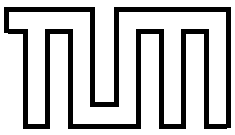


TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR INFORMATIK

LEHRSTUHL FÜR COMPUTER GRAPHIK UND VISUALISIERUNG

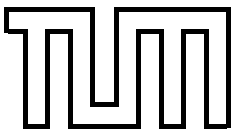
Visual Analysis of Statistical Aspects of Ensemble Data

Mihaela Jarema

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.



TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR INFORMATIK

LEHRSTUHL FÜR COMPUTER GRAPHIK UND VISUALISIERUNG

Visual Analysis of Statistical Aspects of Ensemble Data

Mihaela Jarema

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. B. Rost
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. R. Westermann
	2.	Univ.-Prof. Dr. L. Linsen, Jacobs University Bremen

Die Dissertation wurde am 28.09.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 06.12.2016 angenommen.

Abstract

Complex systems are often approximated using computational models, and simulated to gain more insight into their behavior and forecast their future states. In this context, the produced forecasts are uncertain for two main reasons. Firstly, the constructed models are not ideal representations of the governing physics. Secondly, perfect measurements are not possible and the evolution of the system may be very sensitive to the choice of initial conditions, so that a slight change in the initial conditions may lead to considerably different forecasts. Consequently, forecasting the future state of a system is spatially and temporally limited.

A technique to evaluate and quantify the amount of uncertainty is *ensemble forecasting*. Here, the intrinsic uncertainty is estimated by generating a collection (or ensemble) of forecasts starting from slightly perturbed initial conditions. The amount of uncertainty present depends on the extent to which the forecasts agree with one another: the more the individual forecasts deviate, the greater the uncertainty is. Nevertheless, even in cases of significantly dissimilar behavior, the manner in which the ensemble members differ from one another may still reduce the uncertainty underlying the future state of the system, for instance, by indicating whether certain events are more or less likely to occur. Obtaining this kind of insight is, however, difficult, since ensemble data sets usually comprise multidimensional, multivariate, and multivalue data.

This thesis proposes several novel approaches to model scalar and vector ensembles statistically, in order to capture variability aspects and convey these effectively. For scalar ensembles, the variability in location and shape of prominent features is addressed by analyzing positional and type stability characteristics of the associated critical points. This is achieved by propagating the uncertainty from the scalar data to derived operators like the gradient and the Hessian matrix. For vector ensembles, two frameworks are proposed to allow a comparative visual analysis of local and global variability aspects. Both approaches are based on approximating distributions using mixture models and applying these fitted components to visualize either local directional or global transport trends, as well as their associated variability. The mixture models are also employed to derive variability measures for a number of purposes, such as distinguishing between locations with similar and dissimilar behavior, or clustering ensemble members based on their similarity across the domain. The benefits of the proposed approaches are demonstrated on a number of synthetic and real world weather forecast ensembles.

Zusammenfassung

Komplexe Systeme werden oft durch Computermodelle approximiert und simuliert, um Einblicke in deren Verhalten zu gewinnen und deren künftigen Zustand vorherzusagen. Hierbei sind die erzeugten Vorhersagemodelle in zweierlei Hinsicht unsicher. Erstens können die Computermodelle die zugrundeliegende Physik nicht genau repräsentieren. Zweitens sind perfekte Messungen unmöglich und Systeme reagieren sensibel auf die Wahl des Anfangszustands, sodass eine kleine Veränderung der Anfangswerte zu einer völligen Veränderung der Prognose führen kann. Dementsprechend sind Vorhersagen räumlich und zeitlich begrenzt.

Ensemblevorhersage ist eine Methode um die inhärente Unsicherheit zu evaluieren und zu quantifizieren, indem man ein ganzes Ensemble von Vorhersagen generiert, wobei jedes Ensemble Member basierend auf leicht veränderten Anfangswerten erzeugt wird. Die Prognoseunsicherheit hängt davon ab, wie sehr sich die Member voneinander unterscheiden: je unähnlicher sich die Member verhalten, desto größer ist die Unsicherheit. Nichtdestoweniger kann auch in solchen Fällen, in denen die Unähnlichkeit signifikant ist, die Art und Weise wie sich die Ensemble Member voneinander unterscheiden, die Unsicherheit der Vorhersage immer noch reduzieren, indem angezeigt wird, ob bestimmte Vorkommnisse mehr oder weniger wahrscheinlich sind. Solch einen Einblick zu gewinnen ist allerdings schwierig, da Ensembles normalerweise multidimensionale, multivariate und mehrwertige Datensätze enthalten.

Diese Arbeit leistet neuartige Beiträge in der statistischen Modellierung von Skalarfeld- und Vektorfeld-Ensembles, um Besonderheiten hinsichtlich der Variabilität herauszufinden und diese effizient zu übermitteln. Für Skalarfeld-Ensembles wird die räumliche und strukturelle Variabilität von prominenten Features behandelt, indem räumliche und typbedingte Besonderheiten hinsichtlich der Stabilität der assoziierten kritischen Punkte analysiert werden. Zu diesem Zweck wird die Unsicherheit von Skalarfeldern bezüglich abgeleiteter Operatoren wie dem Gradienten und der Hesse-Matrix propagiert. Für Vektorfeld-Ensembles werden zwei Arten von Frameworks eingeführt, um eine komparative visuelle Analyse von lokalen und globalen Aspekten zu ermöglichen. Beide Vorgehensweisen basieren darauf, Verteilungen anhand von Mischverteilungen zu approximieren und diese Mischverteilungen dazu einzusetzen, um sowohl lokale Richtungs- als auch globale Transporttrends, sowie die damit assoziierte Variabilität zu visualisieren. Mischverteilungen werden auch dazu eingesetzt, um zwischen Regionen mit ähnlichem oder unähnlichem Verhalten zu unterscheiden, oder um ähnliche Member zu gruppieren. Die Vorteile der eingeführten Methoden werden anhand von synthetischen und realen Ensembles aus der Wettervorhersage demonstriert.

Acknowledgments

I would like to express my gratitude to all people who made this thesis possible.

First of all, I would like to acknowledge my adviser, Prof. Dr. Rüdiger Westermann, for giving me the opportunity to embark on this research journey and steering me in the right direction each time.

I would also like to acknowledge all my current and former colleagues at the Chair of Computer Graphics and Visualization, for interesting discussions on and off research topics. They helped me become a better researcher and, more importantly, a better person.

Finally, I am grateful to Florian Ferstl, Tobias Pfaffelmoser, and Marc Rautenhaus for providing me with data sets.

This thesis was funded by the European Union under the ERC Advanced Grant 291372: Safer-Vis – Uncertainty Visualization for Reliable Data Discovery.

Contents

1 Introduction **1**

1.1 Contribution 5

1.2 Thesis Outline 6

1.3 List of Publications 6

2 Approaches to Uncertainty and Ensemble Visualization **9**

2.1 Visualizing Uncertainty 11

2.2 Ensemble Visualization 13

2.3 Evaluation 19

3 Theoretical Background **21**

3.1 Basics of Uncertainty Representation 21

3.2 Statistical Modeling of Distributions 25

3.2.1 Gaussian Mixture Models 25

3.2.2 Non-parametric Models 29

3.3 Multivariate Random Variables 32

4 Visualizing the Stability of Critical Points in Uncertain Scalar Fields **39**

4.1 Critical Points in Ensembles 41

4.1.1 Confidence Intervals 42

4.1.2 Confidence Intervals for Gradients . . . 44
4.1.3 Confidence Intervals for the Hessian Matrix . . . 45
4.1.4 Indicator Functions . . . 46
4.2 Visualization . . . 52
4.2.1 Visualization of Positional Indicators . . . 53
4.2.2 Visualization of Type Indicators . . . 56
4.3 Validation . . . 59
4.4 Further Results . . . 62
4.5 Conclusion and Future Directions . . . 68

5 Comparative Visual Analysis of Vector Field Ensembles 69
5.1 Modeling Directional Data . . . 71
5.1.1 Modeling 2D Directional Data . . . 72
5.1.2 Modeling 3D Directional Data . . . 76
5.2 Comparative Analysis of Ensemble Members Based on Fitted Mixture Models . 78
5.2.1 Local Comparative Measures . . . 78
5.2.2 Global Comparative Measures . . . 81
5.2.3 Comparing Mixture Models . . . 82
5.3 Visual Analysis of Vector Ensembles . . . 83
5.3.1 Member Similarity Visualization . . . 83
5.3.2 Similarity Visualization over the Spatial Domain . . . 85
5.3.3 Querying-driven Visualization . . . 90
5.3.4 3D Extension Outlook . . . 91
5.4 Further Results . . . 99
5.4.1 Similarity Analysis . . . 99
5.4.2 Implementation, Performance and Scalability . . . 107
5.5 Conclusion and Evaluation . . . 109

6 Comparative Visual Analysis of Transport Variability in Flow Ensembles 111
6.1 Transport Variability . . . 113
6.1.1 Pairwise Dissimilarity Analysis . . . 115
6.1.2 Transport Variability over the Ensemble Domain . . . 118
6.1.3 Flow-based Similarity . . . 122
6.2 Visual Analysis of the Transport Variability . . . 122

6.2.1 Flow Variability over the Spatial Domain . . . 122
6.2.2 Flow Variability over Subdomains . . . 124
6.2.3 Flow-based Similarity Visualization . . . 126
6.3 Results . . . 127
6.3.1 Transport Variability Analysis . . . 128
6.3.2 Implementation, Performance Analysis, and Scalability . . . 135
6.4 Conclusion, Evaluation, and Future Work . . . 135

7 Conclusion and Future Directions **137**

Bibliography **141**

1 Introduction

In many scientific disciplines, experimental observations are often complemented by numerical simulations, allowing scientists to gain further insight into the behavior of various systems and predict their future states. For instance, in meteorology, the future state of the atmosphere is predicted numerically based on non-linear differential equations (Palmer (2000)).

A numerical simulation is deterministic: given a certain initial state, rerunning the same computational model will not change the result. However, as the computational models used in simulations cannot completely represent the governing physics and perfectly accurate measurements are not possible, numerical simulations are inherently uncertain.

In other words, any single numerical simulation – albeit potentially the best that could be obtained with the state-of-the-art models and computational resources – is only one of the infinite states that could actually occur. Thus, simulations need to be probabilistic in order to include information on the uncertainty underlying the forecasting process.

A technique to evaluate and quantify the amount of uncertainty is *ensemble forecasting*, whereby the intrinsic uncertainty is estimated by running multiple simulations (ensemble members), started from slightly perturbed initial conditions and using potentially different models (Wilks (2011)).

Reputable organizations have operational ensemble forecast systems to produce ensemble forecasts. For instance, the European Centre for Medium-Range Weather Forecasts (ECMWF), which operates one of the largest supercomputer sites in Europe and is worldwide renowned for the quality of its numerical weather predictions, has the Ensemble Prediction System

(EPS), which produces forecasts twice daily. The ensembles consist each of 51 forecasts: a control run – the unperturbed forecast run from the operational ECWMF analysis without stochastics – and another 50 perturbed forecasts[1]. Thus, instead of using a single deterministic forecast, a probabilistic forecast is performed, where the input parameters of the ensemble simulations represent the probability distribution of the initial state.

Theoretically, this would lead to an ensemble of future states that characterize the forecast uncertainty: low uncertainty when ensemble members behave similarly and high uncertainty when they exhibit different behavior.

Practically, however, the initial distribution is not perfectly known (for instance, because of imprecise measurements or the lack thereof) and the initial conditions are not random samples of the initial distribution[2]. Moreover, errors are introduced in the forecasting process as well, caused, for example, by the simplifications used in modeling systems, the methods employed for numerical integration, the discretization resolutions used, etc. This means that every ensemble forecast produced is just an approximation of the forecast distribution, meant to guide users in assessing the evolution of a system and the likelihood of certain events.

Estimating the inherent uncertainty in a forecast by generating a collection of members produces huge amounts of multidimensional, multivariate, and multivalue data. A straightforward manner to summarize this uncertainty is by computing the ensemble average and using various spread measures. The ensemble average, while often physically implausible, is useful to indicate the main features present in most ensemble members. When the ensemble is characterized by several states, however, its behavior is no longer well-represented by the ensemble average. Spread measures, such as the standard deviation, indicate how much the individual members tend to deviate from the ensemble average. The uncertainty can then be assessed as low or high depending on whether the spread is correspondingly low or high.

Other approaches of gaining insight into the massive amount of complex data is by using graphical displays. Different approaches used, for instance, by ECMWF include (Wilks (2011))

[1] For more information on ECMWF's operational forecast system, please refer, for example, to Leutbecher and Palmer (2008).

[2] The initial conditions are not chosen randomly because the number of ensemble members that may be produced with an appropriate spatial resolution during a reasonable time period is limited. There are therefore several established methods of choosing the initial ensemble members (see Palmer (2000)).

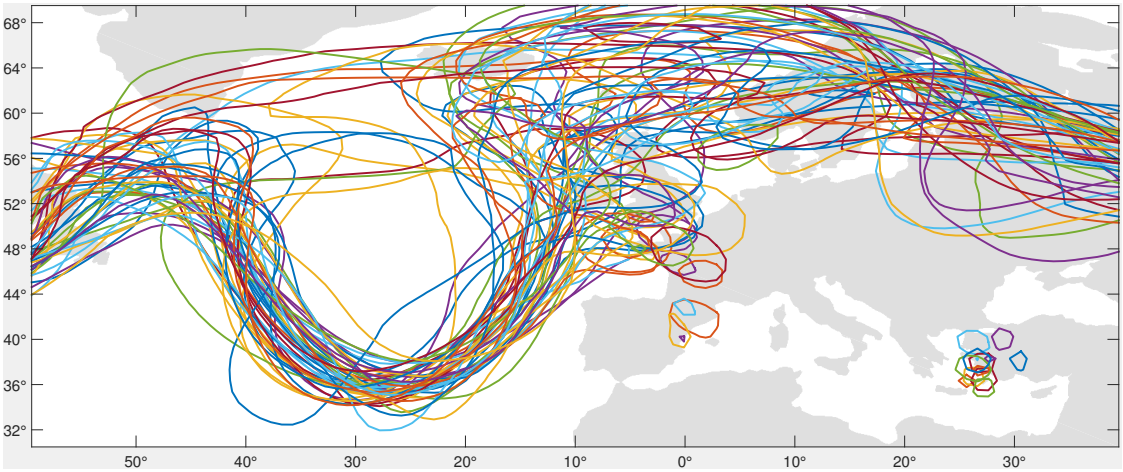


Figure 1.1: Spaghetti plot for the 5500-m contours of the 500 hPa geopotential height field of an ECMWF ensemble.

- *spaghetti plots* (see Figure 1.1 for an example) superpose selected contours of the ensemble members; the visual display may suffer from clutter; furthermore, a limited number of spaghetti plots can be visualized concurrently

- *stamp maps* (see Figure 1.2 for an example) display miniaturized versions of all ensemble members, where just the main features can be conveyed; a simultaneous interpretation of the possible outcomes and the degree to which ensemble members differ is not easy

- *summary plots* show various summary statistics, such as the simultaneous display of the average and standard deviation fields; for single locations, *ensemble meteograms* show time series of box plots for selected variables

- *plume graphs* convey the temporal evolution of the distributions of individual variables

- *ensemble relative frequency maps* show relative frequencies of occurrence for dichotomous events

Most of these techniques are effective only for moderate ensemble sizes or pertain to individual locations. To condense information and reduce the data complexity, clustering algorithms can be used to group together similar members and employ graphical displays for cluster representatives only. Clustering suffers nevertheless the drawback that potentially important information may be discarded.

Figure 1.2: Stamp maps for an ECMWF ensemble showing the geopotential height fields of the perturbed members.

1.1 Contribution

Ensemble data sets pose numerous challenges, due to the sheer amount of multidimensional, multivariate, and multivalue data. Analyzing and interpreting simulation results are thus demanding tasks, and often require new concepts and visual abstractions to convey the complex data and its associated uncertainty in an intuitive manner. This thesis addresses several challenges in the quantification and effective integration into visualization of statistical aspects of scalar and vector field ensembles.

For scalar ensembles, interesting features often relate to critical points, since, depending on their positions and types, critical points indicate salient surface components and their topological changes: surface components emerge or vanish at minima or maxima, join or split at saddle points. However, as ensemble data is inherently affected by uncertainty, the locations and types of critical points vary and can no longer be predicted accurately. Therefore, in this thesis, novel measures are derived to assess the likelihood that critical points of certain types occur in scalar field ensembles. By deriving confidence regions for gradient vector fields, domain points are classified according to whether critical points are likely to occur around certain locations or not; confidence intervals for operators of Hessian matrices are used to indicate whether specific types of critical points should be expected at locations where critical points are likely to occur.

Ensembles of vector fields are much more challenging to explore than scalar ensembles, since even the mere visualization of a vector requires more visual attributes than that of a scalar quantity. Two novel visual analysis approaches are presented to support a comparative exploration of vector-valued ensembles. The first approach addresses local directional data in vector field ensembles; distributions of directional data are approximated using mixture models instead of being summarized by mean and standard deviation. The fitted mixture models are employed both to construct glyphs indicating the main features of potentially multimodal directional distributions and to assess the degree of similarity between ensemble members. This enables users to identify rapidly the spatial locations with high angular variability. Then, the proposed similarity measures are used to cluster ensemble members hierarchically, based on their local angular similarity throughout the domain. A graphical user interface comprising multiple linked views enables a simultaneous visualization of aggregated global and detailed local variations.

The second visual analysis approach addresses the transport variability of vector field ensembles, focusing on identifying the locations where and the time intervals over which a simulation is more or less reliable. To this purpose, distributions of tracer particles are approximated using Gaussian Mixture Models (GMMs); this allows applying the Mahalanobis distance on the identified modes to analyze the time-varying pairwise dissimilarities of member particles relative to the existent variability instead of a synthetic threshold. These dissimilarities are encoded in enhanced spaghetti plots, revealing the main flow trends and variability at selected locations. Furthermore, a novel graphical abstraction based on the visualization of miniaturized variants of enhanced spaghetti plots in a small-multiples layout enables a concurrent visualization of the transport variability across selected subdomains. The downscaling is performed so as to preserve the relevant flow trends and convey the transport variability. A graphical user interface with multiple linked views allows a simultaneous visualization of global and local transport variations, and offers insight into how similar the transport behavior of the ensemble members is across the domain.

1.2 Thesis Outline

The rest of this thesis is structured into six chapters. Chapter 2 surveys existent techniques and encountered challenges when attempting to define, characterize, and visualize uncertainty and ensemble data. Then, Chapter 3 provides fundamental mathematical notions that lie at the basis of the techniques introduced in the following chapters. Means of investigating various stability aspects of prominent features in ensembles of scalar fields by analyzing their associated critical points are presented in Chapter 4. Visual analysis frameworks to explore the variability in vector field ensembles are proposed in Chapter 5 for the directional variability and Chapter 6 for the flow variability. The last chapter presents concluding remarks on the contributions of the proposed approaches, as well as possible directions for future work.

1.3 List of Publications

This thesis comprises both unpublished and published original research work; the latter includes works published in the following peer-reviewed conference papers and journal articles:

- *Mihai, M.*, & Westermann, R. (2014). Visualizing the Stability of Critical Points in Uncertain Scalar Fields. J. Computers & Graphics, 41:13–25 (Chapter 4 is partly based on this paper, ©2014 Elsevier.)
- *Jarema, M.*, Demir, I., Kehrer, J., & Westermann, R. (2015). Comparative Visual Analysis of Vector Field Ensembles. Proc. IEEE VAST 2015, 81–88 (Chapter 5 is partly based on this paper, ©2015 IEEE.)
- *Jarema, M.*, Kehrer, J., & Westermann, R. (2016). Comparative Visual Analysis of Transport Variability in Flow Ensembles. J. WSCG, 24(1):25–34 (Chapter 6 is partly based on this paper, ©2016 WSCG.)

2 Approaches to Uncertainty and Ensemble Visualization

Already as early as 1961, the celebrated John W. Tukey was promoting the use of graphical methods for an exploratory data analysis (Tukey (1962)). Later, Tukey (1977) proposed the *box plot* to obtain fast visual insight into main characteristics of a data set, such as extreme values or quartiles dividing the ordered data set into quarters. Box plots have a simple design (see Figure 2.1 for an example of a traditional box plot) and can help identify outliers (data points located far from the main cluster) or compare distributions of numerical data. An even older example of a graphical tool to examine data sets is the *histogram* (first use credited to Pearson (1894)), a graphical approximation of the distribution of numerical data.

For a long time now, peer-reviewed publications dealing with experimental or simulation data have incorporated uncertainty aspects in the visual representations of their data, most often in the form of 2D graphs. The relatively new field of scientific visualization (McCormick (1988)), however, has only comparatively recently begun to include uncertainty in 2D and 3D visualizations.

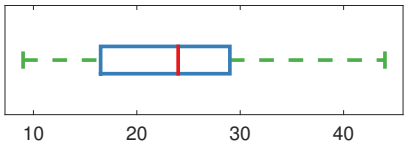


Figure 2.1: A box plot is a simple graphical representation of main characteristics of a data set, such as its extreme values (in this example, min = 9 and max = 44) or its quartiles (in this example, 16.5, 24, and 29).

Johnson and Sanderson (2003) point out the scarcity of 2D and 3D graphical representations that consider uncertainty and argue for their importance. Also, Johnson (2004) includes the visualization of uncertainty in the list of top challenges in scientific visualization.

There are several reasons why incorporating uncertainty in graphical displays is difficult, such as the lack of consensus regarding how to define and represent uncertainty (Pang (2001)), as well as the increase in data amount and dimensionality. Among the few closed definitions, uncertainty is described by Hunter (1999) as the "degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations without caution."

Griethe and Schumann (2006) enumerate a quite heterogeneous list of concepts used when referring to uncertainty, which include the more mathematical notions of *error*, seen as the deviation from a true value, *accuracy*, defined as the size of the interval a value lies in, or *noise*, for the random disturbances affecting the data, but also concepts like *subjectivity*, how subjective the data is, or *non-specificity*, when specific characteristics are absent.

Potter et al. (2012b) distinguish between *epistemic* and *aleatoric* uncertainty. The first is the systematic type of uncertainty, which appears due to the lack of knowledge and missing data that theoretically could be known, but practically is not. The second type is the statistical uncertainty, arising, for instance, when the same experiment is run under the same conditions, but with slightly different results. The rest of the chapter refers primarily to aleatoric or statistical uncertainty, which can be described quantitatively in numerous ways, e.g., using standard deviation, probability density functions, etc.

Not only does uncertainty encompass a whole list of concepts, but it can also affect data in a variety of ways. Pang et al. (1997) review the different sources of uncertainty that may appear during the three main stages of the process of generating a visual representation from raw data – the visualization pipeline. Thus, during *data acquisition*, uncertainty may be introduced, for instance, due to measurement variations or simplifications in numerical models. Further uncertainty is introduced as data is *transformed*, for example, using interpolation, quantization, etc. Finally, uncertainty can also be introduced during the *visualization* step; for instance, rendering isosurfaces requires the use of interpolation, while flow visualization is based on numerical integration techniques.

2.1 Visualizing Uncertainty

Pang et al. (1997) divide approaches of visualizing uncertainty in two main categories: *overloading* considers uncertainty as an additional dimension to the data and visualizes it so that it can be clearly distinguished from the other data dimensions (for instance, using multivalue visualization methods); in contrast, *verity visualization*, e.g., Wittenbrink et al. (1995), uses new visual primitives and abstractions to interpret data and its associated uncertainty holistically.

Techniques for visualizing uncertainty can, nevertheless, be classified according to several other characteristics, for instance value (scalar, vector, tensor, multivariate) and extent (discrete or continuous, or general approach versus application domain (Pang et al. (1997)).

Uncertainty in scalar data is often summarized statistically by quantities such as mean and standard deviation, and encoded graphically along with the data using a variety of techniques (Pang et al. (1997); Griethe and Schumann (2006)), such as:

- modifying various geometry attributes, like color (Hengl (2003); Rhodes et al. (2003); Zehner et al. (2010)), opacity (Djurcilov et al. (2002)), texture (Djurcilov et al. (2002); Rhodes et al. (2003)), etc.

- modifying the geometry itself, e.g., Grigoryan and Rheingans (2004) reveal the uncertainty affecting surfaces by rendering these surfaces using points and displacing them along the surface normals with an amount proportional to the existent uncertainty

- adding geometry, e.g., Zehner et al. (2010) propose adding enveloping surfaces to indicate the confidence volume for extracted isosurfaces

- adding glyphs, e.g., uncertainty in 1D scalar fields is often expressed using error bars or box plots (Potter (2006)); Cedilnik and Rheingans (2000) use procedural annotations to show uncertainty with a minimal conceal of the data; Zehner et al. (2010) emphasize the confidence envelope thickness by adding line glyphs on the extracted isosurface along the surface normals

- using animation, e.g., Brown (2004) proposes visual vibrations techniques to indicate uncertain locations, while Lundström et al. (2007) use probabilistic animation to convey the uncertainty in Direct Volume Rendering

For vector fields, Wittenbrink et al. (1996) present several types of uncertainty vector glyphs to depict the magnitude and angular uncertainty. In addition to uncertainty glyphs, Lodha et al. (1996a) advance also other techniques, such as flow envelopes, animations, twirling batons or rakes. Lodha et al. (1996b) propose incorporating sound when visualizing fluid flow uncertainty, by mapping, e.g., distances between different particle traces to pitch. Pfaffelmoser et al. (2013) propose circular glyphs to describe the uncertainty in mean and orientation of gradients in uncertain 2D scalar fields. Other local methods for uncertain vector fields include semi-Lagrangian texture advection (Botchen et al. (2005, 2006)) and a reaction-diffusion model (Sanderson et al. (2004)). Zuk et al. (2008) propose combining static glyphs with animation to show the uncertainty in bidirectional vector fields. Allendes Osorio and Brodlie (2009) propose adapting the Line Integral Convolution (LIC) algorithm for 2D uncertain steady vectors by modifying the input texture or other visual attributes, i.e., hue or brightness. In the unsteady case, Hlawatsch et al. (2011) introduce flow radar glyphs to visualize with less clutter the temporal evolution of uncertain local directions using static images. To visualize transport uncertainty, Otto et al. integrate particle density functions to determine an uncertain topological segmentation for 2D (Otto et al. (2010)) and 3D (Otto et al. (2011)) normally distributed steady vector fields. To assess the transport of uncertainty for unsteady vector fields, Schneider et al. (2012) use stochastic flow maps, whereas Guo et al. (2016) extend the finite-time Lyapunov exponent (FTLE).

In the case of *bounded uncertainty*, no distributions can be assumed for the data; instead, the only certainty about the data is that the exact value lies within an interval with known lower and upper bounds (Olston and Mackinlay (2002)). Then, the visual outputs should be adjusted, in order to convey the fact that all values within the given interval are equally likely. Olston and Mackinlay (2002) propose applying ambiguation to existent methods such as error bars, for instance by smearing the crisp part between possible boundaries to a graphical fuzz. Another case where statistical summaries like the mean and the standard deviation are not usable is categorical data, where variables take values in a limited number of nominal categories. For such data, Potter et al. (2013) propose the use of entropy to quantify the existent uncertainty, which they visualize using color mapping.

Several of the aforementioned methods for scalar fields, e.g., Brown (2004); Grigoryan and Rheingans (2004); Zehner et al. (2010), indicate not only the amount of uncertainty affecting the data, but also allow inferring on how uncertainty affects specific features in the data, such as level-sets. Pfaffelmoser et al. (2011) investigate the positional and geometrical variability

of level-sets, and, in order to gain insight into potential structural variations, Pfaffelmoser and Westermann (2012) and Pfaffelmoser and Westermann (2013) consider correlation as well. Also to assess the positional uncertainty of level-sets, Pöthkow and Hege (2011) propose using the concept of numerical condition, which they then extend to include spatial correlation (Pöthkow et al. (2011)).

Many other approaches to analyze prominent features stem from topology (for a short survey on scalar topology see Laramee et al. (2007) and Heine et al. (2016), and for overviews of topological methods for both steady and time-dependent vector fields refer to Scheuermann and Tricoche (2005); Laramee et al. (2007); Theisel et al. (2008); Heine et al. (2016)). For scalar fields, Thompson et al. (2011) propose hixels – histograms of scalar values at each domain location – to extract salient topological features of down-sampled data. Wu and Zhang (2013) augment contour trees to visualize several types of uncertainty: the topology variability is shown e.g., by the thickness of the branches, the positional variability of the contours, e.g., by ribbons attached to the contour tree, and the data-level uncertainty, e.g., by graduated circular glyphs.

Examples of aforementioned topological approaches for Gaussian distributed steady vector fields are the works of Otto et al. (Otto et al. (2010, 2011)). Also for Gaussian distributed vector fields, Petz et al. (2012) take spatial correlation into consideration to extract probabilistic local features, such as critical points. Pöthkow and Hege (2013) extend this method to non-parametric uncertain models, such as histograms and kernel density estimates. Bhatia et al. (2011) propose edge-maps as representations of vector fields and generalize the concept of streamlines to streamwaves to compute fuzzy topological decompositions.

For tensor topology (Delmarcelle and Hesselink (1994)), Schultz et al. (2007) use probabilistic fiber tracking on diffusion MRI data to define topological features similar to concepts from flow topology, and propose an algorithm to extract these features.

2.2 Ensemble Visualization

Recently, another paradigm has been gaining popularity for presenting uncertainty in scientific visualization (Brodlie et al. (2012)): instead of describing uncertainty using stochastic (or random) fields, *ensemble visualization* assumes that uncertainty is represented by a set (an ensemble) of possible outcomes.

Ensemble data sets are collections of simulations for the same physical phenomenon or process, run with slightly different input parameter values, initial conditions, numerical models, etc. There are several sources of uncertainty in ensemble data. For instance, the models employed are only approximations of the simulated physical process and, even if the models could perfectly simulate the process, errors in the initial conditions could still not be eliminated.

In this context, Love et al. (2005) introduce *multivalue* data, where a multivalue (or ensemble) data set consists of a collection of n values defined at each location in a given domain. Thus, the data and its associated uncertainty are now given by a set of n fields, each of which is potentially multidimensional and multivariate. Given that even 2D and 3D displays can suffer from cluttering and occlusion, ensemble visualization significantly increases the complexity of the rendering task, since a straightforward extension of any traditional visualization approach, such as isosurface extraction, would now yield n isosurfaces that needed to be visualized in such a manner as to reveal their similarities and dissimilarities.

The uncertainty in the ensemble data at every domain location can be described by a probability density function (PDF). This PDF is typically not known, but it can be approximated using, for instance, a discrete function like a histogram or a continuous one like a kernel density estimator. The uncertainty can also be estimated using parameters like the mean and standard deviation (see Figure 2.2 for an example), as well as other statistical summaries.

In fact, when ensemble data can be described well by a few parameters only, this significantly reduces the amount of data to be visualized. For instance, when data can be assumed to follow a Gaussian or Poisson probabilistic model, Love et al. (2005) propose the use of Tukey's box

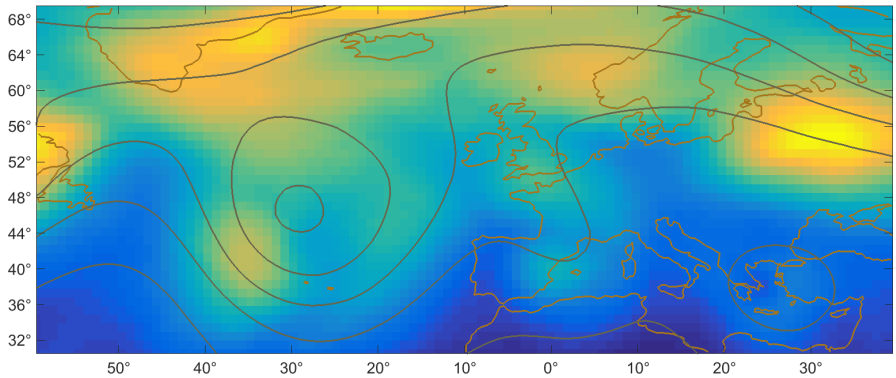


Figure 2.2: Summarizing a scalar ensemble using mean (shown using dark gray contours) and standard deviation (shown using color coding).

plot for a single multivalue datum. For meteorological and climate ensemble data, Potter et al. (2009) display spatial summary views, mapping, e.g., the mean field to color and showing the standard deviation using overlaid contours. Also for weather prediction ensembles, Sanyal et al. (2010) encode uncertainty metrics for weather ensembles using circular glyphs and graduated ribbons. For climate ensembles, Kehrer et al. (2011) propose 2D billboard glyphs based on superellipses, and map statistical aggregates like quartiles to glyph colors and sizes, and skewness to glyph shapes.

The parametric approach of summarizing data just by its mean and standard deviation fails if the data is, for instance, multimodal, because a multimodal distribution may have the same mean and variance as a unimodal one (Love et al. (2005)). Instead, a multimodal distribution can be approximated by a mixture of modes, where each mode in the mixture model can then be summarized by statistical aggregates.

A widely employed method to summarize PDFs is fitting Gaussian Mixture Models (GMMs) to the data. GMMs do not require storing the PDFs, but only three parameters per mode – the mean, standard deviation (or, for higher dimensions, the covariance matrix), and the weight of each mode. Correa et al. (2009) use GMMs in their framework to model, propagate, and aggregate uncertainty in the visual analytics process. Hollister and Pang (2013) apply GMMs to compute interpolations of distributions. Liu et al. (2012) use per voxel GMMs to perform a volume rendering of stochastic fields.

Distributions can also be approximated non-parametrically. For instance, for 1D variables, Kao et al. (2001) propose visualizing entire distributions, by approximating at each 2D location the $1D$ distributions using histograms and visualizing the resulting histogram volume by means of cutting planes. Potter et al. (2010) propose 1D and 2D summary plots that combine traditional box plots, histograms, and also use higher-order descriptive statistics, such as the skewness and kurtosis of a distribution. To approximate distributions more accurately, Kao et al. (2002) use kernel estimators instead of histograms and visualize the 3D density estimate volume via techniques such as probing, cutting planes, extracting isosurfaces, and Direct Volume Rendering. Also for distributions approximated using kernel estimators, Kao et al. (2013) propose visualizing the modality of the distributions and exploring the modes, for example using mode querying and plotting several PDFs side-by-side, overlaid, or directly above the grid cells. Potter et al. (2012a) use difference measures to compare PDFs with an ansatz PDF selected by the user and allow probing PDFs at any physical location. For ensembles of level sets, Whitaker et al. (2013) introduce contour box plots to display statistical

quantities that are conceptually analogous to those shown in conventional box plots. While such methods may give more insight into the shape of the distributions, it is difficult to tailor them for higher dimensions.

To extend traditional approaches to multiple dimensions, Luo et al. (2003) propose the use of various mathematical and procedural operators. For instance, the operator *ToScalar(M)* can be used to reduce distribution data to scalars, which can then be visualized, e.g., by pseudo-coloring. Chen et al. (2015), on the other hand, perform a 2D projection of multidimensional distributions of individual ensemble objects so as to consider the dissimilarities between both the individual means and distributions, and visualize selected multidimensional distributions using ensemble bars – colorbar representations of the deviations of the members.

Other means of assessing the uncertainty of an ensemble without relying on summary statistics like the mean and standard deviation have been proposed. For instance, Gosink et al. (2013) incorporate ground truth observations and propose Bayesian Model Averaging to obtain an aggregate prediction from univariate ensemble data. Bensema et al. (2015) use Hartigan's dip test for unimodality to classify high-variance locations in scalar ensembles as unimodal, bimodal, or neither unimodal, nor bimodal.

Another popular approach to deal with large and complex data is *clustering* (Jain (2010)). Bordoloi et al. (2004) cluster raw ensemble data hierarchically to reduce the amount of information and visualize representative PDFs only. Bruckner and Möller (2010) perform a density-based clustering of time-dependent volume data in physically-based simulations. Reh et al. (2013) use k-means to cluster pores in industrial XCT data to obtain mean objects (MObjects). For time-varying flow ensembles, Hummel et al. (2013) use Minimum Spanning Trees to detect trends in the transport behavior. Ferstl et al. use hierarchical clustering to cluster streamlines (Ferstl et al. (2016a)) and isocontours (Ferstl et al. (2016b)) for meteorological ensembles, whereas Oeltze et al. (2014) cluster streamlines modeling blood flow. Also for CFD ensembles, Hollister and Pang (2016) perform a density-based clustering of streamlines, to yield scalar fields that reveal potential bifurcations along trajectories even when the terminal separation of the members is weak. Density-based clustering is also applied by Liu et al. (2016) to find regions with high variation in vector field ensembles, where variation and similarity fields are determined based on the number of common blocks the ensemble pathlines pass through. Obermaier et al. (2016) cluster features using runt pruning to visualize trends in time-varying ensembles, where a trend is formed from clusters that persist over several time steps. Hao et al. (2016) cluster particle collision ensembles members hierarchically based on

their shape similarity, i.e., the particle density and spatial extent. For 3D scalar ensembles, Demir et al. (2016b) employ hierarchical clustering in their approach to minimize occlusion by rendering silhouettes instead of solid isosurfaces.

Major challenges of ensemble data include the complexity of the data, and also the abundance of information that can be derived and visualized. Depending on the application domain and the interests of the user, these challenges typically require multidimensional, multivariate and multivalue data visualizations, as well as multiple representations (Potter et al. (2009)). Overviews of multivariate visualization techniques in scientific visualization are given by Wong and Bergeron (1997); Fuchs and Hauser (2009); Kehrer and Hauser (2013).

Examples of techniques for visualizing multivariate data include the *parallel coordinates plot*, where individual variables are represented by axes parallel to one another, and individual samples are drawn as segmented lines passing through all axes (Inselberg (1985)), as well as its extension, the *parallel sets plot* (Bendix et al. (2005)). Feng et al. (2010) use kernel density estimation to augment parallel coordinates with density plots and visualize uncertain multivariate data for MR spectroscopy. For a survey on parallel coordinates, the reader is referred to Heinrich and Weiskopf (2013).

Another widespread class of techniques for multivariate data is given by glyphs (the reader should refer to Ward (2008); Ropinski et al. (2011); Borgo et al. (2013)).

Other popular techniques to visualize multidimensional, multivariate and multivalue data include *Multiple Coordinated Views (MCVs)* (see Roberts (2007) for an overview of MCV visualization techniques). MCVs is the general term used when data is shown in *multiple* windows and operations performed on the shown data are *coordinated*.

Such systems are designed to be highly interactive. Users can interact with the system using either *indirect manipulation*, such as *dynamic queries* (Stockinger et al. (2005)) – users employ sliders, menus, buttons, etc. to filter data or modify the manner in which data is shown – or *direct manipulation*, such as *brushing* (Becker and Cleveland (1987)) – users themselves select in a display the elements to be either shown or filtered; all elements inside the brush (usually a mouse-controlled rectangle) are highlighted. Moreover, users can perform operations such as zooming in and out, changing the viewpoint, rotating about a particular point, etc. (Plumlee and Ware (2003)).

Systems employing only two *side-by-side views*, also called *dual systems*, include variants such as (Roberts (2007))

- *Overview + detail*: one (usually aggregated) view shows the entire data set, whereas the other shows only part a of the data set, but in more detail
- *Focus + context*: one detail view, with its context given in a second view
- *Difference views*: two or more views are merged to highlight their differences
- *Master/slave*: dual views where one view (the master) controls the other (the slave)

Moreover, *small-multiples* (Tufte (1983)), while not exactly dual, are glyphs or symbols arranged in a matrix of small visualizations.

Increasingly more works dealing with ensemble data employ multiple coordinated views. For instance, Nocke et al. (2007) use MCVs in their library of visualization techniques for climate ensembles. Kehrer et al. (2010) propose brushing traditional and robust statistical moments, such as mean, variance, skewness, and kurtosis, in multiple views. Matkovic et al. (2010) integrate a simulation model view in an MCV system to visually analyze multiple simulation runs for an application from the automotive industry. Pertinent to the development process of powertrain systems, Piringer et al. (2012) perform a comparative visual analysis of 2D function ensembles on multiple levels of details (member-oriented, domain-oriented, and surface plot). Höllt et al. (2013) use MCVs to visualize heightfield ensemble data for ocean forecasting. Demir et al. (2014) introduce multi-charts to encode statistical information on large 3D scalar ensembles at multiple scales, and link the multi-chart view to a volume visualization. Molchanov and Linsen (2014) propose a framework based on MCVs for the visualization of parameter, data, and attribute spaces, as well as projected views of the data. Fofonov et al. (2016) link spatial data visualizations of isocontours with similarity plots showing the temporal evolution of ensemble members, where the similarity measure considers isocontours at individual time steps. To explore similarity patterns in spatio-temporal ensembles, Shu et al. (2016) propose three types of views: a spatial view showing the domain partitioning according to member similarities, a temporal view showing a graph with the regions of similar behavior over time, and a comparison view for comparing different aspects of the behavior of the members.

2.3 Evaluation

An often overlooked but necessary aspect of uncertainty visualization is the evaluation of how effective the different proposed approaches are. In their work on visualizing geospatial information uncertainty, MacEachren et al. (2005) regard as a key research challenge the need to assess "the usability and utility of uncertainty capture, representation, and interaction methods and tools," and also mention several publications that include empirical evaluations of usability.

Wittenbrink et al. (1995) propose two types of approaches to evaluate uncertainty visualization methods:

- *qualitative methods*, the exact measure used depending on the application domain, but also relating to perceptual issues, learning time, improvements over existing methods, etc.
- *quantitative methods*, as suggested by Tufte (1983), such as data-ink maximization, clutter and Moire pattern minimization, and multi-functionality of graphic elements

Kardos et al. (2003) present a case study to assess several uncertainty visualization techniques according to criteria such as their visual appeal, speed of comprehension, and overall effectiveness. From the seven static methods (adjacent maps, overlay, blurring, fog, pixel mixture, color saturation, and sound) and two dynamic methods (blinking pixels and animation), the survey participants rated as more-usable-than-not primarily techniques such as blinking pixels, adjacent maps or overlay, where displaying uncertainty information did not obscure the original data, like in the case of blurring or fog. Rhodes et al. (2003) also state that the ability to interactively switch uncertainty visualization on and off is a prerequisite that permits the user to understand first the visualized data and only then its associated uncertainty. Brown (2004), on the other hand, considers blurring as a visual feature that is intuitively mapped to uncertainty, since the removal of high frequency information reduces the ability to recognize fine detail, inducing thus a sense of uncertainty. Furthermore, while methods such as side-by-side comparisons or overlay indeed reveal the data, using them, e.g., for a large number of ensemble members may become a daunting task.

Sanyal et al. (2009) describe a user study to compare four visualization techniques (errorbars, glyph size, glyph color, surface color) for 1D and $2D$ data sets, and assess the effectiveness of the techniques as dependent on the particular tasks, with just the errorbars underperforming

consistently. For multiple views, Wang Baldonado et al. (2000) propose usability heuristics for when and how to use such systems.

Zuk and Carpendale (2006) consider three established perceptual and cognitive theoretical approaches, and investigate which principles were pertinent to eight representative uncertainty visualization techniques. The first approach, the Properties of the Graphics System (Bertin (1983)), considers eight visual variables (the six retinal variables: size, value, grain, color, orientation, and shape, as well as the planar dimensions (x, y)) and categorizes them depending on their immediate perceptual group selection, perceptual grouping characteristics, etc. The second approach (Tufte (1983)) – used also by Wittenbrink et al. (1995, 1996) – presents general principles for graphic excellence and integrity that allow a viewer to gain the most insight in the shortest time possible, while still using the minimum amount of ink and space. Finally, the third approach (Ware (2004)) presents principles on visual perception and comprehension essential to creating information visualizations. From these three approaches, the authors sum up a subset of possible heuristics that were found to be most relevant to the considered techniques. They also argue for the use of such "light-weight evaluations" in future uncertainty publications, in order to encourage authors to include evaluations of their proposed approaches.

Increasingly more techniques are presented along with formal evaluations by means of user studies or based on perceptual and cognitive theories. Nevertheless, there is still the need for a formal uncertainty visualization framework to measure the actual effectiveness of proposed methods.

This chapter served to introduce the reader into the various uncertainty visualization methods proposed in the area of scientific visualization, primarily from a point of view relevant to the contributions of this thesis, and it is by no means exhaustive. For further taxonomic schemes and overviews of uncertainty visualization approaches, the reader is referred to, e.g., Thomson et al. (2005); Griethe and Schumann (2006); Potter et al. (2012b); Bonneau et al. (2014); Ristovski et al. (2014).

3 Theoretical Background

Formally, an ensemble is a collection of N ensemble members consisting of observations $x_{ij\,j\in\{1,2,\ldots,N\}}$ for random variables X_i expressed on a grid structure.

A random variable X_i can take on any value in a set of possible values and is called

- *discrete*, if X_i takes on only a countable number of distinct values
- *continuous*, if X_i takes on a continuous (and thus infinite) range of possible values

The forecast output considered in this thesis includes mostly weather elements predicted by the ECMWF EPS, such as temperature, geopotential heights, or wind velocities, which take values in continuous ranges. Therefore, only continuous random variables are addressed in the rest of this chapter.

3.1 Basics of Uncertainty Representation

The uncertainty of a continuous random variable X_i can be represented by its *probability density function (PDF)*. The PDF $f_{X_i}(x)$ describes the probability (which is always a value between 0 and 1) that the random variable X_i takes on a value in a given range $[a, b]$ as the area under the PDF bounded by the two extreme values of the range

$$\Pr[a \leq X_i \leq b] = \int_a^b f_{X_i}(x)\,\mathrm{d}x \tag{3.1}$$

Knowing the PDF of the random variable X_i can reduce the uncertainty regarding the behavior of the random variable. For instance, temperatures forecast by the EPS can be statistically modeled using *Gaussian* distributions (Wilson et al. (1999)), which allows gaining insight into the likely range of values.

A random variable X_i is Gaussian (or normally) distributed if it follows a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ with corresponding PDF

$$f_{X_i}(x) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}, \quad x \in \mathbb{R} \tag{3.2}$$

where μ_i is the *mean* of the distribution and σ_i its *standard deviation* (Casella and Berger (1990)).

The *mean* μ_i of the random variable X_i with PDF $f_{X_i}(x)$ is its average value (weighted according to $f_{X_i}(x)$), a measure of the *central tendency* of the random variable

$$\mu_i = \int_{-\infty}^{\infty} x f_{X_i}(x) \mathrm{d}x \tag{3.3}$$

whereas the *standard deviation* σ_i is a measure of the *spread* of the random variable X_i around the mean μ_i

$$\sigma_i^2 = \int_{-\infty}^{\infty} (x-\mu_i)^2 f_{X_i}(x) \mathrm{d}x \tag{3.4}$$

When the random variable X_i can be assumed normally distributed, then there is less uncertainty regarding its behavior: while the range of possible values x_i that the random variable X_i can take on is theoretically infinite, the values that are likely to occur are those situated close to the mean μ_i, relative to the standard deviation σ_i.

The shape of a Gaussian PDF depends on the two parameters μ_i and σ_i. Two examples are illustrated in Figure 3.1. It can be noticed that both functions have symmetric bell shapes, but of different heights and widths. Actually, all Gaussian PDFs are bell-shaped, because values that fall outside the $[\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$ interval are highly unlikely to occur. In fact, the $68-95-99.7$ *(empirical) rule* tells that about 68.2% of the values taken by a normally distributed random variable occur within one standard deviation from the mean, 95.4% two standard deviations away, and 99.7% three standard deviations away. The likely values are, therefore, those clustered around the mean; the uncertainty characterizing a random variable

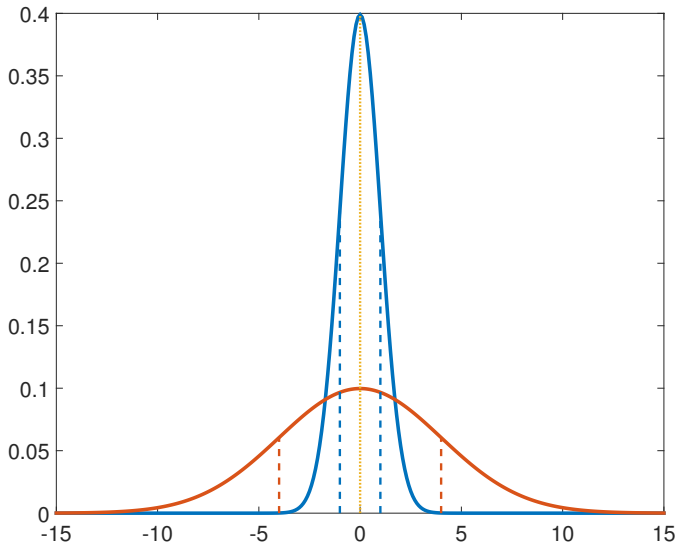


Figure 3.1: Two Gaussian PDFs with mean 0 and standard deviation 1 (blue-colored line) and 4 (red-colored line). The orange-colored dotted line is drawn at the peak of both PDFs, which occurs at the mean. The blue- and red-colored dashed lines are each drawn one standard deviation away from the mean.

depends on how far the values spread out from the mean. This is given by the standard deviation: the amount of uncertainty grows with increasing σ_i. For instance, in Figure 3.1, the blue-colored PDF has a smaller standard deviation than the red-colored PDF. The uncertainty is thus lower, because the range of likely outcomes is narrower.

The normal distribution is an example of a *parametric* statistical model, which can be completely described using the two parameters μ_i and σ_i. Otherwise stated, if the mean μ_i and the standard deviation σ_i are known, then the probability of any value occurring within a defined range can be computed. Moreover, a normal distribution with the specified mean and standard deviation can be used to generate *a random sample* of normally distributed random observations.

Frequently, however, a sample of N observations (or realizations) $x_{ij}{}_{j\in\{1,2,\ldots,N\}}$ is already available for the random variable X_i, for instance, from experiments or numerical simulations. The two parameters of the distribution of the *population* – the complete set containing all possible realizations of X_i – are, however, unknown. The population mean μ_i and standard

deviation σ_i can then be approximated by the arithmetic mean $\overline{\mathrm{x}_i}$

$$\overline{\mathrm{x}_i} = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \tag{3.5}$$

and standard deviation s_i of the sample

$$s_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_{ij} - \overline{\mathrm{x}_i})^2 \tag{3.6}$$

Besides the arithmetic mean and the sample standard deviation, there are several other statistics used to summarize the main characteristics of a sample. For instance, other summary statistics used in this thesis are the *mode* and the *median*, which are, like the mean, measures showing the central tendency. The *mode* is the most likely value, i.e., any value x_i in the allowed interval that maximizes the PDF $f_{X_i}(x)$. The *median* is the "middle value" – separating the higher half of a population/sample from the lower half, i.e., for a population, any value x_i that satisfies $Pr(X_i \leq x_i) \geq .5$ and $Pr(X_i \geq x_i) \geq .5$.

A Gaussian PDF is symmetric about its mean, which is not only the "average" of the distribution, but also its median and most likely value. This, however, does not necessarily also hold for other types of PDFs. Figure 3.2 demonstrates these measures for two skewed distributions. It can be noticed that, in these cases, the mode and the median are more intuitive indications of the central tendency of the skewed distributions. Moreover, compared to the mean, the median is a *robust* measure, because it is affected to a lesser extent by outliers – observations located far from the bulk of the majority of the values.

Nevertheless, if a random variable X_i can be assumed normally distributed, and the mean and the standard deviation are known, then the probability of any future observation is known. The Gaussian distribution is very popular in statistics, because it is analytically tractable, has a simple symmetric bell shape, and can approximate a wide range of distributions in large samples (Casella and Berger (1990)). For instance, any random variable that is a linear function $g(X_i)$ of the random variable X_i, also follows a normal distribution (in particular, $Z_i = (X_i - \mu_i)/\sigma_i$ follows a *standard* normal distribution $\mathcal{N}(0, 1)$; its tabulated probabilities can be used to compute all normal probabilities (Casella and Berger (1990))).

If the Gaussian assumption does not hold, however, the performed statistical analysis might be incorrect. Consequently, erroneous conclusions might be drawn, which may be hazardous.

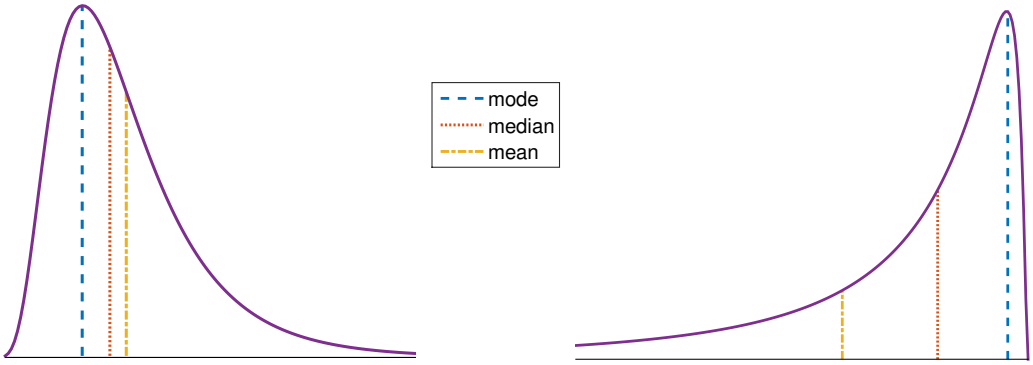


Figure 3.2: Comparison of mode, median, and mean for two skewed distributions.

Before assuming that a random variable follows a Gaussian distribution, it is therefore important to assess whether the normality assumption is valid or not (see Razali and Wah (2011) for more on normality tests).

In addition to the normal distribution, there are several other families of well-known continuous distributions. Another continuous distribution used in this thesis is the *uniform distribution* $\mathcal{U}(a, b)$, which has a constant PDF over the finite interval $[a, b]$

$$f_{X_i}(x) = \frac{1}{b-a}, \quad x \in [a, b] \tag{3.7}$$

3.2 Statistical Modeling of Distributions

The PDF $f_{X_i}(x)$ of a random variable X_i plays a major role in estimating the uncertainty underlying the behavior of X_i, because it allows using Equation 3.1 to compute probabilities that the random variable X_i takes on values in given ranges (Silverman (1986)).

3.2.1 Gaussian Mixture Models

When a random variable cannot be modeled by a common PDF, the distribution function can be approximated by a finite mixture of Gaussian distributions. A *Gaussian Mixture Model* (GMM) is a semi-parametric framework to model complex distributions (McLachlan and Peel

(2004)). Thus, the density function $f_{X_i}(x)$ of the random variable X_i can be approximated by a mixture of n Gaussian component densities (or modes) $\mathcal{N}\left(\mu_{ik}, \sigma_{ik}^2\right),\ k \in \{1, 2, ..., n\}$

$$f_{X_i}(x) = \sum_{k=1}^{n} \alpha_{ik} \mathcal{N}\left(\mu_{ik}, \sigma_{ik}^2\right)(x), \quad \alpha_{ik} > 0, \quad \sum_{k=1}^{n} \alpha_{ik} = 1 \tag{3.8}$$

The parameters $\alpha_{ik},\ k \in \{1, 2, ..., n\}$ are the mixing weights. Modeling distributions using GMMs offers more flexibility than single Gaussian distributions would, without causing a large increase in the dimension of the parameter space: instead of two parameters, (μ_i, σ_i), the PDF is summarized using a set of $3n$ parameters $(\alpha_{ik}, \mu_{ik}, \sigma_{ik})$.

Figure 3.3 shows two examples of fitted GMMs, with two (left figure) and three (right figure) components. The data set comprises 107 eruption lengths (in minutes) of Old Faithful geyser in Yellowstone National Park, USA (see Silverman (1986) for more details on the data set).

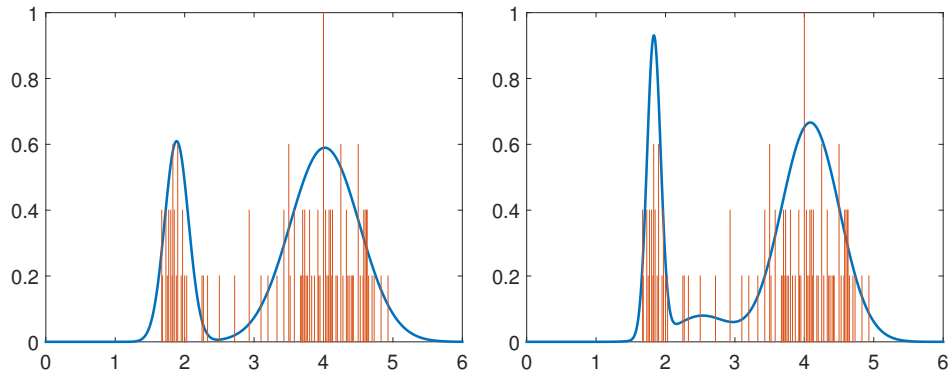


Figure 3.3: GMMs with (left) two and (right) three fitted Gaussian components for the Old Faithful geyser data set. The eruption lengths (in minutes) are shown on the horizontal axis. The red-orange vertical lines show the frequencies of the eruption lengths, normalized by the maximum value.

Mixture models are usually fitted using an iterative *Expectation Maximization* (EM) algorithm (McLachlan and Peel (2004)). To fit a number of n components, the algorithm starts with some initial guess[1] for the $3n$ summary parameters of the mixture $(\alpha_{ik}, \mu_{ik}, \sigma_{ik})$, which are updated iteratively until the prescribed convergence criteria are met.

An iteration consists of two steps

[1]Due to random initial conditions, EM algorithms may converge to local optima and lead to non-repetitive solutions. To alleviate this problem, several fittings with different starting conditions are typically performed.

- *E-Step*: based on the current $(\alpha_{ik}, \mu_{ik}, \sigma_{ik})$ of all components, the algorithm computes membership weights – posterior probabilities of component memberships – for each realization in the sample data

- *M-Step*: based on the membership weights, the algorithm computes for each component new mixing weights and updates the summary parameters

The convergence criteria may include scalar threshold values for the change in log-likelihood after each iteration, the maximum number of iterations that should be performed, etc. Figure 3.4 illustrates different steps in the EM algorithm while fitting two Gaussian mixtures to the Old Faithful geyser data set.

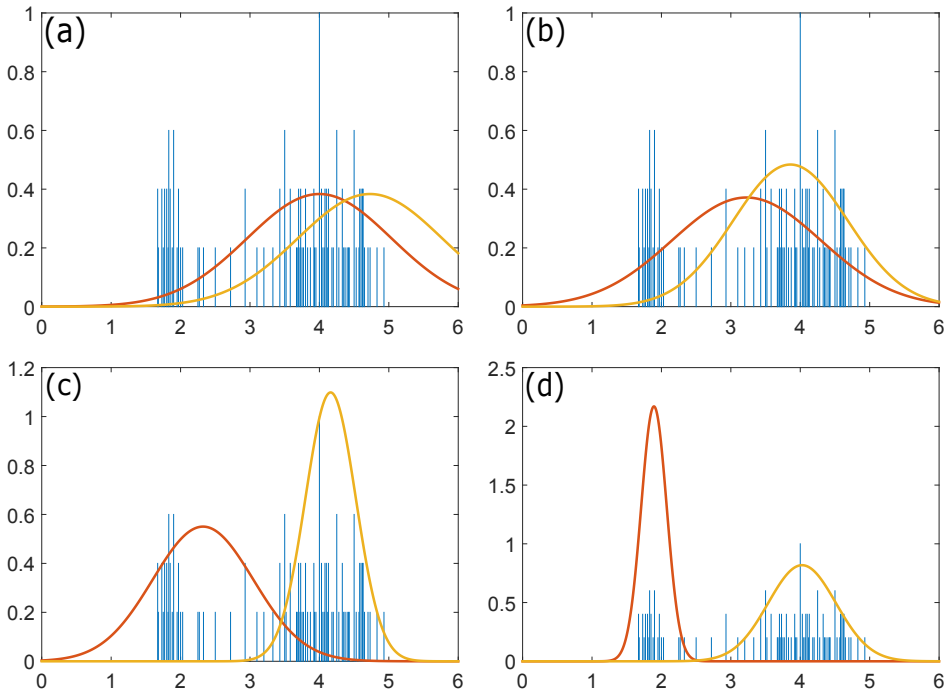


Figure 3.4: For a mixture with two Gaussian components, the two fitted Gaussian components (without scaling according to the mixing weights) for the Old Faithful geyser data set, at various stages of the EM algorithm: (a) initial guess, (b) first iteration, (c) tenth iteration, (d) twentieth iteration. The eruption lengths (in minutes) are shown on the horizontal axis. The blue-colored vertical lines show the frequencies of the eruption lengths, normalized by the maximum value.

Up to this moment, the number n of mixture components has been considered to take on a

fixed value. Typically, however, the correct number of components is unknown and has to be determined from the data sample (McLachlan and Peel (2004)). In spite of a considerable number of publications on assessing the number of components, there is no optimal solution to this problem. Proposed techniques include starting with a large number of modes and merging similar components (Hennig (2010)). Other methods fit the data several times with an increasing number of components and then evaluate the fitted models using different criteria to determine the optimal number of modes (Oliveira-Brochado and Martins (2005)).

The GMM algorithm applied in the methods proposed in this thesis is inspired by the procedure proposed by Hamerly and Elkan (2004) for the k-means algorithm. Their routine has been shown to produce good results at identifying the number of modes in moderately high dimensions. The iterative algorithm determines both the number of modes and their shape automatically, without requiring any prior knowledge on the number of components. It functions by splitting the data observations recursively until the observations assigned to each k-means center can be assumed Gaussian distributed.

The GMM algorithm works similarly, but wraps around GMM instead of k-means. By using the less restrictive EM algorithm, it permits employing a different full covariance matrix for every component, instead of assuming that data points are distributed spherically around the k-means centers. In a nutshell, unless the data can be assumed Gaussian distributed, the algorithm attempts to fit two Gaussian modes to the data. The procedure repeats recursively until no splitting needs to be performed anymore (either because the observations appear Gaussian distributed or their cardinality is too small). The final solution is refined by using the EM algorithm on the entire sample with the identified components as initial conditions.

Distribution tests are employed to decide whether a group of observations needs to be split or the Gaussian assumption holds. For small samples, the Lilliefors test (Lilliefors (1967)) is recommended (Gosling (1995)). The Anderson-Darling test (Anderson and Darling (1954)) (also used by Hamerly and Elkan (2004)) is employed otherwise to test whether the null hypothesis that the sample comes from a normal distribution can be rejected at a specified significance level. Moreover, the Bonferroni correction (Dunn (1961)) is used to address the problem of multiple comparisons, which appears when multiple hypotheses are tested on a single data set[2].

[2]Performing m tests on a single data set increases the likelihood that the null hypothesis is rejected by chance even if it is true. To reduce the chances of obtaining false positives, each of the m tests should be tested at a significance level $1/m$ times as low as the level for testing only one hypothesis. For an ensemble size around 50, where typically just a few splits are performed, the selected significance level is 0.1%.

Once GMMs have been fitted, the components can be used to perform a clustering of the sample data. What a mixture model basically assumes is that the individual realizations $x_{ij\,j\in\{1,2,\ldots,N\}}$ were drawn in various proportions from the n Gaussian components. By computing the posterior probabilities p_{ijk} that each realization x_{ij} was drawn from component k, clustering can be

- *hard* – realizations are assigned to the component with the highest posterior probability
- *soft* – realizations belong to every component with a membership score based on the posterior probability

There are several criteria to validate the obtained cluster structures (Jain and Dubes (1988)). In this thesis, we use *silhouettes* (Rousseeuw (1987)) – internal criteria requiring only the data in order to assess the clustering solution and thus suitable for algorithms starting with random initial guesses like GMM. Silhouette values are computed for every observation in the data set, to indicate how similar the observation is to other realizations in its own cluster, when compared to realizations in other clusters. The distance metric used for the silhouettes is arbitrary. The individual silhouette values can then be aggregated to yield one silhouette value per cluster or even per data sample.

3.2.2 Non-parametric Models

Mixture models do not necessarily have to be applied only for multimodal distributions. GMMs can also be employed for unimodal skewed distributions, where using a single normal distribution would not suffice (McLachlan and Peel (2004))[3]. Other effective ways of modeling skewed or multimodal distributions that require nevertheless a higher parameter space are given by non-parametric approaches. Here, the PDF of a random variable can be estimated directly from the observations (Silverman (1986)).

The *histogram* is the most prevalent density estimator, as it is very simple to construct: given a sample of N observations $x_{ij\,j\in\{1,2,\ldots,N\}}$ for the random variable X_i, a histogram splits an interval containing all observations into a series of (usually) consecutive, non-overlapping, and adjacent intervals (or *bins*) of equal size, and counts the frequencies of observations in each interval. To construct the histogram, it is necessary to select an origin x_o and a bin width h.

[3] Another way to model skewed data is by applying transformations, such as the log transformation, in order to reduce or even remove the asymmetry (McLachlan and Peel (2004)).

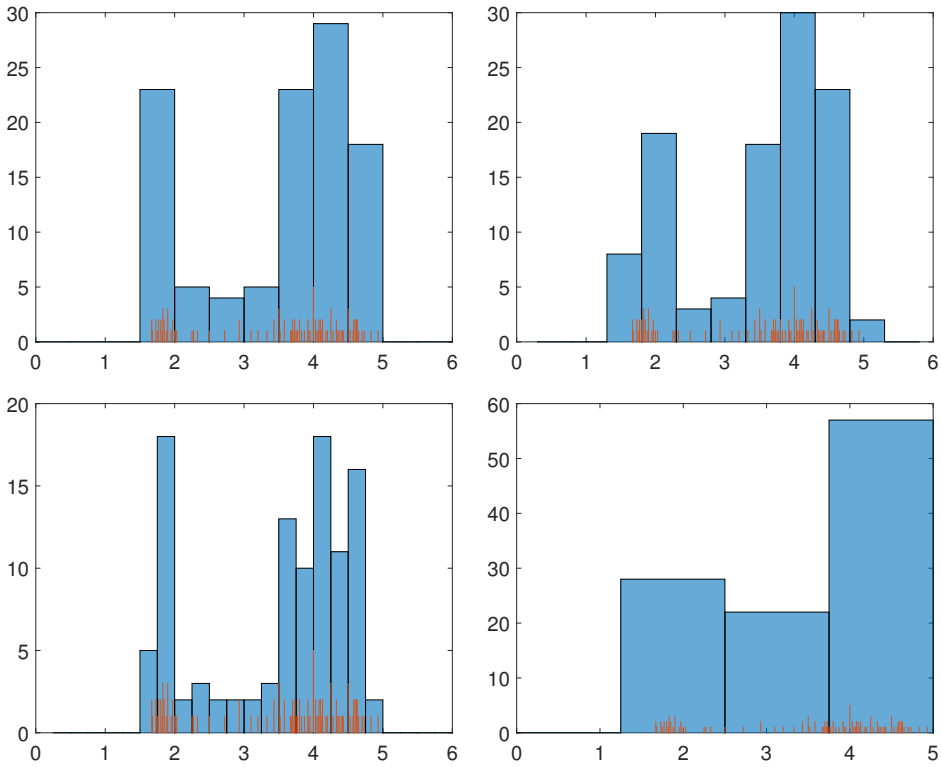


Figure 3.5: Histograms density estimators of eruption lengths, with different choices of origin (top row) and different bin widths (bottom row). The eruption lengths (in minutes) are shown on the horizontal axis, while the frequencies are shown on the vertical one. The red-orange vertical lines show the frequency of the eruption lengths.

The bins can then be defined as $[x_o + mh, x_o + (m+1)h]$, where m stands for positive and/or negative integers.

Despite its simplicity, the histogram suffers major drawbacks: the choice for the starting point and group boundaries is arbitrary, which may potentially distort the presented information. Furthermore, it requires as input the amount of smoothing of the estimate – the bin width. The effect of various choices of origin and bin width is displayed in Figure 3.5 for the Old Faithful geyser data set. For instance, even if the two density estimates in the top row of Figure 3.5 have the same bin width and quite similar shapes, the different origins lead in the left figure to a narrower left peak and a wider separation of the two main peaks than in the right figure. Moreover, the choice of the bin width (see the bottom row of Figure 3.5) can

have an even greater impact on the density estimate: the very small bin-width in the left figure leads to the emergence of several narrow peaks, while the very large bin-width in the right figure creates the wrong impression of a left-skewed unimodal density estimate.

A more accurate density estimate where the choice of the bin positions is no longer relevant is the *kernel density estimator*. Instead of specifying bin boundaries and counting how many observations fall within each bin, the density estimate is obtained by placing as many bins as there are observations (such that each observation is located at the center of its bin), counting the number of observations within each bin, and summing up their contributions. The final shape of the density estimate depends on the shape of the bin (given by its *kernel function*) and the bin width. A density estimate generalizes thus the semi-parametric mixture model, since the number of components is given by the sample size.

Figure 3.6 shows the density estimates for the Old Faithful geyser data set for two different kernel functions, a uniform (left) and a Gaussian function (right), but with the same bin width (0.5). The density estimate with a uniform kernel is reminiscent of a histogram, since the box-shape of the kernel lends the estimate a stepwise aspect. The bell shape of the Gaussian kernel leads to a significantly smoother density estimate, while still preserving the two main peaks.

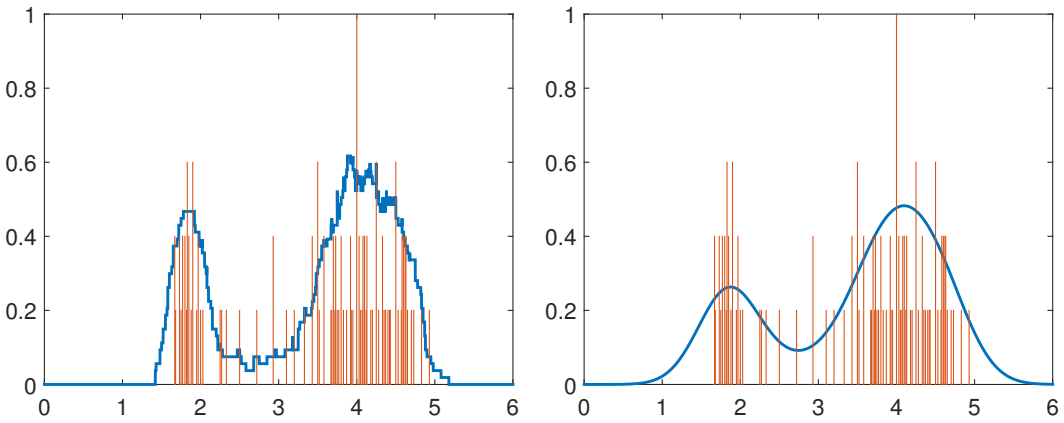


Figure 3.6: Density estimates of eruption lengths with (left) uniform and (right) Gaussian kernel functions. The eruption lengths (in minutes) are shown on the horizontal axis. The red-orange vertical lines show the frequency of the eruption lengths, normalized by the maximum value.

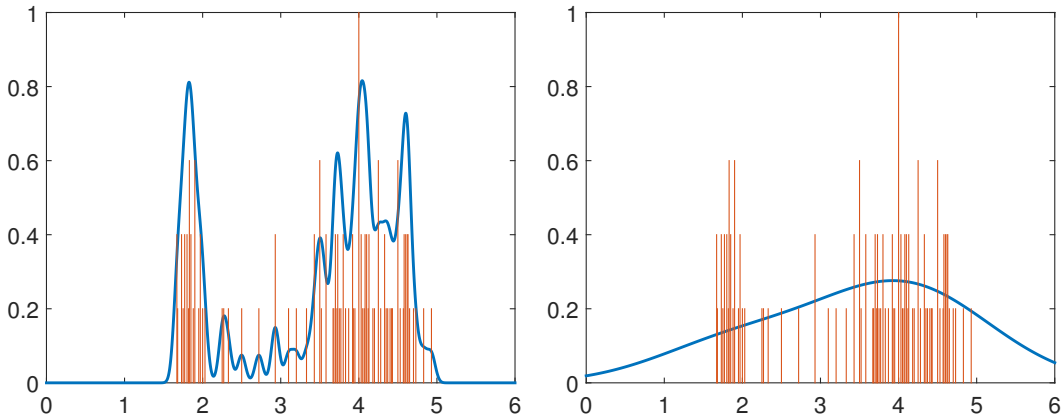


Figure 3.7: (Left) Undersmoothed (bin width 0.1) and (right) oversmoothed (bin width 2) Gaussian kernel density estimates of eruption lengths. The red-orange vertical lines show the frequency of the eruption lengths, normalized by the maximum value.

The smoothing parameter, i.e., the bin width, is also consequential to the density estimate, as can be seen in Figure 3.7: a very small amount of smoothing may lead to spurious fine structures in the estimate (see Figure 3.7 (left)), the number of "bumps" decreasing as the smoothing increases, to the extent that the multimodality of the density estimate can be obscured (see Figure 3.7 (right)).

Choosing the appropriate amount of smoothing is therefore of paramount importance for the resulting PDF. While there are no standard methods for determining the bin width, there are several rules of thumb that have been shown to work effectively for various kernel functions. Nevertheless, deriving automatically additional quantitative information, such as the modes of the PDF, is not trivial. Even if the modes can be assessed visually by the number of "bumps" or the local peaks of the density, it is still necessary to determine the number of significant peaks, their locations and sizes.

3.3 Multivariate Random Variables

The mathematical derivations presented so far were *univariate*, i.e., they involved no more than one random variable X_i. This is, however, often not the case. Ensemble data, even

when referring to observations from a single grid location, are frequently *multivariate* – two or more random variables are involved.

A simple example is an ensemble of N 2D vector fields defined over a Cartesian grid structure. At every domain location, there are $2N$ values, N for the realizations of each of the two random variables (X_{ix}, X_{iy}) modeling the x and y components of the vectors. $X_i = [X_{ix}, X_{iy}]^{\mathrm{T}}$ is then a 2-dimensional random vector (or *bivariate* random variable). This section is an outline of how several of the previously described probability concepts can be extended to include several variables (more details can be found in Mardia et al. (1979)). For the sake of simplicity, bivariate random variables are discussed here. The mathematical extensions to more dimensions are often straightforward.

Consider a bivariate random variable $X_i = [X_{ix}, X_{iy}]^{\mathrm{T}}$. The multivariate density function $f_{X_{ix}X_{iy}}(x, y)$ of the random vector is called the *joint* PDF. As before, the joint (or, for 2D, bivariate) PDF $f_{X_{ix}X_{iy}}(x, y)$ describes the probability for the random variable to take on a value in a given region A in the plane

$$\Pr[(X_{ix}, X_{iy}) \in A] = \int \int_A f_{X_{ix}X_{iy}}(x, y)\,\mathrm{d}x\mathrm{d}y \tag{3.9}$$

The *marginal* PDFs for the x and y components $f_{X_{ix}}(x)$ and $f_{X_{iy}}(y)$ are obtained by integrating over one component only

$$\begin{aligned} f_{X_{ix}}(x) &= \int_{-\infty}^{\infty} f_{X_{ix}X_{iy}}(x, y)\,\mathrm{d}y \\ f_{X_{iy}}(y) &= \int_{-\infty}^{\infty} f_{X_{ix}X_{iy}}(x, y)\,\mathrm{d}x \end{aligned} \tag{3.10}$$

The expected value of $f_{X_{ix}X_{iy}}$ is

$$\mu_i = \mathrm{E}[X_i] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_{ix}X_{iy}}\,\mathrm{d}x\mathrm{d}y \tag{3.11}$$

which translates in block form to

$$\mu_i = \begin{bmatrix} \mu_{ix} \\ \mu_{iy} \end{bmatrix} \tag{3.12}$$

The standard deviation is replaced by the 2×2 *covariance matrix*

$$\Sigma_i = \mathrm{Var}[X_i] = \mathrm{E}[(X_i - \mu_i)(X_i - \mu_i)^{\mathrm{T}}] = \begin{bmatrix} \sigma_{ix}^2 & \rho_{ixy}\sigma_{ix}\sigma_{iy} \\ \rho_{ixy}\sigma_{ix}\sigma_{iy} & \sigma_{iy}^2 \end{bmatrix} \tag{3.13}$$

If for a univariate random variable the standard deviation was enough to characterize the spread around the mean, for a bivariate random variable the covariance matrix includes not only the standard deviations of the two variables X_{ix} and X_{iy}, but also their covariance

$$\mathrm{Cov}[X_{ix}, X_{iy}] = \mathrm{E}[(X_{ix} - \mu_{ix})(X_{iy} - \mu_{iy})^{\mathrm{T}}] = \rho_{ixy}\sigma_{ix}\sigma_{iy} \tag{3.14}$$

which is a measure of the dependence between the two variables.

Thus, ρ_{ixy} is the *Pearson correlation coefficient* between X_{ix} and X_{iy}, measuring their degree of linear dependence. Its values range from -1, when observations are perfectly negatively correlated, to $+1$, when observations are perfectly positively correlated. Observations are positively correlated if they tend to increase (decrease) together and negatively correlated if they tend to exhibit contrary behavior. Perfect correlation means that observations lie on a line. When $\rho_{ixy} = 0$, there is no linear correlation. This, however, does not imply that the variables are independent, but only that their dependence, if existent, is nonlinear and could potentially be approximated better using other types of correlation coefficients or measures.

Examples of the effect of various correlation coefficients on the covariance matrix are shown in Figure 3.8. It should be noted that the covariance matrices in Figures 3.8 (a)-(c) have equal variances (regarding each of the two random variables) and only the correlation coefficients differ. Moreover, it can be observed that, even if Figures 3.8 (c) and (d) show examples of covariance matrices with zero Pearson correlation coefficients, it is only the random variables in Figure 3.8 (c), which are jointly normal, that are also independent. In Figure 3.8 (d), the two random variables are perfectly dependent $X_{iy} = X_{ix}^2$.

Like in the univariate case, for a sample, the mean and covariance matrix can be approximated by the *sample mean vector*

$$\overline{\mathrm{x}_i} = \begin{bmatrix} \overline{\mathrm{x}_{ix}} \\ \overline{\mathrm{x}_{iy}} \end{bmatrix} \tag{3.15}$$

and the *sample covariance matrix*

$$S_i = \begin{bmatrix} s_{ix}^2 & s_{ixy} \\ s_{ixy} & s_{iy}^2 \end{bmatrix} \tag{3.16}$$

where s_{ix}^2 and s_{iy}^2 are sample variances.

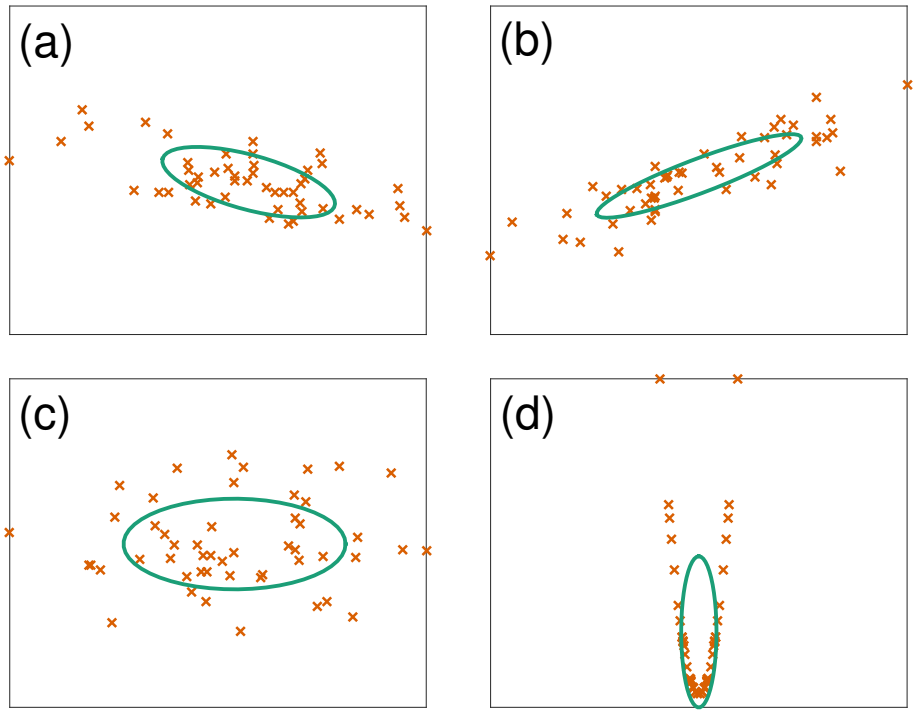


Figure 3.8: Examples of samples drawn from PDFs with various Pearson correlation coefficients: (a) rather strong negative correlation (-.6); (b) strong positive correlation (+.9); (c) zero correlation; (d) zero linear correlation, but perfect non-linear correlation. The PDFs in (a)-(c) have the same variances.

The PDF of a *bivariate normal distribution* with mean μ_i and covariance matrix Σ_i is

$$f_{X_{ix}X_{iy}}(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma_i)}} \exp\left(-\frac{1}{2}(x_i-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x_i-\mu_i)\right) \tag{3.17}$$

The two marginal PDFs of a bivariate normal distribution are also Gaussian, as can be seen in Figure 3.9. The elliptic shapes of the bivariate normal joint density visualized in this figure are given by the descriptive statistic $(x_i-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x_i-\mu_i)$, whose square root is also known as the *Mahalanobis distance* (Mahalanobis (1936)).

The Mahalanobis distance essentially computes how far an observation is from the mean, relative to the shape of the covariance matrix. In the univariate case, this reduces to verifying how far an observation is from the mean, relative to the standard deviation.

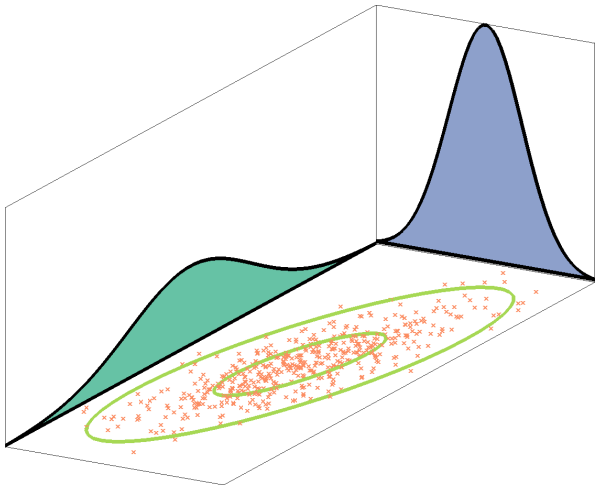


Figure 3.9: Example of 500 observations drawn randomly from a bivariate normal density function, shown using two confidence ellipses (the inner ellipse corresponding to the covariance matrix and the outer ellipse to a 95.4% confidence level) and marginal densities.

By varying the parameter τ on the right hand side of $(x_i - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (x_i - \mu_i) \leqslant \tau$, confidence regions (ellipses) corresponding to various confidence levels can be obtained. In Figure 3.9, the inner green ellipse corresponds to $\tau = 1$ (the covariance matrix), whereas the outer ellipse corresponds to $\tau = 6.17$ (the 95.4% confidence level).

An intuitive description of the Mahalanobis distance is illustrated in Figure 3.10. The left figure presents two observations in 1D. The observation marked with a green plus sign is less than one standard deviation away from the mean, being a more likely realization than the observation marked with a magenta cross, which is more than two standard deviations away from the mean.

In 2D, however, (see Figure 3.10 (right)), the distance of an observation from the mean should take into account that realizations are typically not equally far from the mean, but distributed in an elliptical manner. For instance, the realization marked with a green plus has a greater Euclidean distance to the mean than the realization marked with a magenta cross. Nevertheless, its Mahalanobis distance is smaller, because the axis of the ellipse in the direction of the observation with a plus sign is longer. Observations are thus allowed to vary more in the direction of the major axis than in the direction of the minor axis.

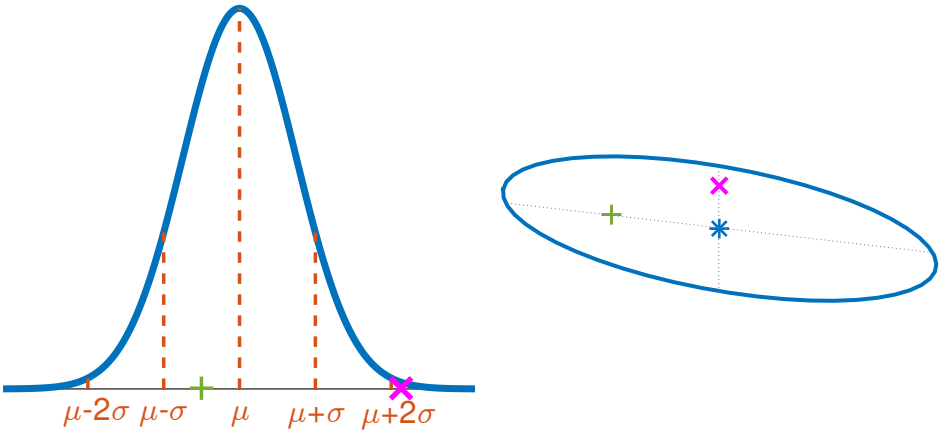


Figure 3.10: (left) 1D Gaussian PDF with two observations; (right) ellipse of covariance matrix for bivariate normal PDF with two realizations and mean.

This chapter laid the fundamental notions that lie at the basis of the techniques proposed in this thesis. Specific details on several of the topics can be found in the following chapters.

4

Visualizing the Stability of Critical Points in Uncertain Scalar Fields

Ensembles are typically large data sets that vary in space and time. The sheer size and complexity of such data sets can make the analysis and visualization of the main characteristics of the data and its inherent variability difficult even for ensembles of 2D scalar fields. Comprehension can be facilitated by removing the inessential information and keeping only representative features of the data, which can then be analyzed, classified, visualized, and, possibly, tracked. What defines a representative feature depends on the particular research problem at hand.

For scalar fields, an important class of features includes *level-sets* or *isocontours*. An isocontour is the set of all points in the domain where the scalar field takes on a specific value, the so-called *isovalue*. Moreover, interesting features frequently pertain to critical points – domain locations where the gradient vector has zero magnitude. This happens because critical points offer insight into salient surface components and their topological changes: surface components may emerge or vanish at minima and maxima, join or split at saddles.

The effect of uncertainty on isocontours has been addressed before in several works (Pfaffelmoser et al. (2011); Pöthkow and Hege (2011); Wu and Zhang (2013)), which focused on the positional uncertainty of level-sets. To investigate the geometrical variability of level-sets, Pfaffelmoser et al. (2013) analyzed gradient vectors of scalar fields and derived, at every domain location, confidence intervals for the gradient magnitudes and orientations. By studying the effect of uncertainty on the variability of gradients in scalar fields, the authors were able to

indicate whether geometric changes of level-sets were likely to occur, and perform a stability analysis of the shape and slope of isocontours.

While numerous methods have been proposed to evaluate how uncertainty acts upon features like level-sets, to the best of the author's knowledge, the approaches presented in this chapter have been the first to address the variability induced by uncertainty on critical points. Exploring various aspects of the stability of critical points and the effect of uncertainty upon them is, nevertheless, of interest, because critical points relate to prominent features in the scalar fields and their topological changes. Such an analysis would offer a different kind of insight into how uncertainty affects level-sets and other related features. Based upon an uncertainty analysis of derived quantities (gradient vectors and Hessian matrices) in uncertain scalar ensembles, the techniques proposed in this chapter classify critical points with respect to several notions of stability. This helps reveal the possible locations of critical points and assess their stability in type across the ensemble.

Even though no such studies have been performed for uncertain scalar fields, several works classify critical points according to various stability or importance measures. Thus, in the context of vector fields, various measures have been proposed to classify their importance, such as the Euclidean distance between critical points (Tricoche et al. (2001)) or the area of corresponding flow regions in the topology graph (De Leeuw and Van Liere (1999)). Moreover, for both stationary and time-varying vector fields, Wang et al. (2013) employed the topological notion of robustness to assess the stability of critical points with respect to perturbations.

For scalar fields, Edelsbrunner et al. (2000) rank critical points according to an importance measure based on the notion of homological persistence, so as to filter out those critical points with a persistence under a prescribed threshold and simplify the topology. Dey and Wenger (2007) generalize this notion to interval persistence, to assess the stability of critical points under small perturbations of the scalar field. Reininghaus et al. (2011) combine the notion of homological persistence with scale space theory to distinguish between minima and maxima with blob-, ridge-, or outlier-like spatial extent. These approaches are based on the Morse-Smale (MS) complex, which analyzes the topology of a surface by decomposing manifolds into regions of uniform gradient flow.

Topological persistence is used in the context of the MS complex, which decomposes manifolds into regions of uniform gradient flow behavior to investigate the topology of the surfaces. This helps reveal both the various features of the surface and the manner in which they are

connected. Critical points, linked together by lines of steepest descent, become the nodes of the MS complex. A hierarchy of MS complexes can be obtained by successively eliminating the critical points having importance measures below prescribed thresholds, e.g., Bremer et al. (2003) or Edelsbrunner et al. (2003). Nevertheless, MS complexes rely on a number of strong prerequisites regarding the scalar fields and their critical points, and also require numerical integration. Moreover, only the stability aspects of the critical points themselves are of interest to the presented approaches. Consequently, MS complexes are not computed here, even though gradient vector fields and Hessian matrices are used in both MS complexes and the analysis described here. The methods proposed in this chapter have thus a considerably lower degree of complexity and depend on less strong requirements.

The rest of this chapter details the proposed approaches to reveal the likely locations of critical points in scalar ensembles, and assess their stability in type by analyzing the associated gradient and Hessian matrix fields. Firstly, the scalar ensembles are summarized statistically and corresponding statistical moments are derived for the gradient vector fields. Critical points occur where the gradient vectors vanish. Therefore, confidence intervals of the gradients can be employed to derive indicator functions reflecting the likelihood of existence of a critical point at a given domain location. Further insight into the characteristics of critical points across the ensemble can be gained by deriving confidence intervals for the trace and determinant of the Hessian matrix; this pertains to the tendency of critical points to behave like minima (sources), maxima (sinks), or saddles at various domain locations.

The proposed methods can thus be applied for scalar ensembles to identify and visualize the potential locations where critical points may emerge, as well as their type stability, providing a corresponding insight into salient level-sets, surface components and surface topology events across the ensembles.

4.1 Critical Points in Ensembles

In scalar fields given on Cartesian grid structures, critical points occur at the locations where the gradient vectors vanish[1]. There are a number of methods that can be used to determine critical points in scalar data sets, such as locating the crossings of the zero-contours of the x-

[1] For data given on triangular grids, critical points can occur only at vertices, where gradients do not necessarily vanish. For Cartesian grid structures, however, critical points may occur also within the cells and are well-characterized by a zero gradient.

and y-components of the gradient field, or the grid points with non-zero Poincaré indices, etc. Data uncertainty affects, nevertheless, the locations of critical points, causing variations in both the positions and types of the critical points that are likely to emerge across the ensemble. Consequently, it is important to indicate how likely it is that a critical point emerges around a given location and whether it shows a certain type of behavior. In this context, there are two types of stability concepts that can be applied when examining the variability at each domain location:

- *positional stability* implies that critical points are likely to occur consistently in the ensemble members around the indicated location
- *type stability* indicates that critical points of the same type (minima, maxima, or saddle points) are likely to emerge consistently around the location

The actual critical points of the individual ensemble members are not used to find the "stable" domain locations, because ensemble data sets do not typically comprise the huge number of members that would be necessary for their critical points to highlight the majority of the "stable" locations. Instead, two types of indicator functions (one for each type of stability) are derived at every domain point; these indicators suggest the likelihood of a critical point of a certain type to occur around a given location. Because gradients and Hessian matrices play a major role in identifying critical points and their types, these operators are summarized statistically via confidence intervals and used to derive the indicators.

4.1.1 Confidence Intervals

At each grid point $x_{i,j}$ of a 2D Cartesian grid structure, the data uncertainty is modeled by means of a random variable $X_{i,j}$. The range of possible data values at each grid point can be expressed using intervals, $[\mu(X_{i,j}) - \sigma(X_{i,j}), \mu(X_{i,j}) + \sigma(X_{i,j})]$, where $\mu(X_{i,j})$ is the mean value and $\sigma(X_{i,j})$ the standard deviation, representing the uncertainty at the vertex.

When the random variables can be assumed to follow a specific probability distribution, confidence intervals of various confidence levels can be constructed. For instance, the aforementioned interval corresponds to a 68% confidence level for a 1D Gaussian distributed variable. Otherwise stated, there is a 68% probability that the true value lies in the confidence interval. In the rest of this chapter, $[\mu(X_{i,j}) - \sigma(X_{i,j}), \mu(X_{i,j}) + \sigma(X_{i,j})]$ is called a confidence interval

irrespective of the probability distribution. Confidence levels are nonetheless assigned for Gaussian distributions only.

To quantify the effect of uncertainty on derived quantities such as partial derivatives, these quantities are expressed in terms of functions of the random variables in a given neighborhood; then, the uncertainty is propagated from the input variables to the outputs. This procedure yields confidence intervals for the derived quantities, the exact routine depending on the function relating the input to the output.

For the first-order partial derivatives, the functions approximating the output quantities are linear combinations of the random variables at the neighboring grid points of a selected vertex. Propagating the uncertainty for linear combinations of random variables has been used before to estimate the variability of gradients in 2D uncertain scalar fields, and obtain confidence intervals for the gradient magnitude and orientation (Pfaffelmoser et al. (2013)). Their proposed approach is used here for those quantities that can be modeled using linear combinations, such as the gradient, the Hessian matrix, and the trace of the Hessian matrix.

Thus, suppose a random variable Z is a linear combination of other random variables X and Y, with means $\mu_X = \mathrm{E}[X]$ and $\mu_Y = \mathrm{E}[Y]$, and standard deviations σ_X and σ_Y

$$Z = aX + bY = AS, \quad \text{where} \quad A = [a, b], \quad S = [X, Y]^{\mathrm{T}} \tag{4.1}$$

To derive the corresponding mean μ_Z and covariance matrix Σ_Z, the uncertainty can be propagated by considering the linearity of the expectation operator and the bilinearity of the covariance

$$\begin{aligned} \mu_Z &= E[aX + bY] = aE[X] + bE[Y] = AE[S] = A\mu_S \\ \Sigma_Z &= \mathrm{cov}(AS) = A\mathrm{cov}(S)A^{\mathrm{T}} = A\Sigma_S A^{\mathrm{T}} \\ \mu_S &= E[S] = [E[X], E[Y]]^{\mathrm{T}} = [\mu_X, \mu_Y]^{\mathrm{T}}, \quad \Sigma_S = \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \end{aligned} \tag{4.2}$$

In Equation 4.2, ρ_{XY} is the correlation between the random variables X and Y.

For non-linear combinations, like the determinant of the Hessian matrix, the uncertainty is propagated by linearizing the function by a first-order Taylor approximation and then proceeding like in the linear case.

4.1.2 Confidence Intervals for Gradients

Gradient vectors can be numerically approximated at every grid point using finite differences. For 2D scalar fields defined on Cartesian grids, this makes the gradient $\nabla X_{i,j}$ at grid vertex (i, j) a linear combination A_∇ of the four[2] random variables at the non-diagonal neighbors of the vertex (see Figure 4.1 for the stencil $s_1(X_{i,j})$ holding the four random variables)

$$\nabla X_{i,j} = A_\nabla s_1(X_{i,j}) \tag{4.3}$$

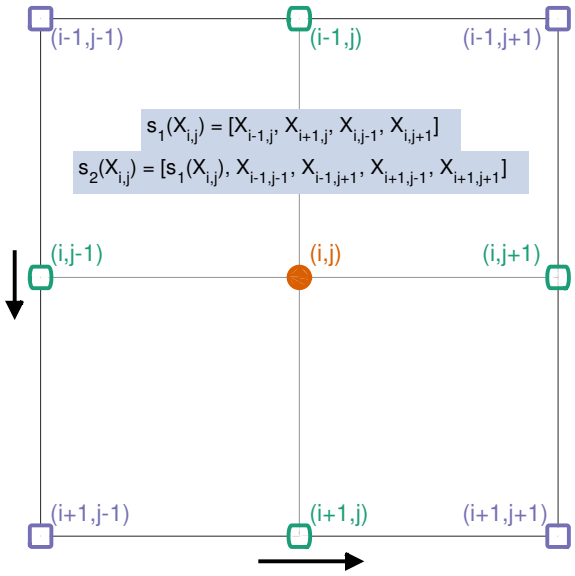


Figure 4.1: Stencils of the random variables used when approximating the gradient (s_1) and the Hessian matrix (s_2).

Introducing the linear operator A_∇ in Equations 4.2 to propagate the uncertainty for the gradient yields a mean $\mu_\nabla(X_{i,j})$ and covariance matrix $\Sigma_\nabla(X_{i,j})$ at every grid point (i, j)

$$\mu_\nabla(X_{i,j}) = A_\nabla \mu(s_1(X_{i,j})) \tag{4.4}$$

$$\Sigma_\nabla(X_{i,j}) = A_\nabla \Sigma(s_1(X_{i,j})) A_\nabla^{\mathrm{T}} \tag{4.5}$$

[2]This number holds for grid points that are not located on the boundary and where central differences can be applied; for boundary vertices, forward or backward differences are used instead, which reduces the number to three for edge points and two for corner points.

In the previous equation, $\mu(s_1(X_{i,j}))$ is a four-element vector comprising the mean values $\mu(s_1(X_{i,j})_k)$ at each element of the stencil, while $\Sigma(s_1(X_{i,j}))$ is a 4x4 covariance matrix, with the squared standard deviations $\sigma(s_1(X_{i,j})_k)^2$ on the diagonal and, as non-diagonal entries, the covariances $\sigma(s_1(X_{i,j})_k)\sigma(s_1(X_{i,j})_l)\rho(s_1(X_{i,j})_k, s_1(X_{i,j})_l)$.

The mean of the gradient $\mu_\nabla(x_{i,j})$ consists of the two mean values of the first-order partial derivatives in the x- and y-directions, whereas the covariance matrix $\Sigma_\nabla(x_{i,j})$ comprises the squared standard deviations and the covariances of the first-order partial derivatives. It can be observed that, for gradient vectors, unlike in the scalar case, standard deviations no longer suffice to express the uncertainty. Instead, a 2x2 covariance matrix is necessary to indicate the variability in the two gradient components in the x- and y-directions, as well as their correlation. Because of the presence of the correlation, two confidence intervals (one for each of the two directions) would not be enough to express the gradient uncertainty. Correspondingly, a confidence region is derived instead

$$\mu_\nabla(X_{i,j})^{\mathrm{T}}\Sigma_\nabla(X_{i,j})^{-1}\mu_\nabla(X_{i,j}) \leq 1$$

4.1.3 Confidence Intervals for the Hessian Matrix

Propagating the uncertainty is done similarly for the second-order derivatives, with the difference that the central differences kernel A_H is now applied on a nine-point stencil $s_2(x_{i,j})$ – the point itself and all of its neighbors (see Figure 4.1)

$$\begin{aligned} \mu_H(X_{i,j}) &= A_H\mu(s_2(X_{i,j})) \\ \Sigma_H(X_{i,j}) &= A_H\Sigma(s_2(X_{i,j}))A_H^{\mathrm{T}} \end{aligned} \tag{4.6}$$

Here, $\mu_H(X_{i,j})$ is a three-element vector, comprising the mean values of the second-order partial derivatives, while $\Sigma_H(X_{i,j})$ is the covariance matrix of dimensions 3x3.

To explore the type stability of critical points, a confidence volume of the Hessian matrix is not of interest. The uncertainty parameters are instead used to derive confidence intervals for the trace and determinant of the Hessian matrix.

For the trace, $\mathrm{tr}(H) = X_{xx} + X_{yy}$, the equations

$$\begin{aligned} \mu_{\mathrm{tr}}(X_{i,j}) &= A_{\mathrm{tr}}\mu(s_3(X_{i,j})) \\ \sigma_{\mathrm{tr}}(X_{i,j}) &= \sqrt{A_{\mathrm{tr}}\Sigma(s_3(X_{i,j}))A_{\mathrm{tr}}^{\mathrm{T}}} \end{aligned} \tag{4.7}$$

result in a mean $\mu_{\mathrm{tr}}(X_{i,j})$ and a standard deviation $\sigma_{\mathrm{tr}}(X_{i,j})$ at every grid vertex. The linear operator $A_{\mathrm{tr}} = [1, 1, 0]$ is applied on the three-element stencil $s_3(X_{i,j}) = [X_{xx}, X_{yy}, X_{xy}]^{\mathrm{T}}$, yielding $[\mu_{\mathrm{tr}} - \sigma_{\mathrm{tr}}, \mu_{\mathrm{tr}} + \sigma_{\mathrm{tr}}]$ as a confidence interval for the trace of the Hessian matrix.

To propagate the uncertainty for the determinant of the Hessian matrix, which is a non-linear combination of random variables $\det(H) = X_{xx} \cdot X_{yy} - X_{xy}^2$, the corresponding function

$$F(X_{xx}, X_{yy}, X_{xy}) = X_{xx} \cdot X_{yy} - X_{xy}^2$$

is linearized by a first-order Taylor series approximation[3]

$$F \approx c + Js_3$$

Here, c is a constant that is disregarded in the propagation and $J = [X_{yy}, X_{xx}, -2X_{xy}]$ is the Jacobian matrix. The Jacobian can now be used to propagate uncertainty like before

$$\sigma_{\det}(X_{i,j}) = \sqrt{J\Sigma(s_3(X_{i,j}))J^{\mathrm{T}}} \tag{4.8}$$

The corresponding mean is simply $\mu_{\det}(X_{i,j}) = F(\mu(s_3(X_{i,j})))$. The confidence interval for the determinant of the Hessian matrix is then $[\mu_{\det} - \sigma_{\det}, \mu_{\det} + \sigma_{\det}]$.

4.1.4 Indicator Functions

In the previous section, uncertainty has been propagated to derive confidence regions/intervals for the gradient, and the trace and determinant of the Hessian matrix, without making any assumption on the probability distribution of the random variables. Using confidence intervals does not require computing probabilities or any specific probability distributions of the random variables. Moreover, these procedures are both deterministic and computationally

[3] For highly non-linear functions, other probabilistic approaches, e.g., a Monte Carlo simulation, would be preferred to a function linearization.

inexpensive; they result in neither the long running times, nor the huge number of ensemble members that Monte Carlo integrations would.

The confidence intervals serve to assess the two types of stability of the critical points (positional and type) around each grid location in the domain. The confidence region of the gradient is used to indicate whether a critical point can occur around the respective location. Since critical points can occur anywhere within a grid cell where the gradient vanishes, positional indicators relate the confidence region of the gradient to a zero magnitude. Then, the confidence intervals of the trace and determinant of the Hessian matrix are used to infer on whether critical points around a given location are likely to be of a certain type.

Positional Indicator Functions

The previously defined confidence region of the gradient $\mu_\nabla(X_{i,j})^{\mathrm{T}}\Sigma_\nabla(X_{i,j})^{-1}\mu_\nabla(X_{i,j}) \leq 1$ corresponds to the covariance matrix, i.e., it corresponds, for the x and y gradient components, to their dispersions around the means and their coupling. Without any further knowledge, a critical point is likely to occur around a grid location if the origin falls within the confidence region corresponding to the covariance matrix.

If a specific distribution can be assumed, such as a Gaussian distribution, for instance, then confidence regions corresponding to different confidence levels can be constructed and used to test whether a zero gradient is possible. For example, the confidence region for a 95.4% confidence level is given by $\mu_\nabla^{\mathrm{T}}\Sigma_\nabla^{-1}\mu_\nabla \leq 6.17$; since there is a 95.4% chance that the true value of the gradient lies within the confidence region, if the origin does not fall within the region, it is quite unlikely that a critical point emerges around the location.

Nevertheless, irrespective of the probability distribution of the gradient vector, the confidence region of the gradient provides at least a binary positional indicator of whether a critical point is likely to occur around a grid vertex

$$\mathrm{ind1}(x_{i,j}) = \begin{cases} 1 & \text{if } \mu_\nabla^{\mathrm{T}}\Sigma_\nabla^{-1}\mu_\nabla \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

However, due to the use of the inverse of the covariance matrix in Equation 4.9, the binary positional indicator is liable to produce spurious results when the covariance matrices are ill-conditioned.

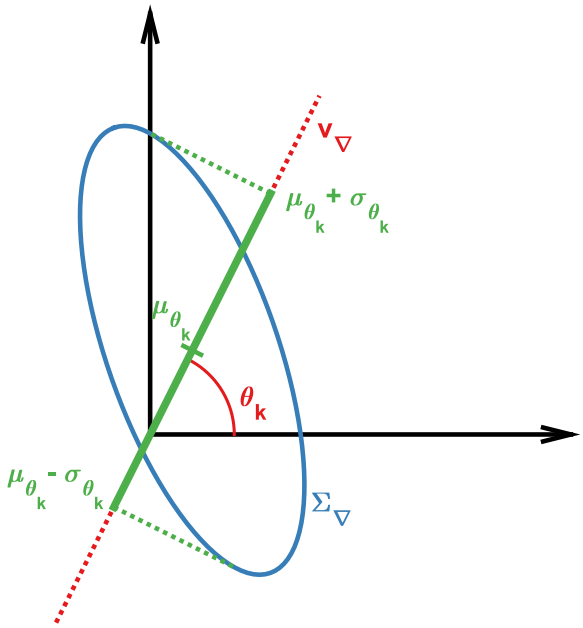


Figure 4.2: Projecting covariance matrix Σ_∇ on direction v_k corresponding to angle θ_k yields mean μ_{θ_k} and standard deviation σ_{θ_k}.

In such cases, instead of using Equation 4.9, the covariance matrix is projected on every direction of a discretization of the unit circle $\theta_k \in [0, 2\pi]$ (see Figure 4.2). This yields a mean μ_{θ_k} and a standard deviation σ_{θ_k} for every direction $v_k = [\cos\theta_k, \sin\theta_k]^\mathrm{T}$

$$\begin{aligned} \mu_{\theta_k} &= v_k^\mathrm{T} \mu_\nabla \\ \sigma_{\theta_k} &= \sqrt{v_k^\mathrm{T} \Sigma_\nabla v_k} \end{aligned} \tag{4.10}$$

The binary positional indicator tests whether every confidence interval contains the origin

$$\mathrm{ind1}(x_{i,j}) = \begin{cases} 1 & \text{if } |\mu_{\theta_k}| \leq \sigma_{\theta_k}, \ \forall \ \theta_k \in [0, 2\pi] \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

For particular PDFs like the Gaussian distribution, where the mean is the most likely value of the gradient, the likelihood of the origin with respect to the covariance matrix can be used in order to refine the positional indicators. Thus, computing the Mahalanobis distance yields how far from the mean the origin lies relative to the covariance matrix. To have a value of one when the mean is at the origin, the complement of the Mahalanobis distance is used in

the refined positional indicators

$$\text{ind1}(x_{i,j}) = \begin{cases} 1 - \sqrt{\mu_\nabla^\mathrm{T} \Sigma_\nabla^{-1} \mu_\nabla} & \text{if } \mu_\nabla^\mathrm{T} \Sigma_\nabla^{-1} \mu_\nabla < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

The refined positional indicators vary from 1, when the origin is at the center of the confidence region, to 0, when the origin is on the boundary of the confidence region or outside of it. In this manner, the likelihood of the origin is assessed based on its position with respect to the confidence region. A mean with a zero magnitude indicates the origin as the most probable realization of the gradient. Critical points are thus likely to emerge around the domain location in most ensemble members. This likelihood decreases as the origin drifts from the center of the confidence region, relative to the shape of the covariance matrix. As a result, critical points occur less frequently around the grid point throughout the ensemble.

The positional indicators have positive values in those domain regions where, according to the variability analysis, critical points tend to occur repeatedly across the ensemble. The binary and refined indicator functions can thus be regarded as measures of the positional stability of critical points. Nonetheless, the mere fact that a positional indicator has a zero value does not imply that a critical point cannot appear around the grid vertex. A critical point may still emerge, but it is less likely to occur, e.g., it may be a transitory state or noise. Moreover, as it will be shown later on (see Section 4.4), the larger the number of ensemble members is, the less likely it becomes that critical points occur around locations with zero-valued indicators. Conversely, indicators may have positive values at grid points around which no critical point appears in any ensemble member. The non-zero-valued indicators reveal those locations where critical points could have occurred in ensemble members that have not been realized (see Section 4.4). Furthermore, in the particular case of Gaussian distributions, the refined indicators also reveal how likely it is that a critical point occurs around a vertex.

Type Indicator Functions

The positional indicators provide information on whether critical points are likely to occur or not around a certain position in most ensemble members, but not on whether the critical points are likely to be of a certain type. For this kind of information, analyzing the variability of the gradient does not suffice. It is necessary to examine the Hessian matrix and its associ-

ated eigenvalues: positive eigenvalues indicate a local minimum, negative eigenvalues a local maximum, and both positive and negative eigenvalues a saddle.

Deriving confidence intervals for the eigenvalues is, however, not desired, because it is less complex and error-prone to analyze the type of the critical points by summarizing the trace and determinant of the Hessian matrix: the trace is a linear operator of the second-order partial derivatives, while the determinant, albeit being non-linear, has a simpler function of the second-order derivatives than the eigenvalues do. The confidence intervals of the trace and determinant of the Hessian matrix are thus used to derive type indicators showing the tendency of critical points to behave consistently like maxima, minima, or saddle points across the ensemble.

The classification of critical points according to the trace and determinant of the Hessian matrix is as follows: depending on the sign of the determinant, critical points can be either saddles, if $\det(H) < 0$, or minima and maxima, if $\det(H) > 0$. For a positive determinant, the sign of the trace helps distinguish between minima, if $\mathrm{tr}(H) > 0$, and maxima, if $\mathrm{tr}(H) < 0$. Consequently, critical points with a stable type behavior are likely to occur around the domain locations where the trace and the determinant of the Hessian can be regarded as either distinctly positive or distinctly negative based on their confidence intervals (see Figure 4.3).

In this context, the trace of the Hessian matrix is said to be

- *distinctly positive* if the lower endpoint of its confidence interval is greater than zero

  $$\mu - \sigma > 0 \text{ or } \mu/\sigma > 1$$

- *distinctly negative* if the upper endpoint of its confidence interval is less than zero

  $$\mu + \sigma < 0 \text{ or } \mu/\sigma < -1$$

The trace of the Hessian matrix is nothing else than the divergence of the gradient vector field. Thus, a clearly positive (negative) value of the trace implies that a critical point emerging around a given location tends to behave like a minimum (maximum) or, potentially, a saddle. In other words, a grid point having a distinctly positive trace indicates that the critical points occurring around the location are unlikely to be maxima. This happens because minima have both eigenvalues positive and, thus, positive divergence values, whereas maxima have both eigenvalues negative and, thus, negative divergence values. Saddles, on the other hand, may exhibit both positive and negative divergence values, since they have both positive and

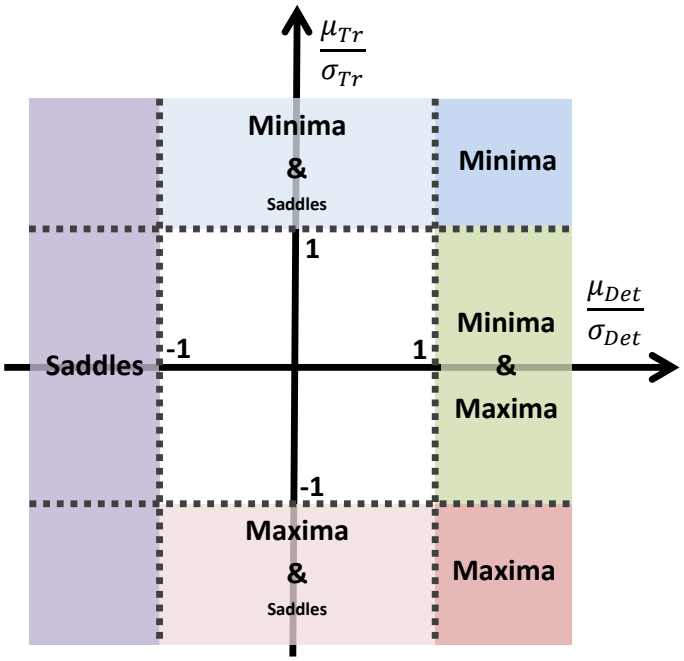


Figure 4.3: Classification of critical points showing stable behavior, depending on the confidence intervals of the trace and determinant of the Hessian matrix.

negative eigenvalues, the divergence depending on which eigenvalue is larger in absolute value. Nevertheless, since saddles may have negative or zero divergence values as well, compared to minima, saddle points are less likely to occur near domain locations with clearly positive divergence values. This holds even if no definite statement can be made about the sign of the determinant. However, when the determinant can be regarded as distinctly positive, emerging critical points are most likely minima. Likewise, a distinctly negative determinant implies a saddle.

A similar analysis can be performed for locations exhibiting clearly negative trace values, revealing whether emerging critical points are more likely to behave like maxima, saddles, or both maxima and saddles.

The analysis of every domain location with positive positional indicators can thus be started from the trace of the Hessian matrix, which differentiates between predominantly minima and maxima (and, potentially, saddle points), and then, based on the determinant, between minima (maxima) and saddles. Alternatively, the analysis can start from the determinant

of the Hessian, which distinguishes between saddle points and minimum/maximum behavior, and then, based on the trace, between minima and maxima. Both ways lead to the same ultimate conclusion.

When specific probability distributions can be assumed, type indicators can also be further refined. For instance, in the particular case of a Gaussian distribution, the type indicators can be extended to reveal locations with divergence values close to zero, i.e., if zero is less than a certain threshold τ away from μ_{Tr} in terms of σ_{Tr}, i.e., $|\mu_{\mathrm{Tr}}|/\sigma_{\mathrm{Tr}} < \tau$. At such locations, saddle points should primarily be expected.

To summarize, positional indicators reveal the locations where ensemble members are likely to have their critical points, while type indicators show whether these critical points are likely to be of a certain type or the type is unstable. Whenever a stable sign can be assumed for both the trace and the determinant of the Hessian matrix, critical points exhibit a stable type behavior. This means that a specific type of critical points (and of corresponding surface behavior) can be expected to occur consistently around the given locations in the ensemble. For instance, a location indicating stable minimum behavior implies a stable structure as well, since the critical point, the region grown around it, and the topological event of a surface component emerging appear frequently among the ensemble members. When only one of the two quantities shows a stable sign, certain variations in type can appear. For instance, if the determinant of the Hessian matrix is clearly positive, but no definite statement can be made about the trace, then both maxima and minima are likely. This implies that minimum/maximum oscillations are possible and associated structures, albeit present in most members, may have their shape inverting from one member to another. Finally, when none of the two quantities shows a stable sign, any type of critical point can be expected to occur. This also indicates a highly unstable surface behavior among the ensemble members.

4.2 Visualization

This section presents techniques that allow a concurrent visualization of the proposed indicators and representative scalar fields of the ensemble. Critical points are also displayed for validation purposes on several occasions, although they bear no relevance to the derivation of the indicators. Nevertheless, placing the indicators and the critical points in space allows

observing the likely locations of critical points and their type stability, along with various related surface components.

Representative scalar fields are thus always shown in the background, either as contour plots or texture maps, with a sequential colormap varying from blue (for low values) to green (for high values). The number of used colors is held small, in order to avoid smooth transitions and thus convey better different surface components shown via texture maps. The visualization techniques lend themselves thus to being extended by integrating further surface components. However, since such additional components are domain specific or dependent on the needs of the user, no other surface components are included here. Critical points, if shown, are displayed using circles. Their color varies depending on the type information they carry:

- black – when the type information is not relevant and for saddles
- orange – for maxima
- pink – for minima

Connected areas with indicators having positive values are denoted in the following as *emphasized* or *marked* regions.

4.2.1 Visualization of Positional Indicators

Binary positional indicators are shown using gray-colored circular glyphs, centered at every grid point of a 2D Cartesian grid where indicators have positive values. To avoid clutter, their radii are equal to half the length of a grid cell. Displaying the positional indicators as circular glyphs, as opposed to using a point-based representation, serves to reflect that critical points are expected to occur around the grid points and not necessarily exactly at the grid vertices. Moreover, it also emphasizes the locations where critical points are likely to appear.

For particular cases, such as a Gaussian distribution, refined positional indicators are encoded in the opacity of the glyph: the more likely it is for critical points to appear consistently in the ensemble members around a location, the more opaque the glyph is. Examples of both binary and refined positional indicators are given in Figures 4.4 (a) and (b) for an ECMWF temperature ensemble simulated for a forecast period of nine days above Europe. The contour lines shown in the background are the isocontours of the mean scalar field. Figure 4.4 (c) shows the actual critical points of the 50 members, superimposed over the binary indicators

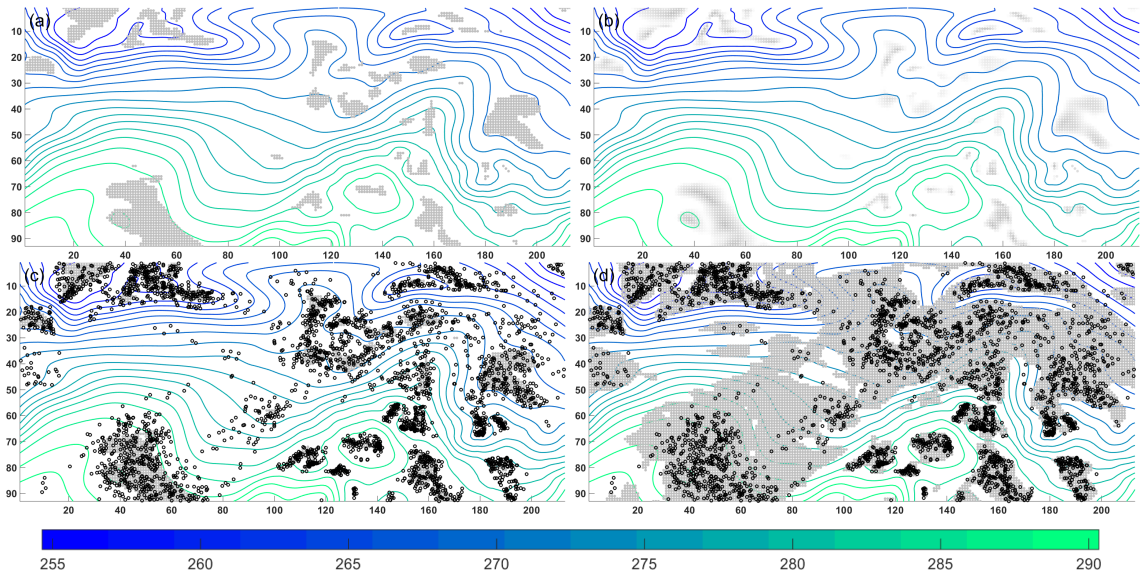


Figure 4.4: Positional indicators for ECMWF temperature ensemble: (a) binary and (b) refined indicators for the confidence region given by the covariance matrix; (c) binary indicators, with all critical points of the ensemble superimposed; (d) binary indicators for a confidence level of 95.4%, with all critical points superimposed. All figures have the isocontours of the mean scalar field in the background.

from Figure 4.4 (a). The critical points of the ensemble members occur primarily in the regions emphasized by the indicators.

Assuming a Gaussian distribution, Figure 4.4 (d) displays the positional indicators for a confidence level of 95.4%. Since the confidence level is higher than when the covariance matrix is considered, emphasized regions occupy larger areas. Therefore, just a few critical points occur in unmarked regions. Also, because of the high confidence level, critical points are unlikely to occur in unmarked areas. Moreover, critical points emerging in regions marked in Figure 4.4 (d), but not in Figure 4.4 (c), are less likely to occur consistently than those appearing in regions marked in both figures.

Using the positional indicators to highlight the domain locations where critical points are likely to emerge in an ensemble has several benefits over simply showing its critical points. First of all, the procedure of deriving the indicators is computationally inexpensive and emphasizes the regions where critical points tend to occur consistently in the ensemble. Then, compared to various existent clustering algorithms, it requires almost no parameter tailoring from the user. Moreover, the indicators express the variability induced by uncertainty on critical points.

Showing just the critical points of the ensemble members would require a huge number of members in order to offer a similar insight into the regions where critical points are likely to appear consistently in the ensemble.

While indicators highlight regions where no critical points occur, these are locations around which critical points could have appeared in ensemble members that have not been realized (see Section 4.4). Conversely, critical points occurring in unmarked regions are likely to be unstable, e.g., noise or transitory states. Refined indicators even serve to distinguish between more stable regions – where critical points are likely to emerge in most ensemble members – and less stable regions – where critical points are likely to appear in a few members only. Since critical points pertain to the occurrence of certain surface events or features, finding regions displaying "positional stability" may be used, for instance, to obtain a fast insight into interesting surface components and isovalues.

Figures 4.4 (a)-(d) show the isocontours of the mean scalar field in the background. When a Gaussian distribution is not appropriate to model the scalar ensemble, however, the mean field is not necessarily suitable to represent the average ensemble behavior. Therefore, any other ensemble member or statistical aggregate could be used instead. For instance, overlaying the circular glyphs of the positional indicators over the isocontours of individual ensemble members and superimposing their critical points can help classify these points as stable or unstable. Furthermore, in the Gaussian case, a slider functionality can be used to classify critical points interactively, from most to least stable. For example, as the confidence level α varies in the equation of the confidence region $\mu_\nabla(x_{i,j})^{\mathrm{T}}\Sigma_\nabla(x_{i,j})^{-1}\mu_\nabla(x_{i,j}) \leq \alpha$ from 0 to 9.21 (corresponding to a confidence level of 99%), increasingly more grid points get positive positional indicators and have correspondingly stable critical points nearby; critical points can be classified as more or less stable depending on the confidence levels when they become stable. Critical points in regions left unmarked for $\alpha \geq 9.21$ would be classified as unstable according to the uncertainty analysis.

Displaying only the critical points of the mean scalar field or of any particular ensemble member would not be as insightful as showing the positional indicators. For instance, even when the mean field is representative for the ensemble (like in the case of a Gaussian distribution), its critical points do not necessarily appear in every region emphasized by the indicators (see Figure 4.5). Moreover, the critical points of the mean field reveal neither the shape, nor the extent of the regions where critical points are likely to occur.

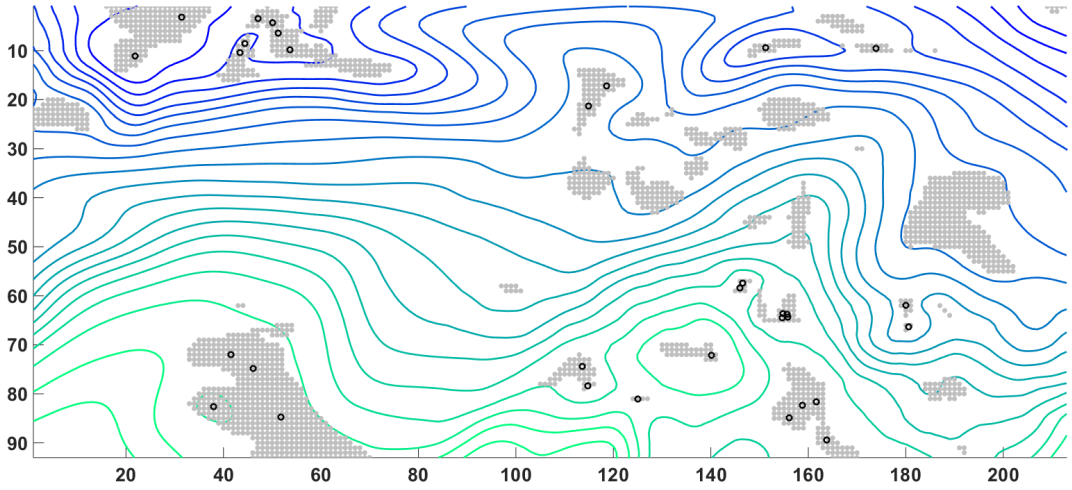


Figure 4.5: Critical points of the mean scalar field superimposed over binary positional indicators for the ECMWF temperature ensemble.

3D Extension for the Positional Indicators

The proposed methods addressing the stability of critical points have been described for the 2D case. The mathematical extension to 3D is, nevertheless, straightforward and thus not described here. Several 3D visualization techniques can be applied to reveal potential domain locations of critical points in 3D.

Two such techniques are demonstrated here for the 3D ECMWF temperature ensemble that had the previous ensemble as its 2D slice at the highest pressure level. For the first one (see Figure 4.6 (top)), spherical glyphs are immersed in the volume of stacked semi-transparent 2D slices. For the second technique, the Mahalanobis distance $\mu_{\nabla}^{\mathrm{T}}\Sigma_{\nabla}^{-1}\mu_{\nabla}$ of the origin from the gradient mean is computed at every grid point. The isosurface of isovalue 1, colored based on the values of the scalar field at the vertices of the isosurface, is shown in Figure 4.6 (bottom), along with several critical points of the ensemble (shown as red spheres).

4.2.2 Visualization of Type Indicators

The glyphs for type indicators are similar to those for positional ones, except that they have a different color, depending on how the specified criteria are fulfilled. For instance, when

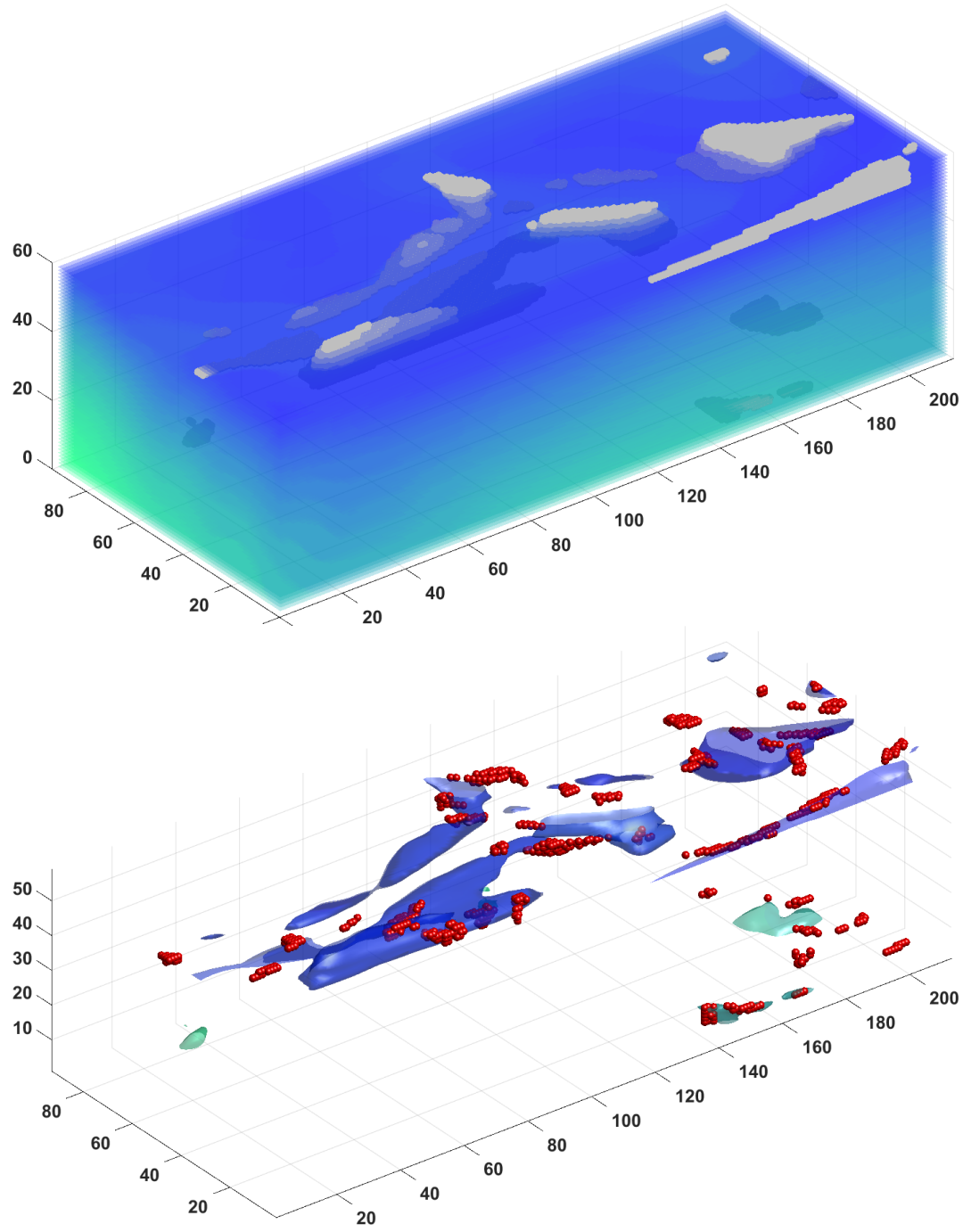


Figure 4.6: 3D ECMWF temperature ensemble showing positional indicators using (top) gray-colored spherical glyphs and (bottom) isosurface of isovalue 1; critical points are shown as red spheres in the bottom figure.

looking for locations where critical points tend to behave like minima, the corresponding criteria exhibit distinctly positive trace and determinant values.

Glyphs are only shown when at least one criterion is met and are colored

- red – both criteria are met (most stable type behavior)
- purple – only the determinant fulfills the specified criterion
- brown – only the trace fulfills the specified criterion
- gray – trace fulfills specified criterion for minimum/maximum behavior, but determinant indicates saddle behavior

For the previous 2D ECMWF temperature ensemble, critical points are displayed in Figure 4.7 (a), colored according to their type. Figure 4.7 (b) highlights locations where primarily maxima tend to appear. It can be observed that regions with stable type behavior are indicated mostly at locations where critical points of the same type cluster together. Nevertheless, the type indicators offer more insight than the naive display of critical points colored according to their types. For instance, both regions numbered 1 and 2 in Figure 4.7 (b) appear to comprise clusters of maxima and saddle points: region 1 shows a cluster of maxima enclosed to the left and to the right by two clusters of saddle points; region 2 shows the opposite, a cluster of saddle points, enclosed by two clusters of maxima. A visual inspection would indicate for both regions stable maximum behavior near the clusters of maxima. According to the variability analysis, however, it is only the second region where both criteria are fulfilled, implying thus consistent maximum behavior. The type indicators for the first region show clearly negative trace values; this means that minima are unlikely to occur, but maxima and, potentially, saddle points can be expected around the indicated regions. Compared to regions 1 and 2, clusters of critical points of different types are more difficult to identify visually in region 3. The colors of the type indicators suggest maxima and, potentially, saddles in the upper half, and maxima and minima in the lower half. Three grid points exhibit stable maximum behavior.

Type indicators for stable saddle behavior are shown in Figure 4.7 (c). Even if critical points of all three types mix to some extent, hindering a visual clustering for both regions numbered 1 and 2, the type indicators reveal the grid vertices around which saddle points are likely to occur in many ensemble members. Finally, locations where minima should predominate are indicated in Figure 4.7 (d). In regions numbered 1 and 2, gray-colored type circular glyphs

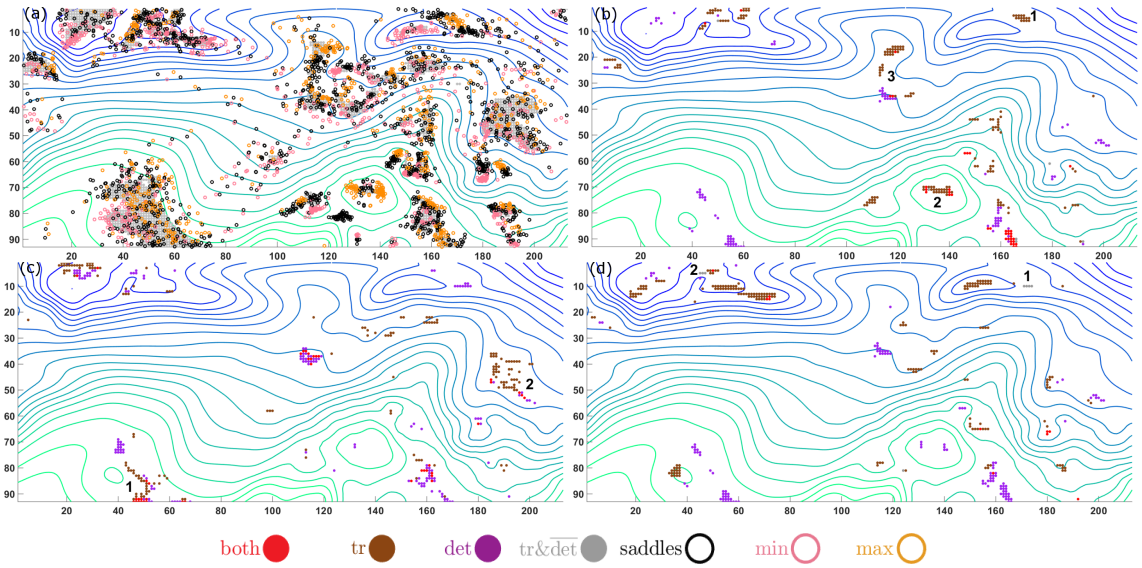


Figure 4.7: ECMWF temperature ensemble with isocontours of mean scalar field in the background. (a) Critical points superimposed over binary positional indicators. Type indicators for (b) maxima, (c) saddle points, and (d) minima.

show grid points with distinctively positive trace values, but clearly negative determinant values. This implies that saddle points are more likely to occur around these locations than minima are. As can be seen in Figure 4.7 (a), saddle points indeed prevail in these regions.

4.3 Validation

Two types of indicator functions were introduced in the previous sections to highlight the domain locations where critical points are likely to emerge (positional stability) and exhibit a certain kind of behavior (type stability). Thus, critical points have been assessed as more likely to appear more frequently in the ensemble near those locations with positive positional indicators. Even if no actual critical points occur in the ensemble members around these locations, critical points have been considered as likely to occur at such locations in members that have not been realized. Moreover, critical points emerging near locations with zero-valued positional indicators have been regarded mostly as numerical noise or transitory states.

To validate the proposed techniques, the number of occurrences of critical points around a certain position should be related to the positional indicator of the grid point. This is,

H	n	r	H	n	r
0.56	1	0	0.22	15	0.33
0.5	1	1	0.2	19	0.58
0.46	2	0.5	0.18	27	0.52
0.4	1	0	0.16	33	0.3
0.38	1	0	0.14	62	0.21
0.34	2	0	0.12	105	0.29
0.32	3	0	0.1	183	0.36
0.3	2	0.5	0.08	352	0.41
0.28	3	0.67	0.06	676	0.54
0.26	5	0.4	0.04	1344	0.66
0.24	12	0.17	0.02	2868	0.84

Table 4.1: False negative analysis for the ECMWF temperature ensemble. H - histogram value, normalized by the number of ensemble members; n - number of grid points attaining the histogram value; r - error rate.

nonetheless, not straightforward, since critical points appear typically neither at the grid vertices, nor at the same spatial locations.

Assigning critical points to grid vertices should also be done with care, because assigning critical points to different neighboring grid points may artificially lower the number of occurrences assigned to a grid point. Assigning a critical point to all neighboring vertices alleviates this problem, but may result in positive numbers of occurrences at grid points that have zero-valued positional indicators, but neighboring grid points with positive positional indicators. In this work, the latter method of assigning critical points to all neighboring vertices is, nevertheless, preferred, even if this may result in some instances of false negatives.

A 2D histogram is constructed, counting at each grid point the number of ensemble members where at least one critical point was assigned to the vertex. For each histogram value in decreasing order, the percentage of *false negatives* is computed by counting, for every grid point attaining the respective histogram value, how many of these grid points have zero-valued positional indicators. Conversely, the percentage of *false positives* counts how many grid points have positive indicator values, but no assigned critical points.

Table 4.1 shows the false negative error rates for the 2D ECMWF temperature ensemble. The six grid points with the five highest histogram values, numbered 1 to 5 in decreasing order, are shown in Figure 4.8.

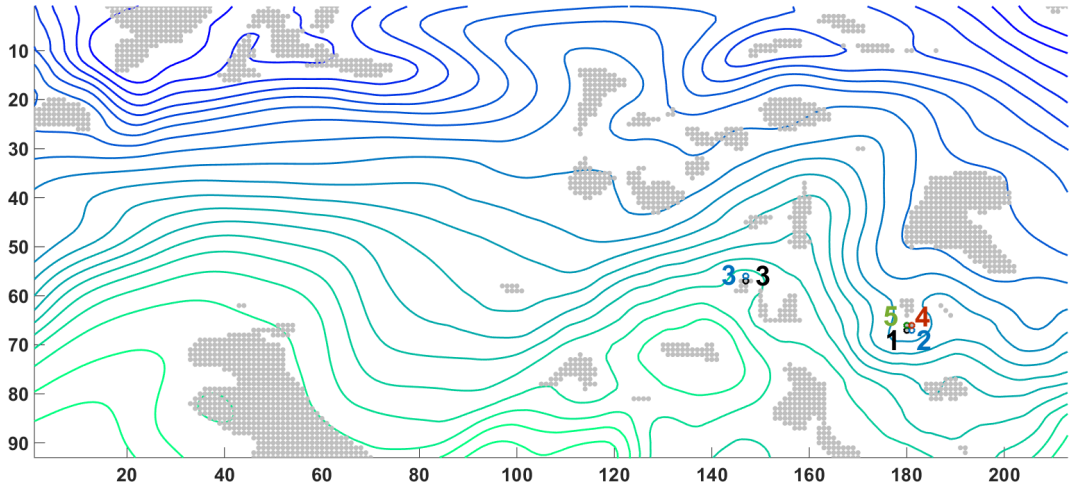


Figure 4.8: ECMWF temperature ensemble with the isocontours of the mean scalar field; superimposed are binary positional indicators and the grid points with the five highest histogram values, numbered 1 to 5.

The grid point at the peak of the histogram (numbered 1 and assigned to in 56% of the ensemble members), has a positive indicator value. Three other of its neighboring vertices also attain high histogram values, but only grid points numbered 4 and 5 have positive indicator values as well. Grid point numbered 2 is an example of a grid point having high histogram values because of critical points being assigned to all enclosing vertices. The high histogram values for these four grid points, of which one of them is a false negative, implies that critical points are likely to occur frequently in the corresponding grid cell, but the frequency is prone to decrease in the lower right direction. Similar observations hold for the two grid points numbered 3: critical points are assigned to both locations in 46% of the ensemble members, but only the lower (black-colored) grid point has a positive indicator value. The frequency of critical points is therefore likely to drop in the upper direction. Such grid points (numbered 2 and 3) are the reason for the high false negative error rates in the upper rows of Table 4.1. High false negative error rates appear also in the last rows of Table 4.1, caused by the grid points around which critical points appear irregularly. Regarding the false positive error rates, no critical points were assigned to 12% of the 2019 grid points with positive indicator values. It should also be taken into consideration that these error rates are upper bounds of the actual values, as critical points are assigned to all of their enclosing grid points. For specific distributions, false negative error rates could be decreased by having higher confidence levels. This could, however, also increase the false positive error rates.

4.4 Further Results

In this section, the proposed techniques are applied on three other ensembles, two synthetic data sets and another ECMWF ensemble.

The first synthetic ensemble has 5000 members and is defined on a 100×100 grid in the $[-2, 2] \times [-2, 2]$ range. The ensemble was generated by drawing random vectors from a multivariate normal distribution with the mean and covariance matrix given by

$$\mu = \begin{bmatrix} 0.5 \\ 1 \\ 0.5 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$$

and inputting these for the three parameters a, b, and c in

$$-x^4/4 - y^4/4 - x^2y^2/2 + ax^2/2 + bxy + cy^2/2$$

The mean scalar field has three critical points, a saddle at $(0, 0)$ and two maximum points at $(\sqrt{0.75}, \sqrt{0.75})$ and $(-\sqrt{0.75}, -\sqrt{0.75})$.

Figure 4.9 (a) illustrates the refined positional indicators when only the first 50 ensemble members are considered. The critical points occur primarily in the regions emphasized by the indicators, especially near the three critical points of the mean field. Regarding the type stability, several grid points situated around the two maxima of the mean field show stable maximum behavior, i.e., distinctly positive determinant and negative trace values (see Figure 4.9 (b)). Other grid points have only definitely negative trace values, which means that saddle points could also emerge. There are no grid points exhibiting stable minimum behavior (see Figure 4.9 (c)). The clearly positive determinant values shown in the figure exclude saddle points only and, as can be seen in Figure 4.9 (b), indicate maximum behavior. In fact, minimum points appear only around the origin, as part of saddle-minimum-saddle configurations, and there are just a few such instances. Although critical points occur around the origin in every ensemble member, which results in positive positional indicators, many of them are saddle points. Since neither distinctly negative determinant values, nor distinctly positive trace values are found, the type behavior at the origin is unstable. Locations with small trace values ($\tau = 0.1$), where saddle points are predominantly expected, are shown in Figure 4.9 (d).

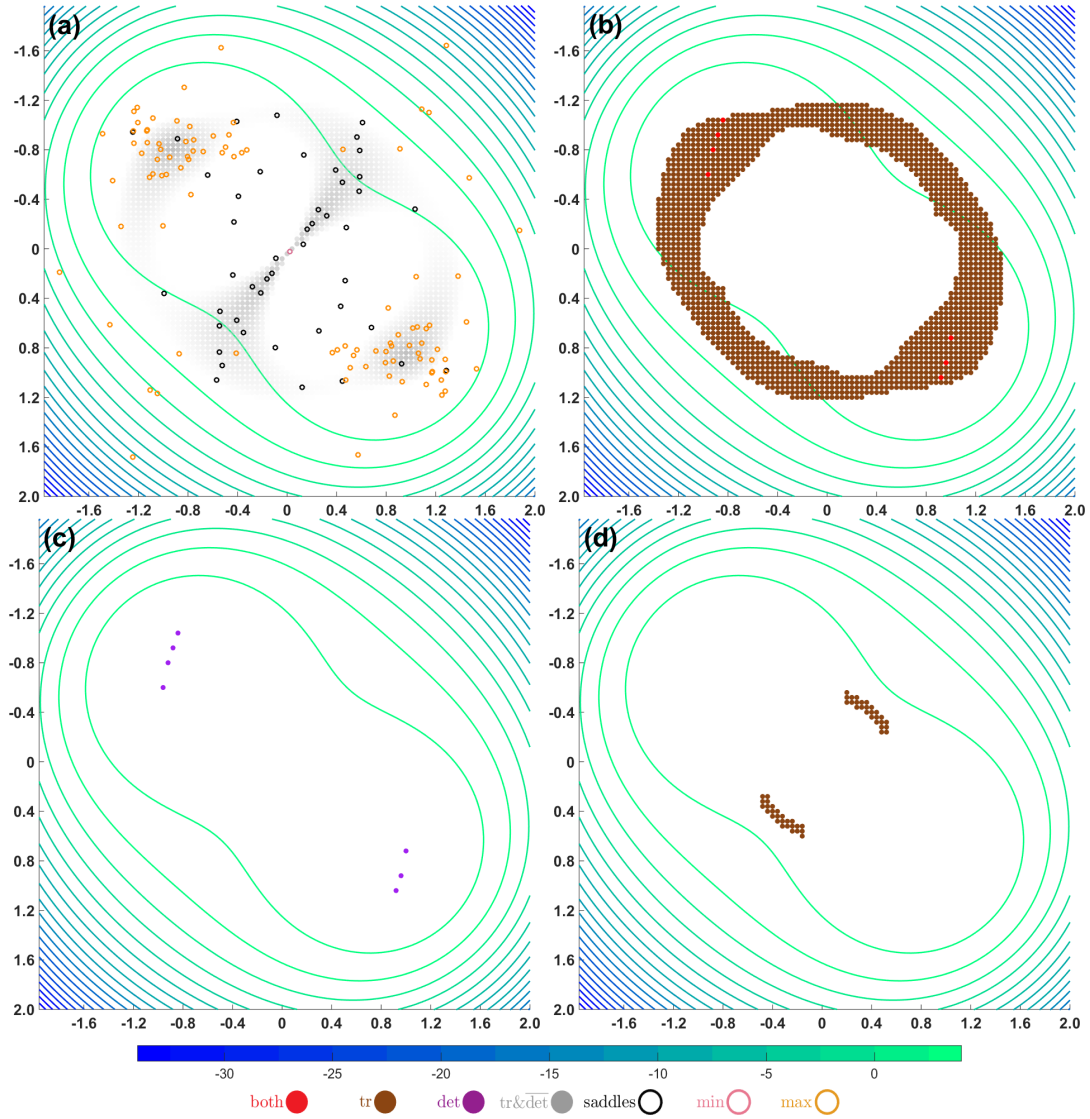


Figure 4.9: First synthetic ensemble (with isocontours of mean scalar field in the background): (a) refined positional indicators with critical points; type indicators for (b) maxima, (c) minima, and (d) saddle points.

Table 4.2 shows the false negative error rates. The four grid points at the peak of the histogram are located near the origin, where critical points emerge in every ensemble member. The two vertices on the secondary diagonal of the cell have positive indicator values, since critical points tend to occur mostly around them (see Figure 4.9 (a)). The other two vertices are assigned critical points because of being part of the grid cell. Other critical points that are not

H	1	0.04	0.02
n	4	50	476
r	0.5	0.48	0.40

Table 4.2: False negative error rates for the first synthetic ensemble (50 ensemble members).

N	50	100	500	1000	2500	5000
R	0.84	0.71	0.33	0.15	0.02	0.001

Table 4.3: False positive error rates for the first synthetic ensemble. N - number of ensemble members; R - false positive error rate.

in the vicinity of the origin are quite scattered, so that no other domain location is assigned a critical point in more than two ensemble members. Even so, grid points around which more critical points appear have positive indicator values.

The false positive error rate is very high (84%) for the first 50 members, because critical points are quite scattered. Nonetheless, the grid vertices with positive positional indicators around which no critical points emerge are suggestive of locations where critical points are likely in further ensemble members that could be realized. This can be observed when considering all the 5000 ensemble members, which brings about a dramatic decrease to 0.1% in the false positive rate. While critical points remain scattered to a certain extent (except for the locations near the origin, no other grid point is assigned a critical point in more than 64 ensemble members), they cover more densely the regions marked by the indicators. Increasing the ensemble cardinality does not result in any modifications at the peak of the histogram; this shows that the tendency of critical points to emerge on the secondary diagonal of the square had been captured well by the first 50 ensemble members. The decreasing trend of the error rate as more ensemble members are considered is given in Table 4.3.

A second synthetic ensemble, also of grid dimensions 100 x 100, was generated by drawing random numbers from a normal distribution with a zero mean and a standard deviation of 2 for the parameter ω in

$$x^3 + y^3 - \omega xy - x - y, \quad (x, y) \in [-2, 2] \times [-2, 2]$$

Depending on the value of ω, there are three types of scenarios for the critical points

- four critical points (two nodes and two saddles) when $-2 < \omega < 2$

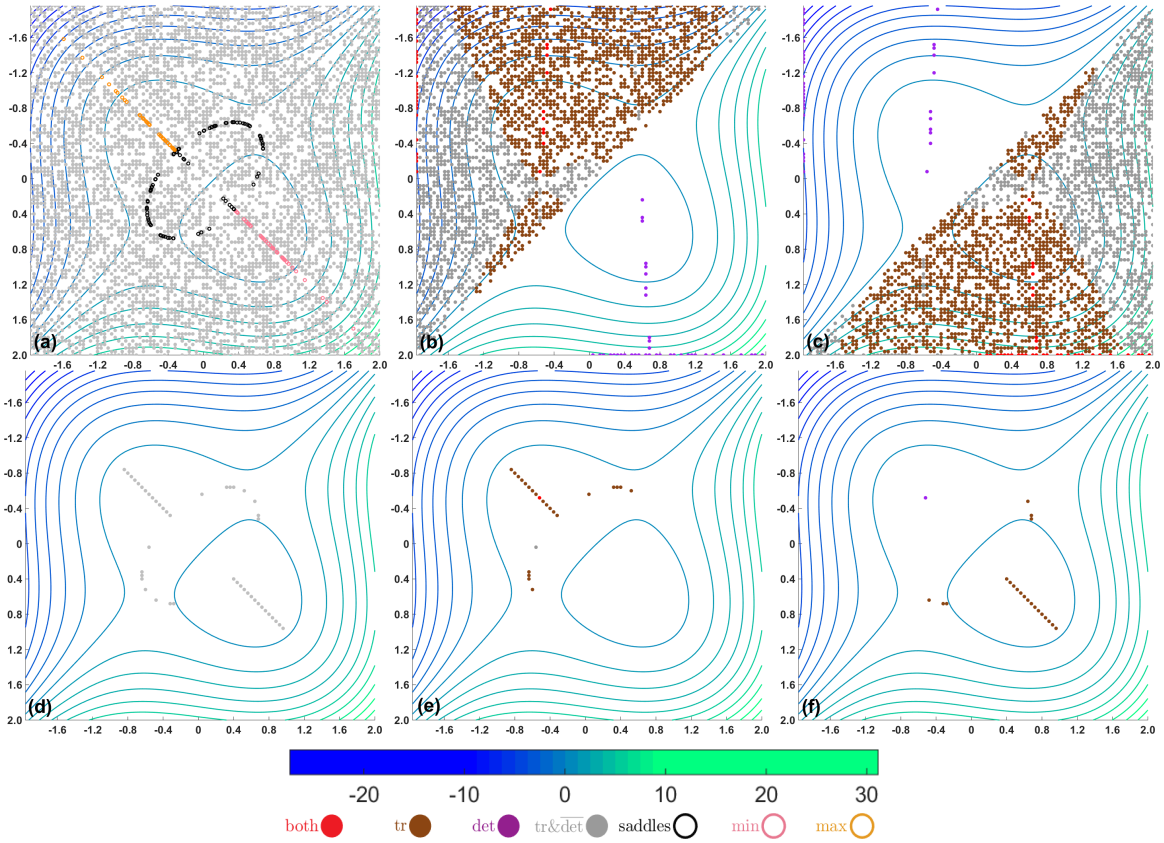


Figure 4.10: Second synthetic ensemble (with isocontours of mean scalar field in the background): (a) positional indicators and all critical points; type indicators for (b) maxima and (c) minima; (d) modified positional indicators; type indicators for (e) maxima and (f) minima after modification.

- three critical points (two nodes and a saddle) when $\omega = \pm 2$
- two critical points (a node and a saddle) when $\omega < -2$ or $2 < \omega$

Figures 4.10 (a)-(c) display the positional and type indicators over the isocontours of the mean scalar field. The critical points of the ensemble are shown in Figure 4.10 (a). It can be noticed that the ill-conditioned covariance matrices of the gradients yield spurious results for the positional indicators, which cover the entire domain with no distinct pattern. However, the type indicators seem to be not affected (see Figures 4.10 (b) and (c)), as these still identify the maximum behavior in the upper triangular part of the domain and the minimum behavior in its lower triangular part.

H	n	r	H	n	r
0.24	1	0	0.12	9	0.44
0.22	1	0	0.1	15	0.87
0.2	3	0.67	0.08	38	0.76
0.18	1	0	0.06	47	0.81
0.16	3	0	0.04	46	0.89
0.14	1	0	0.02	97	0.99

Table 4.4: False negative error rates for the second synthetic ensemble.

Recomputing the positional indicators by projecting the covariance matrix (see Equation 4.11) emphasizes locations where critical points are identified (see Figures 4.10 (d)-(f)). The false positive (13% for 45 grid points to which critical points have been assigned) and negative error rates (see Table 4.4) are also quite low.

The last ensemble is an ECMWF ensemble comprising 50 members and a control run. The forecast, initialized on October 17, 2012, is the 60 hour forecast of the geopotential height field at a pressure level of 1000 hPa, valid on October 19, 2012. Figure 4.11 displays the geolocated mean scalar field, with low altitudes of the pressure surface colored in shades of blue and high altitudes in shades of green. The positional indicators for a confidence region corresponding to a 95.4% confidence level are shown in Figure 4.11 (a), along with the critical points of the ensemble.

In this context, critical points are beneficial in helping identify features pertaining to severe weather events, such as cyclonic centers. Detecting well-defined geopotential minima is one of the techniques used to locate cyclonic features. As meteorological data is inherently affected by uncertainty, it is important to highlight those regions where pressure minima are likely to appear.

In this particular example, numerous critical points cluster south of Greenland, where a distinct low pressure system is visible. The positional indicators also emphasize this region of low pressure, along with several other areas where critical points can be seen to occur. False negative error rates are therefore correspondingly low, most of them under 20%; the false positive error rate is 37%. The regions displaying stable minimum behavior according to the type stability analysis are shown in Figure 4.11 (b). It can be observed that several grid points have distinctly positive trace values in the upper left corner of the low pressure region. However, no clearly positive determinant values are found, while five of the grid points display

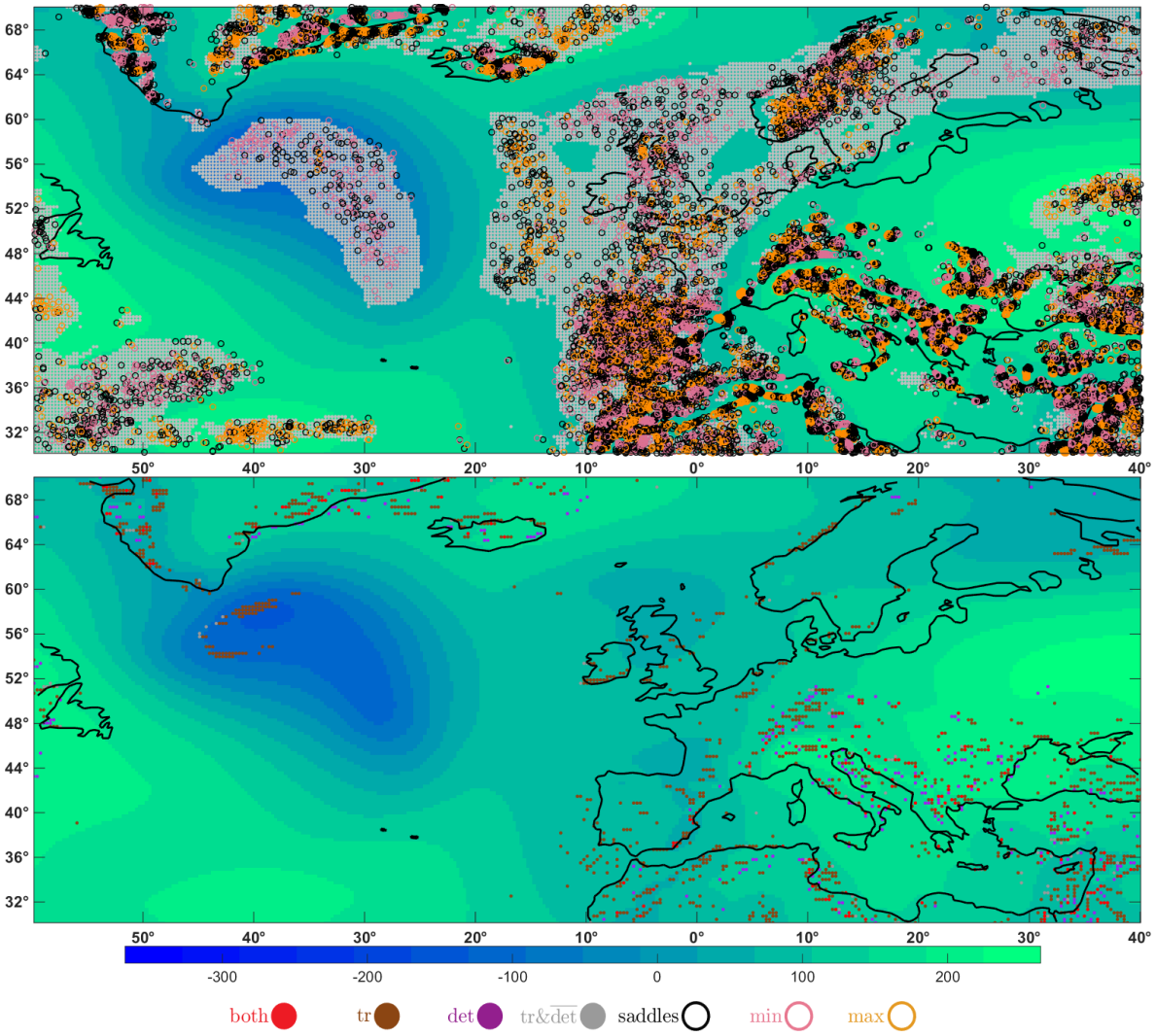


Figure 4.11: Mean scalar field of ECMWF geopotential ensemble with (a) binary indicators and all critical points, and (b) type indicators for minima.

even distinctly negative determinant values. This implies that saddle points are also likely to appear in the region along with minima, especially in the vicinity of the five grid points with negative determinant values.

4.5 Conclusion and Future Directions

Salient features in scalar ensembles are inherently affected by uncertainty, which influences their locations in the domain and their shapes. In this chapter, several indicator functions were proposed to offer insight into the prominent features of scalar ensembles and their stability, by analyzing their associated critical points. To this purpose, the scalar ensembles were summarized statistically and uncertainty was propagated to obtain confidence intervals/regions for the associated gradient fields, and the determinant and trace of the Hessian matrices. The confidence regions of the gradients were used to derive positional indicators highlighting regions where critical points are likely to emerge. The confidence intervals of the trace and determinant of the Hessian matrix helped derive type indicators revealing regions where critical points of certain types are likely to appear. Determining the locations of critical points and their type is important, because critical points are connected to the topological changes of surface components, which appear at minima, disappear at maxima, and split or merge at saddle points. The proposed indicators were then applied and validated on two synthetic and two real world ensembles.

Possible future directions include applying such a variability analysis to uncertain vector fields. Then, the stability concepts for critical points could also be extended, to allow tracking critical points (and their associated features) from one ensemble member to another.

5

Comparative Visual Analysis of Vector Field Ensembles

The previous chapter addressed different aspects of the effect of uncertainty on features in scalar ensembles. Analyzing the variability induced by uncertainty on ensembles of vector fields is more challenging than in the scalar case, since the dimensionality is higher for both the data to be visualized and its associated variability. This chapter deals with the variability of vector field ensembles from a local perspective.

Exploring the variability of ensembles of vector fields is important. For instance, in meteorology, a number of meteorological features, such as cyclones (low pressure systems), anticyclones (high pressure systems), and jet streams (strong upper level winds) become manifest in the wind fields. Variations in wind fields may thus become an important source of uncertainty. For example, poor forecasts of high impact weather events such as cyclones may be traced back to disturbances in wind fields.

When ensemble members no longer behave similarly, it is helpful to determine the locations where disturbances occur and which members, i.e., which physical models or input parameter settings, caused these disturbances. The visual analysis approach proposed in this chapter is intended to support a local exploratory investigation of ensembles of vector fields, which also considers global aspects. The techniques are tailored for the visual analysis of *directional data* in 2D and 3D vector field ensembles, but the main concepts and manner of proceeding are generic and can be applied for ensemble data in general.

The approach builds on approximating directional distributions locally using mixture models, which allows summarizing complex distributions by concise sets of parameters for each mixture

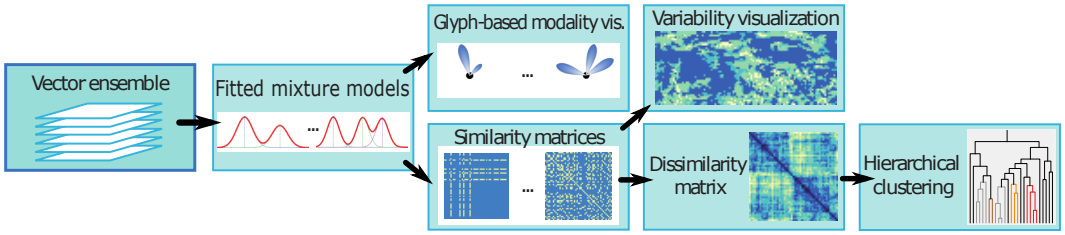


Figure 5.1: Directional distributions of vector-valued ensembles are approximated using mixture models. The resulting compact representations are used locally to design glyphs for visualizing main directional trends, and derive similarity matrices for the ensemble members. Matrices help reveal locations of (dis)similar behavior, and cluster ensemble members based on their similarity extent.

component (its mean direction, variation around its mean, and its weight). These sets of parameters are much easier to analyze and visualize than the distributions themselves. The mixture components can thus be used to perform a local similarity analysis of the ensemble members based on their membership to these modes and distinguish between locations of low or high variability. The initially local analysis can be extended over the whole domain by aggregating the local similarity measures, which allows ensemble members to be clustered depending on their degree of similarity throughout the domain. The different computational stages are illustrated in Figure 5.1.

The mixture models fitted at each domain point of a Cartesian grid structure constitute the basis of the major components of the proposed approaches to explore the variability of vector field ensembles:

- *Glyph-based visualization of directional distributions*: Lobular glyphs are used to visualize in an intuitive manner the main directional trends and their variability over selected subregions. Each glyph consists of a number of lobes equal to the number of components in the corresponding mixture model; the mean direction, angular variation, and component weight are mapped visually to the shape and orientation of the corresponding lobe. To allow displaying glyphs over the entire domain and still permit lobes to be clearly perceived, a Level of Detail (LOD) approach is also available.

- *Clustering*: The mixture models are also applied to derive similarity measures that describe the angular similarity at every domain location; the similarity measures are then used to cluster ensemble members hierarchically based on their degree of similarity

across the domain. This helps reveal outlier members, as well as the relationships between groups of two or more members regarding their membership to local trends.

- *Brushing and linking*: An interactive user interface with multiple linked views enables the simultaneous visualization of aggregated global and detailed local variations. According to the visualization information-seeking mantra "Overview first, zoom and filter, then details-on-demand" (Shneiderman (1996)), global variability information is visualized first, followed by means to interactively zoom into the data and steer the analysis towards interesting regions. Ensemble members can be selected, either individually – to visualize their behavior relative to the main directional trends – or in groups – to highlight regions of similar and dissimilar behavior.

5.1 Modeling Directional Data

As discussed in Section 3.2, there are several established approaches to estimate, e.g., directional distributions:

- *Parametric approaches*: Directional data is modeled using PDFs on the circle, such as the Wrapped Normal (WN) or the von Mises (vM) distributions (Fisher (1995)). This allows summarizing data using circular statistical parameters similar to the mean and standard deviation characteristic of a Gaussian distribution[1]. However, the PDF of directional data is often too complex to be properly approximated by a single such distribution; these statistical parameters are then misleading, e.g., for a multimodal distribution, the mean is no longer representative for the data sample.

- *Non-parametric approaches*: The distribution is estimated directly from the data sample, using angular histograms, such as rose diagrams, or smoother estimates, such as the kernel density estimate. For such approaches, the choice of the smoothing parameter plays a major role in the ultimate shape of the distribution, potentially leading to an under- or oversmoothing of the PDF (see Figures 5.2 (a) and (b)). Although there is no universally accepted rule for determining a "best" smoothing parameter, several methods have been proposed that are effective for a wide range of distributions. Nevertheless, deriving additional quantitative information is not trivial. A visual inspection

[1]The mathematical formulae that need to be evaluated are, however, more complicated than for the Gaussian distribution, requiring often iterative procedures or more involved computations (Fisher (1995)).

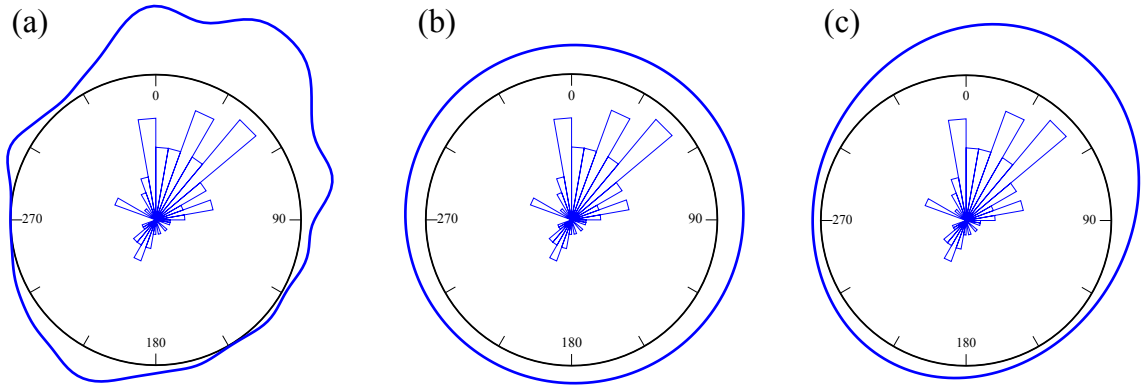


Figure 5.2: Density approximations (wrapped around the unit circle) for a bimodal set of directions taken by 76 turtles after treatment (see Fisher (1995) for more details on the directional data set); rose diagrams are shown inside every circle; (a)-(b) kernel density estimates with different smoothing parameters, where (a) undersmoothing creates several minor modes, while (b) oversmoothing obscures relevant structures; (c) mixture of normal distributions.

of the number of "bumps" or the local peaks of the distribution may allow assessing the modality of the distribution, but determining automatically the actual number of significant peaks, their locations, and widths is not straightforward.

- *Semi-parametric approaches*: Parametric mixture models (see Figure 5.2 (c)) combine the advantages of the two methods, by providing a typically manageable set of parameters that offer good approximations of complex distribution shapes. While care must be taken to identify the appropriate number of mixture components, the mixture modes and their statistical summaries are already available and do not have to be determined additionally. For such reasons, mixture models are the preferred approach in this thesis.

5.1.1 Modeling 2D Directional Data

The parametric models initially considered for the mixture modeling were the vM distribution and the WN – the circular analogue of the normal distribution.

To fit vM mixtures, a stepwise procedure – testing for n components against more than n (Fisher (1995)) – helps determine the number of vM components in the mixture. Once the

modality is known, a recent EM algorithm (Banerjee et al. (2005)) is used to estimate the parameters of the vM modes.

Regarding WN mixtures, because the Wrapped Normal distribution $WN(\mu, \rho)$ can be obtained by wrapping around the circle the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ given on the line

$$\rho = \exp\left(-\frac{1}{2}\sigma^2\right), \quad \text{where} \quad \sigma^2 = -2\log(\rho) \tag{5.1}$$

and fitting mixtures of Gaussians (unlike fitting mixtures of WNs) is tractable, directional data is modeled by using GMMs for unwrapped data and reverting to WNs via Equation 5.1 whenever circular statistics are necessary.

In a first stage, mixtures of both vM and Gaussian distributions were fitted to the data. Because mixture models algorithms start with random initial guesses, depending on the initial configuration, solutions may converge to local optima and yield different structures. Typical solutions to alleviate this problem perform several fittings with different starting values and choose the best solution.

As mentioned in Section 3.2, fitted mixtures are assessed via validation techniques suitable for algorithms having random initial guesses, such as silhouettes. The validation considers only the resulting partitions, independent of the algorithm that generated them. In this case, the best solution was selected as the one with the highest average silhouette from those solutions with the most frequently occurring modalities (number of components).

To discover whether any of the two mixture models were consistently more non-repetitive on the tested data sets, the fitting algorithms (including the repetitions) were applied ten times at every grid point in the domain and corresponding scalar fields holding the modalities of the fitted mixtures were derived for both procedures. While a detailed comparison of the two algorithms is beyond the scope of this thesis, for the tested ensembles, the GMM yielded much more consistent results. The modalities repeated consistently at approximately 98% of the total number of grid points, compared to 50% for the vM mixtures. Given the small number of non-repetitive modality instances and that the spatial neighborhoods of such grid points can also be used to infer on a possible under- or overestimated number of modes, without loss of generality, the GMM was the algorithm of choice for modeling 2D directional data.

Modelling 2D Directional Data using GMMs

Circular data can be regarded either as unit vectors in the plane $\mathbb{R}^2$ or as points on the unit circle (Mardia and Jupp (2000)). In the latter case, selecting an initial direction, e.g., the x-axis, and an orientation, e.g., counterclockwise, allows specifying each data observation as the angle from the axis to the point on the unit circle.

Modeling the PDF of a sample $\theta_1, ..., \theta_N$ using a mixture of n Gaussian components, each parameterized by a mean vector μ_i and standard deviation σ_i^2, and with α_i as weight, yields

$$f(\theta) = \sum_{i=1}^{n} \alpha_i \mathcal{N}\left(\mu_i, \sigma_i^2\right)(\theta), \quad \alpha_i > 0, \ \sum_{i=1}^{n} \alpha_i = 1 \tag{5.2}$$

The GMM provides a mixture of Gaussian modes from which the observations are assumed to be drawn in varying proportions (given by the weights). Such partitioning algorithms, however, tend to find clusters in the data even when the observations might be uniformly distributed. Uniform distributions may occur, for example, around critical points, where the corresponding vectors are zero. Since no directions can be defined in such cases, corresponding observations should be modeled by components of their own and not be considered when fitting mixtures. However, as there were no instances of zero vectors appearing exactly at the grid points for any of the data sets tested in this work, this case is not addressed here in further details. Instead, before fitting a mixture of Gaussians, a statistical test of randomness is performed first. The omnibus test (Fisher (1995)) is used to test against any non-uniform alternative model. If the hypothesis of randomness can be rejected at the 5% significance level, the GMM algorithm described in Section 3.2 is applied to determine the number and shape of the components.

To apply GMMs, however, the directional data needs to be unwrapped on the real line. The straightforward approach would be to cut the circle at a suitable point and unwrap it to an interval of length 2π. Which cutting point is chosen may have a significant impact on the unwrapped data (see Figure 5.3 (a) and (b)): cutting near (b) a mode rather than its (a) antimode creates the wrong impression of two modes at the ends of the interval.

Instead of cutting the unit circle, Mardia and Jupp (2000) suggest repeating a complete cycle of the data to eliminate the ambiguity about the modality. This recommendation is followed by Wu et al. (2012), who pad circular data to fit mixture models to wave direction data using a standard variational Bayesian approach and an initially overestimated number of components.

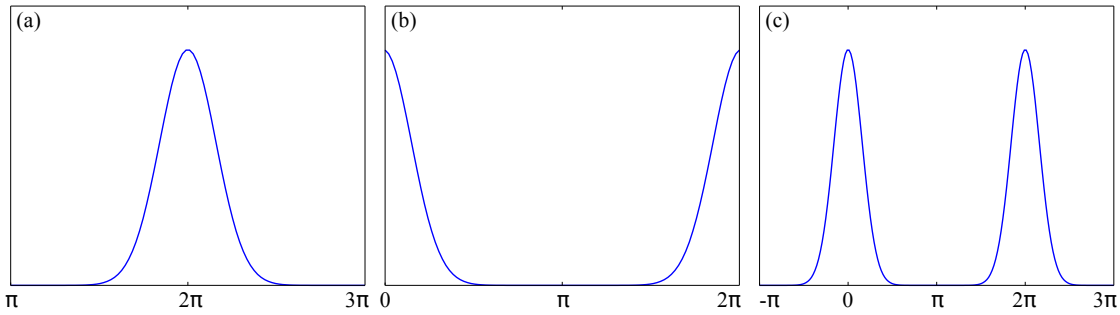


Figure 5.3: Cutting a circular distribution with a mode at 0 (a) at the antimode π conveys the mode clearly, while cutting it (b) at the mode creates the impression of bimodality. The ambiguity about the modality is resolved by (c) repeating the complete cycle and padding each half to the left and right.

By duplicating the data observations in the $(0, \pi]$ interval to $(2\pi, 3\pi]$ and those in $[\pi, 2\pi)$ to $[-\pi, 0)$ (see Figure 5.3 (c)), the original data set defined over $[0, 2\pi]$ is extended to $[-\pi, 3\pi]$. The techniques proposed in this thesis follow the same approach of padding the data at both ends; GMM is then applied on the augmented data set.

Once a GMM has been fitted to the padded data, the extended interval $[-\pi, 3\pi]$ is restricted to the initial interval $[0, 2\pi]$ and the final components are summarized on the unit circle. Restricting to the original interval involves

- grouping initial observations from two modes into one component (like for the two modes in Figure 5.3 (c), where the data values cluster around 0 and 2π on the line)
- discarding the modes around values not originally in the data set and keeping only the modes with values in the restricted data set

To this purpose, a merging step is performed: for every pair of original and duplicated observations, the corresponding pair of components to which both observations can belong with posterior probabilities over a given threshold is determined. From the thus derived pairs of components, merged are only those pairs $(\mathcal{N}(\mu_i, \sigma_i^2), \mathcal{N}(\mu_j, \sigma_j^2))$ with means located close to each other on the unit circle. The distance between the means is determined by

$$d(\mu_i, \mu_j) = \pi - |\pi - |\mu_i - \mu_j|| \tag{5.3}$$

Components that are not merged and do not model any initial observations are discarded. Finally, all means are restricted to values in the interval $[0, 2\pi]$.

The resulting components are summarized via sample trigonometric moments. For a component modeling m angles, the unit vector representation of the observations is used. The mean direction μ is then computed as the direction of the resultant vector after vector addition (Fisher (1995))

$$\mu = \begin{cases} \tan^{-1}(S/C) & \text{if } S > 0 \text{ and } C > 0 \\ \tan^{-1}(S/C) + \pi & \text{if } C < 0 \\ \tan^{-1}(S/C) + 2\pi & \text{if } S < 0 \text{ and } C > 0 \end{cases} \tag{5.4}$$

where

$$C = \Sigma_{i=1}^{m} \cos(\theta_i) \quad \text{and} \quad S = \Sigma_{i=1}^{m} \sin(\theta_i)$$

The mean resultant length ρ associated to the mean direction μ is the length of the resultant vector, normalized by m,

$$\rho = \frac{\sqrt{C^2 + S^2}}{m} \tag{5.5}$$

and takes values in the range $[0, 1]$ – the closer to 1 ρ is, the more concentrated the observations are around the mean direction (Mardia and Jupp (2000)). Equation 5.1 can then be used to compute the corresponding parameters of the associated Gaussian distribution.

5.1.2 Modeling 3D Directional Data

Analogously to the 2D case, spherical data consists of points on the unit sphere, which can be viewed as unit vectors in $\mathbb{R}^3$. If circular data can be unwrapped, this is not possible for spherical data, since a sphere cannot be mapped on the plane without distortion. Instead of fitting GMMs, spherical data is modeled using a mixture of von Mises Fisher (vMF) distributions – the equivalent of the Gaussian distribution on the sphere.

A unit vector v is said to follow a mixture model of vMFs if its PDF is given by

$$f(\mathrm{v}) = \sum_{i=1}^{n} \alpha_i \frac{\kappa_i}{4\pi \sinh(\kappa_i)} \exp\left(\kappa_i \mu_i^T \mathrm{v}\right), \quad \alpha_i > 0, \quad \sum_{i=1}^{n} \alpha_i = 1, \tag{5.6}$$

where n is, as before, the number of vMF components, the unit vectors μ_i are their mean directions, and κ_i the concentration parameters. Similarly to the variation in the Gaussian

case, the concentration parameter determines the shape of the distribution. Thus, $\kappa = 0$ stands for a uniform distribution, whereas higher values indicate stronger concentrations around the mean directions.

The parameters of the vMF mixture are estimated using the EM soft-moVMF algorithm introduced by Banerjee et al. (2005). While there is no analytic expression for the concentration parameter, the approximation proposed by the authors helps avoid iterative schemes for estimating κ, decreasing thus the computational load.

The main steps of fitting a mixture of vMFs to spherical data are similar to the circular case. First a test of randomness is performed, as proposed by Rayleigh (Fisher et al. (1987)). If the hypothesis of uniformity is rejected at the 5% significance level, the performed procedure wraps around the soft-moVMF algorithm, by splitting the data into components in a hierarchical fashion, until every component can be assumed unimodal. When no further splits are performed, the soft-moVMF is run again to refine the solution.

The decision whether a component of size m can be assumed unimodal is based upon an exploratory analysis of the shape of the distribution. For the unit vectors $\mathrm{v}_i = (x_i, y_i, z_i)$, it involves the mean resultant length (denoted here as $\overline{R}$) and the normalized eigenvalues (τ_1, τ_2, τ_3) of the orientation matrix T

$$T = \begin{pmatrix} \sum_i x_i^2 & \sum_i x_i y_i & \sum_i x_i z_i \\ \sum_i x_i y_i & \sum_i y_i^2 & \sum_i y_i z_i \\ \sum_i x_i z_i & \sum_i y_i z_i & \sum_i z_i^2 \end{pmatrix} \tag{5.7}$$

where the eigenvalues are normalized by m (Fisher et al. (1987)).

A component is approximated as unimodal if $\overline{R} \approx 1$, i.e., the vectors cluster together, and τ_3 is large, while τ_1 and τ_2 are small (in the current implementation if the logarithms of the ratios τ_3/τ_2 and τ_3/τ_1 are supraunitary).

Once a mixture of components has been fitted, the last step is to derive confidence cones around the mean directions μ. How the confidence cone is determined depends on the size of the component (Fisher et al. (1987)).

For components of moderate to large size ($m \geq 25$), this amounts to computing the estimated

spherical standard error $\tilde{\sigma}$ of the sample mean direction

$$\tilde{\sigma}^2 = d/\left(m\overline{R}^2\right), \text{where } d = 1 - \frac{1}{m}\sum_{i=1}^{m}(\mu \cdot \mathrm{v}_i)^2 \tag{5.8}$$

A 68% confidence cone for μ has a semi-vertical angle equal to

$$q = \arcsin(1.0674\tilde{\sigma})$$

For smaller sizes ($m < 25$), a bootstrap technique is necessary to derive a confidence interval for μ. To this purpose, it is necessary to simulate a large number, e.g., 200, of samples of size m from the corresponding vMF component, and estimate the mean direction for each sample. Details on how to simulate data from a vMF distribution can be found in Fisher et al. (1987). Sorting the bootstrap estimates in increasing order of their angle to the mean, the 68% confidence cone centered on μ is approximated by the cone containing 68% of the simulated mean directions, in this case 136 estimates. The angle between the 33rd smallest and largest mean is computed as the arccos of their dot product.

5.2 Comparative Analysis of Ensemble Members Based on Fitted Mixture Models

The mixture models of the directional distributions can be used not only to summarize the main directional trends, but also to derive local comparative measures for the ensemble members and reveal the locations of (dis)similar behavior in the spatial domain. These local measures can then be accumulated and used to cluster ensemble members depending on the extent of their similarity across the domain. For the sake of clarity, this section focuses on the 2D case only, but similar derivations are possible in 3D.

5.2.1 Local Comparative Measures

Local comparative measures are intended to assess the direction similarity of two ensemble members at every domain location. Various measures can be used to this purpose, such as the angle between the vectors or their cosine similarity. In this context, ensemble members can be classified as angularly similar if, for instance, the selected measure is below a prescribed

threshold. These prescribed thresholds are usually arbitrary or the result of trial and error. Instead of imposing an arbitrary threshold, the derived mixture models can be used to assess pairwise similarities locally relative to the variability of all ensemble members at the given location.

At every grid point in the domain, the N directional observations $\theta_1, ..., \theta_N$ are modeled by a mixture of n Gaussian components, each observation θ_i having a certain posterior probability p_{ik} of having been drawn from mode k. This can be summarized in a $N \times n$ matrix P.

The first local similarity measure – the *modality measure* – considers two ensemble members as similar (at a given location) if their realizations θ_i and θ_j have posterior probabilities over a given threshold τ of belonging to the same mode k

$$\mathrm{m}(\theta_i, \theta_j) = \begin{cases} 1 & \text{if } p_{ik} \geq \tau \text{ and } p_{jk} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

At every location, the binary similarity values can be summarized into an $N \times N$ symmetric similarity matrix. Soft clustering is used, so that members located at the borders of two modes are not artificially classified as similar to the members of one mode only; the ensemble members belonging clearly to just one cluster are not affected.

This first measure differentiates between members that clearly belong to different modes. It gives, however, no insight into how similar the members are within the modes. For instance, ensemble members where the fitted distributions are unimodal are all classified as similar, even if some observations are far from the main data mass. To refine the pairwise characterizations as similar or dissimilar, a second similarity measure – the *scaled angular measure* – is proposed: for every pair of ensemble members whose observations θ_i and θ_j are assessed as similar according to the modality measure, the mode k which both observations are most likely to belong to is identified as the component with the highest sum of posterior probabilities $p_{ik} + p_{jk}$; the members are considered similar if the smallest angle between them[2] is less than or equal to the maximum sample circular standard deviation[3] in the domain

$$\mathrm{sm}(\theta_i, \theta_i) = \begin{cases} 1 & \text{if } \mathrm{m}(\theta_i, \theta_j) = 1 \text{ and } d(\theta_i, \theta_j) \leq \sigma_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

[2]The smallest angle between two observations is computed like in Equation 5.3.
[3]The sample circular standard deviation is, in this particular case, identical to σ (see Equation 5.1).

In this way, members having large angular deviations are no longer considered similar, even if they are likely to have been drawn from the same component.

To also address the relatively large deviations occurring in regions with low variation, the angular deviation can be computed relative to the local variation instead of to the global one. In other words, for the third measure – the *locally scaled angular measure* – the angular deviation is considered relative to the local sample circular standard deviation

$$\mathrm{lsm}(\theta_i, \theta_i) = \begin{cases} 1 & \text{if } \mathrm{m}(\theta_i, \theta_j) = 1 \text{ and } d(\theta_i, \theta_j) \leq \sigma_k \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

Domain locations exhibiting higher angular variability allow larger deviations than those with lower variability. For uniform distributions, where all directions are equally likely, two realizations are considered dissimilar if their angular deviation is higher than a prescribed threshold, e.g., $\pi/2$.

Three synthetic examples illustrating the use of the proposed similarity measures are demonstrated in Figure 5.4. The first column shows the glyph representations (see Section 5.3 for details on the glyph construction) of the corresponding PDFs, along with three selected ensemble members, colored in red in the figure; the following three columns display the 3×3 similarity matrices for the selected members, one column for each proposed local similarity measure.

Assessing members based on their mode membership (second column), classifies the members of both unimodal PDFs (Figure 5.4 (a)-(b)) as similar. For the bimodal PDF (Figure 5.4 (c)), ensemble members 1 and 3 are clearly modeled by different modes and are thus classified as dissimilar. Member 2, situated at the border of the two clusters, is similar to both 1 and 3. Considering the angular deviations relative to the highest variability (third column), several other pairs of members emerge as dissimilar: in Figure 5.4 (b), member 3 (an outlier) is now dissimilar to ensemble members 1 and 2 (both inliers); moreover, in Figure 5.4 (c), members 2 and 3, located near the opposite ends of the wider mode, have also become dissimilar. Relative to the local variations (fourth column), members 2 and 3 in Figure 5.4 (a), as well as 1 and 2 in (c) are additionally classified as dissimilar. Even though members 1 and 2 in Figure 5.4 (b) have approximately the same angle as 2 and 3 in (a), their angular deviation is low relative to the local variation.

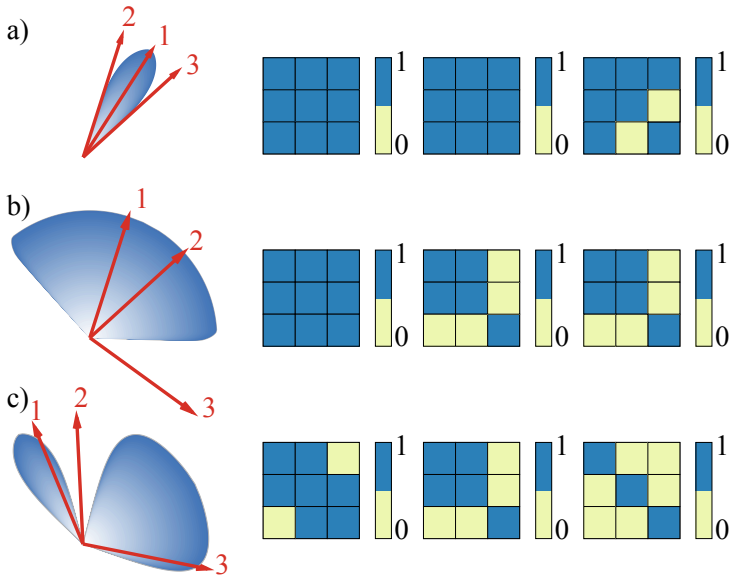


Figure 5.4: Similarity measures for selected members of three synthetic data sets. First column: selected members (red) and glyph representations of (a) a unimodal PDF showing low variation and (b) high variation, and (c) a bimodal PDF. The next three columns depict for each data set the similarity matrices corresponding to the three types of similarity measures.

5.2.2 Global Comparative Measures

The similarity measures can be summarized via $N \times N$ local similarity matrices at every grid point of the domain. These local matrices can be used to reveal the domain locations where two or more members behave dissimilarly and even quantify the global degree of dissimilarity.

To find the regions of different behavior for a pair of members (i, j), the binary value at row i and column j is extracted from the local similarity matrix at every grid point, yielding a binary field of similar and dissimilar locations. When more than two members are considered, values range between 0 and 1; the exact value depends on how many pairs of members are similar, out of the total number of pairs ($1 + k(k-1)/2$ for k members).

To derive global measures of (dis)similarity for any pair of members (i, j), the corresponding binary field can be aggregated to yield the global similarity of these members, as the percentage of similar locations from the total number of grid points. The pairwise global similarities

of every pair of members can be then summarized into a single representation – the *global similarity matrix*.

Based on the global similarity matrix, ensemble members are clustered in a hierarchical manner, building a multilevel hierarchy of clusters. Hierarchical clustering (Jain and Dubes (1988)) starts with the pairs of members that are most similar and links clusters at one level into larger clusters at a next level until a hierarchical cluster tree is formed. Ultimately, any two members are joined at some level, the variability increasing as more clusters are joined. The linkage criterion, based on which the next clusters to be merged are chosen, is given by the similarity between two or more members. For pairs of members, this is their global dissimilarity (the complement of the global similarity), while for more than two members, the complemented global similarity value is obtained by aggregating similarity values (between 0 and 1) at all grid points into one value, which is normalized by the total number of points.

5.2.3 Comparing Mixture Models

In addition to revealing the locations of (dis)similar behavior and clustering similar members, other issues are of interest for ensemble data, namely how ensemble members behave at selected locations compared to the rest of the domain. In this context, queries can be run to answer a series of questions. For instance, querying the modality field – the scalar field holding the number of components at each location – reveals where in the domain the same modality is exhibited as at the selected location.

More specific queries, such as determining domain locations where PDFs have similar shapes, require means of comparing two PDFs. While there is no single best measure to assess the degree to which two PDFs are similar, there are several tools available, such as the Kullback-Leibler (KL) distance. To ensure that such queries are interactive, however, it is important that the selected measure has a closed form for GMMs and is easily computed on the fly. As the KL distance is analytically tractable only for two Gaussian distributions, the measure used here builds upon the *concordance coefficient* (Ray (2003)), which assesses the extent to which two PDFs agree. Thus, for two mixtures of Gaussians, $f_1 = \sum_{i=1}^{n_1} \alpha_{1i} \mathcal{N}(\mu_{1i}, \sigma_{1i}^2)$ and $f_2 = \sum_{j=1}^{n_2} \alpha_{2j} \mathcal{N}(\mu_{2j}, \sigma_{2j}^2)$, the concordance coefficient reads

$$C(f_1, f_2) = \frac{2 \int f_1(x) f_2(x) \mathrm{d}x}{\int f_1^2(x) \mathrm{d}x + \int f_2^2(x) \mathrm{d}x} = \frac{2F(f_1, f_2)}{F(f_1, f_1) + F(f_2, f_2)} \tag{5.12}$$

where

$$F(f_1, f_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{\alpha_{1i}\alpha_{2j}}{\sqrt{\sigma_{1i}^2 + \sigma_{2j}^2}} \exp\left(-\frac{1}{2}\frac{(\mu_{1i} - \mu_{2j})^2}{\sigma_{1i}^2 + \sigma_{2j}^2}\right) \tag{5.13}$$

For circular data, the variance is computed according to Equation 5.1, while the distance between two means can be determined using Equation 5.3. The concordance coefficient takes values between zero – when the two PDFs have completely dissimilar support sets – and one – when they are identical. In order to get values in $[0, \infty)$ instead of $[0, 1]$ and have, for constant variation, the measure grow rather than decrease with the distance between the means, the coefficient $D(f_1, f_2) = -\log(C(f_1, f_2))$ is used as a similarity measure. This similarity measure is nevertheless applied only when the PDFs belong to the same family, i.e., both mixtures of Gaussians. A uniform PDF is regarded as completely dissimilar to a GMM (infinite coefficient), while two uniform PDFs are considered identical (zero coefficient).

Besides querying to determine PDFs with the same modality and a similar shape, other aspects that could be explored include the extent to which ensemble members vary in the same manner elsewhere in the domain. For example, given a bimodal distribution at a selected location, are the subsets of members belonging to the two modes the same at other bimodal PDFs? Such questions can be answered by comparing the local similarity matrices at the other grid points with the matrix at the selected location and determining the percentage of common similarities from the total number of similarities.

5.3 Visual Analysis of Vector Ensembles

The main components of the interactive graphical user interface proposed for the comparative visual analysis of ensembles of directional data are shown in Figure 5.5. Visualizing the degree of (dis)similarity of the ensemble members occurs via four linked views, two over the spatial domain (overview (A) and detail view (B)) and two abstract representations of the similarities of the ensemble members (views (C) and (D)).

5.3.1 Member Similarity Visualization

Ensemble members are clustered hierarchically according to how similar their behavior is across the spatial domain. To highlight how members cluster together and enable a rapid

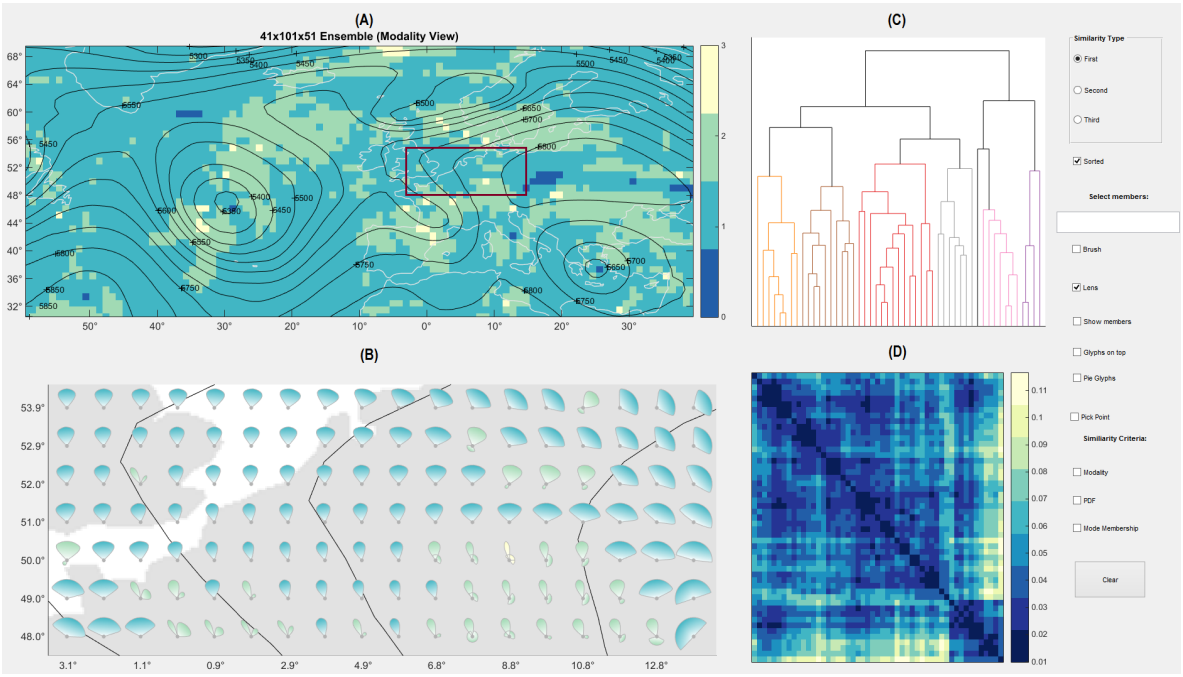


Figure 5.5: Multiple linked views of an ECMWF wind ensemble. View (A) color codes geolocated scalar fields over the entire domain (in this case, the modality of the distributions), while (B) shows a detail view over the region selected by the user (marked in A by a red rectangle), where directional distributions are displayed using lobular glyphs. The isocontours in the background show the geopotential height of the control run. The hierarchical clustering of the ensemble members is shown using a dendrogram in view (C), while the global pairwise dissimilarities between ensemble members are shown in the global dissimilarity matrix in view (D).

identification of outliers, the hierarchical clustering (performed for any of the three proposed similarity measures) is summarized graphically as a dendrogram (see Figure 5.5 (C)).

The joining of subclusters is represented using U-shaped lines, the horizontal axis of dendrogram showing the IDs of the ensemble members, and the vertical axis displaying the levels where subclusters are joined. The order of the members on the horizontal axis is chosen so as to maximize the sum of the similarities between adjacent leaves, but without dividing clusters. Furthermore, to distinguish better between the different groups of members, main neighboring branches are depicted in different colors.

Dendrograms are useful because they do not simply depict a single set of clusters, but show

through nested partitions how subclusters relate to one another. Furthermore, they highlight both similar members – joined at lower-level clusters – as well as outliers – members merged at high levels, where the variability among the members increases. Hierarchical clustering suffers nevertheless the drawback that the merging steps are greedy and constrained by previous choices. Thus, having been once clustered, ensemble members cannot be reclustered. This means that similar members may end up in different main clusters.

To alleviate this problem, a second abstract representation of how members are related is presented (see Figure 5.5 (D)), namely the global dissimilarity matrix (the complement of the global similarity matrix). As a result, even if similar members may be clustered in different groups, their similarity is revealed by the low pairwise dissimilarity value in the global dissimilarity matrix, which shows how similar or dissimilar any pair of ensemble members is. The dissimilarity values are shown using a sequential ColorBrewer scheme (Harrower and Brewer (2003)), ranging from dark blue (most similar) to light yellow (most dissimilar). By default, ensemble members in the global dissimilarity matrix are ordered like in the dendrogram, which makes the clusters stand out as bluish block-diagonal forms. Other orderings are, nonetheless, also supported, such as sorted by member ID.

5.3.2 Similarity Visualization over the Spatial Domain

According to the visualization information-seeking mantra, visualizing the ensemble variability is performed first at the highest level of abstraction. Thus, the default view over the entire domain shows the modality (number of components) of the mixture models. Other derived variability scalar fields can be displayed as well. For instance, by selecting one or more ensemble members, the main view updates instantaneously to show the dissimilarity scalar field of the corresponding members. Member selection is performed by picking a branch in the dendrogram, brushing cells in the matrix view, or specifying members as text input.

Visualizing non-scalar data such as the directional PDFs simultaneously over the entire domain would, however, lead to massive clutter and occlusion. To alleviate these problems and still reveal the main directional trends at several locations concurrently, in a detail view, users can either zoom interactively into a region of fixed dimensions (see Figure 5.5 (B)) or see the entire domain by means of a Level of Detail (LOD) approach (see Figure 5.17) that shows in an aggregated glyph the most representative distribution for several spatial locations.

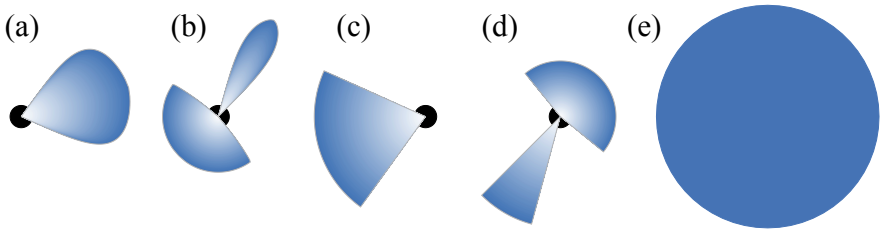


Figure 5.6: Glyphs for various directional distributions: lobular glyphs representing (a) a unimodal PDF, with medium variation, and (b) a bimodal PDF, with low variation (main mode) and high variation (second mode); (c)-(d) show corresponding pie glyphs; (e) glyph representing uniform PDF.

Glyph-based Visualization for 2D Directional Data

When visualizing distributions on the circle, it is important to bear in mind that directional information cannot be mapped intuitively to visual parameters like color or surface deformation. The shape of such circular distributions is easier to perceive when presented using a natural representation, such as glyphs given on the unit circle. Several techniques are available to display directional data in this manner, such as raw data plots around the circle, or various angular histograms and rose diagrams (see Figure 5.2). Nonetheless, as already explained in Section 3.2.2, the choice of boundaries and bin width can seriously distort the shape of the PDF.

Other popular methods estimate the PDFs and wrap them around the circle, as shown in Figure 5.2 for two kernel density estimates and a GMM. However, unless the displayed modes are very narrow, it may be difficult to identify the modes, their variation and contribution to the PDF, especially for simultaneous displays. Perceiving the modes could be achieved easier by a linear representation of the PDF, but the latter is less intuitive, as directional information cannot be obtained without analyzing the annotations.

To alleviate these issues and visualize circular distributions in a compact and intuitive manner, lobular glyphs are proposed to reveals the main characteristics of a potentially multimodal directional PDF: its number of components, as well as their mean directions, widths, and weights. A glyph (see Figures 5.6 (a)-(b)) consists of a number of lobes equal to the number of modes, where the mean direction, weight, and width of each mode are mapped visually to the orientation, length, and opening angle of the corresponding lobe.

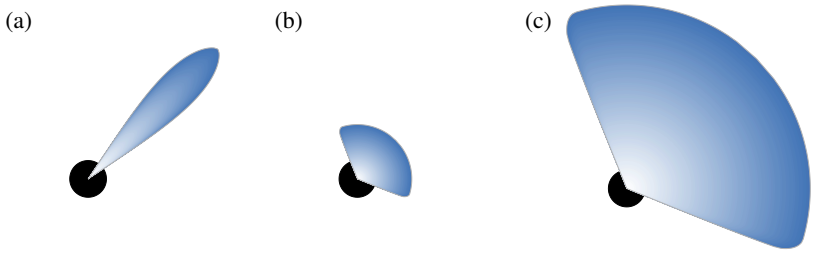


Figure 5.7: Lobular glyphs for unimodal PDF with (a) low variation and (b) high variation; the length of the lobe in (b) has been decreased so that the lobe has the same area as the lobe in (a). (c) Lobular glyph for the same PDF as in (b), where the length of the lobe was set, like in (a), to the maximum allowed, thus conveying the mode properly.

Compared to distributions wrapped around the circle, the length of a lobe encodes the weight of the mixture component, i.e., the percentage of observations modeled by the mode, rather than the height of the peak. This is done because the height of a peak depends not only on the weight of the mixture component, but also on the width of the peak, being higher for thinner peaks and decreasing as the width increases. Setting the length of each lobe to the height of the corresponding peak could be visually misleading for wide peaks, especially in simultaneous displays, since the perceived direction might appear orthogonal to the actual direction. Furthermore, the lobes would become too small for the uncertain regions to be prominent. By mapping the mixture weight to the length of the lobe, lobes of unimodal PDFs have the same length, representing the actual direction adequately and also emphasizing the uncertain regions (see Figure 5.7).

For lobes of multimodal distributions, the mixture weights were initially mapped directly to the lengths of the lobes. However, this made the corresponding glyphs appear significantly smaller relative to the glyphs representing unimodal and uniform PDFs (see Figure 5.8). Therefore, weights are normalized by the local maximum weight before being mapped to the length of the lobes. While this allows only a qualitative comparison between glyphs of distributions with different modality, it makes the glyphs of multimodal PDFs easily noticeable, so that uncertain regions are salient.

The lobular glyphs reveal the main directions in which the majority of the unit vectors point, as well as the variability around these directions. Meteorologists, however, frequently show the directions from which the winds blow as well. To also allow showing the directions from

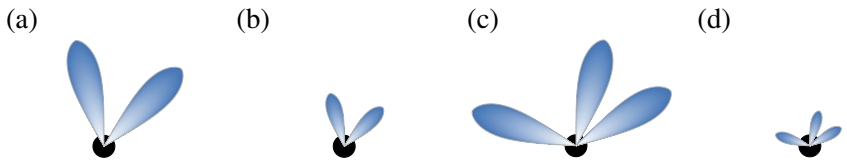


Figure 5.8: Lobular glyphs of a (a) bimodal and (c) trimodal PDF, with lobe lengths given by weights after normalization by the maximum weight. (b),(d) Lobular glyphs without weight normalization are considerably more difficult to distinguish.

rather than into which the vectors point, the lobular glyphs are replaced by pie glyphs (see Figure 5.6 (c)-(d)), where directions are read from the heads of the glyphs to the tails.

Comparing lobular to pie glyphs, the pointy head of the lobes emphasizes the main directions, especially in the case of narrow lobes, where showing a clear direction is meaningful. The main direction is more difficult to read in pie glyphs. However, since pie glyphs are mirrored to show the directions from which the vectors point, reading the direction from the round head to the pointy tail of the pie alleviates this problem. The lobes are, nevertheless, less effective at conveying the variability, since the lobe shape, compared to the pie shape, makes narrow modes look somewhat narrower. Wider modes do not suffer this drawback, since lobes converge to pies as their widths increase.

To emphasize the directions in (from) which the ensemble members point (if such directions exist), small anchor balls are drawn at the tail (head) of each lobe (pie). Furthermore, the directional trends are emphasized by means of a color gradient, the brightness decreasing as the distance from the tail increases. According to the Gestalt laws of pattern perception (Ware (2004)), the brightness contrast makes the head more distinct than the tail and enhances thus the direction information. Examples of both lobular and pie glyphs are shown in Figure 5.6.

Finally, to reveal in the detail view not only the main directional trends, but also potential outliers, the unit vectors of all ensemble members can be shown as gray sticks, either on top of or under the glyphs. Selected members, if any, are represented using black sticks and are always drawn on top (an example is shown in Figure 5.16). Uniform distributions are shown by a full disk (see Figure 5.6 (e)). The width is given by twice the sample circular standard deviation (see Equation 5.1). For the sake of clarity, lobes have a minimum opening angle and length. In the case of overlapping lobes, those with a smaller width are drawn on top. Moreover, to ease the matching process, glyphs have the same color as the corresponding locations in the main spatial view.

Aggregated Glyphs for 2D Directional Data

Displaying glyphs over selected regions in a detail view of fixed dimensions is tailored to allow both a concurrent visualization of many glyphs and still convey the glyphs properly. This, however, shows the main directional trends over limited parts of the domain only. To reveal the individual distributions and provide a global understanding of the ensemble behavior, glyphs should be displayed in a dense representation at regularly spaced points over the whole domain. The representation should nevertheless be still sparse enough to avoid overlapping glyphs. Given the circular structure of the lobular glyphs, each glyph will never occupy an area larger than that of a circle of a specified radius. Nonetheless, glyphs still need a certain spatial extent in order to be easily discernible, and, depending on the selected radius, there may be a number of grid points whose glyphs are not shown. To alleviate this problem, the similarity measure introduced in Section 5.2.3 can be used in a LOD approach to find the most representative glyph for the covered region, if such a glyph exists.

More specifically, for a region of $n \times n$ samples, if the mixture models at these locations have all the same number of components, a representative glyph is determined as follows: first, the $n^2(n^2-1)/2$ similarity coefficients between every pair of mixture models are computed and arranged in a $n^2 \times n^2$ symmetric matrix with zeros on the main diagonal; then, for each one of the n^2 grid points, the median of the n^2-1 similarity coefficients to the other points is determined; finally, the glyph selected to be displayed is that representing the mixture model at the grid point with the smallest median.

When the mixture models do not have the same modality over the entire $n \times n$ region, a GMM is fitted to the directional data coming from all grid points and the corresponding glyph is shown instead. While this does not allow distinguishing between the individual distributions and, moreover, individual multimodal distributions may be aggregated by a unimodal distribution with a large spread, it does reveal the preferred directions, if any, over the whole $n \times n$ region.

To highlight whether the shown glyphs stem from regions of homogeneous or heterogeneous modality, the latter are drawn without a color gradient and a contour. The representativeness of a glyph is encoded also in its color, cobalt blue for glyphs over regions of homogeneous modality and light grayish blue for glyphs over regions of heterogeneous modality. Moreover, for regions of homogeneous modality, the scalar field in the background color codes in grayscale

the similarity coefficients between the PDF at each grid point and the PDF of the representative glyph, drawing attention to locations where the shown glyphs are less representative for the region they cover. Examples are shown in Figure 5.17 (top and middle).

A better understanding of the less well-represented regions can be obtained by allowing more glyphs to be displayed in the same region. Decreasing the radius, however, increases the amount of information the user has to cope with and also makes the glyphs harder to perceive. To make the visualization output less cluttered, the representative glyphs can be further aggregated to reduce the number of shown glyphs.

To this purpose, at locations with representative glyphs, a region growing algorithm is applied where mixture models have the same number of components. To allow a clear distinction between regions of different modalities, these regions are emphasized by their background color: pale green for unimodal distributions, pale pink for bimodal, pale purple for multimodal, and pale gray for uniform. Moreover, regions are displayed with rounded corners (unless the corners connect two parts of a region); this takes the Gestalt principle of continuity (Ware (2004)) into consideration, which states that smooth visual elements are more likely to be perceived as entities than those containing abrupt changes in direction.

Once the regions of homogeneous modality have been determined, the similarity coefficients are used within these regions to search for square subregions where PDFs have similar shapes. In such subregions, glyphs can be replaced with the most representative glyphs for the corresponding subregions; the maximum length allowed for a glyph is half the side of its subregion. Figure 5.17 presents an example of the LOD approach before (middle) and after these modifications (bottom).

5.3.3 Querying-driven Visualization

As explained in Section 5.2.3, users can run queries to verify how members behave at selected locations compared to other grid points. A glyph can be selected by clicking in the spatial views. Depending on the similarity criteria specified by the user, once a glyph has been selected, all dissimilar locations (and their corresponding glyphs) fade out and only similar locations and glyphs maintain their colors. For instance, when looking for the same modality and/or a similar PDF shape (up to a predefined threshold), all dissimilar locations in the main spatial view and the glyphs in the detail view fade out to the same extent. When searching

for locations where ensemble members vary in the same manner, each location (and glyph) exhibits a degree of fading proportional to the extent to which members disagree. Since the color scheme goes over yellow, in order to distinguish between similar yellow-colored glyphs and non-similar gray-colored glyphs, the contours of dissimilar glyphs become light gray, while the contours of similar glyphs maintain their dark gray color. To highlight similar glyphs at locations where members display comparable behavior, the contours of these glyphs become black, whereas the contour of the selected glyph is drawn in a contrasting color. For an example of a query-driven visualization see Figure 5.18.

5.3.4 3D Extension Outlook

Extending 2D visualization methods to 3D is rarely straightforward, because clutter and occlusion are considerably more problematic than in 2D. This section addresses a few of the challenges existent in 3D and suggests potential directions that could be explored in the future.

For visualizing directional trends, 3D glyphs have been applied before to represent uncertainty. Such glyphs have been used, for instance, in the context of HARDI data, to convey PDFs for fiber orientation (Jiao et al. (2012); Schultz et al. (2013)).

The 2D glyphs introduced in Section 5.3.2 are readily extended to 3D, e.g., by drawing the lobes as surfaces of revolution and mapping the parameters describing the PDFs to the lobes as before: for every mixture component, there is a corresponding lobe, and the mean direction, angle of the confidence cone, and mixture weight are mapped to the orientation, opening angle and length of each lobe (see Figure 5.9). Uniform distributions are drawn as spheres. To reduce clutter, glyphs are drawn without anchor balls.

The 3D glyphs convey thus the main characteristics of the PDFs. Dense representations of 3D glyphs, however, lead to massive clutter and occlusion, even if they were based on a hierarchical approach. Therefore, additional visualization methods should be employed to help reveal the variations in the PDFs across the domain. For instance, the 3D domain can be partitioned in connected regions of similar behavior. The partitioning can be done iteratively, segmenting the domain first into regions of the same modality and, then, segmenting regions of the same modality into subregions of similar PDF shape.

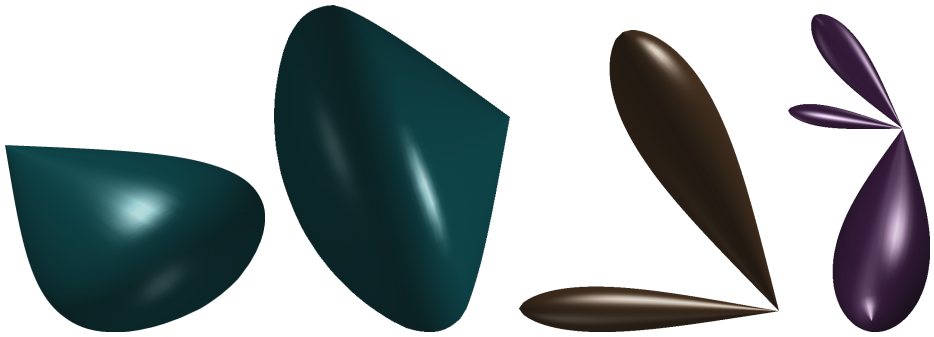


Figure 5.9: Examples of 3D lobular glyphs for PDFs with various modalities and angular variabilities.

A synthetic ensemble of a flow past an ellipsoidal obstacle is used here for illustration purposes. It was created by slightly increasing each time the viscosity input parameter in a 3D Navier-Stokes simulation. The ensemble comprises 56 members and is given on a $145 \times 49 \times 49$ Cartesian grid. The first few ensemble members exhibit laminar flow around the obstacle, but, as the input parameter increases, the flow separates and a recirculation region comprising two symmetric vortices of increasing size forms behind the obstacle.

Figure 5.10 shows a modality partitioning for the flow ensemble. The modality segmentation of the domain partitions the domain into 80 regions, but just a few of these regions have more than one grid point. In fact, two of these regions contain most grid points, especially the region with the isosurface colored in pink in Figure 5.10, which connects locations where unimodal distributions were fitted to the data. The U-shaped isosurface (shown in blue in the same figure) behind the ellipsoidal obstacle (colored in green) connects locations where bimodal distributions were fitted to the ensemble data. It is also where recirculation regions of increasing sizes appear in the ensemble members.

The modality segmentation offers a first partitioning of the domain. Nevertheless, the number of surfaces that can be displayed concurrently is also limited. To gain more understanding into how the domain is partitioned, the domain segmentation can be summarized into an abstract representation of the partitions. Figure 5.11 shows two manners of encoding these abstract representations.

In the top figure, the individual regions are shown in a layered graph layout, obtained using Graphviz (Gansner and North (2000)), an open source graph visualization software. Regions are shown as circular nodes, colored according to their modality (light shades of pink for

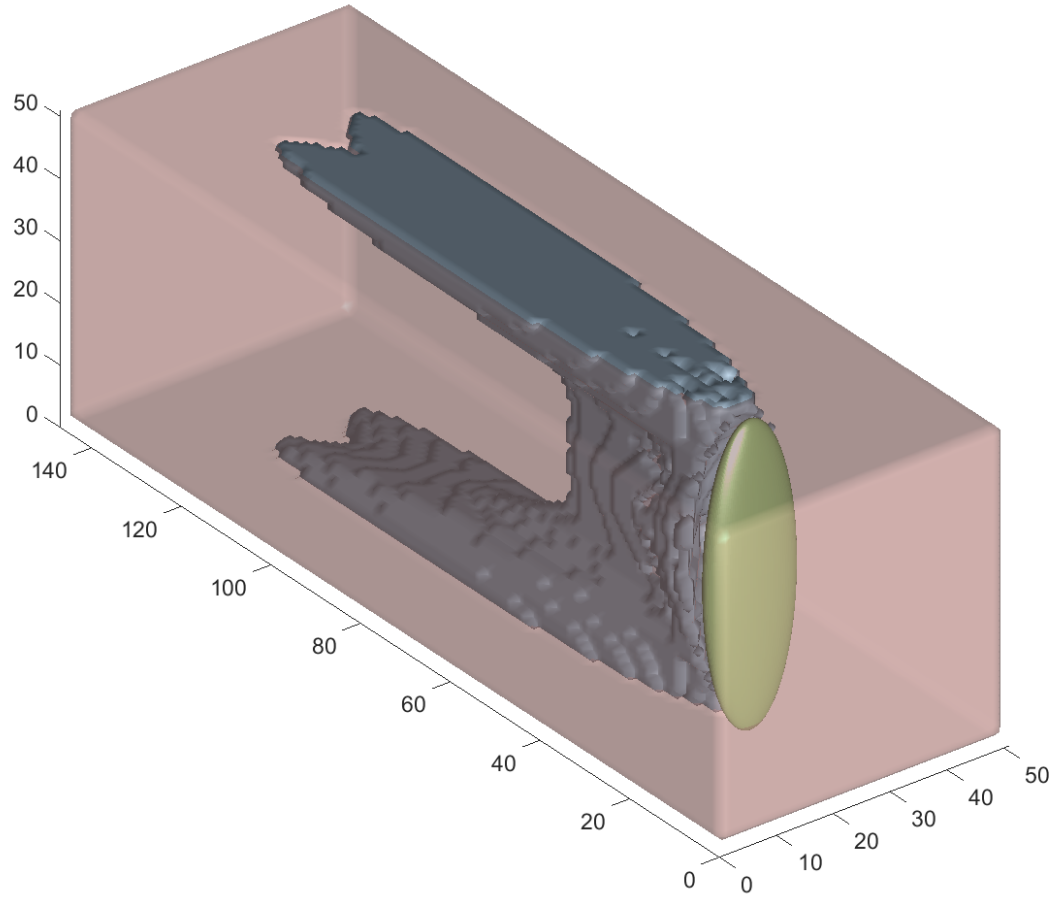


Figure 5.10: Domain segmentation in regions of the same modality for a synthetic flow ensemble. The ellipsoidal obstacle is shown in green, along with the two largest isosurfaces that contain most domain points: the isosurface shown in pink connects locations with unimodal PDFs and is drawn partly transparent, as it covers most of the domain, except primarily in the U-shaped region behind the obstacle (the green ellipsoid), where the blue-colored isosurface connects bimodal regions.

unimodal, blue for bimodal, green for trimodal, and violet for quadromodal) and with radii suggestive of the number of grid points in the surface. Nodes of neighboring regions are connected using edges, drawn using two colors, each half being drawn in the color of the corresponding node. The two aforementioned surfaces are shown as the two largest pink and blue nodes and can be observed to be connected with most of the other regions. The graph layout was designed so as to minimize crossings, but consider the spatial distances between

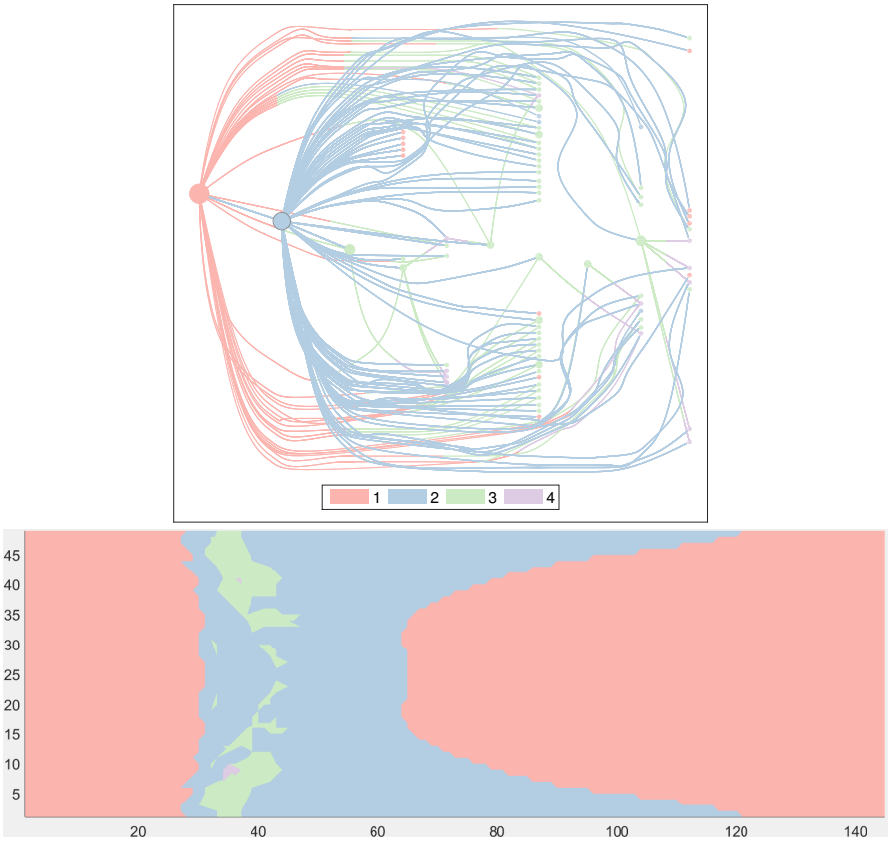


Figure 5.11: Abstract representations of the partitions from the modality-based clustering for the synthetic flow ensemble: (top) regions are shown as nodes (colored according to modality) in a graph layout, with neighboring regions connected by edges; (bottom) regions are projected on a 2D plane dependent on the viewpoint specified by the user.

regions when placing nodes. Spatial coherence is nevertheless lost.

To provide more insight into how regions are situated in the spatial domain, a projected view is shown in the bottom figure. The regions are projected from 3D to 2D depending on the viewpoint selected by the user. Even though this results in some regions being distorted or occluded, changing the viewport helps distinguish between regions.

An interactive GUI was designed to guide the user in the exploration of regions of different behavior. It allows users to select regions in the abstract views and visualize the corresponding isosurfaces with customizable color and opacity values. Regions of the same modality can be

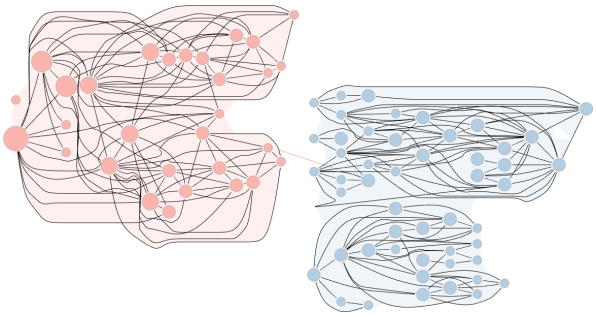


Figure 5.12: Graph layout for the synthetic flow ensemble, displaying subregions showing similar PDFs within the two main unimodal and bimodal regions.

further segmented based on the similarity of the PDF shape. The partitioning algorithm attempts to grow regions where the mixture components have mean directions with angular deviations within the smaller confidence cone, as well as similar angular spreads and mixture weights (below prescribed thresholds). The subregions can be explored in a similar fashion as the modality regions, using abstract views (see the graph in Figure 5.12 for subregions of the two largest modality regions).

Users can then explore both region and subregion isosurfaces in the domain view. However, even if this helps segment the domain in regions of similar behavior, the isosurfaces alone do not offer an intuitive visualization of vector information. More insight into the directional trends can be displayed by placing either single glyphs at user-selected locations or a specified number of glyphs on user-selected isosurfaces. The latter is automatic and places the number of glyphs indicated by the user in such a manner that the distance between glyphs is maximized. If glyphs are placed on isosurfaces representing regions rather than subregions, the algorithm places glyphs sequentially in subregions, in decreasing order of their number of locations. Glyphs are drawn with a color contrasting to the color of the isosurface they are placed on and can have their maximum lengths adjusted by the user.

An example is given in Figure 5.13. The blue-colored isosurface corresponds to the second largest region connecting locations with bimodal PDFs, whereas the pink-colored isosurface corresponds to a subregion of the largest region connecting locations with unimodal PDFs. Even if this region covers a significant part of the spatial domain, when the shape of the PDF is considered, it is subdivided into several subregions. Most of these subregions display low spreads (the flow is quite laminar in front of the obstacle and further away from the recirculation areas), but different directions (for instance, the flow separates because of the

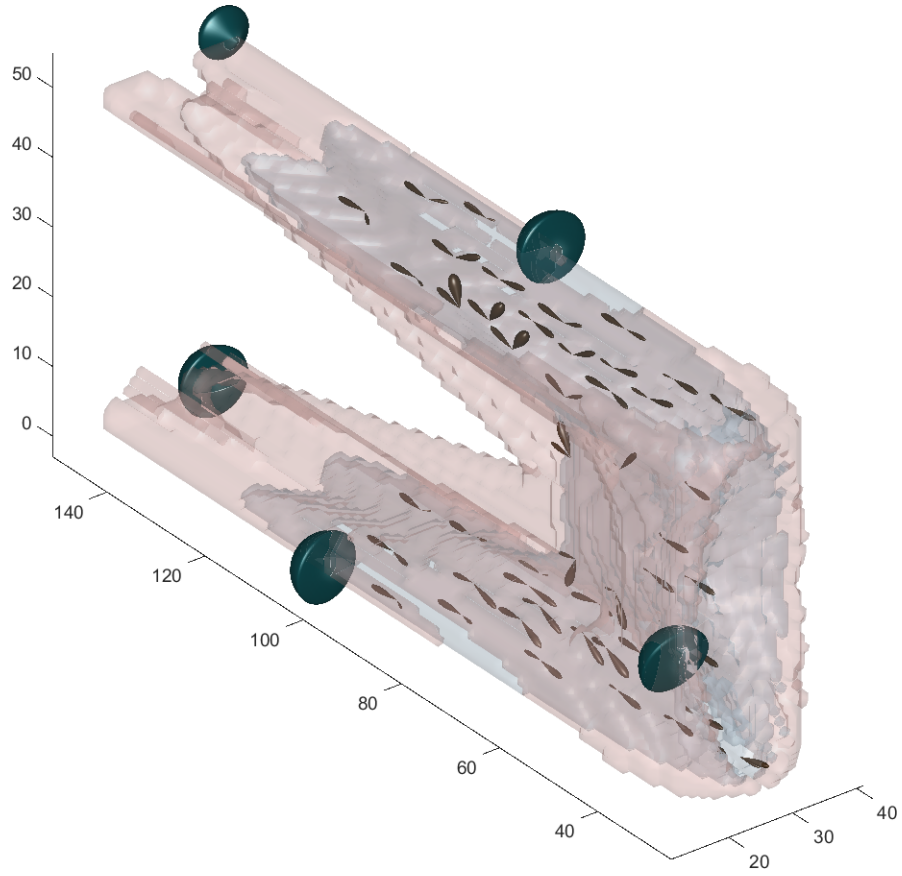


Figure 5.13: Domain view of the synthetic flow ensemble, with user-selected isosurfaces and automatically placed glyphs showing directional trends and their associated variability.

obstacle). Some regions, however, like the pink subregion enclosing the bimodal region in the figure, exhibit large spreads (see the five glyphs that were automatically placed on the surface of the subregion).

50 glyphs were automatically placed on the surface representing the bimodal region. Several of them show clear bimodality, with almost equally long lobes of approximately opposite orientations. The lobes oriented towards the obstacle model those ensemble members where the flow recirculates, as opposed to those having smaller recirculation regions or showing laminar flow. Other glyphs show weaker bimodality or lobes that are quite close to one another, especially near the boundary of the bimodal region.

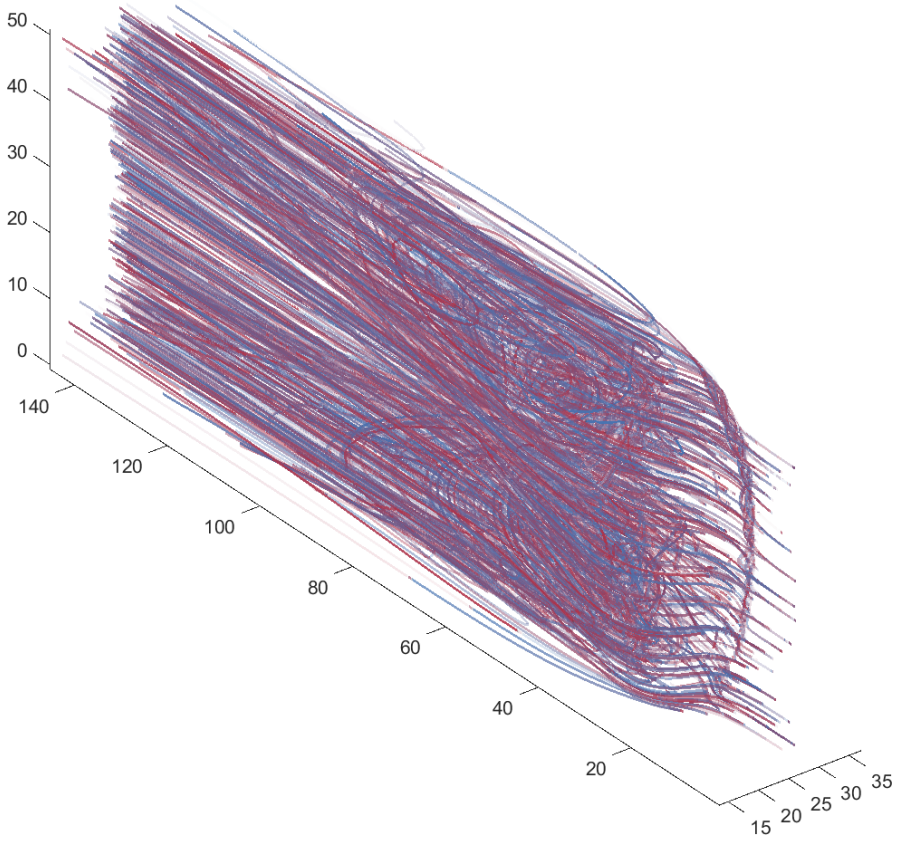


Figure 5.14: Visualizing flow trends for the synthetic flow ensemble using representative streamlines at selected locations.

While the glyphs do reveal the directional trends at selected locations, understanding the main trends in the flow behavior is not always straightforward, since the glyphs are not connected and it is not trivial to understand which lobes should and could be connected.

Understanding the flow behavior could be facilitated, for instance, by displaying streamlines. However, seeding strategies to minimize clutter and occlusion are challenging even for crisp vector fields, let alone for a whole ensemble. Potential solutions could include showing a limited number of representative streamlines at selected locations. For instance, in Figure 5.14, representative streamlines were seeded at those domain locations that are local maxima of the particle divergence. A recent streamline clustering algorithm (Ferstl et al. (2016a)) was applied to group similar streamlines together, and determine cluster representatives.

The color of the streamlines varies from red to blue, encoding their time evolution. The

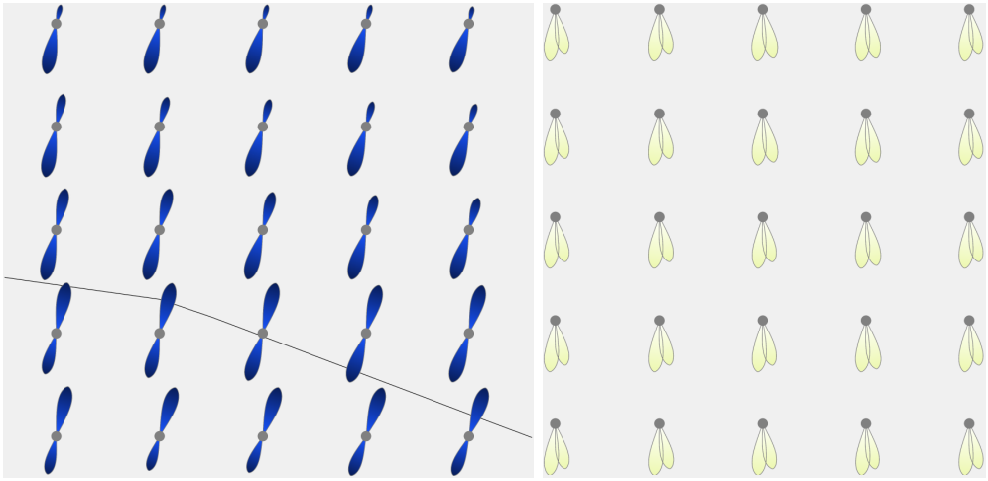


Figure 5.15: Lobular glyphs for directional distributions of vector-to-closest-point ensemble at grid points with (left) high and (right) low centrality values. Median shape contours are likely to be extracted at locations with high centrality values, like in the left figure.

displayed streamlines are suggestive of the flow separation around the ellipsoidal obstacle, the two symmetric regions of flow recirculation, and the rejoining flow. However, the horizontal streamlines are dominating, giving the impression of laminar flow in regions where local fitted mixtures are clearly bimodal (streamlines seeded at such locations are part of laminar flow in several ensemble members and of recirculating regions in others; the impression of laminar flow is stronger when both types of streamlines are almost horizontal and, depending also on the viewing angle and often in spite of color coding the time evolution, create the impression of a single streamline).

It would be interesting to couple the local and global analysis to gain more insight into the ensemble directional and transport trends and variability. Possible directions include using the local analysis to develop seeding strategies, combining streamlines with lobular glyphs, etc. Nevertheless, there are types of vector ensembles where transport is not of interest and a local approach using mixture models (and corresponding glyphs) suffices. This is, for instance, the case for vector-to-closest-point ensembles (VCPE).

For geometric objects represented by polygonal surfaces embedded in 3D Cartesian spaces, the vector-to-closest-point at a given grid location is the vector to the closest surface point (for more on VCP representations, see, for instance, Demir and Westermann (2015)). In this

context (Demir et al. (2016a)), mixture models have been fitted to directional distributions, and used to analyze the central tendency of ensembles of 2D and 3D shapes, by classifying domain points depending on how central their positions with respect to the surrounding closest points are and applying this to determine median shapes. A point is likely to be most central, if, among other requirements, the directions of the closest point vectors can be modeled by two modes with means of approximately the same direction, but contrary orientation. An example is given in Figure 5.15, which shows a detail view from such a 2D ensemble; the left figure displays lobular glyphs at points with high centrality values, especially in the immediate vicinity of the median contour, whereas the right figure shows locations with low centrality values, where a median contour is unlikely to be extracted.

5.4 Further Results

The proposed techniques are tested on three ECMWF ensembles: two wind forecasts, both given on a 101×41 Cartesian grid, and comprising 50 members and a control run, and an ensemble of temperature gradient fields derived from an ECMWF forecast, given on a 221×101 Cartesian grid and comprising 50 members.

5.4.1 Similarity Analysis

The first data set is the 120 hour wind direction forecast at a pressure level of 500 hPa, valid on October 19, 2012. The geolocated modality scalar field (containing the number of mixture components at every grid point) is shown in the upper spatial view (A) in Figure 5.5. The isocontours of the geopotential height field of the control run are superimposed in black.

A hard clustering of the ensemble members, based on an empirical orthogonal function analysis of the 500 hPa geopotential field, was available from ECMWF. It partitioned the members into three main groups of cardinality 21, 20, and 10. Because the intra- and inter-cluster similarities were not available, a detailed analysis of this clustering solution was not possible. However, since at this pressure level winds blow approximately parallel to the contours of the geopotential heights, it was interesting to verify whether the clustering solution obtained using the proposed techniques was consistent with the ECMWF clustering. In fact, examining clustering solutions derived using more than one method or measure is highly important

in meteorology, as there is no single "best" clustering method and the different techniques influence the resulting partitions. Thus, only consistency across techniques serves to increase confidence in the validity of the results (Alhamed et al. (2002)).

The ECMWF clusters were assessed using internal criteria (silhouette coefficients) to verify the degree to which the ensemble members in these clusters were both closely related and well-separated from the members in the other clusters. The silhouette-based analysis revealed that the ensemble members grouped in the third ECMWF partition formed a coherent cluster (with silhouette coefficients around $0.4 - 0.6$) and were well-separated from the majority of the other members. This can also be observed in the two abstract views in Figure 5.5, where nine of the members are placed in the last main group: in the dendrogram, the last main branch is merged at the highest level, while in the global dissimilarity matrix, the bulk of the dissimilarity values to other ensemble members varies in the upper end of the colormap. The ensemble members in the other two ECMWF clusters were not well-separated according to the silhouette analysis. Nevertheless, several of the nested subclusters in the dendrogram comprise primarily subgroups of these ECMWF clusters.

The two abstract views summarizing the relationships between the ensemble members highlight groups of similar members – joined at low levels in the dendrogram and with low pairwise global dissimilarity values colored in shades of blue (primarily on the diagonal of the global dissimilarity matrix). Outlier members can also be rapidly determined, since such members are typically merged at higher levels in the dendrogram. For instance, member 11 is the last ensemble member to be merged in the first (orange-colored) subcluster in the dendrogram. Moreover, the primarily moderate to high global pairwise dissimilarity values are shown using mostly colors from the upper range of the colormap.

Various measures for the variability of the ensemble members over the domain are encoded in the main spatial view. For instance, Figure 5.5 (A) shows the modality scalar field for this ensemble. Interesting regions can then be selected to explore the associated PDFs in a detail view (see Figure 5.5 (B)). Regions of higher variability, with multimodal PDFs and/or higher angular spreads, occur primarily around the pressure centers (shown using concentric closed contours) and at their boundaries.

Figure 5.16 presents a zoom into a region south of Greenland at the boundary between an area of high pressure (showing anticyclonic or clockwise flow) and a low pressure center (showing

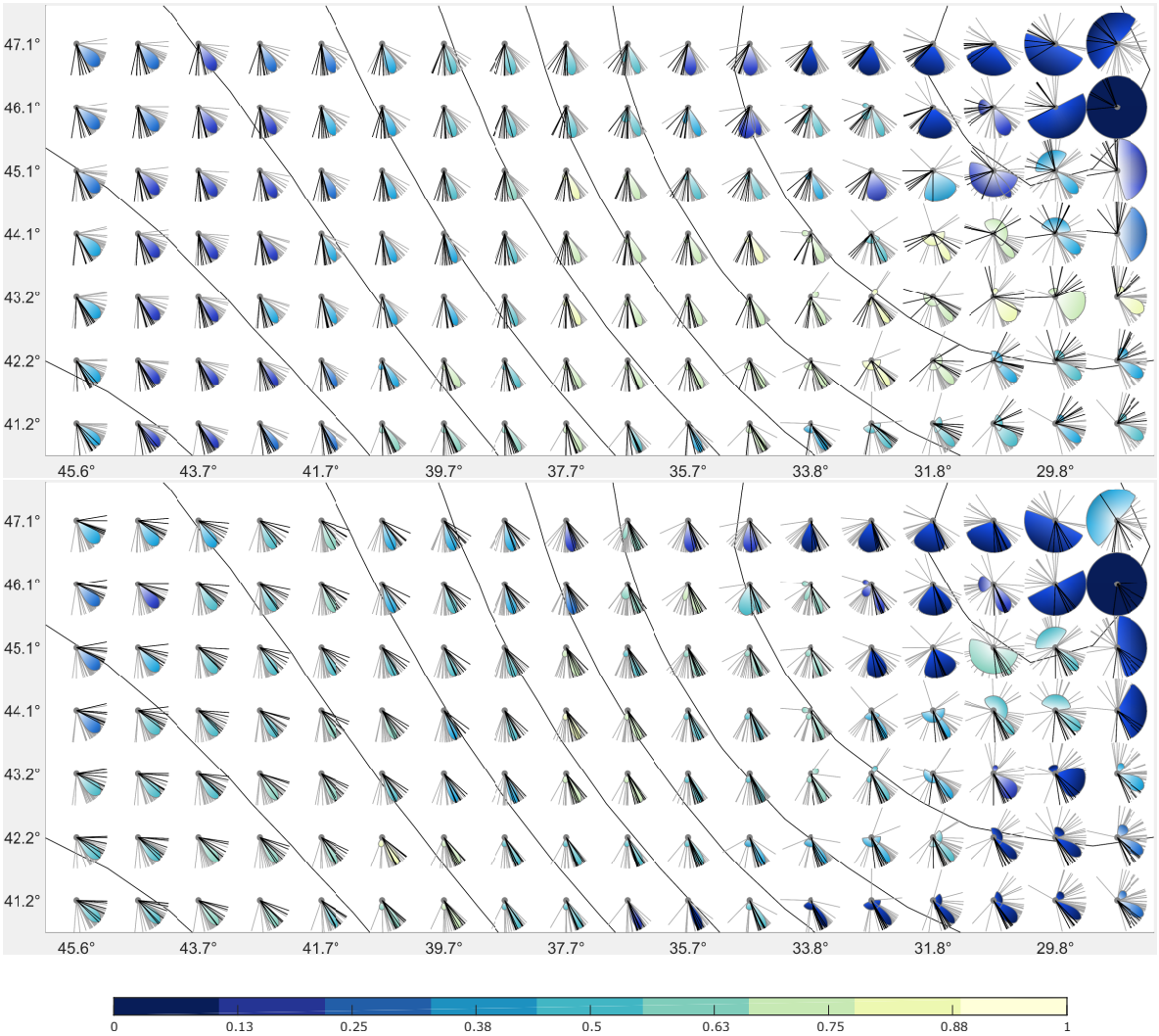


Figure 5.16: ECMWF wind ensemble at a pressure level of 500 hPa: Zoom-in over uncertain region for two subclusters from the ECMWF cluster 1 (top) and 3 (bottom). Color varies from blue to yellow as the member dissimilarity increases.

cyclonic or counterclockwise flow) for two subclusters from the first and third ECWMF clusters. These subclusters were classified as dissimilar according to the locally scaled angular similarity measure (relative to the local variability).

Alternatively, the main directional trends over the entire domain can be displayed at various levels of detail using aggregated glyphs. Figure 5.17 (top) shows an example of glyphs placed at regular intervals, each glyph covering a region of 4×4 samples. The normalized scalar field

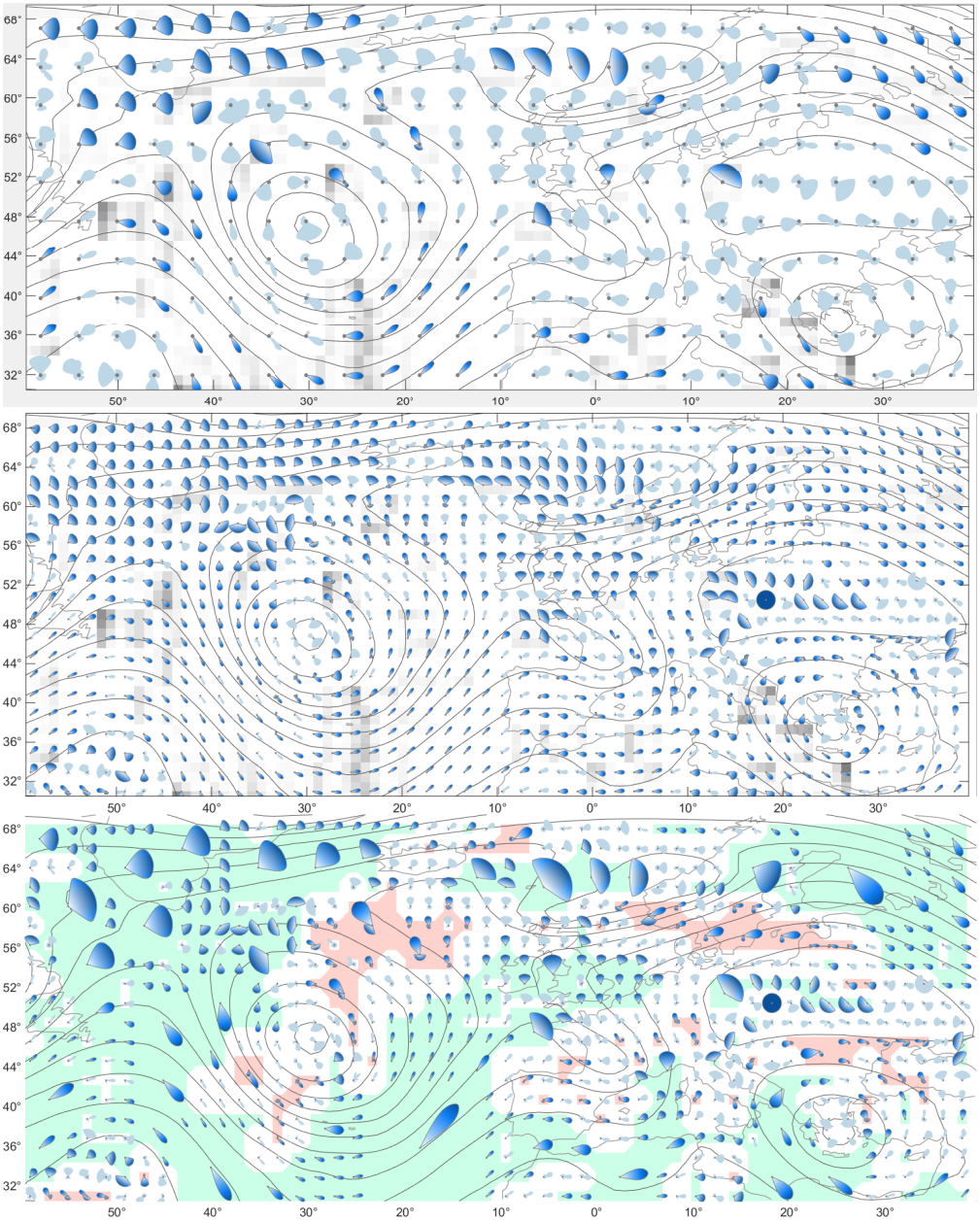


Figure 5.17: ECMWF wind ensemble at a pressure level of 500 hPa with aggregated representative (dark blue) and unrepresentative (light blue) glyphs at various levels of detail: each glyph covers a region of (top) 4×4 samples and (middle and bottom) 2×2 samples; additional glyph aggregations were performed for the representative glyphs in the bottom plot in order to reduce clutter, and regions connecting representative glyphs with the same number of lobes were highlighted (green for unimodal and pink for bimodal).

in the background encodes how dissimilar the PDFs at the hidden locations are, compared to the PDFs of the shown glyphs.

A denser representation (see Figure 5.17 (middle) for glyphs aggregated over 2×2 regions) allows observing the directional trends and variability at more locations, but the visual display is more cluttered and some of the glyphs are harder to distinguish.

These issues are alleviated (see Figure 5.17 (bottom)) by highlighting connected regions where PDFs have the same number of modes and replacing similar glyphs occupying square regions by larger representative glyphs. Clutter is therefore reduced, numerous glyphs are more clearly discernible, and distinguishing between regions of (dis)similar behavior is facilitated. For instance, most of the fitted mixture models in the elongated region in the bottom left corner of Figure 5.17 (bottom) are unimodal PDFs with low angular variability, while at the pressure center to the right, the majority are multimodal and have higher spreads.

Further insight into the ensemble behavior can be gained by running queries to identify, for a selected glyph, locations where PDFs have the same modality, a similar shape (under a prescribed threshold τ), or a similar variation of the ensemble members. An example is shown in Figure 5.18, where the selected bimodal glyph has been emphasized by means of a red outline in the top figure.

The top figure shows a detail view in the neighborhood of the selected glyph, revealing locations where members behave similarly. The glyphs representing PDFs with different modalities are most faded out and contoured with a light gray. As to what the shape of the PDF is concerned, how high the angular deviation needs to be for two locations to be assessed as dissimilar depends on the existent variation: a mode with a higher variation and a higher deviation is assessed as more similar than a mode with a lower variation and a lower deviation, if the latter's deviation is higher relative to the existent variation.

In the top figure, the other glyphs corresponding to bimodal PDFs have their main modes pointing in directions similar to that of the selected glyph; thus, the shape similarity (for a threshold $\tau = 1$) does not exclude any of these glyphs. The final color of these glyphs depends on whether the ensemble members at these locations vary in the same manner. The locally scaled angular similarity measure is used in the top figure; therefore, the percentage of members behaving similarly at other locations is around 70% only in the immediate neighborhood of the selected point (the glyphs at these locations are visually enhanced in the top figure). No other locations exceed 30%.

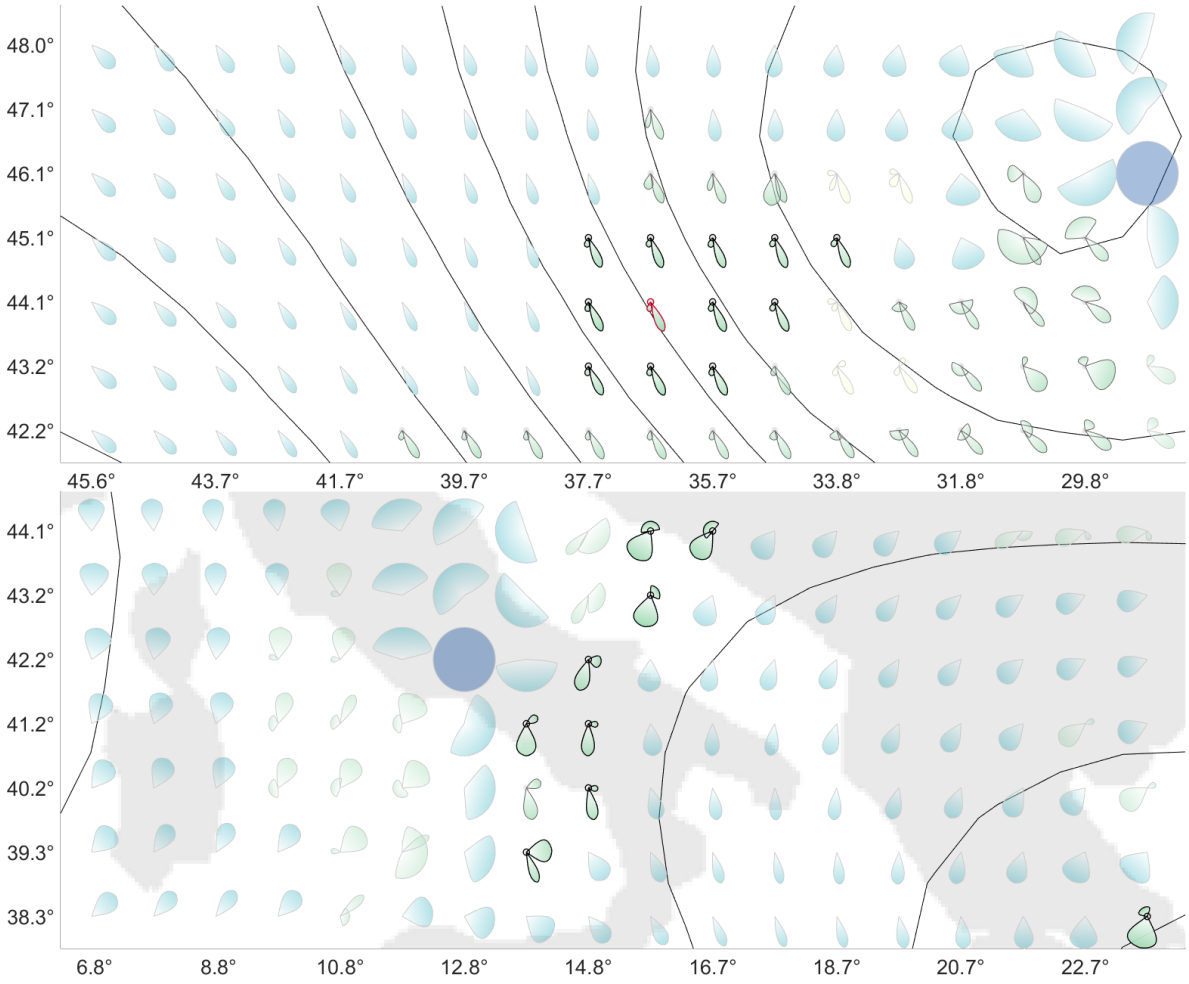


Figure 5.18: ECMWF wind ensemble at a pressure level of 500 hPa: querying for other locations where PDFs have the same modality and a similar shape, as well as similar member behavior, as the PDF at the selected location (corresponding glyph outlined in red); the locally scaled similarity measure (top) identifies only neighboring locations as similar, whereas the scaled measure (bottom) is less restrictive and also indicates remote locations as similar.

If a less restrictive similarity measure is used instead, such as the scaled angular measure, where members are similar if they may belong to the same mode and their angular deviation is within the maximum limit allowed, several remote locations reach values of even 80%, while the neighboring locations are over 90%. Figure 5.18 (bottom) shows an example of remote regions exhibiting similar locations, but also illustrates other glyphs representing bimodal PDFs that have been classified as dissimilar due to their shape dissimilarity.

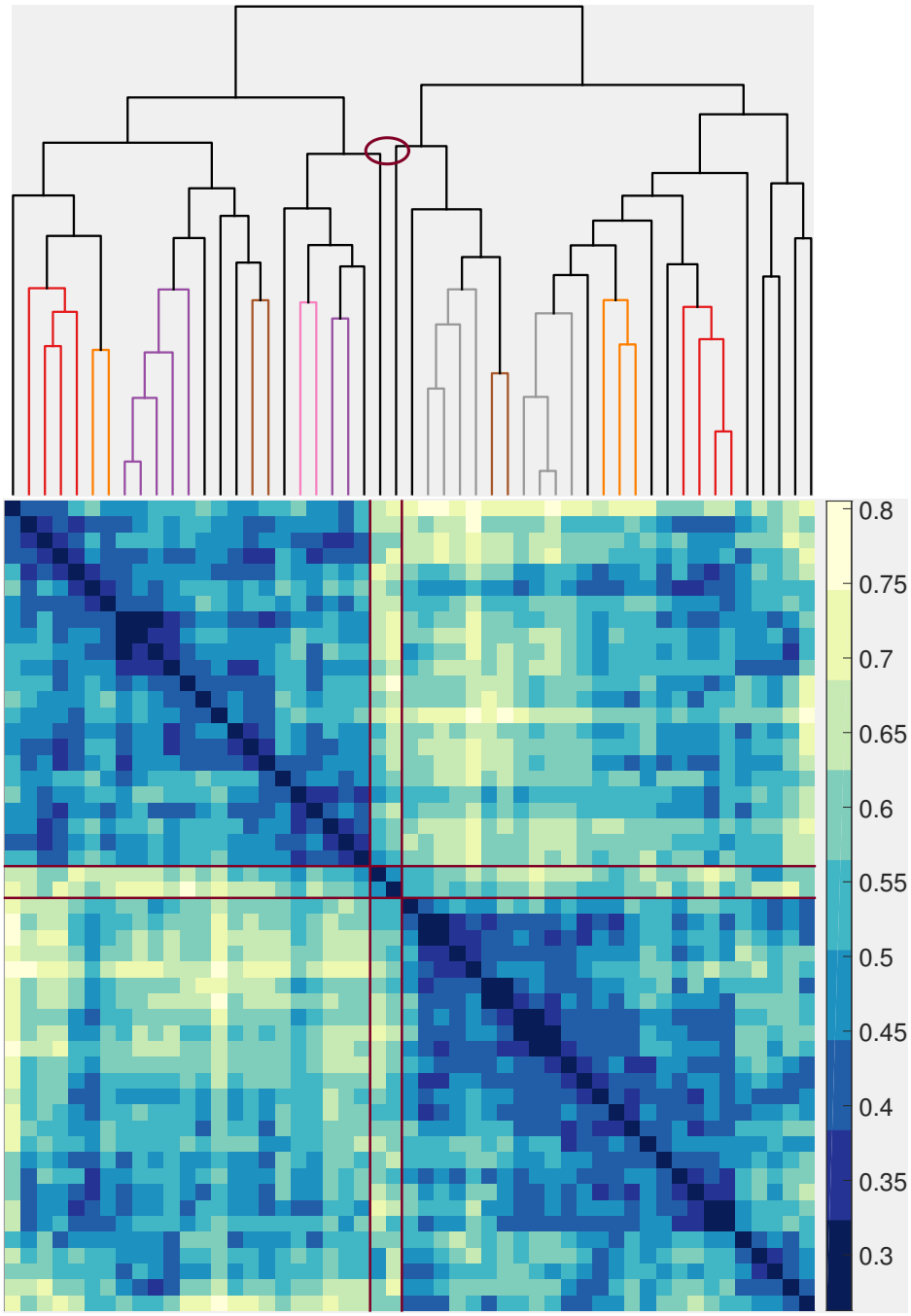


Figure 5.19: ECMWF wind ensemble at a pressure level of 850 hPa: abstract representations (dendrogram view (top) and global dissimilarity matrix (bottom)) of the clustering solution reveal better separated groups than at the pressure level of 500 hPa. The color coding of the pairwise global dissimilarity values varies from blue (low) to yellow (high).

The second ECMWF ensemble is the 120 hour wind direction forecast at a pressure level of 850 hPa. Closer to the ground, the wind direction is influenced by the surface (due to friction). Therefore, regions displaying high variability occupy a much larger area than before, emerging at pressure centers, boundaries between centers, and over irregular terrain.

The clustering solution shows that ensemble members are now better separated than at the 500 hPa level, (see Figure 5.19 for the clustering solution corresponding to the locally scaled angular similarity measure). Only a few ensemble members display similarities to members in other clusters. Outlier members can be identified as those ensemble members merging at high levels in the dendrogram, such as members 16 and 39, which are the last members to be merged. What the dendrogram does not emphasize, however, is that the two members are rather similar to one another. This aspect is, nevertheless, revealed in the global dissimilarity matrix (where their dissimilarity values to all members are outlined in brown), which also confirms their otherwise outlier statuses.

Users can select groups of members to reveal locations of (dis)similar behavior. For instance, Figure 5.20 (top) shows the scalar dissimilarity field for a group of three members that were classified as similar according to the locally scaled angular similarity measure. Their colormap has four sequential discrete colors, corresponding to the number of dissimilar pairs out of the three possible (besides the 0 value occurring when all three members are similar).

Including a fourth member (see Figure 5.20 (top)), dissimilar to the previous three according to the global dissimilarity matrix, reveals that the latter disagrees with the previous three at most locations that were previously shown as similar (colored dark-blue in Figure 5.20 (top), and in a lighter shade of blue in Figure 5.20 (bottom) – showing that three out of the six pairs possible are dissimilar).

The third data set is the temperature gradient ensemble, at a pressure level of 200 hPa, valid on January 12, 2011. The geolocated modality field is shown in Figure 5.21 (top), along with the isotherms of the mean temperature field. It can be noticed that uniform distributions (modality zero), indicative of no preferred direction in the greatest rate of change in temperature, predominate around the centers of air masses of the mean field. Multimodal distributions tend to occur across the boundaries of warm and cold air masses. A detail view of such a region located at the boundaries is illustrated in Figure 5.21 (bottom). Its location within the spatial domain is marked with a brown rectangle in the top figure.

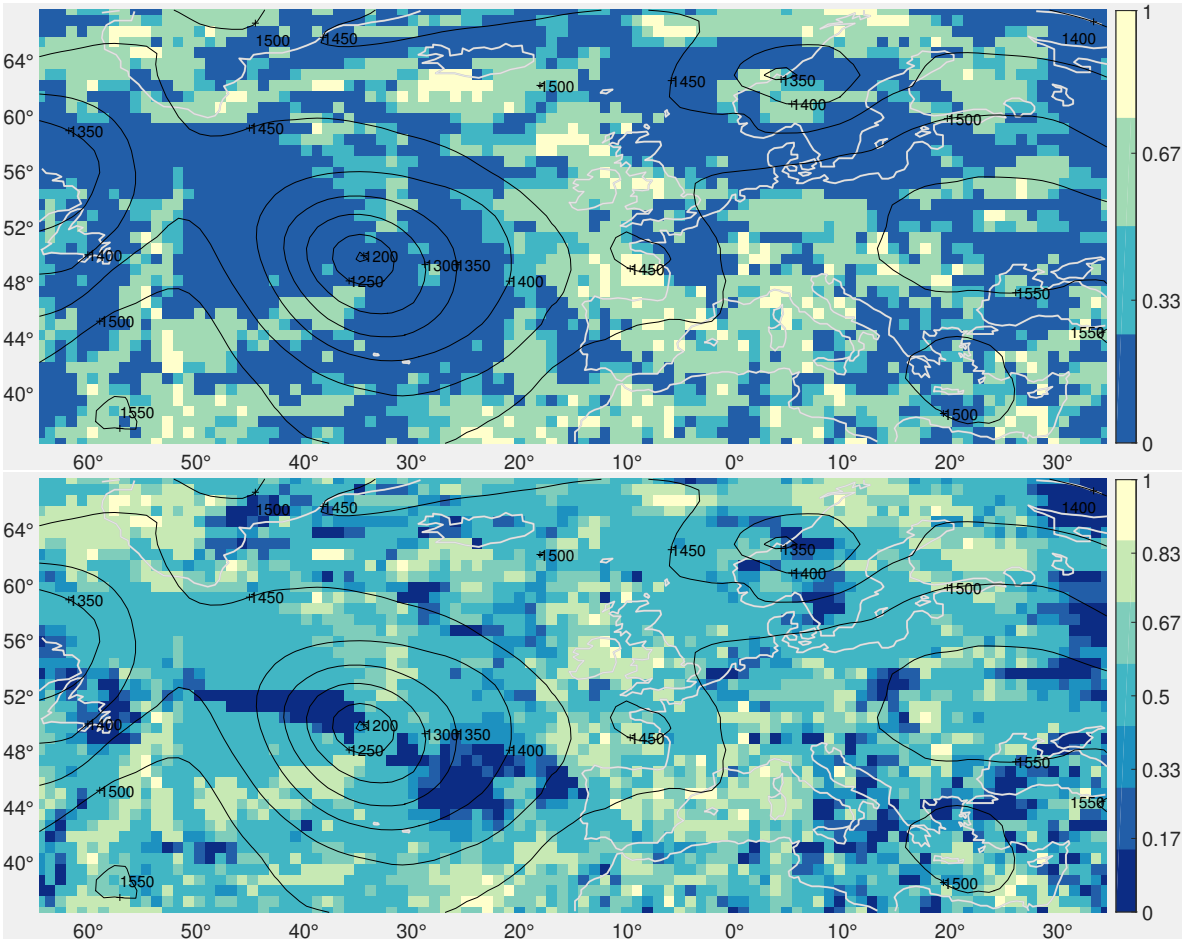


Figure 5.20: ECMWF wind ensemble at a pressure level of 850 hPa: (top) spatial view highlighting dissimilar locations for a group of three similar ensemble members; (bottom) updated dissimilarity field after a fourth member, classified as dissimilar to the first three, is added.

5.4.2 Implementation, Performance and Scalability

The tests performed on the aforementioned ensemble data sets were run on a standard desktop PC, equipped with an Intel i7-4790 quad-core processor running at 3.6GHz and with 12GB RAM. All computational operations were performed in Matlab in a preprocessing step, using available Matlab functionality to fit GMMs on the real line and for the normality tests. For each of the two wind ensembles, performing all computations (including the ten repetitions) took roughly 12 minutes, whereas the larger temperature gradient ensemble required about

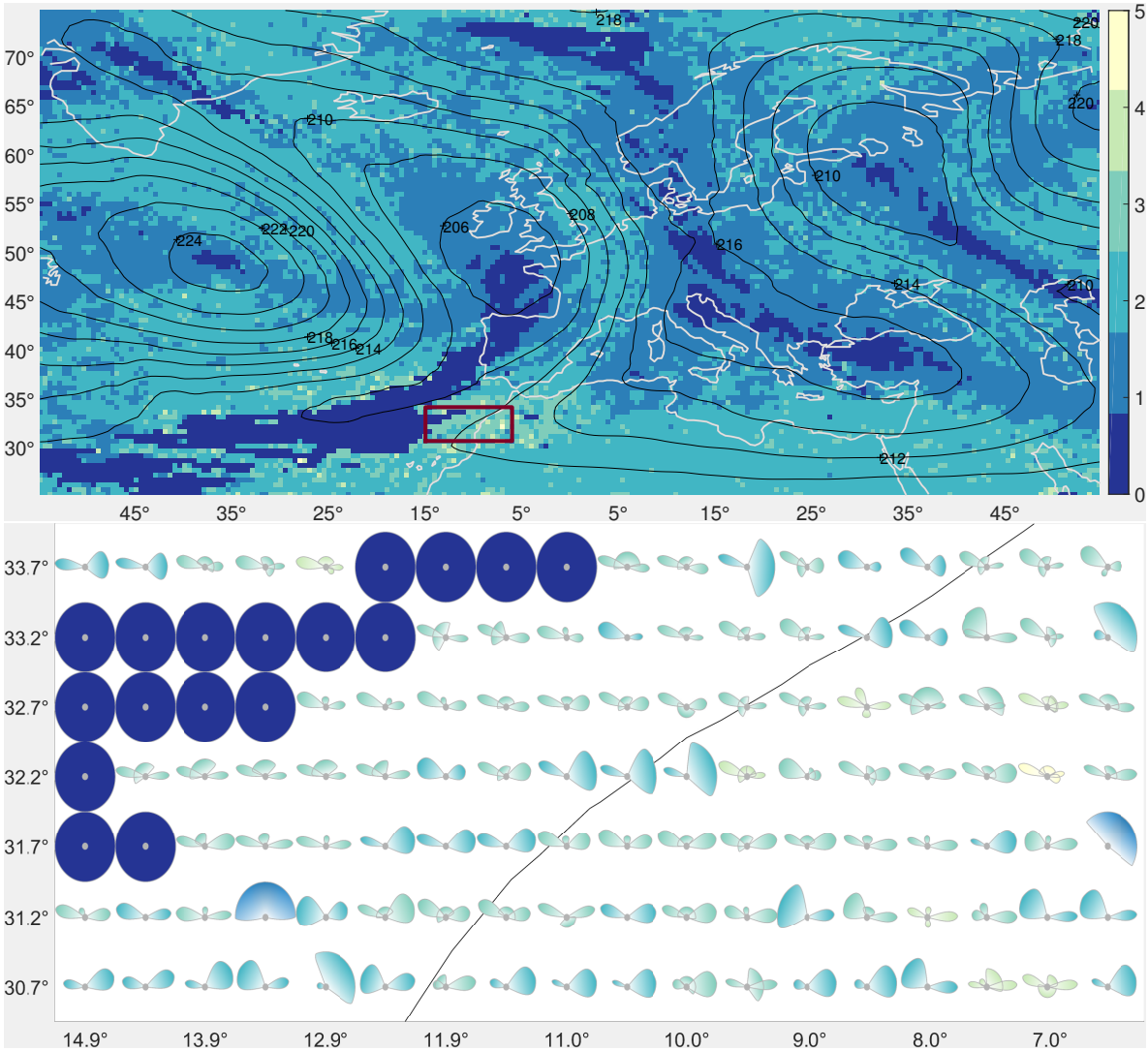


Figure 5.21: ECMWF temperature gradient ensemble: (top) spatial view for the 221×101 modality field and (bottom) a 18×7 detail view.

80 minutes. The most time consuming operation was the fitting, which made up most of the processing time, followed by the similarity computations and the global assembly. Other operations, such as obtaining the clustering solution or the glyph representations, last no longer than a few seconds for all data sets. Due to the compact representation of the fitted mixture models, the memory requirement of all precomputed quantities is moderate. Moreover, the analysis session with multiple linked views can be handled at interactive frame rates.

Regarding scalability, ensemble data of larger sizes and dimensions can still be analyzed with the proposed techniques and GUI. Thus, increasing the grid dimensions reflects only on the overview spatial view, where the size of the pixel for each grid point decreases (see Figure 5.21 (top)). The detail view showing the directional distributions is not affected, because the number of glyphs shown was adjusted to fit the displayed view and is constant for all grid dimensions (see Figure 5.21 (bottom)). For the LOD approach, the displayed glyphs can aggregate square regions of increasing size. However, as finer levels are selected to reveal more detail, depending on the grid dimensions, there is a threshold below which the visual outputs become too cluttered and glyphs can no longer be distinguished clearly. The two abstract views are not influenced in any way by the size of the grid. A significant increase in the number of ensemble members, however, may affect the readability of the abstract views (but not that of the domain views). Nevertheless, the typical number of ensemble members in such meteorological ensembles as those used in this work is approximately 50, so legibility is not an issue. Ensemble data sets of considerably higher sizes would nevertheless require focus and context techniques, or zooming and panning.

5.5 Conclusion and Evaluation

The techniques presented in this chapter are meant to allow an interactive comparative exploration of ensembles of vector fields. To this purpose, mixture models were fitted at every grid point in the domain, to approximate directional distributions using concise set of parameters that are more readily analyzed and visualized. EM algorithms have random initial guesses and are thus prone to converge to local optima. Running these algorithms several times with different starting values helped nevertheless to significantly reduce the percentage of non-repetitive structures. Based on the mixture models, local and global similarity measures were derived, both to reveal locations where two or more ensemble members disagree and to identify similar groups of members.

Informal feedback on the presented techniques from domain experts from meteorology was positive. The experts highly appreciated the interaction facilities that allowed investigating directional distributions in a detail view. They also liked the color coding of the modalities, which helped them identify locations where multimodal distributions were fitted to the data. However, they were initially hesitant about the global dissimilarity matrix, as they were not familiar with such a representation and thus required some effort to understand it. On the

other hand, they found it interesting to quickly grasp which ensemble members disagree with the majority of the other members (as the corresponding columns/rows were mainly yellow). They liked the possibility to brush ensemble members in the matrix and dendrogram, and distinguish between locations of (dis)similar behavior. Moreover, they appreciated the option to blend in the individual ensemble members, which allowed them to see outliers in addition to directional trends. They considered that lobular glyphs convey the directional distributions properly, but preferred the pie glyphs, which resembled other graphical representations used in meteorology.

While this first feedback was positive, the full potential of the proposed approaches has to be evaluated in a more detailed analysis in the future. To gain an initial insight into how the lobular glyphs compare to the pie glyphs, a user study with 65 inexperienced participants (54 males and 11 females, aged 18 to 30 years) was conducted. The participants were given six main tasks (three for the lobular glyphs and another three for the pie glyphs) in a limited amount of time: for each primitive in every figure, they had to estimate either its direction, its spread, or the ratio between its length and the maximum length in the respective figure. Participants were also asked to assess the difficulty of the tasks on a five-level Likert scale. The questions were identical for both lobular and pie glyphs, but in a different order. Furthermore, the subtasks of each main task were split on two different pages and delivered in two separate batches, so that the answers of the participants to the two types of visualizations could be recorded after they had got to interact with both. Upon a preliminary examination of the results, it turned out that participants were equally capable of solving the tasks with either method, the difference in accuracy (measured via mean and standard deviation) being insignificant. In the perceived difficulty, however, the majority of the participants (81%) found the lobular glyphs more intuitive when estimating directions, but preferred the pie glyphs for the other two types of tasks (87% for the spread and 81% for the ratios).

Future work includes gaining a deeper understanding into the ability of the users to assess the different characteristics of directional data using the two proposed methods. Moreover, the 3D extension possibilities which were outlined in Section 5.3.4 should be further explored and refined.

6

Comparative Visual Analysis of Transport Variability in Flow Ensembles

In the previous chapter, the variability of ensembles of vector fields was analyzed based on a mixture modeling of the directional data at every location. However, when the ensemble describes flow phenomena, as is the case for wind ensembles, the initial uncertainty is transported along the flow. Analyzing the temporal variability of an ensemble is important, for instance, to determine when and where divergences occur, and, implicitly, the locations where and the time intervals over which a simulation is more or less reliable.

Performing such an analysis is, however, not straightforward, due to several reasons. First of all, transport divergences among ensemble members occur gradually over time, due to spatial variations of the transport paths and the transport velocity along these. Analyzing the flow variability requires thus new concepts to identify these divergences and, in particular, when the divergences causing a spread in transport started occurring. Moreover, visualizing the temporal variability of an ensemble – locally, at a selected domain point, or globally, to compare transport over the ensemble domain – is challenging, since it requires new graphical abstractions to show the complex spatio-temporal information of the ensemble evolution in an intuitive way.

In this chapter, a novel visual analysis approach is proposed to address these challenges and explore the temporal variability of ensembles of both stationary and time-dependent 2D vector fields. To this purpose, transport deviations are analyzed over time statistically for ensembles of vector fields given on 2D Cartesian grid structures. This is done first by approximating

distributions of tracer particles at every grid point and each time step with mixtures of Gaussian components; then, by applying the Mahalanobis distance on the Gaussian modes to assess the pairwise member (dis)similarities relative to the variability allowed by each mode. These pairwise values can be used in a number of ways to quantify the ensemble divergence over time, revealing the time steps and locations of dissimilar behavior.

To allow an effective local and global visual analysis, the variability information derived from the pairwise dissimilarities is processed and visualized using a number of novel approaches. Like for the local approach described in the previous chapter, the proposed techniques are combined into an interactive visual analysis tool (see Figure 6.1), to give insight into various aspects of the flow variability:

- **Divergence Visualization:** The various derived measures for the transport variability are color coded over the entire domain (Figure 6.1 (A) – background). Moreover, spaghetti plots of ensemble trajectories are enhanced at selected locations (Figure 6.1 (A) – foreground), by additionally encoding the temporal evolution of the transport and its variability, to emphasize when and where ensemble members behave (dis)similarly; this type of encoding is shown to be especially effective in cases when member trajectories exhibit similar geometric shapes, but particle positions considered at the same time instant are dispersed.

- **Small-Multiples:** To compare the flow variability across selected subdomains, a novel graphical abstraction is introduced, based on miniaturized versions of enhanced spaghetti plots in a small-multiples layout (Figure 6.1 (B)). The downscaling method, based on a selection of prominent time steps, is designed to preserve the main transport trends.

- **Similarity visualization:** The pairwise dissimilarity values are also used for a hierarchical clustering of ensemble members with similar flow across the domain. The clustering solution of the ensemble members over the entire forecast interval is shown using dendrograms (Figure 6.1 (C)), whereas the dynamical evolution of the clusters at selected time steps is visualized using parallel sets (Figure 6.1 (D)), conveying split and merge events over time.

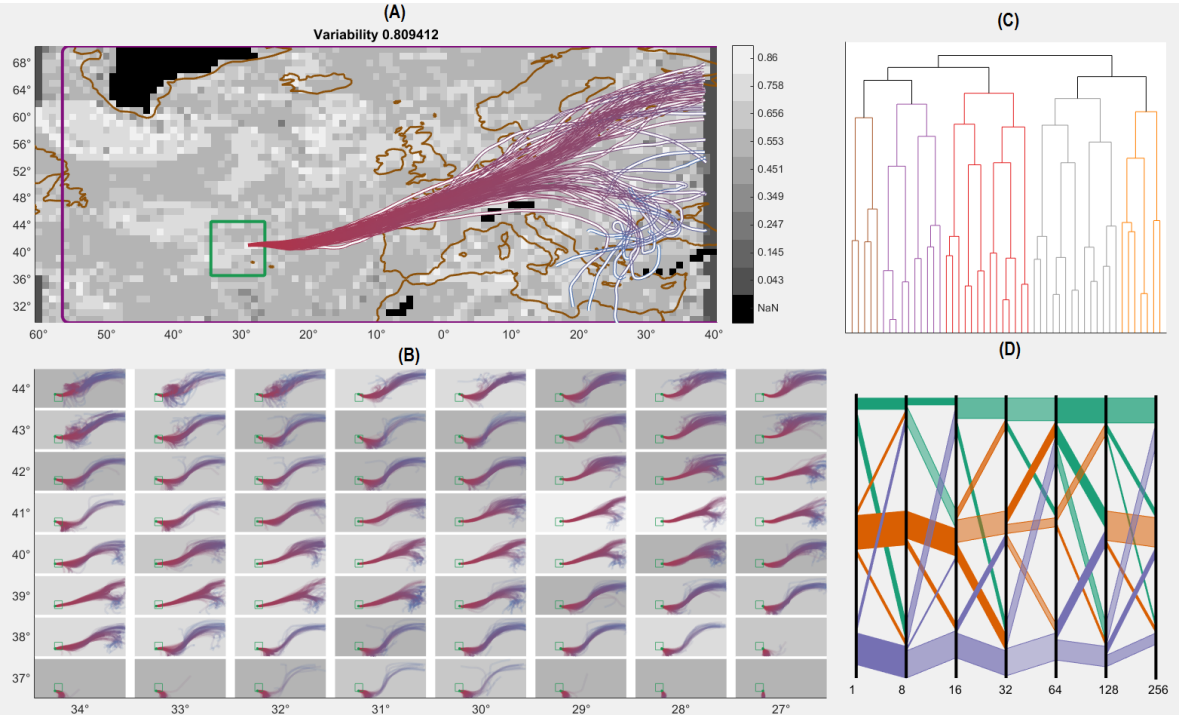


Figure 6.1: Multiple linked views for an ECMWF time-varying ensemble. The geolocated spatial view (A) shows the aggregated flow variability over the whole domain, along with an enhanced spaghetti plot at a user-selected location. For the selected region marked by a green square in (A), the detail view (B) displays downscaled enhanced spaghetti plots. View (C) shows the dendrogram of the hierarchical clustering of the ensemble members with perturbed initial conditions over the entire forecast interval, whereas view (D) reveals the variability of the clustering solution at selected time steps.

6.1 Transport Variability

The method starts with an ensemble of vector fields given on a 2D grid structure, and identifies the member pairwise (dis)similarities at every grid point and every time step. To determine when and where changes in the transport behavior happen, ensemble members should be assessed as (dis)similar based on their deviation at every time step. Existing methods include using synthetic thresholds above which deviations among members are regarded as "diverging" or the spread of the particle positions is considered large. Yet, due to the arbitrariness of the threshold, the uncertainty would be characterized as high, e.g., only when the spread would

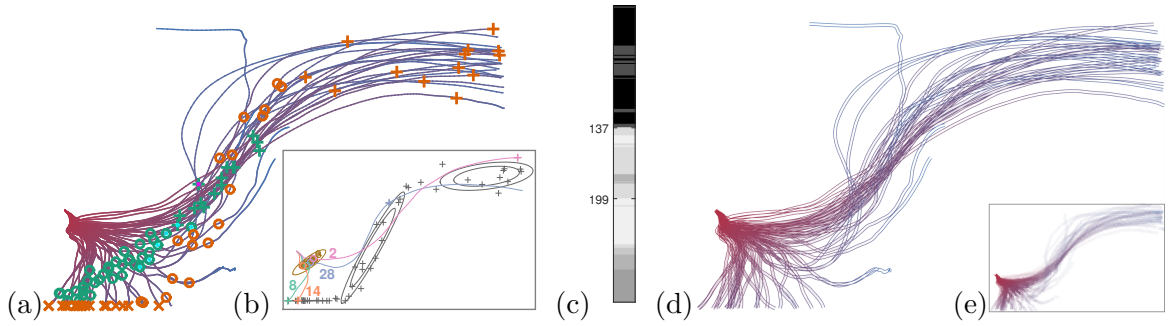


Figure 6.2: Pathline variability: (a) trajectories of all 51 members; particle positions are shown at steps 137 and 199 with green and orange markers, respectively, and different symbols for different clusters; (b) four singled-out members: 8 (green), 14 (orange), 28 (blue), and 2 (pink), with colored markers for all members and ellipses for the Gaussian modes at steps 58 (orange-colored) and 199 (colored in dark gray); (c) grayscale encoding of the time step divergence count; as the divergence counts increase, the stripes (d) fade to white in the enhanced spaghetti plot and (e) become less opaque in the downscaled version.

be "large enough" and not when the divergences causing this spread started occurring.

When data can be assumed Gaussian distributed, the Mahalanobis distance (Mahalanobis (1936)) provides natural thresholds, since, due to its use of the covariance matrix, it is dependent neither on the different scales of the variables, nor affected by correlations between them. The Mahalanobis distance assesses ensemble members as (dis)similar based on their deviation relative to the data variability, instead of an artificial threshold. However, for ensemble data, the Gaussian assumption is often not valid. The complex distributions of 2D particle positions are thus approximated using mixtures of Gaussians. The Mahalanobis distance can then be applied on the individual Gaussian modes, to characterize pairs of members as (dis)similar relative to the variability of the corresponding components.

Figure 6.2 offers some initial insight into the rationale behind the fine-grained pairwise dissimilarity analysis using four selected members of an ensemble. The pathlines of all 51 ensemble members are shown in (a), the color (varying from red to blue) encoding their time evolution, while (b) shows only the trajectories of the four selected ensemble members (2, 8, 14, and 28). GMMs were fitted at all integration time steps and the Mahalanobis distance was applied on each mode to assess pairs of members as (dis)similar based on whether their deviation was within the allowed variability or not.

Thus, until step 137 (corresponding particle positions are shown with green markers in Figure 6.2 (a)), only unimodal distributions were fitted to the particle positions. According to the deviations relative to the local variability, member 8, initially an inlier, is, up to step 58, similar only to member 28 and not to members 2 and 14 as well (the Gaussian component fitted to the particle positions at step 58 is marked in (b) using brown concentric ellipses for the Gaussian mode, whereas the particle positions are shown using circles). This happens because, even if all members are modeled by the same Gaussian component, the latter are initially outliers.

Thereafter, however, member 8 becomes and stays similar to 14, both members following the downward trend. Moreover, as the deviation between members 8 and 28 increases, the two are assessed as dissimilar shortly after step 58; regarding members 8 and 2, they are dissimilar from the very beginning, even though the spread of the tracer particles is initially very low. In fact, member 2 is dissimilar for most of the time even to member 28, even though both pathlines are geometrically similar and follow the upward trend. This happens because particles travel at different speeds. For instance, at step 199, it can be observed that the two positions are modeled by different modes (the three Gaussian modes fitted at this time step are shown by concentric gray ellipses in (b) and all members are marked with pluses).

6.1.1 Pairwise Dissimilarity Analysis

The derivations in this chapter are given for ensembles of 2D vector fields given over Cartesian grid structures, but are readily extended to 3D. The pairwise dissimilarity analysis is performed at every grid point, on the particle positions at every integration time step of the N member trajectories[1].

Initially, the numbers of time steps m_i for each trajectory are not necessarily equal. To make sure all trajectories are of equal length $m = \max_{i=1,N}(m_i)$, the final particle positions of the shorter trajectories are repeated $m - m_i$ times. This step is necessary in order to avoid that ensemble members with similar trajectories, but of slightly different lengths $m_i > m_j$, such as members 8 and 14 in the previous example, are artificially dissimilar in the $m_i - m_j$ interval. The pairwise dissimilarity analysis for ensemble members with trajectories of considerably different length and behavior, such as members 8 and 28 in the same example, is not affected by the repetition; such members are dissimilar as soon as their deviations become large

[1]Member trajectories are obtained by numerical integration, e.g., using 4-th-order Runge-Kutta methods.

relative to the mode variability, or their positions are clearly modeled by different components. Moreover, the (dis)similarity of any two ensemble members (i, j) does not change any longer after step $\max(m_i, m_j)$.

The pairwise analysis comprises two stages: in the first stage, the distribution of the particle positions at each time step is modeled using mixture of Gaussians. GMMs are fitted using the algorithm described in Section 3.2, which identifies both the number of Gaussian modes and their shapes, by partitioning the particle positions iteratively until the positions in each partition can be assumed Gaussian distributed.

Thus, a sample of 2D observations $x_i = [x_{xi}, x_{yi}]$, $i = 1, N$, is modeled by a mixture of n Gaussian modes

$$f(x) = \sum_{j=1}^{n} \alpha_j \mathcal{N}(\mu_j, \Sigma_j), \quad \alpha_j > 0, \quad \sum_{j=1}^{n} \alpha_j = 1 \tag{6.1}$$

where each Gaussian mode $\mathcal{N}(\mu_j, \Sigma_j)$ has a weight α_j and is parameterized by its mean μ_j and covariance matrix Σ_j. Furthermore, the particle position of each ensemble member x_i, $i = 1, N$, has a certain posterior probability p_{ij} of having been drawn from Gaussian mode j, where $\sum_{j=1}^{n} p_{ij} = 1$.

Instances of Gaussian components fitted to particle positions at different time steps are shown in Figure 6.2 (b) for the previous example. Each Gaussian component is represented using concentric ellipses: at step 58, a single Gaussian component (shown using orange-colored concentric ellipses) is enough to model the particle positions, whereas at step 199, three Gaussian components (displayed using ellipses colored in dark gray) are necessary.

For the sake of simplicity, the partitioning of the ensemble members in this example was hard, with members being assigned to single Gaussian components according to their highest posterior probabilities. However, in the following, a *soft* clustering is used, so that a pair of ensemble members (x_i, x_j) is assessed as dissimilar only if their posterior probabilities p_{ik} and p_{jk} of having been drawn from the same component k are very low. Thus, for particle positions located on the boundary of two Gaussian components, members are considered similar to all other members whose posterior probabilities of having been modeled by any of the two modes are not very low.

An example of a bimodal distribution being fitted to the particle positions is shown in Figure 6.2 (a). The hard clustering solution of the particle positions at time step 137 is shown using green markers of two types (corresponding to the two modes): circles and pluses. The

few ensemble members that are located on the boundaries of the two modes (and thus assessed as similar to all members) have been marked with cyan dots in the middle of their markers.

Other pairs of ensemble members, like, for instance, those members located towards the opposite boundaries of the two modes and thus with low probabilities of having been drawn from the other mode, are assessed as dissimilar. In other words, these pairs of members are assessed at time step 137 as dissimilar, because they belong clearly to different modes. However, just because at previous time steps there was not enough proof for a departure from normality, this does not imply that members have started diverging at time step 137 only. Instead, even though all members were modeled by a single Gaussian component up to integration step 137, the deviation of some members may have been large relative to the local variability already earlier. Examples of such cases are, for instance, ensemble members 8 and 28, or 8 and 2 (see Figure 6.2 (b)).

Thus, in a second stage, for every pair of members (x_i, x_j) classified as similar in the first stage, the Mahalanobis distance refines this (dis)similarity classification by considering the deviation of the members relative to the allowed variability (summarized by the covariance matrix Σ_k)

$$MD(x_i, x_j) = \sqrt{(x_i - x_j)\Sigma_k^{-1}(x_i - x_j)^{\mathrm{T}}} \tag{6.2}$$

Ensemble members are assessed as dissimilar if $MD(x_i, x_j) > 2.3$, which corresponds statistically to less likely deviations falling outside one confidence region (68%) for a bivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$. For ensemble members located on the boundary of two Gaussian components, the Mahalanobis distance is computed relative to the variability allowed by both modes and the more restrictive value, i.e., the higher value of the two, is taken.

For the previous example, from the six members situated on the boundary of the two modes fitted to the data, only one member (hard clustered in the upper mode) still exhibits similarities to members in the other mode. This happens because the Mahalanobis distance takes into account the shapes of the covariance matrices, which, in both cases, allow more variation in the vertical direction than in the horizontal one. This property of the Mahalanobis distance reflects on other ensemble members as well: for instance, the ensemble member in the upper mode, marked in Figure 6.2 (a) by a magenta dot in the middle, is dissimilar to all other ensemble members.

6.1.2 Transport Variability over the Ensemble Domain

The fine-grained analysis based on mixtures of Gaussian components and the Mahalanobis distance identifies the pairwise (dis)similarities of the ensemble members at every grid point and time step. Further quantitative divergence measures can be derived from this analysis. For instance, in order to distinguish between inliers and outliers within modes or among all ensemble members, the *member divergence count* (mdc_i^t) of ensemble member i at time step t is defined as the number of members dissimilar to it at that instant, normalized by the total number of members.

While the member divergence counts are almost always positive (except in the rare case where member trajectories are identical), inlier members have lower divergence counts than outliers. For instance, when particles can be assumed Gaussian distributed, the divergence counts of the members increase with the distance of the particles from the mean. For multimodal distributions, divergence counts are expected to increase for both mode inliers and outliers, but the individual increases depend on the mixture weights, as well as on the shapes of the modes and their separation.

To aggregate the member divergence counts and identify time steps and locations with higher transport variability, the *time step divergence count (*tdc^{t_k}*)* is obtained by summing up and normalizing all member divergence counts at time step t_k

$$\mathrm{tdc}^{t_k} = \frac{\sum_{i=1}^{N} \mathrm{mdc}_i^{t_k}}{N} \tag{6.3}$$

The aggregated time step divergence count over a given time interval $[t_1, t_2]$ is

$$\mathrm{tdc}^{t_1,t_2} = \mathrm{median}(\mathrm{tdc}^{t_k})_{t_k \in [t_1,t_2]} \tag{6.4}$$

The median is applied in the previous equation because it is a robust measure of the central tendency of the data. Nevertheless, other summary statistics, like the mean, could be used instead.

Furthermore, since divergence counts are always positive, reference values need to be defined for the divergence counts in order to distinguish between low and high values. Thus, for the member divergence count, the reference value is the median of the minimum member

divergence count aggregated over every grid point (i, j)

$$\mathrm{mdc}_{\mathrm{ref}} = \mathrm{median}(\min(\mathrm{mdc}_{i,j_k^t})_{k=1,N,\, t=1,t_{\max}}) \tag{6.5}$$

while, for the time step divergence count, the reference value is the median of the time step divergence counts at the first time step

$$\mathrm{tdc}_{\mathrm{ref}} = \mathrm{median}(\mathrm{tdc}_{i,j}^1) \tag{6.6}$$

The time step divergence counts for the previous example (taking values in the $[0.55, 0.82]$ range) are encoded in grayscale in Figure 6.2 (c). As long as there is not enough proof for a departure from normality, the time step divergence count values vary in the $[0.55, 0.61]$ range, depending on how the ensemble members deviate from one another relative to the variability allowed by the fitted Gaussian component. When a bimodal distribution is fitted to the data for the first time at step 137, a surge to 0.76 occurs in the divergence count, since the pairwise deviations are now assessed relative to two smaller covariance matrices instead of one larger covariance matrix. Nevertheless, as can be seen in Figure 6.2 (c), the tendency of the particles to separate into two bundles started becoming apparent earlier, since lighter gray shades are visible in the colorbar already before step 137.

The surge at step 137 is succeeded by small gradual decreases, as more particles follow the upward trend. However, as the particles going upwards move at significantly different speeds, this leads gradually to a second surge (0.81) at step 199, when three Gaussian modes are fitted to the particle positions (see the orange markers in Figure 6.2 (a)). Notice that, in spite of the geometrical similarity of the trajectories going upwards, the individual particle positions are modeled using two well-separated Gaussian modes of different orientations. As particles eventually approach their final positions, the time step divergence count values decrease again, although to values no lower than 0.69. This happens because most particles end up clustered quite tightly together, either in the bottom or the upper bundle.

The derived divergence counts can be used in a number of ways, for instance, to determine a selection of time steps that preserves the transport behavior. This is useful to downscale trajectories without obscuring the transport trends and variability, as a straightforward downscaling would (Figure 6.3 (top)).

Thus, instead of keeping all trajectory points, only the integration steps with changes in

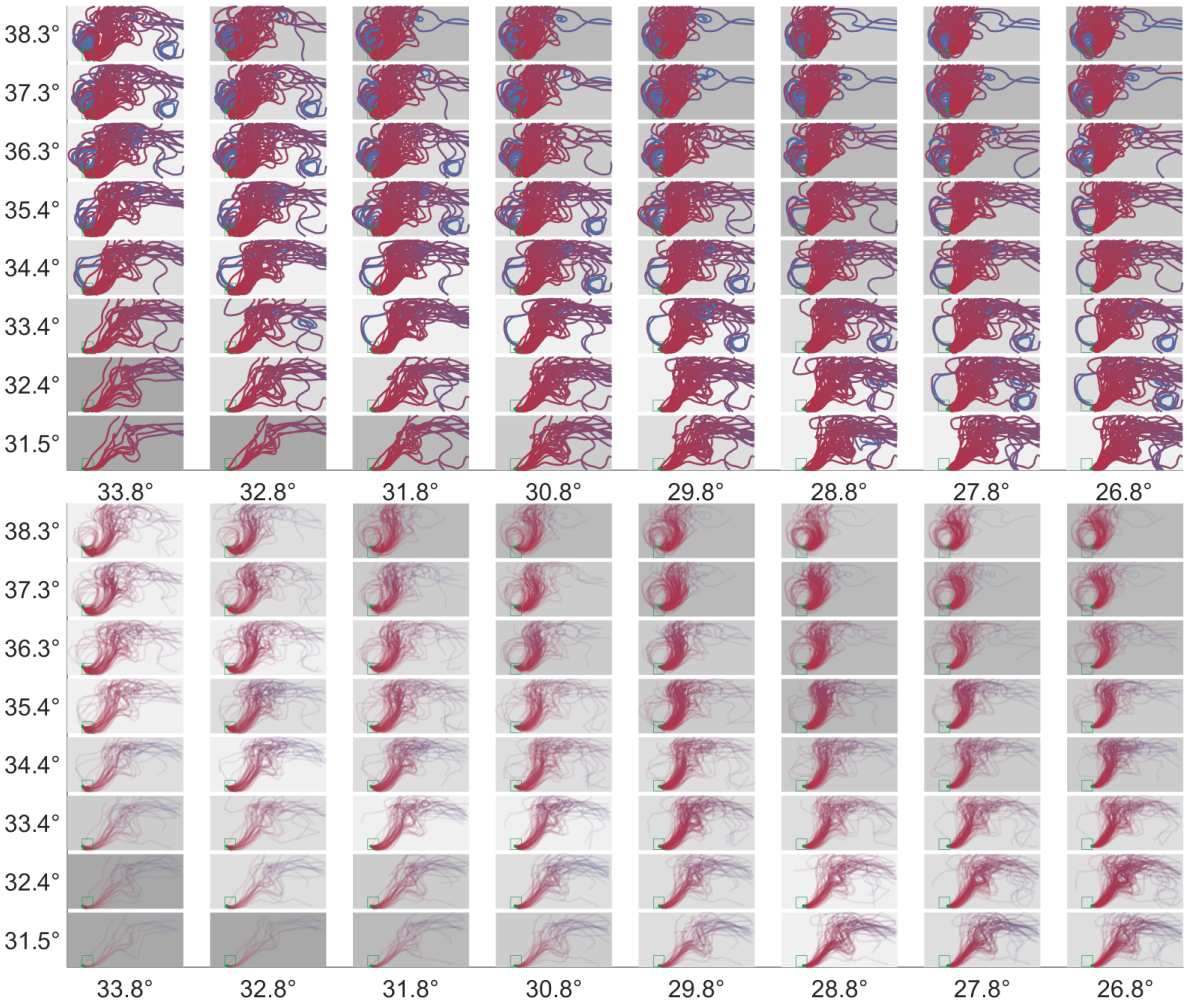


Figure 6.3: Small-multiples displays for ensembles are significantly more cluttered when downscaled in a (top) straightforward manner than (bottom) with sampling and variability encoding in opacity.

(dis)similarities (the salient time steps) are preserved. Other non-salient integration steps are sampled regularly, the sampling rate being proportional to the number of non-salient steps. Although the information loss may lead to a coarser curve approximation, especially in intervals with high curvature, the flow structure is qualitatively well-preserved (see Figure 6.3 (bottom)). For instance, for the previous example, even though around 60% of the trajectory points are discarded, the flow structure and variability are still retained (see Figure 6.2 (e)).

The time step divergence count is nevertheless just an aggregated value of the member divergence counts, meant to highlight important events, such as changes of regimes, and does not reveal any details on how the ensemble members behave. Even the member divergence counts themselves are just a measure of how (dis)similarly ensemble members behave compared to other members at the same or other locations, and not an exact description.

Such a description is contained in the pairwise dissimilarity matrices at every grid point, which are nonetheless too large ($N \times Nt$, where t is the total number of time steps) to be analyzed and visualized concurrently. Nevertheless, the dissimilarity matrices can be compressed (by storing, for instance, just the dissimilarity values from the salient time steps). The dissimilarity matrices can also be recomputed on the fly from the GMMs fitted to the data at every time step, which are also compact representations of the ensemble behavior.

A significant amount of variability information can be derived from the pairwise dissimilarity matrices at the grid points, such as the time steps when new (dis)similarities occur and which members diverge. To determine which members diverge and when they diverge, it is first necessary to identify the time steps when new dissimilarities occur. At each such time step, there is a number of dissimilar pairs of members. Counting the number of occurrences in each unique pair, ensemble members can then be sorted in decreasing order of their frequencies. All members whose frequency values attain the maximum frequency are considered diverging. After removing from the list of pairs all those pairs containing diverging ensemble members, the procedure can be repeated until all pairs of dissimilar members have been considered.

For instance, assuming that at a certain time step, member i were assessed as dissimilar to all other $N-1$ members (but otherwise no other dissimilarities existed), then there would be $N-1$ unique pairs of dissimilar members. Computing the frequencies of occurrence, member i would attain the maximum, $N-1$, while all others would have the same frequency 1. Removing all diverging pairs containing member i would remove all dissimilar pairs of members, leaving only member i as diverging. On the other hand, assuming ensemble members could be partitioned in two halves of the same size, each member in one partition being dissimilar to every member in the other partition, but to no other member in its own partition, then all members would be classified as diverging, since they would all have the same frequency of occurrence.

Classifying ensemble members as diverging can help emphasize regions where dissimilarities appear, that is, domain regions that possibly contribute to the emergence of divergences. While pinpointing the exact locations that lead to divergences is not really possible, potential

regions that are likely to play a role in causing divergences may be identified. For instance, when a member is identified at a certain time step t_k as diverging, the region traversed by the member trajectory between time steps t_{k-1} and t_k could be marked as a region potentially causing divergences. Proceeding in this manner for all diverging members (at every integration step of new dissimilarities) at all grid points, all such regions can be identified (see Section 6.3.1 for an example).

6.1.3 Flow-based Similarity

The fine-grained pairwise dissimilarity analysis lends itself to clustering ensemble members hierarchically, in a manner similar to that described in the previous chapter. Previously, the binary similarity of any two members at a grid point depended on the angular deviation relative to local variation of their mode. Now it is the Mahalanobis distance that is applied to assess the particle deviations with respect to the mode variability.

Clustering can be performed per time step or over a forecast interval. In the first case, the global similarity measure is defined, as before, by the percentage of similar locations out of the total number of grid points. For a forecast interval, dissimilarities are averaged over the given interval. Ensemble members are thus clustered based on their transport similarity, rather than their local angular similarity.

6.2 Visual Analysis of the Transport Variability

The main components for visualizing various aspects of the transport variability have been put together into a graphical user interface (see Figure 6.1) comprising four linked views, an overview and a detail view showing the flow variability over the spatial domain, and two abstract views for the variability of the clustering solution.

6.2.1 Flow Variability over the Spatial Domain

The overview (see Figure 6.1 (A)) encodes various scalar measures of the ensemble variability over the entire domain. In this particular example, the geolocated scalar field color codes the aggregated time step divergence count for the entire forecast interval and all ensemble

members) using a sequential grayscale colormap. Users can nevertheless select between several options. For instance, the displayed variability fields can be derived from single time steps or aggregated over time intervals, and for a selection of one or more members (or even the entire ensemble). Moreover, a time slider functionality allows selecting specific time steps or time intervals.

Based on the chosen variability field, users can select individual locations of interest to superimpose enhanced spaghetti plots and thus gain more insight into the evolution of the member trajectories. The grayscale colormaps used in the background support employing a colorblind-safe red-to-blue variation for the trajectories. This allows member trajectories to stand out against the background. However, grayscale colormaps are known to suffer from simultaneous contrast effects, which hinders a correct interpretation of the displayed data (Ware (2004)). To alleviate this issue and ease the variability evaluation, the user is informed about both the global reference values for the member and time divergence counts, and the exact variability value at the selected location.

Enhanced spaghetti plots encode both the integration time and the flow variability, the latter in the form of member divergence counts. Each trajectory is displayed as a curve stripe, the integration time being encoded in the color of the contour of the stripe (from red to blue), and the member divergence count in the color of the stripe. The color of the stripe depends on the (normalized) member divergence count: when the divergence count is very small relative to the local reference value, the color of the stripe is essentially the same as that of the contour at the corresponding time step; as the member divergence count increases, the color of the stripe gradually fades to white.

The local reference value is the smaller value between the minimum divergence count at the given grid point and the member divergence count reference value across the domain. Therefore, when ensemble members have divergence counts similar to the local reference values, the colors of the stripes can barely be distinguished from the colors of the contours. As the stripe colors fade to white with increasing member divergence counts, the trajectories of members with higher variability are emphasized less against the grayscale background than trajectories with lower variability. The temporal evolution of the ensemble members is, nevertheless, not lost, because it is fully preserved in the color of the contours.

An illustration of an enhanced spaghetti plot for the example introduced in the previous section is shown in Figure 6.2 (d). It can be observed that stripes begin to fade rapidly, because

particles tend to spread out from an early stage. As integration step 137 is approached, the color of most stripes fades further, portending the proximate split. After a bimodal distribution is fitted to the data at step 137, the stripes of several trajectories from the upper bundle fade to white. This happens because the particles going upwards travel at considerably different speeds, being therefore considerably more dispersed than a visual inspection of their geometrical shape would reveal. For instance, from the geometry information alone, the trajectories in the lower bundle appear to be more dispersed than those in the upper bundle. Nevertheless, their stripes fade to a lesser extent, because their flow behavior is more similar.

Enhancing spaghetti plots by encoding information derived from the pairwise dissimilarity analysis based on GMMs and the Mahalanobis distance helps thus highlight not only the temporal evolution of the ensemble members, but also the flow variability. It should be noted that GMMs are used to cluster particle positions at consecutive time steps and not the trajectories themselves. GMMs could nevertheless be applied to produce an actual clustering of trajectories[2]. This would also lead to a less cluttered representation (since the flow patterns and the transport variability could be visualized by a few cluster representatives instead of the whole spaghetti plot), but extracting the time steps when and locations where divergences occur would not be a straightforward procedure.

Moreover, while spaghetti plots are indeed more cluttered, enhancing the variability helps bring out the main flow patterns, if existent, and the associated flow variability. Even if trajectories of ensemble members with higher divergence counts are emphasized less than those with lower values, the flow trends and the outliers are shown concurrently. Visualizing just cluster representatives would not offer the same insight, even when accompanied by confidence envelopes showing the variability around these representatives. Furthermore, enhanced spaghetti plots are significantly less affected by a potentially suboptimal clustering solution, since all trajectories are shown to the user.

6.2.2 Flow Variability over Subdomains

The enhanced spaghetti plots allow visualizing the main flow patterns and the variability. Displaying several enhanced spaghetti plots to explore and compare the flow variability at

[2]For instance, by applying GMM on the member trajectories given as $N \times m$ matrices, where N is the number of ensemble members and m the number of time steps. Other approaches are also possible, for instance by applying first a dimensionality reduction technique, such as Principal Component Analysis, and applying GMM on the reduced data.

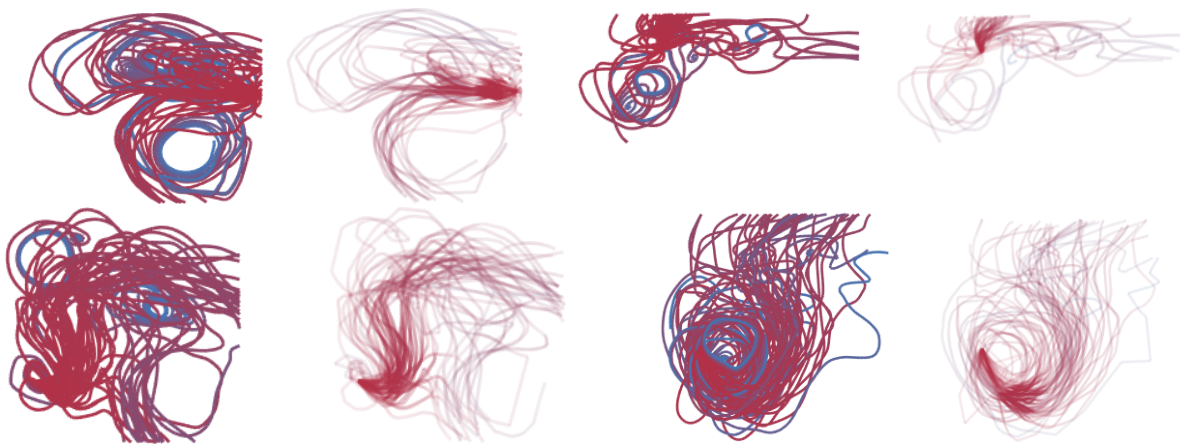


Figure 6.4: In a straightforward downscaling (odd plots – the first plot is the leftmost top plot) clutter and occlusion obscure the flow trends and variability; downscaling with downsampling and variability encoding (even plots) conveys the main flow patterns and variability.

multiple locations concurrently, would, however, lead to massive clutter and occlusion. To alleviate such problems, enhanced spaghetti plots across selected subdomains are downscaled and visualized in a small-multiples layout with miniaturized plots.

Downscaling pathlines as glyphs in zoomed-in scales has been used before to reduce clutter in time-varying vector data (Hlawatsch et al. (2014)), but for individual pathlines of crisp vector fields and not for entire ensembles. The technique proposed here builds upon the concept of small-multiples (Tufte (1983)), but includes the derived transport variability to construct small-multiples preserving the main trends of member trajectories and the flow variability. To this purpose, downscaling trajectories is based on a selection of salient time steps (see Section 6.1.2).

The detail view (see Figure 6.1 (B)) shows a predefined number of downscaled trajectory plots, computed at regular intervals in the spatial domain. As already mentioned in Section 6.1.2 and can be seen in Figure 6.4, a straightforward downscaling could obstruct the flow trends and the associated variability. For instance, in the upper leftmost plot, the ending spiraling structures of several pathlines (especially in the upper bundle), obscure the stable flow pattern at the beginning.

Rather than simply downscaling the trajectory points, a selection of salient time steps is

determined, which preserves the main trends of the transport behavior. Moreover, since downscaled plots are too small to distinguish between the colors of the contours and those of the stripes, the integration time is encoded only in the color of the stripes. The flow variability (given by the member divergence counts) is mapped to the opacity of the stripes, making the main flow trends clearly discernible in the miniaturized trajectory plots.

The benefits of downsampling and opacity encoding can be seen in the counterpart miniaturized trajectory plot for the previous example: the dense sampling of the spiraling parts (where there are hardly any changes in the flow variability), along with encoding the variability in the opacity of the stripes, helps emphasize both the flow patterns and the variability. Figure 6.4 presents several other examples of straightforward downscaling instances and their improved counterparts.

To facilitate a comparative visual analysis of the downscaled trajectory plots, all plots have the aspect ratio of the original domain and the same domain section in the background; the corresponding domain section is displayed in the main spatial view by a purple rectangle, whereas the selected area over which miniaturized trajectory plots are shown is marked by a green square (see Figure 6.1 (A)). The initial domain section shown as a purple rectangle is the bounding box that encompasses all spaghetti plots displayed in the detail view; it can nevertheless be adjusted interactively (as was done in this example). Moreover, to preserve context information, the green square is also displayed in each downscaled plot; the seeding point of the trajectories in each miniaturized plot is marked by a green anchor ball.

6.2.3 Flow-based Similarity Visualization

Like in the previous chapter, the hierarchical clustering of the ensemble members is summarized as a dendrogram (see Figure 6.1 (C)). The dendrogram view shows the clustering solution that considers the global flow similarity over the whole time interval. One or more ensemble members can be interactively selected in the dendrogram (or as text input) to visualize different measures of their spatial variability in the two spatial views.

Further insight into the dynamical evolution of the clustering solution is offered using parallel sets (Bendix et al. (2005)), where a selection of time steps is shown along the horizontal axis and the cluster IDs along the vertical one (see Figure 6.1 (D)). A hierarchical clustering is performed at every selected time step. For the sake of uniformity and simplicity, the

members are partitioned into at most three clusters, whose componence variability can be tracked over the selected time steps. The selected time steps are in powers of two, in order to have a denser sampling at the beginning, since, for the investigated ensembles, changes in regimes occur often quite early in the integration time, making the divergence counts and local clustering solutions vary more earlier in the integration.

Clusters at consecutive selected steps are connected using branches, the thickness of each branch depending on the corresponding number of ensemble members. Since the partition IDs in the clustering solution do not have any meaning, the order of the clusters supports ID continuity from one selected time step to another. The ordering thus depends on the number of common members for every pair of clusters at consecutive time steps, cluster IDs being assigned in decreasing order of the intersection cardinalities. Moreover, to ease tracking split and join events, branches originating from the same cluster have the same color: green, orange, and purple for the first, second, and third cluster, respectively.

This encoding reveals the clustering variability from one time step to another. It offers, however, no insight into how the cluster memberships vary. To this purpose, for the ensemble members in each branch, their cluster IDs at the previous step are determined and the cardinalities of these cluster groups are computed; the ratios of the highest cardinality to the number of ensemble members in the branch is mapped to the opacity of each branch. Thus, the more opaque a branch is, the less the cluster componence has changed from the previous selected time step. In order to see how the cluster componence has changed as well, users can select a branch to visualize the temporal evolution of the clustering solution for the members in the selected branch over all selected steps (see Figure 6.9 for an example).

6.3 Results

The proposed techniques are demonstrated here using two ECMWF wind ensembles of dimensions 101×41, each comprising a control run and another 50 members with perturbed initial conditions.

6.3.1 Transport Variability Analysis

The first ensemble is a time-varying wind forecast at a pressure level of 850 hPa, initiated on October 15, 2012. The geolocated time step divergence count field, aggregated over the entire forecast interval, is shown in Figure 6.1 (A). Spatial locations where data was not available at this pressure level are marked as "Not a Number (NaN)" in the figure. Upon examining the enhanced spaghetti plots over several subdomains, it turned out that pathlines seeded over heterogeneous color regions (such as in Figure 6.1 (B)) exhibit dissimilar flow behavior, as opposed to those seeded over homogeneous areas.

Displaying multiple downscaled enhanced spaghetti plots in the detail view enables a comparative visual analysis of the pathlines seeded over several locations and reveals the numerous changes in flow patterns and variability across the chosen subdomain. For instance, in Figure 6.1 (B), trajectories seeded in the bottom line (to the left), where most of the particles go southwards, show aggregated time step divergence counts similar to the reference value (0.55). The variability increases to the right (up to 0.65), as more particles go northeastwards, but decreases again further to the right, where most particles show no change of regime. In the row directly above (again to the left), the majority of the particles follow a northeastward trend, but with various speeds and increasing dispersion; thus, in spite of the geometrical similarity of the trajectories, the aggregated time step divergence values are around 0.7. In the third row (also to the left), where member trajectories bifurcate geometrically as well, the variability increases further (close to 0.8).

Regarding the transport similarity, Figure 6.1 (C) shows the clustering solution over the entire forecast interval for the ensemble members with perturbed initial conditions. The control run was excluded from this view, because it was very similar to the majority of the other ensemble members and thus caused a large cluster to form around it in the dendrogram. This hindered the identification of natural divisions in the hierarchical tree.

Upon disregarding the control run in the hierarchical clustering, natural groupings were detected in the dendrogram, which revealed an interesting similarity pattern that should be further explored together with domain experts: ensemble member i was typically considerably more similar to ensemble members $i+20$ and, if available, $i+40$, than to other ensemble members. This can also be observed in the dendrogram, where subclusters of two or three ensemble members predominate at low levels, and larger subclusters form at distinctly higher levels.

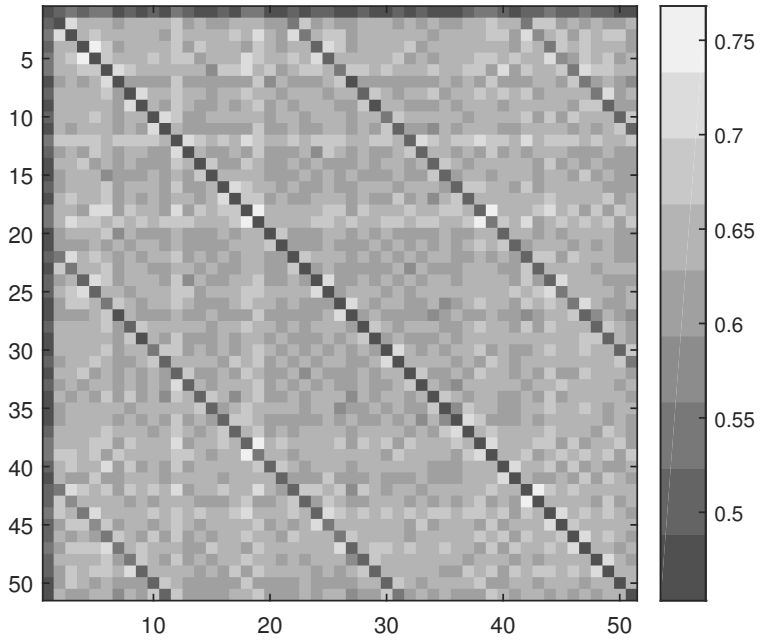


Figure 6.5: Symmetric dissimilarity matrix for the clustering solution of the time-varying ECMWF ensemble.

The pattern can also be seen in Figure 6.5, where the global pairwise dissimilarities were arranged in a symmetric matrix with zeros on the main diagonal – since members are identical to themselves. For the sake of visibility, however, diagonal values were set to the minimum non-zero dissimilarity value in the matrix. The similarity of the control run to the other members is visible in the values in the upmost line, which vary in the lower range of the colormap, while the pattern is noticeable in the darker diagonal lines parallel to the main diagonal.

The second ensemble is a stationary ECMWF wind forecast ensemble at a pressure level of 500 hPa, valid on October 19, 2012. Two geolocated variability fields are displayed over the entire domain in Figure 6.6, aggregated for all members at the first time step (top) and over the whole forecast interval (middle). The isocontours of the geopotential height field of the control run were superimposed as black contours.

Initially, regions of high variability are determined mostly at the pressure centers. Their extent increases nonetheless over time, as particles seeded in regions of lower variability enter regions of higher variability. This tendency can also be noticed in the detail view (bottom),

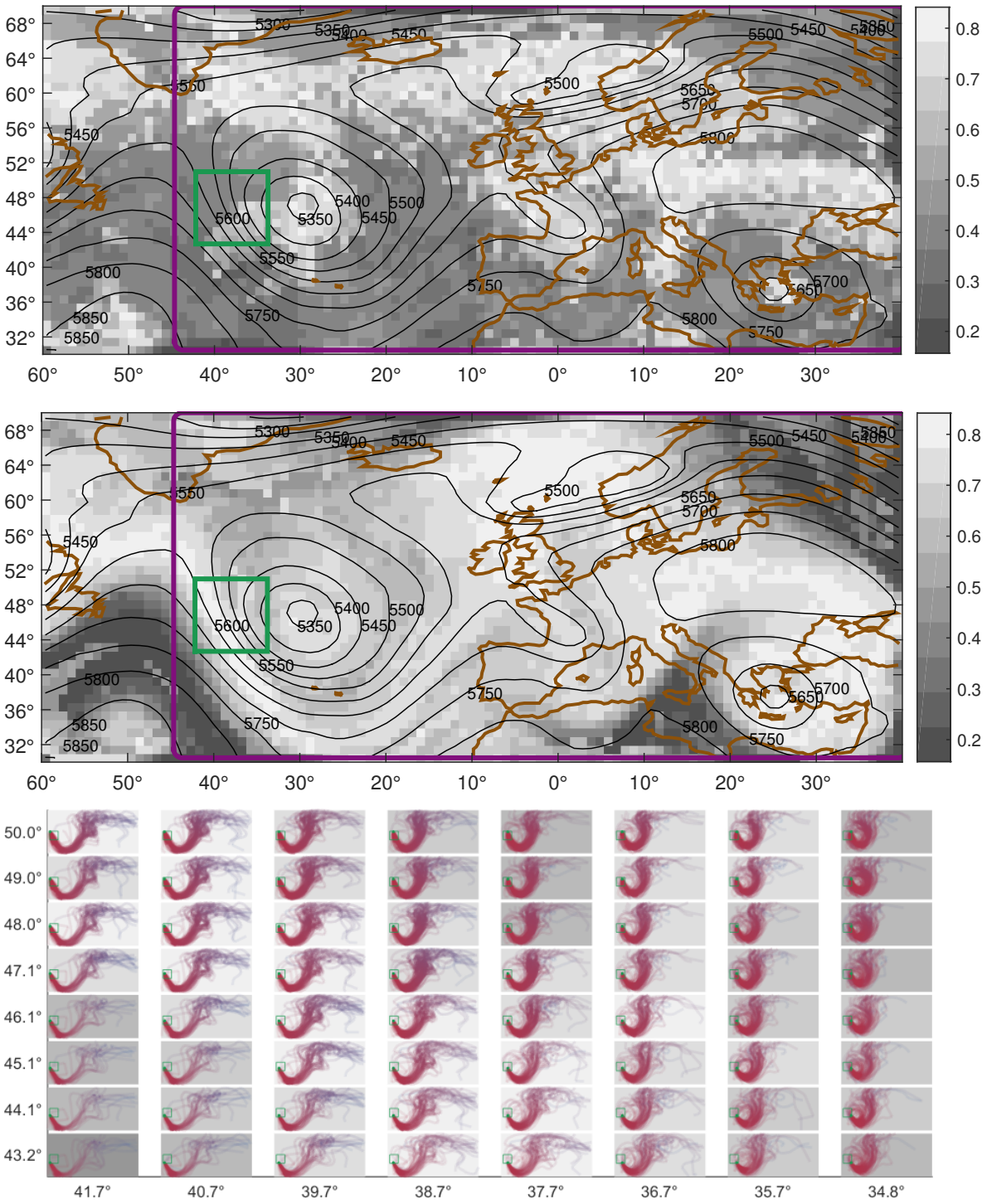


Figure 6.6: Spatial view for the stationary ECMWF ensemble showing the variability for all ensemble members at the (top) first time step and (middle) over the whole time interval; (bottom) detail view with downscaled enhanced trajectory plots.

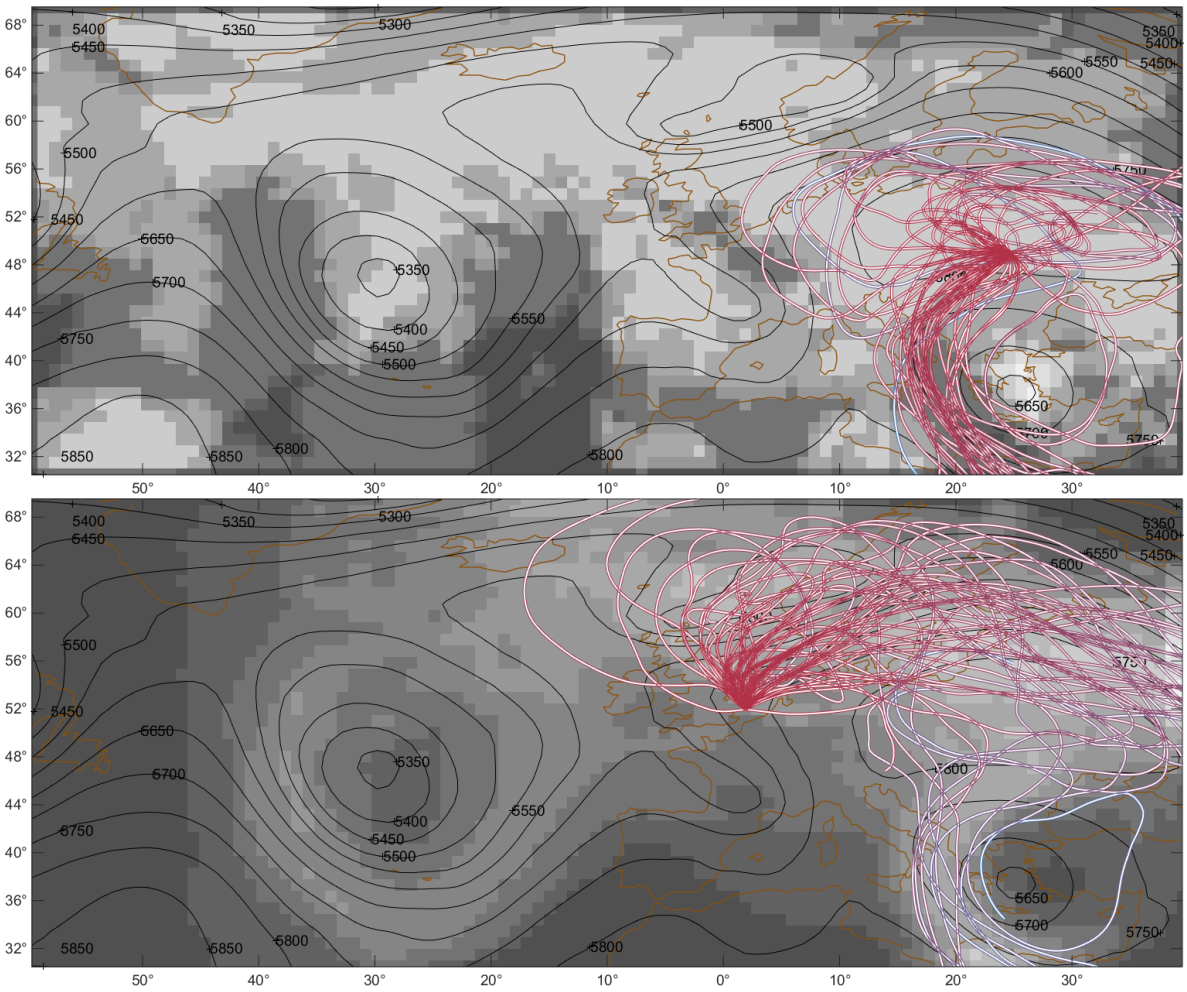


Figure 6.7: Scalar variability fields along with examples of enhanced spaghetti plots for the stationary ECMWF ensemble, showing regions where (top) the first 10% of the dissimilarities and (bottom) all dissimilarities emerge.

where, at all locations over the selected domain, most streamlines follow the southeastwards-northeastwards trend of the isocontours before beginning to disperse. In fact, the variability in this region remains relatively low for much longer than elsewhere in the domain. Eventually, the regions with low variability occupy much smaller extents, e.g., the low-frequency elongated region of high pressure in the bottom left corner.

Figure 6.7 highlights regions potentially causing particles traversing them to diverge. The

scalar fields were obtained by determining, at every grid point, the divergent members and the corresponding regions traversed when the divergent behavior was identified. The scalar values were incremented at all grid points comprising traversed regions, with at most one increment for each grid point from where trajectories were seeded.

In the top figure, the procedure was interrupted as soon as at least 10% of all unique dissimilarities identified at every point for all integration time steps were considered; in the bottom figure, all dissimilarities were considered. The early termination was performed in order to determine regions that are responsible for early splits occurring when particle positions are still likely to be approximated by single Gaussian components.

Since the procedure was terminated quite rapidly, the regions highlighted in the top figure are reminiscent of the regions of high variability in Figure 6.6 (top). This happens despite the initial spreads of the particles being small relative to spreads observed towards the end of the integration, because, for instance, the directional variability is high. An enhanced spaghetti plot illustrating such a situation is displayed in Figure 6.7 (top). It can be seen that the directional variation is high from the very beginning of the integration.

When all dissimilarities are considered, regions are highlighted most in the upper right part of the domain, as can be seen in Figure 6.7 (bottom); an enhanced spaghetti plot where trajectories traverse such regions is also shown in this figure; it can be observed that, even though the particles spread out quite early, several dissimilarities still occur later on during the integration process. While Figure 6.7 (bottom) does highlight those regions where divergences were identified most frequently, it would be interesting to know whether the emphasized regions exhibit the same behavior when divergences are regarded relative to the number of trajectories that traverse these regions.

Thus, Figure 6.8 (top) color codes at each grid point the total number of trajectories that have passed near the grid point, whereas the bottom figure color codes the number of dissimilarities relative to the number of trajectories (considering at most one dissimilarity per trajectory). This reveals that the regions previously highlighted as locations where dissimilarities occur most frequently were emphasized primarily because of the high number of trajectories traversing these regions.

Relatively numerous dissimilarities occur, for instance, at the pressure centers. It can also be observed that, while many trajectories traverse the region detailed in Figure 6.6 (bottom), relatively few diverging trajectories are identified there; moreover, according to Figure 6.7

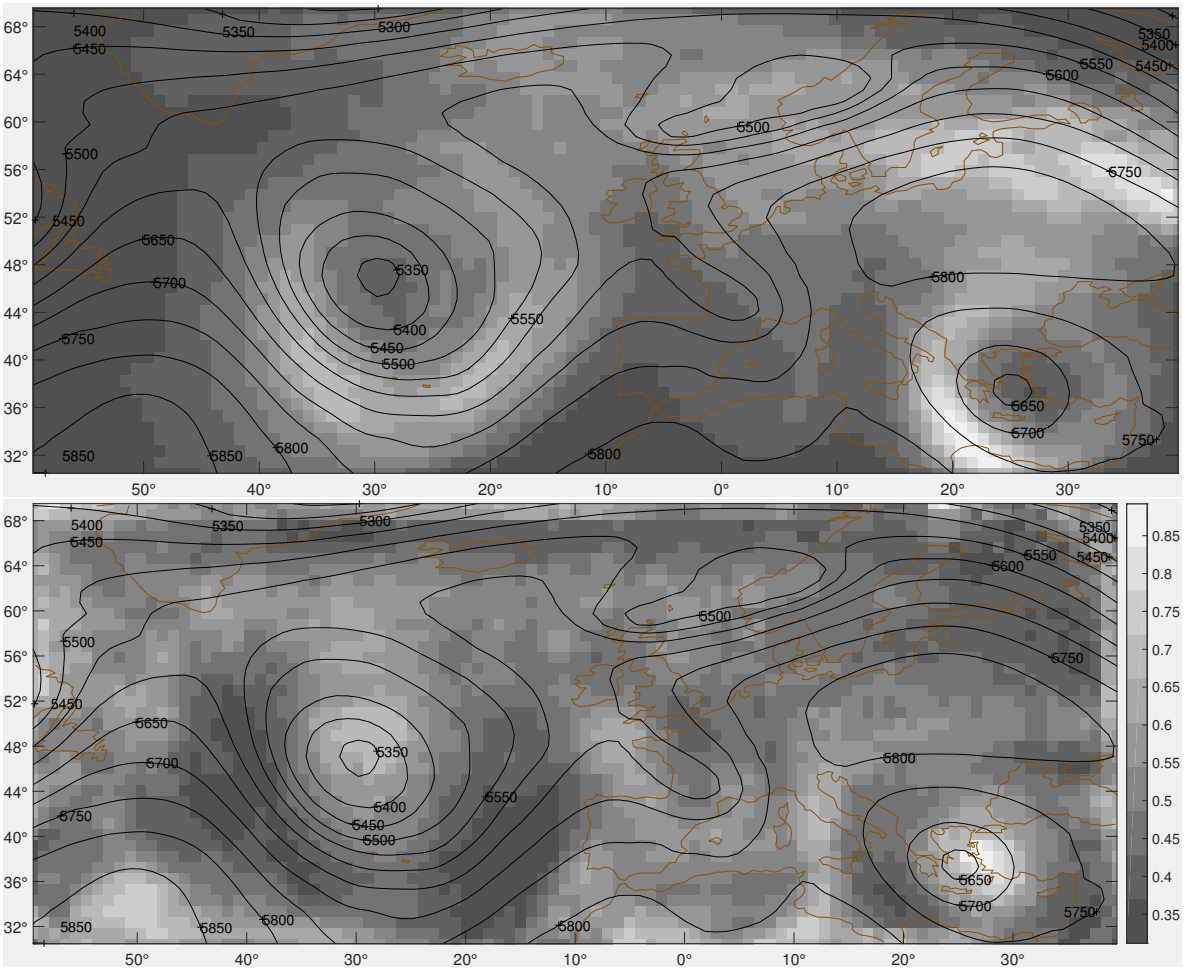


Figure 6.8: For the stationary ECMWF ensemble, the geolocated scalar fields encode at each grid point (top) how many member trajectories from the whole ensemble pass near the grid point and (bottom) at how many of these trajectories new dissimilarities are identified.

(top), most of these divergences occur later on in the integration. For the upper left regions, on the other hand, significantly fewer trajectories traverse the regions, but many of them are diverged meanwhile.

The clustering solution over the entire forecast interval takes this time all ensemble members into consideration. Its dynamical evolution at seven selected time steps is shown in Figure 6.9 (top). The clustering solution looks quite stable at the first two time steps, but the variability

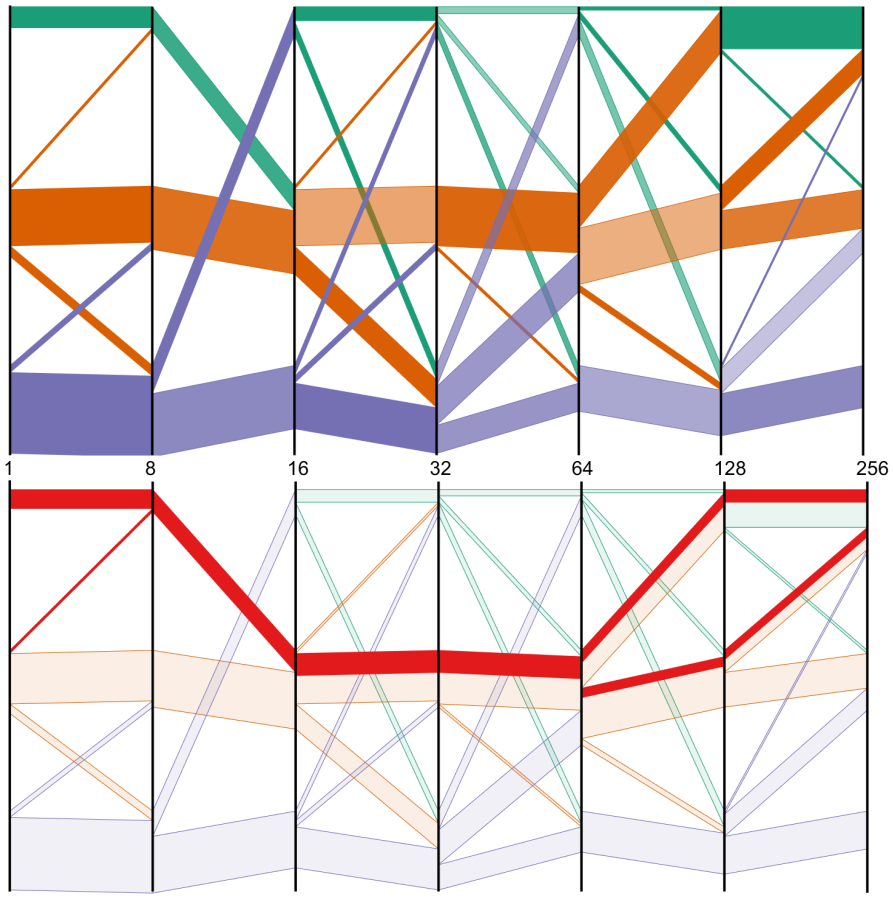


Figure 6.9: (Top) Clustering variability for a stationary ECMWF ensemble at selected time steps (the cluster IDs are shown on the vertical axis, the selected time steps on the horizontal axis); (bottom) variability of cluster componence for selected branches.

increases thereafter.

An initial insight into how the cluster componence of the individual branches varies is given by the opacity of the branches. For example, the thickest of the orange branches joining the second clusters at selected time steps 16 and 32 is quite transparent. This implies that the ensemble members in the branch are a mixture of the members in the green and orange branches that joined the second cluster at step 16. On the other hand, the thinner orange branch that joins the second and third clusters is opaque, meaning that the componence of the members in the branch has not changed from the previous step.

For further insight into the membership variability, users can select branches to see how the

cluster memberships of the ensemble members in these branches vary over the selected steps. Figure 6.9 (bottom) shows an example, where the selected branch is the green branch joining the second cluster at step 16. The branch is quite opaque in the top figure, because, except for a few ensemble members that joined from the second cluster, the members in the branch are the same as at the previous step. In fact, but for a temporary split from time step 64 to step 128, the members remain clustered together over all other selected time steps.

6.3.2 Implementation, Performance Analysis, and Scalability

The tests performed on the ECMWF ensemble data were ran on a standard desktop PC, equipped with an Intel i7-4790 quad-core processor running at 3.6 GHz and with 12 GB RAM. Since fitting GMMs and computing the pairwise (dis)similarities using the Mahalanobis distance can be performed in parallel at every time step and every grid point, these operations can be parallelized in a straightforward way on the GPU. This yields computation times of less than five seconds for all datasets. Moreover, the comparative analysis session can be handled at interactive frame rates.

The proposed approaches are scalable for ensembles larger than those used here. As described in the previous chapter as well, ensembles of larger sizes could require techniques like focus and context for the dendrogram. The same holds for the readability of the parallel sets, which depends on the number of selected time steps. These abstract views are, nonetheless, not affected by an increase in the grid dimensions. The detail view is also not affected, since the number of downscaled trajectory plots displayed is fixed; the spatial view over the entire domain is, however, influenced, because the size of the pixel of each grid point decreases.

6.4 Conclusion, Evaluation, and Future Work

The framework proposed in this chapter enabled an interactive comparative visual analysis of the transport variability of 2D vector field ensembles. By modeling particle positions at every time step and grid point using mixtures of Gaussians and applying the Mahalanobis distance on the fitted Gaussian modes, the pairwise (dis)similarities of the ensemble members were determined. This enabled assessing pairs of ensemble members as (dis)similar based on their deviations relative to the variability exhibited by the modes, instead of prescribed

artificial thresholds. Then, several measures were derived from this fine-grained analysis, based on which several visualization techniques were proposed to convey the variability of the spatio-temporal evolution of an ensemble and its clustering solution.

The informal feedback provided by domain experts from meteorology was positive. The experts found the proposed methods effective in gaining a fast insight into the flow predictability over the spatial domain, the time interval over which the forecasts can be trusted, and the regions where numerous divergences are likely to be detected. They also described the techniques as useful in identifying regions of different qualitative flow, and highly appreciated the small-multiples approach that allowed displaying the flow behavior at several locations concurrently with less clutter and occlusion. Nonetheless, they were initially puzzled about some neighboring locations where the flow appeared to be similar, but the variability values were quite different. Detailed inspections of such cases revealed the dissimilarities in flow behavior that had led to the different variability values; occasionally, however, the differences were caused by a suboptimal partitioning of the mixture algorithm. The experts were also surprised by the spread of particle positions, given the geometrically similar trajectories, and commended the enhanced spaghetti plots for highlighting the flow variability. Regarding the time-varying ensemble, they were interested to further investigate the reason for the pattern present in the clustering solution over the entire forecast interval.

Potential directions to be explored in the future include enriching the possibilities of analyzing the variability of the clustering solution and the dynamics of the ensemble. Further means to determine regions that are likely to cause important changes of regime, as well as the effect of these regions on particles traversing them, are also of interest; for instance, local methods (such as the directional approach from the previous chapter) could be coupled with global methods, in order to combine advantages of both. Extending the analysis to 3D vector data is also an important direction. While the mathematical extension is straightforward, clutter and occlusion are significantly more challenging than in 2D, and require novel graphical abstractions to reveal the ensemble variability.

7

Conclusion and Future Directions

In this thesis, several novel approaches were proposed to extract variability aspects of scalar and vector ensembles that had not been addressed before. To this purpose, data samples were summarized statistically at every domain location, in order to allow deriving measures for specific variability characteristics of ensembles.

In the case of scalar field ensembles, the introduced methods assessed the stability of critical points with respect to potential variations in position and type. This was achieved by propagating the summarized uncertainty from scalar fields to derived quantities, such as gradient vectors and Hessian matrices, and deriving functions indicating at every grid point whether critical points of certain types can be expected to occur in the vicinity of the grid point. Critical points are important for scalar ensembles, because they relate to salient surface features and associated topology events, such as features merging or splitting.

The methods were validated on several synthetic and real-world data sets, which demonstrated the benefits of the approach. For instance, the methods helped distinguish between regions where critical points were likely to appear and regions where emerging critical points were probably transitory states or noise. Moreover, they allowed inferring on whether the behavior of features around certain locations was likely to be stable (for instance, frequent instances of surface elements emerging at minima) or unstable (for example, oscillating structures, such as surface elements emerging at minima alternating with surface elements vanishing at maxima).

Compared to displaying the actual critical points of the ensemble (or a few representative points), the methods produced more insightful results, since they highlighted regions where critical points could have emerged in ensemble members that had not been realized.

For vector ensembles, the proposed approaches went beyond standard techniques of summarizing ensembles using statistical parameters like the mean and standard deviation, which offer an accurate representation of the typical ensemble behavior and associated variability only when Gaussian assumptions hold. Distributions were approximated instead using mixture models, which are effective at modeling complex distributions by relatively compact sets of parameters. Based on these parameters, the pairwise (dis)similarities of the ensemble members were assessed relative to the variability of the fitted mixtures, rather than to artificially prescribed thresholds.

The derived measures were shown to be effective for performing a local dissimilarity analysis of the directional variability, as well as a global investigation of the transport variability. This allowed revealing various aspects of the ensemble variability, such as distinguishing between regions of high and low local or global variability, and clustering ensemble members based on their local angular or global transport similarity across the domain. Moreover, the mixture models were employed to allow a concurrent comparative visualization over selected domains, both of local aspects – displaying in an intuitive manner the main features of potentially multimodal directional distributions – and global aspects – exhibiting the main flow patterns and the temporal evolution of the flow variability. The methods were applied on meteorological ensembles and the results were discussed with domain experts, who responded positively to the proposed techniques.

It should be noted that the proposed methods are, to a great extent, generic and can be used for other types of ensemble data as well. This holds, in particular, for the frameworks developed for vector field ensembles; nevertheless, the methods developed to assess stability characteristics of critical points in scalar fields lend themselves to being extended to vector fields as well. Thus, modeling the data samples at every grid location with mixture models and applying similarity measures that consider member deviations relative to the variability allowed by each mixture component, as well as aggregating these measures over the domain in order to group similar members together, can also be applied to scalar or higher dimension data.

Possible future directions are numerous. To start with, 3D extensions are inherently prone

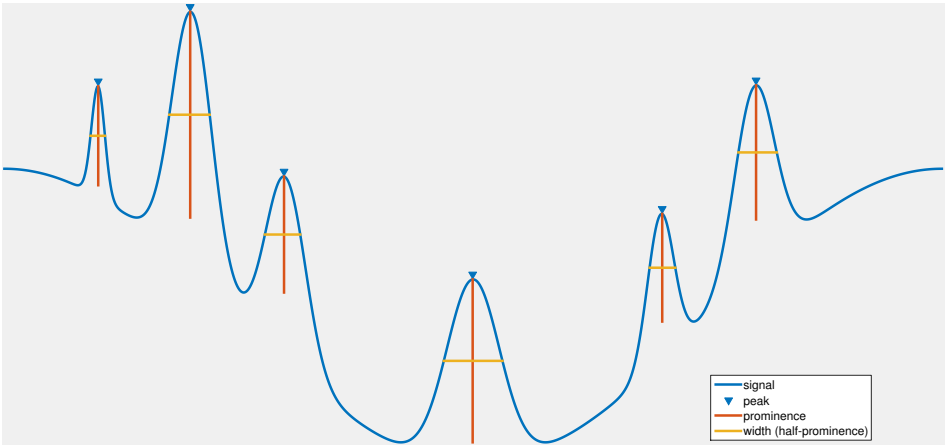


Figure 7.1: Extracting prominent peaks from contours and summarizing the peaks by their locations, heights, and widths to facilitate tracking events such as splits and merges.

to clutter and occlusions, particularly when time-varying data is considered; future work investigating effective means of alleviating these issues is thus indispensable. A smart coupling of local and global techniques is worthy investigating in this context.

Time-varying ensemble data is also an important topic that needs to be explored in more detail, for instance, to extract features of interest and follow their temporal evolution across the ensemble. A potential idea to simplify tracking would be to summarize features by means of a few compact parameters, which could be used to facilitate tracking. Figure 7.1 shows an example, where salient peaks in a contour are extracted and described using their locations, heights, and widths. Other aspects regarding the predictability of features that could be extracted and encoded visually include: How long is a certain regime expected to persist for? Is there a change of regime likely and, if so, when is it expected to occur? Moreover, what patterns can be expected to develop?

Last, but not least, a framework to assess and evaluate the effectiveness of communicating the variability in ensemble data is necessary. It should be borne in mind that the ultimate goal of analyzing and visualizing ensemble forecasts is supporting users in understanding the evolution of systems and helping them make informed decisions.

Bibliography

Alhamed, A., Lakshmivarahan, S., and Stensrud, D. J. (2002). Cluster Analysis of Multimodel Ensemble Data from SAMEX. *Mon. Weather Rev.*, 130(2):226–256.

Allendes Osorio, R. and Brodlie, K. W. (2009). Uncertain Flow Visualization Using LIC. *Proc. TPCG'09*, pages 215–222.

Anderson, T. W. and Darling, D. A. (1954). A Test of Goodness of Fit. *JASA*, 49(268):765–769.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions. *J. Mach. Learn. Res.*, 6:1345–1382.

Becker, R. A. and Cleveland, W. S. (1987). Brushing Scatterplots. *Technometrics*, 29(2):127–142.

Bendix, F., Kosara, R., and Hauser, H. (2005). Parallel Sets: Visual Analysis of Categorical Data. *IEEE INFOVIS*, pages 133–140.

Bensema, K., Gosink, L., Obermaier, H., and Joy, K. (2015). Modality-driven Classification and Visualization of Ensemble Variance. *IEEE TVCG*, pages 2289–2299.

Bertin, J. (1983). *Semiology of Graphics*. The University of Wisconsin Press.

Bhatia, H., Jadhav, S., Bremer, P.-T., Chen, G., Levine, J. A., Nonato, L. G., and Pascucci, V. (2011). Edge maps: Representing Flow with Bounded Error. In *Proc. PacificVis'11*, pages 75–82.

Bonneau, G.-P., Hege, H.-C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., and Schultz, T. (2014). Overview and State-of-the-art of Uncertainty Visualization. In *Scientific Visualization*, pages 3–27. Springer.

Bordoloi, U. D., Kao, D. L., and Shen, H.-W. (2004). Visualization Techniques for Spatial Probability Density Function Data. *JDS*, 3:153–162.

Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramee, R. S., Hauser, H., Ward, M., and Chen, M. (2013). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In *STAR Proc. Eurographics'13*, pages 39–63.

Botchen, R. P., Weiskopf, D., and Ertl, T. (2005). Texture-based Visualization of Uncertainty in Flow Fields. *Proc. IEEE Visualization*, pages 647–654.

Botchen, R. P., Weiskopf, D., and Ertl, T. (2006). Interactive Visualization of Uncertainty in Flow Fields Using Texture-based Techniques. In *Proc. Intl. Symp. on Flow Visualization*.

Bremer, P.-T., Edelsbrunner, H., Hamann, B., and Pascucci, V. (2003). A Multi-resolution Data Structure for 2D Morse-Smale Functions. *Proc. IEEE Visualization*, pages 139–146.

Brodlie, K., Osorio, R. A., and Lopes, A. (2012). A Review of Uncertainty in Data Visualization. In *Expanding the frontiers of visual analytics and visualization*, pages 81–109. Springer.

Brown, R. (2004). Animated Visual Vibrations as an Uncertainty Visualisation Technique. In *Proc. GRAPHITE'04*, pages 84–89.

Bruckner, S. and Möller, T. (2010). Result-Driven Exploration of Simulation Parameter Spaces for Visual Effects Design. *IEEE TVCG*, 16(6):1468–1476.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Wadsworth Publishing Company, Belmont, California.

Cedilnik, A. and Rheingans, P. (2000). Procedural Annotation of Uncertain Information. In *Proc. VIS'00*, pages 77–84.

Chen, H., Zhang, S., Chen, W., Mei, H., Zhang, J., Mercer, A., Liang, R., and Qu, H. (2015). Uncertainty-aware Multidimensional Ensemble Data Visualization and Exploration. *IEEE TVCG*, 21(9):1072–1086.

Correa, C., Chan, Y.-H., and Ma, K.-L. (2009). A Framework for Uncertainty-aware Visual Analytics. In *Proc. IEEE VAST*, pages 51–58.

De Leeuw, W. and Van Liere, R. (1999). Collapsing Flow Topology Using Area Metrics. In *Proc. VIS'99*, pages 349–354.

Delmarcelle, T. and Hesselink, L. (1994). The Topology of Symmetric, Second-order Tensor Fields. In *Proc. VIS'94*, pages 140–147.

Demir, I., Dick, C., and Westermann, R. (2014). Multi-charts for Comparative 3D Ensemble Visualization. *IEEE TVCG*, 20(12):2694–2703.

Demir, I., Jarema, M., and Westermann, R. (2016a). Visualizing the Central Tendency of Ensembles of Shapes. In *Proc. ACM SA16VIS*.

Demir, I., Kehrer, J., and Westermann, R. (2016b). Screen-space Silhouettes for Visualizing Ensembles of 3D Isosurfaces. In *Proc. PacificVis'16*, pages 204–208.

Demir, I. and Westermann, R. (2015). Vector-to-Closest-Point Octree for Surface Ray-Casting. In Bommes, D., Ritschel, T., and Schultz, T., editors, *Proc. VMV'15*, pages 65–72.

Dey, T. K. and Wenger, R. (2007). Stability of Critical Points with Interval Persistence. *DCG*, 38(3):479–512.

Djurcilov, S., Kim, K., Lermusiaux, P., and Pang, A. (2002). Visualizing Scalar Volumetric Data with Uncertainty. *Comput. & Graph.*, 26(2):239–248.

Dunn, O. J. (1961). Multiple Comparisons Among Means. *JASA*, 56(293):52–64.

Edelsbrunner, H., Harer, J., and Zomorodian, A. (2003). Hierarchical Morse Complexes for Piecewise Linear 2-Manifolds. *Discrete Comput. Geom.*, 30:87–107.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological Persistence and Simplification. In *Proc. FOCS'00*, pages 454–463.

Feng, D., Kwock, L., Lee, Y., and Taylor, R. (2010). Matching Visual Saliency to Confidence in Plots of Uncertain Data. *IEEE TVCG*, 16(6):980–989.

Ferstl, F., Bürger, K., and Westermann, R. (2016a). Streamline Variability Plots for Characterizing the Uncertainty in Vector Field Ensembles. *IEEE TVCG*, 22(1):767–776.

Ferstl, F., Kanzler, M., Rautenhaus, M., and Westermann, R. (2016b). Visual Analysis of Spatial Variability and Global Correlations in Ensembles of Iso-Contours. *Comput. Graph. Forum*, 35(3):221–230.

Fisher, N. (1995). *Statistical Analysis of Circular Data*. Cambridge Univ. Press.

Fisher, N., Lewis, T., and Embleton, B. (1987). *Statistical Analysis of Spherical Data*. Cambridge Univ. Press.

Fofonov, A., Molchanov, V., and Linsen, L. (2016). Visual Analysis of Multi-Run Spatio-Temporal Simulations Using Isocontour Similarity for Projected Views. *IEEE TVCG*, 22(8):2037–2050.

Fuchs, R. and Hauser, H. (2009). Visualization of Multi-Variate Scientific Data. *Comput. Graph. Forum*, 28(6):1670–1690.

Gansner, E. R. and North, S. C. (2000). An Open Graph Visualization System and its Applications to Software Engineering. *Softw. Pract. Exper.*, 30(11):1203–1233.

Gosink, L., Bensema, K., Pulsipher, T., Obermaier, H., Henry, M., Childs, H., and Joy, K. I. (2013). Characterizing and Visualizing Predictive Uncertainty in Numerical Ensembles through Bayesian Model Averaging. *IEEE TVCG*, 19(12):2703–2712.

Gosling, J. (1995). *Introductory Statistics*. Pascal Press.

Griethe, H. and Schumann, H. (2006). The Visualization of Uncertain Data: Methods and Problems. In *SimVis*, pages 143–156.

Grigoryan, G. and Rheingans, P. (2004). Point-based Probabilistic Surfaces to Show Surface Uncertainty. *IEEE TVCG*, 10(5):564–573.

Guo, H., He, W., Peterka, T., Shen, H.-W., Collis, S. M., and Helmus, J. J. (2016). Finite-Time Lyapunov Exponents and Lagrangian Coherent Structures in Uncertain Unsteady Flows. *IEEE TVCG*, 22(6):1672–1682.

Hamerly, G. and Elkan, C. (2004). Learning the K in K-Means. In *Proc. NIST'03*, pages 281–288.

Hao, L., Healey, C. G., and Bass, S. A. (2016). Effective Visualization of Temporal Ensembles. *IEEE TVCG*, 22(1):787–796.

Harrower, M. and Brewer, C. A. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *Cartogr. J.*, 40(1):27–37.

Heine, C., Leitte, H., Hlawitschka, M., Iuricich, F., Floriani, L. D., Scheuermann, G., Hagen, H., and Garth, C. (2016). A Survey of Topology-based Methods in Visualization. *Comput. Graph. Forum*, 35(3):643–667.

Heinrich, J. and Weiskopf, D. (2013). State of the Art of Parallel Coordinates. *STAR Proc. Eurographics'13*, 2013:95–116.

Hengl, T. (2003). Visualisation of Uncertainty Using the HSI Colour Model: Computations with Colours. In *Proc. GeoComputation'03 (on CD-ROM)*.

Hennig, C. (2010). Methods for Merging Gaussian Mixture Components. *ADAC*, 4(1):3–34.

Hlawatsch, M., Leube, P., Nowak, W., and Weiskopf, D. (2011). Flow Radar Glyphs – Static Visualization of Unsteady Flow with Uncertainty. *IEEE TVCG*, 17(12):1949–1958.

Hlawatsch, M., Sadlo, F., Jang, H., and Weiskopf, D. (2014). Pathline Glyphs. *Comput. Graph. Forum*, 33(2):497–506.

Hollister, B. E. and Pang, A. (2013). Interpolation of Non-Gaussian Probability Distributions for Ensemble Visualization. Technical report.

Hollister, B. E. and Pang, A. (2016). Visual Analysis of Transport Similarity in 2D CFD Ensembles. In *IS&T Electronic Imaging Conference on Visualization and Data Analysis*.

Höllt, T., Magdy, A., Chen, G., Gopalakrishnan, G., Hoteit, I., Hansen, C. D., and Hadwiger, M. (2013). Visual Analysis of Uncertainties in Ocean Forecasts for Planning and Operation of Off-shore Structures. In *Proc. PacificVis'13*, pages 185–192.

Hummel, M., Obermaier, H., Garth, C., and Joy, K. (2013). Comparative Visual Analysis of Lagrangian Transport in CFD Ensembles. *IEEE TVCG*, 19(12):2743–2752.

Hunter, G. J. (1999). Managing Uncertainty in GIS. *JGIS*, 2:633–641.

Inselberg, A. (1985). The Plane with Parallel Coordinates. *Visual Comput.*, 1(2):69–91.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-means. *Pattern Recogn. Lett.*, 31(8):651 – 666.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc.

Jiao, F., Phillips, J., Gur, Y., and Johnson, C. (2012). Uncertainty Visualization in HARDI Based on Ensembles of ODFs. In *Proc. PacificVis'12*, pages 193–200.

Johnson, C. (2004). Top Scientific Visualization Research Problems. *IEEE Comput. Graph. Appl.*, 24(4):13–17.

Johnson, C. R. and Sanderson, A. R. (2003). A Next Step: Visualizing Errors and Uncertainty. *IEEE Comput. Graph. Appl.*, 23(5):6–10.

Kao, D., Dungan, J. L., and Pang, A. (2001). Visualizing 2D Probability Distributions from EOS Satellite Image-derived Data Sets: a Case Study. In *Proc. VIS'01*, pages 457–589.

Kao, D., Luo, A., Dungan, J. L., and Pang, A. (2002). Visualizing Spatially Varying Distribution Data. In *Proc. Int'l. Conf. Information Visualization (IV)*, pages 219–225.

Kao, D. L., Kramer, M. G., Love, A. L., Dungan, J. L., and Pang, A. T. (2013). Visualizing Distributions from Multi-return LIDAR Data to Understand Forest Structure. *Cartogr. J.*

Kardos, J., Moore, A., and Benwell, G. L. (2003). The Visualisation of Uncertainty in Spatially-referenced Attribute Data Using TRUSTworthy Data Structures. In *Proc. SIRC'03*, pages 11–26.

Kehrer, J., Filzmoser, P., and Hauser, H. (2010). Brushing Moments in Interactive Visual Analysis. *Comput. Graph. Forum*, 29(3):813–822.

Kehrer, J. and Hauser, H. (2013). Visualization and Visual Analysis of Multi-faceted Scientific Data: A Survey. *IEEE TVCG*, 19(3):495–513.

Kehrer, J., Muigg, P., Doleisch, H., and Hauser, H. (2011). Interactive Visual Analysis of Heterogeneous Scientific Data across an Interface. *IEEE TVCG*, 17(7):934–946.

Laramee, R. S., Hauser, H., Zhao, L., and Post, F. H. (2007). Topology-based Flow Visualization, the State of the Art. In *Topology-based methods in visualization*, pages 1–19. Springer.

Leutbecher, M. and Palmer, T. (2008). Ensemble Forecasting. *J. Comput. Phys.*, 227(7):3515–3539.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *JASA*, 62(318):399–402.

Liu, R., Guo, H., Zhang, J., and Yuan, X. (2016). Comparative Visualization of Vector Field Ensembles Based on Longest Common Subsequence. *Proc. PacificVis'16*, pages 96–103.

Liu, S., Levine, J., Bremer, P., and Pascucci, V. (2012). Gaussian Mixture Model Based Volume Visualization. In *Proc. LDAV'12*, pages 73–77.

Lodha, S. K., Pang, A., Sheehan, R. E., and Wittenbrink, C. M. (1996a). UFLOW: Visualizing Uncertainty in Fluid Flow. In *Proc. VIS'96*, pages 249–ff.

Lodha, S. K., Wilson, C. M., and Sheehan, R. E. (1996b). LISTEN: Sounding Uncertainty Visualization. In *Proc. VIS'96*, pages 189–ff.

Love, A. L., Pang, A., and Kao, D. L. (2005). Visualizing Spatial Multivalue Data. *IEEE Comput. Graph. Appl.*, 25(3):69–79.

Lundström, C., Ljung, P., Persson, A., and Ynnerman, A. (2007). Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. *IEEE TVCG*, 13(6):1648–1655.

Luo, A., Kao, D., and Pang, A. (2003). Visualizing Spatial Distribution Data Sets. In *VisSym*, volume 3, pages 29–38.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. (2005). Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *CaGIS*, 32(3):139–160.

Mahalanobis, P. C. (1936). On the Generalised Distance in Statistics. *Proc. Nat. Instit. Sci. India*, 2(1):49–55.

Mardia, K. and Jupp, P. (2000). *Directional statistics*. Wiley & Sons.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press Inc.

Matkovic, K., Gracanin, D., Jelovic, M., Ammer, A., Lez, A., and Hauser, H. (2010). Interactive Visual Analysis of Multiple Simulation Runs Using the Simulation Model View: Understanding and Tuning of an Electronic Unit Injector. *IEEE TVCG*, 16(6):1449–1457.

McCormick, B. H. (1988). Visualization in Scientific Computing. *SIGBIO Newsl.*, 10(1):15–21.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Molchanov, V. and Linsen, L. (2014). Visual Exploration of Patterns in Multi-run Time-varying Multi-field Simulation Data Using Projected Views. In *Proc. WSCG'14*, volume 21, pages 49–58.

Nocke, T., Flechsig, M., and Böhm, U. (2007). Visual Exploration and Evaluation of Climate-related Simulation Data. In *Proc. WSC'07*, pages 703–711.

Obermaier, H., Bensema, K., and Joy, K. I. (2016). Visual Trends Analysis in Time-Varying Ensembles. *IEEE TVCG*, 22(10):2331–2342.

Oeltze, S., Lehmann, D. J., Kuhn, A., Janiga, G., Theisel, H., and Preim, B. (2014). Blood Flow Clustering and Applications inVirtual Stenting of Intracranial Aneurysms. *IEEE TVCG*, 20(5):686–701.

Oliveira-Brochado, A. and Martins, F. V. (2005). Assessing the Number of Components in Mixture Models: a Review. Technical report, Universidade do Porto, Faculdade de Economia do Porto.

Olston, C. and Mackinlay, J. D. (2002). Visualizing Data with Bounded Uncertainty. In *Proc. InfoVis'02*, pages 37–40.

Otto, M., Germer, T., Hege, H.-C., and Theisel, H. (2010). Uncertain 2D Vector Field Topology. *Comput. Graph. Forum*, 29(2):347–356.

Otto, M., Germer, T., and Theisel, H. (2011). Uncertain Topology of 3D Vector Fields. In *Proc. PacificVis'11*, pages 67–74.

Palmer, T. N. (2000). Predicting Uncertainty in Forecasts of Weather and Climate. *Rep. Prog. Phys*, 63(2):71.

Pang, A. (2001). Visualizing Uncertainty in Geo-spatial Data. In *Proc. Intersections between Geospatial Information and Information Technology*, pages 1–14.

Pang, A. T., Wittenbrink, C. M., and Lodha, S. K. (1997). Approaches to Uncertainty Visualization. *Visual Comput.*, 13(8):370–390.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Phil. Trans.*, 185:71–110.

Petz, C., Pöthkow, K., and Hege, H.-C. (2012). Probabilistic Local Features in Uncertain Vector Fields with Spatial Correlation. *Comput. Graph. Forum*, 31(3pt2):1045–1054.

Pfaffelmoser, T., Mihai, M., and Westermann, R. (2013). Visualizing the Variability of Gradients in Uncertain 2D Scalar Fields. *IEEE TVCG*, 19(11):1948–1961.

Pfaffelmoser, T., Reitinger, M., and Westermann, R. (2011). Visualizing the Positional and Geometrical Variability of Isosurfaces in Uncertain Scalar Fields. *Comput. Graph. Forum*, 30(3):951–960.

Pfaffelmoser, T. and Westermann, R. (2012). Visualization of Global Correlation Structures in Uncertain 2D Scalar Fields. *Comput. Graph. Forum*, 31(3pt2):1025–1034.

Pfaffelmoser, T. and Westermann, R. (2013). Correlation Visualization for Structural Uncertainty Analysis. *Int. J. Uncertain Quantif.*, 3(2):171–186.

Piringer, H., Pajer, S., Berger, W., and Teichmann, H. (2012). Comparative Visual Analysis of 2D Function Ensembles. *Comput. Graph. Forum*, 31(3pt3):1195–1204.

Plumlee, M. and Ware, C. (2003). Integrating Multiple 3D Views Through Frame-of-reference Interaction. In *Proc. CMV'03*, pages 34–43.

Pöthkow, K. and Hege, H.-C. (2011). Positional Uncertainty of Isocontours: Condition Analysis and Probabilistic Measures. *IEEE TVCG*, 17(10):1393–1406.

Pöthkow, K. and Hege, H.-C. (2013). Nonparametric Models for Uncertainty Visualization. *Comput. Graph. Forum*, 32(3pt2):131–140.

Pöthkow, K., Weber, B., and Hege, H.-C. (2011). Probabilistic Marching Cubes. *Comput. Graph. Forum*, 30(3):931–940.

Potter, K. (2006). Methods for Presenting Statistical Information: The Box Plot. *Visualization of Large and Unstructured Data Sets*, S-4:97–106.

Potter, K., Gerber, S., and Anderson, E. W. (2013). Visualization of Uncertainty without a Mean. *IEEE Comput. Graph. Appl.*, 33(1):75–79.

Potter, K., Kirby, M., Xiu, D., and Johnson, C. R. (2012a). Interactive Visualization of Probability and Cumulative Density Functions. *Int. J. Uncertainty Quantif.*, 2(4):397–412.

Potter, K., Kniss, J., Riesenfeld, R., and Johnson, C. R. (2010). Visualizing Summary Statistics and Uncertainty. *Comput. Graph. Forum*, 29(3):823–832.

Potter, K., Rosen, P., and Johnson, C. R. (2012b). From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer.

Potter, K., Wilson, A., Bremer, P.-T., Williams, D., Doutriaux, C., Pascucci, V., and Johnson, C. R. (2009). Ensemble-vis: A Framework for the Statistical Visualization of Ensemble Data. In *Proc. ICDMW'09*, pages 233–240.

Ray, S. (2003). *Distance-based Model-selection with Application to the Analysis of Gene Expression Data*. PhD thesis, PSU.

Razali, N. M. and Wah, Y. B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Statistical Modeling and Analytics*, 2(1):21–33.

Reh, A., Gusenbauer, C., Kastner, J., Gröller, M. E., and Heinzl, C. (2013). MObjects: A Novel Method for Visualization and Interactive Exploration of Defects in Industrial XCT Data. *IEEE TVCG*, 19(12):2906–2915.

Reininghaus, J., Kotava, N., Gunther, D., Kasten, J., Hagen, H., and Hotz, I. (2011). A Scale Space Based Persistence Measure for Critical Points in 2D Scalar Fields. *IEEE TVCG*, 17(12):2045–2052.

Rhodes, P. J., Laramee, R. S., Bergeron, R. D., Sparr, T. M., et al. (2003). Uncertainty Visualization Methods in Isosurface Rendering. In *Proc. Eurographics'03*, volume 2003, pages 83–88.

Ristovski, G., Preusser, T., Hahn, H. K., and Linsen, L. (2014). Uncertainty in Medical Visualization: Towards a Taxonomy. *Comput. & Graph.*, 39:60–73.

Roberts, J. C. (2007). State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *Proc. CMV'07*, pages 61–71.

Ropinski, T., Oeltze, S., and Preim, B. (2011). Survey of Glyph-based Visualization Techniques for Spatial Multivariate Medical Data. *Comput. & Graph.*, 35(2):392–401.

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.*, 20:53–65.

Sanderson, A. R., Johnson, C. R., and Kirby, R. M. (2004). Display of Vector Fields Using a Reaction-diffusion Model. *Proc. IEEE Visualization*, pages 115–122.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., and Moorhead, R. (2009). A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets. *IEEE TVCG*, 15(6):1209–1218.

Sanyal, J., Zhang, S., Dyer, J., Mercer, A., Amburn, P., and Moorhead, R. (2010). Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty. *IEEE TVCG*, 16(6):1421–1430.

Scheuermann, G. and Tricoche, X. (2005). Topological Methods for Flow. *The Visualization Handbook*, pages 331–346.

Schneider, D., Fuhrmann, J., Reich, W., and Scheuermann, G. (2012). A Variance Based FTLE-Like Method for Unsteady Uncertain Vector Fields. In *Topological Methods in Data Analysis and Visualization II*, pages 255–268. Springer.

Schultz, T., Schlaffke, L., Schölkopf, B., and Schmidt-Wilcke, T. (2013). HiFiVE: A Hilbert Space Embedding of Fiber Variability Estimates for Uncertainty Modeling and Visualization. *Comput. Graph. Forum*, 32(3):121–130.

Schultz, T., Theisel, H., and Seidel, H.-P. (2007). Topological Visualization of Brain Diffusion MRI Data. *IEEE TVCG*, 13(6):1496–1503.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. VL'96*, pages 336–343.

Shu, Q., Guo, H., Liang, J., Che, L., Liu, J., and Yuan, X. (2016). EnsembleGraph: Interactive Visual Analysis of Spatiotemporal Behaviors in Ensemble Simulation Data. In *Proc. PacificVis'16*, pages 56–63.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Stockinger, K., Shalf, J., Wu, K., and Bethel, E. W. (2005). Query-Driven Visualization of Large Data Sets. *Proc. IEEE Visualization*, 5:167–174.

Theisel, H., Rössl, C., and Weinkauf, T. (2008). Topological Representations of Vector Fields. In *Shape Analysis and Structuring*, pages 215–240. Springer.

Thompson, D., Levine, J., Bennett, J., Bremer, P.-T., Gyulassy, A., Pascucci, V., and Pebay, P. (2011). Analysis of Large-scale Scalar Data Using Hixels. In *Proc. LDAV'11*, pages 23–30.

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., and Pavel, M. (2005). A Typology for Visualizing Uncertainty. In *Electronic Imaging*, pages 146–157.

Tricoche, X., Scheuermann, G., and Hagen, H. (2001). Continuous Topology Simplification of Planar Vector Fields. In *Proc. VIS'01*, pages 159–166.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press.

Tukey, J. W. (1962). The Future of Data Analysis. *Ann. Math. Statist.*, 33(1):1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wang, B., Rosen, P., Skraba, P., Bhatia, H., and Pascucci, V. (2013). Visualizing Robustness of Critical Points for 2D Time-Varying Vector Fields. *Comput. Graph. Forum*, 32(3pt2):221–230.

Wang Baldonado, M. Q., Woodruff, A., and Kuchinsky, A. (2000). Guidelines for Using Multiple Views in Information Visualization. In *Proc. AVI'00*, pages 110–119.

Ward, O. M. (2008). *Multivariate Data Glyphs: Principles and Practice*, pages 179–198. Springer.

Ware, C. (2004). *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, 2nd edition.

Whitaker, R. T., Mirzargar, M., and Kirby, R. M. (2013). Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE TVCG*, 19(12):2713–2722.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Elsevier, 3rd edition.

Wilson, L. J., Burrows, W. R., and Lanzinger, A. (1999). A Strategy for Verification of Weather Element Forecasts from an Ensemble Prediction System. *Mon. Weather Rev.*, 127(6):956–970.

Wittenbrink, C., Pang, A., and Lodha, S. (1996). Glyphs for Visualizing Uncertainty in Vector Fields. *IEEE TVCG*, 2(3):266–279.

Wittenbrink, C. M., Pang, A. T., and Lodha, S. K. (1995). *Verity Visualization: Visual mappings*. Computer Research Laboratory [University of California, Santa Cruz].

Wong, P. C. and Bergeron, R. D. (1997). 30 Years of Multidimensional Multivariate Visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33.

Wu, B., McGrory, C. A., and Pettitt, A. N. (2012). The Variational Bayesian Approach to Fitting Mixture Models to Circular Wave Direction Data. *JAMC*, 51(10):1750–1762.

Wu, K. and Zhang, S. (2013). A Contour Tree Based Visualization for Exploring Data with Uncertainty. *Int. J. Uncertain Quantif.*, 3(3):203–223.

Zehner, B., Watanabe, N., and Kolditz, O. (2010). Visualization of Gridded Scalar Data with Uncertainty in Geosciences. *Comput. Geosci.*, 36(10):1268–1275.

Zuk, T. and Carpendale, S. (2006). Theoretical Analysis of Uncertainty Visualizations. In *Electronic Imaging 2006*, pages 606007–606007.

Zuk, T., Downton, J., Gray, D., Carpendale, S., and Liang, J. (2008). Exploration of Uncertainty in Bidirectional Vector Fields. In *Electronic Imaging 2008*, pages 68090B–68090B.