceeds a certain threshold concentration and if it does, the four grid points that surround the cell are marked to be used for clustering. A threshold concentration makes sense in a way that parts of the plume which are already dissolving are ignored. This threshold was set to $C_{thresh} = 10$ ng/m^3 as a first approximation. The method to select grid points with valuable information is visualised in figure (4.3).

Of course this method would require the knowledge of all concentration forecasts of the ensemble since the involved area indicated by the HRES forecast will not be valid for the other ensemble members. The different perturbations of the meteorological input

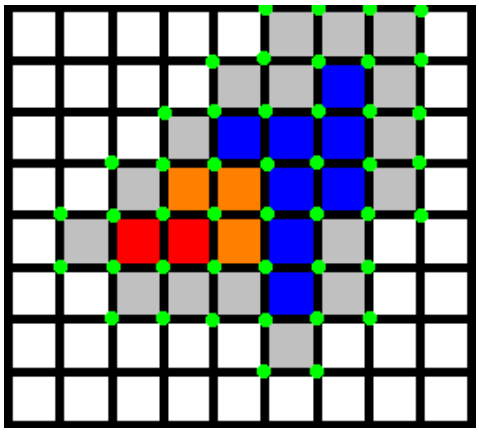


Figure 4.3: Selecting grid points (green dots) for clustering with the help of a concentration mask. Red, orange and blue grid cells represent different concentration values. Grey grid cells indicate the uncertainty area for the mask.

data will lead to different dispersions of the plumes. Therefore individual masks for every member would be needed. To overcome this problem an uncertainty area is added to the mask of the HRES forecast described above. The target is to widen the area in such a way that all grid cells affected by either one of the members of the dispersion forecasts lie within the extended mask. It is not trivial to estimate by how much the area should be enlarged without the knowledge of all members of the concentration forecasts. As a first approximation for the spread of the area the simulations of the three test scenarios were analysed. It was found that the spread starts to grow after the first 12 h and has a maximum variation of 5-8° on the LL grid. According to these results a linear function was defined which grows by 1° every 12 h after a forecast time of 12 h and is added to the borders of the plume. A plot of this function is shown in figure (4.4).

Additionally it must be noted that the area of an ash plume will increase and vary with forecast time. As a consequence, no static mask should be used. It is desirable that the mask describes the uncertainty area as well as possible for every time step. Therefore, for every time step a mask is built to account for the temporal variability. For better understanding, the clustering masks for two different forecast times are shown in figure (4.5).

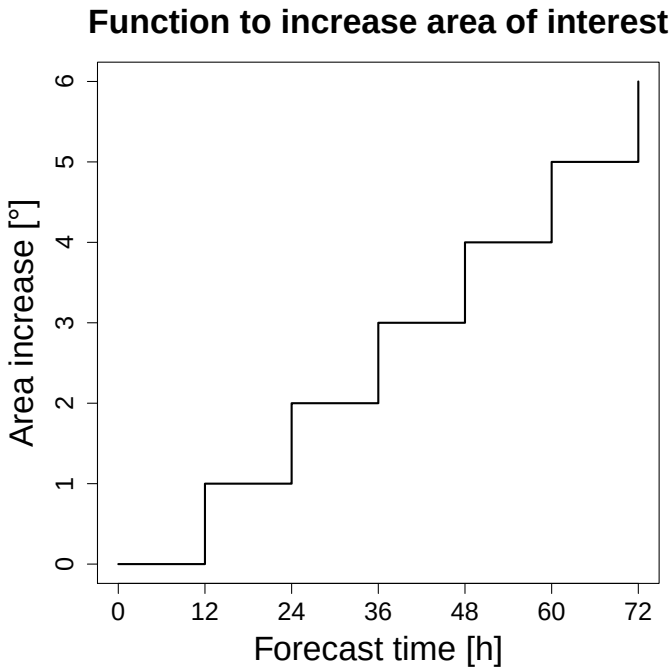
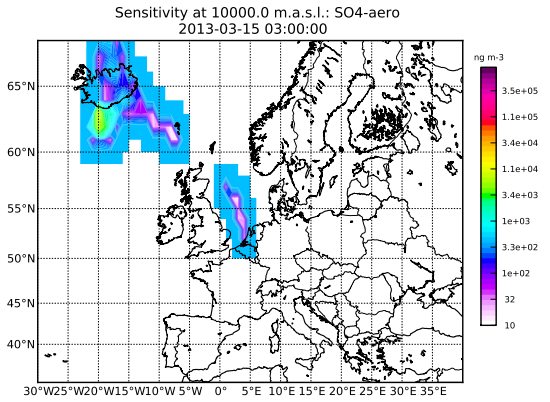
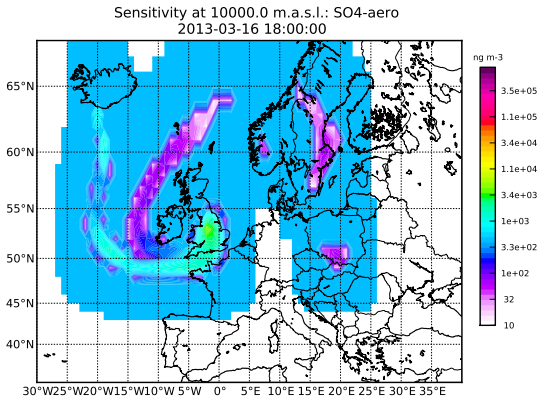


Figure 4.4: Function to increase the initial clustering area given by the first guess FLEXPART simulation to account for the uncertainty of the dispersion ensemble



(a) Mask for forecast time t+27 h. This corresponds to an area increase of 2° according to function (4.4)



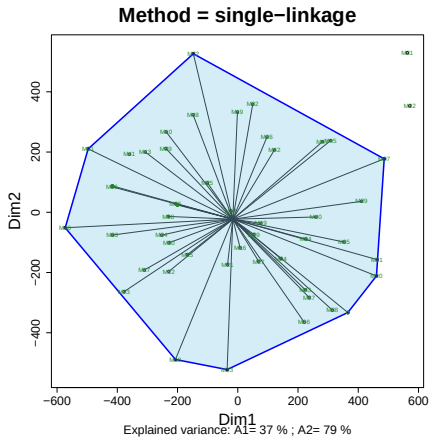
(b) Mask for forecast time t+66 h. This corresponds to an area increase of 5° according to function (4.4)

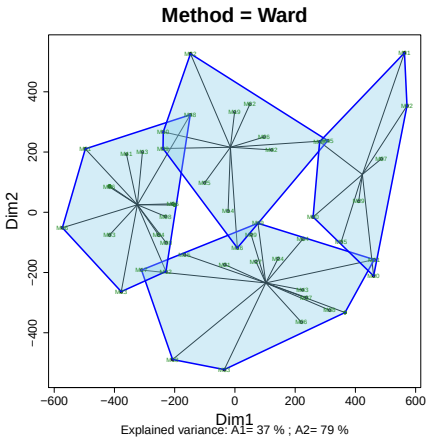
Figure 4.5: Examples for the addition of an uncertainty area (light blue) to the initial area given by the first guess run. The plume itself combined with the uncertainty area defines the area of interest for clustering.

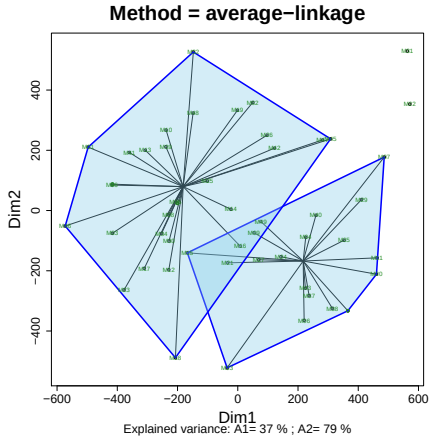
4.3.3 Configuration of the clustering setup

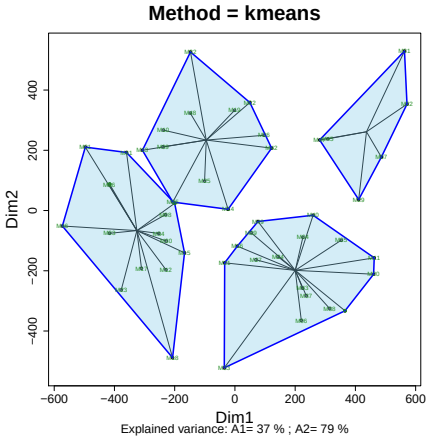
To configure a clustering setup, three basic questions have to be investigated, namely, the distance measure, the cluster algorithm and the number of clusters into which the data should be separated. For the decision which distance measure should be used, one has to look at the characteristics of the data and it must be determined which aspects should be pointed out. It seems reasonable to stick with the commonly used Euclidean distance for the horizontal wind field ensemble because it is compact and interval-scaled. If, for example, one wants to put emphasis on objects that are further apart, the squared Euclidean distance would be a good choice.

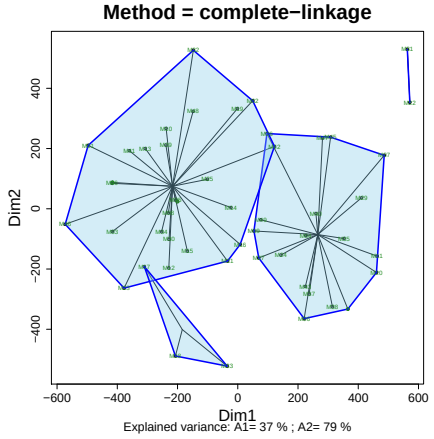
When CA is used for new applications it is often not known in advance which cluster algorithm fits best. Therefore, also in this study several algorithms were explored by applying them to example data sets. Five algorithms were picked from the hierarchical category (single-linkage, complete-linkage, average-linkage, centroid-linkage and Ward's method (2.1.2)) and two common methods from the partitioning category (K-means (2.1.3) and PAM (2.1.4)). The clustering results for one example data set (horizontal wind field at 10 km) can be seen in figure (4.6). As mentioned in (2.1.1), the linkage methods are used to define how the distance of a merged pair of the distance matrix is calculated. For single-linkage, the minimum distance of the two distances is used, for complete-linkage the maximum distance is used, for average-linkage the mean of the two distances is calculated and for centroid-linkage the Euclidean distance between the two merged clusters is calculated (Wilks, 2006). The characteristics of these three algorithms are described by Wilks (2006) as follows. Single-linkage is prone to chaining and therefore it produces few large clusters. Complete-linkage is the other extreme and yields more numerous groups as the criterion to fuse the clusters is more stringent. Average-linkage can be seen in the middle of the latter two. These properties are true for the examined example data set as can be seen in figure (4.6a). Centroid-linkage is not shown in figure (4.6) because the negative features mentioned by Kaufman and Rousseeuw (2005) were found for the example data set and, as a consequence, it was excluded at the very beginning. In the clustering process of this method it can happen that the distance between a new fusion is smaller than a distance of a previous fusion which makes the results hard to interpret. In contrast to the other hierarchical methods, Ward's method was able to produce reasonably compact groups. This is in agreement with theory where it is commented that this method aims to find compact and equally

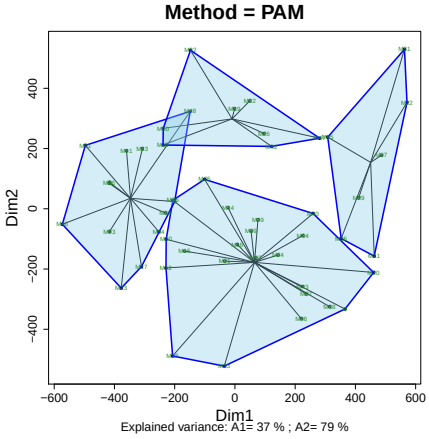












(a) algorithms which do not find a deeper structure in the data

(b) algorithms which are able to find reasonable groupings

Figure 4.6: Comparing properties of different cluster algorithms with multidimensional scaling (MDS). The investigated data set is a 72 h ensemble forecast of the masked (4.3.2) horizontal wind field at 10 km and it should be separated into four clusters by the algorithms. Clusters are indicated as light blue areas with a blue border and black lines point from each object to the centre of the cluster.

sized clusters (Mooi and Sarstedt, 2011). K-means (Morissette and Chartier, 2013) and PAM (Kaufman and Rousseeuw, 2005) yield similar groupings. As these clusters seem most appropriate to find representative members for a reduced ensemble, that is able to reproduce the ensemble spread of the full ensemble, the algorithms Ward's method, K-means and PAM are used for this study. The advantage over the other algorithms cannot be underlined with the Rand Index (2.2.3) and Variation of Information (2.2.4) but this is only due to the fact that these measures yield very good values for single-linkage and centroid-linkage as there is one cluster that contains almost every member and three clusters with only one member (see figure 4.7).

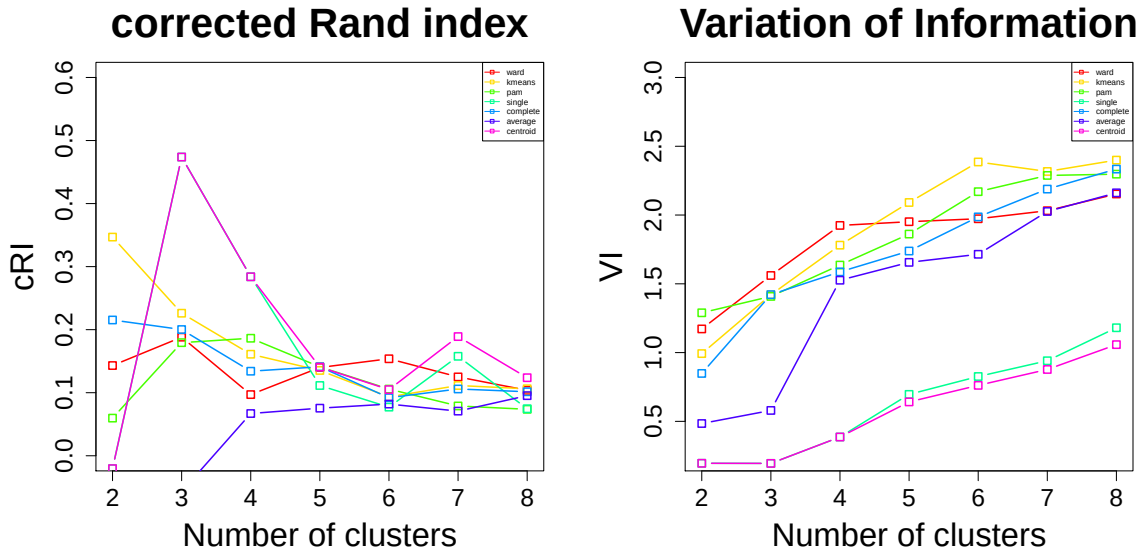


Figure 4.7: Comparison of all seven investigated clustering algorithms concerning the corrected Rand Index (RI) and Variation of Information (VI). The clustering is applied to the Eyjafjallajökull 2013-03-14 test scenario at 10 km (horizontal wind field) / 9.5-10 km (concentrations). Average-linkage and especially single-linkage and centroid-linkage yield good values because their solution contains one big cluster and many groups with only one member but this behaviour is not desirable. Due to very bad agreement, the average-linkage algorithm gets negative RI values for low numbers of clusters.

The last question mentioned at the beginning of this section is concerned with the number of clusters which should be used. As described in (2.1.1), one can analyse the evolving distance between groups during the clustering process for hierarchical methods or use the optimum average silhouette width that can be calculated from silhouette plots for partitioning methods. It was tried to apply these methods to the given data sets

and the results indicate for all three case studies numbers of 1-3. The MDS plots were also examined by eye and here, the same numbers were found too. These numbers and the variation are reasonable findings when clustering meteorological data. The results are comparable with the range of clusters that is defined by ECMWF (1-6, (1.2)) and it was already mentioned that some synoptic regimes are easier to predict than others (2.3.2). At this point it has to be noted that the solution to the given problem of the study is not necessarily to identify these regimes, as it is done by ECMWF for the geopotential field (1.2). Since resources would allow about 4-6 dispersion forecasts we can further concentrate on finding a reduced ensemble that has similar characteristics as the full ensemble, i.e. on finding a reduced ensemble that can represent the ensemble spread of the full ensemble. So in this thesis, the number of clusters is rather defined by the operational limitations than by natural group identities. To examine if it is possible to represent the ensemble spread of the full ensemble with the 4-6 possible forecasts, the number of clusters is varied from 2-8 for the test scenarios.

4.4 Cluster validation

Cluster validation is a difficult task because one does not know the *correct* solution, and therefore, often no reference solution exists. Furthermore, it is possible that one or more solutions describe the data in a suitable way. For clustering meteorological forecast data one has the additional problem that every forecast has different characteristics and this aggravates the issue of finding a reference. Since in this study it is the aim to find a reduced concentration forecast ensemble according to information contained in meteorological input data it seems reasonable to use separately calculated clustering solutions of the full concentration ensemble as reference. How this concentration ensemble is gained is explained in section (4.1). It has to be noted that this reference can be used for this study but will not be available for the operational application. The following sections describe the developed methods to validate the cluster solutions and the reduced ensemble. It is explained how they can be interpreted and examples for desirable results are presented.

4.4.1 Multidimensional scaling

As shown in figure (2.1a), MDS can be used to visualise the distances between the ensemble members in a plot and if a cluster solution is available, groups can be marked in such a plot (figure (2.1b)). This method is applied to the data matrix of the concentration values and to the corresponding horizontal wind field data which is also arranged in a data matrix. Both data matrices have 50 rows, one for each ensemble member. For the concentration data matrix the grid values for each forecast time are arranged as a 1D-array and a row contains 24 such arrays to hold the 3-hourly information of the 72 h forecast. This is also true for the data matrix of the horizontal wind field, only that for one forecast time two arrays have to be included, one for the u-component and one for the v-component.

Figure (4.8) presents examples of MDS plots that hold cluster solutions where the left one shows the MDS of a concentration data matrix and the one in the middle shows the MDS of a horizontal wind field. The right plot in this figure shows again the set of points gained from the MDS of the concentration data matrix but the clusters are marked according to the clustering solution of the horizontal wind field. If the wind clustering yielded the same group assignments for each member as the concentration clustering (perfect match), the left and the right plot would be equal. As a consequence

of two different variables being clustered and of much information being neglected by comparing a part of the meteorological input data and the dispersion model output a perfect clustering will never be the result. However, if the two variables have a strong linkage the clustering will identify similar groupings and then comparable patterns should be found by analysing these two plots. This is true for the case shown in figure (4.8). The concentration clustering shows two clusters that are far away from each other (the one to the left and the one to right) and two that are close to each other (the two in the middle). We can see that the two clusters that are near to each other also can be seen in the plot to the right but they are not clearly separated anymore. The solution yielded from the wind clustering also clearly separates the two cluster that have a great distance in the left plot. Moreover, the cluster to the right nearly has been perfectly reproduced. If all groupings in the right plot more or less overlapped, the clustering, and therefore the representative members that are found, would be weakly linked to the concentration information and show low quality. This method can be used to qualitatively analyse the linkage between the two cluster solutions.

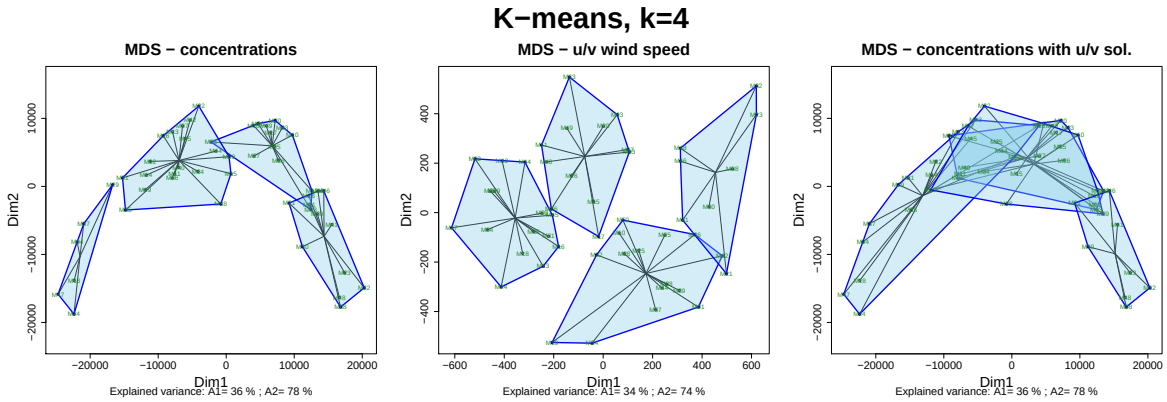


Figure 4.8: Example for validating and comparing cluster solutions with MDS.

4.4.2 Rand Index and Variation of Information

Two measures that were developed to compare cluster solutions are the Rand Index (RI) (2.2.3) and Variation of Information (VI) (2.2.4). With the help of these measures the solution of different cluster algorithms can additionally be compared among themselves. For the RI the two solutions are analysed and a contingency table is set up according to the same or different group assignment of each member. The corrected Rand Index (cRI) is used for validation instead of RI because it is more stringent by accounting

also for the agreement by chance. It ranges from 0-1, where 1 indicates identical cluster solutions and the value is often presented in percent. For strong disagreement of the solution cRI can yield negative values. In contrast to cRI, VI measures the mutual information between the two solutions in bit. It ranges from 0-log n where 0 indicates identical classifications. If the number of objects to be clustered is the same in each data set, it can be normalised to range from 0-1. An example of these two indices is shown in figure (4.9). cRI numbers the similarity between 10 and 20% for nearly all numbers of clusters. The value will be higher at a lower number of groups for most cases. nVI shows that the identical information contained in the two solutions is greater than or equal to a half of what is possible. As stated above, the clustering solution gained from the concentration and the one from the horizontal wind field data matrix are not expected to be identical. However, they should show similarities to some extent which is indicated by the two measures for this example.

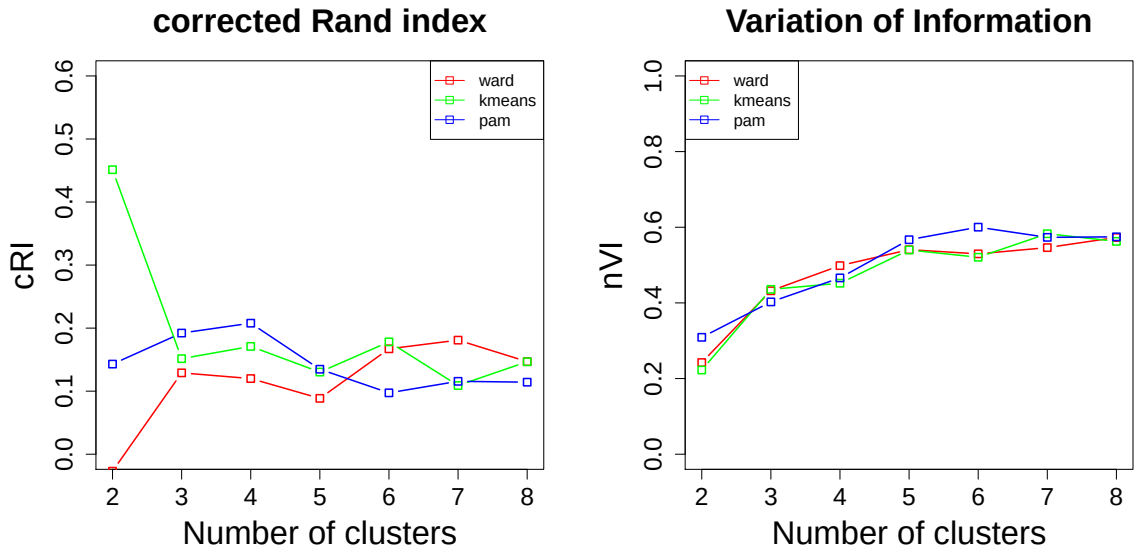


Figure 4.9: Example for validating and comparing cluster solutions with RI and VI.

4.4.3 Ensemble spread

The main aim of this thesis is the reduction of ensembles and these reduced ensembles should represent the spread given by the full ensembles as well as possible. Therefore, a method to validate the quality is to analyse the ensemble spread. As spread measure the variance (VAR) compared to the ensemble mean (EM) is chosen among several

possibilities (see table (2.1)). For a qualitative approach, one could plot the variance for each member of the whole concentration ensemble over forecast time and mark the representative members that are found by applying CA to the horizontal wind field. If the representative members are distributed over the whole range of the ensemble while simultaneously accounting for the relative frequencies of the forecasts indicated by the full ensemble, the clustering can be seen as successful. This is the case for the example presented in the left plot in figure (4.10). Additionally, to get more specific information, a less qualitative approach is developed. To quantify the spread of an ensemble one can average the individual variances given by each member. This is done for the full ensemble, the reduced ensemble and each sample of cluster members where the EM is always calculated for the respective group. An example is the right plot in figure (4.10). The spread of the reduced ensemble should be very similar to the spread of the full ensemble and since a cluster should contain similar members, the cluster spreads should be smaller than the other two. This is true for the given example because the variance of the reduced and the full ensemble (blue and black line) overlap for almost every forecast time and the cluster variances (dashed red lines) are always below them.

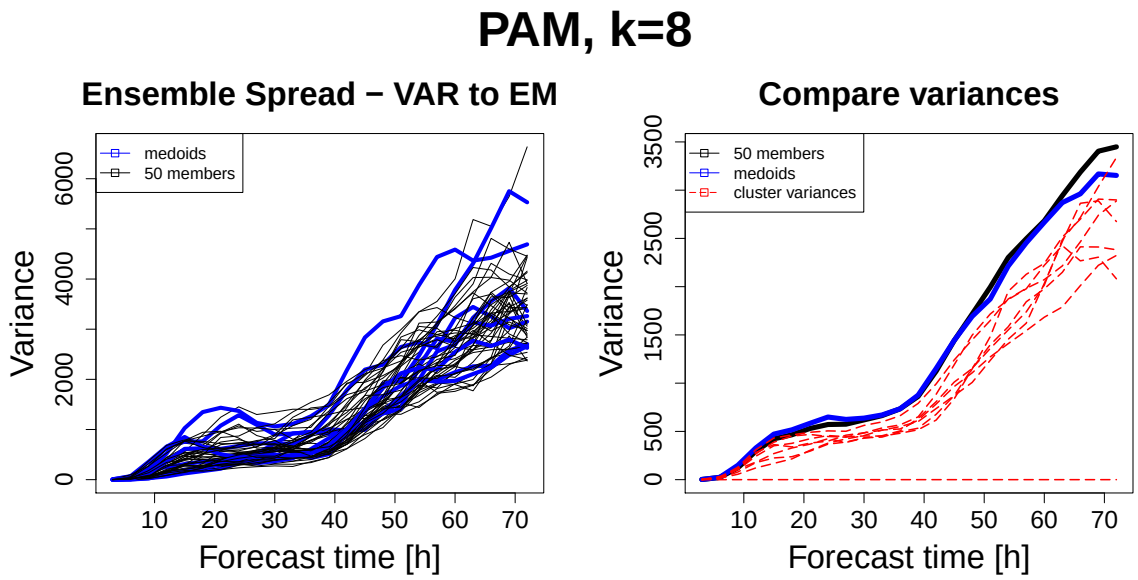


Figure 4.10: Example for analysing the spread of a reduced ensemble. The left plot shows the concentration ensemble (black) and representative members (medoids, blue). In the second plot the variance given by the full ensemble (black), the medoids (blue) and each cluster (dashed red) are compared.

5 Results

5.1 Test scenarios and investigations

To test the developed concept shown in figure (4.1a), three test scenarios are defined to simulate volcanic eruptions. Table (5.1) provides an overview of theses scenarios. The main differences between them are the synoptic situations and the different handling of the surface fluxes of the first two time steps, see section (3.2). An additional difference is that the pressure level data for clustering is only available for data set 2 (3.1).

	Volcano	Coordinates	Data set (Table 3.1)
Scenario 1	Eyjafjallajökull, Iceland	N 63.630, E -19.620	1
Scenario 2	Tambora, Indonesia	N -8.250, E 118.000	1
Scenario 3	Etna, Italy	N 37.734, E 15.004	2

Table 5.1: Overview of simulated test scenarios. Coordinates are taken from Smithsonian Institution (2013).

The events are discussed according to the synoptic situation and a chosen sample of concentration forecasts is given to visualise the characteristics of the forecasts. Three different altitudes are investigated, namely 8, 10 and 12 km and model level and pressure level results are compared for events simulated with data set 2. As a consequence of the huge amount of data the presentation mainly concentrates on altitude 10 km. Furthermore, the results of the introduced validation methods in section (4.4) are presented and analysed for the individual scenarios.

5.2 Scenario 1

The first scenario represents a volcanic eruption of the Eyjafjallajökull in Iceland. At the start of the forecast at 2013-03-14 00 UTC, the SO_4 tracer is released with constant emission (4.2) for the whole forecast time of 72 h. Figure (5.1) describes the synoptic situation at the beginning of the forecast. This GFS weather chart shows the geopotential at 500 hPa, surface pressure and the relative topography H500 - H1000. We have a huge trough over Europe ranging from Scandinavia to North Africa. The rear of the trough is strongly influencing the upper air flow in the region of the volcano. This indicates an intense forcing for the transport of volcanic ash through the horizontal wind field. Additionally, a short-wave cyclone is embedded near the volcano which is often difficult to predict. In that case, this fact is represented by the ensemble and can be observed in the ensemble spread which shows a peak in the middle of the forecast (see figure (5.4)).

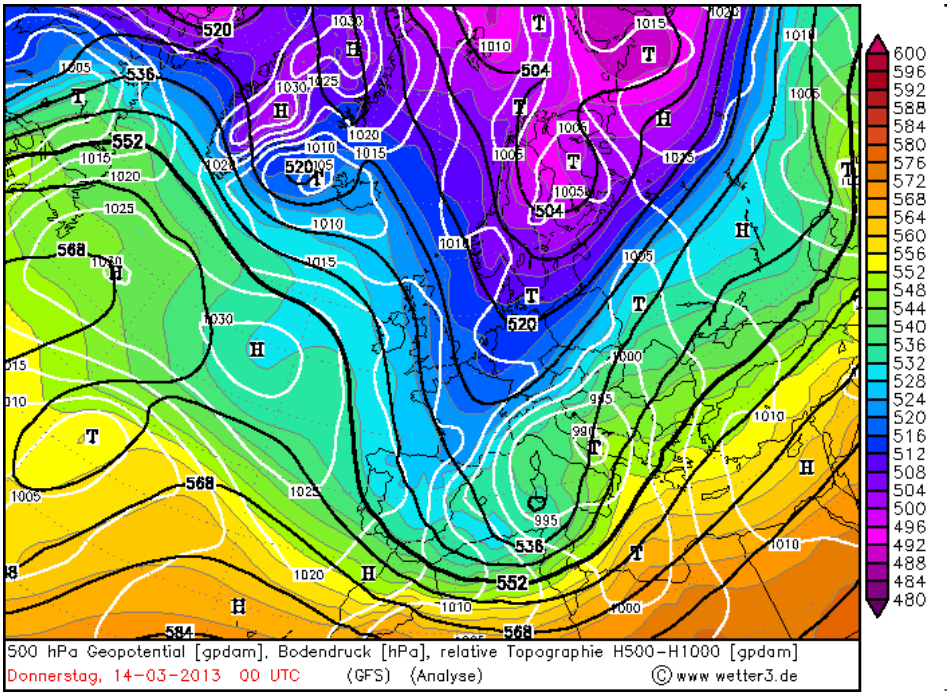


Figure 5.1: Synoptic situation of scenario 1 (Wetter3 archive, 2013)

At the beginning of the study it was not clear if an ensemble of concentration forecasts with a time range of 72 h would show a considerable variation since the perturbations of the ECMWF ENS are growing only after forecast day 1-2. A sample of 6 out of 50 possible concentration forecasts presented in figure (5.2) demonstrates that

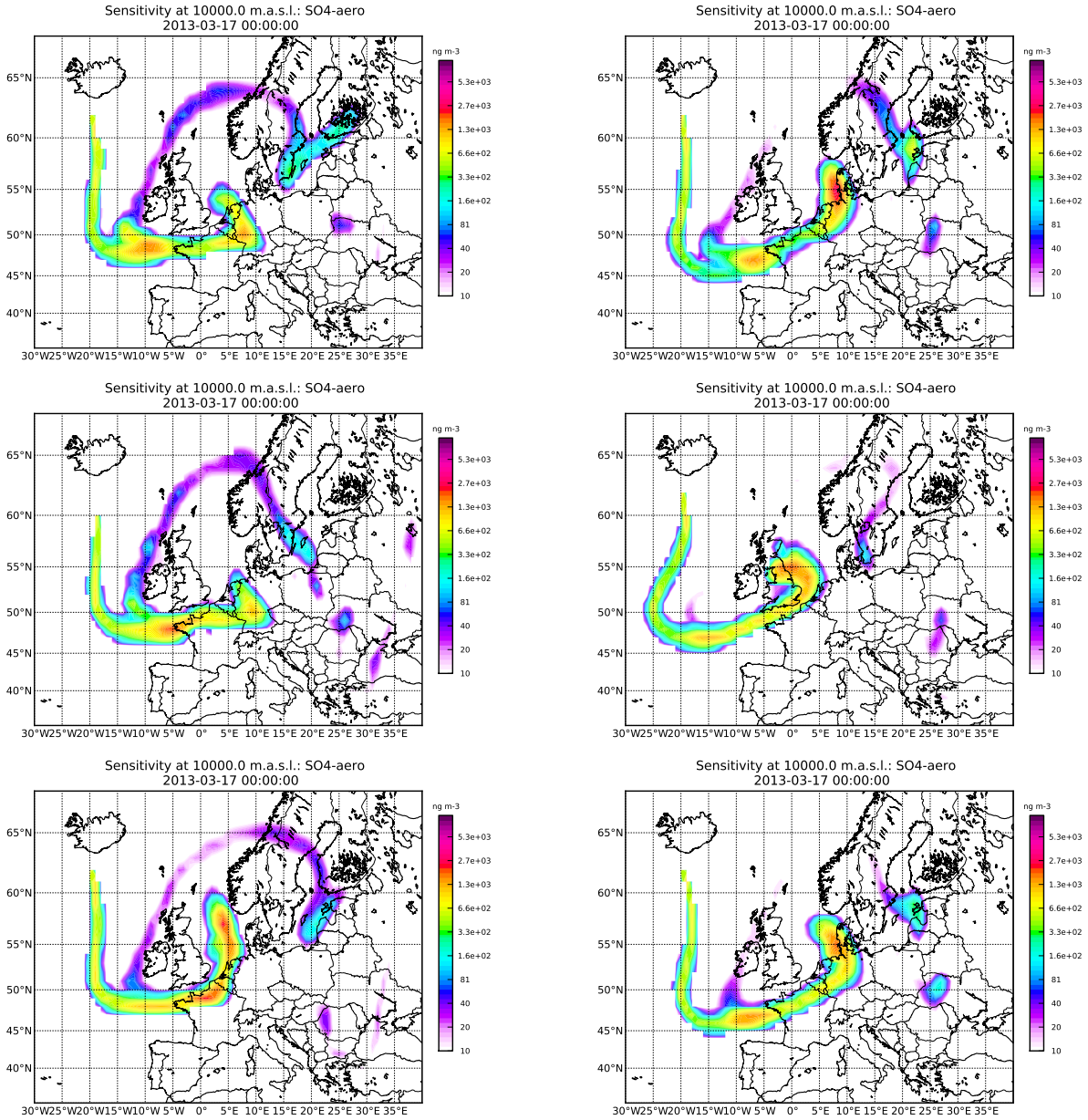


Figure 5.2: Sample of concentration forecasts to characterise scenario 1 with the eruption of Eyjafjallajökull on Iceland. The plots show the concentrations between 9.5-10 km at forecast time t+72 h.

it already makes sense to use a multiple-input ensemble for dispersion simulations for a 3 day forecast range. This is also observed for the other two test scenarios. For the forecasts in the left column we see that some members are more affected by the mentioned short-wave in the upper air flow and that the plume splits at the south-west of Ireland and makes a northward curve towards Scandinavia. Another characteristic is that some plumes have two maxima of concentration (first row and the lower right plot). The right forecast in the middle has a curved plume towards the west at latitude 50° and it is the only forecast that does not affect northern France and Germany.

After clustering the horizontal wind field at different heights and also the corresponding concentration fields one can validate the clustering solutions with the comparison measures cRI and nVI. Figure (5.3) presents the results for this scenario. These measures indicate a linkage between the horizontal wind field that can be numbered with 10-20% according to the cRI. Both indices denote an increase of similarity when going from 8 km altitude to 10-12 km. This is reasonable since 10-12 km defines the top of the troposphere where the horizontal wind field dominates and other processes that influence the ash transport have a minimum contribution. For lower numbers of clusters, cRI is a bit jumpy when it is compared among the algorithms. nVI instead indicates a constant decrease of similarity when increasing the number of groups up to 5-6. This is what would be expected since the linkage between the two different variables is present but moderate and for finer groupings clusters start to get more fuzzy.

When investigating the ensemble spread we want to determine for which number of representative members the ensemble spread can be reproduced. The upper plots in figure (5.4a) show that two members provide far too little information and suggest that more members are needed. The spread of the reduced ensemble converges to the spread of the original one when the number of members is increased to 50 in this case. So the question is: By which number of medoids is enough information available? When looking at the reproduced spread in the lower plots in figure (5.4a) it can be noticed that this result is sufficient. Although on average, all three algorithms perform almost equally well, there are differences for each case and number of clusters. While the PAM algorithm yields acceptable results for k=8, Ward's method underestimates the spread by choosing too many members from the main solution. The opposite is true for K-means. It nicely covers the whole spread with the chosen members but overestimates the value around forecast time 35 h.

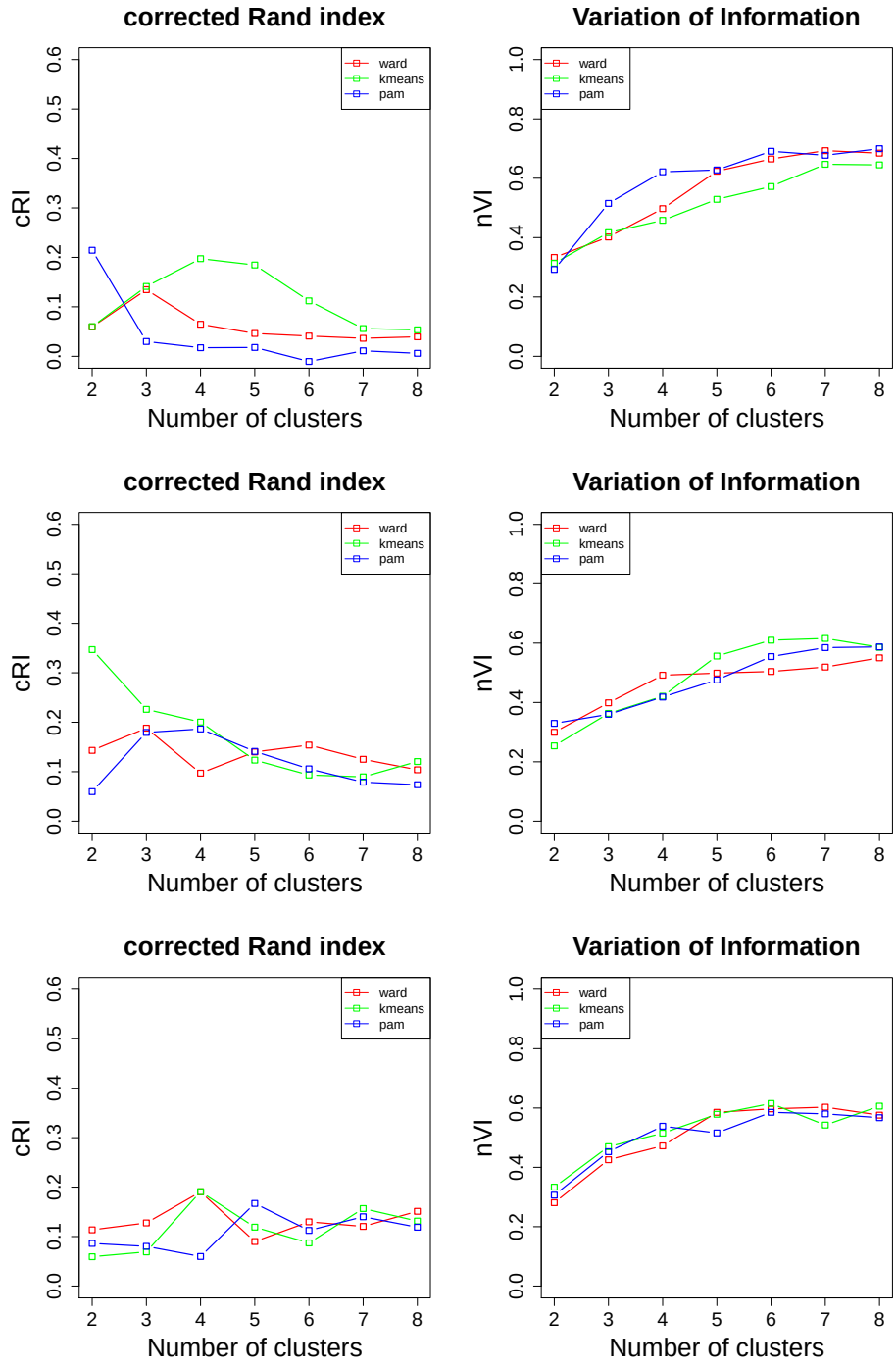
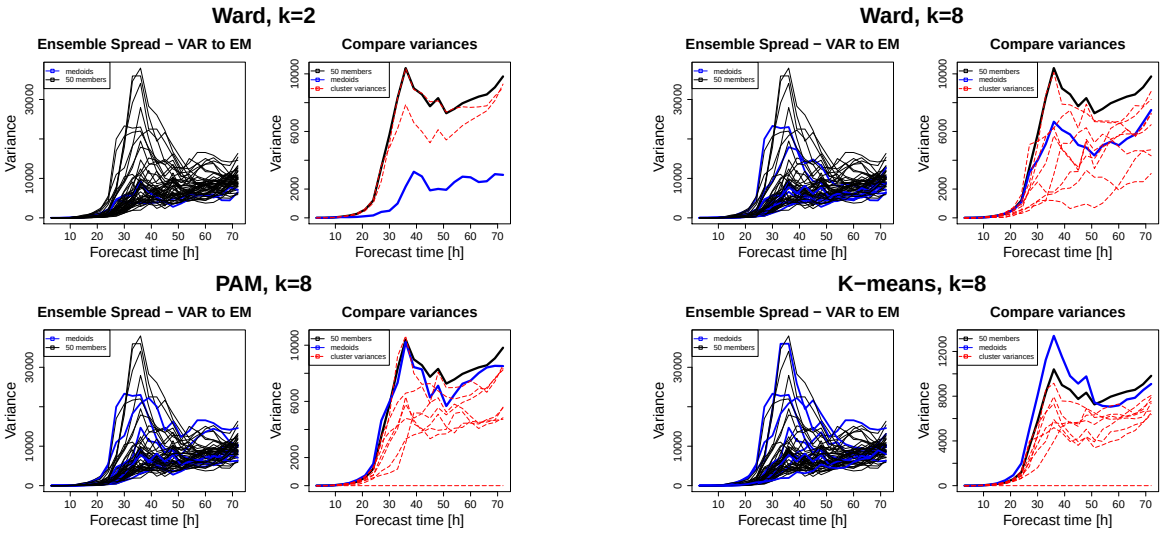


Figure 5.3: Comparing clustering results of different heights for scenario 1 according to cRI and nVI. The plots show the results of two measures for 8, 10 and 12 km starting from the top.



(a) Relation between reproduced spread and the number of clusters

(b) Under and over estimating the real ensemble spread

Figure 5.4: Examples to show properties of the reproduced ensemble spread. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.

For this scenario, the PAM algorithm with 6 clusters yields sufficient results while keeping the number of clusters to a minimum (lower plots in figure (5.5)). Also, almost every cluster variance is kept low. K-means (upper plots in figure (5.5)) also gained good results but overestimates the spread a little bit. Both solutions have the problem that they underestimate the spread near the end of the forecast period. For this event, the variance of several members is not growing constantly with time but has a great peak in the middle of the forecast. This circumstance makes it not easy to approximate the variance of the full ensemble equally well for all time steps. Due to the fact that the clustering distance measure is applied over the whole time range at once, areas where bigger differences occur have more weight. Thus, the original spread will be better explained in areas with relatively high differences in variance.

To compare the performance of the developed method we can compare the results for the ensemble spread against the reference clustering solution, i.e. the classification according to the concentration fields. In figure (5.6) the effects of neglecting a big amount of information by only using the horizontal wind field can be seen for the cluster variances. In contrast to the results in figure (5.5) the spread of the clusters is

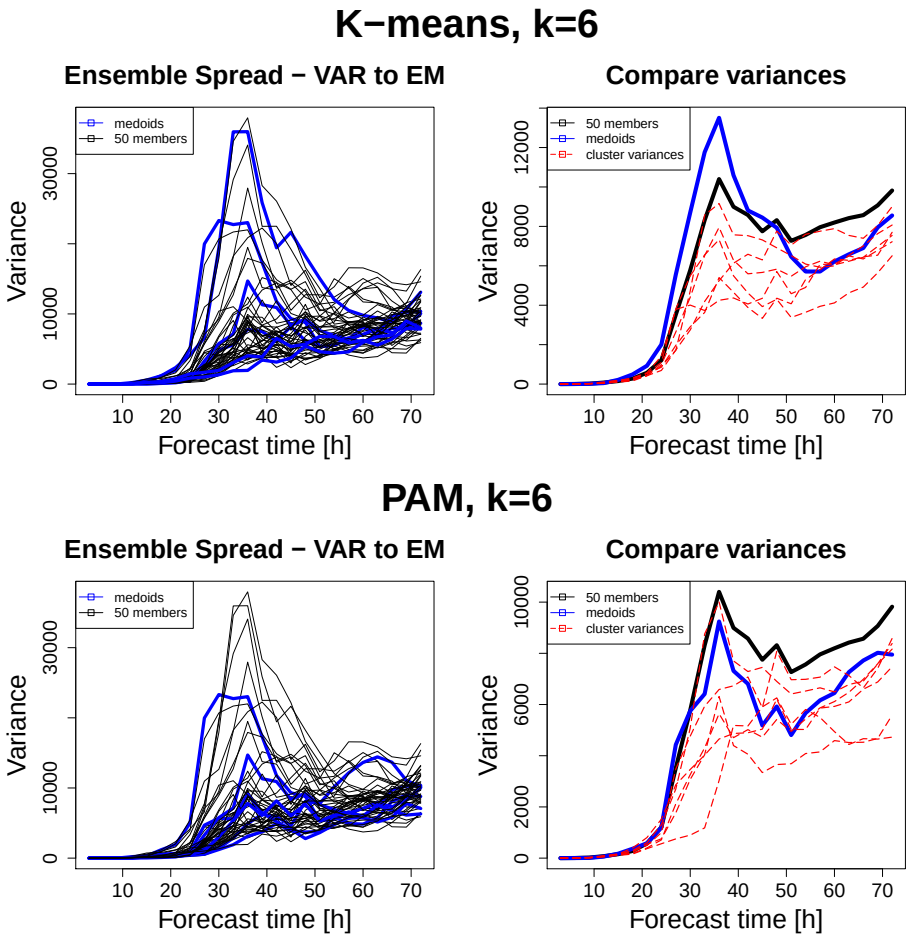


Figure 5.5: Minimum number of clusters to sufficiently represent the ensemble spread for scenario 1 at 10 km.

kept very low after forecast hour 30. The mean medoid variance is nearly always above the cluster mean variance and gives a good approximation to the full ensemble spread. However, the underestimation at the end is also present here.

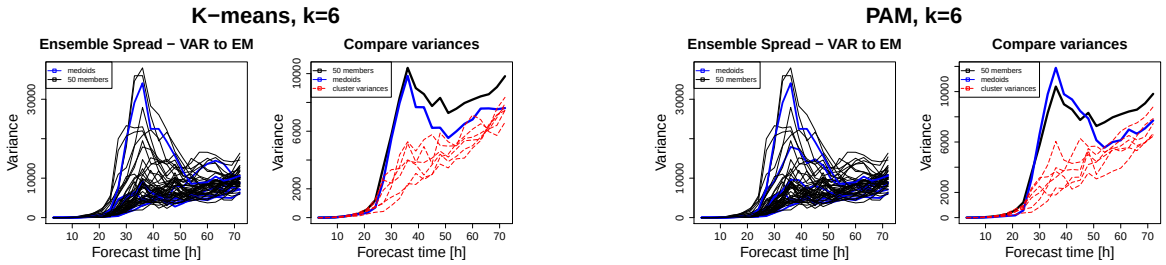


Figure 5.6: Reproduced ensemble spread by clustering concentration fields and not horizontal wind fields for scenario 1 at 10 km. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.

When visualising the results for 6 clusters, MDS plots start to get unclear. Nevertheless, figure (5.7) shows clusters of moderate size and members that are far away from each other are still separated. This confirms the findings that the PAM algorithm with k=6 groups is a good clustering solution for this case.

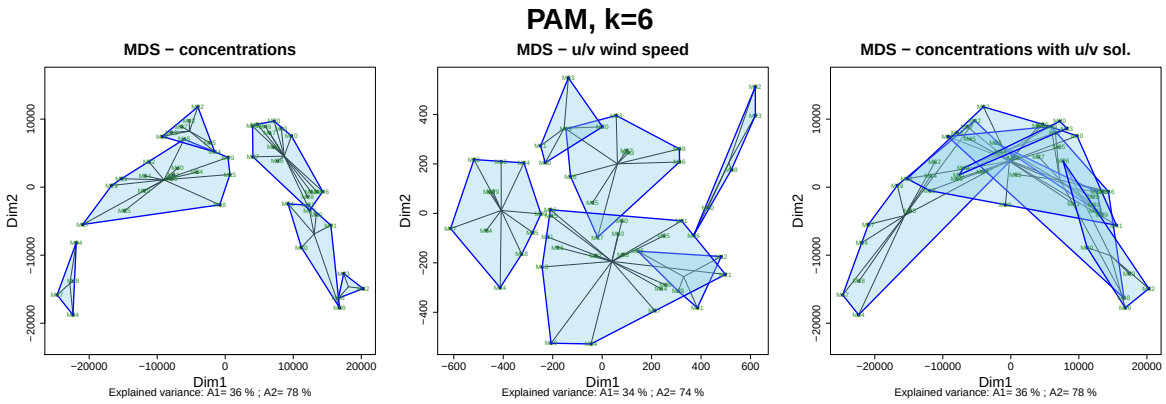


Figure 5.7: MDS for the PAM algorithm with k=6 of scenario 1 at 10 km

5.3 Scenario 2

Scenario 2 describes an eruption of the Tambora volcano in Indonesia. The forecasts also start at 2013-03-14 00 UTC for a time range of 72 h. Compared to scenario 1 we have a completely different synoptic situation. Since the volcano is located in the tropics the transport of ash is less influenced by a distinct horizontal wind field. Figure (5.8) shows the very weak gradients of geopotential.

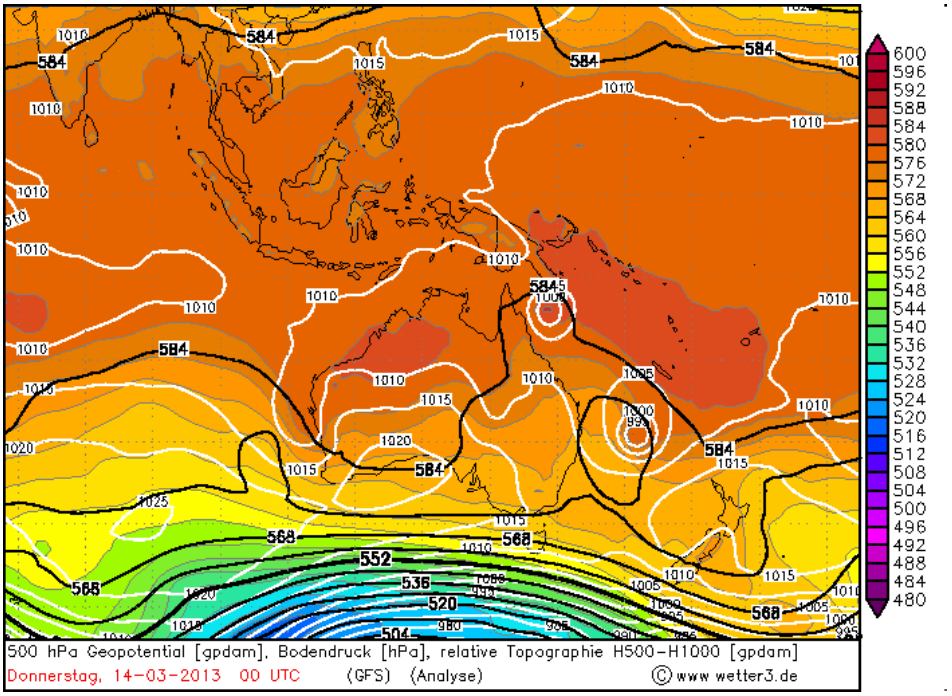


Figure 5.8: Synoptic situation of scenario 2 (Wetter3 archive, 2013)

In this region, convection has a major influence on the transport process. This can be seen by examining the sample of forecasts of concentration fields in figure (5.9). The ash plume stays near the volcano and is distributed randomly around the source. So the results of CA are expected to yield not clearly separated groups. Additionally, this scenario should present a weakness of the developed method, namely, that no clear clustering is possible when the forcing through the horizontal wind field is very weak. Further is has to be noted that the quality of the ECMWF data for the tropics is not as good as for the other two case studies located in or near Europe.

Figure (5.10) presents the results for the cluster comparison measures cRI and nVI. cRI numbers the similarity of the cluster solutions for 8 and 10 km below 10 %. Also,

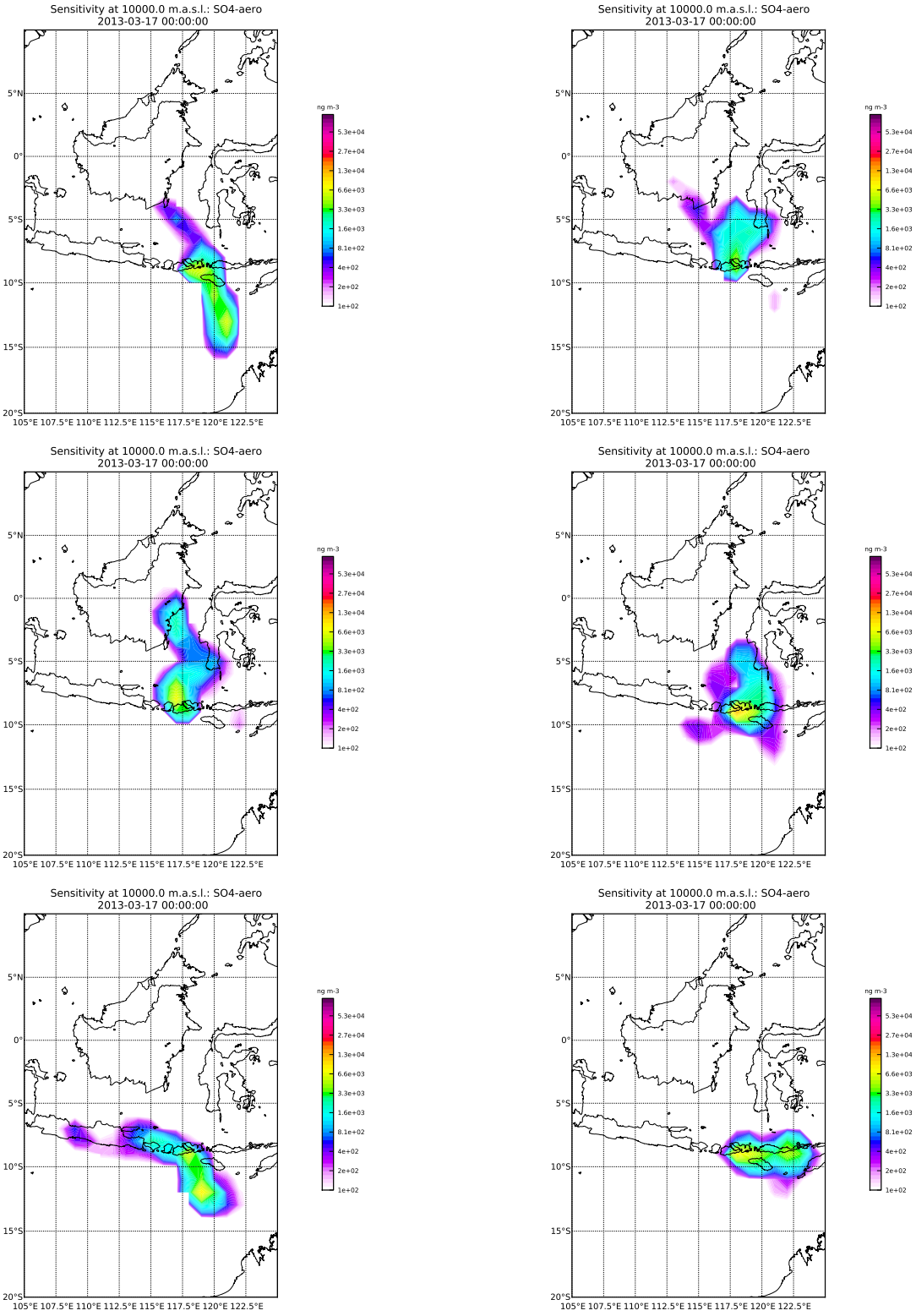


Figure 5.9: Sample of concentration forecasts to characterise scenario 2 with the eruption of Tambora in Indonesia. The plots show the concentrations between 9.5-10 km at forecast time t+72 h.

nVI indicates, compared to the other scenarios, a rather poor clustering at altitude 10 km with a value of 0.7 for 8 clusters. For the level of 12 km K-means yields better results according to cRI; PAM and Ward's method indicate a slightly better agreement for numbers of clusters 5-7.

Through the investigation of the ensemble spreads in figure (5.11) we see that these results are in accordance with cRI and nVI. As a consequence of the high variability of the concentration fields it makes only sense to examine the solutions with 8 clusters. Although the underlying cluster solutions of the chosen medoids are not exceedingly similar to the concentration cluster solutions, the results for 12 km are sufficient (figure (5.11a)). The quality is comparable with the findings of scenario 1 in figure (5.5) but there, 6 instead of 8 clusters are enough. It has to be noted that with 8 representative members we are already beyond the upper limit of operational capabilities. We also see in figure (5.11b) that the clustering at altitude 10 km does not work well for this case. The spread is poorly reproduced for the last third of the forecast. Moreover, the variance of the individual clusters is almost every time higher than the mean medoid variance and is sometimes exceeding the spread of the full ensemble to a certain extent. These are undesirable results.

To show that a big part of the bad results for this scenario at 10 km is due to the lack of information when doing the clustering based on the horizontal wind field we can compare them with the results based on the concentration fields presented in figure (5.12). The spread of the full ensemble is acceptably reproduced and cluster variances are low, especially for the PAM algorithm. Additionally, we see that the convection influenced dispersion of ash is more difficult to classify because the cluster variances gained by concentration clustering are not as low as for scenario 1 (see figure (5.6)).

The MDS plots also underline the findings mentioned above. Figure (5.13) shows a clustering solution for 10 km gained by the K-means algorithm. Although the fields should be separated only into 4 groups the areas of the clusters extremely overlap in the right plot. This indicates low physical linkage between the horizontal wind fields and the concentration fields.

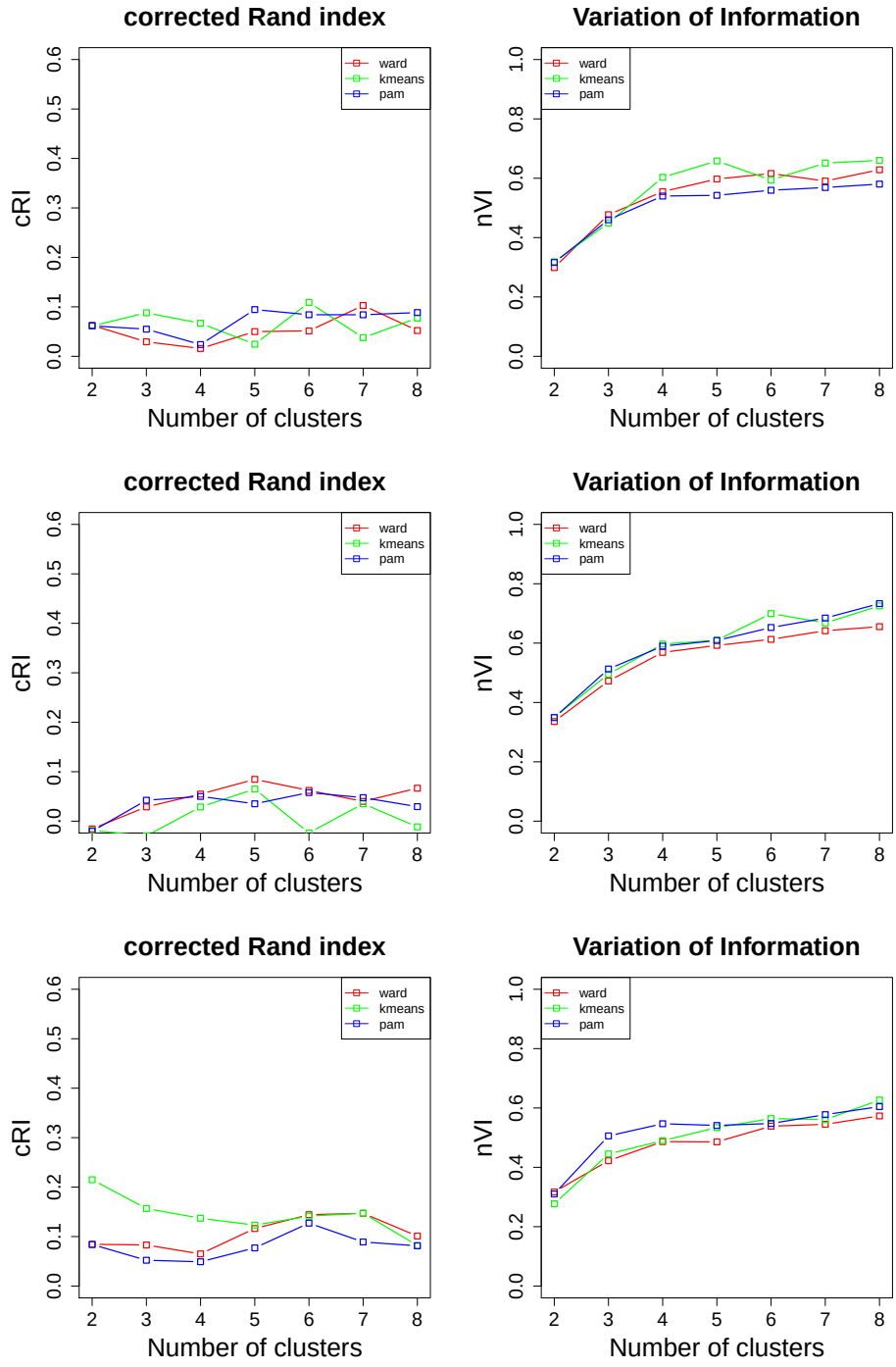
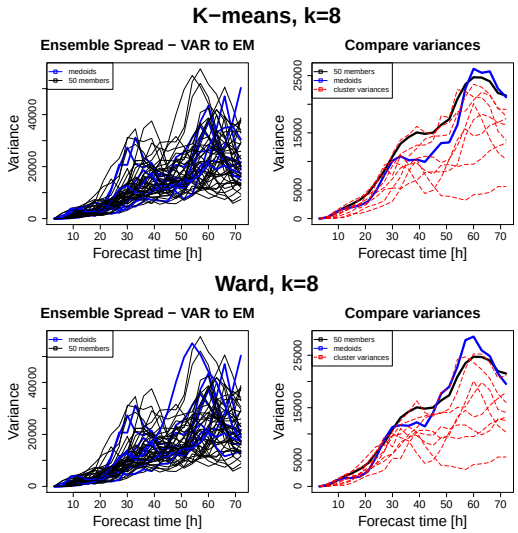
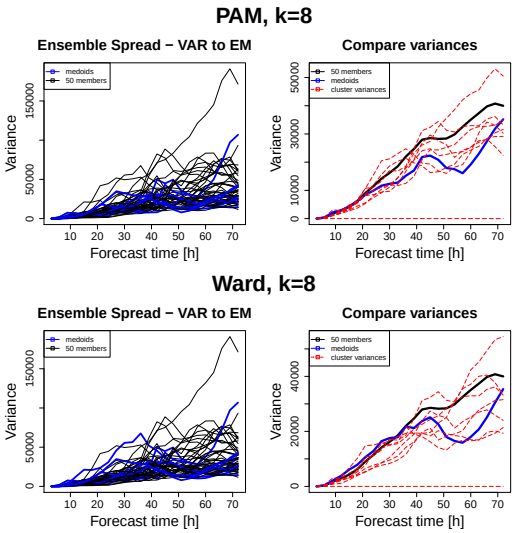


Figure 5.10: Comparing clustering results of different heights for scenario 2 according to cRI and nVI. The plots show the results of two measures for 8, 10 and 12 km starting from the top.

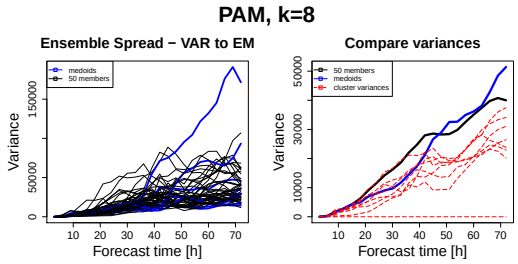


(a) Reproduced ensemble spread for 12 km



(b) Reproduced ensemble spread for 10 km

Figure 5.11: Examples to show the reproduced ensemble spread for scenario 2. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.



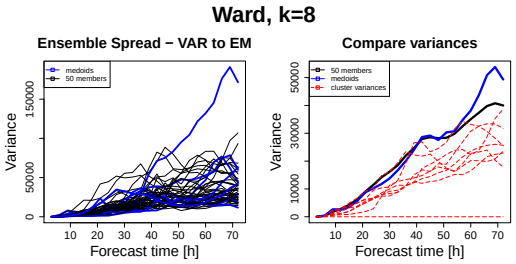


Figure 5.12: Reproduced ensemble spread by clustering concentration fields and not horizontal wind fields for scenario 2 at 10 km. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.

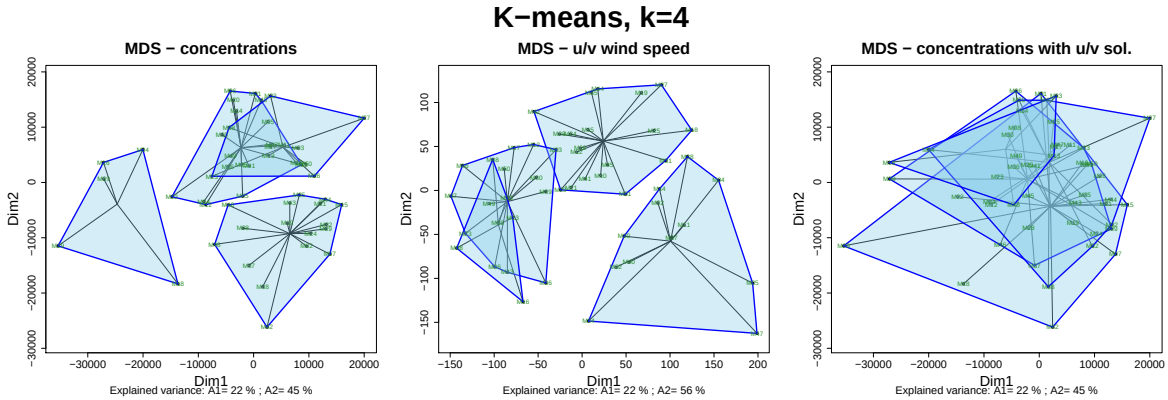


Figure 5.13: MDS for the K-means algorithm with k=4 of scenario 2 at 10 km.

5.4 Scenario 3

The third scenario that is discussed describes a simulated eruption of the volcano Etna in Italy. The forecast has a time range of 72 h and is initialised at 2013-03-21 00 UTC. To analyse the synoptic situation the geopotential field is shown in figure (5.14). The area around Sicily is influenced by a trough of a large Rossby wave that is moving eastward due to the large mean zonal wind. Such waves can be predicted with not too much uncertainty and therefore, the variance of the ensemble forecast for the investigated heights will be relatively small.

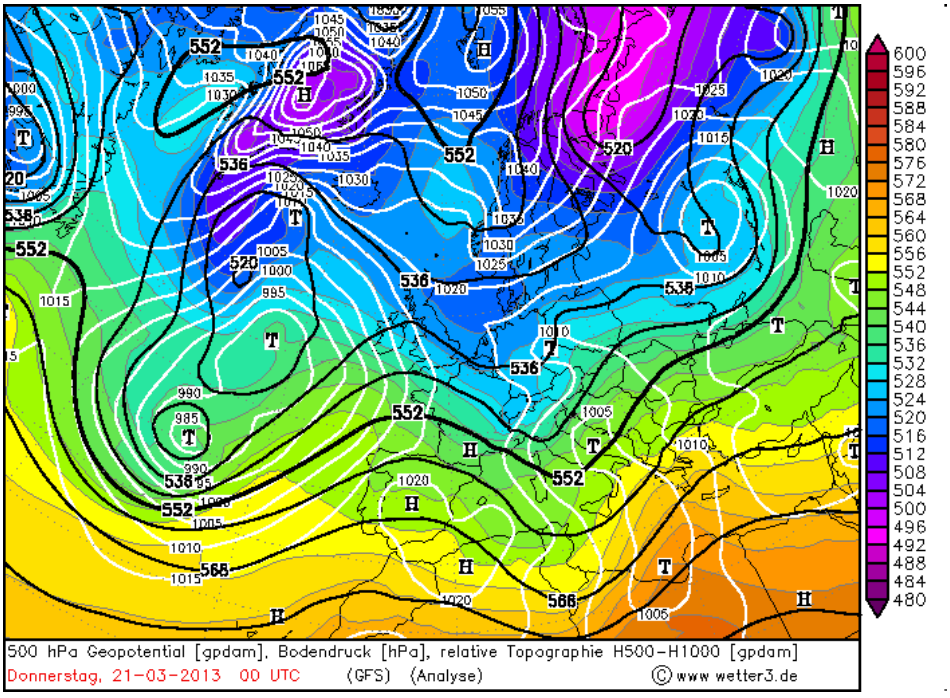


Figure 5.14: Synoptic situation of scenario 3 (Wetter3 archive, 2013)

When analysing the sample of concentration forecasts in figure (5.15) we see that two different groups develop. These groups are similar to each other because both patterns of the ash plume mainly follow the Rossby wave. The main difference is that the second wave crest of the plume is reaching latitude 40° again for the forecasts in the left column, whereas none of the forecasts in the right column have such a strong northward deflection.

This scenario is simulated with the meteorological data set 2 (see table 3.1) where the horizontal wind field for 8 and 12 km is available on both model and pressure levels.

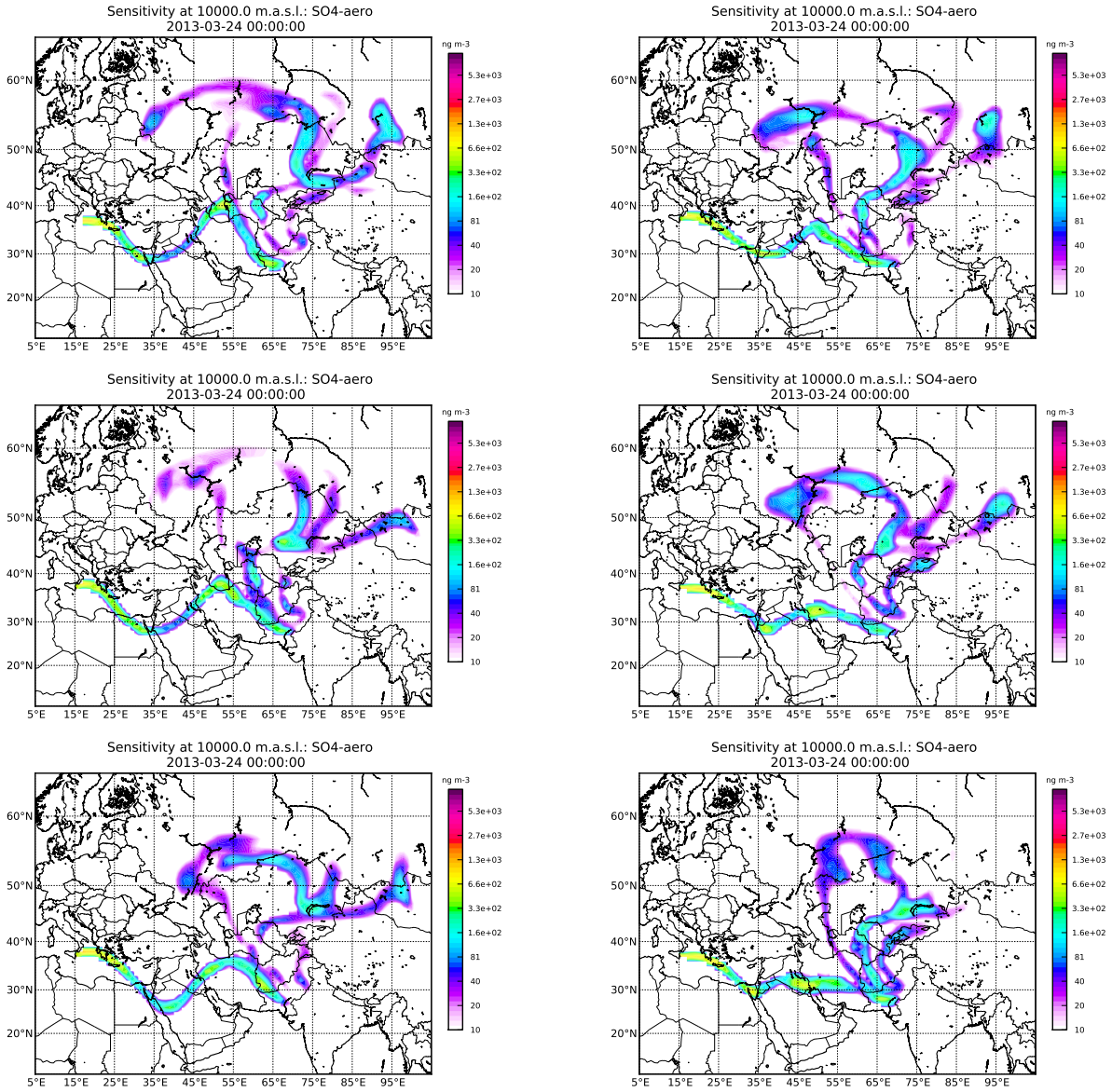
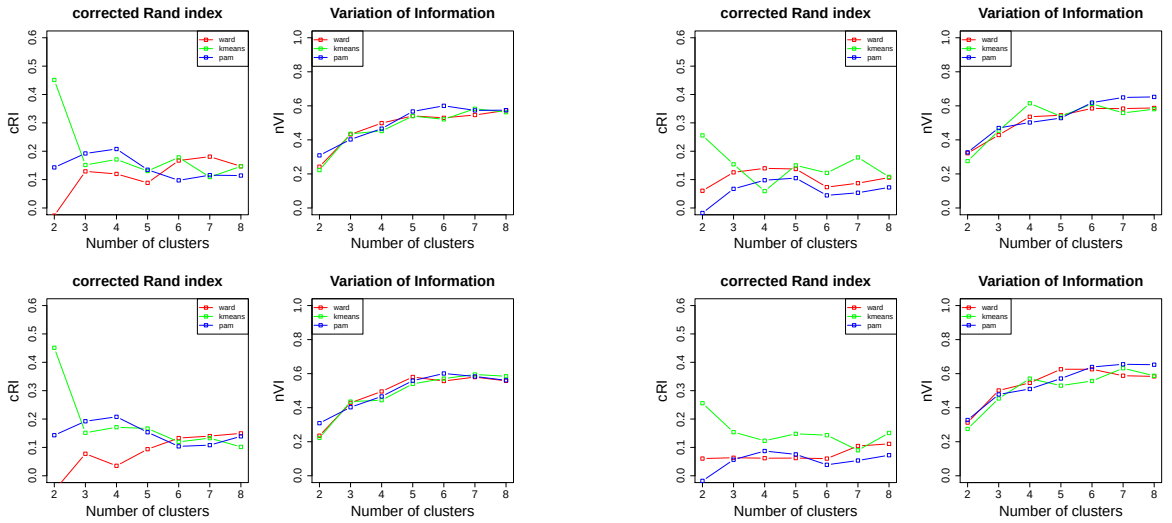


Figure 5.15: Sample of concentration forecasts to characterise scenario 3 with the eruption of Etna in Italy. The plots show the concentrations between 9.5-10 km at forecast time t+72 h.

Since the pressure of the two levels differs only slightly it is interesting to see how sensitive the clustering is according to height. This is especially true with the approximation of height with the help of the barometric formula because this is a strong assumption. Furthermore, the wind fields differ to a small extent due to the different interpolation steps. The model level data are interpolated from T319 ($\approx 0.5°$) out of vorticity and divergence to a 1° LL grid. The pressure level data is interpolated directly from the provided ECMWF 0.5° LL grid for which an initial resolution of T639 ($\approx 0.25°$) was used and it is therefore more accurate. This fact can additionally influence the clustering sensibility. Figure (5.16) shows the cluster comparison measures for model and pressure levels at two different heights. For 8 km the results are nearly the same but at 12 km we can notice bigger differences for Ward's method and K-means according to cRI and nVI.



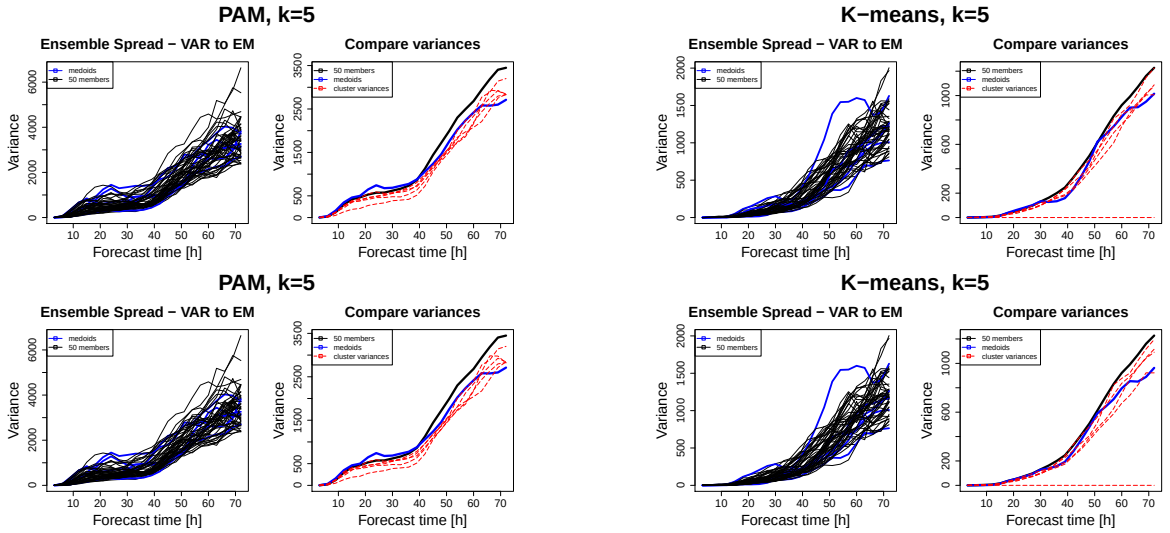
(a) Comparison at 8 km altitude

(b) Comparison at 12 km altitude

Figure 5.16: Comparing clustering results for model and pressure levels at 8 and 12 km according to cRI and nVI. The upper row shows the model level results and the lower row the pressure level results.

To analyse the influence on the reproduced spread, figure (5.17a) shows the results at 8 km for the PAM algorithm and the K-means method at 12 km. The PAM algorithm is not as much affected by the differences as the other two algorithms. Cluster variances are almost identical and the same medoids are chosen from the clustered groups. This is true for both heights. As indicated by cRI and nVI, K-means behaves more different and therefore yields somewhat diverse results for cluster variances when

examining figure (5.17b). There is also a disagreement about one medoid which affects the mean medoid variance at forecast time t+40 h.



(a) Reproduced ensemble spread for 8 km

(b) Reproduced ensemble spread for 12 km

Figure 5.17: Comparing reproduced ensemble spread for model and pressure levels at 8 and 12 km. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.

To compare the performance with scenario 1 and 2 at 10 km, figure (5.18) shows the results for cRI, nVI and the reproduced spread for all three algorithms. It can be noticed that the comparison measures indicate constant performance of the clustering methods against each other throughout all numbers of clusters. Most of the time, cRI and nVI provide acceptable information about the quality of the clustering and therefore also for the chosen members. When analysing the findings of the three algorithms we see that the best clustering solution does not yield the best reduced ensemble in every case. Although for K-means twice as much similarity is demonstrated by cRI, the mean average medoid variance shows a less satisfying result.

It has to be noted that for all examples the criterion of low cluster variance is not fulfilled. This fact is no issue and is in agreement with the discussed synoptic situation. Since all forecasts are very similar, the variance of the individual groups will not be much lower than the variance of the full ensemble. Figure (5.19) presents the reproduced spreads for the clustering solution due to the concentration fields. There,

the mean medoid variance can be lowered after forecast hour 45 but also not to a large extent. Moreover, it is shown that the PAM algorithm needs 6 medoids to reproduce the full ensemble variance (compare the plots in figure 5.19) in contrast to the clustering of the wind field data (see figure 5.18) where 5 representative members are suitable. Overall, it can be concluded that 5 medoids are needed for this scenario to acceptably reproduce the ensemble spread.

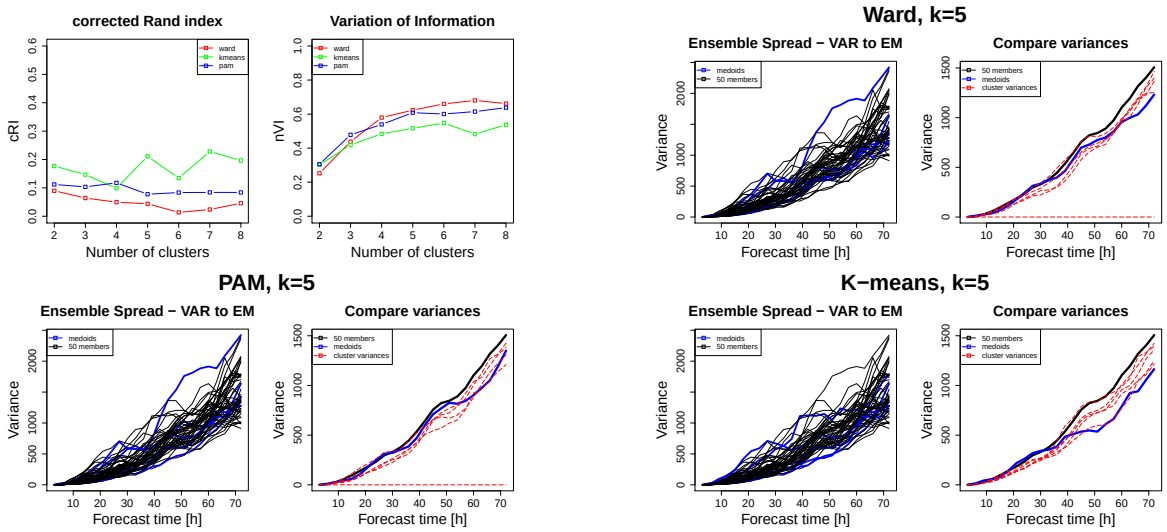


Figure 5.18: Comparing reproduced ensemble spread at 10 km for scenario 2 by showing the results of all three algorithms for the minimum number of medoids (k=5). Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances. Additionally, the cluster comparison measures cRI and VI are presented.

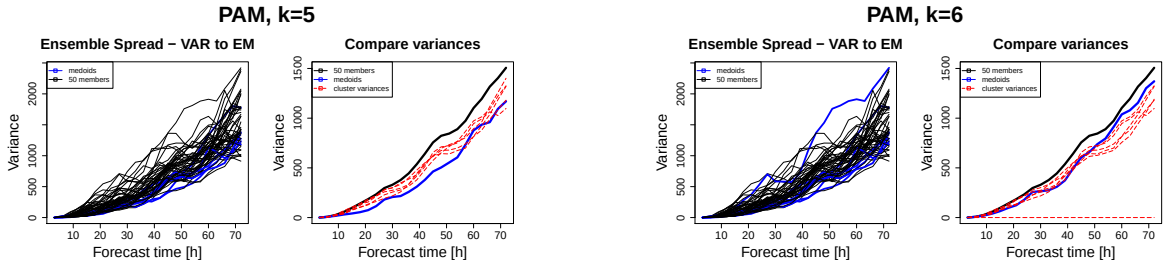


Figure 5.19: Reproduced ensemble spread by clustering concentration fields at 10 km for the PAM algorithm with 5 and 6 medoids. Black lines denote members of the full ensemble, blue lines the medoids of the clusters and the red dashed lines show the average cluster variances.

6 Discussion

Results of the simulated volcanic events in this thesis show that a linkage between clusters of horizontal wind fields and ash concentration fields can be found. This is underlined by the cluster comparison measures corrected Rand Index (cRI) and normalised Variation of Information (nVI). cRI values the similarity for the cluster solutions with 10-30% and nVI indicates that for numbers of clusters between 5 and 8 still one third to one half of the mutual information is explained. With the use of multidimensional scaling it is possible to visualise the properties of cluster solutions. These plots demonstrate that with the classification of wind fields the main patterns of concentration clusters are enclosed. Findings further confirm that reduced ensembles, composed of 5-8 representative group members based on wind clusters, can acceptably reproduce the spread of the initial ensemble. Through the investigation of the test scenarios it can be shown that the number of medoids needed for the reduced ensemble strongly depends on prevailing synoptic situations. It is not easy to carry out further verification on the reduced ensembles because there is no real measurement data available to apply typical probability skill scores.

The main advantage of the method can be seen when comparing the timings of the two concepts presented in figure (4.1). An ensemble forecast with 4 members takes about 3 h with the current setup when using the clustering of the horizontal wind field. To yield the same ensemble with the classification of 50 concentration forecasts 14.5 h are needed. It is found that the most time intensive part of the reference concept is data retrieval and preprocessing for the dispersion model.

The similarity found between horizontal wind fields and concentration would probably not be classified as sufficient from a theoretical point of view when not knowing the properties of the variables and the simplifications. However, one has to keep in mind that the aim of the thesis is to quickly produce concentration ensemble forecasts after a volcanic event with limited operational resources. To reach this target it is necessary to neglect information and make assumptions. A part of these trade-offs concerns the input data and transport processes. Although important parameters for dispersion in higher levels are used other parameters must be disregarded. On the one hand, this is due to limitations of the amount of data available for daily retrieval. On the other hand, it makes no sense to combine all meteorological parameters in the distance func-

tion for CA without knowledge of the equations that describe the interaction between them and the influences on the ash concentration. It is the task of the dispersion model to describe turbulent wind fluctuations and convection. This fact highlights the strong assumption that is made when comparing meteorological input data with model output. Other trade-offs are related to the definition of the spatial domain for clustering. To choose vertical levels for comparison a standard isothermal atmosphere is assumed. With the barometric formula pressure levels are assigned to investigated altitudes. The clustering is always optimised for one altitude. Combining different levels yielded fuzzy clusters. Moreover, it is not reasonable to make a forecast for all possible cruising altitudes under the condition of such strong negligence of information because then the forecast will not be useful for any level. To optimise the linkage between the fields it is essential to adapt the wind field area to the area affected by the ash plume. Otherwise, features would be classified that are not primarily important for the ash transport. This area of interest is determined with a first guess simulation based on the available ECMWF HRES data. For accurate selection of important grid points first guess runs for all ensemble members would be needed. Since data are not available for these runs an uncertainty area is added to the first guess simulation. The resulting area is used for all members and it accounts for the variability of the ensemble. As a consequence, there are still unimportant areas included or, even worse, some areas are not selected due to underestimation of the ensemble spread. In the worst case, the plume simulated with the operational run is deflected in another direction than all ensemble members. This is not very probable but a very strong limitation.

The developed method is, in principle, applicable for all regions where the horizontal wind field determines the transport. If the method is used for cases where other processes dominate the transport, the reduced ensemble based on the wind clusters will lose its justification. It is tried to include all characteristics of the complete ensemble as well as possible. Although the ensemble spread, measured according to the variance of the ensemble mean, is sufficiently approximated one has to keep in mind that it is probable that the reduced ensemble does not include the forecast with the highest concentrations. This is a problem that one has to accept when using downsized ensembles.

7 Conclusions and recommendations

In this thesis a quite general method for reducing the ensemble size through clustering has been applied to probabilistic forecasts of volcanic ash. The concept is not limited to such forecasts and can also be used for other applications that deal with dispersion simulations. Further, it is useful for institutions with limited resources. In the future, with more computational capabilities, the simulation and clustering of the full ensemble will be the method of choice. Then the reduced ensemble will include information about all available meteorological parameters and both atmospheric and transport processes. Additionally, the full set of forecasts can be provided to the user to analyse them for outliers and extreme values.

Through this study only a very small sample of nearly infinite weather regimes is analysed and more practical investigations should be conducted. The dependence of the number of clusters on predominant synoptic situations is shown. This fact could be investigated in subsequent studies to vary the size of the reduced ensemble in an appropriate range, like it is done at ECMWF (1.2). Furthermore, it may be possible to define a variable distance function according to the strength of the forcing due to the wind field. As shown for Scenario 2 (5.3), a less pronounced linkage between the variables is found for such situations and the clustering of the vertical velocity may be a better choice for the tropics. In following studies, attention regarding classification could be drawn to the distribution of the underlying variables. A histogram of the components of the horizontal wind field shows a normal distribution whereas the ash concentration fields have a skewed distribution. The concentration fields were transformed to follow a log-normal distribution for short examination and then slightly increased values for comparison measures cRI and VI were found. Since the linkage is already proven with the original distributions this topic was not pursued. Moreover, there is a need for better estimation of the uncertainty area that is added to the first guess simulation to account for the variability of the ensemble. The constant enlargement is a first approximation but the function that describes this increase should contain some information about the variance of the meteorological forecasts. Including current spread measures of the ECMWF ENS forecasts or information of their clustering product for the geopotential field could possibly yield some improvement.

As a last point, the preprocessing methods for the meteorological input should

be discussed. Some experts share the opinion that the de-accumulation scheme (3.3.2) described by James (2000) is a bit overloaded. For some cases it shows patterns that are not real and fields appear to provide more informations than there actually are. Suggestions are made to simply scale the variables and subtract the previous field from the actual time step. A problem that also concerns preprocessing is the calculation of the vertical velocity $\dot{\eta}\,(\partial p/\partial \eta)$ (3.3.3). For better accuracy it is currently calculated on a reduced Gaussian grid where no subarea extractions can be done with the ECMWF EMOSLIB. This means that the variables required for the calculation must be retrieved globally which increases the amount of data, and therefore, extends the overall process time to gain the forecasts. The problem that concerns the vertical velocity arises from the different coordinate systems. A possibility could be to rewrite the FLEXPART preprocessor but this implicates other issues. Then, many different versions of FLEXPART will be developed which makes it hard to keep the model up-to-date. This circumstance is not desirable, especially for an operational system. Another approach could deal with the reduction of transferred data, for example, by setting up a forecast system directly at a meteorological centre. This, in turn, includes other problems like financial costs and software incompatibilities.

Overall, the presented results show that informations gained from the horizontal wind fields are meaningful and that reduced ensembles based on this informations can be used for practical dispersion applications. For weather regimes that are appropriate for the usage of the proposed concept, a minimum number of 5 to 7 representative members can be recommended. The three investigated clustering algorithms perform slightly different for the examined scenarios. However, K-means and Ward's method tend to be in the upper range according to the validation methods and can be recommended.

A Model settings

This appendix contains the three important configuration files (COMMAND, OUTGRID, RELEASES) of FLEXPART for the test scenario Eyjafjallajökull 2013-03-14. The first one provides general settings, the second one defines the FLEXPART output grid and the last one describes the source.

A.1 General settings - COMMAND file

```
*******************************************************************************
*                                                                             *
*      Input file for the Lagrangian particle dispersion model FLEXPART       *
*                           Please select your options                        *
*                                                                             *
*******************************************************************************

1. __                3X, I2
    1
   LDIRECT           1 FOR FORWARD SIMULATION, -1 FOR BACKWARD SIMULATION

2. ________ ______   3X, I8, 1X, I6
   20130314 000000
   YYYYMMDD HHMISS   BEGINNING DATE OF SIMULATION

3. ________ ______   3X, I8, 1X, I6
   20130317 210000
   YYYYMMDD HHMISS   ENDING DATE OF SIMULATION

4. _____             3X, I5
   10800
   SSSSS             OUTPUT EVERY SSSSS SECONDS

5. _____             3X, I5
   10800
   SSSSS             TIME AVERAGE OF OUTPUT (IN SSSSS SECONDS)

6. _____             3X, I5
     900
   SSSSS             SAMPLING RATE OF OUTPUT (IN SSSSS SECONDS)

7. _________         3X, I9
   999999999
   SSSSSSSSS         TIME CONSTANT FOR PARTICLE SPLITTING (IN SECONDS)

8. _____             3X, I5
     900
   SSSSS             SYNCHRONISATION INTERVAL OF FLEXPART (IN SECONDS)

9.  ---.--           4X, F6.4
     -5.0
    CTL               FACTOR, BY WHICH TIME STEP MUST BE SMALLER THAN TL

10. ---              4X, I3
      4
    IFINE             DECREASE OF TIME STEP FOR VERTICAL MOTION BY FACTOR IFINE

11. -                4X, I1
    1
    IOUT              1 CONCENTRATION (RESIDENCE TIME FOR BACKWARD RUNS) OUTPUT,
                      2 MIXING RATIO OUTPUT, 3 BOTH,4 PLUME TRAJECT., 5=1+4

12. -                4X, I1
    0
```

```
    IPOUT            PARTICLE DUMP: 0 NO, 1 EVERY OUTPUT INTERVAL, 2 ONLY AT END

13. _                4X, I1
    1
    LSUBGRID         SUBGRID TERRAIN EFFECT PARAMETERIZATION: 1 YES, 0 NO

14. _                4X, I1
    1
    LCONVECTION      CONVECTION: 1 YES, 0 NO

15. _                4X, I1
    0
    LAGESPECTRA      AGE SPECTRA: 1 YES, 0 NO

16. _                4X, I1
    0
    IPIN             CONTINUE SIMULATION WITH DUMPED PARTICLE DATA: 1 YES, 0 NO

17. _
    0                4X,I1
    IOFR             IOUTPUTFOREACHREL CREATE AN OUPUT FILE FOR EACH RELEASE LOCATION:
                     1 YES, 0 NO

18. _                4X, I1
    0
    IFLUX            CALCULATE FLUXES: 1 YES, 0 NO

19. _                4X, I1
    0
    MDOMAINFILL      DOMAIN-FILLING TRAJECTORY OPTION: 1 YES, 0 NO, 2 STRAT. O3 TRACER

20. _                4X, I1
    1
    IND_SOURCE       1=MASS UNIT , 2=MASS MIXING RATIO UNIT

21. _                4X, I1
    1
    IND_RECEPTOR     1=MASS UNIT , 2=MASS MIXING RATIO UNIT

22. _                4X, I1
    0
    MQUASILAG        QUASILAGRANGIAN MODE TO TRACK INDIVIDUAL PARTICLES: 1 YES, 0 NO

23. _                4X, I1
    0
    NESTED_OUTPUT    SHALL NESTED OUTPUT BE USED? 1 YES, 0 NO

24. _                4X, I1
    0
    LINIT_COND       INITIAL COND. FOR BW RUNS:
                     0=NO,1=MASS UNIT,2=MASS MIXING RATIO UNIT
```

A.2 Output grid - OUTGRID file

```
*******************************************************************************
*                                                                             *
*      Input file for the Lagrangian particle dispersion model FLEXPART       *
*                       Please specify your output grid                       *
*                                                                             *
*******************************************************************************

1.  ------.----       4X,F11.4
       -50.0000       GEOGRAFICAL LONGITUDE OF LOWER LEFT CORNER OF OUTPUT GRID
    OUTLONLEFT        (left boundary of the first grid cell - not its centre)

2.  ------.----       4X,F11.4
        30.0000       GEOGRAFICAL LATITUDE OF LOWER LEFT CORNER OF OUTPUT GRID
    OUTLATLOWER       (lower boundary of the first grid cell - not its centre)

3.  -----             4X,I5
       90              NUMBER OF GRID POINTS IN X DIRECTION (= No. of cells + 1)
    NUMXGRID

4.  -----             4X,I5
       50             NUMBER OF GRID POINTS IN Y DIRECTION (= No. of cells + 1)
    NUMYGRID

5.  ------.---        4X,F10.3
         1.000        GRID DISTANCE IN X DIRECTION
    DXOUTLON

6.  ------.---        4X,F10.3
         1.000        GRID DISTANCE IN Y DIRECTION
    DYOUTLAT

7. -----.-           4X, F7.1
    5000.0
    LEVEL 1           HEIGHT OF LEVEL (UPPER BOUNDARY)

8. -----.-           4X, F7.1
    5500.0
    LEVEL 2           HEIGHT OF LEVEL (UPPER BOUNDARY)

9. -----.-           4X, F7.1
    6000.0
    LEVEL 3           HEIGHT OF LEVEL (UPPER BOUNDARY)

10. -----.-           4X, F7.1
     6500.0
    LEVEL 4           HEIGHT OF LEVEL (UPPER BOUNDARY)

11. -----.-           4X, F7.1
     7000.0
    LEVEL 5           HEIGHT OF LEVEL (UPPER BOUNDARY)

12. -----.-           4X, F7.1
    7500.0
    LEVEL 6           HEIGHT OF LEVEL (UPPER BOUNDARY)

13. -----.-           4X, F7.1
    8000.0
    LEVEL 7           HEIGHT OF LEVEL (UPPER BOUNDARY)

14. -----.-           4X, F7.1
    8500.0
    LEVEL 8           HEIGHT OF LEVEL (UPPER BOUNDARY)

15. -----.-           4X, F7.1
     9000.0
    LEVEL 9           HEIGHT OF LEVEL (UPPER BOUNDARY)

16. -----.-           4X, F7.1
```

```
     9500.0
    LEVEL 10           HEIGHT OF LEVEL (UPPER BOUNDARY)

17. -----.-           4X, F7.1
    10000.0
    LEVEL 11           HEIGHT OF LEVEL (UPPER BOUNDARY)

18. -----.-           4X, F7.1
    10500.0
    LEVEL 12           HEIGHT OF LEVEL (UPPER BOUNDARY)

19. -----.-           4X, F7.1
    11000.0
    LEVEL 13           HEIGHT OF LEVEL (UPPER BOUNDARY)

20. -----.-           4X, F7.1
    11500.0
    LEVEL 14           HEIGHT OF LEVEL (UPPER BOUNDARY)

21. -----.-           4X, F7.1
    12000.0
    LEVEL 15           HEIGHT OF LEVEL (UPPER BOUNDARY)

22. -----.-           4X, F7.1
    12500.0
    LEVEL 16           HEIGHT OF LEVEL (UPPER BOUNDARY)

23. -----.-           4X, F7.1
    13000.0
    LEVEL 17           HEIGHT OF LEVEL (UPPER BOUNDARY)

24. -----.-           4X, F7.1
    13500.0
    LEVEL 18           HEIGHT OF LEVEL (UPPER BOUNDARY)

25. -----.-           4X, F7.1
    14000.0
    LEVEL 19           HEIGHT OF LEVEL (UPPER BOUNDARY)
```

A.3 Source of volcanic ash - RELEASES file

```
*****************************************************************************
*                                                                           *
*   Input file for the Lagrangian particle dispersion model FLEXPART        *
*                        Please select your options                         *
*                                                                           *
*****************************************************************************
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
  1
---                         i3    Total number of species emitted

 12
---                         i3    Index of species in file SPECIES

=============================================================================
20130314 000000
-------- ------             i8,1x,i6 Beginning date and time of release

20130317 000000
-------- ------             i8,1x,i6 Ending date and time of release

 -19.6200
----.----                   f9.4  Longitude [DEG] of lower left corner

  63.6300
----.----                   f9.4  Latitude [DEG] of lower left corner

 -19.6200
----.----                   f9.4  Longitude [DEG] of upper right corner

  63.6300
----.----                   f9.4  Latitude [DEG] of upper right corner

        2
---------                   i9    1 for m above ground, 2 for m above sea level

 6000.00
-----.---                   f10.3 Lower z-level (in m agl or m asl)

14000.00
-----.---                   f10.3 Upper z-level (in m agl or m asl)

  1440000
---------                   i9    Total number of particles to be released

6.1600E+10
-.----E--                   e9.4  Total mass emitted

VOLCANO_1
----------------------------------------   character*40 comment
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

Bibliography

Arnold, D., 2013: FLEXPART training course 2013. URL `http://www.flexpart.eu/downloads/28`, last visited 2013-06-21.

Barkmeijer, J., R. Buizza, E. Källén, F. Molteni, R. Mureau, T. N. Palmer, S. Tibaldi, and J. Tribbia, 2012: 20 years of ensemble prediction at ECMWF. *ECMWF Newsletter*, **134**, 16–32.

Beaver, S. and A. Palazoglu, 2006: Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality. *Journal of Applied Meteorology and Climatology*, **45**, 1710–1726.

Buizza, R., 1997: The singular vector approach to the analysis of perturbation growth in the atmosphere. Ph.D. thesis, University of London.

Buizza, R. and T. N. Palmer, 1998: Impact of ensemble size on ensemble precipitation. *Monthly Weather Review*, **126**, 2503–2518.

Burkhart, J. F., 2011: The pflexible module. URL `http://niflheim.nilu.no/~burkhart/pflexpart/pflexible.html`, last visited 2013-06-30.

Challa, V. S., et al., 2008: Sensitivity of atmospheric dispersion simulations by HYSPLIT to the meteorological predictions from a meso-scale model. *Environmental Fluid Mechanics*, **8**, 367–387.

Cheng, X. and J. M. Wallace, 1993: Cluster analysis of the northern hemisphere wintertime 500-hpa height field: Spatial patterns. *Journal of the Atmospheric Sciences*, **50**, 2674–2696.

Dando, P., 2013: Interpolation - Introduction and basic concepts. URL `http://www.ecmwf.int/services/computing/training/material/interpolation.pdf`, last visited 2013-06-19, ECMWF EMOSLIB, computer user training course 2013.

ECMWF Directorate, 2012: Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, **133**, 11–13.

Ehrendorfer, M., 1994: The liouville equation and its potential usefulness for the prediction of forecast skil. part i: Theory. *Monthly Weather Review*, **122**, 703–713.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteorologische Zeitschrift*, **6**, 147–183.

Farrell, B., 1982: The initial growth of disturbances in a baroclinic flow. *Journal of the Atmospheric Sciences*, **39**, 1663–1686.

Ferranti, L. and S. Corti, 2011: New clustering products. *ECMWF Newsletter*, **127**, 6–11.

Forgy, E. W., 1965: Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics*, **21**, 768–769.

Galmarini, S., et al., 2004: Ensemble dispersion forecasting-Part I: Concept, approach and indicators. *Atmospheric Environment*, **38**, 4607–4617.

Gordon, A. D., 1999: *Classification*. Chapman & Hall/CRC.

Gower, J. C., 1966: Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Grimit, E. P. and C. F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Monthly Weather Review*, **135**, 203–221.

Haimberger, L., 2009: Ecmwfdata v3.0 update. Tech. rep., Institut für Meteorologie und Geophysik.

Hartigan, J. A. and M. A. Wong, 1979: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 100–108.

Holton, J. R., 2004: *An introduction to dynamic meteorology*, chap. Introduction, 20–21. El.

Hubert, L. and P. Arabie, 1985: Comparing partitions. *Journal of Classification*, **2**, 193–218.

Jain, A. K. and R. C. Dubes, 1988: *Algorithms for clustering data*, chap. Cluster validity. Prentice Hall, Englewood Cliffs, New Jersey.

James, P., 2000: De-accumulation of ECMWF surface flux data. URL `http://zardoz.nilu.no/~andreas/flextra/ecmwf_extr.html`, last visited 2013-06-18, Software documentation.

Kaufman, L. and P. J. Rousseeuw, 1987: *Statistical data analysis based on the L1-norm and related methods*, chap. Clustering by means of medoids, 405–416. Elsevier.

Kaufman, L. and P. J. Rousseeuw, 2005: *Finding groups in data*. Wiley-Interscience.

Lee, J. A., W. C. Kolczynski, T. C. McCandless, and S. Haupt, 2012: An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Monthly Weather Review*, **140**, 2270–2286.

Lloyd, S. P., 1957, 1982: Least squares quantization in PCM. *Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory*, **28**, 128–137.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281–297.

Marida, K. V., J. T. Kent, and J. M. Bibby, 1997: *Multivariate analysis*, chap. Multidimensional scaling, 394–410. Academic Press.

Marzban, C. and S. Sandgathe, 2005: Cluster analysis for verification of precipitation fields. *Weather and Forecasting*, **21**, 824–838.

Meila, M., 2007: Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, **98**, 873–895.

Milligan, G. W. and M. C. Cooper, 1986: A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, **21**, 441–458.

Mooi, E. and M. Sarstedt, 2011: *A concise guide to market research*, chap. Cluster analysis, 237–284. Springer-Verlag Berlin Heidelberg.

Morissette, L. and S. Chartier, 2013: The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, **9**, 15 pp.

Oxford Economics, 2010: The economic impacts of of air travel restrictions due to volcanic ash, report prepared for airbus. URL `http://www.airbus.com/company/environment/documentation/?docID=10262&eID=dam_frontend_push`, last visited 2013-06-22.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bulletin of the American Meteorological Society*, **74**, 49–65.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, **63**, 71–116.

Rand, W. M., 1971: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Rousseeuw, P. J., 1987: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematrics*, **20**, 53–65.

Schoenberg, I. J., 1935: Remarks to Maurice Fréchet's article 'Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert'. *Annals of Mathematics (2nd Series)*, **36**, 724–732.

Seibert, P., A. Frank, and H. Formayer, 2007: Synoptic and regional patterns of heavy precipitation in Austia. *Theoretical and Applied Climatology*, **87**, 139–153.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **612**, 3079–3102.

Simmons, A. J. and D. M. Burridge, 1981: An energy and momentum conserving finite difference scheme and hybrid vertical coordinates. *Monthly Weather Review*, **109**, 758–766.

Smithsonian Institution, 2013: Global volcanism program. URL `http://www.volcano.si.edu/index.cfm`, last visited 2013-07-09.

Steinhaus, H., 1956: Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Science*, **4**, 801–804.

Stohl, A., M. Hittenberger, and G. Wotawa, 1998: Validation of the Lagrangian particle dispersion model FLEXPART against large scale tracer experiment data. *Atmospheric Environment*, **32**, 4245–4264.

Stohl, A., H. Sodemann, S. Eckardt, P. Frank, A. Seibert, and G. Wotawa, 2010: The Lagrangian particle dispersion model FLEXPART version 8.2. URL `http://www.flexpart.eu/downloads/26`, last visited 2013-06-21.

Torgerson, W. S., 1952: Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**, 401–419.

VAST technical proposal, 2012: Technical proposal to the european space agency in response to: Enhancing use of eo in volcanic monitoring and forecasting - volcanic ash strategic-initiative team. Revision 1, NILU, FMI, NUIG, ZAMG, S&T Corp AS.

Ward, J. H., 1963: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Wernli, H., 2012: Wetter, Chaos und probabilistische Wettervorhersagen. *promet*, **37, Nr. 3/4**, 3–11.

Wetter3 archive, 2013: GFS weather chart archive. URL `http://www.wetter3.de/Archiv/`, last visited 2013-07-18.

Wilks, D. S., 2006: *Statistical methods in the Atmospheric Sciences*, chap. Cluster Analysis, 549–560. Elsevier.

Wishart, 1969: An algorithm for hierarchical classifications. *Biometrics*, **25**, 165–170.

Young, G. and A. S. Householder, 1938: Discussion of a set of points in terms of mutual distances. *Psychometrika*, **3**, 19–22.

Acknowledgements

First of all, I would like to give special thanks to Prof. Leopold Haimberger for supervising this thesis, his great support, the useful comments and remarks, the good communication and the basic subroutines for preprocessing the meteorological data. I also would like to express my gratitude to my supervisor at ZAMG, Gerhard Wotawa, for offering me this interesting thesis topic, giving me the chance to work within the VAST project and the possibility to join two very appealing training courses. At this point I also want to thank the institution ZAMG for employment, the provided meteorological data and financial support for the training courses. I also owe Delia Arnold a debt of gratitude for several discussion sessions that really pushed this thesis forward, her support about questions related to the dispersion model FLEXPART and for always motivating me.

Furthermore, I want to thank Christian Maurer for his contributions to the preprocessing routines. Special thanks are given to Rainer Stowasser for technical and administrative support. I also want to thank Matthias Langer for retrieving the meteorological data and for discussing theoretical and technical questions with me. Moreover, I want to thank Paul Skomorowski for sharing his experience about the Fortran library EMOSLIB which saved me a lot of time. For very valuable comments during presentations of my thesis I want to thank Prof. Petra Seibert. Additionally, I want to thank John Burkhart for providing his Python plotting routines to the FLEXPART community. I also owe special thanks to Maria Hahn-Hahn for proofreading this thesis.

In the end, I am deeply grateful to my loved ones. Without the encouragement and the financial support of my family, especially my parents Gertrude and Josef, throughout my studies and life, it would have been impossible for me to finish my graduate education seamlessly. Furthermore, I would like to acknowledge my girlfriend Sandra for listening to my thoughts and concerns, her patience and always being there for me.

Curriculum vitae

Personal Data

Name:	Robert Klonner, BSc
Place of Birth:	Zwettl, Lower Austria, Austria
Current Residence:	Vienna, Austria

Education

Since 2011	Study of Meteorology (Masters course)
2008 - 2011	Study of Meteorology (Bachelors course)
2002 - 2007	Secondary technical school
1992 - 2002	Primary and lower secondary school

Work Experience

11/2012 - 04/2013	Master thesis, Meteorologist Employment within the Volcanic Ash Strategic-initiative Team (VAST) project at ZAMG
08/2012 - 09/2012	Placement, Meteorologist, Software engineer Environmental department of ZAMG
05/2012 - 07/2012	Placement, Meteorologist, Software engineer Climatological department of ZAMG
03/2012 - 06/2012	Tutor at a meteorological - climatological excursion to Death Valley University of Vienna
08/2011 - 09/2011	Placement, Meteorologist, Software engineer Climatological department of ZAMG
03/2011 - 06/2011	Tutor for Micrometeorology University of Vienna