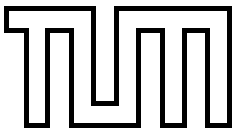

TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR INFORMATIK

LEHRSTUHL FÜR COMPUTER GRAPHIK UND VISUALISIERUNG

Visualization of Uncertain Scalar Data Sets

Tobias Josef Pfaffelmoser

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. H.-J. Bungartz
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. R. Westermann
	2.	Univ.-Prof. Dr.-Ing. H. Theisel, Otto-von-Guericke-Universität Magdeburg

Die Dissertation wurde am 13.06.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 24.07.2013 angenommen.

To my family

Abstract

As scientific computing typically encompasses applications of computer simulation, numerical analysis and parallelized high-performance computing to solve various physical or technical problems in diverse disciplines, adequate techniques are necessary for analyzing and interpreting the usually very large resulting data sets. Therefore, suitable visualization algorithms are commonly used for obtaining the necessary insight with respect to the intrinsic characteristics of such data sets.

As numerical simulations or physical computations often suffer from varying, erroneous or incomplete physical measurement input data or error sources in the used algorithm, the resulting data sets are often neither deterministic, nor trustworthy. Therefore, one has to be aware that the information contained in the data is affected by inherent uncertainty and cannot be regarded as a reliable source for correct interpretations and assumptions. The data is often expected to help answer questions on an initially defined problem and mathematical model formulation. However, this can be done only by taking into consideration both the computed data and the derived uncertainty measures. In order to allow a comprehensive interpretation of the data, uncertainty measures and indicators have to be integrated into adequate visual representations. Due to the high importance of allowing reliable interpretations of data affected by uncertainty, uncertainty visualization is regarded as one of the grand challenges in visual data exploration.

This PhD thesis provides significant contributions to this research direction. While introducing several approaches for mathematically modeling uncertainty for visualization purposes, a variety of algorithms is proposed, allowing a more comprehensive visualization and interpretation of multidimensional uncertain scalar data sets. The thesis presents two novel concepts addressing the problem of visualizing the positional and geometrical variability of prominent features like isocontours and iso-surfaces in 2D and 3D uncertain scalar data sets. Furthermore, a completely new direction in uncertainty visualization is addressed: Novel visual algorithms are presented for revealing relative uncertainties, i.e., possible data variations at different spatial points relative to each other. Inferring such stochastic dependences allows analyzing variations of prominent structures in the data. Two algorithms are presented for visualizing local and global correlations, which serve as indicators for structural uncertainty in Gaussian distributed data sets. In addition, a novel approach is discussed for analyzing the variability of derived higher-order quantities like gradients in uncertain scalar data sets. The advantages of the proposed novel techniques are demonstrated for several synthetic and real-world data sets in geophysical and weather forecast scenarios.

Zusammenfassung

Techniken der numerischen Computersimulation als auch des parallelisierten Hochleistungsrechnens werden oft eingesetzt, um physikalische oder technische Probleme aus verschiedenen Disziplinen zu lösen. Um die dabei generierten sehr großen Datensätze besser interpretieren zu können, sind spezielle Analysemethoden notwendig. Dabei werden oft Visualisierungsalgorithmen verwendet, um einen intuitiven Einblick in die spezifischen Eigenschaften eines Datensatzes zu erlangen.

Numerische Simulationen als auch physikalische Berechnungen basieren oft auf fehlerbehafteten Messungen oder werden durch Fehlerquellen im Rechenprozess beeinflusst. Aus diesem Grund sind die dabei gewonnen Daten oft weder deterministisch noch vertrauenswürdig. Des Weiteren muss man sich bewusst sein, dass die in den Daten enthaltenen Informationen mit Unsicherheiten behaftet sind und nicht als verlässliche Quelle für korrekt abgeleitete Annahmen oder Interpretationen dienen können. Nur die gleichzeitige Betrachtung von abgeleiteten Unsicherheitsgrößen und des eigentlichen Datensatzes erlaubt eine verlässliche Beantwortung der dem Datengenerierungsprozess zugrunde liegenden physikalischen oder mathematischen Fragestellungen. Dabei ist oft eine umfassende Interpretation der Daten nur dann möglich, wenn entsprechende Unsicherheitsgrößen oder Indikatoren in die Visualisierung integriert werden. Aus diesem Grund wird die Unsicherheitsvisualisierung als eine der zentralen Herausforderungen der visuellen Datenanalyse angesehen.

Die vorliegende Dissertation leistet mehrere signifikante Beiträge zum Forschungsgebiet der Unsicherheitsvisualisierung. Neben mehreren mathematischen Ansätzen zur Unsicherheitsmodelierung werden vor allem neue Algorithmen vorgeschlagen, die eine umfassende Darstellung und Interpretation von mehrdimensionalen unsicheren Datensätzen ermöglichen. Die Arbeit präsentiert dabei zwei neue Konzepte zur visuellen Darstellung der positionellen und geometrischen Variabilität von Iso-Konturen und Iso-Flächen in zwei- und drei-dimensionalen Datensätzen. Des Weiteren werden neue Algorithmen zur Analyse von relativen Unsicherheitsgrößen präsentiert, die die visuelle Darstellung der Datenvariabilität an verschiedenen Ortspunkten relativ zueinander zum Ziel haben. Im Rahmen dieser komplett neuen Forschungsrichtung werden auch zwei neue Ansätze zur Visualisierung von lokalen und globalen Korrelationswerten präsentiert, die als verlässlicher Indikator für die strukturelle Unsicherheit in Datensätzen mit zugrundeliegender Gauss-Normalverteilung dienen können. Des Weiteren wird ein neuer Ansatz zur visuellen Darstellung der Variabilität von Gradienten in unsicheren Datensätzen diskutiert. Die Vorteile der präsentierten neuen Methoden werden anhand mehrerer synthetischer als auch realer Daten aus den Bereichen der Geophysik und der numerischen Wettervorhersage erläutert.

Acknowledgments

I gratefully acknowledge the support of all people who made this thesis possible.

First and foremost, I would like to express my sincere gratitude to my adviser Prof. Dr. Rüdiger Westermann for offering me the possibility to pursue research in the highly interesting but still quite young field of uncertainty visualization. I am very grateful for his motivating supervision and his time commitment to my work, paired with a constant abundance of patience. Above all, his openness and keenness to discuss various ideas and opinions on diverse problems have made an essential contribution to the success of this work. I am thankful to all of my current and former colleagues that have always been open for discussions: Dr. Stefan Auer, Dr. Kai Bürger, Shunting Cao, Matthäus Chajdas, Ismail Demir, Dr. Christian Dick, Florian Ferstl, Dr. Roland Fraedrich, Raymund Fülöp, Dr. Joachim Georgii, Andreas Klein, Hans-Georg Menz, Mihaela Mihai, Marc Rautenhaus, Florian Reichl, Dr. Jens Schneider, Marc Treib, Mikael Vaaraniemi and Jun Wu.

I want to thank Prof. Dr. Heiner Igel and Karin Sigloch, PhD from the Department of Earth and Environmental Sciences (LMU München) for introducing me to the field of geophysics, for answering numerous questions and discussing various topics related to visualization applications in geophysical scenarios. I am very grateful for their invitation to various international workshops on inversion methods in the earth sciences. Karin Sigloch introduced me to Prof. Dietmar Müller, PhD from the University of Sydney, where I could spend two months for a research project in 2011. In particular, I want to thank John Cannon from the EarthByte Group in Sydney for the very fruitful cooperation on the development of the GPlates 3D framework. I also want to thank my fellow students Christian Boehm, Simon Stähler and Ran Zhang in the research project on "Efficient Inversion Methods for Parameter Identification in the Earth Sciences" for the fruitful cooperation during the last years.

Furthermore, I want to thank Prof. Malcolm Sambridge, PhD and Thomas Bodin, PhD from the Australian National University in Canberra, as well as Paul Käufl and Ilaria Mosca for providing me with several geophysical data sets. I am thankful to Michael Kern, Mihaela Mihai and Matthias Reitinger for their contribution to the development of several techniques presented in this thesis.

I want to thank the Munich Centre of Advanced Computing and the International Graduate School of Science and Engineering at the TU München for funding my work by a doctoral scholarship.

Finally, I am enormously thankful to my parents Gabriele and Siegfried, my family and friends for giving me all the support I needed during the last years.

Contents

1 Introduction

In his 1962 book on numerical methods [Ham62], Richard Hamming stated his philosophy on scientific computing: "The purpose of computing is insight, not numbers." As scientific computing typically encompasses applications of computer simulation, numerical analysis and parallelized high-performance computing to solve various physical or technical problems in diverse disciplines, adequate techniques are necessary for analyzing and interpreting the usually very large resulting data sets. Therefore, suitable visualization algorithms are commonly used for gaining the necessary insight into the intrinsic characteristics of such data sets.

As numerical simulations or physical computations often suffer from varying, erroneous, or incomplete physical measurement input data or error sources in the used algorithm, the resulting data sets are often neither deterministic, nor trustworthy. Therefore, one has to be aware that the information contained in the data is affected by inherent *uncertainty* and cannot be regarded as a source for reliable interpretations and assumptions. The data is often expected to help answer questions on an initial problem and mathematical model formulation. However, this can be done only by taking into consideration *both* the computed data *and* the derived uncertainty measures. In order to allow a comprehensive interpretation of data, uncertainty measures and indicators have to be integrated into visual representations of the data.

Due to the high importance of a reliable interpretation of the data affected by uncertainty, *uncertainty visualization* is regarded as one of the grand challenges in visual data exploration [Joh04]. The PhD thesis at hand provides significant contributions to this research direction: Besides introducing several approaches for mathematically modeling uncertainty for visualization purposes, a variety of algorithms are proposed, allowing a more comprehensive visualization and interpretation of multi-dimensional uncertain scalar data sets.

In this chapter, the data generation process with possible sources of uncertainty, as well as intrinsic characteristics of uncertain data sets are covered in more detail. The term 'uncertainty' is defined and classified with respect to a common topology. Furthermore, several approaches and possible

problems regarding uncertainty quantification and interpretation are discussed. This chapter also contains an introduction to several application domains, where uncertainty analysis and its visualization play a significant role. Finally, the contributions of this thesis are set into the context of the current main challenges in uncertainty visualization.

1.1 Uncertain Data Sets

This section provides an overview on different types of numerical concepts and algorithms will be provided, which typically produce data sets affected by inherent uncertainty. Furthermore, possible uncertainty sources are discussed and the term 'uncertainty' is defined in this context.

Uncertain data sets often result from *numerical simulations*. Thereby, the term 'simulation' is used according to its definition as the imitation of an operation of a real-world process or system over time [BCN*05]. Numerical simulations are commonly used for modeling complex natural systems in engineering, chemistry, physics and biology. Although such systems can often be fully described using adequate mathematical models, a closed-form analytic solution cannot be easily derived in most cases. In such cases, numerical simulations substitute analytic solutions in order to predict the behavior of the modeled (physical) system based on an initially given set of parameters, observations or other conditions. The simulation problem of predicting the result of measurements based on a complete description of a parameterized (physical) system is also known as the *forward problem* [Tar05].

Typical applications for numerical simulations can be found in the domain of *numerical weather predictions*. Such algorithms simulate the dynamic behavior of weather conditions (temperature distribution, wind velocities, air pressure, etc.) based on complex mathematical models of the earth's atmosphere and oceans. Furthermore, current weather conditions are used as the initial setup for the simulation process. This data can be obtained from a variety of different sources, like terrain maps, radiosondes in weather balloons, as well as weather satellites [Ste07]. In addition to measured initial weather conditions and mathematical models, certain meteorological processes are represented by adequate parameterized models. The *parameterization* is used for physical phenomena that are either too small-scale or too complex to be explicitly included in the mathematical weather prediction model. For instance, the modeling of certain cloud formations on a local scale can be done using appropriate parameter setups [NO07]. Based on the mathematical prediction model, initial weather conditions and possible parameter settings, numerical weather prediction simulations are able to compute a weather scenario for any future point in time.

Numerical forward simulations also have a strong impact on geophysical computations. *Seismic wave simulations* model the propagation of seismic pressure and shear waves through the earth's mantle. Thereby, certain parameters characterizing the physical properties of geological subsurface

structures (e.g., elasticity and material density parameters) are used as initial conditions for the simulation run. In addition, the location and orientation of a wave source (e.g., earthquake epicenter) has to be defined and included in the propagation model. Based on this setup, the (global) wave propagation over time is computed using, for instance, finite differences or spectral element methods [Gra96, KT99, KRT02].

Numerical simulations in weather prediction and seismic wave propagation are both based on the setup of initial physical conditions and an eventual parameter setting. Therefore, simulation results can be highly dependent on certain parameters and can also reveal a variation in spatial sensitivity with respect to the initial conditions.

Another group of data sets, which are affected by uncertainty, result from *inverse problems*. Inverse problems represent the complement of the simulation forward computations. For a given measurable physical (dynamic) observation, an inverse problems asks for the parameters causing the observed phenomenon. In general, an inverse problem framework converts the observed measurements to obtain parameters characterizing the underlying (physical) object or system. A typical example for inverse problems can be found in *seismic tomography*. For a given observation of seismic waves (recorded seismograms), originating from earthquake epicenters, an inverse problem computes the parameters describing the material characteristics of subsurface structures. Seismic waves are traveling through these structures from the epicenter to the receivers on the earth's surface. Compared to the seismic wave simulation discussed above, this concept follows the exact opposite or inverse direction, as parameters are derived from the wave propagation. A detailed introduction into the inverse theory and its application in seismic tomography can be found in [Tar05]. The forward or simulation problem usually has a unique solution in deterministic physics. However, for inverse problems, the solution is often not unique and highly dependent on the amount and quality of the measured and observed input data. Therefore, inverse problems are often modeled and resolved using probabilistic theory and represent accordingly a class of data generation processes where uncertainty plays a significant role during the computation and the interpretation phase.

1.1.1 Sources of Uncertainty

The previous section outlined that most data generation processes in simulation and inverse computation rely on the initial setting of certain observed physical conditions or specific parameters modeling critical parts of the computation process. As this information is often not deterministic and affected by diverse types of errors, the resulting data of the computation process is also not reliable and should not be used for any type of analysis or interpretation without taking into account the inherent uncertainties. This subsection will shed light onto different sources that can introduce uncertainty in the already discussed types of data generation processes. In [PWL97, Pan08] sources of uncertainty are classified in three groups: *acquisition*, *transformation* and *visualization*.

The data *acquisition* process covers the generation of initial input data for the respective simulation or inverse computation algorithm. For instance, in numerical weather prediction, the initial weather conditions are obtained by physical measurements (e.g., temperature, air pressure, etc.) using different kinds of sensors in the atmosphere. In seismic tomography, the acquisition process covers the recording of seismic wave amplitudes and frequencies using seismometers all over the world. In all acquisition scenarios, uncertainty can be introduced due to mis-calibrated instruments and erroneous measurements. Furthermore, background noise can significantly influence the correct recording of physical data. Apart from defective measuring instruments, a non-sufficient data coverage can introduce a considerable amount of uncertainty in the acquisition stage. For instance, if only a small number of seismic data is available for a certain part of the earth, e.g., due to a small number of recorded earthquakes, the physical properties of subsurface structure cannot be reliably resolved for the respective region, due to the ill-posed nature of the inverse problem. Note that quantifying the uncertainty introduced by missing or incompletely input data is often even more challenging than modeling the error range of instruments or the environmental background noise corrupting the data during the measuring process.

After the acquisition stage, the obtained data is further processed according to the respective simulation or inverse algorithm. During this *transformation* phase further uncertainties can be introduced. The loss of reliability is often caused by the computational accuracy or different filtering techniques applied to the original data, such as down-sampling, re-scaling, quantization, condition effects or interpolation. The introduced amount and type of uncertainty is highly dependent on the overall algorithm design. As the transformation of the original data follows strict deterministic rules, the introduced error can be quantified more easily than during the acquisition stage. Therefore, this type of error is also called *systematic uncertainty* [Pan08].

Both the acquisition and transformation stage introduce uncertainties, which directly affect the resulting data of the respective algorithm. For analyzing and interpreting the results, a final *visualization* stage is necessary. Although the visualization is generally meant to reveal hidden information in often rather abstract data sets, it is important to note that applied visual algorithms can even introduce uncertainties themselves. They often result from approximations (e.g., radiosity algorithms in 3D illumination) or filtering stages like interpolation. Apart from errors at the data level, the use of non-adequate rendering techniques (e.g., direct volume rendering vs. iso-surface extraction in 3D) can also lead to false assumptions or misinterpretations.

Although all three stages can introduce different kinds of uncertainty, which can negatively influence the interpretation of the data, the transformation and visualization uncertainty are usually negligible compared to the uncertainty introduced during the acquisition process. Therefore, only this source is considered in the literature regarding uncertainty visualization. This thesis will not differentiate between uncertainty sources and only consider the acquisition process for developing adequate visual uncertainty representation techniques.

1.1.2 Typology of Uncertainty

Although the terms 'uncertainty' and 'error' have been used interchangeable throughout this thesis, a clear definition is still missing. However, in general, the term 'uncertainty' is affected by a highly ambiguous terminology and compasses different concepts, which are used for expressing certain characteristics in different contexts [GS06]: *Error* expresses the deviation from a true values, *imprecision* indicates the resolution of a value compared to the needed resolution, *accuracy* models the size of an interval a value lies in, *lineage* is used for indicating if the data is provided at first or second hand, *subjectivity* expresses the degree of subjective influence in the data, *non-specificity* can be regarded as a lack of distinctions for objects and *noise* indicates undesired background influence.

Furthermore, the National Institute of Standards and Technology (NIST) distinguishes four different concepts of uncertainty [TK94, PWL97]: *Statistical* uncertainty is represented by a mean and standard deviation value, *error* indicates the difference between a correct true value and an estimate, a *range* is used for expressing an interval in which the data must exist and *scientific judgment* based on all relevant known general knowledge. Other definitions include reliability, ignorance, clearness, distinctivenes and many more. A more detailed overview on the typology of uncertainty can be found in [THM*05].

All definitions and classifications of 'uncertainty' have in common that they always express some degree of imperfection of the user's knowledge about the resulting data, the initial conditions or the numerical process. In this thesis, the term 'uncertainty' is differentiated from 'error' following the definitions given in [HG93, MRH*05, DK97]: The deviation from a known true value is called *error*. In this case, the inaccuracy is known objectively. However, *uncertainty* covers a broader range of inaccuracy than error, as it refers to knowledge of possible deviations from the true value, but without knowing either the exact value or the magnitude of the deviations. Therefore, the concept of uncertainty does not allow any assumption on the exact true value. However, uncertainty quantification techniques can determine intervals or distributions that provide hints on the unknown exact value.

In addition to the more general notion of uncertainty, the term of *probabilistic resolution* is introduced and used in this thesis several times. It is closely related to the idea of resolution of data resulting from inverse problem solutions [Tar05]. Probabilistic resolution refers to the appearance of certain pre-defined events in the data under uncertainty. A *feature* in the data (e.g., an isocontour, a critical point, a sharp transition, etc.) will be called *well-resolved*, if its appearance can be guaranteed with a certain credibility or confidence based on the given uncertainty. Thereby, a stochastic or statistic model of the underlying uncertainty is often necessary. Where uncertainty describes the lack of knowledge at the data level, probabilistic resolution can be used to assess the appearance of certain prominent features or structures in the data according to their reliability

under uncertainty. This concept can directly influence the analysis and interpretation process of the data by the respective user.

1.2 Uncertainty Analysis

The consideration of possible uncertainties in the interpretation process of data sets can help to draw more reliable conclusions or even increase the degree of insight into the contained information. This section will provide an overview on uncertainty quantification techniques and will present prominent application areas where a profound uncertainty analysis can be of fundamental importance.

Scientific data sets, as discussed in section 1.1, serve typically one of two main purposes. The first motivation for the data generation process could be the desire to gain new insights into a pre-defined more abstract problem or a concrete physical phenomenon. In this case, the main purpose of the data generation process is a desired *increase in knowledge* or the validation of a pre-defined hypotheses. For uncertain data sets, the ignorance of the underlying uncertainty could result in false assumptions or misinterpretations. This would result in an unreliable increase of knowledge. A typical example would be the already discussed seismic tomography: the ignorance of uncertainty data could result in false assumptions on the material and structural characteristics of the earth's interior. Wrong interpretations would also influence other research areas. This can only be avoided by using a fundamental uncertainty analysis. As this examination would mainly assess if the generated knowledge is reliable, it will be called *reliability analysis*.

Apart from the simple generation of new knowledge, data sets can be also used as a basis for *decision making processes*. Complex problems in diverse disciplines often demand decision making under conditions of uncertainty, lack of knowledge and ignorance by both engineers and scientists [Ayy06]. As decisions based on uncertain information can have far more severe implications than the simple generation of unreliable knowledge, an extensive uncertainty analysis is even more essential. The numerical weather prediction, as well as the seimsic wave propagation can serve as examples. Surface wave simulations are often run to examine the impact of earthquakes on certain areas on the earth's surface in order to identify potential high-risk regions. If certain actions are derived from this information, their potential inherent risk is directly linked with the underlying uncertainty of the data. Therefore, the uncertainty analysis for data sets, used for decision making processes, will be classified as *risk analysis*.

1.2.1 Uncertainty Quantification

Understanding both the reliability of data and making sound decisions require the knowledge of which parts of the data sets are affected by uncertainty and adequate quantitative computations thereof [Ger98]. This can be achieved using suitable *uncertainty quantification* techniques.

For simulation processes, the question to be approached is how uncertainties in the input data (initial conditions, parameters, etc.) are propagated by the simulation algorithm to the final result. If a complete mathematical description of the simulation process is available and if sensitivity measures can be derived for all input variables without large computational efforts, the impact of variability in the input on the output can be determined analytically. For this scenario, uncertainty quantification is directly related to the area of sensitivity analysis, which focuses primarily on the relation between input and output data [AH07]. However, in most scenarios, the overall problem and the used algorithm are far too complex for an analytical sensitivity analysis. Then, the effect of variations in the input on the output data has to be quantified by performing several simulation runs. A typical example for uncertainty quantification in simulation processes is the generation of *ensemble data sets*. It is assumed that the initial conditions or the parameter setting are affected by uncertainty. This is typically the case for e.g., weather prediction simulations. The uncertain initial setup is considered by performing multiple simulation runs for slightly changed initial parameters or conditions. The resulting collection of simulated data sets is called *ensemble*. The local or global variation of data present in the ensemble is directly related to the parameter uncertainty and the sensitivity.

For inverse problems, the quantification of uncertainties via direct computations is far more complex and mostly only possible for linear scenarios [FWA*11, ABT12, Tar05, ZT07]. For non-linear or large-scale problems (e.g., in seismic tomography), *Bayesian frameworks* and *Markov Chain Monte Carlo* techniques (MCMC) are used for both the inverse computation and uncertainty quantification [MS02, ABT12, GFWG10, GCN*09, MT02, Tar05]. Opposite to simulation runs, inverse problems ask for the initial parameter setting, which generates an observed data set (cf. 1.1). MCMC techniques sample the parameter space, generate forward solutions and compare this data to the given observation using a fitting function. They can generate an ensemble of parameter sets that solve the inverse problem equally and result in the same final fitting error (cf. section on Monte Carlo sampling in [Tar05]). The resulting ensemble data set and possible inherent data variations directly represent the underlying uncertainty and can be further processed using either statistical processes or stochastic models. Note that the number of ensemble members needed for a reliable uncertainty quantification increases exponentially with the total number of parameters.

Apart from sensitivity analysis and Monte Carlo techniques, there are several other methods available for quantifying uncertainties in simulation or inverse processes. For instance, insight in uncertain data structures can be obtained by comparing the results of the data generation process with former results or common knowledge on the expected resulting information [TK94]. Further approaches could include statistical hypothesis tests, which also try to ask and answer certain questions on the data [SBSS06, PKRJ10] and interactive techniques where the user can make the uncertainty quantification dependent on, e.g., spatial or temporal conditions. Common approaches use concepts of information theory for quantitatively identifying uncertainties in data sets. Typically, *information entropy* is used for showing and analyzing uncertain structures in the data

[WRL12, XLS10].

Although most uncertainty quantification techniques aim at analyzing the propagation process between input and output data, the impact of the algorithm itself must not be ignored. In most inverse computation algorithms, specific regularization techniques are applied in order 'smooth out' regions with low data coverage and to stabilize the overall numerical process [Tar05]. However, regularization effects can have a strong impact on the propagation by suppressing the input uncertainty and can lead to misinterpretations. Therefore, additional knowledge on the algorithm design and possible side effects can significantly improve the overall uncertainty quantification and interpretation process. Detailed information on *regularization effects* and prior vs. posterior uncertainty concepts can be found in [FWA*11, Tar05].

1.2.2 Application Domains

Uncertainty analysis is a crucial task in many application areas, where an adequate risk or reliability assessment is necessary. In this subsection, a small selection of applications is presented.

This thesis will focus on the already introduced main application areas of uncertainties in weather prediction and seismic tomography (cf. [SBSS06]) scenarios. Throughout the following chapters, specific characteristics, interpretations, as well as challenges with respect to uncertainty quantification in these areas will be discussed.

In general, uncertain information arises typically in environmental sciences and especially in scenarios involving natural hazards like wind, flood or earthquakes [Ayy06, Pan08, WFR*10]. Such scenarios often involve business or political decision-making processes based on an extensive risk assessment [BAF08]. Furthermore, uncertain data can be found in other geo-statistical or cartography and GIS domains, often based on vision operations or image processing [BFW02, Kyr03, MRH*05].

Further areas can comprise medical applications (e.g., MR spectroscopy [FKLTI10] or guided volume segmentation [PRH10]), architectural reconstructions [GS05], air quality studies [PdJB07], astrophysics [LFLH07] and many more.

1.3 Challenges for Uncertainty Visualization

For making a comprehensive analysis and reliable interpretations of uncertain data sets possible, adequate visualization techniques are necessary in order to communicate the relevant information to data analysts or decision makers. Uncertainty visualization is still regarded as one of the grand challenges in visual computing [Joh04], as most published techniques do not address a comprehensive communication of uncertainty characteristics on a satisfactory level. Furthermore, the existing

literature is inconsistent on the best approaches and concepts for specific problems in uncertainty visualization [BAF08]. In the following, an overview of grand challenges in uncertainty visualization is provided.

Reliable visualizations of uncertain data sets require suitable mathematical models for providing necessary parameter values or probabilities as input for the rendering stage. However, most current visualization techniques do not include sophisticated stochastic or statistic formulations of the underlying uncertainty in scalar data sets. They often limit themselves to abstract indicators or heuristic models without profound mathematical models. Using a correct and clear formulation of uncertainty as basis for further rendering techniques does not only allow a more intuitive and comprehensive interpretation of the generated visual uncertainty effects, but could even offer new insight related to specific uncertainty structures (e.g., stochastic dependences).

Uncertainties in data sets are often modeled via specific probability distributions (e.g., Gaussian distribution), which can either be used directly for visualization purposes or serve as a basis for the derivation of certain uncertainty parameters (e.g., standard deviation). In this context, interpretations of resulting visualizations would highly benefit from further research on the impact of different distribution assumptions on the visual result. Furthermore, techniques have to be developed for both cases where uncertainty is represented as distribution or by derived parameters, as both approaches have different pros and cons in different contexts. Reliable and comprehensive uncertainty visualization frameworks have to take this into consideration.

Typically, uncertainty is given, or can be computed, as additional auxiliary information to the actual resulting data of a simulation or inverse process. In order to allow a comprehensive interpretation, in most cases both the actual data and the uncertainty information have to be integrated in one visual representation. This is a challenging task due to a limited number of visualization channels (e.g., color, opacity, glyphs, etc.) or due to severe cluttering effects or artifacts. Uncertainty information has to be linked to the actual data without overloading the final visual representation. This challenge of a context-aware visualization could also include the correct uncertainty representation with respect to further entities like spatial position and time.

Uncertainty information is typically integrated in visual representations at a data level. For instance, parameters like standard deviation are mapped to color at certain spatial positions. However, in many approaches, data sets are visualized by the extraction of specific features. For instance, 3D volume data can be visualized by an intermediate iso-surface representation. Visualizing uncertainty on a data level would not reveal any information on possible variations (e.g., in position) of the iso-surface. Therefore, specific transformation models and algorithms are necessary in order to derive uncertainty measures for multi-dimensional data features (e.g., iso-surfaces, critical points, vortices, etc.) from the given data uncertainty model.

Current approaches only visualize uncertainties with respect to the actual data values at certain

spatial positions of a data set. However, the effect of uncertainty on differential quantities like gradients, Hessian data, curvature, etc. is neglected. In order to analyze the variability of certain structures in the data or the geometric shape of features (e.g., surfaces), uncertainty visualization of higher-order data entities would be necessary. Furthermore, auxiliary uncertainty parameters (e.g., covariance, correlation, etc.) could help to analyze stochastic dependences or structural instabilities in the data with respect to its variability. For integrating this kind of information, new models and techniques for both the uncertainty quantification, representation and rendering stage would be necessary. This has not been addressed for scalar data sets in common uncertainty visualization approaches.

The integration of uncertainty information in visualization algorithms could significantly increase the overall data amount. Furthermore, additional visual effects could slow down the rendering performance. Therefore, an efficient implementation (e.g., based on GPU computations), as well as suitable data models, possible approximations and storage algorithms are necessary in order to allow an interactive and comprehensive exploration of uncertain data sets.

1.4 Contribution

This thesis focuses on uncertain 2D and 3D scalar data sets, originating from simulation processes and inverse problems. By introducing novel techniques for addressing several aforementioned challenges, the following chapters make a significant contribution to an improved and more intuitive visualization and understanding of uncertain information. Furthermore, the presented techniques address the main visualization challenges mentioned before.

In order to provide more accurate and reliable visualizations, all presented algorithms are based on more sophisticated stochastic or statistic models. The extended integration of suitable mathematical formulations does allow a more intuitive understanding of the presented visual effects and provide additional insights with respect to the possible impacts of uncertainty on the actual data.

This thesis presents two novel concepts addressing the problem of visualizing the positional and geometrical variability of iso-contours and iso-surfaces in 2D and 3D uncertain data sets. The first framework introduces mathematical concepts for translating data- to positional iso-surface uncertainty and suggests an incremental update scheme for efficiently integrating uncertainty information into front-to-back volume ray-casting. Thereby, a Gaussian probability distribution is assumed for modeling the underlying uncertainty. Although the positional variability of iso-surfaces has already been covered in the literature, the presented framework considers correlation structures and therefore stochastic dependences in the data and accounts both for homogenous and anisotropic correlation. The use of this auxiliary data results in a more accurate computation of the positional

probability distribution of iso-surfaces. Furthermore, the update scheme and the correlation integration with the ray-casting process makes the visualization results independent of the ray-sampling rate and allows for highly interactive frame rates.

The second novel algorithm focuses on the positional uncertainty of iso-contours in 2D ensemble data sets. The presented approach introduces a statistical framework for deriving spatial probability density and distribution functions for arbitrary non-parametric iso-contour distributions in uncertain 2D data sets. Furthermore, the technique is able to analyze the modality of positional contour distributions and allows assumptions on the iso-contour uncertainty topology in the respective ensemble data sets. Although many uncertainty visualization approaches rely on the more simple assumption of underlying parameterized Gaussian probability distributions, the presented framework can be regraded as a first step into the direction of distribution-independent non-parametric uncertainty visualizations.

Apart from uncertain iso-surfaces and iso-contours, this thesis address a complete new direction in uncertainty visualization. Typical approaches take only local uncertainty measures (e.g., standard deviation) into consideration, which are related to one single entity, like a value at a discrete spatial position in the data set. This thesis presents novel visual algorithms for revealing relative uncertainties, i.e., possible data variations at different points relative to each other. Inferring such stochastic dependences allows to analyze variations of certain prominent geometric structures (e.g., curvature, orientation, etc. of surfaces/lines). On the one hand, two algorithms are presented for visualizing local and global correlations, which serve as indicators for structural uncertainty in Gaussian distributed data sets. Furthermore, a comprehensive introduction to correlation structures and their importance for reliable uncertainty visualization techniques is provided. On the other hand, a novel approach is discussed for analyzing the variability of higher-order quantities like gradients in scalar data sets. As only uncertainty at the data level for single values has been covered by many algorithms so far, the presented visualization framework introduces both suitable mathematical models for deriving the necessary quantities and rendering techniques for displaying variations in gradient magnitude and orientation for the first time. The resulting visualizations allow a better assessment of the stability of certain features (e.g., contours, ridges, valleys, etc.) in the uncertain data sets.

1.5 Outline

In the following chapter 2, diverse mathematical concepts are presented for modeling uncertain data sets stochastically and for obtaining necessary parameters and indicators, which serve as data basis for further uncertainty visualization algorithms. Chapter 3 provides an overview on a selection of the most prominent methods and introduces the main concepts and principles with regard to uncertainty visualization. Chapter 4 introduces several novel approaches for visualizing

the positional and geometrical uncertainty of iso-contours and iso-surfaces in 2D and 3D uncertain scalar data sets. In chapter 5, several novel techniques for visualizing the structural uncertainty of specific features or data distributions in 2D and 3D scalar data fields are proposed. The presented approaches make a significant contribution to the field of uncertainty visualization, as they clearly show how the structural uncertainty analysis can reveal information about uncertain data sets that cannot be obtained by traditional methods. Chapter 6 summarizes the main contribution presented in this thesis and provides an overview on interesting directions for future research.

1.6 List of Publications

The research results presented in this thesis have been originally published in the following peer-reviewed conference papers and journal articles (listed in chronological order):

1. T. Pfaffelmoser, M. Reitinger, and R. Westermann, *Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields*, Computer Graphics Forum (Proceedings of EuroVis 2011), 2011, pp. 951–960 [PRW11]
2. T. Pfaffelmoser and R. Westermann, *Visualization of Global Correlation Structures in Uncertain 2D Scalar Fields*, Computer Graphics Forum (Proceedings of EuroVis 2012), 2012, pp. 1025–1034 [PW12]
3. T. Pfaffelmoser and R. Westermann, *Correlation Visualization for Structural Uncertainty Analysis*, International Journal for Uncertainty Quantification, 2013, vol. 3, no. 2, [PW13a]
4. T. Pfaffelmoser and R. Westermann, *Visualizing Contour Distributions in 2D Ensemble Data*, EuroVis Short Papers, 2013, pp. 55–59 [PW13b]
5. T. Pfaffelmoser, M. Mihai, and R. Westermann, *Visualizing the Variability of Gradients in Uncertain 2D Scalar Fields*, IEEE Transactions on Visualization and Computer Graphics, 2013 [PMW13]

2

Mathematical Foundations

For achieving reliable visualization result, a thorough mathematical modeling of inherent uncertainties in data sets is essential. As stated above, suitable stochastic or statistic concepts can both support the interpretation process and reveal novel information with respect to uncertainty structures. In chapter 1 the term 'uncertainty' with respect to scalar data values was defined as deviation of data samples from a true or assumed value without exact knowledge of its magnitude. In the following, diverse mathematical concepts are presented for modeling uncertain data sets stochastically and obtaining necessary parameters and indicators, which serve as data basis for further uncertainty visualization algorithms.

In the following, it is assumed that uncertain scalar data sets are given as discrete sampling of mapping from a continuous, bounded spatial domain $\mathbb{S} \subset \mathbb{R}^3$ into $\mathbb{R}$. The *sampled domain* is represented by a finite set of n *spatial points*

$$\mathbb{S}_n := \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{S}, i \in \{1, 2, ..., n\}\} \subset \mathbb{S}. \tag{2.1}$$

Note, that for defining an uncertainty model no explicit assumption on the sampling structure of the domain and the position of the sampling points $\mathbf{x}_i$ is needed. For instance, it could represent either regular or irregular grid types. Throughout this thesis, the discrete sampling $\mathbb{S}_n$ is mainly identified with a regular Cartesian grid, but clarified in detail in the respective chapters.

The scalar output data of different generation processes (e.g., simulation, inverse computation, etc.) can be modeled as *data function*

$$y \ : \ \mathbb{S}_n \longrightarrow \mathbb{R}, \ \mathbf{x} \longmapsto y(\mathbf{x}). \tag{2.2}$$

For each discrete spatial point $\mathbf{x} \in \mathbb{S}_n$, one data value $y(\mathbf{x})$ is given. Simulation runs can Monte Carlo methods can produce data ensembles which represent a certain variation in the output values.

An ensemble $\mathbb{E}_m$ with m *ensemble members* is defined as set of data functions

$$\mathbb{E}_m := \{y_i \mid y_i : \mathbb{S}_n \longrightarrow \mathbb{R}, i \ \in \{1, 2, ..., m\}\}. \tag{2.3}$$

Each ensemble member y_i represents one mapping from the discrete domain points in $\mathbb{S}_n$ to real scalar values.

2.1 Uncertainty Modeling

For uncertain data sets, the mapping from the domain points in $\mathbb{S}_n$ to real numbers is non-deterministic. Therefore, it is not possible to assign to each spatial domain point one fixed scalar value, as the exact or true value is unknown. This section covers concepts for modeling this uncertainty at different spatial points.

2.1.1 Random Variables

The uncertainty in assignment of scalar values to spatial points can be modeled by a *random function*

$$Y \ : \ \mathbb{S}_n \longrightarrow \mathbb{R}, \ \mathbf{x} \longmapsto Y(\mathbf{x}), \tag{2.4}$$

where for each sampling point $\mathbf{x}$ the mapping $Y(\mathbf{x})$ is considered as *random variable*. A random variable represents a variable whose possible values are non-deterministic and subject to variation due to uncertainty. Conceptually, it can take on different possible values, which can be assigned to specific probability values. Note, that random variables are originally defined as mappings from abstract *stochastic events* to real numbers. However, in the context of numerical data sets, the stochastic 'event' at each spatial domain point is the assignment of a real value to this point. Therefore, no differentiation is made between stochastic events and potentially assigned real values and consequently the random variables directly represent the uncertainty of the data values.

Furthermore, the random variables may also represent the outcome variability of an uncertain process. For instance, in this context, the random variable $Y(\mathbf{x})$ can serve as representative of a possible data ensemble $y_1(\mathbf{x}), y_2(\mathbf{x}), ..., y_m(\mathbf{x})$. Note, that random functions can be treated the same way as deterministic functions. It is possible to derive higher-order quantities (e.g., gradients) or to apply further transformations. However, the respective results are also non-deterministic and therefore random variables.

Note, that random variables serve only as abstract model for the presence of uncertainty. They do not refer to neither any information on the true value nor to any kind of prescribed distribution function. Furthermore, a random function, defined on a discrete set $\mathbb{S}_n$ of spatial points, can be interpreted as *random vector* $\mathbf{Y}$ with m components $Y_i := Y(\mathbf{x}_i)$. This random vector is also called

multi-variate random variable, where 'multi-variate' expresses the aggregation of several single random variables as components in one vector.

2.1.2 Probability Space

The introduced random variable concepts indicate that the scalar values at the spatial points are uncertain and non-deterministic. As next step, a mathematical measure has to be introduced to assess the probability of possible outcomes of a (multi-variate) random variable. For instance, for numerical uncertain scalar data set with modeled uncertainty at the spatial domain points by random variables, a suitable measure would indicate the occurrence probability of certain value ranges among the real numbers for respective random variables.

In the following, the mathematical construct of a *probability space* is introduced, which quantitatively models a real-world process consisting of states affected by uncertainty. In the following, the concept is applied to a random variable modeling real values at a specific spatial point affect by uncertainty. The probability space is often also called probability triple (Ω,$\mathcal{F}$,$\mathcal{P}$), consisting of three parts [KMBR56]:

The *sample space* Ω is a set containing all possible outcomes of the modeled process. For a random variable representing possible real scalar values, the setting $\Omega = \mathbb{R}$ would hold true.

The *σ-algebra* $\mathcal{F}$ is a set containing all possible measurable *stochastic events*, which might be considered for measuring their respective probability. An event is a subset of the sample space Ω. If a real-valued random variable is measured, the σ-algebra contains, for instance, all possible open and closed intervals in $\mathbb{R}$. Note, that probability values according to chosen measure can be only assigned to elements in $\mathcal{F}$. For instance, by selecting an interval $[a, b] \in \mathcal{F}$ one could computes a probability value that the random variable assumes a real value in this interval.

The probability measure $\mathcal{P}$ is function returning a probability value for a particular event. The probability value is a real number between zero and one:

$$\mathcal{P} \ : \ \mathcal{F} \longrightarrow [0, 1], \ A \longmapsto \mathcal{P}(A) \tag{2.5}$$

A probability measure must satisfy the two requirements $\mathcal{P}(\Omega) = 1$ and that the probability of the union of several countable disjoint events in $\mathcal{F}$ is equal to the sum of the probabilities of each of the respective events:

$$\mathcal{P}\left(\bigcup_i A_i\right) = \sum_i \mathcal{P}(A_i), \quad A_i \in \mathcal{F}, \quad A_i \cap A_j = \emptyset, \quad i \neq j \tag{2.6}$$

Thus, it is guaranteed that the definite event, that the value of the random variable is contained in the sample space, is equal to 1 and the probability measure is countably additive.

Note, that not necessarily every subset of Ω can be considered as measurable event in $\mathcal{F}$. For instance, for $\Omega = \mathbb{R}$, the event that a random variable assumes a irrational number is not measurable and the respective probability would be zero. In general, only events with a non-zero *Borel measure* (cf. e.g., [Wei74]) can receive a non-zero probability value. For instance, the typical Borel measure μ for the real line $\Omega = \mathbb{R}$ would be $\mu([a,b]) = b - a$. Therefore, for the event that the random variable Y assumes a particular value x in $\mathbb{R}$, $\mathcal{P}(Y = x) = 0$ holds always true, as $\mu(x) = 0$. Note, that the Borel measure coincides with the Lebesgue measure on every measurable set.

2.2 Probability Distributions

As this thesis focuses on uncertain numerical data sets, resulting from simulation or inverse process, it can be assumed that the resulting data values at the given spatial points are real numbers. Without loss of generality, it is assumed that the stochastic sample space for a data set, given on a discrete spatial domain $\mathbb{S}_n$, can be written as $\Omega = \mathbb{R}^n$ for a multi-variate random vector $\mathbf{Y} = (Y(\mathbf{x}_1), Y(\mathbf{x}_2), ..., Y(\mathbf{x}_n))^\top$. Thus, each component $Y(\mathbf{x}_i)$ has its own sample space $\Omega_i = \mathbb{R}$. Therefore, in the following, continuous probability distribution functions are introduced for characterizing probability measures for stochastic events related to the real-valued random variables.

2.2.1 Distribution Functions

In general, a *probability distribution* $\mathcal{P}$ assigns to each Borel measurable event in $\mathcal{F}$ a probability value according to Equ. (2.5). For a random variable with $\Omega = \mathbb{R}$, the σ-algebra contains all open and closed intervals as well as disjoint unions of them on the real line. As the probability measure is countably additive by definition, a probability value can be derived for every element in $\mathcal{F}$ by only defining probabilities for all open intervals $]\infty, a]$. For instance, the probability for an arbitrary closed interval $[a, b]$ in $\mathbb{R}$ can be computed as

$$\mathcal{P}(Y \in [a,b]) = \mathcal{P}(Y \in]\infty, b]) - \mathcal{P}(Y \in]\infty, a]) \tag{2.7}$$

The probability distribution of the random variable Y is completely characterized by the definition of the *cumulative distribution function* (CDF)

$$F_Y(y) := \mathcal{P}(Y \leq y), \quad y \in \Omega = \mathbb{R}. \tag{2.8}$$

The cumulative distribution function expresses the probability that a real-valued random variable Y assumes a value less than or equal to y.

Although it is possible to characterize the probability distribution for Y by defining a suitable cumulative distribution function, typically the *probability density function* (PDF) f_Y is utilized for

this purpose. The relation between the CDF and PDF is given by

$$F_Y(y) = \int_{-\infty}^{y} f_Y(u)\,\mathrm{d}u. \tag{2.9}$$

The probability for the random variable Y to fall within a certain range in $\mathbb{R}$ can be computed by integrating the PDF over the respective range. In general, the density function describes the *relative likelihood* for the random variable to take on a certain value in $\mathbb{R}$. Note, that the single likelihood values can be only used for a relative comparison of the probability weights or densities between certain regions in the sample space. However, they do not express the probability that Y takes on a specific single value in $\mathbb{R}$. For this probability the equation $\mathcal{P}(Y = y) = 0$ always holds true as Borel or Lebesgue measure of the interval $[y, y]$ in $\mathbb{R}$ is always zero.

Note, that probability density functions are typically utilized as initial characterization for the probability distribution of a random variable. The cumulative distribution function is than derived by using Equ. (2.9). As PDF every positive and integrable function can be used with the only requirement

$$\int_{-\infty}^{\infty} f_Y(u)\,\mathrm{d}u = 1. \tag{2.10}$$

So far, only cumulative distribution and density function have been introduced for the 1D sample space $\omega = \mathbb{R}$. However, similar definitions exists for multi-variate random variable $\mathbf{Y}$. For a given n-dimensional PDF over the domain $\Omega = \mathbb{R}^n$, the respective cumulative distribution function is obtained as

$$F_{\mathbf{Y}}(\mathbf{y}) = \mathcal{P}(\mathbf{Y} \leq \mathbf{y}) = \mathcal{P}(Y_1 \leq y_1, Y_2 \leq y_2, ...) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} ... f_{\mathbf{Y}}(u_1, u_2, ...)\,\mathrm{d}u_1\,\mathrm{d}u_2 ... \ . \tag{2.11}$$

Note, that by integrating over all except one components of $\mathbf{Y}$, one obtains again a probability density function

$$f_{Y_i}(y_i) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{\mathbf{Y}}(u_1, ..., u_{i-1}, y_i, u_{i+1}, ..., u_n)\,\mathrm{d}u_1 ... \,\mathrm{d}u_{i-1}\,\mathrm{d}u_{i+1} ... \,\mathrm{d}u_n. \tag{2.12}$$

This PDF is called *marginal probability density* and models the probability distribution of the random variable $Y_i = Y(\mathbf{x}_i)$ at the single spatial domain $\mathbf{x}_i$.

2.2.2 Characteristic Values

The probability density function over the domain $\Omega = \mathbb{R}$ provides the full information on the probability distribution of a real-valued random variable Y. However, as in many cases handling the full function is not feasible, certain critical values are utilized for describing the main characteristics of the probability density:

The probability density function returns *likelihood* values. For a continuous sample space, they are not equal to probability values due to the zero Lebesgue measure of single real values (cf. 2.2.1).

The *mode* represents the value in Ω where the PDF reaches its global maximum. It is often also called *maximum likelihood* value. Note, that the mode indicates the region in the sample space where the most likely values are position but, however, is not necessarily equal to the stochastic expected value.

The *modality* of a density function counts its number of local maximums. For instance, an uni-modal distribution has one maximum and one major peak, a bi-modal desnity function two main peaks and so on.

The *expected value* or *mean* represents the weighted average of the sample space values using the probability density values as weight. For a given real-valued random variable Y with known probability density function f_Y the expected value $\mathbb{E}(Y)$ is defined as

$$\mathbb{E}(Y) := \int_{-\infty}^{\infty} u f_Y(u)\,\mathrm{d}u. \tag{2.13}$$

The *variance* measures the spread or dispersion of a probability distribution and is defined as

$$\mathbb{V}\mathrm{ar}(Y) := \int_{-\infty}^{\infty} (u - \mathbb{E}(Y))^2 f_Y(u)\,\mathrm{d}u. \tag{2.14}$$

It represents the mean squared distance between all sample space values and the expected value $\mathbb{E}(Y)$. The mean distance is given by the *standard deviation* which is defined as square root of the variance

$$\mathbb{S}\mathrm{td}(Y) := \sqrt{\mathbb{V}\mathrm{ar}(Y)}. \tag{2.15}$$

Note, that the standard deviation is a direct indicator of the dispersion of the probability distribution and therefore of the data uncertainty modeled by the random variable.

The *median* represents the value in the sample space, which there is a probability of 0.5 that the random variable takes on either a value above or below the median. The median m is defined by the equation

$$\int_{-\infty}^{m} f_Y(u)\,\mathrm{d}u = 0.5. \tag{2.16}$$

For multi-variate distributions, the covariance between two random variable or components of a random vector is an important characteristic, whose impact on uncertainty analysis will be discussed in more detail throughout the following chapters. The covariance $\mathbb{C}\mathrm{ov}(Y, Z)$ between two random variables Y and Z with given bi-variate probability density function f_{YZ} is defined as

$$\mathbb{C}\mathrm{ov}(Y, Z) := \mathbb{E}((Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (u - \mathbb{E}(Y))(w - \mathbb{E}(Z))\,\mathrm{d}u\,\mathrm{d}w. \tag{2.17}$$

Closely related to the covariance is the *correlation* between two random variables. The *correlation coefficient*, also known as normalized covariance, is defined as

$$\mathbb{C}\mathrm{orr}(Y,Z) := \frac{\mathbb{C}\mathrm{ov}(Y,Z)}{\mathbb{S}\mathrm{td}(Y)\,\mathbb{S}\mathrm{td}(Z)}. \tag{2.18}$$

The importance of the correlation for the uncertainty analysis and possible significant contributions for a more reliable uncertainty visualization is discussed in chapter ...

Note, that distribution functions can be characterized by diverse other parameters like skewness, kurtosis, symmetry, etc. However, as these measures are not considered in the stochastic concepts introduced in this thesis, they are not discussed in detail in this section.

2.2.3 Normal Distribution

Probability distribution functions are means for quantitatively describing the uncertainty modeled by a certain (multi-variate) random variable. They directly reveal which values or ranges in the sample space are more likely than others and allows a quantitatively measurement of certain events by assigning probability values to them.

Different analytic probability density and respective cumulative distribution functions exist for modeling probability distributions for diverse phenomena or processes in natural and social sciences. In the context of uncertain data sets generated by simulation or inverse computation processes, as introduced in 1.1, on the one hand, distribution functions are used for modeling the uncertainty with respect to the initial conditions or the parameter setup. In this case, a priori knowledge is utilized for selecting the correct probability density function, representing, for instance, the uncertainty or error introduced by a measuring instrument. The chosen distribution functions are often only represented by their characterizing parameters during the computation process, which are used for propagating the input uncertainty to the output. On the other hand, distribution functions are used for modeling the uncertainty in the output data of a generation process. Thereby, a concrete probability density function is often only assumed or derived from the knowledge on the uncertainty propagation and transformation process during the computation.

A popular distribution function for modeling errors in physical measurements, a priori parameter setups or the underlying uncertainty in the resulting data set is the *normal* (or *Gaussian*) *distribution*. Its probability density g and cumulative distribution function G for the continuous real-valued one-dimensional sample space is defined as

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R} \tag{2.19}$$

$$G(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \quad x \in \mathbb{R}. \tag{2.20}$$

The parameters μ and σ represent the mean (expected value) and the standard deviation of the distribution. For the special case $\mu = 0$ and $\sigma 1$, the Gaussian distribution is called *standard normal distribution* and its function are denoted by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R} \tag{2.21}$$

$$\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \quad x \in \mathbb{R}. \tag{2.22}$$

The Gaussian distribution is also defined for a multi-variate and multi-dimensional random variable $\mathbf{Y}$. The multi-variate probability density function is given as

$$g(\mathbf{y}) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right), \tag{2.23}$$

with mean vector μ and covariance matrix Σ. The covariance matrix contains covariance values for all component pairs of $\mathbf{Y}$. Note, that for a given multi-variate Gaussian distribution for $\mathbf{Y}$, also all components and subsets thereof are Gaussian distributed.

Gaussian distributions are extremely important in statistics, uncertainty modeling and its quantification. In particular, the error and the uncertainty in the measurement of physical phenomena can very often be modeled correctly by a Gaussian distribution. One reason for this is the *central limit theorem*: Under certain conditions, the mean of a sufficiently large number of independently distributed random variables is approximately normal distributed [Sil86]. Thus, due to the fact that uncertain physical quantities are often the sum of many independent erroneous processes, they have a distribution very close to the Gaussian one. Therefore, normal distributions are often utilized for modeling the initial conditions or physical parameters of a simulation or inverse computation process.

Another reason for the popularity of Gaussian distributions in uncertainty modeling is that the uncertainty propagation in the numerical process can often be derived even analytically and therefore more reliable models can be obtained for the final resulting data uncertainty. For instance, for widely used linear inverse problems, the Gaussian distribution is invariant under linear transformations: For normal distributed initial parameters or conditions the assumption of Gaussian distributed uncertainties in the output data is valid.

Gaussian distributions are fully characterized by the mean and standard deviation (covariance) parameters. In many applications, the user is only interested in these primary uncertainty indicators. In this context, Gaussian distributions are often utilized as an assumption, especially when the actual distribution of a random variable is unknown or its exact derivation is not desired or necessary.

2.3 Uncertainty Indicators

For analyzing uncertain data sets and especially for visualization purposes, so-called uncertainty indicators are necessary for quantitatively assessing the degree of uncertainty and its distribution over the spatial domain.

If multi-variate distribution functions or the marginals with respect to single spatial domain points can be analytically derived or computed during the data generation process, they serve as most comprehensive uncertainty indicator, as probability density functions directly reveal the likelihoods of possible values in the sample space and indicate the spread over the whole sample domain. Furthermore, the density function can reveal the distribution's modality (e.g., one, two, or more maximums) and possible outlier value ranges. If the distribution type is known beforehand or follows a specific assumption (e.g., Gaussian distribution), it is sufficient to only derive characterizing parameters (e.g., mean and standard deviation).

2.3.1 Mean and Standard Deviation

The standard parameters, which are used for characterizing probability distributions, are the mean and the standard deviation (square root of variance). The mean is often used as representative of the distribution as it indicates the expected value by computing the sample space average weighted by the probability values. However, using the mean can result is false assumptions. This is especially the case for multi-modal distribution, where the mean is not positioned close to the maximum likelihood region of the probability density. For instance, for a bi-modal probability density function with two equally high main peaks, the mean is positioned between the peaks at a location with low likelihood values. Therefore, the use of the mean as representative could be misleading, as it does not represent a very likely result. The standard deviation, however, is a reliable direct indicator for the uncertainty represented by a certain probability density function. It indicates the deviation from the mean value and the dispersion of the probability distribution. A low or high standard deviation represents a low or high degree of uncertainty.

The mean and standard deviation are often combined to obtain further uncertainty indicators. Commonly used constructs are so-called *confidence intervals*. The confidence interval definition used in this thesis describes an interval $\mathcal{I}(Y)$ in the sample space for a random variable Y containing all values, whose distance to the mean value $\mu(Y)$ is less than the respective standard deviation $\sigma(Y)$

$$\mathcal{I}(Y) := \{y \in \Omega = \mathbb{R} \mid \ |y - \mu(Y)| \leq \sigma(Y)\}. \tag{2.24}$$

This definition can be re-written as

$$\mathcal{I}(Y) := \{y \in \Omega = \mathbb{R} \mid \mathcal{M}_Y(y) \leq 1\}, \tag{2.25}$$

by defining the so-called Mahalanobis distance [Mah36]

$$\mathcal{M}_Y(y) := \frac{|y - \mu(Y)|}{\sigma Y}, \quad y \in \Omega = \mathbb{R}. \tag{2.26}$$

Both the confidence interval and the Mahalanobis distance can be defined for arbitrary (multi-variate) random variables $\mathbf{Y}$ as

$$\mathcal{I}(\mathbf{Y}) := \{\mathbf{y} \in \Omega = \mathbb{R}^n \mid \mathcal{M}_{\mathbf{Y}}(\mathbf{y}) \leq 1\}, \tag{2.27}$$

$$\mathcal{M}_{\mathbf{Y}}(\mathbf{y}) := \sqrt{(\mathbf{y} - \mu(\mathbf{Y}))^\top \Sigma^{-1} (\mathbf{y} - \mu(\mathbf{Y}))}, \quad \mathbf{y} \in \Omega = \mathbb{R}^n. \tag{2.28}$$

Note, that the one-dimensional definition is equal to its multi-variate counterpart for $n = 1$. Confidence intervals are commonly used as uncertainty indicator, especially for constructing so-called *box-plots* in statistical analysis. If the probability distribution of a random variable is known or assumed, the respective confidence interval defines a region in the sample space, in which the true value is positioned with a certain probability. For instance, for a Gaussian probability distribution $\mathcal{P}$, one obtains $\mathcal{P}(\mathcal{I}(Y)) = 0.68$ for $\Omega = \mathbb{R}$.

For non-zero standard deviations, the Mahalanobis distance provides a well-defined metric on uncertain data sets. It is commonly used in classification algorithms and cluster analysis for uncertain data [DMJRM00]. For Gaussian distributed data sets, where the mean is equal to the maximum likelihood point and the likelihood is exponentially decreasing with increasing distance from the mean in the sample space, the Mahalanobis distance is a direct indicator for the likelihood given by the probability density function.

2.3.2 Stochastic Dependence

So far, only the *absolute uncertainty* of data value at a particular spatial domain has been considered. The *relative* behavior of two random variables Y and Z with respect to each other can be described by the concept of *stochastic dependence*. Stochastic dependence refers to any stochastic relationship between two random variables. Two random variables are dependent when the probability distribution of one variable does dependent on a particular setting of the other random variables. For instance, if Y and Z are stochastically dependent, the probability distributions on Z are not necessarily equal for two different assumptions $Y = y_1$ and $Y = y_2$.

The stochastic dependence can be also described using the definition of the *conditional probability*. For two random variables Y and Z the probability of a certain event $E_Y \in \mathcal{F}_Y$ under the condition of the true event $E_Z \in \mathcal{F}_Z$ is

$$\mathcal{P}(E_Y | E_Z) = \frac{\mathcal{P}(E_Y \cap E_Z)}{\mathcal{P}(E_Z)}. \tag{2.29}$$

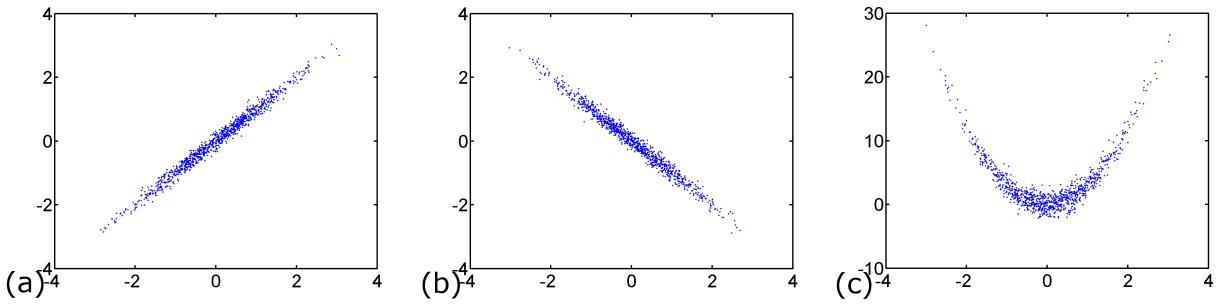

Figure 2.1: Three scatterplots for a bi-variate distributed random variable are shown. In (a) and (b), the realizations follow a standard Gaussian distribution with a strong positive and negative correlation. In (c), a strong non-linear stochastic dependence is illustrated.

For independent variable the equation $\mathcal{P}(E_Y \cap E_Z) = \mathcal{P}(E_Y)\mathcal{P}(E_Z)$ holds true which results in the equality $\mathcal{P}(E_Y|E_Z) = \mathcal{P}(E_Z)$. Therefore, the probability distribution on Y does not dependent on any assumption on a true event on Z.

In this thesis, the stochastic dependence is classified into two different types: The *linear dependence* indicates a high probability of a linear positive or negative relationship between two random variables. On the contrary, also non-linear relationships between random variables are possible, resulting in a *non-linear dependence*. Note that two bi-variate Gaussian distributed random variables are either independent or linear dependent.

Fig. 2.1 shows three scatterplots for a bi-variate distributed two-dimensional random variable. In (a) and (b), a bi-variate standard Gaussian distribution with a strong positive and negative correlation between the two components of the random vector is illustrated, respectively. The two scatterplots directly reveal the strong linear stochastic dependence between the component random variables. In (c), the dependence between the components of the bi-variate random variable is highly non-linear and therefore cannot be characterized by a correlation value.

The correlation, as introduced in Equ. (2.18) is a direct indicator for the linear stochastic dependence between two random variables. Positive and negative correlation values represent positive and negative linear relationships, respectively. For instance, for a positive relationship, a positive deviation from a reference value (e.g., mean value) for one random variable is likely to result in also a positive deviation for the other random variable.

2.3.3 Indicator Estimation

However, in most cases, the full probability distribution cannot be derived analytically. For instance, this is not possible for ensemble data sets, where for each spatial point in the data domain a collection of values is given (cf. Equ. (2.3)). If the probability density function for one random

variable Y is only represented by an ensemble of m scalar values $\mathbb{E}_m = \{y_1, y_2, ..., y_m\}$, these values are called *realizations* of the probability distributions. The realization can be either used for deriving a discrete probability function, also called *histogram*, or for a direct estimation of the indicator values discussed above.

The mean and the standard deviation can be obtained from the realizations in $\mathbb{E}_m$ by applying the unbiased estimators

$$\mu(Y) = \frac{1}{m} \sum_{i=1}^{m} y_i, \tag{2.30}$$

$$\sigma(Y) = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (y_i - \mu(Y))^2}. \tag{2.31}$$

The correlation $\rho(Y, Z)$ between the random variable Y and Z can be obtained by applying the estimator

$$\rho(Y, Z) = \frac{1}{(m-1)\sigma(Y)\sigma(Z)} \sum_{i=1}^{m} (y_i - \mu(Y))(z_i - \mu(Z)) \tag{2.32}$$

to the respective realizations. Thereby, $\mu(Y)$, $\mu(Z)$, $\sigma(Y)$ and $\sigma(Z)$ represent the estimated mean values and standard deviations.

The parameters are often estimated with respect to a certain level of significance, based on the theory of statistical hypothesis tests. Therefore, the reliability of the estimation with respect to the significance level is highly dependent on the number of realizations.

3

General Approaches to Uncertainty Visualization

Although the desire for a comprehensive analysis of errors and uncertainties in numerical simulations, inverse problem computations or other numerical process has been present over several decades, the importance of uncertainty visualization and its regard as one of the grand challenges in visual computing [Joh04] has been realized only throughout the last 15 years. However, in the young research field of uncertainty visualization several techniques have been proposed for visually representing uncertainty measures for diverse types of data sets. This chapter will provide an overview on a selection of the most prominent methods and introduce the main concepts and principles with regard to uncertainty visualization. The following chapters of this thesis will point out specific challenges for visualizing uncertain data sets and propose suitable novel solutions. Related work and supporting literature will be discussed there and is not part of this overview chapter.

One of the first publications, providing an overview on sources of uncertain data sets, general challenges in uncertainty visualization and a classification of suitable techniques, can be in [PWL97]. This 1997 work has been followed by a variety of publications, introducing an updated overview on more recent advances in this field and new concepts for approaching diverse challenges in uncertainty visualization. For further details, the reader is referred to the introductory publications [GS05, GS06, JS03, LPK05, LKP03, MRH*05, Pan08, PRJ12, ZC06].

In the following, the presented visualization approaches are classified according to the chosen modality of visually communication uncertain data to the viewer. Therefore, the two classes of *extrinsic* and *intrinsic* uncertainty visualization methods are used, as suggested i.e. in [HM96, BAF08].

An *extrinsic* approach introduces additional symbols or specific entities (e.g., color, texture, noise, etc.) to the visualization to visually encode uncertainty information. Often the degree of uncertainty is only communicated qualitatively to visually separate spatial regions in the data domain, affected by lower and higher degrees of uncertainty. Extrinsic visualization do not allow any conclusions on possible variations in the visual presentation of the data under uncertainty. For instance, an extrinsic method could use a color map to encode the standard deviation of data values in 2D

scalar data set. However, it could not reveal the positional spatial variation of an iso-contour in the 2D domain. While extrinsic techniques are able to efficiently and intuitively communicate the degree of the underlying uncertainty in scalar data sets on a more qualitative and global level, they fail in visually describing the spatial (or temporal) variation of certain features (e.g., critical points, iso-contours, surfaces, etc.) in the data set.

On the other hand, *intrinsic* visualization methods focus on visually indicating the variation of certain objects or features in the data under uncertainty. They primarily reveal, how the underlying data uncertainty affect specific characteristics of the respective data. Intrinsic methods are especially utilized for simultaneously integrating a suitable number of members of an ensemble data set in one visualization to convey the data variation in certain parts of the spatial domain. They allow to quantitatively analyze the spatial uncertainty of (local) features in the data, but are suitable for revealing the uncertainty distribution over the domain to a much lesser extent. For instance, intrinsic methods are used for visualizing the variability of streamlines in uncertain vector fields, spatial confidence regions for uncertain iso-surfaces or directly show several members of an ensemble data sets utilizing animation effects.

3.1 Extrinsic Methods

3.1.1 Color

The most common technique for encoding uncertainty information is the use of certain color effects. Thereby, a suitable uncertainty indicator (e.g., standard deviation values for single data domain points) is derived from the uncertain data set. This indicator is then mapped to a pre-defined color range, used in an interactively specified transfer function or translated into other advanced color effects. This mapping is then used for highlighting uncertainty information over the total data domain (e.g., volume rendering) or only restricted to certain data features (e.g., coloring of surfaces).

In [DKLP02], the visualization of uncertain information is integrated in standard 3D volume rendering. Thereby, the uncertainty indicators are used as free parameter in the transfer function and mapped to, e.g., opacity. Furthermore, 2D transfer function are used to classify domain regions according to data values and uncertainty indicators. The degree of uncertainty is communicated by color and opacity to the viewer. In addition, the authors suggest to use noise texture to qualitatively highlight uncertain regions in 3D volume rendering. Although the presented approaches can provide a good first overview on the relative distribution of uncertainties over the 3D data domain, they are not based on a mathematical uncertainty model and lack in providing qualitative information on possible data or feature variations on the visual level.

In [Hen03] and [HT06], for a 2D data domain, the data and uncertainty values are encoded into the hue and saturation channel of the HSV color space. The resulting visualization shows different colors over the 2D domain with varying saturation, caused by a varying degree of uncertainty. Furthermore, the approach is also used for classification purposes of domain regions affected by uncertainty Although this visualization technique is very intuitive and easy to communicate, it suffers from the major limitation that data values and uncertainty information cannot be analyzed separately from each other. In regions with maximum uncertainty, the saturation is almost equal to zero and hue values cannot be seen any more.

In [KVUS*05], color is used for visualizing statistical segmentation in MRI scans, based on fuzzy classification techniques. Furthermore, a user controlled transfer function maps sensitivity values to color in 3D surface visualization of the human skull. The presented techniques put a strong focus on the user interaction with data, both for classification purposes as well as for the transfer function design.

The integration of noise textures for indicating uncertainty is applied to uncertain flow fields in [OB09]. Noise textures with varying frequency and amplitude are added to the visualization to indicate uncertain regions in the LIC representation of 2D vector fields. Furthermore, color is used to visualize the vector magnitudes. Similar to other approaches, the approach does provide an intuitive overview on the relative uncertainties over the domain but lack in communicating quantitative information on the effect of uncertainties on vector directions and magnitudes in the field.

Further examples for mapping uncertainty indicator like standard deviation to color, attached either to 3D surfaces or directly shown in the respective data domain, can be found in [PKJ08, PWB*09b, PWB*09a]. Apart from integrating both uncertainty information and data values in one visual representation, the presented scenarios also discuss the possibility of separating both entities to individual view-ports and visualizations. Especially for highly interactive tools or for scenarios with additional integration of further data (e.g., geographic data), this approach can be regarded as valid option.

In [WRL12] and [XLS10], concepts of the field of information theory are used for revealing uncertainty information in geological models and uncertain flow fields. Thereby, entropy measures are used as main uncertainty indicators communicated using specific color effects. Information entropy is a very abstract indicator and does not allow to directly analyze the variability of features or data values on a quantitative level. However, they it can be easily computed, does often not have to rely on a mathematical uncertainty model and can provide an intuitive first overview on the overall uncertainty distribution in the data domain.

3.1.2 Glyph Representation

Another option for adding uncertainty information to visualizations is the integration of specific glyph or other objects, whose shape or coloring is directly related to uncertainty indicator values. Although glyphs are commonly used for uncertain vector fields, they provide also certain advantages for normal scalar data sets. For instance, if the data values are color coded in a 2D domain, uncertainty information often cannot added easily due to a limitation of visualization channels. By adding uncertainty glyphs to the data visualization, the degree of uncertainty in certain spatial domain regions can be communication simultaneously to the data values. This allows an integrated view on both entities and consequently a more comprehensive analysis.

In [BWE05], uncertain vector fields are visualized by applying Gaussian error diffusion and a modifies LIC approach to small objected, displayed in the 2D data domain. The introduced particles/objects simultaneously reveal the vector field data (orientation and magnitude of vectors) and local variations in uncertainty. By slightly varying the appearance of the objects, regions with low and high vector uncertainty can be clearly identified. However, the visualization communicates more qualitative information and does not allow to analyze the vector field topology under uncertainty.

In [SCB*04], box-, sphere- and arrow-glyphs are used for the visual representation of underwater environmental uncertainty. Thereby, the glyphs placed placed in the vicinity of colored surfaces, representing, for instance, the underwater ground level. In this context, the glyphs can represent the uncertainty with respect the ground level depths but are not limited to this scenario. The work shows, that glyphs can also be utilized for depicting uncertainty information with respect to diverse entities. Multidimensional data is used for analyzing local oceanic effects like internal waves, thermal currents or soil structures. Thereby, glyphs, colors and surface representations interact with each other to create an integrated view on both data and uncertainty values in one combined visualization.

In [CR00], deformable annotations are added to the color-based visualization of 2D scar fields. The integrated objects are either straight lines or circular glyphs. Their shape is distorted (e.g., noise, sharpness, frequency, amplitude) according to the local degree of uncertainty. Although this visualization allows an integrated view on both data values as well as uncertainty information, the presented uncertainty representation is very coarse. Local variations in uncertainty strengths cannot be displayed easily. Furthermore, no qualitative information on the data variability can be derived.

3.2 Intrinsic Methods

3.2.1 Uncertain Vector Fields

Intrinsic uncertainty visualization techniques display the variability of specific features in uncertain data sets. Especially for vector field data sets, the visualization of features like streamlines, streaklines, critical points or vortices allows often an improved analysis of the data set's most prominent characteristics or topological structures. The visualization of the variability of such features in uncertain data sets has been approached in several publications.

In [BFMW12], a GPU based Monte Carlo particle tracing is carried out in order to construct so-called visitation maps for uncertain 3D vector fields, indicating envelopes enclosing particle pathways with a certain percentage. The pathways envelopes are then visualized using means of direct volume rendering. This approach clearly reveals the variability of particle trajectories in uncertain vector fields and also allows to quantitatively derive probabilities for the occurrence likelihoods of stream-, streak- and path-lines.

In [HLNW11], so-called flow radar glyphs are introduced for visualizing time-dependent uncertain flow fields. The glyphs indicate ranges the orientation of a vector at a particular position in the domain is positioned in with a certain probability. These angular confidence ranges are time-dependent and displayed for each radial glyph with increasing distance from the glyph's center. The chosen visualization techniques allows a clear quantitative analysis of the local variations of individual vectors in the domain. However, a global differentiation between regions with low and high uncertainties is rather complicated as the confidence intervals are only shown locally.

The work in [OGHT10, OGT11a, OGT11b, OT12] presents visual methods for conveying how uncertainties affect the topology of 2D and 3D vector fields. The main focus lies on the analyzes of the positional distribution of features like critical points, sinks, sources and vortices. Furthermore, the presented algorithms reveal the probability density of the vector field skeleton and therefore can help analyzing the effect of uncertainties on the overall vector field topology. In addition, the introduced techniques are based on a mathematical uncertainty formulation and a Gaussian assumption on the underlying vector probability distributions. Therefore, uncertainties can be presented by full probability distribution functions instead of abstract confidence ranges for individual vectors.

3.2.2 Ensemble Exploration

Apart from features in vector fields, intrinsic uncertainty visualization methods are also utilized for conveying variations in data values among individual members of an ensemble data set. In this context, the ensemble members are often directly used for visualization purposes instead of a prior derivation of specific uncertainty indicators.

In [AWH*12], the technique of ensemble surface slicing is introduced for visualizing and identifying spatial similarities among members of 3D surface ensemble data sets. Thereby, surfaces of all ensemble members are displayed simultaneously in one visual representation. However, the rendering of each surface is restricted to a certain spatial region. Therefore, the rendered object consists of a collection of sliced ensemble member surfaces. The approach suffers from two main limitations: The ensemble surface slicing does only allow a global analysis of surface variations. It is not possible to analyze the variability of surface position and geometry on a local basis. In addition, the presented approach is only suitable for a low number of ensemble members, as otherwise the surfaces slices could not be visually separated from each other.

In [PPA*12], members of 3D scalar ensemble data sets are also rendered simultaneously in one image. However, the data in each single representation is represented by small objects in the 3D domain. These objects are colored according to their membership number. Therefore, an integrated view on all ensemble members is possible. Furthermore, a techniques is introduced which restricts the rendering of the ensemble member to certain disjoint region of the screen space. Therefore, a similar effect is achieved compared to the ensemble slicing algorithm discussed above. However, this approach suffers from the same limitations.

In [BPFG11], an interactive visual approach is presented for the analysis of a sampled parameter space with respect to a multi-dimensional target space. The suggested technique focus on the uncertainty propagation and sensitivity between input parameters and target values. Thereby, the user is guided interactively to interesting parameter regions with respect to specific target settings. The inherent uncertainty is displayed as scalar ensemble simultaneously utilizing scatterplots and parallel coordinates. The presented approach is highly suitable for an improved local sensitivity analysis with respect to multiple target dimensions and for the interactive exploration of intersting parameter regions.

Instead of showing all ensemble members in one image and restricting the rendering to certain spatial regions or screen space areas, other approaches distribute members of scalar data sets along the time axis. In [Bro04], the ensemble members are represented by single frames of a movie. For instance, with advancing time, the viewer can observed visual vibrations of a surface, affected by uncertainty. The degree of visual instability is directly related to the degree of uncertainty. However, utilizing animation for representing individual members of an ensemble data set has a severe drawback. As multi-dimensional ensemble members can be regarded as realizations of a multi-variate random variable, they also represent a multi-variate probability distribution. The quality of this representation depends highly on the number of ensemble members and the dimensionality of the random variable. For increasing dimensions the number of needed realization is growing exponentially. For instance, Monte Carlo algorithms typically generate several hundred thousand ensemble members. Transforming this amount of data into frames of an animation, would require a significant amount of viewing time for a complete analysis of the represented probability

distribution. This is impracticable in most scenarios. In [ESG97], a special interpolation scheme for ensemble data sets with low number of members is proposed, resulting in a much smoother animation.

In [LLPY07], the technique of ensemble animation is used for stenosis assessment in medical volume rendering. Thereby, the position and shape of a stenosis in a blood vessel is simulated over time. The resulting data ensemble and the underlying uncertainty is presented to the analyst using an animation of the stenosis. The single frames represent the ensemble members. Due to a low number of realizations the probabilistic animation provides an intuitive intrinsic visualization of the underlying data uncertainty with respect to the stenosis assessment.

Apart from the automatic visualization of all members in an ensemble using animation techniques, other visualization tools allow an interactive exploration of the data sets by the user. In [PWB*09b] and [PWB*09a], an ensemble exploration tool is presented for analyzing numerical weather forecast data sets. Thereby, the user can flip through the individual ensemble members and interactively analyze the variability of specific features in the data. A similar approach is utilized in the work by [WFR*10]. The presented technique allows a highly interactive visual exploration of flooding scenarios for varying parameter setups in a related simulation process. Thereby, the simulation results can be mapped to various visual variables in order to highlight the most compelling solutions.

3.2.3 Confidence and Probability Analysis

Another area of application for intrinsic uncertainty visualization methods is the direct analysis of probability distribution function or confidence regions modeling the variability of data values or specific data features.

In [CBDT11], a new method is proposed to visualize uncertain scalar fields by integrating perceptually adapted Perlin noise into a color scale visualization of the data values. The parameters of the noise pattern is directly controlled by the underlying uncertainty information. The noise perturbs the color presentation within in the local confidence intervals of the respective data values. Therefore, it is possible to simultaneously shown the data values and their uncertainty while providing a hint on the respective confidence intervals due to the color noise pattern. However, the presented technique can analyze scalar value uncertainties only locally and do not allow to derive any assumptions on stochastic dependences between spatial regions in the data domain.

In [CCM09], several techniques are presented for integrating uncertainty analysis in visual analytics. A special focus is put on the uncertainty propagation in data transformation processes like regression, principal component analysis and k-means clustering. The uncertainty information is communicated using scatterplots and confidence regions for single data points.

In [LFLH07], techniques are presented for the visualization of uncertainties in large-scale astrophysical environments. A strong focus is put on the perception and comprehension of uncertainty across different scale ranges. The introduced visual approaches include an ellipsoid model to represent the positional uncertainty of certain astrophysical events and an ellipsoid envelope model for exposing trajectory uncertainty. In both cases the spatial variability can be assessed quantitatively from the displayed visualizations.

The idea of displaying confidence intervals for single data values has also influence the visualizations in [PKR07] and [PKRJ10]. The presented techniques extend the canonical box plot concept to 2D data. In addition, a new summary plot is suggested, incorporating a collection of descriptive statistics to highlight specific features in the data. The techniques are primarily used for obtaining a high-level overview on the data and its underlying uncertainty.

Apart from the visualization of confidence intervals as uncertainty indicators, sometimes the direct display of the actual underlying probability distribution with respect to the data values is possible. In [FKLT10], the integration of kernel density estimates for revealing probability density functions in parallel coordinate displays and scatter plots is proposed. The related visualizations allow an improved uncertainty-aware tumor segmentation in MR spectroscopy data sets. In [KLDP02], the probability density function for data values in a 2D domain is integrated in the third dimension of the visual representation. For interactively selected lines in the 2D domain the data distribution are displayed an walls, position along the coordinate axes and perpendicular to the domain. This allows a direct interactive analysis of the data uncertainty and characterizing distribution functions.

4

Positional Feature Analysis

A common challenge for uncertainty visualization techniques is the visual analysis of the variability of specific *features* in a data set. Such features can be critical points, contours, surfaces, streamlines, etc. For obtaining reliable visualizations, specific mathematical models are necessary for describing the transformation process between the data level uncertainty and the feature's positional or occurrence uncertainty. Although uncertainty visualization is regarded as one of the grand challenges in visual data exploration [Joh04], it is fair to say that standardized procedures for modeling and visualizing the effect of uncertainty on features in multi-dimensional data are rare. This chapter introduces several novel approaches for visualizing the positional uncertainty of iso-contours and iso-surfaces in 2D and 3D scalar data sets.

The first section focuses on data sets where the underlying uncertainty can be modeled by a multivariate Gaussian probability distribution and can be characterized by standard parameters like mean, standard deviations and correlation values. A novel computation scheme is presented for obtaining probability values for crossing an iso-surface during a ray-casting process. The probability values are then used as input for further opacity and color effects in a 3D visualization.

The second section covers data sets where no prior assumption on the uncertainty model can be made and introduces techniques for the visualization of the iso-contour variability in arbitrary distributed 2D scalar (ensemble) data sets. A novel computation scheme is proposed for deriving positional iso-contour occurrence probabilities. A color scheme is presented for highlighting gradient strengths and topological variabilities in the data set. Furthermore, a visual approach for analyzing the iso-contour distribution modality is proposed.

4.1 Gaussian Distributed Data Sets

One of the recent approaches for visualizing the uncertainty of particular features in 3D scalar data sets in [PH11] suggests to model the uncertainty stochastically, and to derive probability distributions for particular events that correspond to relevant features, e.g., the crossing of iso-surfaces in volume ray-casting. This allows quantifying the spatial distribution of uncertain features, enabling a statistical analysis of the effect of uncertainties in the input data on the uncertainty of these features.

Inspired by [PH11], the motivation behind the work, presented in this section, is twofold: Firstly, it is aiming for the integration of data correlations into the stochastic uncertainty model to enable a more reliable computation of iso-surface crossing probabilities along the view rays during the ray-casting process. When ignoring correlation information in the data, a zero correlation between very close sample points is assumed. This contradicts the assumption of local data continuity and results in vastly overestimated probabilities with respect to certain stochastic events related to the occurrence of features like iso-surfaces in the data.

The second goal is to develop new strategies for mapping uncertainty to optical properties in a way that allows visualizing the positional and geometrical variability of features in the data domain independently of the viewing direction. In this way, we are addressing the problem that positional uncertainties of iso-surfaces can often be revealed only in 2D cross-sections or if the viewer is *not* looking along the surface normal direction. If the iso-surface position varies along this direction, it can be perceived clearly only when the viewing direction is orthogonal to the surface normal. To overcome this limitation, an Euclidean distance measure is proposed that assigns to each possible iso-surface position the spatial deviation from the mean surface. This measure is used in a novel color mapping scheme to assess the uncertainty that is related to the shape of an iso-surface even under viewing directions parallel to the surface normal.

In Fig. 4.1 (a), the mean iso-surface is shown for an uncertain atmospheric temperature forecast data set. In (b), the novel visualization effects are shown for highlighting the confidence interval with respect to the Gaussian distributed iso-surface in the 3D domain. The effect of the proposed visual Euclidean distance variability encoding (c) is compared to an encoding that only considers the deviation in probability space (d).

4.1.1 Related Work

Major efforts have been put on the intrinsic visualization of the variability of iso-surfaces in uncertain 3D scalar fields. [PWL97, JS03] proposed to augment a mean surface by additional surfaces that enclose areas around the mean surface of high confidence. Such techniques provide a good first impression on the positional variability of the iso-surface. However, this impression is often

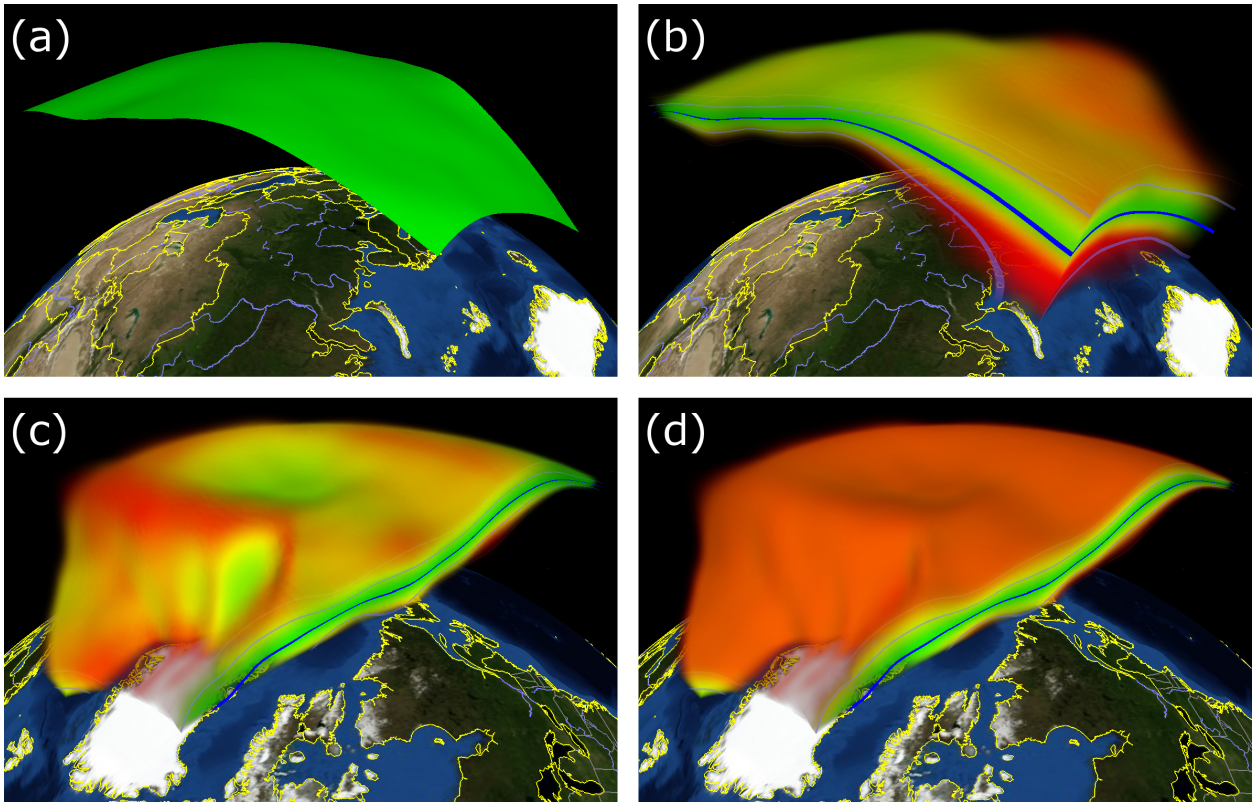

Figure 4.1: (a) The mean surface for a temperature iso-value in the ATMOS data set is shown. (b) The novel IFCP algorithm in combination with distance dependent color mapping indicates the confidence volume. (c) The spatial distance variability is indicated independently of the viewpoint by utilizing a specific novel color scheme. (d) Traditional color mapping scheme based on stochastic distance is shown.

dependent on the viewpoint as no color indicate the variability in Euclidean distance around the mean surface.

The use of opacity to show spatial contiguity and iso-surface confidence regions was demonstrated in [ZWK10]. Flow-lines were introduced in [KWTM03] to visualize the uncertainty of material boundaries. In [GR04] an iso-surface was modeled as a point set, and points were displaced from their original position by an amount proportional to the local uncertainty. The animation of possible iso-surface positions over time was demonstrated in [Bro04]. [RLBS03] uses an extrinsic approach for indicating data uncertainty by mapping color and texture on iso-surfaces. Recently, [PH11] presented a method for visualizing the positional variability around a mean iso-surface using direct volume rendering. Based on probability theory, they introduced mathematical formulations for the positional uncertainty of iso-surfaces and employed the concept of numerical condition for visually presenting how errors in the input data are amplified during the iso-surface extraction.

Most methods do not utilize any mathematical model for describing the underlying positional

iso-surface uncertainty or represent uncertainty indicators (e.g., confidence regions) or probability values using only an abstract mapping to color or opacity effects. The techniques presented in this section will extend such approaches by using a more sophisticated mathematical framework, integrating also correlation information for obtaining more accurate results. Furthermore, an automatic mapping between color and opacity effects and computed positional probabilities will be presented, allowing a more intuitive and less abstract visualization.

4.1.2 Stochastic Modeling of Uncertainties

In the context of 3D scalar data sets, by uncertainty one understands the mean deviation of the data samples from a true or assumed value without precise knowledge of the magnitudes of these deviations. The data samples are assumed to be attributed by parameterized uncertainty, which will be considered in the visualization of the positional and geometrical variability of iso-surfaces in the data.

Uncertainty Representation

The 3D scalar field is assumed a discrete sampling of a mapping from the continuous spatial domain $\mathbb{S} \subseteq \mathbb{R}^3$ into $\mathbb{R}$. The sampling is represented by a finite set of n *spatial points*

$$\mathbb{S}_n = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{S}, i \in \{1, 2, ..., n\}\} \subset \mathbb{S}. \tag{4.1}$$

The mapping and its uncertainty is modeled as a *random function*

$$Y \ : \ \mathbb{S}_n \longrightarrow \mathbb{R}, \ \mathbf{x} \longmapsto Y(\mathbf{x}), \tag{4.2}$$

where for each spatial point $\mathbf{x}_i$ the mapping $Y(\mathbf{x}_i)$ is considered as single *random variable*. As the random function is defined on a finite set of discrete spatial points, it can be interpreted as *random vector* $\mathbf{Y}$ with components $Y_i := Y(\mathbf{x}_i)$.

The random function is characterized by a n-dimensional probability density function $f(y_1, y_2, ..., y_n)$, where y_i is a *realization* or *observed value* of the random variable $Y(\mathbf{x}_i)$. Throughout this section, f is assumed a multi-variate *normal* (or Gaussian) probability density function (MNPDF). MNPDFs are commonly used for modeling probability densities since they have often shown to adequately represent random fluctuations in measured values of deterministic quantities (cf. 2.2.3).

In order to fully characterize f, for every sample point the mean value $\mu_i := \mu(Y_i)$ and the standard deviation $\sigma_i := \sigma(Y_i)$ have to be known, and pair-wise correlations $\rho_{ij} := \rho(Y_i, Y_j)$ are required to build the *covariance matrix* of a MNPDF. With given mean values and the respective covariance matrix, the assumed Gaussian probability function is fully characterized.

Notably, a MNPDF is completely characterized if the m-dimensional probability density function $f(y_1, y_2, ..., y_m)$ is known for realizations y_i at *any points* $x_1, x_2, ..., x_m \in \mathbb{S}$ and *any value* of m [Tar05]. Thus, as the data set is comprised of values at a discrete set of spatial points in $\mathbb{S}_n$, one only has to specify f for any $m \leq n$ points in $\mathbb{S}_n$. For instance, this is the case if f is known for $m = n$ as all low-order *marginals* $f(y_{t_1}, y_{t_2}, ..., y_{t_p})$ with $t_i \in \{1, 2..., n\}$ and $p < n$ are automatically given by integration (cf. (2.12)). A MNPDF, in particular, can be completely described by its 2-order marginals [Tar05]. For uncertainty visualizations using correlations, this is in particular appealing since the computation of these 2-order marginals can be incorporated directly into front-to-back ray-casting.

Correlation

There are many ways to understand the meaning and effect of correlation [LRN88]. An overview and further details have already been introduced in section 2.3.2. In the following, the correlation is interpreted as a measure for *stochastic dependence* and as a modeling tool for smoothness and continuity. For MNPDFs, correlation is specified in the form of a symmetric correlation matrix, which contains *correlation coefficients* $-1 \leq \rho(Y_i, Y_j) \leq 1$ between two components of the multidimensional random vector $\mathbf{Y}$. These coefficients are a direct measure of the stochastic dependence of two components of $\mathbf{Y}$.

If Y_i deviates positively from μ_i by the magnitude $\Delta\mu$, a large correlation value indicates that Y_j also deviates positively ($\rho_{ij} \approx 1$) or negatively ($\rho_{ij} \approx -1$) around $\Delta\mu \cdot \frac{\sigma_j}{\sigma_i}$ from μ_j. For $\rho_{ij} \approx 0$ the realizations of Y_j are considered uncorrelated and stochastically independent from Y_i.

Discrete samplings of a continuous mapping usually assume a certain local smoothness and at least local continuity of the sampled quantity. To achieve this, random function correlation is described by *spatial distance dependent correlation functions* [Tar05]. In the case of a MNPDF, one typically uses the *exponential correlation function* (ECF)

$$\rho(Y_i, Y_j) = \exp(-\tau \left\|\mathbf{x}_i - \mathbf{x}_j\right\|), \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbb{S}_n, \tag{4.3}$$

which assigns higher correlations to random variables of points with smaller Euclidean distance. If the correlation strength τ is defined locally for each point in $\mathbb{S}_n$, the ECF becomes

$$\rho(Y_i, Y_j) = \exp(-0.5(\tau(\mathbf{x}_i) + \tau(\mathbf{x}_j)) \left\|\mathbf{x}_i - \mathbf{x}_j\right\|). \tag{4.4}$$

To model anisotropic correlations, the parameter τ can be made dependent on a specific direction. For a unit vector $\mathbf{r}$, the parameter τ at point $\mathbf{x}_i$ in direction $\mathbf{r}$ is then given by

$$\tau(\mathbf{x}_i, \mathbf{r}) = \mathbf{r}^\top \mathbf{T}(\mathbf{x}_i)\mathbf{r}, \tag{4.5}$$

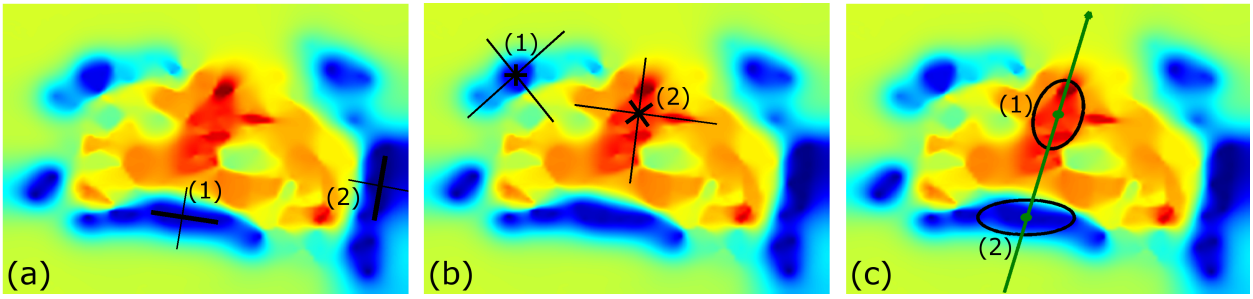

Figure 4.2: Mean values of a seismic tomography model below Australia are shown. (a), (b) Strong and weak correlation strengths between cross center points and other spatial points are illustrated by thick and thin lines. (c) Anisotropic correlations strengths are represented by 2D tensor ellipses.

where $\mathbf{T}$ is a rank-2 tensor that models the anisotropy. This tensor can either be derived from the correlations in the data samples, or it can be specified based on prior knowledge. The six different components of $\mathbf{T}$ can be obtained by solving an (overdetermined) linear system, utilizing correlation values between $Y(\mathbf{x}_i)$ and an 1-ring of neighboring random variables as input parameters.

For instance, if one assumes at every point $\mathbf{x}_i$ different parameters $\tau_1(\mathbf{x}_i)$, $\tau_2(\mathbf{x}_i)$, and $\tau_3(\mathbf{x}_i)$ along the three major spatial directions, the tensor coefficients are given as $T_{kl} = 0$ for $k \neq l$ and $T_{kk} = \tau_k$. The parameters τ_k can be computed by solving the linear equation system

$$\tau_1(\mathbf{x}_i)(x_{h1}-x_{i1})^2+\tau_2(x_i)(x_{h2}-x_{i2})^2+\tau_3(x_i)(x_{h3}-x_{i3})^2 = -\log(|\rho(Y_i,Y_h)|)\,\|\mathbf{x}_h-\mathbf{x}_i\|\,,\quad h\in\{k,l,m\} \tag{4.6}$$

for at least three neighboring points $\mathbf{x}_k$, $\mathbf{x}_l$, and $\mathbf{x}_m$ of $\mathbf{x}_i$ and correlation values $\rho(Y_i, Y_k)$, $\rho(Y_i, Y_l)$ and $\rho(Y_i, Y_m)$ usually given by the data as correlation matrix. In this formulation, x_{is} denotes the s-th component of vector $\mathbf{x}_i$. For a homogeneous correlation model, $\tau(\mathbf{x}_i)$ can be defined as the mean of $\tau_1(\mathbf{x}_i)$, $\tau_2(\mathbf{x}_i)$, and $\tau_3(\mathbf{x}_i)$. Note that only the magnitude of the local correlation is modeled by the ECF.

In Fig. 4.2, mean values of a seismic tomography model below Australia are shown. The color indicates velocity values for seismic pressure waves. Image (a) illustrates the anisotropic nature of correlation information. The black crosses (1) and (2) indicate the correlation between the random variables at the crosses' center points and values in the respective two orthogonal direction. The thin and thick black lines represent a low and high correlation strength in the respective directions. This observation is also supported by the underlying mean data field. The data values in the high correlation direction are much more similar to the cross center point than values in the low correlation direction. In this case, the correlation anisotropy indicates the preferential orientation of structures and features in the data set.

In (b), the distance dependence of correlation is illustrated. One can observe that values at spatial

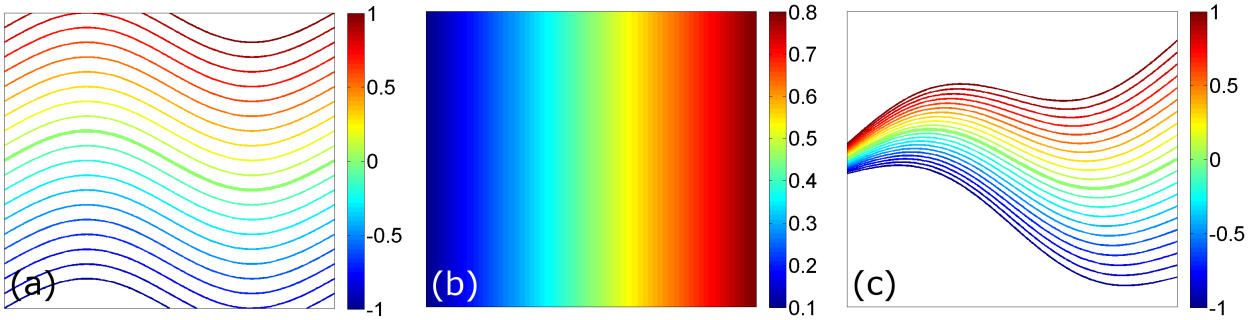

Figure 4.3: (a) Mean iso-contours for values in $\{-1.0, -0.9, -0.8, ..., 1.0\}$ are shown for an uncertain 2D scalar data set. (b) Standard deviation values are color-coded over the 2D domain. (c) SDF iso-contours for SDF values in $\{-1.0, -0.9, -0.8, ..., 1.0\}$ are illustrated.

locations close to the center points of the crosses (1) and (2) are more similar and therefore higher correlated (thick lines) than values at spatial points which are positioned more distant from the cross centers (thin lines). This assumption is valid for most numerical simulation or inverse computation data sets characterized by a certain local data continuity and smoothness.

In (c), 2D ellipses are shown at two spatial positions (1) and (2), illustrating the local anisotropic correlation strengths, modeled by the 2-rank parameter tensors. For the given direction (green arrow) and by using the tensor model, one obtains strong and low local correlation strengths for the two spatial points at (1) and (2), respectively.

Stochastic Distance Function

To relate the possible occurrence of an iso-surface to the local uncertainty, the *stochastic distance function* (SDF)

$$\Psi_\theta(\mathbf{x}_i) := \frac{\mu_i - \theta}{\max(\sigma_i, \sigma_{min})}, \quad \mathbf{x}_i \in \mathbb{S}_n, \quad \theta \in \mathbb{R}, \tag{4.7}$$

is often used. Here, θ refers to a specified iso-value, and a minimum standard deviation σ_{min} is assumed to avoid numerical problems.

The SDF corresponds to the 1-dimensional formulation of the Mahalanobis distance as introduce in section 2.3.1 (cf. [Mah36, DMJRM00]), and it indicates at point $\mathbf{x}_i$ the distance of the mean value μ_i to the iso-value in number of σ_i. SDF fields are often used to depict the *confidence volume* containing the level-θ iso-surface with a certain probability [ZWK10], or the SDF values are used for color-coding uncertain iso-surfaces as in [PH11]. The presented novel algorithm computes the SDF for all grid vertices on-the-fly and uses these values either for computing iso-surface crossing probabilities or for determining probability gradient magnitudes.

In Fig. 4.3, the importance of the SDF for visualizing uncertain data is illustrated in an synthetic example. Mean scalar values in the range $[-1, 1]$ are given on a 2D domain. In (a), the mean data

set is represented by iso-contours for iso-values θ in the equidistant set $\{-1.0, -0.9, -0.8, ..., 1.0\}$. The iso-contours are colored according to the respective iso-value. In the following, the positional uncertainty in the 2D domain of the $\theta = 0$ iso-contour (green thick line in the middle) is analyzed.

In (b), the standard deviation values as primary uncertainty indicators are visualized in the 2D domain, linearly ranging from low to high values from the left to the right part of the domain. Solely by looking at the mean and the standard deviation values in (a) and (b), it is a hard challenge to conclude on the positional uncertainty of the $\theta = 0$ iso-contour.

In (c), the SDF field is shown for the 2D domain, using the respective local mean and standard deviation values and $\theta = 0$ as input. The SDF field is represented by discrete SDF-iso-contours for SDF values in the setting $\{-1.0, -0.9, -0.8, ..., 1.0\}$. The colors represent the respective SDF values. The illustrated SDF range $[-1, 1]$ is directly related to a confidence region with maximum deviation from a mean representation equal to the standard deviation (cf. 2.3.1). Image (c) shows exactly this confidence region for the uncertain $\theta = 0$ iso-contour. It is narrower and wider and therefore more certain and uncertain in the left and right part of the domain, respectively.

For instance, at the left domain border, the standard deviation is equal to 0.1. This indicates a data variability in the range of ± 0.1. As the mean iso-contours in (a) are shown for a value spacing equal to 0.1, it can be assumed that the $\theta = 0$ iso-contours varies in the region between the upper and lower neighboring contours for iso-values equal to 0.1 and -0.1 at the left domain border. This assumptions is validated by the SDF field as the ± 1 SDF region covers exactly this range at the left domain border. Therefore, the encoding of SDF information in visualization algorithms is a good first step towards a reliable uncertainty analysis.

4.1.3 Probabilistic Iso-surface Extraction

In the following, the novel approach for computing positional probabilities of iso-surfaces in uncertain 3D scalar fields via volume ray-casting is described. A 3D grid structure is assumed, attributed by a mean μ_i and a standard deviation σ_i at every grid vertex $\mathbf{x}_i$. For a given iso-value, this allows computing per-vertex SDF values as described in 4.1.2. In addition, for every cell, a rank-2 tensor is stored according to the ECF model as discussed in 4.1.2. The 6 distinct tensor values are stored as per-cell attributes. The grid structure can be arbitrary but is supposed to be equipped with a local cell-wise interpolation scheme to reconstruct SDF values at any point in the 3D domain.

Iso-surface ray-casing is performed by sampling the scalar field along the view rays in front-to-back order. The presented approach for uncertainty visualization computes along each ray and in each sampling interval the probability of crossing the iso-surface *for the first time*. The technique is intertwined with the front-to-back traversal in that it provides an incremental update-scheme for determining these probabilities solely based on local evaluations. Finally, a discrete probability

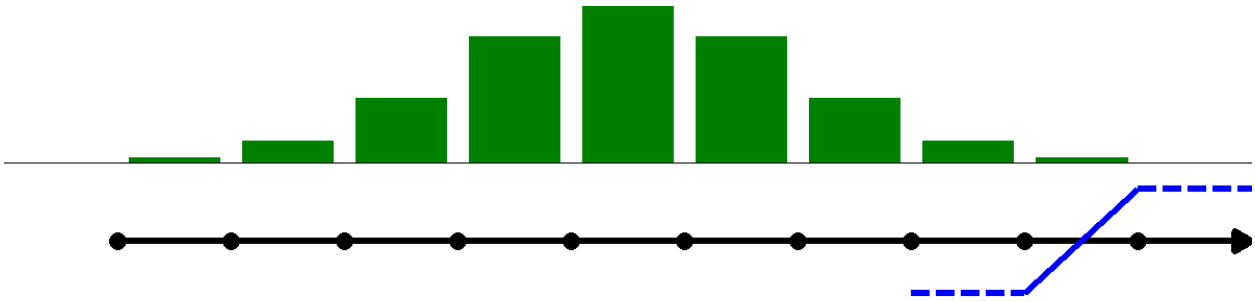

Figure 4.4: An illustration for the iso-surface crossing probability in uncertain ray-casting is given. The bar chart reveals crossing probabilities between a ray (black) and an iso-surface (blue) for equidistant sampling intervals.

density function on the sampling intervals along each ray is obtained with respect to the stochastic event that the iso-surface is crossed for the first time.

Fig. 4.4 provides a basic overview on the overall concept. A ray with discrete sampling points is illustrated in the lower part of the image. The stochastic event of crossing an iso-surface (blue line indicates an iso-contour) is shown in the last sampling interval. The introduced algorithm aims for computing a discrete probability distribution along each ray through the volume, providing crossing probability values for each sampling interval. This probability density is illustrated as green bar chart in the upper part of Fig. 4.4. Note that the sum of all probability values for each ray has to be less or equal to one. The probability values are then used for further rendering purposes.

Iso-surface Crossing Probability

The continuous volume $\mathbb{S}$ is sampled along each ray at equidistant discrete linearly ordered sample points $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\} \subset \mathbb{S}$. At each point $\mathbf{s}_i$ a SDF value ψ_i is obtained via the respective interpolation scheme, and for two consecutive sample points $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$ a correlation value $\rho_i := \rho(Y(\mathbf{s}_i), Y(\mathbf{s}_{i+1}))$ is computed as described in 4.1.2. From this data, the probability p_i for crossing the iso-surface in the sampling interval $I_i = [\mathbf{s}_i, \mathbf{s}_{i+1}]$ is calculated.

Notably the probability p_i *cannot* be computed by only considering the current interval. In this case, the probability $p_i + p_{i+1}$ would also consider the event of crossing the surface in I_i *and* I_{i+1}. The probability of crossing the iso-surface in an arbitrary interval is called local *marginal* crossing probability. However, to guarantee a reliable positional probability estimation, the event of crossing the iso-surface *either* in I_i *or* in I_{i+1} has to be considered, i.e., based on a XOR combination of disjoint crossing events. Therefore, the events

$$Y_i^+ := (Y(\mathbf{s}_i) \geq \theta) \tag{4.8}$$

$$Y_i^- := (Y(\mathbf{s}_i) < \theta), \tag{4.9}$$

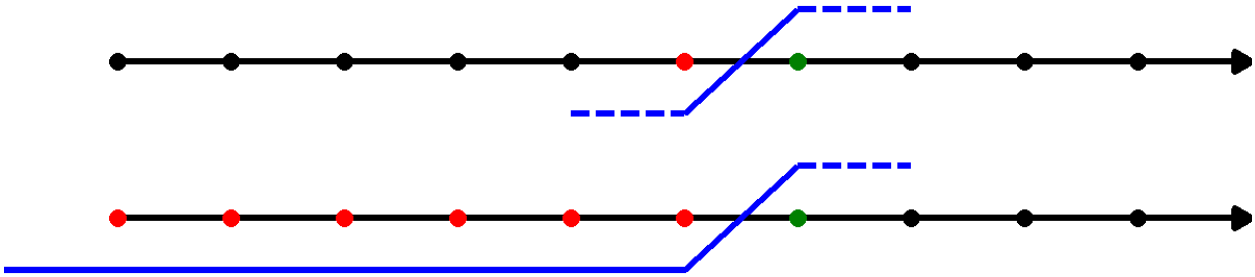

Figure 4.5: The concepts of local marginal crossing probability (upper row) and first time crossing probability (lower row) is illustrated. Sampling points with value lower and high than the iso-value are marked red and green.

are introduced, which compare the value of the random variable Y at the sample points to the iso-value. By using these events, the *positive first crossing event*

$$C_i^+ := Y_1^- \cap Y_2^- \cap ... \cap Y_i^- \cap Y_{i+1}^+ \tag{4.10}$$

can be defined; it describes the incident that the iso-surface is crossed (from lower to higher values) in interval I_i *for the first time* and no crossing has happened before this interval is reached. The negative crossing event C_i^- is defined as

$$C_i^- := Y_1^+ \cap Y_2^+ \cap ... \cap Y_i^+ \cap Y_{i+1}^-. \tag{4.11}$$

For simplicity, in the following, only the event C^+ is investigated — all results apply in the same way for C^-.

In Fig. 4.5, the two concepts of the first time crossing probability and the local marginal crossing probability are illustrated. Similar to Fig. 4.4, two rays with equidistant sampling points are shown. The iso-surface is represented by blue iso-contours, which are crossed by the black sampling rays.

The upper illustration illustrates the stochastic event for the local marginal crossing probability, where only the crossing in a specific interval is considered. The crossing event assumes that the scalar values of two consecutive sampling points are below (red) and above (green) the specified iso-value. Sampling points before and after the respective interval are not taken into consideration.

The lower illustrations illustrates the stochastic event for the first time crossing probability. This event assumes that *all* scalar values before the respective interval are positioned below (red sampling points) the specified iso-value and therefore no crossing has happened before the interval is reached. Note that only the event C^+ is illustrated.

The proposed event formulation guarantees that C_i^+ and C_j^+ for $i \neq j$ are disjoint events that cannot be both true at the same time, i.e., XOR combinations of crossing events are considered.

Note that the local marginal crossing events are *not* disjoint as a crossing in two consecutive sampling intervals is also possible. With a probability measure $\mathcal{P}$, the total positive iso-surface crossing probability along a ray can then be computed as $\sum \mathcal{P}(C_i^+)$. $\mathcal{P}(C_i^+)$, and respectively $\mathcal{P}(C_i^-)$, provide an indication of the positional variability around the most likely position of the iso-surface. In the presented algorithm, the total first time crossing probability is set to

$$p_i := \mathcal{P}(C_i^+) + \mathcal{P}(C_i^-) \tag{4.12}$$

to account for both positive and negative first time crossing events. Here, the fact is exploited that C_i^+ and C_i^- are also disjoint stochastic events.

For computing the positive first time crossing probability

$$\mathcal{P}(C_i^+) = \mathcal{P}(Y(\mathbf{s}_1) < \theta, Y(\mathbf{s}_2) < \theta, ..., Y(\mathbf{s}_i) < \theta, Y(\mathbf{s}_{i+1}) \geq \theta) \tag{4.13}$$

efficiently, the fact is employed that the underlying probability density function of the random function Y is assumed to be of MNPDF type (cf. 4.1.2). In general, one can evaluate the multivariate normal cumulative distribution function (MNCDF) of dimension $i+1$ for each interval I_i. Since this is by far too costly, an efficient method is proposed for incrementally updating $\mathcal{P}(C_i^+)$ with a minimum of additional operations per interval. Furthermore, the presented work shows that it is possible to compute $\mathcal{P}(C_i^+)$ using at most 2-dimensional MNCDFs and, thus, to avoid costly evaluations of high-dimensional MNCDFs.

By using the theory of conditional probability, the probability $\mathcal{P}(C_i^+)$ can be rewritten in the following way:

$$\mathcal{P}(C_i^+) = \mathcal{P}(Y_1^- \cap Y_{i+1}^+ | Y_2^- \cap ... \cap Y_i^-) \mathcal{P}(Y_2^- \cap ... \cap Y_i^-) \tag{4.14}$$

For a MNPDF with the special requirement $\rho(A, C) = \rho(A, B)\rho(B, C)$ on the correlation coefficients the following rule applies:

$$\mathcal{P}(A \cap C|B) = \mathcal{P}(A|B)\ \mathcal{P}(C|B) \tag{4.15}$$

Since an ECF (cf. 4.1.2) is utilized for the modeling of correlations, pair-wise correlations can be written as

$$\rho(Y(\mathbf{s}_i), Y(\mathbf{s}_{i+m})) = \prod_{j=i}^{i+m-1} \rho(Y(\mathbf{s}_j), Y(\mathbf{s}_{j+1})). \tag{4.16}$$

This means that the aforementioned requirement is met and Equ. (4.15) can be applied several times to Equ. (4.14) to arrive at the following equation for $\mathcal{P}(C_i^+)$:

$$\mathcal{P}(C_i^+) = \mathcal{P}(Y_1^-) \frac{\mathcal{P}(Y_i^- \cap Y_{i+1}^+)}{\mathcal{P}(Y_i^-)} \prod_{j=1}^{i-1} \frac{\mathcal{P}(Y_j^- \cap Y_{j+1}^-)}{\mathcal{P}(Y_j^-)}. \tag{4.17}$$

The high-dimensional MNCDF value $\mathcal{P}(C_i^+)$ can be expressed solely by 1- and 2-dimensional

MNCDFs of random variables of *consecutive* sample points. Thus, $\mathcal{P}(C_i^+)$ can be computed incrementally by considering only consecutive sample points along a view ray. This incremental algorithm will be subsequently called the *isosurface-first-crossing-probability* (IFCP) algorithm.

By introducing the negative no-crossing event $N_i^- := Y_1^- \cap Y_2^- \cap ... \cap Y_i^-$, the following rule is proposed for updating $\mathcal{P}(C_i^+)$ in interval I_i along a ray:

$$\mathcal{P}(C_i^+) = \mathcal{P}(N_i^-)\frac{\mathcal{P}(Y_i^- \cap Y_{i+1}^+)}{\mathcal{P}(Y_i^-)}, \tag{4.18}$$

$$\mathcal{P}(N_{i+1}^-) = \mathcal{P}(N_i^-)\frac{\mathcal{P}(Y_i^- \cap Y_{i+1}^-)}{\mathcal{P}(Y_i^-)}. \tag{4.19}$$

Here, $\mathcal{P}(N_i^-)$ can be computed in-place and $\mathcal{P}(N_1^-) = \mathcal{P}(Y_1^-)$ applies. Keeping in mind that $\mathcal{P}(Y_i^- \cap Y_{i+1}^+) = \mathcal{P}(Y_i^-) - \mathcal{P}(Y_i^- \cap Y_{i+1}^-)$, it is sufficient to compute the probabilities $\mathcal{P}(Y_i^-)$ and $\mathcal{P}(Y_i^- \cap Y_{i+1}^-)$ for each interval I_i. This is done using the SDF values ψ_i and ψ_{i+1}, the correlation ρ_i in interval I_i, as well as the uni-variate (Φ_1) and bi-variate (Φ_2) standard normal cumulative distribution functions:

$$\mathcal{P}(Y_i^-) = \Phi_1(-\psi_i), \tag{4.20}$$

$$\mathcal{P}(Y_i^- \cap Y_{i+1}^-) = \Phi_2(-\psi_i, -\psi_{i+1}; \rho_i). \tag{4.21}$$

Probability Mapping

The iso-surface first time crossing probability p_i is then mapped to opacity according to the following considerations. In volume rendering an optical emission-absorption model for accumulating color and opacity information along each ray of sight is employed. The overall opacity along a single ray is written as $\alpha = 1 - \exp(-\lambda)$, where λ can be interpreted as the number of small opaque particles the ray hits on its way through the volume. For each ray through the volume an interactively specified maximum number of particles λ_{max} is distributed among the sample intervals I_i according to their probabilities p_i. This results in the local and global opacities

$$\alpha(I_i) = 1 - \exp(-\lambda_{max} p_i), \tag{4.22}$$

$$\alpha\left(\bigcup_{i=1}^{n-1} I_i\right) = 1 - \exp\left(-\lambda_{max} \sum_{i=1}^{n-1} p_i\right), \tag{4.23}$$

where the global opacity represents the overall probability that the ray hits an iso-surface along its way through the volume. While special attention has been put on the fact that only a linear relation between p_i and $\lambda_i = \lambda_{max} p_i$ is feasible. A linear relation between p_i and α_i is not possible as the probabilities are accumulated in an additive manner whereas opacities are updated multiplicatively along the ray.

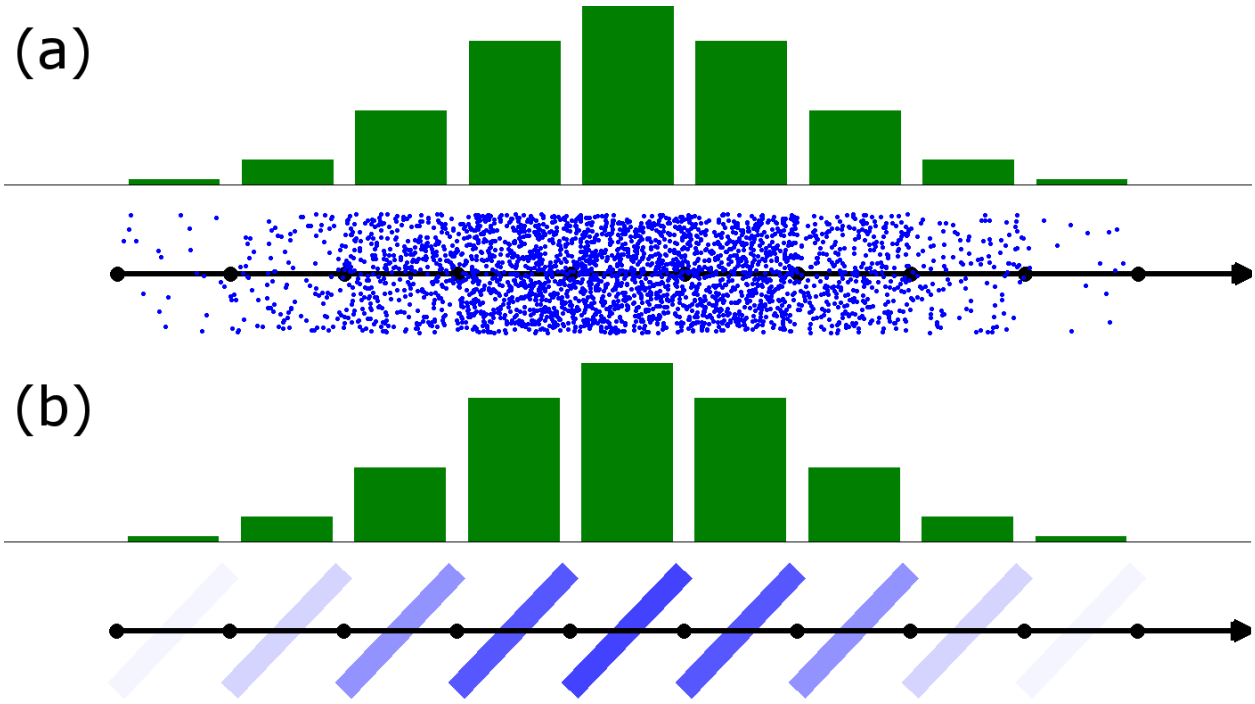

Figure 4.6: (a) Light absorbing particles are distributed along the ray according to crossing probabilities. (b) Absorption strength for opacity effects is linearly related to crossing probabilities.

The concept of relating opacity values at individual sampling intervals along the ray to iso-surface crossing probabilities is illustrated in Fig. 4.6. Similar to Fig. 4.5, rays with sampling points (black) as well as probability values (green bar charts) for the first time iso-contour crossing event in individual sampling intervals are shown. In (a), the distribution of the specified maximum number of particles λ_{max} along the ray according to the crossing probability distribution is illustrated. In (b), the individual number of particles λ_i in each sampling interval is transformed to the respective opacity value utilizing Equ. (4.22).

Probability Types

An open question is whether the IFCP algorithm can also be applied when the ray hits two iso-surfaces that have no stochastic dependence (e.g., because they are too far away from each other). Let us assume that the ray hits an iso-surface with a significant probability (e.g., 0.5) in the sample interval $I = [\mathbf{s}_i, \mathbf{s}_{i+u}]$ and also in interval $J = [\mathbf{s}_j, \mathbf{s}_{j+v}]$ with $i + u \ll j$ and $\rho(Y(\mathbf{s}_{i+u}), Y(\mathbf{s}_j)) \approx 0$. The probabilities that are computed for all sub-intervals in J are *all* conditioned under the *same* assumption that there was no crossing in I. Thus, relative to each other, all sub-probabilities for J describe the positional variability in J. Furthermore, all values grow by the same factor if the true crossing probability for I approaches 0.

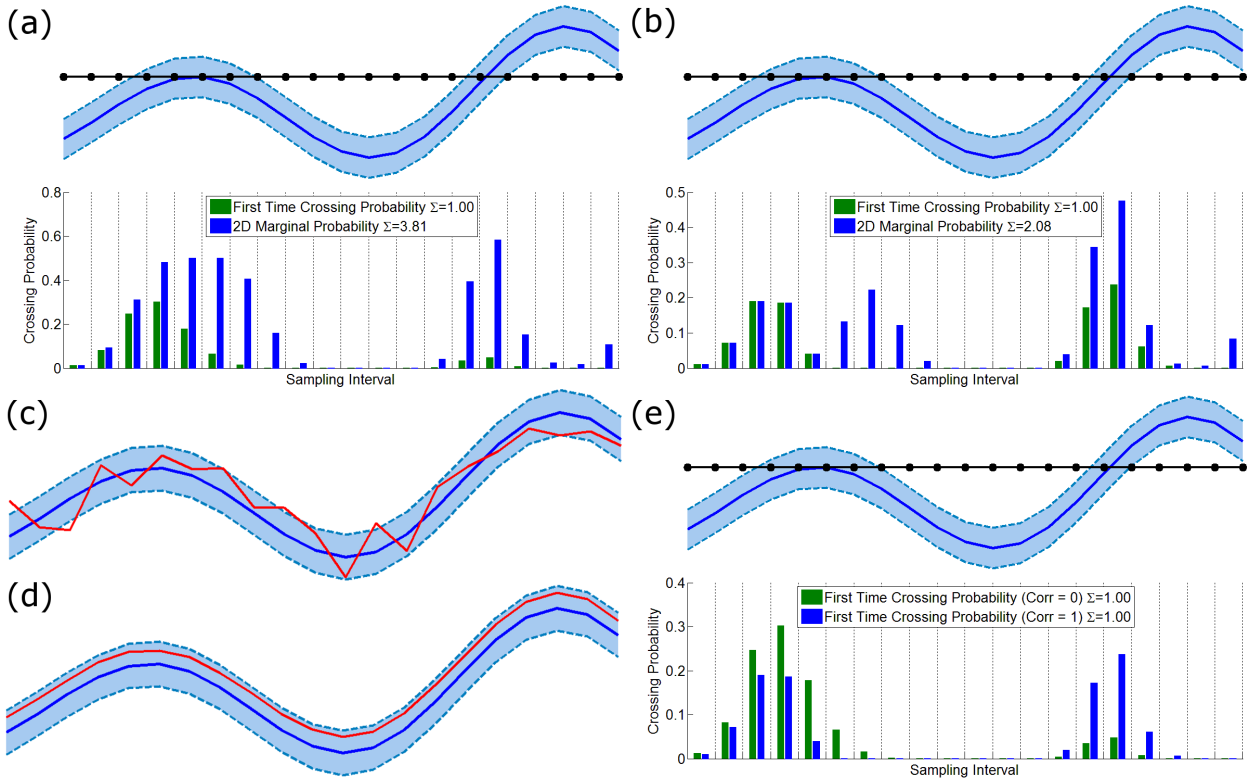

Figure 4.7: Crossing probabilities between ray and iso-contour are revealed for both probability concepts in uncertain 2D data set, represented by a mean iso-contour (blue line) and a confidence region (light blue). Crossing probabilities are shown for equidistant sampling intervals and low (a) and high (b) local correlation. (c), (d) Realizations of contour probability distribution are shown for data set (a) and (b), respectively. (e) Difference in first time crossing probabilities for low and high correlation is shown.

The overall probabilities $\sum_{k=i}^{i+u-1} p_k$ and $\sum_{k=j}^{j+v-1} p_k$ for regions with low correlation between each other are called *probabilities of occurrence* (PO). The sub-probabilities within regions of high correlation (e.g., around the maximum likelihood position in I and respectively in J) will be called *probabilities of position* (PP). The IFCP algorithm has the nice property that while computing the PPs, it simultaneously combines the POs in the correct visibility order. It further guarantees that the overall crossing probability along the entire ray never exceeds 1.

Fig. 4.7 illustrates the concept above, the difference between the first time crossing events and the marginal probabilities and highlights the role of correlation values in an synthetic 2D example. The sampling along a ray (black) in a 2D domain is shown. The iso-contour for a specified iso-value is shown as continuous blue line. Its positional uncertainty is indicated by the blue area, representing the respective confidence region. Domain points with lower and higher data values than the iso-value are positioned above and below the iso-contour, respectively. The shown iso-contour (blue) and the confidence interval represent the mean and the standard deviation of an

synthetically generated ensemble of 2D data sets.

For (a) and (b), low and high local correlation values have been used during the generation process, respectively. The red curves in (c) and (d) show iso-contours for two ensemble realizations of the data sets used in (a) and (b). The low local correlation used for data sets (a) results in a significant higher structural uncertainty of the iso-contour and a less smooth representation.

For the low (a) and high (b) local correlation case and for a ray traveling through the 2D domain and equidistant sampling points (black line), the probability values are shown for the two stochastic events of crossing the iso-contour for the first time (green) or of a local marginal crossing (blue) (cf. Fig. 4.5) in the individual intervals. One can observe that the use of the local marginal crossing events does not result in a crossing probability distribution along the ray, as the events for the individual sampling intervals are not disjoint and therefore the total probability sum is larger than one. Only the disjoint first time crossing events can provide a correct discrete probability distribution for each individual ray.

In (e), the first time crossing probability distribution along the ray are compared for the synthetic setting of zero and maximum local correlation strengths. If independence is assumed, the ray crosses the iso-contour with much higher probability for the first time at an earlier stage compared to the assumption of maximum correlation. It can be seen clearly that the IFCP values for the second crossing are significantly lower if there is a high PO for the first crossing and vice versa. Secondly, the PPs for each crossing do almost not change. The illustration also reveals the importance of incorporating correlation information in the computation of probabilities for obtaining correct PP and PO values.

4.1.4 Geometric Variability

A major problem in visualizing iso-surface variability is that the effectiveness of the visual perception of the variability depends on the viewing direction. If the uncertainty only reflects in the opacity variation, possible iso-surface displacements, in general, can be only visualized if the viewing direction is nearly orthogonal to the surface normal. Similarly, since the SDF does not contain any information on the spatial distance variability, for a normal-parallel view the accumulation of colors that are mapped from SDF values cannot reveal the spatial iso-surface variation in viewing direction.

Thus, a method is required that preserves the positional uncertainty information independently of the viewing direction. To achieve this, a measure of the spatial iso-surface variation due to the uncertainty is proposed. This measure is then used for assigning colors that emphasize the geometric surface variability.

SDF Surfaces

To reveal the uncertainty that is related to the shape of an iso-surface, shading effects are incorporated into uncertainty volume rendering. Therefore, the vector pointing towards the direction of maximum increase in iso-surface crossing probability is used. Except for the sign, this vector is equivalent to the gradient $\nabla\Psi_\theta(\mathbf{x}_i)$ of the stochastic distance function. Setting $t = \Psi_\theta(\mathbf{x}_i)$, this gradient is located orthogonal to the set

$$\vartheta_\theta(t) := \{\mathbf{x} \in \mathbb{S} : \Psi_\theta(\mathbf{x}) = t\}, \tag{4.24}$$

which contains all points having the same SDF value and, thus, the same stochastic distance from the iso-value. This set is commonly referred to as *SDF surface*.

SDF surface rendering is a common approach for visualizing positional iso-surface uncertainty. For instance, in [ZWK10] $\vartheta_\theta(1)$ and $\vartheta_\theta(-1)$ were used to indicate the spatial region that contains the iso-surface with a probability of 0.68. In the presented novel algorithm the gradient $\nabla\Psi_\theta$ is computed on-the-fly from the SDF values at the grid points, and it is then used for revealing the geometric variability of iso-surfaces via shading effects and for estimating Euclidean distances between the mean surface and possible iso-surface positions.

Spatial Distance Estimation

The positional uncertainty can be perceived in normal-orthogonal viewing direction because the opacity decreases with increasing spatial distance from the *mean surface* $\vartheta_\theta(0)$. This distance mainly varies orthogonal to the ray direction. In normal-parallel direction, however, the spatial distance varies in viewing direction, requiring to use an additional visual representation to show this variation.

To visually encode the spatial distance of a point on a SDF surface $\vartheta_\theta(t)$ to the mean surface $\vartheta_\theta(0)$, first, have a measure has to derived that estimates this distance. Therefore, so called *SDF normal curves* $\gamma_\theta^x : \mathbb{R} \to \mathbb{S}$ are defined for a point $\mathbf{x} \in \mathbb{S}$ using the differential equation

$$\frac{\mathrm{d}\gamma_\theta^x}{\mathrm{d}t}(t) = \nabla\Psi_\theta(\gamma_\theta^x(t)), \quad \gamma_\theta^x(0) = \mathbf{x}. \tag{4.25}$$

Each normal curve crosses *all* SDF surfaces orthogonal. The magnitude of its derivative describes the amount of change in SDF value for an infinitesimal change in the spatial domain. The distance $d(\mathbf{x})$ of point $\mathbf{x} \in \vartheta_\theta(t)$ for $t \neq 0$ from the mean surface is now defined as the length of the normal curve γ_θ^x between $\mathbf{x}$ and its intersection point with $\vartheta_\theta(0)$ in 3D space.

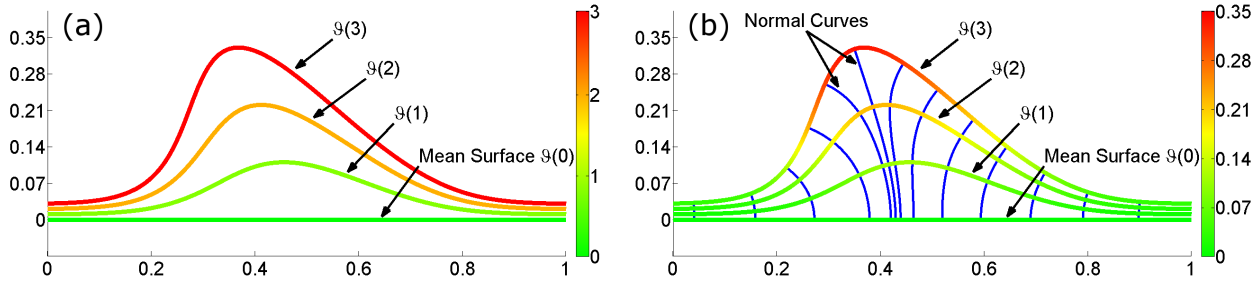

Figure 4.8: This illustration shows the mean iso-contour (green) of a 2D data set, as well as three positive SDF contours. SDF normal curves (blue) are displayed for several points on the $\vartheta_\theta(3)$ contour. (a) Points on the SDF contours are color-coded with respect to the respective SDF value. (b) Points on the SDF contours are color-coded with respect to the length of their normal curve to the intersection point with $\vartheta_\theta(0)$ — from green (small distance) to red (large distance). The magnitude of the numbers on the axes and on the color bar are related to Euclidean distances in the 2D domain.

The "length" of the curve in SDF space is actually the SDF difference $|\Psi_\theta(\mathbf{x})|$ between $\vartheta_\theta(\mathbf{x})$ and $\vartheta_\theta(0)$. Thus, one obtains the equation

$$\int_0^{d(\mathbf{x})} \|\nabla\Psi_\theta(\gamma_\theta^x(t))\| \, \mathrm{d}t = |\Psi_\theta(\mathbf{x})| \,. \tag{4.26}$$

For simplicity, it is assumed that $\|\nabla\Psi_\theta\|$ is constant along the considered curve segment, which means a linear increase/decrease in iso-surface crossing probability along the SDF normal curve. Even though this assumption could be violated for large values of $|\Psi_\theta(\mathbf{x})|$, it should be quite reasonable for SDF surfaces close to the respective mean surface. Finally, the distance can be estimated as

$$d(\mathbf{x}) = \frac{|\Psi_\theta(\mathbf{x})|}{\|\nabla\Psi_\theta(\mathbf{x})\|}, \tag{4.27}$$

which provides a good indication of how strong a point $\mathbf{x}$ is deviated from the mean surface $\vartheta_\theta(0)$.

The introduced concept of utilizing normal curve lengths for a spatial variability indication is illustrated in Fig. 4.8. This illustration shows a flat mean iso-surface (green) of a 2D data set, as well as three positive SDF surfaces. SDF normal curves (blue) are displayed in (b) for several points on the $\vartheta_\theta(3)$ surface.

In (a), a traditional color mapping scheme was used, assigning colors to SDF surfaces according to the respective SDF value. The variability in iso-surface positioned can be only seen in orthogonal viewing direction with respect to the mean surface normals. In normal viewing direction the color order (red, orange, light green, green) is the same for all regions of the mean surface, independent of the local spatial variability.

In (b), points on the SDF surfaces are color coded with respect to the length of their normal curve

to the intersection point with $\vartheta_\theta(0)$ — from green (small distance) to red (large distance). The magnitude of the numbers on the axes and on the color bar are related to Euclidean distances in the 2D domain. This visualization clearly highlights the region with the largest spatial variability (red central part) — independent of the viewing direction.

4.1.5 Visualization

For visualizing the positional and geometrical variability of a particular iso-surface in an uncertain 3D scalar field, front-to-back volume ray-casting is used, while computing the opacity according to the IFCP approach, in which the iso-value is interactively specified by the user using a slider. At each sample point $\mathbf{s}_i$, the spatial distance estimate $d(\mathbf{s}_i)$ is mapped to a HSV color value, with the distance being encoded into the *hue channel*. Based on a user-defined maximum distance d_{max}, the range $[0, d_{max}]$ is linearly mapped to the color map [green $\rightarrow$ yellow $\rightarrow$ red], and values greater than d_{max} are clamped to d_{max}. To integrate shading effects and, thus, to highlight the shape of the SDF surface, the cosine between the ray direction and the SDF gradient $\nabla\Psi_\theta$ is used to modulate the *value* of the HSV color sample. Here a value reduction of at most 50% is allowed.

Color-coding the spatial distance from the mean surface along the SDF normal curves allows for an intuitive perception of the iso-surface variability in normal-parallel viewing direction. However, this approach has a drawback when the iso-surface is cut by a slicing plane. In this case the color distribution is somewhat misleading since "color iso-contours" do not necessarily represent a single SDF value. Therefore, on slicing planes a color scheme based on SDF related measures, e.g., as proposed in [PH11], gives a better impression of the spatial variability.

In order to overcome this limitation, SDF iso-contours are displayed on slicing planes, indicating its intersection with the SDF surfaces $\vartheta(0)$, $\vartheta(\pm1)$, $\vartheta(\pm2)$, etc. with a ±0.1 tolerance. The SDF magnitude is encoded into the blue channel using the LCP value introduced in [PH11]. As illustrated in Fig. 4.9, this gives a clear impression of the relation between SDF values and spatial distances on slicing planes.

In section 4.1.3, the concept of probabilities of occasion and position were discussed. In order to integrate this concept into the visualization, areas with high PO but low correlation to the mean surface are desaturated. For each sample point, the spatial distance $d(\mathbf{s}_i)$ is used for computing a correlation coefficient in $\nabla\Psi_\theta$ direction using the ECF. This coefficient is linearly mapped to *saturation*. Thus, areas with a high spatial distance and a low correlation to values on the mean surface receive a lower saturation, indicating their independence and significance as PO area. In this way the user can visually differentiate between possible iso-surface positions with high positional distribution on the one hand, and regions which might contain an iso-surface but are stochastically independent from those PP areas on the other hand.

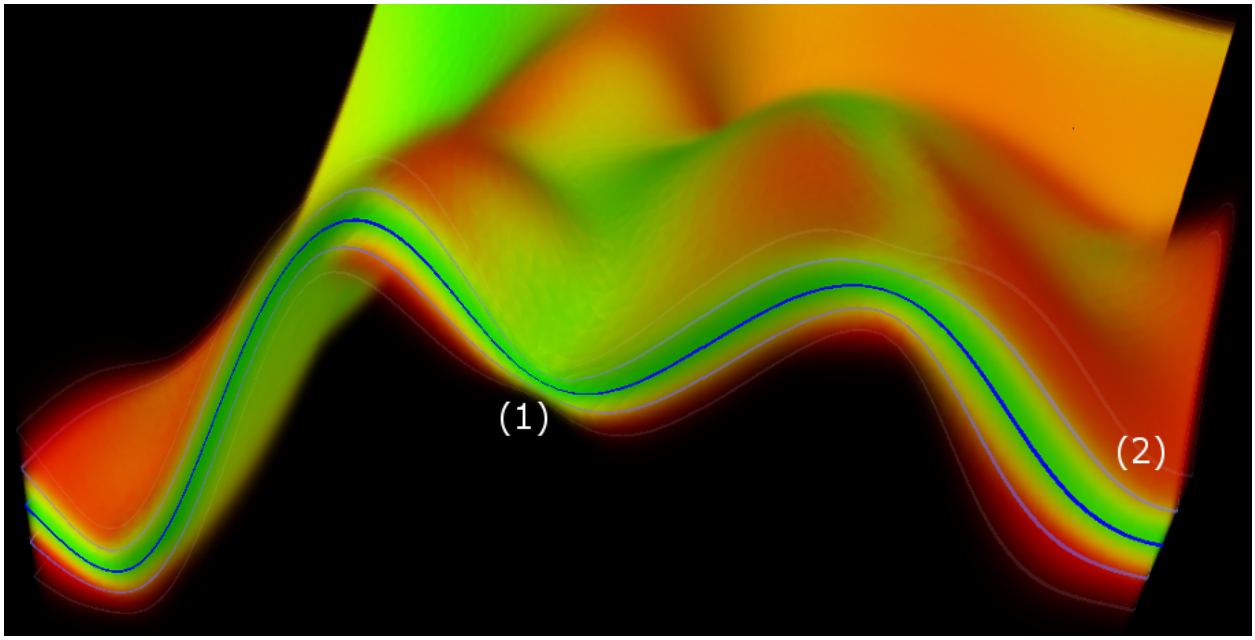

Figure 4.9: An uncertain 3D signed distance field to a 2D topographic height map is shown. Multiple instances to a randomly displaced height map were generated, and the mean and standard deviations where computed from these instances. SDF iso-contours on slicing planes indicate the set $\vartheta_\theta(0)$ (blue) and the sets $\vartheta_\theta(\pm i)$ with decreasing opacity and saturation. Note the relation between converging iso-contours and low spatial distance (green) in (1) and between diverging contours and high spatial distance (red) in (2).

In addition, SDF surfaces can be rendered as add-on to the IFCP approach. A SDF value can be specified interactively by the user and the visualized surface can help to analyze stochastic geometrical and topological changes or to visually link stochastic and spatial distances. Further interaction mechanisms include the specification of the maximum opacity by controlling the respective number of particles λ_{max} (cf. 4.1.3) as well as the range $[0, d_{max}]$ of the spatial distance, which is mapped to the given color map.

4.1.6 Results and Analysis

In the following, the validity and effectiveness of the proposed IFCP algorithm is demonstrated in combination with the uncertainty visualization methods that have been introduced in the last sections. Experiments have been performed on synthetic and real-world data sets given on tetrahedral grids.

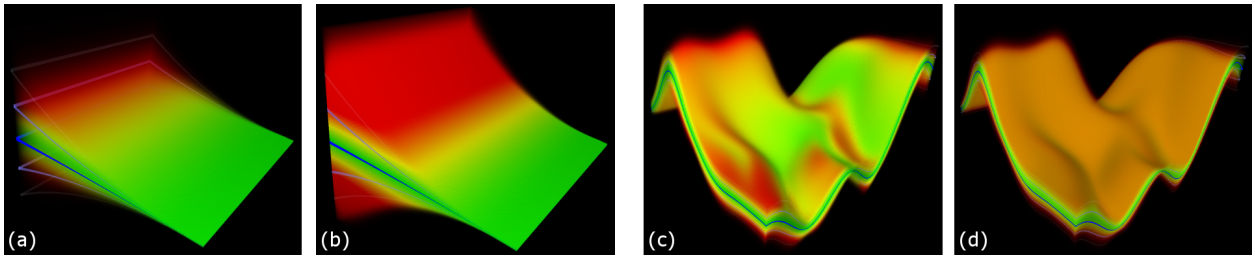

Figure 4.10: Visualizations of iso-surface crossing probabilities in uncertain 3D scalar fields are shown. (a) The novel IFCP algorithm accounts for correlations in the data (absorption coefficient is proportional to probability, increasing distance from mean-surface is color-coded from green to red). (b) If correlations are not considered, probabilities are vastly overestimated. (c) The proposed color-coding reveals spatial distance along normal curves to the mean surface (increasing distance is coded from green to red). (d) Conventional color-coding does not respect spatial distance but stochastic distance measured in number of standard deviations.

Synthetic Simulation Data Sets

The PLATE data set was generated by simulating a vibrating metal sheet, which was fixed at one end. The plate was modeled as cosh-function with one degree of freedom. Several different realizations of the plate were generated by assigning different random values to the free parameter. The overall mean and standard deviations where estimated using all generated samples. In Fig. 4.10 (a) the PLATE dataset was visualized using the IFCP algorithm (including distance dependent correlation) in combination with distance-based color coloring and additional SDF contours. In (b) the same visualization is shown, but the data was supposed to be uncorrelated ($\rho = 0$). It can be seen clearly, that in (b) the iso-surface crossing probability is strongly overestimated. This result would have been achieved using a ray-casting approach with high sampling rate and distance independent correlation. Only the correlation-aware IFCP algorithm allows a visualization of the positional iso-surface uncertainty independent of the chosen sampling rate along the rays.

The TOPO data set consists of a 3D scalar field in which an implicit iso-surface represents a height surface (topography) over a 2D domain. Multiple instances, in which the iso-surface was randomly deviated, were generated, and the mean and standard deviations were computed from these instances. Fig. 4.10 (c) shows a visualization of the TOPO data set using the IFCP algorithm and spatial distance coloring. In (d) the same approach was used, but the SDF values were linearly mapped to color. In both examples, the visualization of SDF iso-contours reveal the positional variability of the iso-surface on the slicing planes. In the inner part and in more gradient-parallel viewing direction the novel approach can clearly emphasize the amount of spatial deviation from the mean surface in convex as well as concave areas. It is worth noting that methods which make use only of the SDF values to determine confidence regions around the mean surface or use SDF-related norms for color-coding cannot achieve a similar result, as the SDF contains no spatial

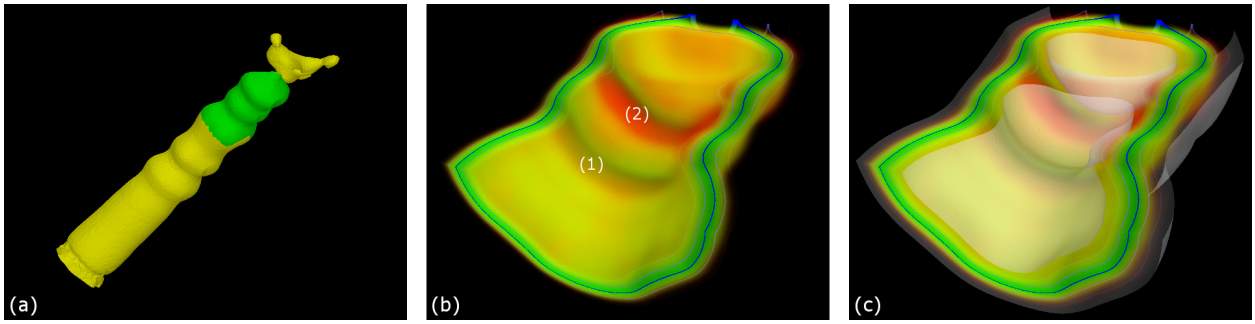

Figure 4.11: (a) Direct volume rendering of the fuel injection data set. The green marked section is analyzed regarding to iso-surface uncertainty. (b) Compared to (1), a higher spatial deviation from the mean surface is revealed in the convex deflection (2). (c) The $\vartheta_\theta(3)$ SDF surface emphasizes topological uncertainty due to the separation of iso-surfaces.

distance information. Especially in gradient-parallel viewing direction such methods typically fail in visualizing the uncertainty with respect to shape. The relationship between the SDF iso-contours and spatial distance colors is illustrated in Fig. 4.9.

Fuel Injection Data Set

In this experiment, a specific part of the fuel injection data set [Ger] shown in Fig. 4.11 (a) is analyzed. In particular, the visualization is meant to demonstrate the suitability of the presented approach for detecting topology changes due to uncertainty. Fig. 4.11 (b) shows the iso-surface crossing probability for the selected green part in (a), which results from applying a constant standard deviation to all data points. The red-colored second convex deflection (2) reveals a much stronger spatial deviation of the iso-surface from its mean position than the first deflection (1). The reason for this difference can be deduced from (c), where the $\vartheta_\theta(\pm 3)$ SDF surfaces are simultaneously visualized. As can be seen, the uncertainty in the surface topology results from the separation of the iso-surface into two distinct structures. Thus, the proposed novel color mapping scheme in combination with SDF surface rendering can be used effectively to indicate how reliable an extracted iso-surface is.

Atmospheric Temperature Forecast Ensemble Data Set

The ATMOS data set shows a 3D temperature field in the exosphere above Europe and the North Atlantic Ocean. It contains the mean values of multiple fields that were simulated by the European Centre for Medium-Range Weather Forecasts (ECMWF) using different input parameters [Eur].

Fig. 4.1 (d) shows a visualization of an uncertain iso-surface in the ATMOS data-set using the IFCP algorithm. SDF values were linearly mapped to color. In (c) the same approach was used,

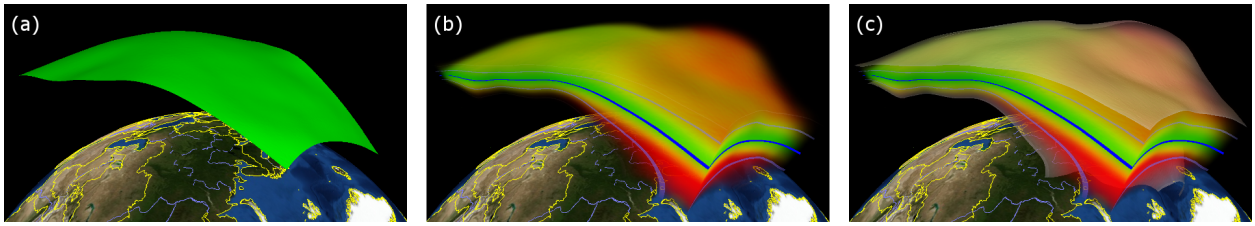

Figure 4.12: (a) The mean surface for a temperature iso-value in the ATMOS data set is shown. (b) The IFCP algorithm in combination with distance dependent color mapping is applied. (c) SDF surfaces $\vartheta(\pm 2)$ emphasize the uncertainty in iso-surface shape.

but SDF values were mapped to color based on their spatial deviation from the mean surface along normal curves. By visualizing SDF iso-contours on a slicing plane, the positional variability of the iso-surface can be revealed locally in either case. Globally, and where the viewing direction is parallel to the normal of the iso-surface, the novel color mapping approach (c) emphasizes the strength of the spatial deviation from the mean surface in convex *and* concave regions. Since SDF values do not contain any spatial distance information, the visualization in (d) fails in depicting the uncertainty with respect to the shape of the mean surface.

The potential of the proposed uncertainty visualization techniques for analyzing the effect of uncertainty on specific data features is demonstrated in Fig. 4.12. In (a), the surface for a given iso-value is shown. In (b), the IFCP algorithm and the spatial distance coloring were used under the assumption of homogeneous correlation ($\rho_i = 1$). The confidence volume containing the iso-surface with a certain probability is enclosed by two stochastic distance surfaces in (c). The uncertainty visualization highlights the region where the simulated temperature field is rather sensitive to the input parameters of the simulation, indicating that the forecast for this region is not reliable. This information suggests to improve either on the physical forecast model, the used initial conditions, or the employed computational scheme to obtain a less sensitive and, thus, more reliable result.

Geophysics Data Set

This data set was generated by seismic tomography [KÏ0], where recorded seismic waves are used for inferring structures in the earth's mantle. In the visualization examples, the boundary surface between two specific earth materials is analyzed. As the computations are performed on incomplete and noisy data, the position of the iso-surface is affected by uncertainty.

Fig. 4.13 illustrates the difference between reliable (a) and overestimated (b) crossing probabilities depending on whether correlation has been considered or not. Note that the visualization in (a) turns into (b) if the sampling rate in the ray-casting approach is increased without using an adequate distance dependent correlation model. Only the correlation-aware IFCP algorithm allows a visualization of positional iso-surface uncertainty independent of the chosen sampling rate along

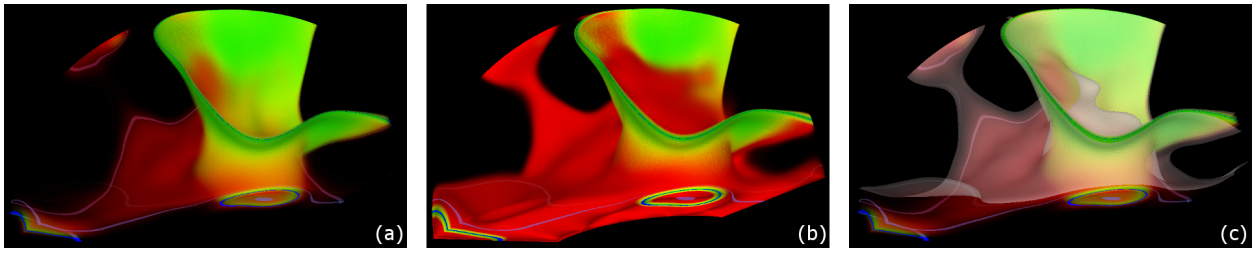

Figure 4.13: (a) A seismic tomography data set is shown. (a) Reliable positional probabilities for the crossing of iso-surfaces along the rays of sight are computed using distance dependent correlations. (b) If correlations are not considered, crossing probabilities are vastly overestimated. (c) The uncertainty related to shape is revealed by SDF surface integration.

the rays. In (a), the lower probability of crossing an iso-surface in the red area is clearly visible due to the assignment of lower opacity. The $\vartheta_\theta(1)$ SDF iso-contour (violet) gives a good impression of the geometrical shape of the related SDF surface. In (c), the perception of the geometrical variability is enhanced by visualizing the $\vartheta_\theta(2)$ SDF surface.

Fig. 4.14 shows the same data set from a different viewpoint and for a slightly different iso-value. In (a), the IFCP algorithm and the spatial distance color mapping were applied under the assumption of homogeneous correlation ($\rho = 1$). The visualization in (b) shows the same setting but using an anisotropic correlation with a low ρ-value in a certain direction. In areas (1) and (2) an iso-surface crossing is now detected with a much higher probability than in (a), clearly demonstrating the benefit of using saturation to encode the correlation. Although a high crossing probability is visualized, the low saturation indicates that this is caused by a low correlation in this area and, therefore, high stochastic independence and high probability of occurrence (PO) but *not* a high positional variability (PP).

An example which demonstrates the suitability of the presented approach for detecting possible topology changes due to uncertainty is shown in the visualization of the GEO data set in Fig. 4.15. In (a), the mean iso-surface, separating subsurface structures with high and low seismic wave velocities in the earth mantle below Australia, is visualized. If the geometric variability of the separating surface due to uncertainty (constant correlation of $\rho_i = 1$) is visualized (b), additional "bridging" structures occur with a certain probability (1). Notably these structures and their geometrical appearance cannot be detected if uncertainty is ignored or only used to color the mean surface. To facilitate an improved assessment of the uncertainty related to the shape of an iso-surface, and to employ the possibilities of a surface structure for integrating shading effects, the SDF surfaces $\vartheta(\pm 2)$ are incorporated into the visualization in (c).

In (d), with respect to (b), an anisotropic correlation decrease is assumed. The result is a significantly higher surface crossing probability along the view rays in area (2). In this example,

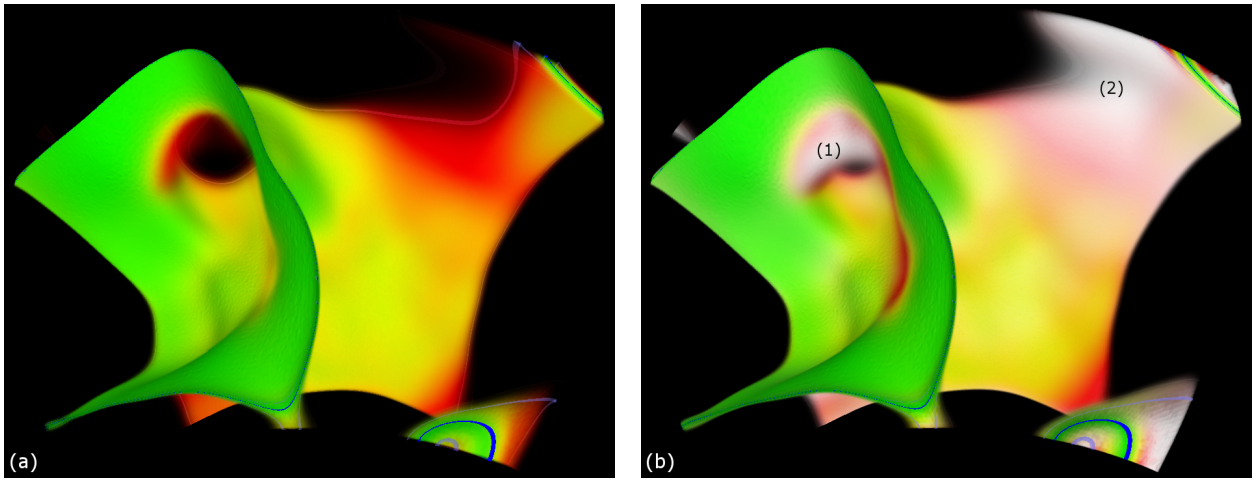

Figure 4.14: A seismic tomography data set is rendered with IFCP opacity mapping and spatial distance estimation (correlation of $\rho = 1$ is assumed). (b) An anisotropic correlation distribution is assumed (with respect to (a) correlation decrease in one spatial direction). Higher crossing probability are determined in (1) and (2) compared to (a), but correlation-based saturation reveals high local stochastic independence rather than high local positional iso-surface variability as major cause.

saturation was chosen as an indication of high distance from the mean surface and low correlation. Thus, the specific color-coding reveals high local stochastic independence rather than high local positional surface variability as major cause of the crossing probability in (2).

This information is important for analyzing the relationship between uncertainty parameters, like standard deviations and correlation structures. Especially in seismic tomography, high correlations are assigned to areas with high uncertainty for regularization purposes, and these correlations are then incorporated into a prior stochastic model. In this way, areas with high PO concurrency, like in (d), can be avoided and uncertain data sets obtained, clearly showing the positional variability like in (b). The visual analysis of the effect of such prior stochastic models is of great importance for steering the tomography process towards reduced uncertainty. The visual integration of the PO/PP concept, as discussed in 4.1.3, is another new strong benefit of the proposed methods, since it enables differentiating between regions with high correlation and, therefore, high positional uncertainty, and areas with high stochastic independence and, therefore, high occurrence uncertainty.

4.1.7 Implementation and Performance Details

All of the presented results were rendered into a 1K $\times$ 768 view-port using volume ray-casting on the GPU as proposed in [Wei05]. Tetrahedral elements — represented by four indices to their vertices — were stored in an element buffer. Each element was accompanied by a correlation tensor and four additional links to their neighboring elements. Shared vertex buffers were used to store the

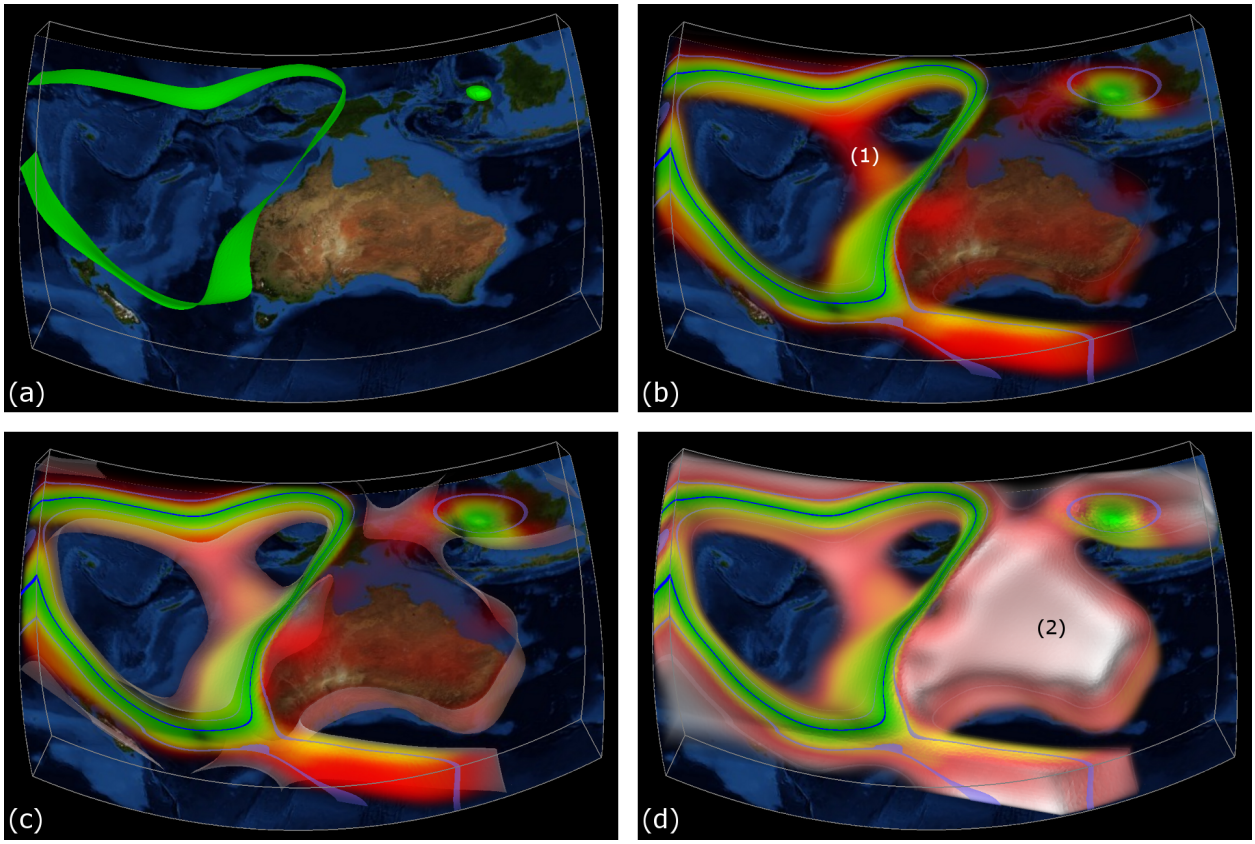

Figure 4.15: (a) A separating iso-surface in a seismic tomography data set is shown. (b) The IFCP algorithm in combination with distance dependent color mapping (homogeneous correlation of $\rho_i = 1$) reveals a possible topological link in (1). (c) SDF surfaces $\vartheta(\pm 2)$ emphasis the iso-surface uncertainty with respect to shape. (d) An anisotropic correlation decrease is assumed. Compared to (b), higher crossing probabilities are determined in (2), but correlation based saturation reveals high stochastic independence rather than high local positional iso-surface variability as major cause.

vertex coordinates as well as the mean values and standard deviations. View rays were traversed with a given sampling distance by subsequently computing ray-cell intersections and following the respective link to the next element. Barycentric interpolation was used to reconstruct a continuous field, and early-ray termination was performed at an optical attenuation above 0.95.

When executing the IFCP algorithm, the update operations in Equ. (4.18) are performed at every sample interval along the view rays. This requires evaluating and interpolating per-vertex SDF values, and evaluating the probabilities $\mathcal{P}(Y_i^-)$ and $\mathcal{P}(Y_i^- \cap Y_{i+1}^-)$ as stated in Equ. (4.20). To reduce the computations required for evaluating the distribution functions Φ_1 and Φ_2, and to avoid precision problems caused by real number divisions in Equ. (4.18), function values $\Phi[a, b, \rho] = \Phi_2(a, b; \rho)/\Phi_1(a)$ were pre-computed, e.g., as described in [AS64, DW90], and stored in a 256^3 texture map with $a, b \in [-5.08, 5.08]$ and $\rho \in [0, 1]$. This texture map is used at run-time as a lookup

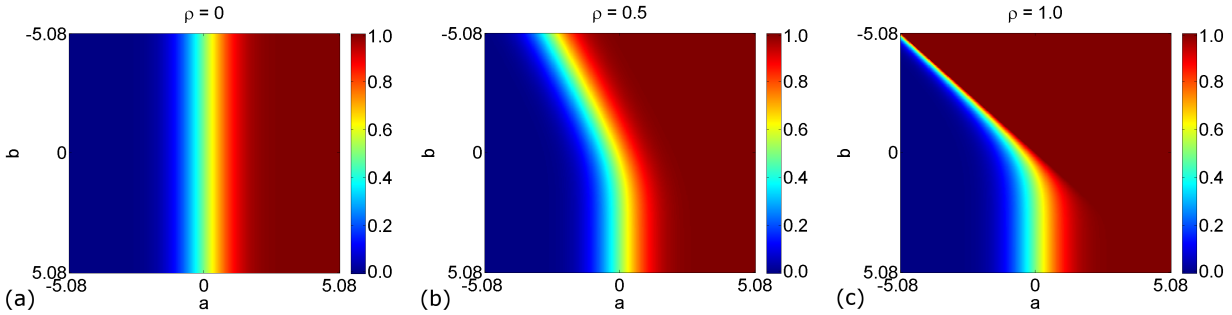

Figure 4.16: Three sample levels of 3D probability quotient lookup texture are shown for correlation values $\rho = 0$ (a), $\rho = 0.5$ (b) and $\rho = 1$ (c).

table, resulting in an approximation error below floating point precision. In Fig. 4.16, three sample levels of the 3D texture are shown, displaying in 2D maps the values $\Phi[a, b, \rho] = \Phi_2(a, b; \rho)/\Phi_1(a)$ for correlation settings $\rho = 0$, $\rho = 0.5$ and $\rho = 1$.

On the used target architecture, a 2.83 GHz Core 2 Quad processor equipped with a NVIDIA Quadro FX5800, the IFCP algorithm roughly doubles the visualization time per frame compared to iso-surface ray-casting. For instance, the visualization of the iso-surface in the GEO data set (3.3 million tetrahedral elements) requires 85 milliseconds (ms), while the IFCP algorithm takes about 150 ms. 80 MBytes are required to store the mean values, the standard deviations, and the correlation tensors, plus additional 65 MBytes consumed by the pre-computed 3D look-up table. The loss of performance in the IFCP algorithm is mainly due to the enlarged memory footprint for computing and interpolating SDF values and gradients on the fly, and for accessing the pre-computed distribution functions.

4.1.8 Conclusion

In the preceding sections, a novel approach has been presented for computing reliable probabilities of position and occurrence for iso-surface crossings in uncertain 3D scalar fields. This has been achieved by incorporating distance dependent correlations into the suggested approach. An efficient update-scheme allows integrating the proposed algorithm into front-to-back ray-casting. Due to the integration of the distance dependent correlation model, the crossing probability computation is independent of the specified sampling rate along the rays. A new measure for estimating the distance between possible iso-surface variations and the mean surface has been developed. This measure is used as input for a color mapping scheme, which allows for an effective visualization of iso-surface variability independent of the viewing direction. It has been demonstrated that the novel approach results in an intuitive understanding of the effect of uncertainty on iso-surfaces in 3D scalar fields.

4.2 Arbitrary Distributed Ensemble Data Sets

Ensemble data sets consist of different instances of the outcome of an experiment. They can be produced by repeated simulations of a parameterized model using slightly modified parameter settings, or by simply running multiple times a procedure which is affected by inherent randomness. When the experiment generates outcomes over a spatial domain, each instance, or ensemble member, captures the possible data values at all locations in space.

Analyzing ensemble data sets is important, because they give answers about the uncertainty of the captured data, i.e., the degree of variation of that data. These answers are usually derived by assuming a parametric uncertainty model, meaning that the instances are realizations of a multi-variate random variable with a *known* probability distribution model. Most often a multi-variate Gaussian probability distribution is assumed, for which the parameters like mean, standard deviation, and correlation are estimated from the ensemble data set. For Gaussian distributed random fields with well-defined parameters, a number of techniques can effectively visualize the possible variations of specific features like iso-contours or critical points in such fields. Although the assumption of a Gaussian distribution is often valid, various types of distributions can occur in general. This asks for a methodology to a) test the hypothesis of a Gaussian distribution, and b) estimate the variability of features for non-Gaussian, or, even more general, non-parametric distributions.

It is clear, on the other hand, that an uncertainty visualization of any feature can always be performed by simply overlaying plots of features in individual ensemble members. One popular such approach is the so-called *spaghetti plot*, which shows simultaneously all iso-contours in the members of a 2D scalar ensemble data set. Especially in meteorology and geoscience, such plots have been positioned as a key uncertainty metaphor (cf. [PWB*09b, SZD*10]). In Fig. 4.17 (a) and (d) the spaghetti plots of two different scalar ensembles are shown for the same iso-value θ. The plots do not indicate any difference between the two data sets.

Images (b) and (c) show two instances of the first ensemble in (a). The data is color-coded from blue (below θ) over white (equal to θ) to red (above θ). The iso-contour "jumps" from the right branch in (b) to the left one in (c), and the data values increase from right to left in both instances. By looking at all ensemble members, one would see that in the lower part of the domain the iso-contour is positioned *either* in the left *or* in the right branch, but never in both. Thus, the contour distribution is uni-modal in the upper part (one major peak) and bi-modal in the lower part. Images (e) and (f) show two members of the second ensemble. Now, in all ensemble members, the iso-contour occurs *simultaneously* in both the left and the right branch, as a v-shaped valley exists in the middle of the domain. These examples make clear that visualizing features in individual ensemble members in one image can be greatly misleading and does not allow for a reliable estimation of the feature

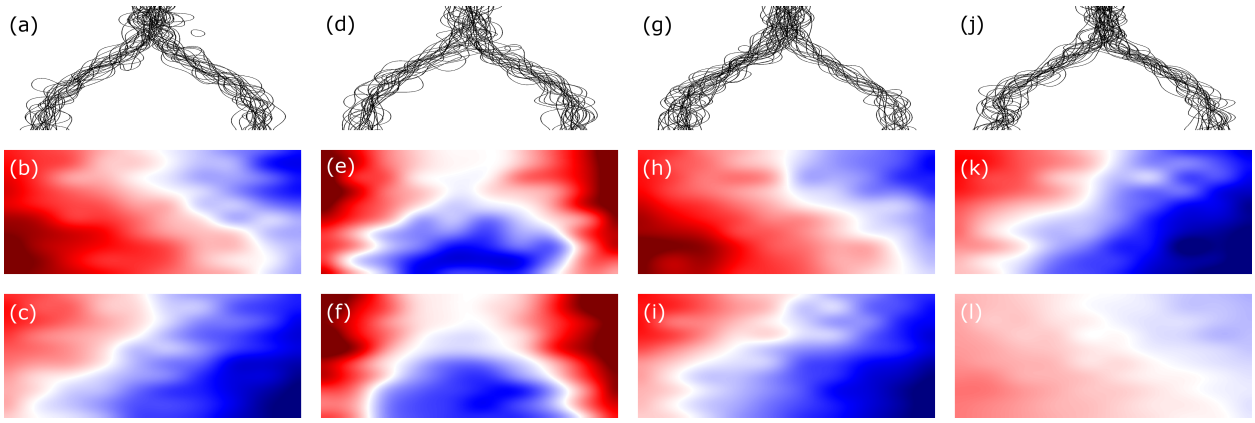

Figure 4.17: Top row: Spaghetti plots of iso-contours for four ensemble data sets are shown. Bottom rows: Two color-coded scalar ensemble members are shown for each of the four ensemble data sets in the top row. For a specified iso-value θ, the data is color-coded from blue (below θ) over white (equal to θ) to red (above θ).

uncertainty in general. These limitations are due to the fact that each individual feature is taken out of the data without respecting any stochastic model of the feature distribution.

Furthermore, as shown by a further example in images (g), (h), and (i), a quantitative differentiation of the local spatial probability density of the contours is not possible. In this example, the iso-contour is twice as many times located in the left than in the right branch. It is clear, that the little visible change in (g) becomes more and more subtle due to cluttering when more ensemble members are used. Therefore, no quantitative analysis with respect to the contour distribution can be made solely based on the spaghetti plot visualization in (g).

In image (j), the spaghetti plot for a similar data set like in (a) and (g) is shown. The iso-contour permutes again between the right and left branch. However, the data gradient strengths are much lower along the iso-contours in the right branch, as the transition in scalar value is much smoother and shallower in the ensemble members, whose iso-contour position belongs to the right branch. The difference between strong and weak gradient strengths is shown in the two ensemble members (k) and (l), respectively.

This section presents a new visualization technique for iso-contours in 2D scalar ensemble fields which overcomes the depicted limitations of spaghetti plots. It makes no assumption about a stochastic uncertainty model, yet by computing a statistical summary of the ensemble over the spatial domain it generates point-wise measures for the likelihood of occurrence of iso-contours. The approach is similar in spirit to kernel density estimates (KDE). KDE is a non-parametric approach for estimating a continuous *probability density function* (PDF) from a finite set of data values [Ros56, Par62]. Instead of creating a PDF for *scalar values* in a 1D data domain, the concept underlying KDE is used to obtain a PDF for *iso-contours* in a 2D spatial domain. In

this novel approach, single iso-contours are represented by Gaussian kernels. However, instead of modeling the kernels directly using the analytic KDE approach, the spatial PDF is obtained indirectly using spatial derivatives of a *cumulative distribution function* (CDF). In this way, any explicit computation and storage of point-wise data histograms can be avoided.

The novel technique is suitable for arbitrary ensemble data sets, and it can be used to identify whether the contours follow a Gaussian-like or a multi-modal distribution. Furthermore, a visual concept is introduced for depicting the modality of the iso-contour distribution and indicating in which parts of the domain the uncertainty cannot faithfully be represented by a Gaussian-like distribution. In combination with a special color scheme, the computed PDFs can be used to enable a quantitative assessment of the iso-contour distribution as well as the variations in gradient magnitude around these contours. Since the entire visualization is carried out on the GPU, the approach allows the user to visually analyze even very large ensemble data sets at interactive rates.

4.2.1 Related Work

The visual exploration of ensemble data sets has been discussed in several publications: [LPK05] give an overview of general approaches on visualizing spatial multi-value data. In [BKS04], a special clustering approach is used for reducing the information content in ensemble data sets. This condensed data is then used for color-coding purposes. [PPA*12] use a pairwise sequential animation method and a screen space subdivision approach for visually analyzing ensemble data sets. In [AWH*12], the Ensemble Surface Slicing algorithm was introduced, which integrates several 3D surfaces simultaneously in one single-image view.

The visualization of uncertain iso-contours in 2D ensemble data sets was explicitly addressed in [PWB*09b]. Although no assumption on any distribution is made (e.g., Gaussian distribution), they propose a framework for interactively flipping through the ensemble members. Iso-contours are visualized using the spaghetti plot concept. This approach was improved in [SZD*10]. The spaghetti plots were enhanced by additional glyphs and confidence ribbons, highlighting the Euclidean spread of the contour ensemble. However, the use of parameters like standard deviation and inter-quartile ranges for describing the uncertainty indirectly assumes a certain stochastic probability model. In [FKLT10], kernel density estimates are used for visualizing the contour density in parallel coordinate plots. In [AOB08], uncertain contours in 2D scalar data sets are qualitatively visualized using visual noise and fuzzy effects.

4.2.2 Spatial Probability Distribution

In the following, the stochastic functions are introduced, which are required for characterizing the spatial variability of iso-contours in uncertain scalar fields, i.e., the *spatial* CDF and PDF. Since no parametric uncertainty model is assumed, first, a new concept is presented for deriving these functions for iso-contours in scalar ensembles. These functions are then used to improve spaghetti plots towards a more reliable visualization of uncertain iso-contours.

Spatial Distribution Functions

A discrete sampling of a continuous 2D domain $\mathbb{S} \subset \mathbb{R}^2$ on a Cartesian grid structure with grid points $\mathbb{S}_{a,b} = \{\mathbf{x}_{i,j} : 1 \leq i \leq a, 1 \leq j \leq b\}$ is assumed. However, since all computations are local at each point and do not require any information about neighboring points, the grid type can be arbitrary in principle. A Cartesian grid structure is chosen, so that the generated data can be stored in a 2D texture map and rendered efficiently on the GPU. It is further assumed that an ensemble has n members, the k-th member containing scalar values $y_k(\mathbf{x}_{i,j})$ and gradients $\nabla y_k(\mathbf{x}_{i,j})$. At each grid point, the *data uncertainty* is given by the variability of the n possible data values. The data uncertainty can be modeled by assigning a *random variable* $Y(\mathbf{x}_{i,j})$ to each grid point $\mathbf{x}_{i,j}$.

In the following, it is assumed that iso-contours for a selected iso-value θ in the ensemble members should be analyzed. Instead of drawing a spaghetti plot, one is interested in determining for every grid point the probability that a contour is located exactly at this point. At a point $\mathbf{x}_{i,j}$, the probability that the data takes on the value θ is $\mathcal{P}(Y(\mathbf{x}_{i,j}) = \theta)$. Unfortunately, this probability vanishes if the data is real-valued and smooth. This is due to the fact that a 1D contour (or 1-manifold) in a 2D domain has a zero Lebesgue (or Borel) measure, i.e. its area is zero (cf. 2.1.2). Since a non-zero probability for the occurrence of a contour is only possible across a non-zero area, it must hold that $\mathcal{P}(Y(\mathbf{x}_{i,j}) = \theta) = 0$.

Since it is not possible to directly compute probabilities of the occurrence of iso-contours at the domain points, a novel alternative is introduced. Instead of interpreting an iso-contour as a 1-manifold embedded into the 2D domain, it is regarded as the boundary between the regions containing all points with scalar values above and below the iso-value. These regions are called the *superlevel* and *sublevel sets*, respectively, and have non-zero Lebesgue measures in general. They can thus be used to derive probability measures.

For this purpose, the spatial CDF is introduced as

$$\Psi_\theta(\mathbf{x}_{i,j}) := \mathcal{P}(Y(\mathbf{x}_{i,j}) \geq \theta) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{P}_k(Y(\mathbf{x}_{i,j}) \geq \theta), \tag{4.28}$$

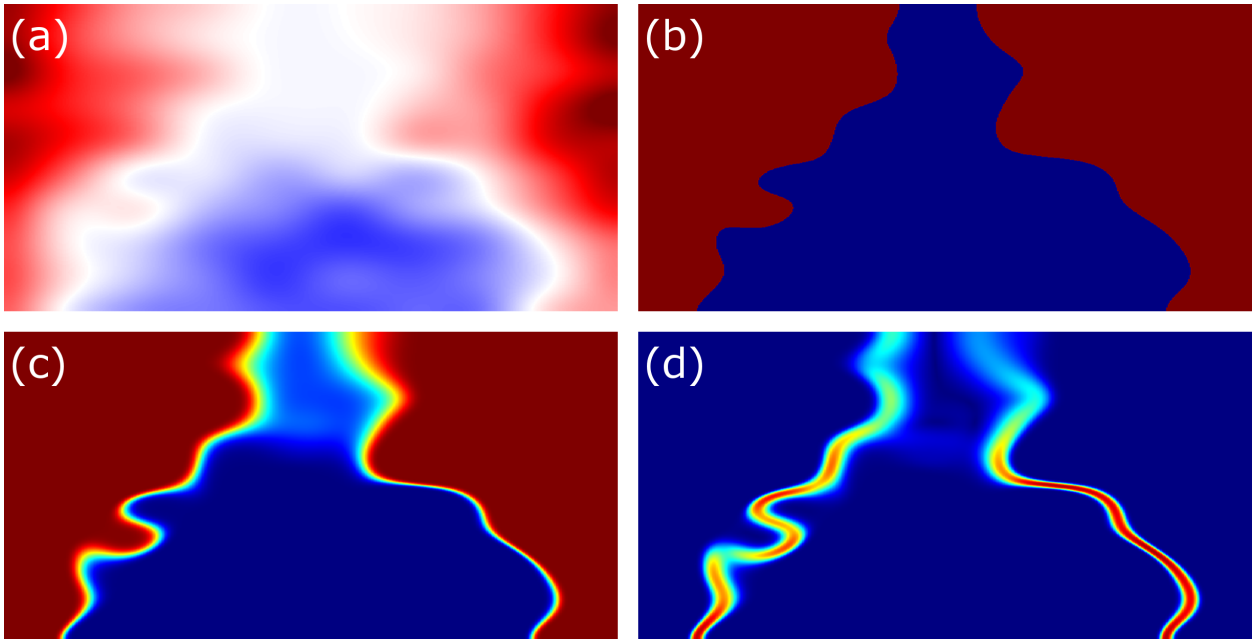

Figure 4.18: (a) One ensemble member of the data set in Fig. 4.17 (d) is shown. (b) Spatial CDF is shown for single ensemble member in (a) using the binary indicator transition function. (c) Spatial CDF using the Gaussian transition function. (d) Spatial PDF as derivative of the spatial CDF.

where $\mathcal{P}_k$ is a boolean indicator function $\mathbb{1}_k$ for each ensemble member k:

$$\mathcal{P}_k(Y(\mathbf{x}_{i,j}) \geq \theta) = \mathbb{1}_k(\mathbf{x}_{i,j}) := \begin{cases} 0 & \text{if} \quad y_k(\mathbf{x}_{i,j}) < \theta \\ 1 & \text{if} \quad y_k(\mathbf{x}_{i,j}) \geq \theta \end{cases} \tag{4.29}$$

The spatial CDF expresses the probability that the scalar value at a particular grid point is greater than the chosen iso-value θ with respect to the data uncertainty at that point. It is defined for any number n of ensemble members. In Fig. 4.18 (b), Equ. (4.29) was applied to the single ensemble member shown in (a). A color table was used to map values in the interval $[0, 1]$ to a color range from blue to red.

Since a CDF is obtained by integrating the respective PDF, the spatial PDF can be computed via differentiation as

$$\psi_\theta(\mathbf{x}_{i,j}) := \|\nabla \Psi_\theta(\mathbf{x}_{i,j})\| \,. \tag{4.30}$$

The following paragraphs will show for the spatial PDF that a) it can be computed at each grid point by using only the ensemble data at this point, and b) it can be used directly to encode the positional uncertainty of the iso-contours.

The concept of the spatial PDF and CDF is illustrated in Fig. 4.19. In (a), the spaghetti plot for iso-contours in a 2D scalar ensemble data set is shown. The different colors of the contours

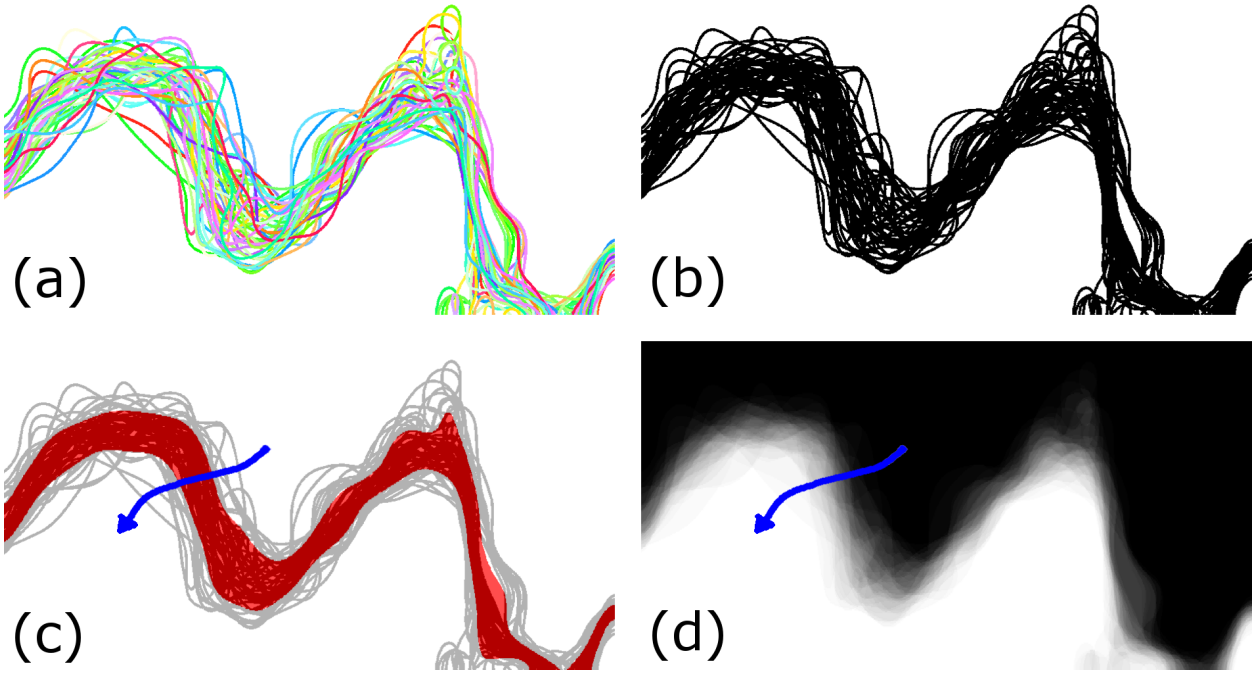

Figure 4.19: (a), (b) Spaghetti plots are shown for an ensemble of iso-contours. Different contour colors are used for different ensemble members in (a). (c) The confidence region for the contour variability is shown in red. (d) Spatial CDF values are encoded from $[0, 1]$ to gray-scales from black to white. A normal curve (blue) is shown in (c) and (d).

represent the different ensemble members. In (b), only the single color black is used for all iso-contours. This visualization shows that for an increasing number of contours a quantitative analysis of the contour distribution becomes more and more difficult due to severe cluttering effects. Using the estimation of traditional uncertainty indicators like and mean and standard deviation allows the computation of a confidence region for the iso-contour distribution. This region (red) is illustrated in (c). Although it provides a good indicator for the overall uncertainty strengths, a more detailed analysis of the positional occurrence probabilities and the distribution modality is not possible.

The spatial CDF and PDF only model the distribution of the iso-contour along the so-called *normal curves* of the CDF field (cf. Fig. 4.19 (d)). These normal curves can be obtained for any spatial domain point, are always aligned orthogonally to the iso-contour distribution and oriented in increasing spatial CDF value direction. The normal curve $\gamma_\theta^{\mathbf{x}} : \mathbb{R} \to \mathbb{S}$ for a point $\mathbf{x} \in \mathbb{S}$ is defined using the following differential equation:

$$\frac{\mathrm{d}\gamma_\theta^x}{\mathrm{d}t}(t) = \nabla\Psi_\theta(\gamma_\theta^x(t)), \quad \gamma_\theta^x(0) = \mathbf{x}. \tag{4.31}$$

Two normal curves (blue) are shown in (c) and (d).

Note that both the spatial PDF and CDF are defined on a 2D *spatial domain* and not on a 2D

data or *parameter domain* like traditional 2D distribution functions. They model the contour distribution along the 1D normal curves. For instance, according to the normal curve definition, the spatial CDF values along a single curve can be obtained by integrating the spatial PDF values along the respective curve in its positive direction.

Gaussian Contour Representation

The spatial PDF expresses the probability density of the iso-contours in the 2D domain. However, it cannot be evaluated directly because the indicator function in Equ. (4.29) is neither continuous nor differentiable. Thus, the same idea underlying KDE is used to obtain a smooth non-parametric PDF from a discrete histogram: The binary indicator function is replaced by a function which generates a smooth and differentiable transition between the sublevel and superlevel sets. In particular, a function with a closed-form first derivative is chosen such that a closed form of Equ. (4.30) can be obtained.

One possible choice is the well-known CDF Φ of the standard normal distribution, yielding

$$\mathcal{P}_k(Y(\mathbf{x}_{i,j}) \geq \theta) = \Phi\left(\frac{y_k(\mathbf{x}_{i,j}) - \theta}{\sigma_s}\right). \tag{4.32}$$

Since Φ is computed as an integral of the Gaussian PDF, it serves the requirements. For the ensemble member in (a), Fig. 4.18 (c) shows the values of $\mathcal{P}_k$ for a small positive *sharpness parameter* σ_s.

Due to the chosen smoothing function, with increasing/decreasing values $y_k(\mathbf{x}_{i,j})$ above/below the iso-value in one ensemble member k, one obtains $\mathcal{P}_k(Y(\mathbf{x}_{i,j}) \geq \theta) \to 1$ and $\mathcal{P}_k(Y(\mathbf{x}_{i,j}) \geq \theta) \to 0$, respectively. The transition rate between the superlevel and sublevel region is controlled by the sharpness parameter. For a fixed sharpness parameter, the spatial CDF has a strong gradient magnitude in those regions along the contour where the data gradient is strong, too. Thus, it provides a measure for the gradient strength along the contours.

By substituting the smoothing function Equ. (4.32) into Equ. (4.28) for one individual member k, and using the resulting CDF in Equ. (4.30), we arrive at the spatial PDF for the selected member:

$$p_k\left(\mathbf{x}_{i,j}\right) = \phi\left(\frac{y_k(\mathbf{x}_{i,j}) - \theta}{\sigma_s}\right) \frac{\|\nabla y_k(\mathbf{x}_{i,j})\|}{\sigma_s}. \tag{4.33}$$

Here, ϕ is the bell-shaped PDF obtained by differentiating the standard normal distribution function. Note that for each ensemble member k, Equ. (4.33) models a Gaussian "uncertainty region" around each single iso-contour, but not for the distribution of the set of iso-contours. The degree of this kind of "uncertainty" is directly related to the data gradient and can be interpreted as *condition indicator* (cf. visual condition analysis of iso-contours in [PH11]). The sharpness parameter σ_s can be interpreted as maximum allowed error for each ensemble member.

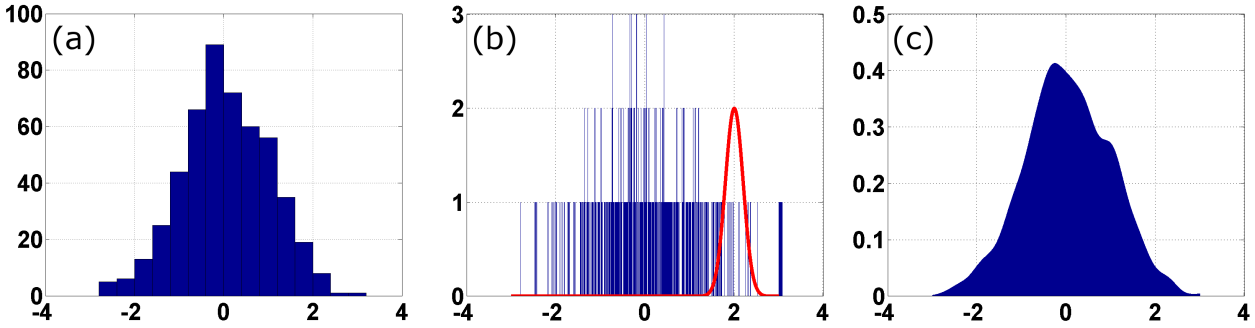

Figure 4.20: Two histograms with 16 bins (a) and 3000 bins (b) are shown for a data set with 500 single scalar values. (c) A continuous histogram representation (3000 bins) is generated by representing each scalar value as Gaussian probability density function (red).

Besides integrating the data gradient information, the novel Gaussian iso-contour representation provides a second strong advantage. This is illustrated by using an example for an analogues problem setting.

An ensemble of 500 single scalar values is assumed. The goal is to illustrate the value distribution by plotting a 1D histogram. For plotting the histogram, the number of bins in a certain range has to be defined. In Fig. 4.20 (a), the histogram for 16 bins is shown, indicating the number of scalar values in the ensemble belonging to each bin. In (b), the histogram resolution is increased to 3000 bins. For increasing bin numbers the distribution becomes more and more discontinuous and its shape more fuzzy. However, the most significant problem for a high bin number appears, when a probability value for an arbitrary point in the value range should be determined. In (b), it could happen that one extracts a probability value of zero close to the maximum likelihood range value around zero, as the histogram (b) is not continuous and has several gaps. This limitation can be approached by representing each ensemble scalar value not as single point in the range but as Gaussian probability density function around this point. This is shown for the value 2 in (b) as red curve. These Gaussian representation are added up and divided by the number of ensemble values. In (c), the result is shown for the same histogram resolution as in (b). The histogram becomes smooth and continuous an no gaps exist any more. Now, it is possible to extract probability density values for arbitrary points in the value range.

This consideration is exactly the same for iso-contours in 2D. Using only a geometric model of crisp contours, one could not build any model which describes the local density of contours for an arbitrary grid point in the 2D domain with arbitrary resolution. This would require a special binning strategy. Using the Gaussian representation and the sharpness parameter, local iso-contour probability density values can be easily obtained, independent of the domain resolution.

In Fig. 4.18 (d), for the ensemble member in (a), the values of p_k are first transformed to $[0, 1]$ via $1 - \exp(-p_k(\mathbf{x}_{i,j}))$, and then to color. The color transition between blue and red allows clearly

distinguishing between low and strong gradient regions along the iso-contour.

The spatial PDF modeling the probability density of the contours in the whole ensemble can now be written as

$$\psi_\theta(\mathbf{x}_{i,j}) = \left\| \frac{1}{n} \sum_{k=1}^{n} \phi\left(\frac{y_k(\mathbf{x}_{i,j}) - \theta}{\sigma_s} \right) \frac{\nabla y_k(\mathbf{x}_{i,j})}{\sigma_s} \right\|. \tag{4.34}$$

The spatial PDF gives rise to a quantitative assessment of the local probability density of iso-contours, which is caused by their positional variation in the ensemble. The distribution puts into relation the strength of the spatial variations and is independent of the domain resolution. As discussed in the following, the sharpness factor is fixed over the entire domain for the purpose of a better visualization of the uncertainty. This is in contrast to KDE, where smoothing filters of adaptive width — or bandwidth — are used to optimize the properties of the obtained non-parametric distribution. In this way, the strength of the local iso-contour variations can be clearly identified. It is worth noting, however, that the spatial PDF covers only the positional variation of an oriented boundary contour. Variations in topology, e.g., flip of the superlevel and sublevel regions across one and the same contour, are not taken into account. However, this is not the case for most ensemble data sets, affected by moderate uncertainty.

Visualization

The following paragraph describes, how to visually incorporate the spatial CDF and PDF into the standard spaghetti plot. Rather than representing iso-contours as separate geometric primitives, the uncertainty information is encoded as color at each domain point.

The spatial CDF computes for every grid point the probability that this point belongs to the region in which the scalar values are greater than the iso-value. These probability values range from 0 to 1, and they are mapped linearly to gray-scales from black to white. The resulting colors serve as background colors C_b, which allow the user to distinguish between the sublevel and superlevel regions, as well as the transition zone in between. It is worth noting here, that this is not possible by just drawing a spaghetti plot.

Next, the color-coded spatial PDFs of all ensemble members are displayed simultaneously. Due to the Gaussian representation, the color-coded contours appear wider and more diffuse in regions showing low gradients, and more narrow and sharper in regions with strong gradients. By drawing all spatial PDFs simultaneously, wider regions could completely hide narrower ones. Therefore, instead of averaging the spatial PDF values of all members (cf. Equ. (4.34)), the largest value at each domain point is always selected first. This results in a new variant of the spatial PDF:

$$\psi_\theta^{\max}(\mathbf{x}_{i,j}) = \max_{k=1,2,\ldots,n} p_k\left(\mathbf{x}_{i,j}\right). \tag{4.35}$$

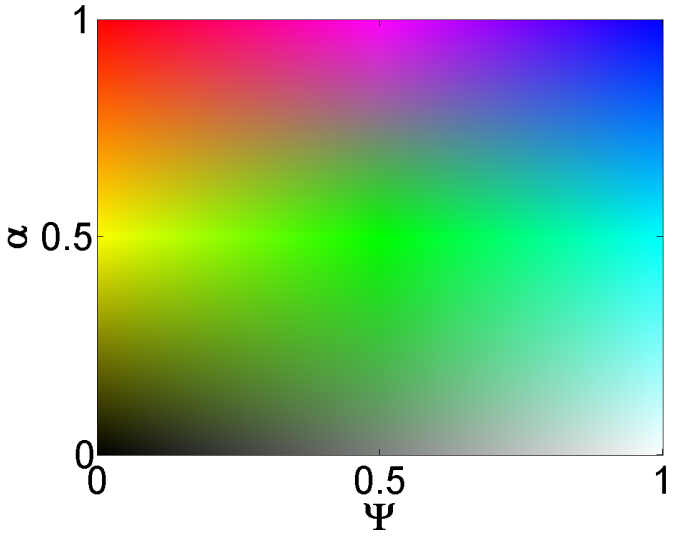

Figure 4.21: 2D color map for the parameters α and the spatial CDF Ψ.

Note that both ψ_θ and $\psi_\theta^{\max}$ range between 0 and ∞. These values are first mapped to $[0, 1]$ using the transfer function

$$\alpha(\mathbf{x}_{i,j}) = 1 - \exp\left(-\tau \cdot \mathrm{T}(\mathbf{x}_{i,j})\right), \tag{4.36}$$

where $\mathrm{T}(\cdot)$ is either replaced by ψ_θ or $\psi_\theta^{\max}$. The scaling parameter τ is used to control the color contrast.

Both α and Ψ are used for color-coding the Gaussian contour representation. Ψ has already been used for constructing a background color C_b. Furthermore, two foreground colors are constructed: The first one, the *lower color* C_l, is obtained by linearly mapping Ψ ($\in [0, 1]$) to the color map [yellow $\rightarrow$ green $\rightarrow$ cyan]. The second one, the *upper color* C_u, is constructed by mapping Ψ to [red $\rightarrow$ magenta $\rightarrow$ blue]. The final color at each grid point is obtained by linearly mapping α from $[0 \rightarrow 0.5 \rightarrow 1]$ to $[C_b \rightarrow C_l \rightarrow C_u]$. The terms *lower color* and *upper color* indicate relatively low and high gradients along the iso-contours. The described mapping of the parameter pair Ψ and α to color is illustrated in Fig. 4.21.

The color scheme allows the simultaneous encoding of the values of the spatial CDF — shows the transition between the sublevel and superlevel sets — and the spatial maximum PDF — indicates regions with high and low gradients.

In Fig. 4.22, the novel color scheme was applied to the ensemble data sets in Fig. 4.17. Regions colored white/black contain those points which belong to the superlevel/sublevel region in all ensemble members. In (a), a gray value of 0.5 indicates an equal number of ensemble members in which the respective point is in the superlevel or sublevel set. Thus, from the location of the gray region it can be concluded directly on a multi-modal distribution of the iso-contours, i.e. the

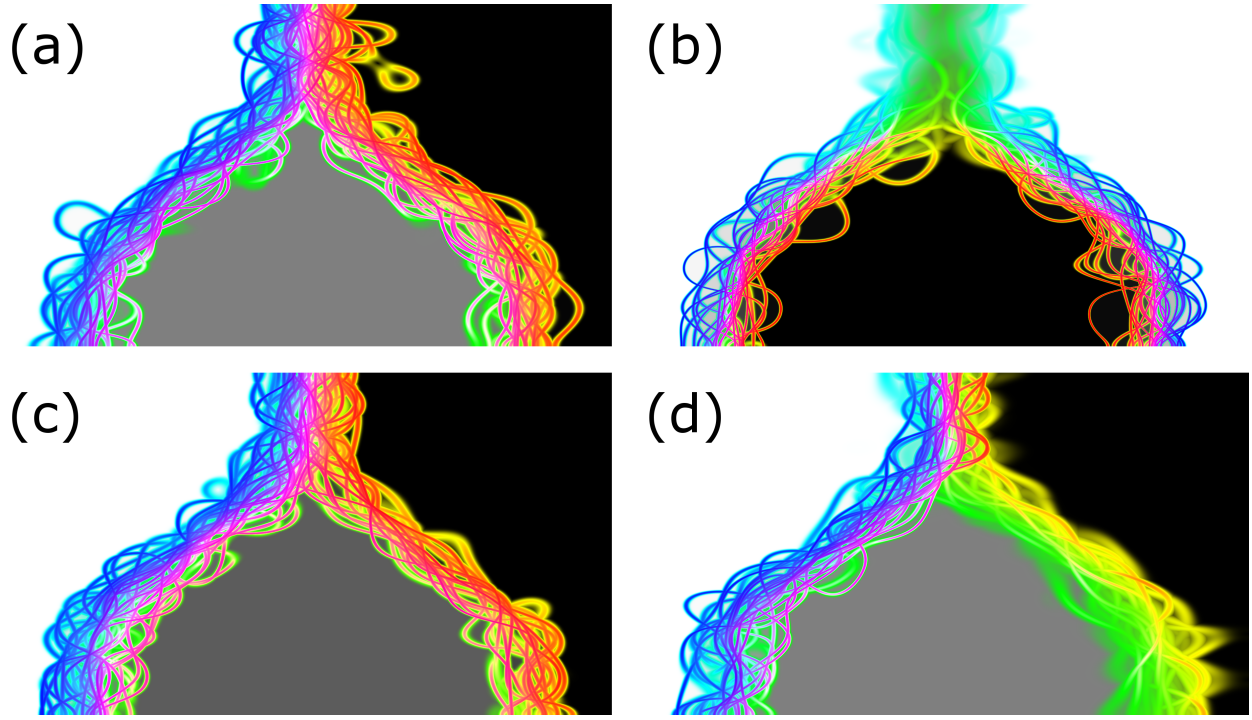

Figure 4.22: Spatial CDF/PDF values for the 4 ensembles in Fig. 4.17 are visualized. Gray-scales encode the point-wise probability (black=0 to white=1) that scalar values are greater than a specified iso-value.

contours are positioned either left or right of the gray region. The color represents the spatial PDF. The strong presence of C_u indicates almost constant gradient magnitude along the iso-contours and no contrast between low and high gradient regions.

The visualization in (b) shows the ensemble data set in Fig. 4.17 (d)-(f). Now, the background color indicates that there are no larger regions belonging *either* to the upper *or* lower iso-regions. Compared to (a), no strong multi-modality in the iso-contour distribution can be concluded; black and white regions reveal a clear separation between the superlevel and sublevel region. The presence of C_l in the upper part of the domain indicates a much lower gradient strength than in the lower part.

In (c), the data set in Fig. 4.17 (g)-(i) is visualized. The different ratios of the frequency of occurrence of the iso-contours in the left and right branch can be slightly observed from the values of the spatial CDF, i.e. the gray region is slighter darker in (c). To emphasize this difference more clearly, an improved visualization approach will be presented in the following sections.

In (d), the data set in Fig. 4.17 (j)-(l) is visualized. The presence of the upper color C_u and lower color C_l in the left and right branch, respectively, indicates the difference in gradient strengths along the iso-contours in both branches (cf. Fig. 4.17 (k) and(l)).

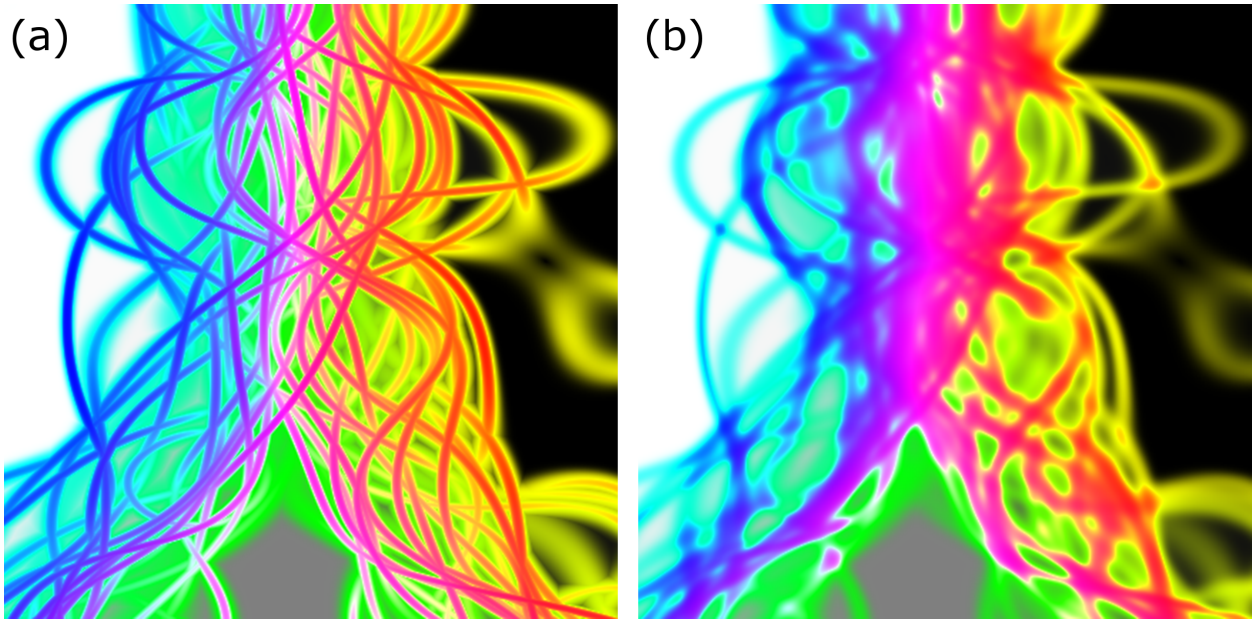

Figure 4.23: Visualization of spatial PDF values for $\psi_\theta^{\max}$ (a) and ψ_θ (b).

Fig. 4.23 illustrates the differences between using $\psi_\theta^{\max}$ and ψ_θ for visualizing the contour distributions. (a) and (b) show the upper domain part of the data set in Fig. 4.17 (a), with $\psi_\theta^{\max}$ and ψ_θ as probability measures, respectively. The individual contours can be visualized much more effectively in (a), and the visual focus is always put on the contours along which the gradients are most prominent. Due to the averaging of values in (b), contour points where several contours intersect each other receive higher values than points where no crossing occurs. This results in a shift of the visual focus from the contours to the intersection regions, and an increasing loss of the contours' shapes.

4.2.3 Statistical Modality Analysis

So far, the presented concepts build upon the stochastic distribution of iso-contours in scalar ensembles, and they take into account the gradient strength to estimate the contours' sharpness. In the following, a closer look is taken at the shape of the spatial distribution of iso-contours. The goal is to depict those regions where the distribution follows a uni-modal probability density, i.e. a Gaussian-like density function with one prominent maximum probability contour, or is represented by a multi-modal density function. In the context of the proposed 2D spatial PDFs, the mentioned PDF modality in a certain spatial region is equal to the modality of the spatial PDF along a 1D normal curve (cf. Equ. (4.31)) in this region.

Gaussian Reference Distribution

To validate whether the contours follow a Gaussian-like distribution, the underlying data distribution is assumed as such and the Gaussian probability density function is computed from the ensemble. If the Gaussian assumption is wrong, the computed ensemble density function will show a clear mismatch with it.

The Gaussian density function is computed by first deriving at every grid point the mean μ and standard deviation σ from the ensemble data set:

$$\mu(Y(\mathbf{x}_{i,j})) = \frac{1}{n}\sum_{k=1}^{n} y_k(\mathbf{x}_{i,j}), \tag{4.37}$$

$$\sigma_e(Y(\mathbf{x}_{i,j})) = \sqrt{\left(\sum_{k=1}^{n}\frac{y_k(\mathbf{x}_{i,j})^2}{n-1}\right) - \left(\sum_{k=1}^{n}\frac{y_k(\mathbf{x}_{i,j})}{\sqrt{n(n-1)}}\right)^2}. \tag{4.38}$$

The sharpness parameter in Equ. 4.32 is added to the standard deviation as an additional degree of uncertainty:

$$\sigma(Y(\mathbf{x}_{i,j})) = \sqrt{\sigma_e(Y(\mathbf{x}_{i,j}))^2 + \sigma_s^2}. \tag{4.39}$$

The rational behind this is that the distribution of each separate iso-contour is modeled via a Gaussian function, and, thus, the variance of this Gaussian function has to be added[1].

By means of these parameters, a corresponding Gaussian spatial CDF (Γ_θ) and PDF (γ_θ) can be computed at every point in the same way as in Equ. (4.28) and (4.30):

$$\Gamma_\theta(\mathbf{x}_{i,j}) := \mathcal{P}(Y(\mathbf{x}_{i,j}) \geq \theta) = \Phi\left(\frac{\mu(Y(\mathbf{x}_{i,j})) - \theta}{\sigma(Y(\mathbf{x}_{i,j}))}\right), \tag{4.40}$$

$$\gamma_\theta(\mathbf{x}_{i,j}) := \left\|\nabla\Gamma_\theta(\mathbf{x}_{i,j})\right\|. \tag{4.41}$$

Since the Gaussian PDF reaches its maximum at its mean, the most likely position of the iso-contours is at grid points where $\mu(Y(\mathbf{x}_{i,j})) = \theta$. Thus, for Gaussian distributed contours the mean contour is a good representation. However, along the mean contour, the Gaussian PDF does not necessarily obey a constant value. The value is higher in regions where the spatial spread of the contours is lower. This can be compared to a simple 1D Gaussian PDF, where the peak is also narrower and higher if the standard deviation is lower.

As the presented concept seeks for an analysis of the modality of the Gaussian distribution and not of its maximum height, all scaling factors are eliminated and a normalized version of the PDF is used:

$$\gamma_\theta^N(\mathbf{x}_{i,j}) = \exp\left(-\frac{(\mu(Y(\mathbf{x}_{i,j})) - \theta)^2}{2\sigma(Y(\mathbf{x}_{i,j}))^2}\right). \tag{4.42}$$

[1] The variance of the sum of two Gaussians equals the sum of the variances of each of them.

The maximum value of this function is 1, and it is obtained at the maximum likelihood positions of the contours where $\mu(Y(\mathbf{x}_{i,j})) = \theta$. This normalized function will serve as reference density function against which the actual distribution of the contours is compared.

Normalized Spatial Probability Density

For comparing the actual non-parametric contour distribution with the Gaussian reference distribution, the spatial PDF of the contour distribution is normalized, too. This is performed by normalizing the spatial PDFs of the individual ensemble members (cf. Equ. (4.33)), i.e. by transforming the individual distribution functions into the normalized versions

$$p_k^N(\mathbf{x}_{i,j}) = \exp\left(-\frac{(y_k(\mathbf{x}_{i,j}) - \theta)^2}{2\sigma_s^2}\right). \tag{4.43}$$

The final total normalized spatial PDF is then obtained by averaging the individual spatial PDF values as

$$\psi_\theta^N(\mathbf{x}_{i,j}) = \frac{1}{n}\sum_{k=1}^{n} p_k^N(\mathbf{x}_{i,j}). \tag{4.44}$$

This normalization process has two major advantages: Firstly, if the contours are Gaussian distributed in a certain region, then for a sufficiently large sharpness parameter σ_s, ψ_θ^N becomes uni-modal and almost equal to the normalized Gaussian reference distribution. By interactively adjusting σ_s, the user can now test for which sharpness parameter the Gaussian assumption is valid. However, this may be dependent on the spatial location. Secondly, as the normalized PDF in each ensemble member has a maximum value of 1 and fades out with increasing distance from the contour, the individual normalized Gaussian representations can be regarded as weighting functions. By computing the average over all ensemble members, the total normalized spatial PDF value is obtained. These values indicate the local probability of the occurrence of a contour with respect to the chosen sharpness parameter.

Visualization

The normalized Gaussian reference PDF γ_θ^N and the normalized total ensemble PDF ψ_θ^N are visualized in one image to provide a visual indication whether the assumption of a Gaussian contour distribution can be justified. To be able to clearly distinguish between the two functions, different mappings are used for either function. The Gaussian reference PDF is encoded by mapping values linearly to the color table shown in Fig. 4.24.

In addition, the values of the distribution ψ_θ^N are interpreted as a height field over the 2D domain, and this height field is shaded based on a local diffuse illumination model. Therefore, at every grid

Figure 4.24: Colormap for normalized spatial PDF values.

point a normal vector is computed analytically from the spatial PDF and the diffuse reflection of a point light source is simulated. The obtained values determine the color intensity at the respective position, creating the impression of a *relief structure* on the 2D domain. By scaling the slope of the height field via a parameter ρ, the relief effect can be decreased or increased interactively. In the visualization, the illumination values are 1 where the height field topography is orthogonal to the viewing direction, and it approaches 0 in regions with very steep slopes. These values are used to scale the color values of the Gaussian reference distribution.

In Fig. 4.25, the proposed visualization technique is applied to the 4 ensembles in Fig. 4.17. In (a), the values of the normalized Gaussian reference PDF are maximal between the two branches (white region). This is the Gaussian maximum likelihood region, where the mean iso-contour would be positioned. By looking at the relief visualization, however, one sees that the ensemble does not match the Gaussian assumption in the lower domain region. Here, the iso-contours exhibit a bi-modal distribution around the Gaussian maximum likelihood region. Although the matching is much better in the upper part, the high-frequent relief structure induced by strong gradients all across the domain does not match with the smooth Gaussian reference distribution.

The situation is different for the second ensemble data set in (b). In the upper region, low gradients result in a much smoother contour distribution, now clearly matching the Gaussian assumption. In the lower part, the contours are still too sharp to allow assuming a uni-modal distribution. However, because the contours are better aligned with the Gaussian mean/maximum likelihood region, the strong bi-modality in (a) is not present any more.

In (c), the Gaussian reference density is slightly deviated towards the left branch, which indicates the higher probability of the occurrence of iso-contours in the left branch.

The relief representation in (d) clearly reveals stronger gradients along the iso-contours in the left branch compared to the right branch. Furthermore, the Gaussian maximum likelihood region (white) is almost aligned with the left branch. Although the iso-contours occur equally in both branches, ensemble members with iso-contours in the right branch are affected by significantly lower gradients. Therefore, the multi-modal positional variability of left branch iso-contours is much lower than for the more uni-modal distribution in the right branch.

Fig. 4.26 shows visualizations where only the normalized ensemble spatial PDF was used. In this case, the function values were displayed using both the relief *and* the color mapping. In addition, the sharpness value σ_s was significantly increased to "smear out" the contours and to concentrate on

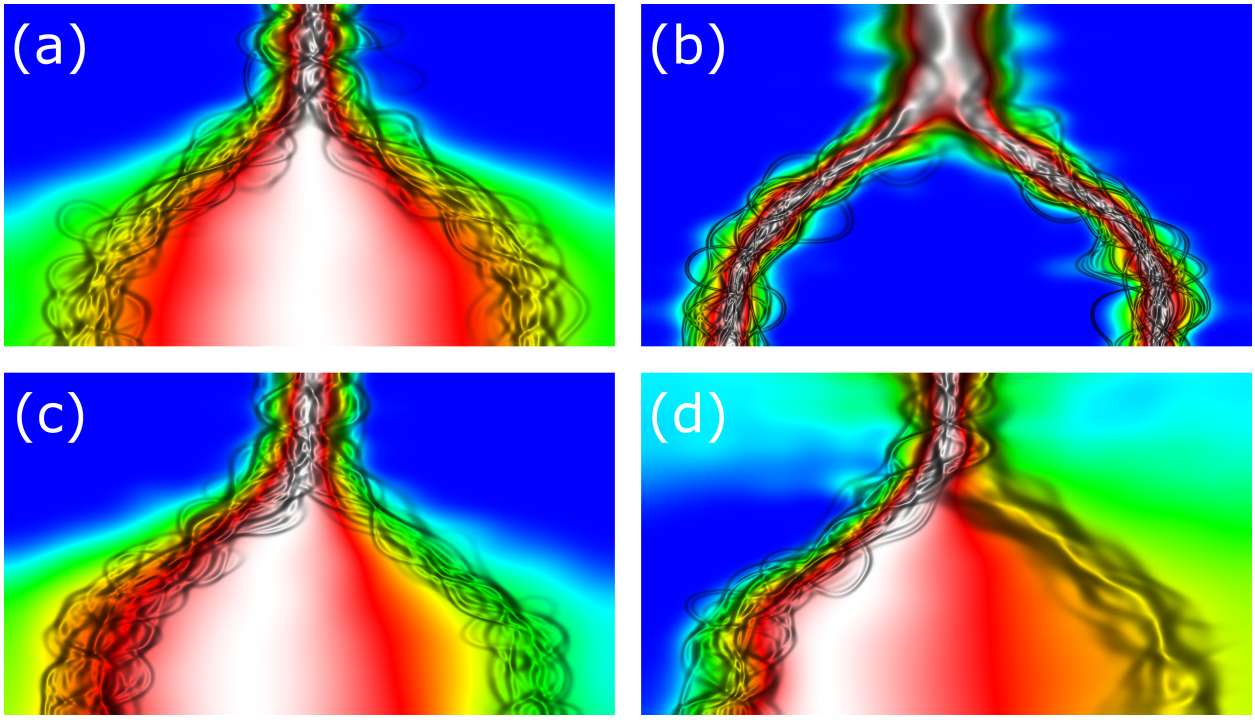

Figure 4.25: Visual modality analysis for ensemble data sets in Fig. 4.17. Normalized Gaussian reference PDF is mapped to color, normalized ensemble PDF is mapped to the relief representation.

the more coarse distribution. The strongest motivation for this visualization mode is shown in (c): A linear re-scaling of the color map as shown in Fig. 4.24 was used so that white color is attached to the ensemble maximum likelihood area where the two branches unite. This normalization results in a red and green coloring of the maximum regions in the left and right branch. The color map in Fig. 4.24 reveals that the values assigned to red and green have a numerical ratio of 2:1. This matches exactly the ratio of the probabilities of the contour occurrences in the left and right branch of the third ensemble data set (cf. Fig. 4.17 (g)-(i)).

4.2.4 Implementation

The presented visualization techniques can be implemented very efficiently because all computations can be performed at every grid point in parallel. This gives rise to a highly efficient implementation on the GPU, with their particular design for massively parallel workloads. The ability to simultaneously use many processing units is employed and the thread level parallelism is exploited to hide latency operations in order to achieve the possibility to change all parameters affecting the estimated distribution function and their visual representation interactively.

In the current implementation, the ensemble of 2D data sets is stored in a 2D texture array on

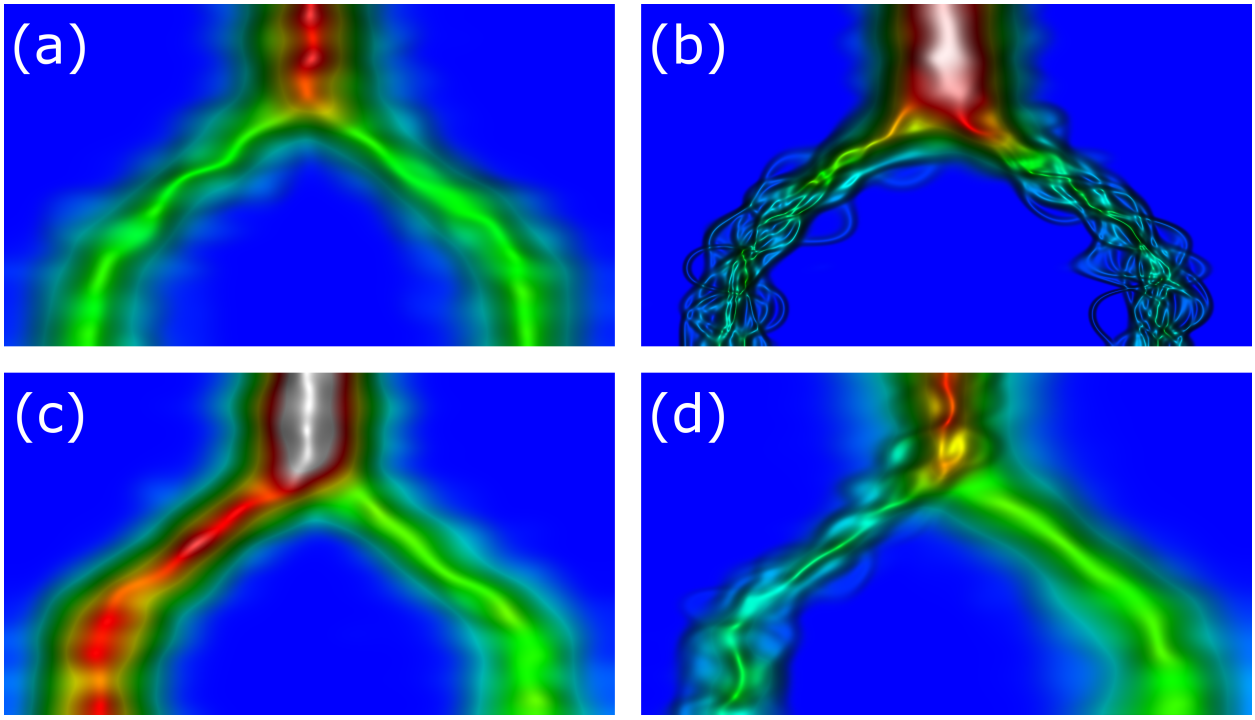

Figure 4.26: Visual modality analysis for ensemble data sets in Fig. 4.17. Normalized ensemble PDF is mapped to both color and relief representation.

the GPU. At each texel, one scalar value and two components for the respective data gradient are stored. The algorithm uses a rectangular area in the screen-space for displaying the visualization results. For each screen-space pixel the spatial CDF and PDF values are computed and mapped to color as described. To do so, both scalar and gradient values are extracted from the texture array and consecutively combined for all ensemble members. As the screen-space resolution is usually higher than the data resolution, values between grid points of the data domain are obtained by using bi-linear texture interpolation. All computations are performed for each individual pixel in the screen-space using the introduced formulas. However, for computing the standard normal cumulative distribution function Φ, the numerical approximation given in [AS64] is used.

The most time consuming computation is caused when changing the iso-value as well as the Gaussian sharpness parameter. This requires a re-computation of both the spatial CDF and PDF values at every grid point. Even for a very high resolution pixel raster of 1860×1040 and a 2D scalar ensemble data set with resolution 1060×460 and comprised of 50 members this operation takes below 20 ms on a NVIDIA GeForce GTX 660 Ti graphics adapter.

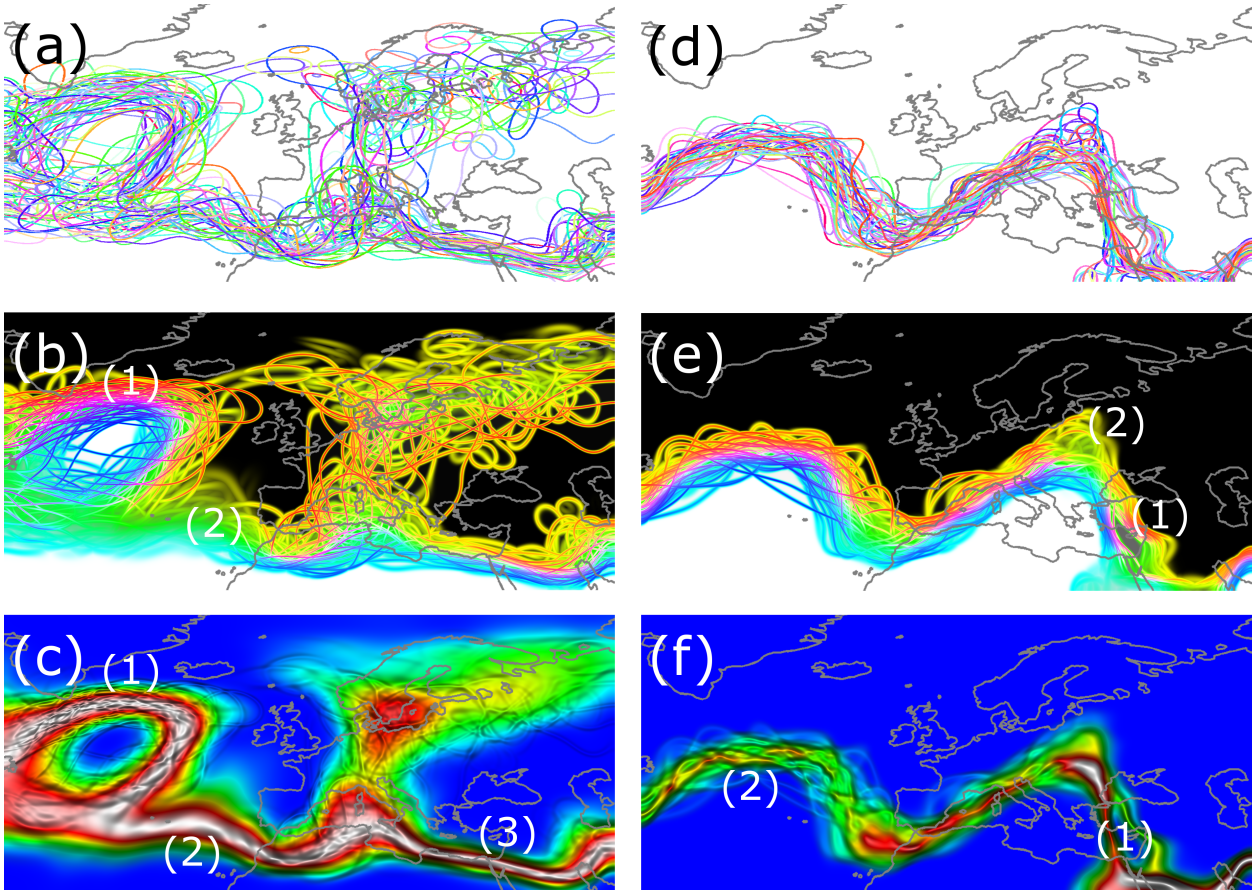

Figure 4.27: Left column: ECMWF temperature ensemble A. (a) Spaghetti plot. (b) Spatial CDF and PDF. (c) Visual modality analysis using normalized Gaussian reference PDF (color) and ensemble PDF (relief). Right column: ECMWF temperature ensemble B. (d) Spaghetti plot. (e) Spatial CDF and PDF. (f) Visual analysis for probability ratio between bi-modal branches in iso-contour distribution. Normalized ensemble PDF is mapped to color and relief structure.

4.2.5 Results

The novel approach is now applied to visualize uncertain iso-contours in two real-world ensembles. Scalar temperature ensembles A and B, with 50 members each, generated by the European Center for Medium-Range Weather Forecast (ECMWF) for two different forecast periods and pressure levels above Europe. Several visualizations are shown in Fig. 4.27 and are put into a geo-spatial context by integrating an overlay of the European coast-lines in gray. The user can change several parameters interactively to control the visual results: The iso-value θ, the sharpness parameter σ_s, the color contrast parameter τ and the relief scaling factor ρ.

In (a), temperature iso-contours are plotted using the standard spaghetti plot approach, with different colors being assigned to contours in different ensemble members. Although only 50 members

are plotted, the visualization quickly becomes cluttered, and due to the aforementioned limitations of spaghetti plots a detailed statistical analysis is difficult.

In (b), the spatial CDF and the maximum values of the spatial PDF are displayed. The gray-valued background allows one to clearly segment the domain into regions with temperature values above (white) and below (black) the selected iso-value. The "sharpness" of the iso-contours and the presence of color C_u in region (1) identify a sharp temperature transition with low gradient uncertainty towards the Greenland border. In region (2), the iso-contours have a smooth appearance and do not show a clear preferential direction. Together with the presence of the lower color C_l this indicates lower gradients and a much smoother temperature transition. These observations cannot be made by just looking at the spaghetti plot in (a).

Image (c) gives hints on the statistical distribution of the iso-contours. The normalized values of the spatial PDF ψ_θ^N of the ensemble contours are visualized via the relief representation. Values of the normalized Gaussian reference distribution are mapped to color according to the color map in Fig. 4.24. One can observe a high-frequent relief structure in region (1), i.e. it has a huge number of peaks. This indicates a highly multi-modal spatial iso-contour probability function in this region. Thus, this region cannot be represented by a Gaussian distribution.

It can be further seen that, despite the reliable sharp temperature transition, the positional uncertainty of the iso-contours is rather high. In 4.2.3, it was stated that the Gaussian representation for each individual iso-contour can be regarded as conditional error bound depending on the gradient strength. The visualization indicates that the spatial positional uncertainty of the iso-contour is significantly higher than the conditional error bounds in the individual ensemble members. Thus, in region (1) it would not make sense to represent the ensemble by a mean iso-contour, as the distribution is highly multi-modal and no maximum likelihood region exists. However, in regions (2) and (3) the relief structure matches with the colors of the Gaussian reference PDF. In (2), the positional uncertainty is high but the individual gradients are rather low. In (3), the gradients are stronger, but the positional uncertainty is quite low. In both cases the uncertainty can be modeled by a Gaussian distribution and a representative mean iso-contour.

Image (d) shows the spaghetti plot for a different iso-value and pressure level in the second temperature ensemble data set B. Compared to the first ensemble, the visualization in (e) reveals a new statistical feature. In region (1) (also shown in Fig. 4.28 (a)), the iso-contours split up into two branches, which indicates a bi-modal distribution. Both branches enclose a gray-valued background area. In addition, the presence of the upper color C_u indicates strong gradients in this region, compared to, for instance, region (2).

Although the spaghetti plot in (d) cannot reveal the gradient strengths, the split of the iso-contours into two branches in region (1) can be clearly observed from it. This works well for a low number of ensemble members, but might be harder to observe for a larger number. However, the visualizations

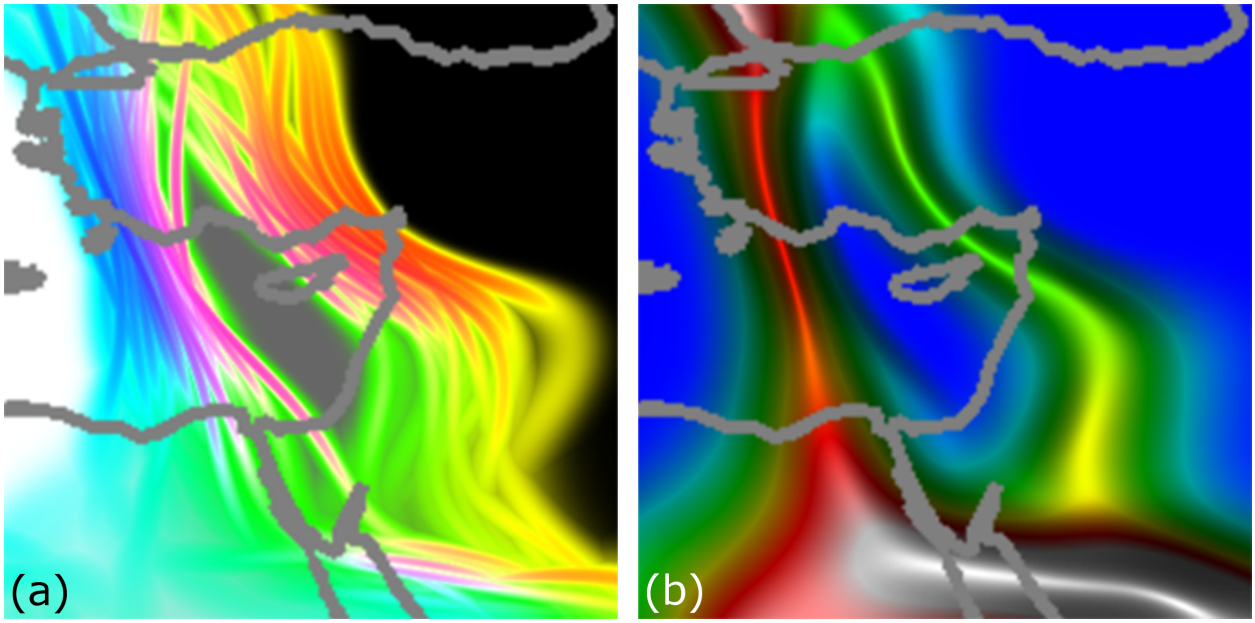

Figure 4.28: Magnifications of region (1) in Fig. 4.27 (e) and (f).

in (d) and (e) do not allow for a quantitative analysis of the iso-contour distribution. For instance, the relative probability ratio between the two branches in region (1) cannot be communicated. To achieve this, in image (f) the normalized ensemble distribution is encoded using colors and via the relief representation. The color values have been rescaled so that the maximum value of the spatial PDF is mapped to white. Now, one can observe that one branch is colored red and the other one green (also shown in Fig. 4.28 (b)). This indicates that the iso-contour is present two times more often in the left than in the right branch. Furthermore, the relief shows only two main branches, but no sub-branches. Thus, the distribution can be assumed bi-modal. On the other hand, the high frequent structure of the relief representation in region (2) indicates a high multi-modal distribution and strong gradients in this region.

4.2.6 Conclusion

In this section, it has been shown how to derive probability distributions for iso-contours in scalar ensemble fields, and how to visually convey this information to allow for an improved uncertainty analysis. Instead of assuming a parametric uncertainty model, non-parametric spatial distributions have been derived by computing statistical summaries and generating continuous distribution functions thereof. It has been demonstrated how the novel technique can be used to improve standard spaghetti plots towards a more reliable and quantitative uncertainty estimation. Furthermore, a special color mapping scheme has been proposed for effectively analyzing probability distribution modalities of the iso-contour uncertainty in different spatial regions.

5

Structural Feature Analysis

Most uncertainty visualization approaches do only consider the local *absolute uncertainty* of single data values or specific features in a data set. Thereby, visual effects encode the variations of scalar data values, critical points, iso-surfaces, etc. utilizing a mathematical model for describing the underlying uncertainty at particular locations of the respective spatial domain. The relative uncertainty, e.g., describing the relative variability of two or more random variables in the spatial domain with respect to each other, is not considered at all in common uncertainty visualization approaches. However, analyzing such relative variabilities can provide substantial insight into the stability of certain structures or features in a data set.

This chapter introduces several novel techniques for visualizing the *structural uncertainty* of specific features or data distributions in 2D and 3D scalar data fields. The presented approaches make a significant contribution to the field of uncertainty visualization, as they clearly show how the structural uncertainty analysis can reveal information about uncertain data sets that cannot be obtained by traditional methods.

The first section of this chapter introduces *correlation* information as primary indicator for analyzing stochastic dependences and therefore relative uncertainties in Gaussian distributed uncertain data sets. Furthermore, a comprehensive introduction to the definition and interpretation of correlation data is provided while illustrating the importance of its visual analysis for uncertain scalar data sets. In a number of examples, some general conclusions, one can draw from correlation information in uncertain data regarding its structural variability, are demonstrated. In addition, a visualization approach for *local* and usually anisotropic correlation structures in the vicinity of uncertain iso-surfaces in 3D data sets is presented.

The second section extends the previously introduced approach to the visualization of *global* correlation structures and therefore to global stochastic dependences between random variables in distinct regions of Gaussian distributed uncertain scalar data sets. A novel spatial clustering approach is

presented for the visual analysis of both positive and inverse correlation relationships within and between spatial regions.

The last section concentrates on the visualization of the variability of *gradients* in uncertain 2D scalar fields. As gradients depend on the rate of change of data values, they are highly affected by relative uncertainties between two or more random variables. Besides the analysis of the effect of uncertainty on differential quantities like gradients, a mathematical framework is developed for computing confidence intervals for both the gradient orientation and magnitude variability. Furthermore, a novel color diffusion scheme and a special family of circular glyphs are introduced for visually communicating these two separate entities of gradient uncertainty. The presented methods are utilized for analyzing the stability of certain features in uncertain 2D scalar fields, with respect to both local derivatives and feature orientations. In contrast to correlation visualization algorithms, the visual analysis of uncertain gradients allows a more quantitative and less abstract assessment of structural uncertainty.

5.1 Local Correlation Visualization

It is quite common to model the uncertainty at discrete spatial points $\mathbf{x}_i$ in an uncertain scalar field by a *multi-variate random variable* $\mathbf{Y}$, i.e., a vector consisting of multiple scalar random variables $Y(\mathbf{x}_i)$. The uncertainty at a point $\mathbf{x}_i$ is then given by a set of *realizations* of the random variable $Y(\mathbf{x}_i)$. The *standard deviation* of one of the random variables indicates the strength of the deviation of possible realizations from their respective *mean value*, independently of the means and deviations of random variables at any other point in the domain. Consequently, the standard deviation describes the *local uncertainty* but does not allow inferring on possible variations at different positions *relative to each other*.

For instance, let us assume that the mean and standard deviation values at two adjacent points are identical, but the realizations of the random variables at both points are *stochastically independent*. In this case, it cannot be predicted whether there is a positive, zero, or negative derivative of the data between the two points, and consequently the values at these points relative to each other cannot be resolved. This kind of uncertainty will be subsequently called *structural uncertainty*, since it is associated with the occurrence of particular structures in the data which are affected by the degree of dependence between the values at two or more data points. In particular, structural uncertainty covers all variabilities of geometric features (e.g., orientation, curvature, etc.) of multi-dimensional structures (e.g., surfaces). To avoid confusion, it should be noted here that in other communities the term "structural uncertainty" is used to indicate the uncertainty in the structure of a model.

In the analysis of structural uncertainty it is thus necessary to first predict the stochastic dependence between the random variables used to model the uncertainty, and then to analyze the possible structural variations of features in this field taking into account these dependences. To achieve this, the random variables used to model the uncertainty are assumed to exhibit a *multi-variate Gaussian distribution* so that mean values and standard deviations exist, and the linear stochastic dependence is given by the *correlation*. It is worth noting here, that non-linear dependences between random variables are not present for multi-variate Gaussian distributions, and, thus, they are not covered by the methods proposed in this work.

For two Gaussian distributed random variables X and Y the mutual stochastic dependence is given by their correlation $\rho(X,Y)$, which ranges from -1 to 1 and characterizes the *linear relation* between the two variables. It is computed as $\mathbb{C}\mathrm{ov}(X,Y)/\sqrt{\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(Y)}$, where $\mathbb{V}\mathrm{ar}(X)$ and $\mathbb{V}\mathrm{ar}(Y)$, respectively, denote the *variance* of X and Y, and $\mathbb{C}\mathrm{ov}(X,Y)$ is the *covariance* between X and Y. For a thorough discussion of the concepts of stochastic dependence and correlation, as well as approaches to compute correlation values from given realization sets, the reader is referred to section 2.3.2 and [Fel08, Geo08, LRN88].

A high positive correlation between two random variables indicates that realizations of both variables are likely to deviate into the same direction from their mean values. A strong negative or *inverse* correlation indicates that a strong positive deviation for one realization is likely to result in a strong negative deviation of the other realization, and vice versa.

In Fig. 5.1, the concept of relative uncertainty between two random variables is illustrated for a strong positive correlation (upper row) and a strong inverse correlation (lower row). The confidence intervals (green intervals along one-dimensional value axis) are shown for two random variables with the same mean (small black line) and standard deviation values. For two correlation values, four realizations are shown for each random variable, indicated by the red dots and linked with a red line. The upper row shows the relative behavior of the realizations for a strong positive correlation between the two random variables. The values always deviate in the same direction with respect to the mean value. The lower row reveals the behavior for a strong inverse stochastic dependence and correlation. The realizations deviate always in opposite direction with respect to the mean value. For instance, compared to the positive correlation case, the derivative or slope between the realizations is highly variable and therefore uncertain.

As the realizations of random variables in regions exhibiting very low correlation can be assumed to be independent of each other, the effect of uncertainty on a mean structure in such a region is to a large extent arbitrary. As a consequence, in such regions, a high structural variation is caused by high uncertainty, and the resulting mean structures are not reliable. Contrarily, regions exhibiting high correlations are affected by "smooth" structural uncertainty, meaning that the uncertainty causes low frequent variations to occur more likely. Thus, high frequencies in the mean structures are not affected strongly and one can argue that they are, therefore, stochastically

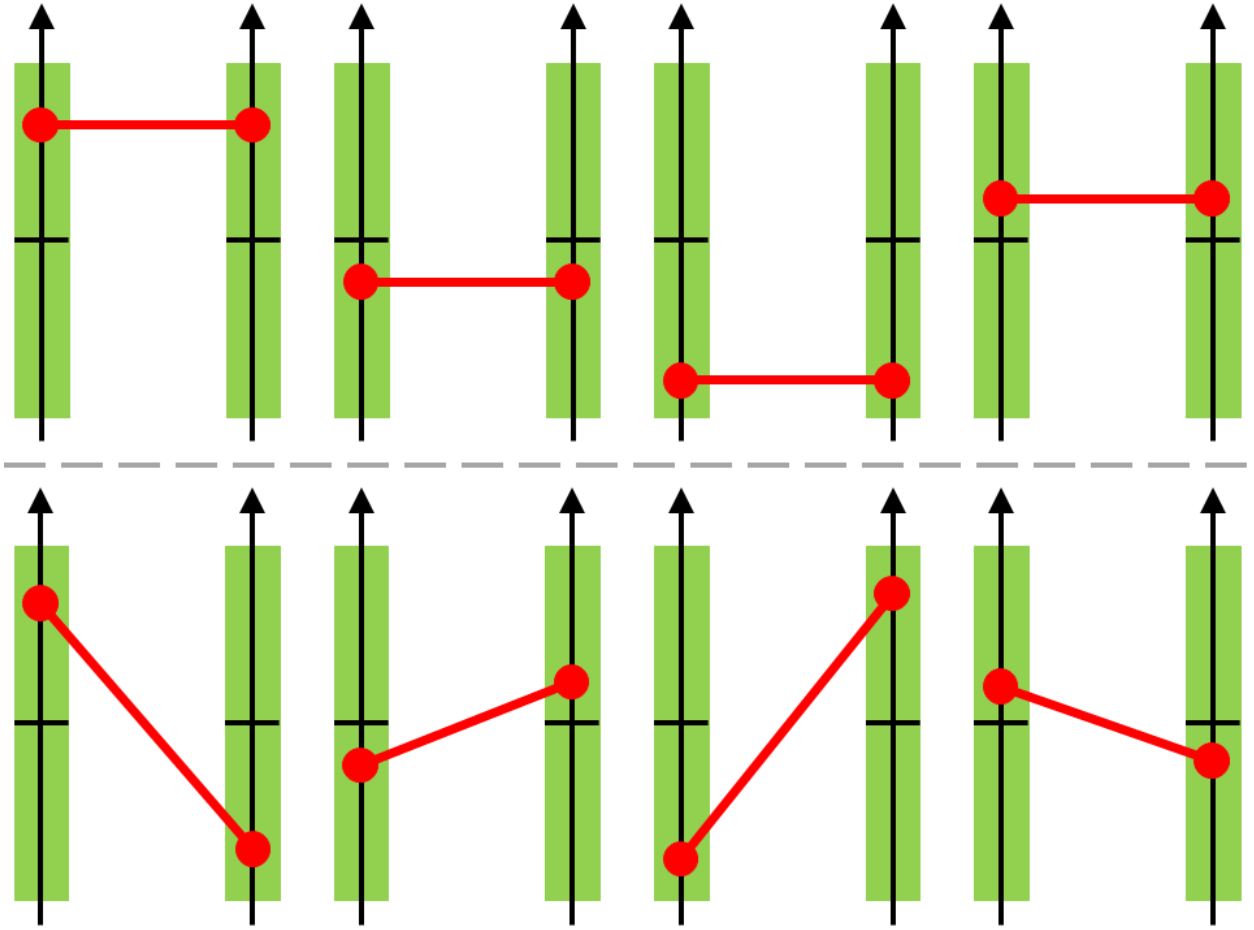

Figure 5.1: The realizations of a bi-variate distributed random variable are shown for equal mean values (horizontal lines) and confidence intervals (green) in both components. The realizations were generated with a strong positive (upper row) and negative (lower row) correlation setting between the components.

more stable. Correlation is thus a very important means to analyze the structural variability of particular features in uncertain data fields, and this property of correlation forms the basis of the investigations presented in this section.

The first goal is to demonstrate the use of correlation as an indicator for structural uncertainty. For this purpose, a number of examples have been designed, ranging from rather simple to quite complicated ones, which clearly show the inter-dependencies between (anisotropic) local correlation structures and the variability of specific features in the data. These examples are utilized to strengthen the awareness of the relevance of correlation analysis for estimating possible structural changes of relevant features due to uncertainty.

The second goal is to investigate methods for visualizing the structural variability of iso-surfaces in uncertain 3D scalar fields. To achieve this, a novel method is proposed for visualizing the correlation

structures in the vicinity of such surfaces based on a local anisotropic correlation model. By visually distinguishing between the local correlations between points on the surface and along the surface's normal directions, an improved understanding of the geometric and topological variability of uncertain iso-surfaces is enabled. The proposed visualization techniques can be used in addition to techniques which directly visualize the standard deviation [JS03, PWL97], and, thus, strive for a comprehensive visualization of the full Gaussian covariance information.

The relevance of the presented work is demonstrated in synthetic and real-world examples using standard random distributions and ensemble simulations. To effectively show the influence of correlation on the structural variability of features in multi-dimensional data sets, a number of synthetic data sets have been generated, using specific random distributions to model uncertainty. In addition, the proposed methods are used to visualize the structural uncertainty in an ensemble of 3D temperature fields in the exosphere which was simulated by the European Centre for Medium-Range Weather Forecasts [Eur]. The proposed technique for correlation visualization allows concluding on the stochastic stability of the mean iso-surface in this field, revealing regions which are strongly and weakly affected by structural uncertainty.

5.1.1 Related Work

So far, only very few approaches have explicitly addressed the visualization of data correlations due to uncertainty. [JPR*04] proposed a tool for the visualization of correlation between two 3D scalar fields via color mapping and slicing. The use of glyphs for the visualization of local covariance information was demonstrated in [KWL*04]. However, in none of these approaches a correlation model was considered, and the used glyphs were not adapted to particular surface structures in 3D data. The approach in [YXK13] suggests a numerical technique for visualizing covariance and cross-covariance fields of 2-dimensional results of stochastic simulations. In [STS06], Multifield-Graphs have been suggested to analyze correlations between several field data. In [SWMW09], data clustering and segmentation techniques are used for identifying correlations in time-varying multi-variate volumetric data sets.

5.1.2 Structural Uncertainty

In this paragraph, several examples show the inter-dependencies between correlation information and the structural variability of features in multi-dimensional scalar fields. In particular, light is shed onto the interpretation of (anisotropic) correlation structures and, thus, the importance of correlation analysis is emphasized for the prediction of possible effects of uncertainty on features in the data. These investigations demonstrate the importance of a correlation analysis and its visualization for making reliable assumptions on the uncertainty of geometric properties of certain features, e.g., position, orientation and curvature for surfaces.

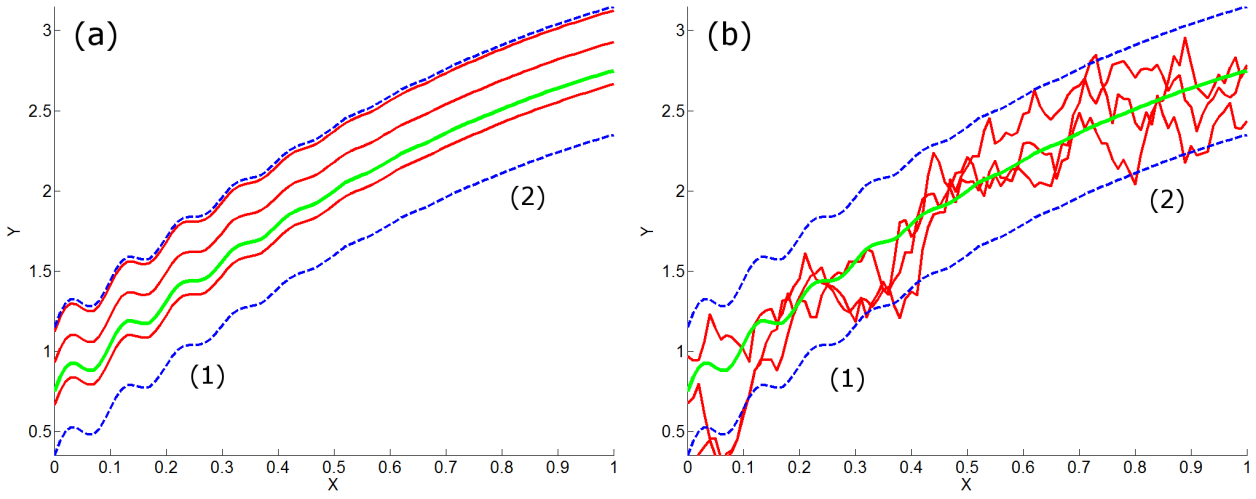

Figure 5.2: Illustration of the structural variability of a 2D curve (green) given by a set of points (x_i, y_i), where each y_i represents the mean of the realizations of a multi-variate normal distributed random variable at x_i. While the mean values vary for different x_i, the standard deviation remains constant. Blue dashed lines enclose the confidence region around the mean curve. Three possible line realizations (red) are shown for very strong (a) and weak (b) local correlation between random variables at adjacent points.

Experiment 1. Fig. 5.2 illustrates a first example using 2D curves that are affected by uncertainty to demonstrate how correlation information can be used for analyzing structural uncertainty. In this example, at every position x_i in a set of equally spaced positions in the x domain, multi-variate normal distributed random variables $Y(x_i)$ with smoothly varying means but the same standard deviation were used to generate sets of y-values for every point x_i. The green curve shows the mean y-values at every x_i. The confidence region for $\pm\sigma$ is enclosed by two blue dashed lines. The red curves always show lines for three possible realizations of the y-values at the points x_i for very strong positive (a) and weak (b) local correlations between adjacent random variables $Y(x_i)$ and $Y(x_{i+1})$.

In region (1), the mean values of the normal distributions used to generate the y-values were chosen such that the mean curve shows higher frequencies than in region (2). Due to the strong correlation between the random values at adjacent positions, the three possible occurrences of curves in (a) show the same structural behavior like the mean curve in (a), while in (b) these structures disappear completely due to the low correlation. Strong correlation effectively keeps adjacent realizations of the random variables used to model the curves close together, while they can occur rather unconstrained with respect to each other in regions of low correlations. It should also be noted here, that the knowledge of the mean and the standard deviation of the normal distributions alone does not allow inferring on the structural behavior of the resulting realizations.

Experiment 2. The next example demonstrates the use of correlation analysis for predicting the

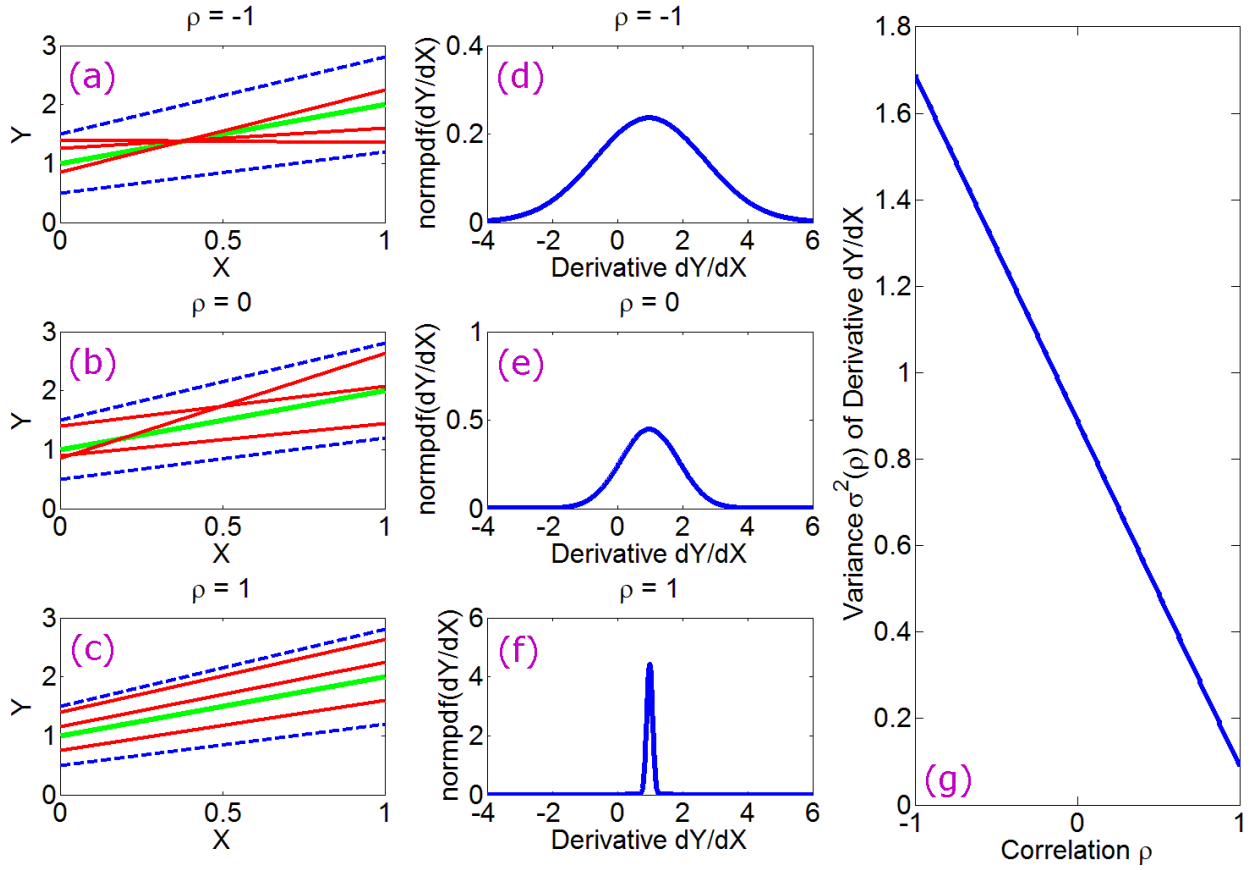

Figure 5.3: In (a), (b) and (c), for two bi-variate normal distributed random variables Y(0) and Y(1) the means (connected by a green line) and the confidence region (enclosed by two blue dashed lines) are shown. Red lines show three realizations for correlations -1, 0, and 1, respectively. In (d), (e) and (f) the Gaussian probability density for the derivative between Y(0) and Y(1) is illustrated. (g) illustrates the relation between correlation values and variance of the derivative distribution.

variability of the orientation of curves that are affected by uncertainty. In a first experiment, at two points positioned at $x = 0$ and $x = 1$ on the x-coordinate axis a bi-variate normal distributed random variable was used to compute sets of y-values. The green line in Fig. 5.3 (a), (b), and (c) shows the linear interpolation between the mean y-values at the two points. The linearly interpolated confidence region is enclosed by the blue dashed lines. The red curves always show the linear interpolation between two possible realizations of the data values at the two points. However, in (a), (b), and (c), respectively, random values were generated using a correlation of $\rho = -1$, $\rho = 0$, and $\rho = 1$ between these values.

To analyze the variability of the orientation of the line connecting the values at $x = 0$ and $x = 1$, Fig. 5.3 (d), (e), and (f) show the probability density functions of the derivative of this line for the different correlation values. The probability density function was computed by taking the

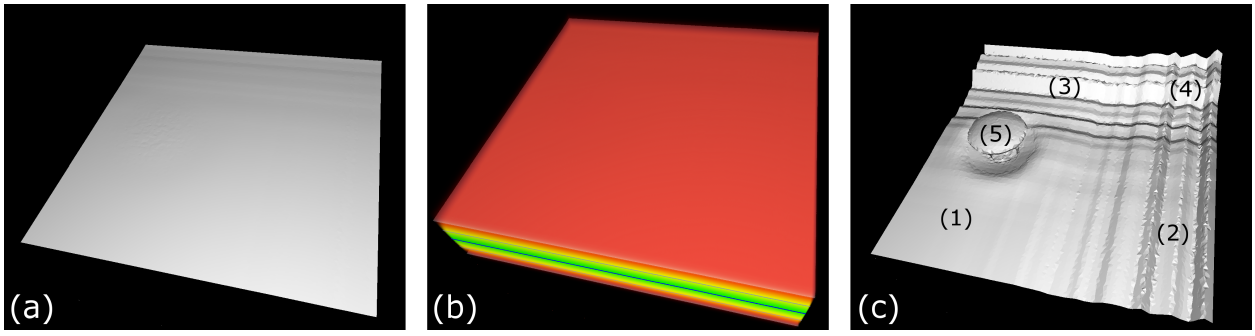

Figure 5.4: A mean iso-surface situated in the x-y-coordinate plane is shown in (a). For an assigned constant standard deviation the confidence region around the mean surface is visualized in (b). The structural variation, caused by the assigned correlation structure, is visible in the surface realization shown in (c). Strong homogeneous correlation was assigned to region (1). In (2) and (3) the correlation magnitude was lowered in x- and y- direction, respectively. In region (4), low correlation in both tangential direction was applied. Around (5), low correlation along the surface normal direction was modeled.

derivatives for all pairs of realizations at the two points and counting their occurrence. In (g), the relation between the correlation and the variance of the distribution of the derivative is shown. The maximum possible variation in the derivative is observed for $\rho = -1$ (inverse correlation), the minimum for $\rho = 1$.

From the experiment above it becomes clear that the correlation has a strong impact on the variability of structural properties of certain features in the data, such as a feature's orientation. The reason is that correlation describes the behavior of random variables at different sample points relative to each other and, thus, constraints the amount of variation between the possible data values at these points.

Experiment 3. The aforementioned considerations can easily be transferred to higher dimensions to analyze the structural variability of iso-surfaces in 3D scalar fields. To demonstrate this, a further experiment is pursued where at every vertex of a Cartesian 3D grid a set of random scalar values was generated using a multi-variate normal distributed random variable with constant standard deviation and linearly increasing mean (starting at 0) along the z- grid dimension. The random values were generated with respect to a specific anisotropic correlation structure. The computed mean values were then stored at the grid vertices and an iso-surface for a selected value was rendered (cf. Fig. 5.4 (a)). In (b), the confidence volume containing all points that belong to the surface with a certain probability is enclosed by two stochastic distance surfaces (transparent white). The blue line represents the contour of the mean surface, the colors encode the distance from the mean position in numbers of standard deviations σ — from 0σ (green) to $\pm 1\sigma$ (red). For details on this visualization mode, the reader is referred to paragraph 4.1.2. Notably, the visualization in (b) does not allow inferring on the possible structural variations of the surface, but it can only tell in which

region of the domain the surface can be expected with a certain probability.

In (c), the occurrence of the surface in (a) for one possible realization of the random values at the grid points is shown. In region (1), the correlations between the random variables in x-, y- and z-direction were all set to 1. Consequently, the structure of the mean surface is not changed (cf. Fig. 5.2 (a)). In region (2), the correlation in x-direction was lowered with respect to region (1). This results in oscillations of the surface in x-direction. As the correlation in y-direction is still 1, the structure does not change along the y-direction and stiff ridges occur along this direction. The same correlation setting was applied to region (3), but this time with a lower value in y-direction. In region (4), the correlations along x and y were lowered simultaneously, but the correlation in z-direction was kept at 1. This means that the surface can possibly oscillate, but since the values in the 3D scalar field are still highly constraint to each other along the z-direction, the surface remains connected. Moreover, the surface still indicates the exact location of the transition between scalar values less and greater than the selected iso-value.

The surface structure changed completely when the correlation in z-direction was made significantly smaller than the correlation in the other two directions. This is shown in region (5). The lower correlation in the surface normal direction can result in several closely located but not necessarily connected surfaces[1], i.e., the transition surface becomes fuzzy and does not have the separating property any more as it had before. The example shows that a visualization that is purely based on uncertainty parameters like standard deviations (cf. confidence volume in (b)) does not allow inferring on the geometric structures, which are likely to occur in realizations of the underlying probability distribution.

Experiment 4. The last experiment outlines the importance of correlation analysis in the context of *inverse problems*, e.g., 2D seismic travel-time tomography where physical measurements are used to infer on the media through which seismic wave forms are traveling. In this experiment, a geophysical setup is modeled, where receivers close to the earth's surface measure the arrival of pressure waves originating from earthquake sources with known positions in the earth's interior. The goal is to determine parameters describing the travel velocity of the waves in certain parts of the interior, based on measured wave travel times between each source-receiver pair.

According to geophysical realities, the seismic travel-time inverse problem is non-linear, as seismic waves do not follow straight paths in a heterogeneous medium and bend continuously as the velocity parameters of the medium vary. For simplification, these real physical conditions are abstracted and a linear inverse problem with straight pressure wave travel paths between source and receiver is assumed. This abstraction complies with the interpretation of the inversion as a first-order correction to a homogeneous medium. Thus, even though real non-linear physical conditions are

[1] The regular spherical surface structure results from the specific random number generation process that have been used to assign anisotropic correlation structure.

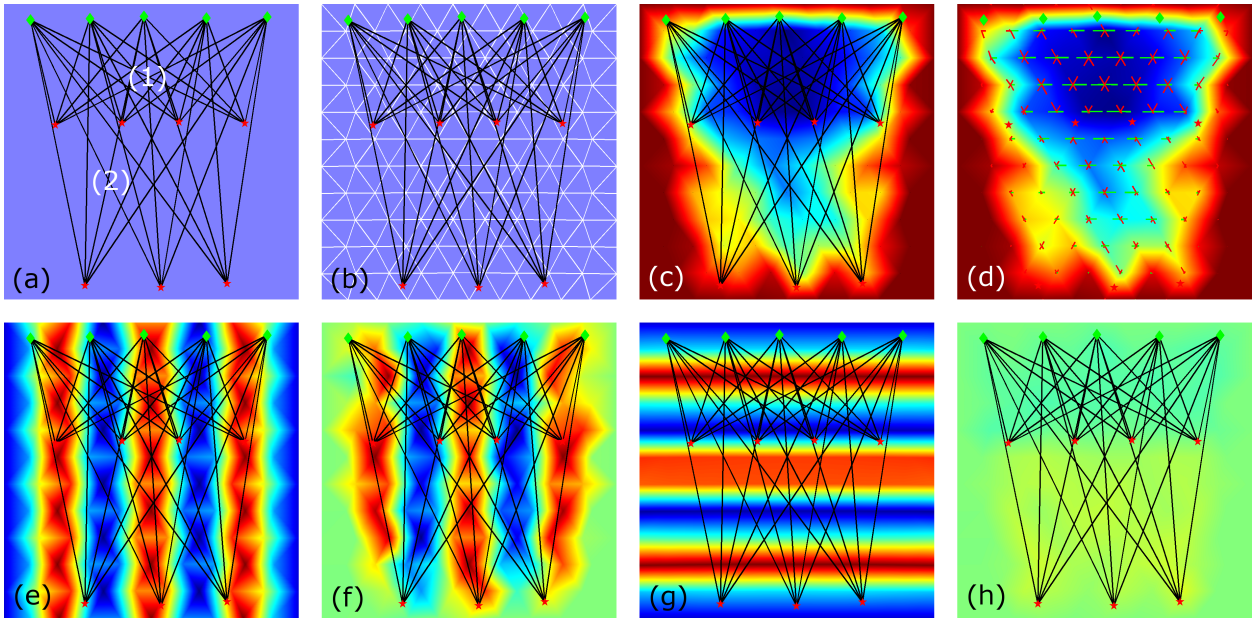

Figure 5.5: Travel-time tomography experiment. (a), (b) Emitter (red stars)-receiver (green diamonds) pairs and domain triangulation are shown. (c) Normalized standard deviations to mean solution of inverse problem is mapped from $[0, 1]$ to [blue $\rightarrow$ red]. (d) Visualization of correlation between adjacent grid points (line length corresponds to correlation strength, green/red indicate positive/negative correlation). (e), (g) Two different parameter configurations (color-coded) used in a forward computation. (f), (h) Inverse computation of parameters from the results of forward computation.

abstracted, the following considerations and interpretations can be adapted to every kind of inverse process step where a linearization takes place.

Mathematically, the abstract problem can be formulated by the system of equations

$$\mathbf{Gm} = \mathbf{d}, \tag{5.1}$$

were $\mathbf{G}$ represents the problem-specific linear operator, $\mathbf{d}$ is the linearized vector of measured data (travel-times), and $\mathbf{m}$ is the linearized vector of unknown parameters. The goal is to compute $\mathbf{m}$ for a given $\mathbf{d}$.

In travel-time tomography, the inverse problem is often under-determined and does not have a unique solution. For linear inverse problems and by using *Bayesian probability theory*, however, the solution $\mathbf{m}$ can be described by a multi-variate normal probability density function. This function is characterized by a mean $\bar{\mathbf{m}}$ and a covariance matrix $\mathbf{C_m}$. For more details on the physical theory and model setup, the reader is referred to [Nol08] and [Tar05].

Fig. 5.5 (a) shows a rectangular domain discretized by a triangular grid (shown in (b)), which models a cross section of the earth. The components of the parameter vector $\mathbf{m}$ are placed at the grid

vertices. The parameters define for each vertex a slowness value (inverse of velocity: time/distance). In the upper and lower part of the domain, respectively, 5 receivers (green diamonds) and 7 sources (red stars) are placed. The travel paths are illustrated by black lines. The travel time for each path is obtained by integrating the slowness parameters along the path (linear interpolation in each triangle). As all parameters in $\mathbf{m}$ appear linear in the integrals (Simpson quadrature), the integrals can be expressed by a matrix $\mathbf{G}$ representing the operator in equation (5.1). The vector $\mathbf{d}$ contains the travel times for all paths.

Intuitively, one would expect that in regions which are covered by a large number of rays, e.g., in (1), there is a lower uncertainty in the computed slowness values than in regions which are poorly sampled by rays, e.g., in (2), since less information is obtained from there. To confirm this, first, the covariance matrix $\mathbf{C_m}$ from the given operator $\mathbf{G}$ is computed (for details, the reader is referred to [Tar05]), and the standard deviation values are derived from the probabilistic solutions for the slowness values from this matrix, i.e., the square roots of entries on the main diagonal of the covariance matrix. In (c), these normalized standard deviations are color-coded via the mapping $[0, 1] \rightarrow$ [blue $\rightarrow$ red]. The coloring clearly indicates that regions with higher ray density are affected by lower uncertainty (close to zero).

To demonstrate the importance of correlation for the analysis of the uncertainty in the computed results, the following experiment is performed: Firstly, two artificial slowness configurations $\mathbf{m}_h$ and $\mathbf{m}_v$ with periodic horizontal and vertical change of slowness values are specified, respectively, and forward computations are performed to obtain data vectors $\mathbf{d}_h$ and $\mathbf{d}_v$. In (e) and (g), these slowness parameters are color-coded. Secondly, with $\mathbf{d}_h$ and $\mathbf{d}_v$ two inverse problems are solved, resulting in two probabilistic solutions for the slowness values. Their means, $\bar{\mathbf{m}}_h$ and $\bar{\mathbf{m}}_v$, are color-coded in (f) and (h) utilizing the same color mapping as in (e) and (g). As can be observed, only $\bar{\mathbf{m}}_h$ is close to its initial distribution, while $\bar{\mathbf{m}}_v$ shows almost no similarity or relevant structure, although the uncertainty (c) in both solutions is the same.

In order to understand this phenomenon, one has to look at the correlations between the values at adjacent points, which are contained in the covariance matrix $\mathbf{C_m}$. Therefore, in (d), the pairwise correlations between the values at every grid point and its one-ring neighbors are visualized. Visualization is performed by drawing lines from the vertex to every adjacent vertices, where the lengths of these lines are linearly related to the magnitude of the respective correlation value. For $|\rho| = 1$ the lines on either side of an edge meet at the edge's center. The lines are colored green for $\rho \geq 0$ and red for $\rho < 0$. The visualization clearly indicates strong positive correlations in horizontal direction and strong negative correlations in vertical direction.

As discussed above, for strong positive correlations, high frequent structures in the data are better resolved, as they are only affected by low frequent uncertainty. However, for inverse correlated regions, no reliable assumptions can be made. Thus, the example clearly shows that only structural changes in horizontal direction can be resolved by the inverse computation. A variation of

parameters in vertical direction cannot by "seen" by this tomographic setup. Thus, the correlation visualization in (d) already provides a clear indication that the data set is affected by a strong vertical but only a low horizontal *structural uncertainty*. This information is not revealed by visualizing only absolute uncertainty values (cf. standard deviations in (c)). As a consequence, it can be concluded that the consideration of correlation information and its visualization can significantly help to avoid misinterpretations that would occur otherwise.

5.1.3 Correlation Visualization

In the previous section, the relevance of correlation for uncertainty analysis have been motivated. As a consequence, the correlation visualization is believed to be a necessary ingredient in uncertainty visualization to facilitate a more reliable prediction of the possible effects of uncertainty on specific features in uncertain data. In this section, light is shed onto the particular requirements for correlation visualization, and a correlation visualization technique is proposed that addresses some of these requirements.

Requirements

In real-world applications, correlation data is often given as correlation matrices which can be either computed directly (cf. Example 4 in section 5.1.2) or estimated from an ensemble data set. If the data is given on a grid with n grid points, a correlation matrix would have n^2 entries, of which $0.5n(n-1)+n$ would be different. For large multi-dimensional grids, this $O(n^2)$ memory and computational complexity has to be reduced significantly before correlation visualization becomes feasible.

However, one difficulty in correlation analysis is that both local and global correlation effects have to be considered and the anisotropic nature of correlation makes it difficult in general to represent it with only a few values. By restricting the analysis to local effects, the first problem can be addressed. A possible solution for the second problem builds upon the development of a correlation model that can efficiently represent anisotropic structures, for instance, by trying to approximate these structures via a suitable basis transformation into a compact representation. Such a model can then be employed for the visualization of correlation, but additional requirements have to be considered. In particular, the direction dependent differences in correlation strength have to be depicted by an appropriate visual mapping. Such a mapping, besides being able to indicate the correlation ratios into different directions, should also allow a comparison of the absolute correlation strengths at different regions in the underlying domain.

Another challenge in correlation visualization arises from the general difficulty of finding a suitable visual mapping for complicated multi-dimensional structures. Even without considering uncer-

tainty, the creation of approaches for mapping such structures onto 2D pixel images in an intuitive way is extremely difficult, and it is by far not clear how additional correlation structures can be integrated into conventional uncertainty visualization approaches. Therefore, as it have been demonstrated in the third experiment in section 5.1.2, one strategy is to restrict the correlation analysis to specific features in the data, such as iso-surfaces in 3D scalar fields. In such a scenario, one can restrict the analysis to the local correlations along and in close vicinity to the feature, possibly augmented by contextual visualizations of the surrounding structures.

Distance Dependent Correlation

In the following, a *local anisotropic distance dependent correlation model* is introduced, which forms the basis of the proposed visualization approach. This model has some specific advantages, and it allows transforming correlations that are expressed in other models into this model straight forwardly. It considers the strength of positive local correlations for analyzing a given data set with respect to structural uncertainty as discussed in section 5.1.2.

A 3D Cartesian grid structure $\mathbb{C}_{pqr} = (\mathbf{x})_{ijk}$ with $1 \leq i \leq p, 1 \leq j \leq q, 1 \leq k \leq r$ is assumed, which is attributed by a mean value and mutual correlation values for neighboring vertices at every grid vertex. This data can be linearly interpolated from values in any arbitrary grid structure. The grid spacing can be specified interactively by the user in order to achieve different glyph resolutions.

This visualization approach is aiming for an analysis of correlation effects in 3D scalar data sets. Usually, such data sets are assumed to be at least continuous and smooth up to a certain degree. This requires that the closer two spatial points in the underlying domain are, the higher the correlation should be between the random variables that are used to model the uncertainty at these points. To achieve this, the correlation is often modeled by a *spatial distance dependent correlation function* [Tar05]. One typically uses the *exponential correlation function* (ECF)

$$\rho(Y(\mathbf{x}_{ijk}), Y(\mathbf{x}_{lmn})) = \exp(-\tau \left\| \mathbf{x}_{ijk} - \mathbf{x}_{lmn} \right\|) \quad , \quad \mathbf{x}_{ijk}, \mathbf{x}_{lmn} \in \mathbb{C}_{pqr}, \tag{5.2}$$

which assigns higher correlations to random variables of points with smaller Euclidean distance and lower correlations to points which are more distant from each other. If the correlation strength parameter τ is defined locally for each point in $\mathbb{C}_{pqr}$, the ECF becomes

$$\rho(Y(\mathbf{x}_{ijk}), Y(\mathbf{x}_{lmn})) = \exp(-0.5(\tau(\mathbf{x}_{ijk}) + \tau(\mathbf{x}_{lmn})) \left\| \mathbf{x}_{ijk} - \mathbf{x}_{lmn} \right\|). \tag{5.3}$$

The ECF meets the intuitive assumption that scalar realizations behave more similar if the respective sample points are closer in Euclidean space. This is considered as a kind of local stochastic continuity.

The distance dependent correlation model assumes isotropic correlations at every grid point. To model anisotropic correlations, τ can be made dependent on a specific direction. For a unit vector $\mathbf{r}$, the parameter τ at point $\mathbf{x}_{ijk}$ in direction $\mathbf{r}$ is then given by $\tau(\mathbf{x}_{ijk}, \mathbf{r}) = \mathbf{r}^\top \mathbf{T}(\mathbf{x}_{ijk})\mathbf{r}$, where $\mathbf{T}$ is the rank-2 *parameter tensor* that models the anisotropy. This tensor can either be derived from the correlations in the data samples, or it can be specified based on a priori knowledge. The adapted correlation model becomes

$$\rho(Y(\mathbf{x}_{ijk}), Y(\mathbf{x}_{lmn})) = \exp\left(-\frac{(\mathbf{x}_{ijk} - \mathbf{x}_{lmn})^\top \overline{\mathbf{T}}(\mathbf{x}_{ijk} - \mathbf{x}_{lmn})}{\|\mathbf{x}_{ijk} - \mathbf{x}_{lmn}\|}\right), \tag{5.4}$$
$$\overline{\mathbf{T}} = 0.5(\mathbf{T}(\mathbf{x}_{ijk}) + \mathbf{T}(\mathbf{x}_{lmn})).$$

One advantageous use of the distance dependent tensor model is the possibility to transform correlation data into this model and, thus, to avoid the explicit storage of a correlation matrix. Therefore, at every grid point $\mathbf{x}_{ijk}$ in $\mathbb{C}_{pqr}$ a tensor is computed from the correlations to its (at most) 26 neighbors

$$\mathbb{N}(\mathbf{x}_{ijk}) := \{\mathbf{x}_{lmn} \in \mathbb{C}_{pqr} | \mathbf{x}_{lmn} \neq \mathbf{x}_{ijk}, \max(|i-l|, |j-m|, |k-m|) = 1\}. \tag{5.5}$$

If $\mathbf{x}_{ijk}$ is not a border point of the grid $\mathbb{C}_{pqr}$, then for every point $\mathbf{n}_h \in \mathbb{N}(\mathbf{x}_{ijk})$ there exists an opposing point $\hat{\mathbf{n}}_h \in \mathbb{N}(\mathbf{x}_{ijk})$ such that $\mathbf{x}_{ijk} = 0.5(\mathbf{n}_h + \hat{\mathbf{n}}_h)$. At most 13 such pairs of $\mathbf{n}_h$ and $\hat{\mathbf{n}}_h$ with $h \in \{1, 2, ..., 13\}$ can be built.

In Fig. 5.6, the 26 neighbors of a non-border point $\mathbf{x}_{ijk}$ (red) are illustrated. Furthermore, two opposing neighbor points $\mathbf{n}_h$ and $\hat{\mathbf{n}}_h$ (green) are highlighted.

By using the correlation model (5.2) and averaging the correlation against opposing neighbors one obtains 13 equations:

$$\mathbf{r}^\top \mathbf{T} \mathbf{r} = \frac{-\log\left[0.5(|\rho(Y(\mathbf{x}_{ijk}), Y(\mathbf{n}_h))| + |\rho(Y(\mathbf{x}_{ijk}), Y(\hat{\mathbf{n}}_h))|)\right]}{\|\mathbf{n}_h - \mathbf{x}_{ijk}\|}, \tag{5.6}$$
$$\mathbf{r} = \frac{\mathbf{n}_h - \hat{\mathbf{n}}_h}{\|\mathbf{n}_h - \hat{\mathbf{n}}_h\|} \quad , \quad \mathbf{n}_h \in \mathbb{N}(\mathbf{x}_{ijk}), \quad h \in \{1, 2, ..., 13\}.$$

If $\mathbf{x}_{ijk}$ is a border point, not for every $\mathbf{n}_h \in \mathbb{N}(\mathbf{x}_{ijk})$ exists an opposing point. In this case, averaging is not performed in (5.6) and only the logarithm $\log(|\rho(Y(\mathbf{x}_{ijk}), Y(\mathbf{n}_h))|)$ is used. Note that only the magnitudes of the correlations are used as the ECF only models the correlation strength. For integrating inverse correlation, another (global) model would be necessary, which will be discussed in section 5.2.

The entries of the 3×3 matrix $\mathbf{T}$ are denoted as t_{ij}. As $\mathbf{T}$ is symmetric and represents a rank-2 tensor, only the 6 values $t_{11}, t_{22}, t_{33}, t_{12}, t_{13}, t_{23}$ have to be determined. These values appear linearly in the 13 (or less if $\mathbf{x}_{ijk}$ is a border point) equations (5.6). So one can build a 13×6

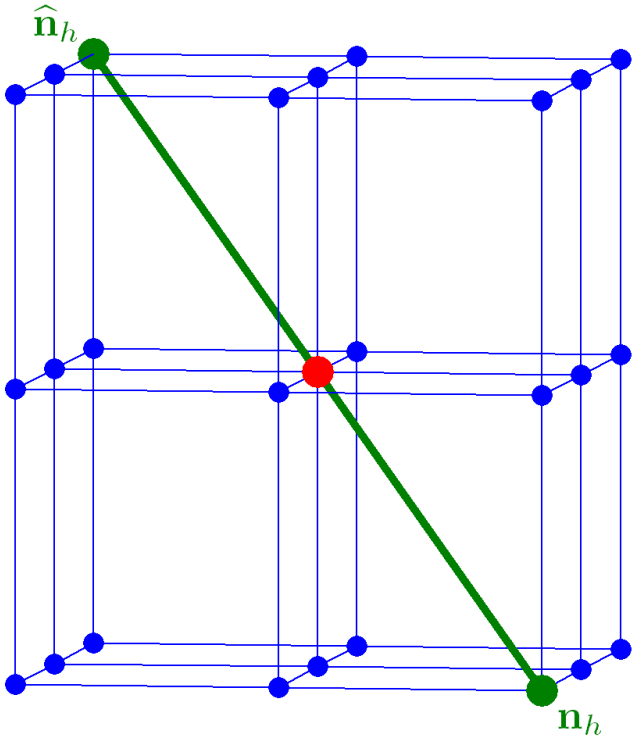

Figure 5.6: The 26-neighborhood of a chosen spatial grid point (red) is shown, highlighting two opposite neighbor points (green).

over-determined linear system, which can be solved using a least squares approach and the normal equation [Bjo96].

The ECF model provides several advantages, amongst others also for visualization purposes. First of all, using a local approach reduces the memory requirement for storing correlation information to some extend: If the correlation grid has $m = |\mathbb{C}_{pqr}| = pqr$ entries, the respective correlation matrix has to store m^2 local and global correlation values. If only local values are stored, approximately 13 values per point in $\mathbb{C}_{pqr}$ would be needed for the 26-element neighborhood system described above. This amounts to a total of $13m$ correlation values. The rank-2 tensor model halves this amount to a total of $6m$ values. This meets the proposed requirement from section 5.1.3 for a memory consumptions of $O(m)$.

An additional reason for using the ECF model is its integration of Euclidean distances. To understand this, let us assume that instead of the local τ tensor, the local correlation values are used at every grid vertex. In this case, the local correlation in every direction depends on the resolution of the correlation grid $\mathbb{C}_{pqr}$, and if this grid has a high resolution, the correlation between neighboring points would approach 1 everywhere. Notably, the computation of $\mathbf{T}$ is independent of the grid resolution, as the Euclidean distances to the neighbors are taken into account for computing $\mathbf{T}$. Thus, correlation is set in relation to Euclidean distances. Furthermore, using the parameter tensor

allows the computation of correlations for a user-specified distance (cf. next paragraph). With a stored correlation tensor for a fixed distance this would not be possible.

Anisotropic Correlation Glyphs

The rank-2 correlation tensors are always positive semi-definite and symmetric. Thus, they can be decomposed into three orthonormal eigenvectors, which point into the directions of the first, second and third principal axes of the associated correlation ellipsoid. Regarding the model described in (5.2), the eigenvalues give the values taken by τ in the respective eigen-direction. The decomposition can be written as

$$\mathbf{T} = \mathbf{V}\mathbf{S}\mathbf{V}^\top, \tag{5.7}$$

with matrix $\mathbf{V}$ containing the orthonormal eigenvectors (principal components of the ellipsoid) in the columns and the diagonal matrix $\mathbf{S}$ containing the eigenvalues on the diagonal.

As the eigenvectors are not related in general to an iso-surface structure in the data, the tensor $\mathbf{T}$ is not directly suitable for analyzing the structural uncertainty of iso-surfaces. To overcome this limitation, a basis transformation $\mathbf{T} = \widehat{\mathbf{V}}\widetilde{\mathbf{S}}\widehat{\mathbf{V}}^\top$ is performed first. The tensor $\mathbf{T}(\mathbf{x}_{ijk})$ at grid point $\mathbf{x}_{ijk}$ is expressed using the normalized data gradient $\mathbf{g}(\mathbf{x}_{ijk})$ and two orthonormal vectors $\mathbf{w}_1(\mathbf{x}_{ijk})$ and $\mathbf{w}_2(\mathbf{x}_{ijk})$. These two vector are elements of unit vectors in the respective tangent plane

$$\mathbb{T}(\mathbf{x}_{ijk}) := \{\mathbf{w} \in \mathbb{R}^3 | \mathbf{w}^\top \mathbf{g}(\mathbf{x}_{ijk}) = 0, \|\mathbf{w}\| = 1\}. \tag{5.8}$$

The vectors $\mathbf{g}$, $\mathbf{w}_1$, and $\mathbf{w}_2$ form the columns of $\widehat{\mathbf{V}}$ and represent the new basis. It is worth noting here, that the new matrix $\widetilde{\mathbf{S}}$ is *not* diagonal.

With the fixed vector $\mathbf{g}$ there is still one degree of freedom for the orientation of the orthonormal pair $\mathbf{w}_1$ and $\mathbf{w}_2$ in the tangent plane at point $\mathbf{x}_{ijk}$. By definition, $\mathbf{w}_1$ and $\mathbf{w}_2$ should point into the direction of maximum and minimum τ in the tangent plane:

$$\mathbf{w}_1(\mathbf{x}_{ijk}) := \underset{\mathbf{w}\in\mathbb{T}(\mathbf{x}_{ijk})}{\arg\max} \left(\mathbf{w}^\top \mathbf{T}\mathbf{w}\right), \tag{5.9}$$

$$\mathbf{w}_2(\mathbf{x}_{ijk}) := \underset{\mathbf{w}\in\mathbb{T}(\mathbf{x}_{ijk})}{\arg\min} \left(\mathbf{w}^\top \mathbf{T}\mathbf{w}\right). \tag{5.10}$$

In Fig. 5.7, a tangent plane (green) for a particular point $\mathbf{x}_{ijk}$ on an iso-surface (red) is illustrated. The tensor $\mathbf{T}(\mathbf{x}_{ijk})$ is represented by a 3D ellipsoid (blue). The maximum and minimum τ-directions are shown as magenta and cyan lines in the tangent plane. The orthogonal gradient direction is illustrated by a yellow line.

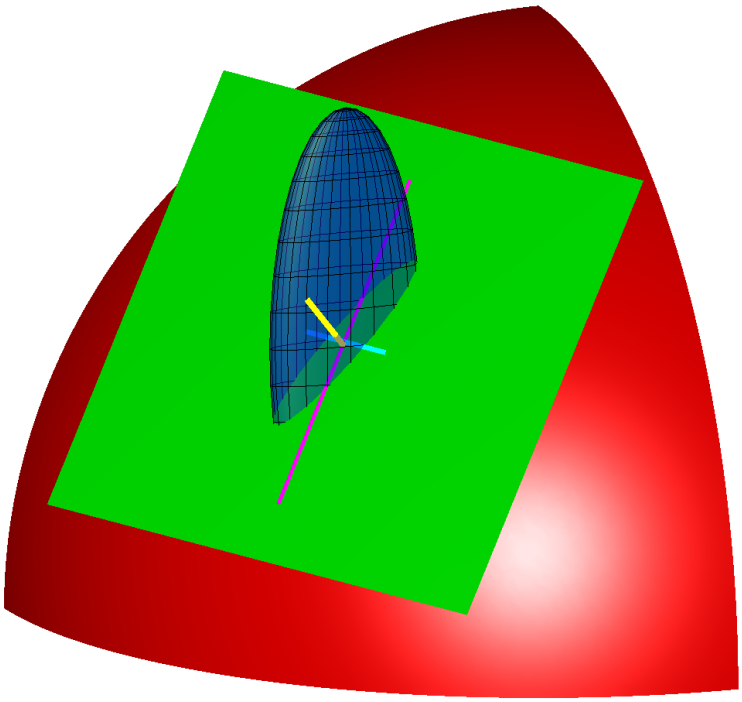

Figure 5.7: The tangent plane (green) for a particular point on an iso-surface (red) is shown. The correlation strength tensor is illustrated by a 3D ellipsoid (blue). The maximum and minimum τ-directions are shown as magenta and cyan lines in the tangent plane. The orthogonal gradient direction is illustrated by a yellow line.

In order to compute $\mathbf{w}_1$ and $\mathbf{w}_2$, first, two arbitrary orthonormal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ are defined, which span the tangent plane at $\mathbf{x}_{ijk}$, as follows:

$$\mathbf{n}_1 := \frac{(g_2, -g_1, 0)^\top}{\sqrt{g_1^2 + g_2^2}}, \quad \mathbf{n}_2 := \mathbf{g} \times \mathbf{n}_1. \tag{5.11}$$

Here, g_i denotes the i-th component of the normalized gradient $\mathbf{g}$. The two spanning vectors are put into the columns of the 3×2 projection matrix $\mathbf{N} := [\mathbf{n}_1|\mathbf{n}_2]$. The tensor $\mathbf{T}$ is projected into the tangent plane by $\mathbf{N}^\top\mathbf{T}\mathbf{N}$. A singular value decomposition of the resulting matrix results in two 2D orthonormal singular vectors $\mathbf{s}_1$ and $\mathbf{s}_2$. The vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ can now be obtained as

$$\mathbf{w}_1 = \mathbf{N}\mathbf{s}_1, \quad \mathbf{w}_2 = \mathbf{N}\mathbf{s}_2. \tag{5.12}$$

The values of τ into the three directions are obtained as

$$\tau_g = \mathbf{g}^\top\mathbf{T}\mathbf{g}, \quad \tau_1 = \mathbf{w}_1^\top\mathbf{T}\mathbf{w}_1, \quad \tau_2 = \mathbf{w}_2^\top\mathbf{T}\mathbf{w}_2. \tag{5.13}$$

The setting is such that $\mathbf{w}_1$ is pointing into the direction of maximum τ. According to (5.2), this

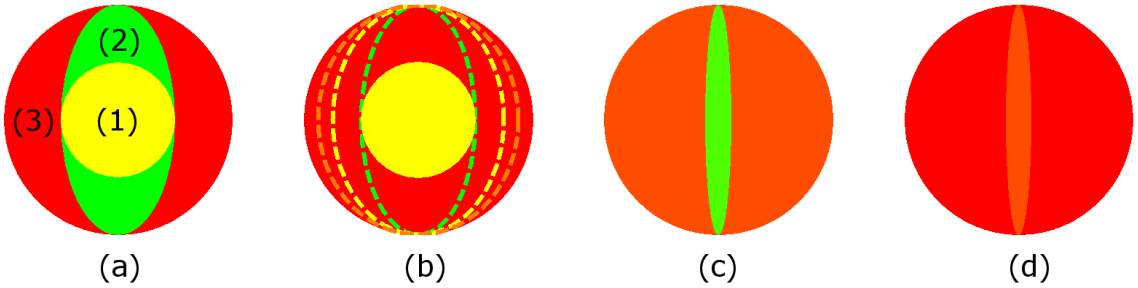

Figure 5.8: (a) The used correlation glyph is shown. Absolute correlation strengths are mapped from $[0, 1]$ to [red $\rightarrow$ green] for surface normal (1), first (2) and second (3) principal tangential correlation directions in three zones. (b) Geometry of zone (2) visualizes the correlation ratio between the two tangential direction. (c) Glyph for high contrast tangential correlation values 0.9 (2) and 0.1 (3) is shown. (d) Glyph for low contrast tangential correlation values 0.09 (2) and 0.01 (3) is shown.

is the direction of minimum correlation. Vector $\mathbf{w}_2$ is pointing into the direction of maximum correlation. These two tangent vectors, together with $\mathbf{g}$ and the parameters τ_g, τ_1, and τ_2 are used for visualizing the anisotropic correlation relations as well as their absolute strengths.

To visualize anisotropic correlation structures, at every grid point a correlation glyph is constructed from the distinct 3D orientation given by the iso-surface normal $\mathbf{g}$ and the two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. This glyph is then used to visualize the correlation ratio in the tangent plane (anisotropy) as well as absolute correlation values for the three directions.

To construct the correlation glyphs, first, correlation values along the three directions $\mathbf{g}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ have to be derived. Therefore, the user selects a specific Euclidean distance d interactively from the interval $[0, d_{max}]$, and at every grid point the correlation values are computed as

$$\rho_g = \exp(-\tau_g d) \quad , \quad \rho_1 = \exp(-\tau_1 d) \quad , \quad \rho_2 = \exp(-\tau_2 d). \tag{5.14}$$

So each glyph visualizes the correlation distribution around its center for the radius d. For the construction of the glyph, at every grid vertex a circle is centered at this vertex and aligned with the respective tangent plane. To avoid overlapping glyphs, the circle's radius is set to half the grid spacing.

Within each circle a glyph is placed as shown in Fig. 5.8 (a). Each glyph consists of three zones, where each zone receives a color from the range [red $\rightarrow$ green], indicating an absolute correlation magnitude value from 0 to 1. The mentioned color map is utilized in order to enhance the contrast for absolute correlation values and to relate "stable" regions to "green" and more "unstable" or independent regions to "red".

Zone (1) is a circle with an interactively adjustable radius r_1. Its color represents the correlation ρ_g in surface normal direction. The colors of zone (2) and (3) encode the values ρ_2 for maximum and ρ_1 for minimum correlation in the tangent plane. Zone (3) forms always a circle with a fixed radius r_3. Zone (2) is an ellipse, which is oriented along the maximum correlation axis $\mathbf{w}_2$ and has an radius in this direction always equal to r_3. The radius in direction $\mathbf{w}_1$ is equal to $(1 - \frac{\rho_1}{\rho_2})r_1 + \frac{\rho_1}{\rho_2}r_3$. So the alignment of the ellipse with respect to the radii r_1 and r_3 directly visualizes the correlation ratio in the tangent plane. Compare the ellipse variations in Fig. 5.8 (b). Note that if $\rho_1 = \rho_2$, then zone (2) fully covers zone (3).

The introduced correlation glyphs have several strong advantages, oriented on the requirements presented in 5.1.3. First of all, it is possible to show the correlation anisotropy in the tangent plane as well as absolute correlation values (coded in color in the three zones) simultaneously in one picture. Showing only correlation ratios would not be sufficient. For instance, the two correlation values 0.9 and 0.1 have a ratio of 9:1, but 0.09 and 0.01 would also have a ratio of 9:1. As the user is predominantly interested in ratios where the *absolute difference* between both values is also large (0.8 compared to 0.08 in our example), the second case is not really interesting.

The two cases are shown with the glyph-visualization in Fig. 5.8 (c) and (d). In both cases, the geometry and the radii of the elliptic zone (2) are the same, but due to the lower color contrast in (d) (colors indicate correlations 0.01 and 0.09), the ellipse is almost vanishing. So the anisotropy is only visible in regions, where $|\rho_1 - \rho_2|$ is considerable large and the user's focus is not distracted by glyphs indicating high ratios of very low correlation values. This is a strong advantage in comparison to other tensor-glyph visualization approaches.

Furthermore, the user is able to adjust the radius r_1 of zone (1) between 0 and r_3. For $r_1 = 0$ the elliptic zone (2) has the largest range of variations. This is best for illustrating the correlation anisotropy in the tangent plane. For $r_1 = r_3$ zone (1) covers zone (3) completely, zone (2) is not perceivable any more and only the absolute correlation value in surface normal direction is visualized. So, while interactively changing the radius r_1, the user has the possibility of putting the focus either on the analysis of the correlation anisotropy along the surface or the strengths in normal direction — or a mixture of both.

If the structural variability of an iso-surface in a 3D scalar field for a particular iso-value θ has to be analyzed, not for every grid vertex a correlation glyph should be visualized. Instead, the visualization should be restricted to only those glyphs close to the surface. Therefore, the visualization of a glyph is made dependent on the mean value $\mu(\mathbf{x}_{ijk})$ at the respective grid vertex: The glyph at position $\mathbf{x}_{ijk}$ is rendered if

$$\begin{aligned} \theta &\geq \min(\{\mu(\mathbf{x}_{ijk}), \mu(\mathbf{x}_{i\pm1jk}), \mu(\mathbf{x}_{ij\pm1k}), \mu(\mathbf{x}_{ijk\pm1})\}), \\ \theta &\leq \max(\{\mu(\mathbf{x}_{ijk}), \mu(\mathbf{x}_{i\pm1jk}), \mu(\mathbf{x}_{ij\pm1k}), \mu(\mathbf{x}_{ijk\pm1})\}). \end{aligned} \tag{5.15}$$

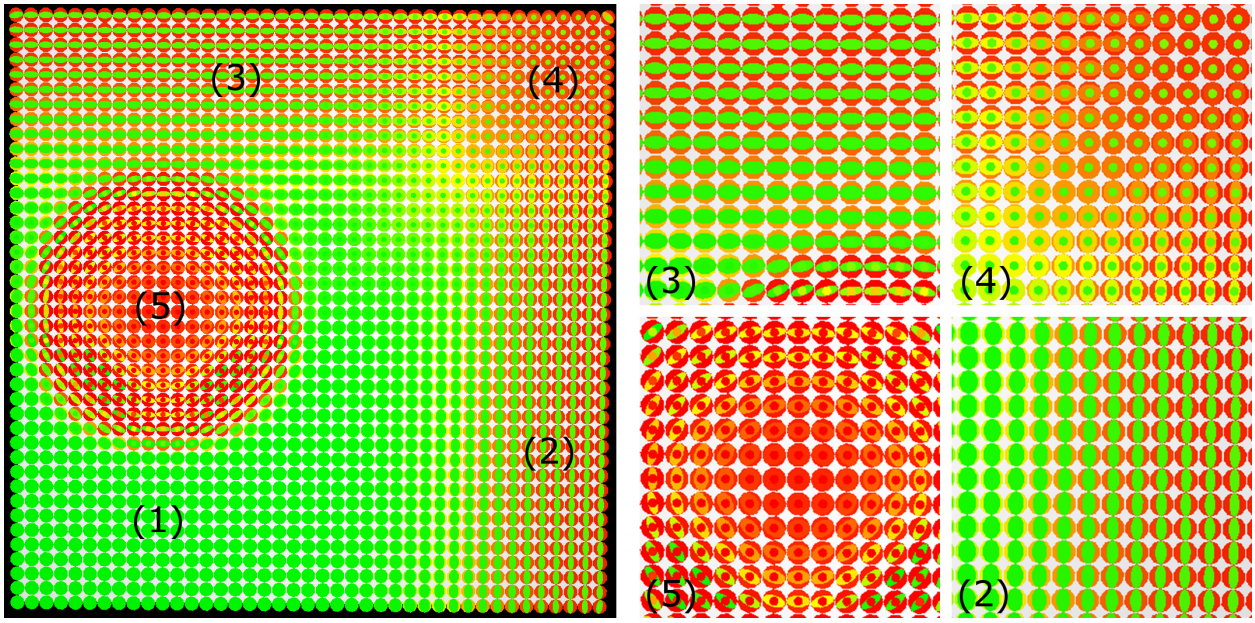

Figure 5.9: Top view on correlation visualization on mean surface from example in Fig. 5.4. Strong local correlations in (1) indicate low structural uncertainty. Glyph geometries reveal high anisotropic correlation difference for regions (2) and (3). High contrast between strong correlation strengths in normal direction and lower ones in tangential directions are shown in (4). Region (5) is affected by low normal correlation, indicated by red glyph center zones. The images on the right show magnifications of four regions.

This rule was chosen to reduce the overlapping of glyphs. The regular placing of the glyphs at the vertices of the 3D grid facilitates the visual perception of anisotropic correlation contrasts compared to approaches where the glyphs have variable distances between each other.

Discussion

To validate the effectiveness of the proposed novel visualization technique for analyzing the structural variability of iso-surfaces in uncertain 3D scalar fields, two experiments using different data sets have been conducted: A stochastically simulated data set with a synthetic setting (cf. Fig. 5.4) and a real-world data set comprised of an ensemble of simulated weather forecasts.

Fig. 5.4 demonstrates the structural variability of an iso-surface depending on different correlation settings. Fig. 5.9 shows the visualization of the assigned correlation structures using the glyph-based visualization approach. Here, the glyph geometry clearly reveals the correlation ratios. The four images on the right show magnifications of four interesting regions. In region (1), the glyphs indicate an isotropic correlation distribution. In (2) and (3), the directions along which the correlation is high can be clearly perceived. In region (4), the occurrence of glyphs having

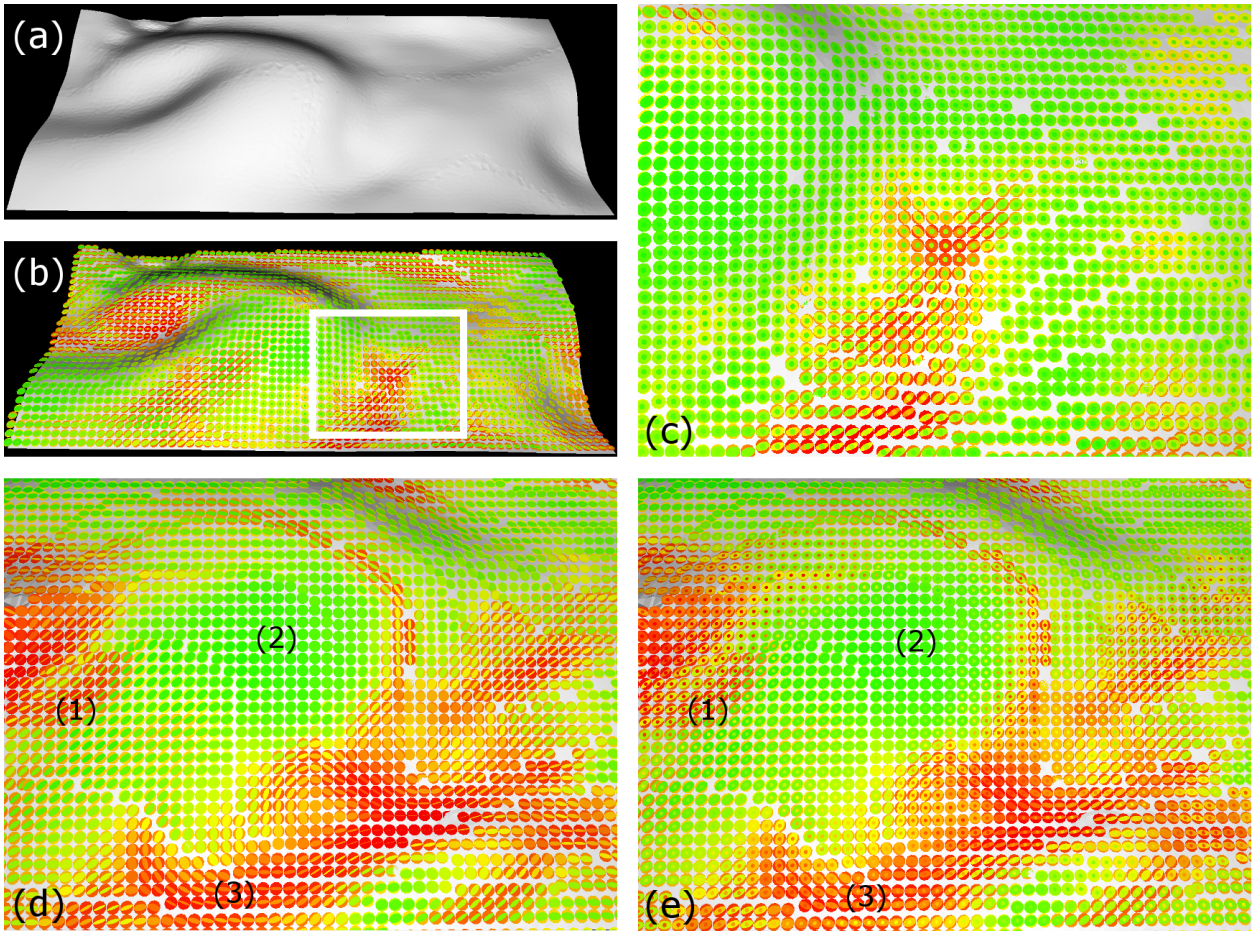

Figure 5.10: (a) A mean temperature iso-surface in a 3D scalar ensemble weather forecast data set is shown. (b) Correlation glyphs reveal anisotropic correlation structures in the surface's vicinity. (c) Close-up view on rectangular region shows region with low tangential correlation (red outer glyph zones) and strong normal correlation values (green inner zone). (d) A region with strong homogeneous correlation and low structural uncertainty thereof (2) is visually separated from low and anisotropic correlation structures in (1) and (3). (e) Additional correlation values in surface normal direction are coded in color in glyph center zones.

green center zones surrounded by red zones indicates a high ratio between the correlations in x-/y-direction (tangent plane, low correlation) and surface normal direction (high correlation). The reverse behavior is visualized in region (5), i.e., low correlation along the normal direction.

The visualization in Fig. 5.10 shows a 3D temperature field in the exosphere above Europe and the North Atlantic Ocean in spherical coordinates. It contains the mean values of multiple fields that were simulated by the European Center for Medium-Range Weather Forecasts using slightly different input parameters. The image (a) shows an iso-surface in the mean temperature field. In (b), the glyph-based correlation visualization is used to depict structural variation effects. In (c),

a close-up view on the region marked by the white rectangle is shown. The green glyph center and the red outer zones indicate a high ratio between the correlation strengths in tangential and normal directions. In such regions, a higher stochastic independence between points in the tangential plane can be concluded. The surrounding regions, which are covered by homogeneous green glyphs, are much less likely affected by structural uncertainty.

Fig. 5.10 (d) and (e) reveal the interplay between correlation ratios and color-based coding of correlation strengths. In (d), only the correlation anisotropy in the tangent planes is visualized, but the correlation in normal direction is ignored. For a tongue-shaped region around (2), strong homogeneous local correlation close to 1 and low structural uncertainty is indicated by uniformly colored green glyphs. In region (1), a high contrast between the two glyph zones (green and red) indicate a strong correlation anisotropy (almost 1 for the first tangential direction (green) and almost 0 for the second tangential direction (red)). Although the glyph geometry reveals the same correlation ratio as, for instance, in region (3), the color contrast between the two glyph zones emphasizes the absolute correlation differences. In (e), the interior glyph zones for the absolute correlation values in normal direction are shown. One can observe that there is very low correlation along the normal direction around the stable homogeneous region (2), indicated by the red center zones. This low correlation in normal direction indicates rather high stochastic instability, meaning that iso-surfaces in these regions are very likely to separate into multiple disconnected parts (cf. region (5) in Fig. 5.4 (c)).

From the given examples some general advantages of the proposed glyph-based approach (based on the anisotropic correlation model) for correlation visualization can by concluded:

- The local distance dependent correlation model, used to parameterize the glyph shape and color, allows for a memory reduction to $O(n)$ correlation/tensor values for n grid points.
- The correlation parameter tensor model accounts for the anisotropic nature of local correlation strengths.
- The glyph geometry clearly indicates correlation ratios in the surface tangent planes and effectively emphasizes the predominant correlation direction.
- The color mapping scheme intuitively reveals the absolute correlation values in surface normal and the two tangential directions. It further allows differentiating between interesting anisotropic correlation regions (low and high correlation in tangent plane) and less interesting regions where the correlation in both directions is low, which however could have the same ratios (cf. Fig. 5.8 (c) and (d)).
- Since the glyph design (variable radius for normal correlation zone) can be changed interactively, the user can flexibly change the focus between a visualization, more concentrated on the anisotropic structures in the tangent planes or on anisotropic structures between the

tangent plane and the surface normal direction. So the structural uncertainty can be assessed interactively for different spatial directions.

Despite the advantages, however, the presented approach also suffers from some limitations and leaves several open questions and requirements:

- The approach does not show any absolute uncertainty information like standard deviation, which is mandatory for a more comprehensive uncertainty analysis. Here a thorough analysis of an adequate visualization channel (color, opacity, etc.) would be necessary, as a straightforward merge with an intuitive uncertainty visualization approach is often not possible.
- The glyph-based approach is based on a local correlation model and, thus, can only show the correlation to regions close to the position where a glyph is placed. Global correlation structures are ignored.
- Inverse correlation structures cannot be visualized, as the chosen correlation model only accounts for correlations strengths (magnitudes). Here, a global correlation model and a different color scheme would be necessary.
- Both the glyph geometry and the color-coding refer to the correlation between the values in a certain distance to each other. Even though this distance can be selected interactively and arbitrarily, it does not appear in the visualization.

From the analysis of the advantages and limitations of the glyph-based approach for correlation visualization it can be concluded that the design of techniques allowing for an intuitive and effective correlation visualization is not straightforward. The integration of correlation information into visualizations of complicated multi-dimensional structures is extremely challenging due to the limited number of visualization channels (e.g., color, opacity, geometry, etc.). Although the presented approach cannot fulfill all requirements on a correlation visualization technique, it can be seen as a first step towards such an integration and is meant to stimulate further research into this direction.

5.1.4 Conclusion

In this section, basic insights into the concept of correlations in uncertain scalar fields are provided. Furthermore, the importance for integrating correlation data into uncertainty visualization approaches is stressed in order to make more reliable assumptions on the structural uncertainty of features and to get a more comprehensive understanding of the effects uncertainty exerts on the underlying data set. In addition, an overview on important features, which should be integrated into a correlation visualization concept, is provided.

In order to approach some of the proposed requirements, a glyph-based correlation visualization approach is presented. Therefore, a tensor-based distance dependent correlation model is suggested for modeling local anisotropic correlation strengths. This model was used for generating glyphs, specially adapted to iso-surface structures and indicating correlation ratios among tangential and between tangential and normal surface directions. A color mapping scheme was introduced, allowing an interactive perception of absolute correlation strengths in various spatial directions. In a synthetic and one real-world data set, the effectiveness of the approach for identifying regions with high and low structural uncertainty is revealed.

5.2 Global Correlation Visualization

After the introduction of novel approaches for visualizing local anisotropic correlation structures, the following sections address the challenge of visualizing global correlation structures, indicating stochastic dependences between random variables at distant spatial points in multi-dimensional data domains.

In a discrete spatial scalar field, the uncertainty can be modeled by a *multi-variate random variable* $\mathbf{Y}$ with scalar-valued components $Y(\mathbf{x}_i)$, where each component models the uncertainty at the respective domain point $\mathbf{x}_i$. In the following, it is assumed that the random variables exhibit a *multi-variate Gaussian distribution*, so that the uncertainty at a point $\mathbf{x}_i$ is indicated by a standard deviation σ_i. The standard deviation is often visualized directly, for instance, via confidence regions, uncertainty glyphs, or specific color or opacity mappings.

Besides analyzing the possible local variations of a quantity via the standard deviation, it is also interesting to investigate the possible variations at different points *relative* to each other. This analysis allows inferring on the possible occurrences of structures, which are determined by the data values at two or more points. The variation of a structure's shape is not only affected by the data values, but also by the degree of stochastic dependence between these values. The property of a structure to vary in shape due to uncertainty is denoted by *structural variability*.

For instance, an *uncertain* 2D height field is considered, where the variability of height values is modeled via Gaussian distributed random variables. For two such variables X and Y the mutual *stochastic dependence* is given by their *correlation* $\rho(X,Y)$, which ranges from -1 to 1 and characterizes the linear relation between the two variables. It is computed as $\mathbb{C}\mathrm{ov}(X,Y)/\sqrt{\mathbb{V}\mathrm{ar}(X)\,\mathbb{V}\mathrm{ar}(Y)}$, where $\mathbb{V}\mathrm{ar}(X)$ and $\mathbb{V}\mathrm{ar}(Y)$ denote the *variances* of X and Y, while $\mathbb{C}\mathrm{ov}(X,Y)$ is the *covariance* between X and Y.

In this example, if the random variables have significantly different standard deviations, the shape of the height field is very likely to change from one realization to another. If the variables have a constant standard deviation, however, it cannot be concluded directly on the probability of shape

variations. In this case, if the random variables have a high positive correlation, i.e., the height values are very likely to either go all up or all down simultaneously, there is low probability that the shape of the height field is strongly affected by the uncertainty. This means that the structural variability is low, even though the entire height field might shift up or down. Contrarily, in regions exhibiting very low or even inverse correlation, the height values can change arbitrarily with respect to each other. In this case, uncertainty can have a strong effect on the height field's shape, causing a high structural variability. Correlation is thus a very important means to analyze the structural variability in uncertain data fields, and this property of correlation forms the basis of the following investigations.

In the following sections, a new approach for visualizing positive and inverse correlation structures in uncertain 2D scalar fields is introduced, where the uncertainty is modeled via multi-variate Gaussian distributions. The approach allows for a local *and* global analysis of the structural variability of a 2D scalar field. By visualizing the uncertain scalar field as a height field, the correlation information can be mapped to the color of surface points and simultaneously integrate common visualizations of the standard deviation on 3D structures.

Since the amount of memory that is required for storing all correlation values is quadratic in the number of spatial sample points, a novel approach is proposed for filtering the correlation information. It seeks for the most prominent spatial correlation structures and represents them as individual clusters. Therefore, so-called *correlation neighborhoods* are introduced and their size is utilized as a new measure for the degree of dependence of a random variable on its local and global surroundings. Correlation neighborhoods are build via spatial clustering of random variables based on mutual correlation strengths. To simultaneously visualize local and global correlations, a subdivision scheme is proposed for breaking clusters indicating long-range dependences into clusters showing ever shorter, yet stronger interactions between their member variables. Since the visualization works solely on the generated clusters, the memory requirement of the novel visualization approach is linear in the number of spatial domain points.

The proposed correlation clusters are associated with random variables at certain spatial positions and their spatial surroundings. Thus, they can be embedded directly into visualizations of the spatial data itself. Fig. 5.11 (a) shows color-coded mean values of an uncertain 2D scalar field, representing a temperature ensemble of data sets that were simulated by the European Center for Medium-Range Weather Forecast (ECMWF). The mean data values are rendered as height surface over the 2D domain in (d). The standard deviations of the ensemble data set are visualized in (b). Although standard deviation values are direct indicators for the underlying uncertainty, they cannot reveal global stochastic dependences described by positive (blue) and inverse (magenta) correlations as illustrated in (c). In (e), color-coded correlation clusters are visualized on the same mean surface. A visualization showing the subdivision of clusters into sub-clusters of ever higher correlation strength and the extrusion of clusters according to standard deviation is shown in (f).

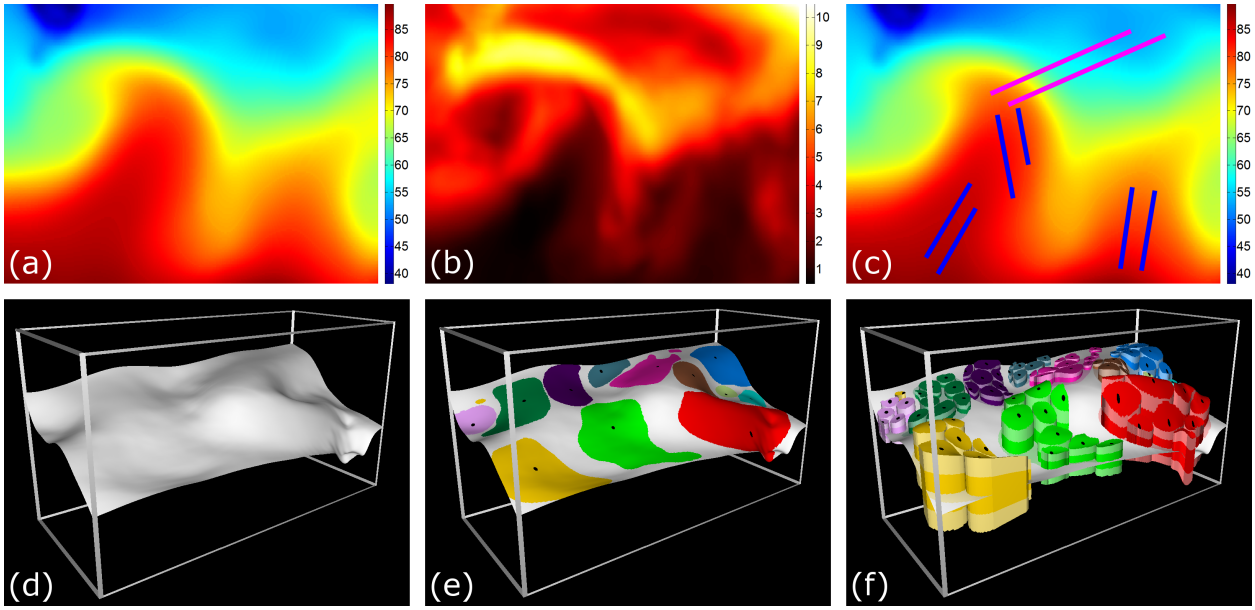

Figure 5.11: (a) Mean height values are color-coded for a temperature iso-surface in an uncertain 2D temperature ensemble data set. (b) Standard deviations are color-coded. (c) Global positive (blue links) and inverse (magenta links) stochastic dependences between different spatial regions are illustrated. (d) Height surface is shown for mean values in (a). (e) Disjoint clusters contain surface points where the uncertainty has a correlation higher than $\rho_1 = 0.4$ to the uncertainty at the cluster centroids (black dots). (f) Clusters are subdivided using $\rho_2 = 0.9$ and extruded along the third dimension according to the standard deviation at the member points.

5.2.1 Related Work

Only very few approaches have explicitly addressed the visualization of data correlations. For instance, a tool for visualizing correlations between two scalar fields via color mapping and slicing was proposed in [JPR*04]. For a similar purpose, [STS06] employed correlation fields and multi-field graphs. Glyph-based visualization of local covariance structures was presented in [KWL*04]. In [SWMW09], correlations in time-varying data have been investigated, and [YXK13] suggests a numerical technique for visualizing (cross-)covariance fields of stochastic 2D simulation results. A sampling scheme for analyzing temporal correlations in 3D time-varying volume data was presented in [CWMW11].

Especially in machine learning applications, correlation clustering as introduced in [BBC04] has been employed to group objects for which pair-wise probabilities about their memberships to common categories are given. Correlation clustering operates on weighted graphs and tries finding a partition of nodes such that the weights of cut positive edges and uncut negative edges is minimized. Since the problem is NP-complete, approximation algorithms using random and local pivoting for selecting cluster centroids have been proposed in [BBC04, ACN08, Zim08, BKKZ04, KKZ09].

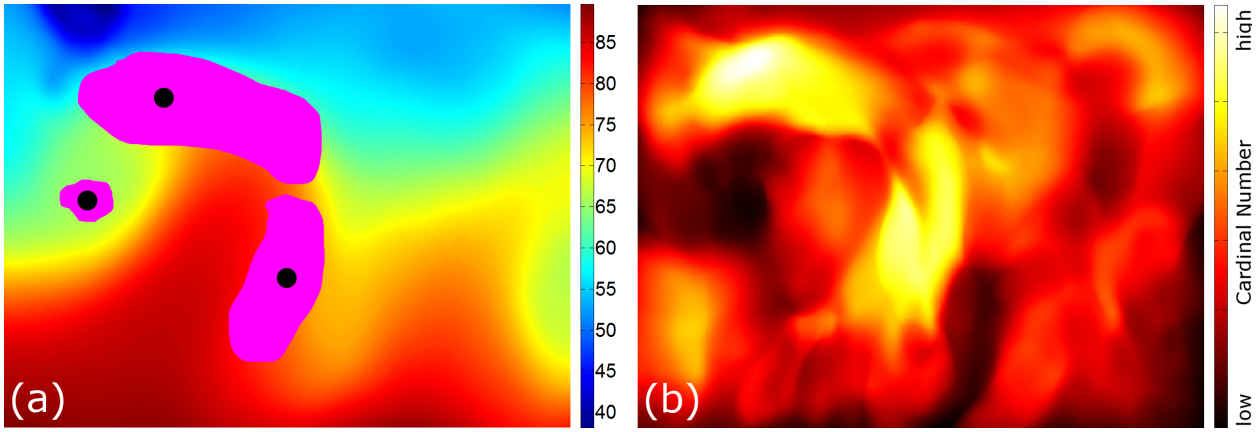

Figure 5.12: (a) Positive correlation neighborhoods (magenta) are shown for three specified cluster centers (black dots) over the ECMWF ensemble mean values. (b) Cardinal numbers of correlation neighborhoods are color-coded for all domain points.

5.2.2 Positive Correlation Clustering

In the following, an uncertain 2D scalar field is assumed. The scalar values are given at the vertices of a 2D grid structure $\mathbb{C}$. At every vertex $\mathbf{x}_i \in \mathbb{C}$, the mean μ_i and the standard deviation σ_i of $Y(\mathbf{x}_i)$ are known, as are the correlation values for every pair $(Y(\mathbf{x}_i), Y(\mathbf{x}_j))$.

To avoid storing the correlation values for every vertex pair during visualization, the correlation information is filtered in a pre-process so as to keep the most relevant correlation structures, but significantly reduce the memory requirements. Therefore, a correlation-based importance measure is introduced, which forms the basis of a novel clustering algorithm.

Correlation Strength Model

For each vertex $\mathbf{x}_i$, the number of vertices $\mathbf{x}_j$, at which the random variables $Y(\mathbf{x}_j)$ have a higher correlation to $Y(\mathbf{x}_i)$ than a pre-defined threshold ρ_1, is computed. The set

$$\eta_{\rho_1}(\mathbf{x}_i) := \{\mathbf{x}_j \in \mathbb{C} \mid \rho(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) \geq \rho_1\} \tag{5.16}$$

is called the *correlation neighborhood* of $\mathbf{x}_i$ for level $\rho_1 \geq 0$, and $|\eta_{\rho_1}(\mathbf{x}_i)|$ the *cardinal number* of this neighborhood. For a given level and vertex $\mathbf{x}_i$, the cardinal number indicates the degree of dependence between the random variable $Y(\mathbf{x}_i)$ and its local and global spatial surroundings, as it counts the most prominent "correlation partners" of $\mathbf{x}_i$, independently of their position in $\mathbb{C}$.

In Fig. 5.12 (a), correlation neighborhoods (magenta) for three different vertices are shown for the temperature ensemble data set introduced in Fig. 5.11. The size of the respective neighborhoods is directly described by the cardinal numbers shown for all domain points in (b). Note that,

for instance, in the region with the smallest neighborhood, the cardinal numbers are very low, indicating also a very low local correlation strength.

The particular choice of the proposed measure is motivated by the assumption of a *distance dependent correlation model* [Tar05]. It is quite common in many applications to assume correlations to be higher/lower between random variables at points with smaller/larger Euclidean distance. For instance, the *Gaussian Correlation Function* (GCF)

$$\rho(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = \exp(-\tau \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) \tag{5.17}$$

models this kind of distance dependent correlation decrease.

The GCF controls the correlation strength by the parameter τ, and, by assigning to each vertex $\mathbf{x}_i$ a specific $\tau(\mathbf{x}_i)$, a particular correlation strength between each $Y(\mathbf{x}_i)$ and its surroundings can be modeled. However, the distance dependent correlation model assumes an isotropic correlation decrease around each vertex and cannot easily be used to account for anisotropic correlation structures. Therefore, instead of assuming a distance dependent correlation decrease and estimating for every vertex a specific τ as proposed in section 5.1, the size of the correlation neighborhood for a given threshold, i.e., the cardinal number, is utilized to measure the local correlation strength of each random variable.

Clustering Algorithm

The cardinal numbers impose an ordering on the vertices that is employed to filter the correlation information. Therefore, the vertices are first ranked in descending order of their cardinal numbers. From this ordered sequence S_{ρ_1}, the algorithm selects the vertex with the largest cardinal number. If and only if the intersection between the correlation neighborhood of this vertex and the correlation neighborhood of any previously selected vertex is empty, the vertex is inserted into a new sequence Ψ_{ρ_1} and removed from S_{ρ_1}. Simultaneously, all vertices belonging to the correlation neighborhood of this vertex are removed from S_{ρ_1}. The algorithm is then applied recursively to the remaining vertices in S_{ρ_1}. This process generates the sequence

$$\Psi_{\rho_1} := \{\mathbf{c}_0, \mathbf{c}_1, ...\} \quad \subset \quad \mathbb{C}, \tag{5.18}$$

which consists of the selected vertices $\mathbf{c}_i$ in descending order of theri cardinal numbers. These vertices are the *centroids* of the correlation clusters that contain all vertices in the corresponding correlation neighborhoods.

The clustering algorithm computes $|\Psi_{\rho_1}|$ clusters, each cluster containing vertices with a correlation to the centroid that is larger than the selected correlation level. The requirement of not allowing intersecting clusters guarantees that every vertex either belongs to exactly one cluster, or does not

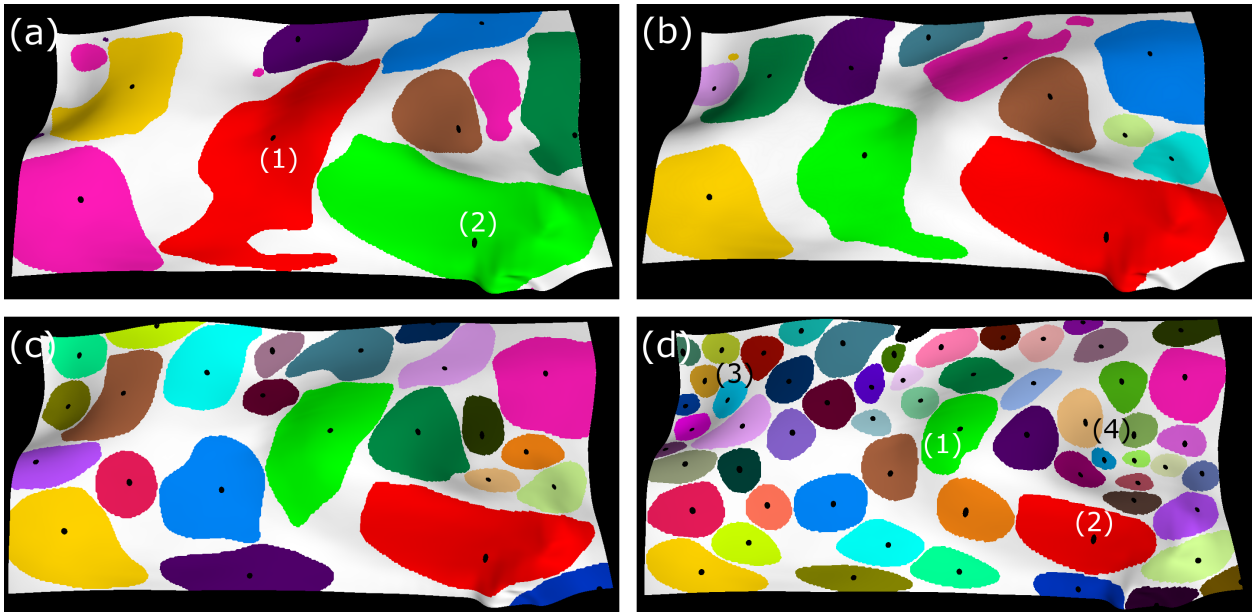

Figure 5.13: Positive correlation clusters for $\rho_1 = 0.3$ (a), $\rho_1 = 0.4$ (b), $\rho_1 = 0.5$ (c) and $\rho_1 = 0.7$ (d) are visualized on the ECMWF mean surface. The clustering for $\rho_1 = 0.7$ (d) indicates strong local correlation in regions (1) and (2), and weak local stochastic dependence in regions (3) and (4).

belong to any cluster. Each cluster gets assigned a unique color using the algorithm proposed in [Hol11] for generating the N perceptually most distinguishable colors. The coloring ensures that also disconnected clusters, indicating so-called *bridging correlations*, can be identified. Fig. 5.13 (a) shows the clusters for $\rho_1 = 0.3$ in the ECMWF data set. The pink clusters show long-range bridging correlations.

The strategy to select the centroids in descending order of cardinality has the preferable property that the largest clusters are always selected first. Correlation clustering algorithms using random or local pivoting strategies for centroid selection, such as the randomized 3-approximation algorithm proposed in [ACN08], cannot achieve this. In addition, a region with strong mutual correlations between the contained points is represented by one single cluster using the presented novel approach, while a randomized algorithm might split up this cluster into multiple ones.

Since the proposed selection strategy represents regions of high and low local correlation by large and small clusters, respectively, it also allows a clear distinction between these regions in the visualization. Furthermore, in the novel approach, the size of a cluster and its expansion in different directions is directly related to the local correlation strength and the correlation distribution in the respective region. A random or local selection of cluster centroids and a complete partitioning of the domain cannot guarantee this and lets the cluster sizes be dependent on the selection order rather than the correlation strength.

Multilevel Clustering

To enable the user to interactively analyze clusters at different correlation levels, multiple sets of clusters are pre-computed for different values of ρ_1. With increasing ρ_1, in regions with low correlation strengths the clusters quickly shrink and the number of clusters increases. Where clusters remain spatially extended, they indicate strongly correlated regions. Fig. 5.13 (b), (c) and (d) show these effects for the initial clusters in (a) and $\rho_1 = 0.4$, $\rho_1 = 0.5$ and $\rho_1 = 0.7$.

For large values of ρ_1, the clusters provide a good impression of the local correlations in the data. For smaller correlation levels, the clusters tend to cover ever larger regions. Although this provides a better focus on global correlation structures and large-range interactions, local correlation structures, as well as the distribution of the correlation structure within the clusters, increasingly disappear.

By providing the user with the possibility to interactively increase and decrease the correlation level, and to examine the evolution of the clusters over multiple levels, these internal structures of spatially extended clusters become apparent.

Cluster Subdivision

To allow the visualization of local correlations within a global context, a subdivision scheme is introduced, splitting the initial clusters at a certain level into disjoint sub-clusters. This is performed by applying the proposed clustering algorithm separately to every initial cluster. For a cluster with centroid $\mathbf{c}_i$, this generates sub-clusters with centroids

$$\Psi^{\mathbf{c}_i}_{\rho_1\rho_2} := \{\mathbf{c}_{i0}, \mathbf{c}_{i1}, ...\} \quad \subset \quad \eta_{\rho_1}(\mathbf{c}_i). \tag{5.19}$$

All sets of sub-clusters are created for a correlation level $\rho_2 > \rho_1$ according to:

$$\eta_{\rho_1\rho_2}(\mathbf{c}_{ij}) := \{\mathbf{x}_k \in \eta_{\rho_1}(\mathbf{c}_i) \mid \rho(Y(\mathbf{c}_{ij}), Y(\mathbf{x}_k)) \geq \rho_2\}. \tag{5.20}$$

By increasing the level ρ_2 for a fixed level ρ_1, the initial clusters are subdivided into ever smaller sub-clusters. Here, the requirement $\rho_2 > \rho_1$ has to be met, because no subdivision will be performed otherwise.

The proposed two-stage approach allows for a simultaneous view on both global correlation structures (selected by ρ_1) and local correlation distributions within these structures (controlled by ρ_2). The sub-clusters indicate in which regions of the initial clusters the random variables are more or less stochastically independent. This effect is shown in Fig. 5.14, were positive correlation clusters for $\rho_1 = 0.4$ in (a) were subdivided using $\rho_2 = 0.9$ in (b). Compared to region (1), the sub-clusters

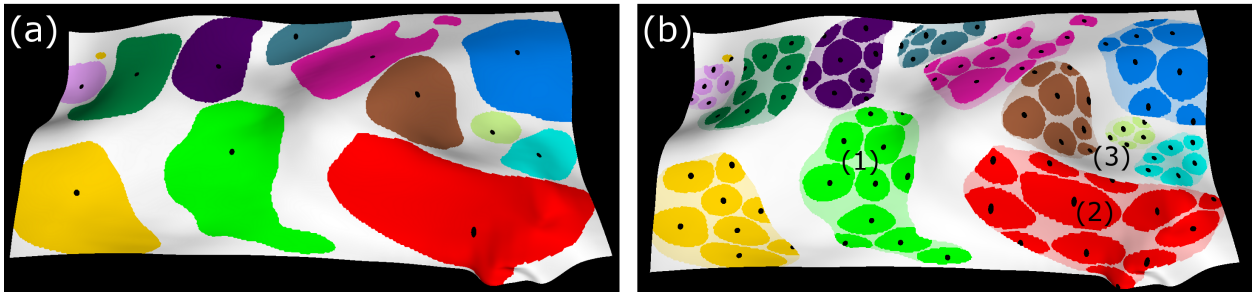

Figure 5.14: Clusters for $\rho_1 = 0.4$ (a) and corresponding sub-clusters for $\rho_2 = 0.9$ (b) show isotropic and anisotropic correlation structures in (1) and (2), respectively. Severe cluster shrinkage indicates low local correlation strength in (3).

in region (2) indicate strong local anisotropic correlation structures with clear preferential directions. In region (3), the severe shrinkage of the clusters indicates a sudden drop in local correlation strength.

Anisotropy Coloring

The proposed clustering algorithm can be used effectively to show the sets of points at which the random variables have a correlation to the random variable at the respective centroid that is larger than a selected correlation level. However, due to the uniform coloring of all points belonging to the same cluster, it can not be seen which sub-regions of a cluster are correlated stronger or weaker to the centroid.

This problem can be approached by mapping the correlation values at every correlation level to color via a specific color table, e.g., as shown in Fig. 5.15 (a). Especially for large clusters the coloring clearly indicates high correlation values around the center points and a distance dependent correlation decrease. For example, the large cluster in the center has significantly lower correlations to points in sub-region (2) than to points in sub-region (1).

However, the used coloring does not pronounce the specific correlation anisotropy within a cluster very well. To overcome this problem, a color mapping scheme is introduced that emphasizes inner-cluster anisotropy.

Note that at every vertex $\mathbf{x}_j$, the correlation to the respective cluster centroid, as well as the Euclidean distance to the centroid are known. This information is used to classify a vertex regarding its correlation decrease from the centroid per unit distance. To this purpose, the GCF in Equ. (5.17) is employed, which controls the correlation strength by the parameter τ. The known correlation and Euclidean distance are used in Equ. (5.17) to rearrange for the unknown τ:

$$\tau(\mathbf{x}_j) := -\frac{\log(\rho(Y(\mathbf{c}_i), Y(\mathbf{x}_j)))}{\|\mathbf{c}_i - \mathbf{x}_j\|^2}, \quad \mathbf{x}_j \in \eta_{\rho_1}(\mathbf{c}_i). \tag{5.21}$$

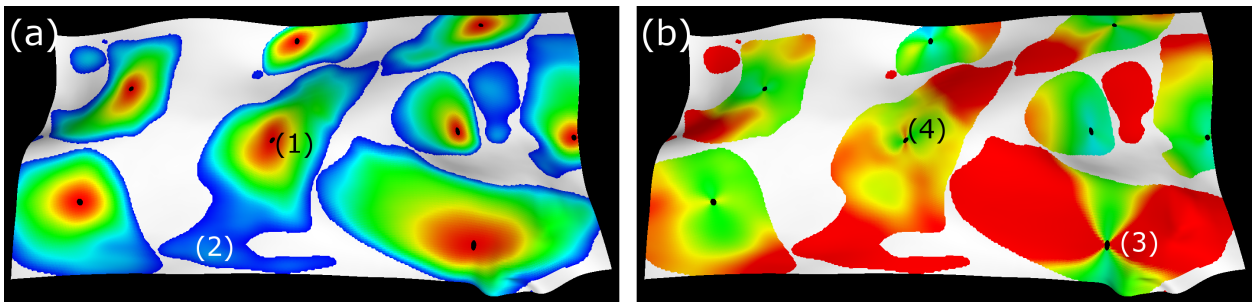

Figure 5.15: (a) Correlation between cluster points and centers is mapped from $[\rho_1, 1]$ to the rainbow color map [blue → red]. (b) CHVD is mapped from $[0, \delta_{\max}]$ to the rainbow color map [blue → red]. High correlation anisotropy around the center point is shown in (3). In (4), high anisotropy is only present close to the center point, while correlation is more isotropic with increasing distance.

From the value of τ it can be concluded on a) a more isotropic correlation structure around a cluster center, i.e., points on concentric circles around the center have the same value, b) an anisotropic correlation structure, i.e., the values along a certain direction have a significantly stronger or weaker decrease, or c) bridging correlations, i.e., isolated groups of points with the same correlation, but at different distance from the center.

Unfortunately, τ has no unit, which prohibits an intuitive interpretation and comparison of different grid points. To alleviate this problem, the *correlation half-value distance* (CHVD)

$$\delta(\tau) := \frac{\log(2)}{\tau}, \tag{5.22}$$

is introduced, indicating the distance after which the correlation drops below 0.5. The CHVD is computed for every point in a cluster (except the centroid), and is then mapped linearly and clamped to a selected range $[0, \delta_{\max}]$, and finally mapped to a specific color table.

In Fig. 5.15 (b), the effectiveness of the proposed color mapping for distinguishing between isotropic and anisotropic correlation structures is demonstrated. In region (3), strong anisotropy can be observed for several radii around the center point. In region (4), anisotropic correlation structures are only present in the vicinity of the centroid. For larger radii the correlations are much more isotropic. In contrast to (a), directions along which the correlation is higher or lower can now clearly be perceived.

5.2.3 Inverse Correlation Clustering

So far, the discussion was restricted to the visualization of clusters where the correlations between the random values at the member points and the cluster centroids exceed a selected *positive* value. In the following, the construction and the visualization of inverse correlated clusters are addressed.

Linear inverse correlation between two random variables indicates that the realization of one random variable deviates positively from its mean when the realization of the other variable deviates negatively, and vice versa. In inverse correlation clustering, one tries to find clusters that consists of two *inverse partners*, i.e., the clusters covering the regions that are inverse correlated to each other.

As for positive correlation clustering, one defines

$$\kappa_{\hat{\rho}}(\mathbf{x}_i) := \{\mathbf{x}_j \in \mathbb{C} \mid \rho(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) \leq \hat{\rho}\} \tag{5.23}$$

as the *inverse correlation neighborhood* of a point $\mathbf{x}_i$ for a negative correlation level $\hat{\rho}$. The cardinal number $|\kappa_{\hat{\rho}}(\mathbf{x}_i)|$ serves again as an indicator for the degree of inverse dependence of a random variable $Y(\mathbf{x}_i)$ to a spatial region. This region, however, is not necessarily a direct spatial neighborhood.

In inverse correlation clustering, one seeks for the most prominent pairs of centroids of inverse partners

$$\Phi_{\hat{\rho}} := \{(\mathbf{a}_0, \mathbf{b}_0), (\mathbf{a}_1, \mathbf{b}_1), ...\} \quad \subset \quad \mathbb{C} \times \mathbb{C}, \tag{5.24}$$

ordered by descending cluster size. These pairs are called *inverse centroids*, and are defined recursively as follows:

$$\mathbf{a}_0 := \arg\max_{\mathbf{x}_j \in \mathbb{C}} |\kappa_{\hat{\rho}}(\mathbf{x}_j)|, \tag{5.25}$$

$$\mathbf{b}_0 := \arg\max_{\mathbf{x}_j \in \kappa_{\hat{\rho}}(\mathbf{a}_0)} |\kappa_{\hat{\rho}}(\mathbf{x}_j)|, \tag{5.26}$$

$$\mathbf{a}_i := \arg\max_{\mathbf{x}_j \in \mathbb{C}, \kappa_{\hat{\rho}}(\mathbf{x}_j) \cap \bigcup_{k<i} (\kappa_{\hat{\rho}}(\mathbf{a}_k) \cup \kappa_{\hat{\rho}}(\mathbf{b}_k)) = \emptyset} |\kappa_{\hat{\rho}}(\mathbf{x}_j)|\,, \tag{5.27}$$

$$\mathbf{b}_i := \arg\max_{\mathbf{x}_j \in \kappa_{\hat{\rho}}(\mathbf{a}_i), \kappa_{\hat{\rho}}(\mathbf{x}_j) \cap \bigcup_{k<i} (\kappa_{\hat{\rho}}(\mathbf{a}_k) \cup \kappa_{\hat{\rho}}(\mathbf{b}_k)) = \emptyset} |\kappa_{\hat{\rho}}(\mathbf{x}_j)|\,. \tag{5.28}$$

The clustering algorithm works much the same way as the one used to construct positive correlation clusters, but now it is necessary to look at two centroids simultaneously in order to create one single cluster $\kappa_{\hat{\rho}}(\mathbf{a}_i) \cup \kappa_{\hat{\rho}}(\mathbf{b}_i)$. Since it always holds that $\mathbf{b}_i \in \kappa_{\hat{\rho}}(\mathbf{a}_i)$ and $\mathbf{a}_i \in \kappa_{\hat{\rho}}(\mathbf{b}_i)$, every inverse partner contains exactly one centroid that is inverse correlated to the other partner and vice versa. Clusters are computed for different negative levels $\hat{\rho}_i$.

In inverse correlation clustering, the coloring of clusters is important to separate pairs of inverse partners from each other. This is accomplished by assigning to each inverse partner $\kappa_{\hat{\rho}}(\mathbf{a}_i)$ and $\kappa_{\hat{\rho}}(\mathbf{b}_i)$ of a cluster a unique distinct color.

It can happen, however, that the partners $\kappa_{\hat{\rho}}(\mathbf{a}_i)$ and $\kappa_{\hat{\rho}}(\mathbf{b}_i)$ each split up into multiple disconnected sub-regions (cf. region (1) in Fig. 5.16 (a)). In this case, the visualization has to indicate that the

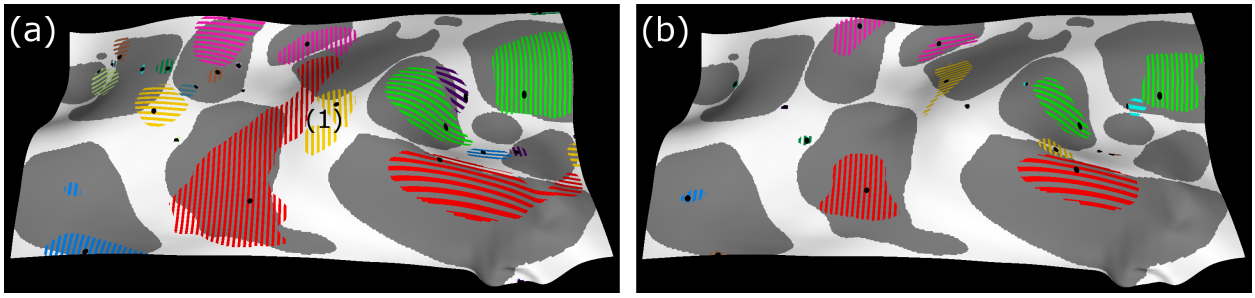

Figure 5.16: Positive clusters are shown in gray, inverse clusters for $\hat{\rho} = -0.3$ (a) and $\hat{\rho} = -0.5$ (b) are color-coded. Clusters with the same color but different stripe orientation are inverse correlated.

sub-regions belonging to the same partner are inverse correlated to the respective other partner, but not to each other. This is achieved by hatching the clusters using different patterns. In particular, vertical and horizontal stripes are used for hatching, uniquely colored to emphasize the cluster. The sub-regions are hatched in the same style than the partner they belong to.

In Fig. 5.16, inverse clusters (in color) are visualized for $\hat{\rho} = -0.3$ (a) and $\hat{\rho} = -0.5$ (b) in relation to the positive clusters (in gray) for $\rho_1 = 0.4$. Each color represents one inverse correlation pair, and the stripe orientation indicates the respective inverse partners within each pair. Note that the size of the inverse clusters are shrinking with decreasing negative correlation level $\hat{\rho}$ and only the most prominent inverse pairs remain present.

5.2.4 Uncertainty Integration

The use of correlation as an indicator for the structural variability in uncertain data sets is only meaningful in regions where a significant standard deviation is present. For instance, if the random variables at two points show a strong inverse correlation, but their standard deviation is low, the effect of the structural variability is also low. Contrarily, a strong effect is very likely if the standard deviation at the two points is high. Consequently, a combined visualization of both the standard deviation and the correlation structures is necessary.

Since in this work uncertain scalar fields over a 2D domain are modeled stochastically via multivariate Gaussian distributions, *stochastic distance functions* (SDF) can be used to visualize the standard deviation. A SDF is defined as (cf. Equ. (4.7))

$$\vartheta(v, \mathbf{x}_i) := \frac{v - \mu(\mathbf{x}_i)}{\max(\sigma(\mathbf{x}_i), \sigma_{\min})}. \tag{5.29}$$

It assigns to every grid vertex $\mathbf{x}_i$ the distance in stochastic data space between the selected scalar value v and the mean value $\mu(\mathbf{x}_i)$ in number of standard deviations $\sigma(\mathbf{x}_i)$. For further details, the

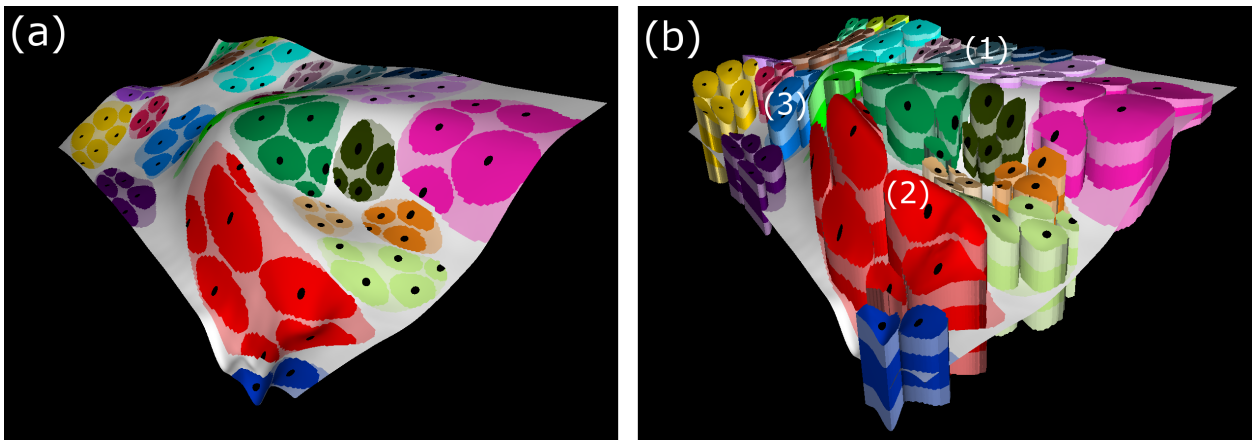

Figure 5.17: (a) Positive, subdivided correlation clusters are shown for $\rho_1 = 0.5$ and $\rho_2 = 0.9$. (b) Cluster extrusion until a selected SDF level reveals low positional variability in (1) and strong positional variabilities in (2) and (3).

reader is referred to paragraph 4.1.2. The region enclosed by the SDF values $|\vartheta(v, \mathbf{x}_i)| \leq 1$ forms the confidence region for $\pm\sigma$.

To visualize the standard deviation, the user first selects a SDF level ϑ^*. Then, all clusters are extruded along the third dimension, both in positive and negative direction, until their height is equal to the selected SDF level at the cluster centroid. The side walls and caps of these "towers" have the same color as the clusters, but, along the side walls, the saturation is decreased by a factor of 0.5 for every second integral change in SDF value.

The integration of standard deviation into correlation visualization is demonstrated in Fig. 5.17. In (a), positive correlation clusters for $\rho_1 = 0.5$, including subdivision for $\rho_2 = 0.9$, are shown. These cluster are extruded to the confidence level $\vartheta(v, \mathbf{x}_i) = 1$ in (b). The small towers in region (1) indicate significantly lower standard deviations compared to regions (2) and (3). Fig. 5.17 shows extruded inverse clusters for $\hat{\rho} = -0.5$.

5.2.5 Implementation and Visualization

The clustering algorithm is performed using a parallelized MATLAB implementation on a shared memory system with two quad-core Opteron 2.6 GHz CPUs. The algorithm has a run-time complexity that is quadratic in the number of grid vertices. For the ECMWF data set that is given on a Cartesian grid of size 185×425, the cluster generation for all levels takes about 19h. The current implementation is relatively unoptimized, as the correlations for many pairs of random variables are re-computed multiple times. This is due to memory limitations, prohibiting the pre-storage of the full correlation matrix. With all correlation data available beforehand, the pre-process would require about 3h.

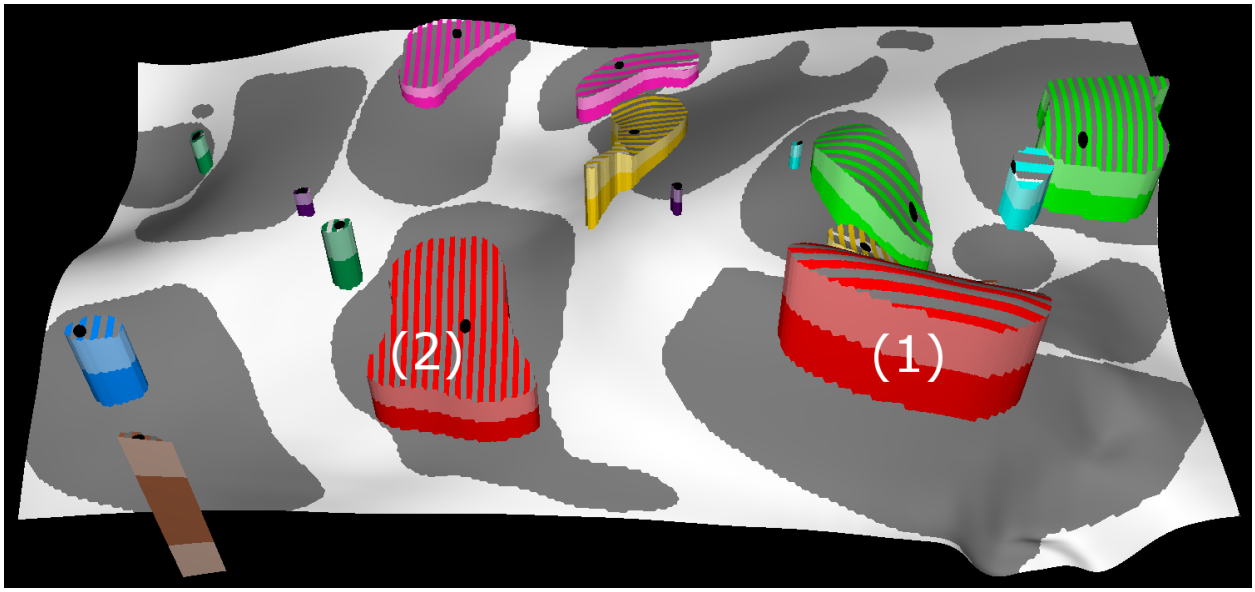

Figure 5.18: Inverse cluster partners for $\hat{\rho} = -0.5$ are shown. Cluster extrusion until a selected SDF level reveals strong inverse positional surface variability between region (1) and (2).

After the pre-process is completed, the resulting data is stored in 3D textures on the GPU. Two textures store the data required to represent the positive correlation clusters. Their (u, v) texture size is the same as the size of the 2D domain over which the height field is given. Thus, every row along the third texture dimension represents the values associated with a particular domain point. The w texture size is equal to the number of possible (ρ_1, ρ_2) pairs, and every texel stores the cluster IDs for exactly one pair. Since ρ_1 and ρ_2 are both $\in \{0.1, 0.2, ..., 0.9\}$, and because $\rho_2 > \rho_1$, 36 pairs have to be stored. Every (ρ_1, ρ_2) tuple gets assigned a unique index at which it is stored in the respective texture row. For every index and every 2D vertex, the IDs of the initial clusters (computed for ρ_1) and corresponding sub-clusters (computed for ρ_2) are stored in the first and second 3D texture, respectively. The IDs are stored in the red color channel, and the correlation and CHVD values to the respective cluster centroids are stored in the green and blue color channels, respectively.

A third texture stores the data computed by inverse correlation clustering. As these clusters are not subdivided, for m negative correlation levels, the texture size along the third dimension is m. For each grid vertex and level a cluster ID is stored. The cluster ID is attributed by a sign which indicates which of the respective two inverse partners the grid vertex belongs to. In an additional texture, the positions of the centroids of all clusters are stored. The cluster IDs that are stored in the 3D textures are chosen such that they can be used directly to reference the respective centroids in this texture. Overall, the memory requirement of the algorithm at run-time is linear in the number of grid points.

For the visualization of the standard deviation, a fourth 3D texture stores the SDF field with respect to the mean and standard deviation in the data. The (u, v) texture size is equal to the size of the 2D domain. The size of the texture in the third dimension was set heuristically depending on the data resolution. Each texture slice along the third dimension is associated to a height value h, ranging from $\min_i(\mu_i) - \max_i(\sigma_i)$ to $\max_i(\mu_i) + \max_i(\sigma_i)$ in m steps. For each grid point $\mathbf{x}_i$ in the (u, v) texture domain and each height layer h_j, a SDF value $\vartheta(h_j, \mathbf{x}_i)$ is stored in the red color channel.

Rendering the correlation clusters in a 2D height field is performed via parallel ray-casting on the GPU. Rays are cast through the 3D SDF field until an intersection with the level-0 iso-surface in this field, i.e., the mean surface, is determined. The projection of the intersection point into the 2D domain and the selected (ρ_1, ρ_2) combination are used to look-up the cluster ID in the pre-computed 3D textures. This ID is then used to color the corresponding pixel. The stripe patterns, indicating the membership to the inverse correlation clusters, are generated procedurally in a pixel shader. For rendering the cluster towers, it is tested at every sampling point along the rays whether the ray has already entered into the selected SDF confidence region. In this case, ray traversal is stopped as soon as the ray enters into a cluster. At the intersection point, the cluster color is looked up and the height dependent saturation is computed. Normals for shading are computed on-the-fly from the SDF field.

5.2.6 Analysis and Discussion

First, the plausibility of the proposed correlation clustering approach is validated using a simple synthetic data set (see Fig. 5.19 (a)). As input for a simulation process, a set of seismic pressure sources are positioned at the grid vertices of a 2D Cartesian grid (white lines) on the earth ground (green plane). Each source generates pressure waves traveling into the earth along the black lines. The waves are reflected from a material discontinuity (red structure at the bottom) and registered at the source locations. The red surface is broken at three fault lines where a discontinuous change between two depth layers is perceivable. By making assumptions on the wave speed and measuring the travel-time, one can estimate the discontinuity's depth at every grid position.

Since the material in-between the source and the discontinuity (second layer) introduces errors in the assumed wave speed, the estimated depth of the discontinuity is uncertain. In the presented model, the "error layer" is subdivided into 34 zones (uniquely colored), each of which introduces a Gaussian distributed zero mean error that are stochastically independent from each other. Within the three quadrants (1), (2), and (3), the error model has the same standard deviation. In quadrant (4), the standard deviation is significantly lower.

Based on the probabilistic error model, 100 possible solutions were computed. Positive correlation clusters in the ensemble field are color coded on the mean surface in Fig. 5.19 (b). It can be seen that

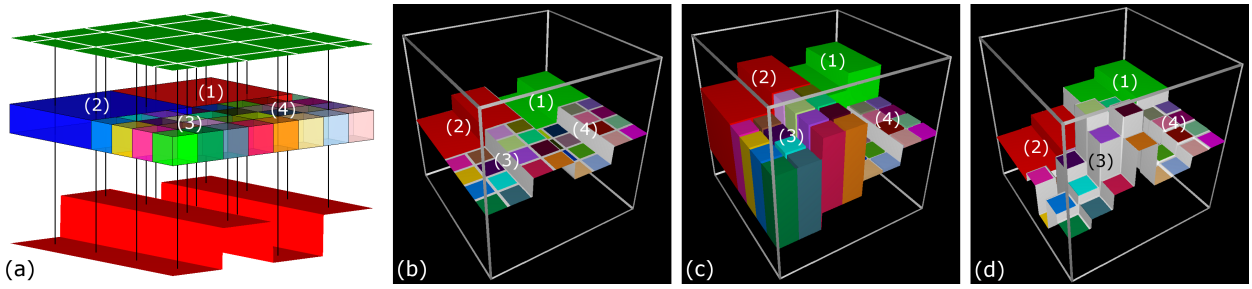

Figure 5.19: (a) Geophysical setup to determine the depth of a material discontinuity (red) in the earth's interior by measuring travel times of artificial pressure waves (emitted along black lines). A material layer between emitters and discontinuity structure simulates a Gaussian error distribution in wave velocities. (b) Correlation clusters are color coded on the mean surface in the simulated data ensemble. (c) Visualization of standard deviation shows equally strong uncertainty in quadrants (1), (2), (3) and low uncertainty in (4). (d) One possible solution (realization) of the depth structure. High structural variability is seen in quadrant (3), which is indicated by low correlations and high standard deviations in (b) and (c).

the independent error structures shown in (a) are correctly grouped by the clustering algorithm. In (c), clusters are extruded to visualize the standard deviation. A significantly lower error is shown in quadrant (4). In (d), one of the possible solutions (realizations) is shown. The faults in (1) and (2) can be well resolved, because correlations are high in both regions, i.e., large correlation clusters exist. The fault between (1) and (2) cannot be resolved, because there is no correlation between (1) and (2). In quadrant (4), the correlation is low, but the fault can be well resolved, because the standard deviation is low, too. In region (3), however, correlation clustering reveals highly uncorrelated sub-regions and high standard deviations. Consequently, the fault cannot be resolved here, because of high structural uncertainty. This example illustrates that an integrated visualization of uncertainty *and* correlation is very important, as the single, detached visualization of each of them could result in false interpretations.

For the ECMWF data set, the proposed correlation visualization is used to conclude on the following peculiarities: Fig. 5.11 (d) and (e) show a long ridge-like surface structure covered by the large red cluster. This shape feature can be assumed to be stable, because it resides in a highly correlated region, indicated by large clusters even for increasing correlation levels (cf. Fig. 5.13). Furthermore, cluster subdivision in Fig. 5.14 and anisotropy coloring in Fig. 5.15 (b) clearly show the alignment of the prominent correlation direction with the ridge orientation, meaning that the structural variability is low along the ridge. In Fig. 5.13 (d), small clusters in region (4) indicate low correlation strength (see also region (3) in Fig. 5.14 (b)). It can be concluded that this area is prone to structural variability, and that the particular occurrence of the mean surface is stochastically unstable.

Inverse correlation visualization reveals very interesting long-range dependences between data val-

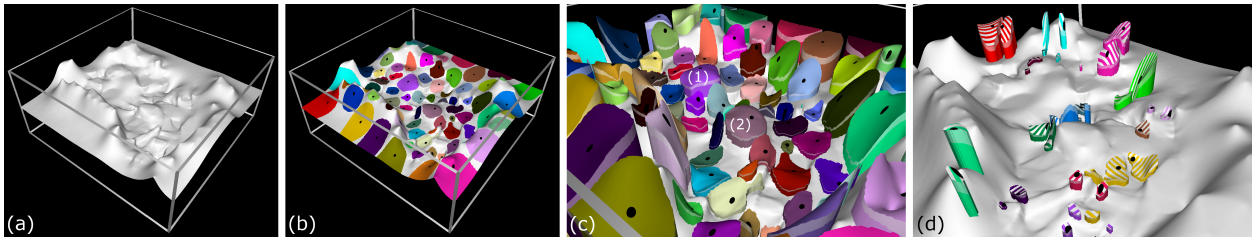

Figure 5.20: (a) Mohorovičić discontinuity below Australia is shown. (b) Strong correlation clusters close to the domain boundaries indicate strong regularizations in the simulation algorithm. (c) Close-up view reveals high and low uncertainties, respectively, at the boundaries and in the center, as well as high local correlations at (1) and (2). (d) Inverse clustering shows that strong inverse correlation takes place on a local rather than a global scale.

ues in the ECMWF data set. Fig. 5.18 shows multiple inverse correlated cluster pairs. In combination with the uncertainty towers, the red clusters turns out to be the most prominent. The visualization shows a strong stochastic dependence between spatial locations over long distances. For instance, a temperature decrease in region (1) is likely to cause a temperature increase in region (2), and vice versa.

In a third example, correlation visualization is used to analyze the Mohorovičić discontinuity — the boundary surface between the Earth's crust and mantle — below Australia (see Fig. 5.20 (a)). The data was acquired using a similar geophysical setup as described in the first example. Positive correlation clusters in (b) show a rather homogeneous correlation distribution in the domain interior and high correlation strengths at the boundaries. A close-up view in (c) also reveals high standard deviation in the outer parts. The reason is that less measurements were performed in these regions and, thus, the data coverage is too low to allow resolving high frequencies in the data. As a consequence, such regions are automatically regularized (smoothed) by the data generation algorithm, resulting in high correlations and standard deviations. In (1) and (2), the same uncertainty and dependence structures are visualized. Correlation visualization supports domain experts in discovering whether smooth structures arise from the real physical material characteristics in the discontinuity or are due to regularization effects.

Besides the visualization of positive correlations, domain experts in geophysics are interested in the location of inverse correlated regions. From this information, one can conclude on regions that cannot be resolved against each other and have a rather uncertain relative position. The visualization in (d) shows that inverse correlated regions are located close to each other. Large-range inverse correlations do not seem to exist. This indicates that strong structural variabilities are restricted to small spatial regions. Note that this is completely different to the situation in the ECMWF data set, where inverse correlated clusters are far more distant to each other and cover significantly larger regions.

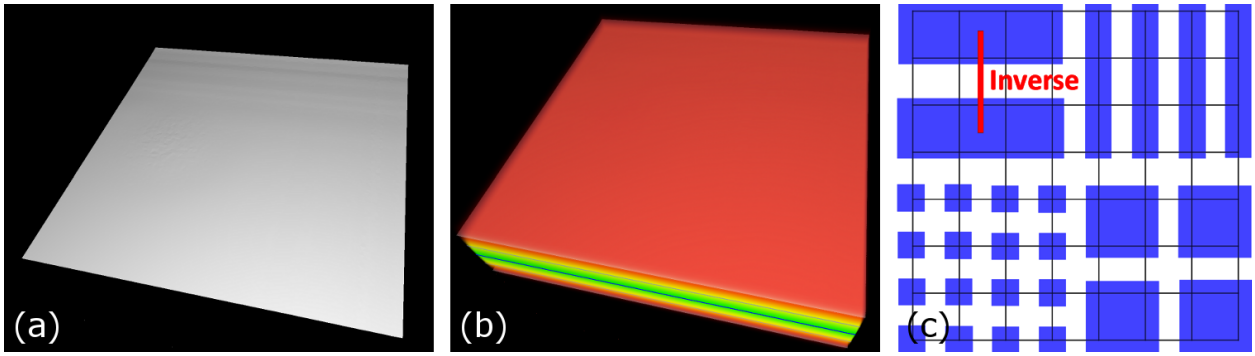

Figure 5.21: The setup for a synthetically generated 2D ensemble data set is illustrated. (a) The height surface for constant mean values over a 2D domain is shown. (b) The confidence volume around the mean surface is visualized for a setting of constant non-zero standard deviation values for all domain points. (c) The synthetic correlation setting for all domain points is illustrated. Grid points within the same blue cell are strong positive correlated. Points between different blue cells are not correlated. The two upper left blue cells (and their member domain points) are inverse correlated with respect to each other.

As final proof of concept, a validation example is presented for demonstrating that the proposed algorithm is able to generate correct results with respect to revealing prior synthetically designed correlation structures. Therefore, first, a synthetic ensemble data set is generated. The data is given on a regular Cartesian 2D grid. At all grid points, a constant value is used as mean of the respective random variables. The standard deviations are also fixed to a certain non-zero value. The height surface of the mean values over the 2D domain is shown in Fig. 5.21 (a). In (b), the respective confidence volume around the mean surface is rendered by mapping SDF values to color and opacity. Furthermore, correlation values are defined between the respective 2D grid points. This setup is illustrated in Fig. 5.21 (c). The 2D grid is shown as the black lines. Correlation values between grid points within a blue cell are strong positive. Grid points between different blue cells have zero correlation. The random variables at grid points in the upper left two blue cells are inverse correlated.

The defined mean, standard deviation and correlation values are used as input for a multi-variate Gaussian random number generator, which is used for generating an ensemble data set of scalar values at the different spatial grid points. Eight ensemble members are show in Fig. 5.22. The scalar values from low to high are mapped to the rainbow color map from blue to red. For obtaining scalar values between the grid points, a bi-linear interpolation scheme has been utilized. The realizations show a high structural uncertainty in the lower left region where the grid points are not correlated with respect to each other. The oblong features in the upper right and left part of the ensemble members directly reflect the respective positive correlation structures shown in Fig. 5.21 (c).

The 2D ensemble data set is used as input for the novel correlation clustering algorithm. The

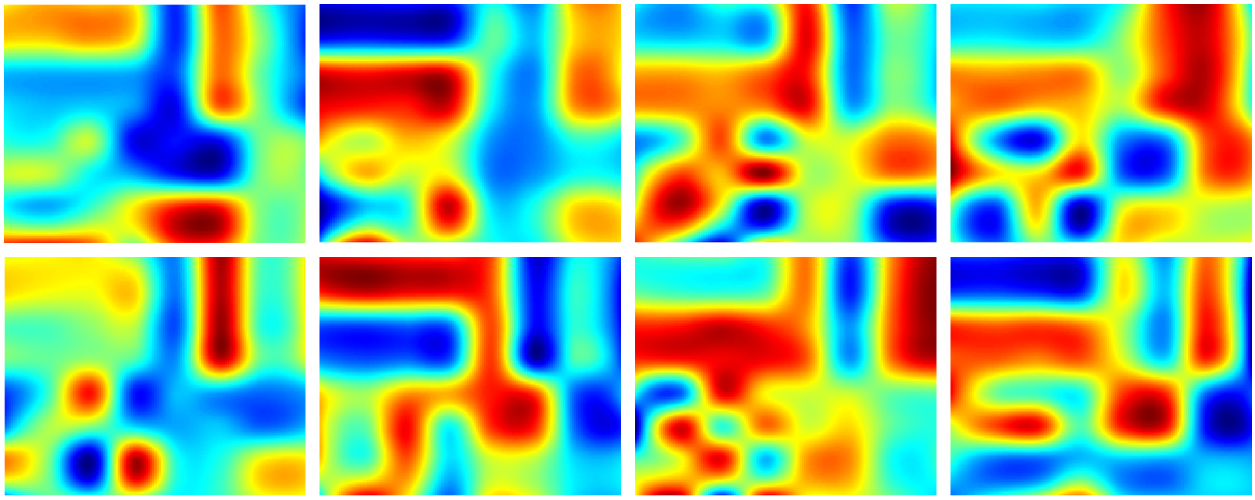

Figure 5.22: Ensemble members (realizations) are shown for the synthetic 2D ensemble data set introduced in Fig. 5.21. Height values are mapped from low to high to the rainbow color map from blue to red.

results are presented in Fig. 5.23. The clustering in (a) directly reveals the synthetically defined correlation structures in Fig. 5.21 (c). The fuzzy borders of the individual clusters are caused by the used interpolation scheme and its impact on the clustering process. The cluster extrusion in (b) illustrate the constant standard deviation values over the 2D domain. The inverse clustering approach in (c) correctly identifies the inverse correlated regions in the upper left part of the domain.

This example reveals that the proposed novel correlation clustering approach is able to correctly visualize correlation structures in a 2D domain. Note that this additional information helps to better predict the degree of structural uncertainty for diverse spatial regions. Structural variability, as illustrated by the ensemble members in Fig. 5.22, cannot be visually communicated by just using standard methods based on traditional uncertainty indicators like mean and standard deviation (cf. confidence visualization in Fig. 5.21 (b)).

5.2.7 Conclusion

The contribution to the field of uncertainty visualization, presented in this section, is a new approach for visualizing positive and inverse global correlation structures in uncertain 2D scalar fields. The novel correlation clustering algorithm is built upon the concept of correlation neighborhoods and their cardinal numbers. The organization of data points into groups takes into account a selected correlation strength, giving rise to an interactive visual analysis of short- and long-range stochastic dependences in the data. The cluster representation which is build in a pre-process requires an amount of memory that is only linear in the number of initial data points.

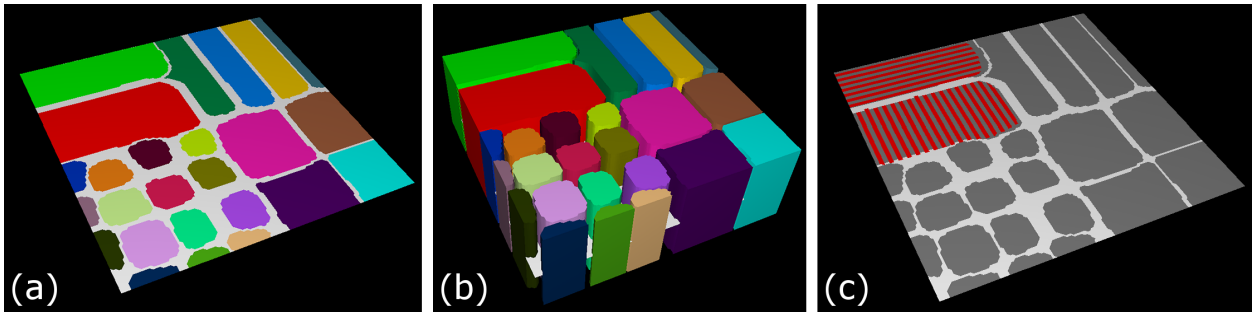

Figure 5.23: The results of the clustering algorithm applied to the synthetic ensemble data set are shown. (a) Algorithm identifies positive correlation clusters at the correct spatial domain regions. (b) Clusters are extruded within the constant confidence volume. (c) Inverse correlated regions are identified at the correct spatial positions.

5.3 Gradient Uncertainty Visualization

Scalar ensemble fields consist of different scalar instances of the outcome of an experiment. They can be produced by repeated simulations of a parameterized model using slightly modified parameter settings, or by simply running a procedure that is affected by inherent randomness multiple times. When the experiment generates outcomes over a spatial domain, each instance captures possible data values at all locations in space. If the variation of instances can be described stochastically, each instance can be assumed a realization of a multi-variate random variable.

In scalar ensemble fields, standard deviations are primary indicators for the degree of variation of the values. Thus, standard deviations are a means to classify the confidence in the data values, and, to this purpose, are often visualized directly, for instance, via confidence regions, uncertainty glyphs, or specific color or opacity mappings [JS03, PWL97].

The standard deviation by itself, however, does not allow a rigorous analysis of uncertain data, because it does not allow inferring on the *relative variability* of data values at different points. This means, in particular, that the effect of uncertainty on quantities that depend on the rate of change of the data, such as *differential quantities*, cannot be analyzed. Such quantities, however, play a major role in data analysis, for instance, to indicate the location and orientation of important geometric features, such as object boundaries or iso-contours.

In the following, the effect of uncertainty on the *variability of gradients* in 2D scalar ensemble fields is analyzed, with respect to both magnitude and orientation. Such an investigation helps answer primary questions on the stability of features in such fields. For instance, if a feature classifier depends on the gradient magnitude and a point has been classified as belonging to the feature, a low variability of the gradient magnitude indicates with a high level of certainty the point's membership to the feature. Then, a low variability in gradient orientation shows that the shape of the iso-contour passing through the respective spatial point is very likely to remain unchanged,

even if the standard deviation shows a high spread in the data values. On the contrary, a high uncertainty in gradient orientation indicates a likely change in the orientation of the contour, even though a low standard deviation might be observed.

Assessing the variability of gradients can therefore reveal the stability of features and their geometric structures in scalar fields. It is, however, considerably more complicated than determining the data variability via the standard deviation, because no stochastic model describing the spread in magnitude and orientation is available initially. Thus, one first has to derive the random variations of both quantities due to the uncertainty in the data, before techniques for analyzing these variations can be developed.

The novel concepts presented in the following section take a first step into this direction: Based on the stochastic modeling of uncertainty via multi-variate random variables, uncertainty parameters, such as the mean and the covariance, are derived for gradients in uncertain scalar fields. Thus, the presented work supplements previous uncertainty visualization techniques by providing a means to analyze the effects of uncertainty on derived quantities that depend on data values at more than one spatial point. It is worth mentioning that no assumption are made about the distributions of the random variables used to model the uncertainty and only the computation of the statistical entities mean and covariance is required.

Building upon the uncertainty parameters, a mathematical framework is developed to analytically derive confidence intervals for the strength of the derivative in any prescribed direction and the gradient orientation. The confidence intervals are utilized because they can be computed solely from the uncertainty parameters, making the introduced approach independent of any assumed initial probability distribution. This is the first time that analytic expressions for confidence intervals of gradients in uncertain scalar fields have been derived. In case that the data uncertainty is modeled via Gaussian distributed multi-variate random functions, even analytical expressions of the distributions of the gradient magnitudes and orientations can be derived.

In the following, novel visualization techniques are presented for qualitatively assessing the gradient variability in uncertain 2D scalar fields. For visualizing the derivative strength uncertainty, a novel approach is introduced using color diffusion in addition to color to simultaneously show the data values and the gradient's variations. The basic idea is to continually diffuse the colors of a base pattern, revealing the variation of data values in the initial field, with preselected colors representing different degrees of uncertainty. By controlling the diffusion strength via the degree of derivative uncertainty, a clear differentiation between geometric features of high and low stability is obtained. For the latter, a special family of circular glyphs is proposed, where a glyph's pattern and color convey the spread in direction and the uncertainty degree, respectively.

In summary, the main contributions of the presented concepts are:

- A derivation of uncertainty parameters such as mean and covariance for gradients in uncertain scalar fields given on discrete grid structures.
- Analytic expressions of the confidence intervals quantifying the variation of the gradients' magnitude and orientation in uncertain scalar fields.
- A visualization technique using color diffusion to indicate the stability of the slope along the gradient direction in 2D scalar fields.
- A family of patterned and colored glyphs to quantitatively depict the uncertainty in the orientation of iso-contours in 2D scalar fields.

To validate the proposed techniques, they are applied to visualize the gradient uncertainty in a number of data sets. A synthetic data set is used to illustrate how the approaches convey additional information that cannot be obtained from the mean and standard deviation values alone. Furthermore, by using several real-world data sets, the strength of the presented approaches to analyze important geometric features with respect to their possible changes due to uncertainty is confirmed.

5.3.1 Related Work

Several visualization techniques have been developed to represent uncertainty in magnitude and orientation of the individual vectors in vector fields generally. Different glyph techniques are presented in [WPL96], where the authors experiment with various arrow glyphs that use the width of the arrow head to indicate uni-modal angular uncertainty and additional arrow heads for the range of possible magnitudes. Rectangular glyphs, together with additional less emphasized lines to encode the uncertainty, are used for bidirectional vector fields in [ZDG*08].

In the tractography domain, [Jon03] introduces a so-called "cone of uncertainty" 3D glyphs to visualize the orientation and corresponding uncertainty of brain fibers. The directional information is approximated by the principal eigenvector of the diffusion tensor, while the associated uncertainty is estimated via bootstrap methods. [SJK04] combines patterns with glyphs to map an anisotropic reaction-diffusion model to vector magnitude and orientation, by producing spot patterns of various shapes, sizes, orientations, and densities, where uncertainty in orientation can be qualitatively incorporated in the amount of anisotropy that gives the shape of the spot.

[BWE05] uses cross-advection and error diffusion in a texture-based flow visualization, where uncertainty in flow direction arising during data acquisition is revealed by changing the spatial frequency orthogonal to the flow direction. Distributions of directions are used in [XLS10] to quantitatively measure the information content in the input flow field and then generate streamlines that appropriately represent the vector field. Uncertainty is modeled here by a random variable having the directions of all vectors in the flow field as components.

A modified LIC approach is presented in [OB09], to convey uncertainty in 2D steady flow fields. Here, the magnitude and direction of vector fields are taken to be described by presumably existent probability density functions. Probabilistic numerical integration in uncertain vector fields is also performed in [OGHT10, OGT11b], where the vector fields are assumed to be Gaussian distributed. The normal distribution supposition, however, does not extend to the vector magnitude and orientation, and no distinction is made between the two quantities. The methods are extended to detect closed orbits in 2D uncertain vector fields in [OGT11a]. Then, in [OT12], the authors introduce an approach for vortex structures in Gaussian distributed vector fields. The method uses derived uncertain vector fields, but the derivations lead to product distributions that are no longer Gaussian and for which there are typically no closed-form solutions.

In [PPH12], the local distribution functions of uncertain vector quantities are computed via Monte Carlo sampling, the stochastic properties being derived from a set of realizations of the uncertain vector data via a computationally expensive process. These methods differ from the presented approach, in that it analyzes the local variations of vectors via an analytical mathematical derivation of the probability density functions and distinguish between vector magnitude and orientation.

5.3.2 Gradient Uncertainty

This section introduces the mathematical foundations necessary to define gradients in uncertain scalar fields, along with their statistical parameters that are required to model the associated uncertainties in magnitude and orientation.

In the following, a discrete sampling of a 2D domain on a Cartesian grid structure is assumed with grid points $\mathbb{S}_{m,n} = \{\mathbf{x}_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n\}$. The data uncertainty is modeled by a *multivariate random variable* $\mathbf{Y}$, i.e., a random vector with scalar-valued components $Y(\mathbf{x}_{i,j})$, based on a linearized order of the elements in $\mathbb{S}_{m,n}$. It is not assumed that the random variables follow any particular distribution, but only that the mean values $\mu(Y(\mathbf{x}_{i,j}))$, the standard deviations $\sigma(Y(\mathbf{x}_{i,j}))$, and the correlation values $\rho(Y(\mathbf{x}_{i,j}), Y(\mathbf{x}_{k,l}))$ between any pair of random variables can be computed. For example, if the underlying data uncertainty is represented by an ensemble of scalar data values at all spatial points in $\mathbb{S}_{m,n}$, the mean values, standard deviations, and correlations can be derived using standard statistical estimators (cf. [Fel08]).

Uncertainty Parameters

In an uncertain scalar field, the variation of the scalar values causes a probability distribution of the gradient. The confidence intervals for the gradient uncertainty are derived by first approximating the gradients from the given random variables via a linear operator, and then using this operator to

approximate the uncertainty parameters, i.e., the means and covariance matrices of the gradients, from the uncertainty parameters of the random variables.

The gradient at a particular point $\mathbf{x}_{i,j}$ can be approximated via central differences (one-sided differences at the domain boundaries) on the random variables as

$$\nabla Y(\mathbf{x}_{i,j}) = \mathbf{A}\mathbf{s}(\mathbf{x}_{i,j}). \tag{5.30}$$

Here, the 4-element stencil $\mathbf{s}$ contains the random variables

$$\mathbf{s}(\mathbf{x}_{i,j}) = [Y(\mathbf{x}_{i+1,j}), Y(\mathbf{x}_{i-1,j}), Y(\mathbf{x}_{i,j+1}), Y(\mathbf{x}_{i,j-1})]^\top, \tag{5.31}$$

and the 2×4 matrix $\mathbf{A}$ contains the inverse point distances

$$\begin{aligned}
&\mathbf{A}_{1,1} = \|\mathbf{x}_{i+1,j} - \mathbf{x}_{i-1,j}\|^{-1}, \mathbf{A}_{1,2} = -\mathbf{A}_{1,1},\\
&\mathbf{A}_{1,3} = \mathbf{A}_{1,4} = \mathbf{A}_{2,1} = \mathbf{A}_{2,2} = 0,\\
&\mathbf{A}_{2,3} = \|\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j-1}\|^{-1}, \mathbf{A}_{2,4} = -\mathbf{A}_{2,3}.
\end{aligned}$$

The stencil $\mathbf{s}(\mathbf{x}_{i,j})$ forms a 4-component subset of the multi-variate random variable $\mathbf{Y}$. In order to quantify the gradient's variation, the mean gradient μ_∇ and the covariance matrix Σ_∇ of ∇Y are computed. From Equ. (5.30), these quantities relate to the mean and covariance values of the random variables via

$$\mu_\nabla(\mathbf{x}_{i,j}) = \mathbf{A}\mu_\mathbf{s}(\mathbf{x}_{i,j}), \tag{5.32}$$

$$\Sigma_\nabla = \mathbf{A}\Sigma_\mathbf{s}\mathbf{A}^\top, \tag{5.33}$$

where the k-th component of $\mu_\mathbf{s}(\mathbf{x}_{i,j})$ contains the mean of the k-th component of $\mathbf{s}(\mathbf{x}_{i,j})$, i.e., $(\mu_\mathbf{s}(\mathbf{x}_{i,j}))_k = \mu(\mathbf{s}(\mathbf{x}_{i,j})_k)$, and the components of the covariance matrix $\Sigma_\mathbf{s}$ of the random stencil vector are $(\Sigma_{\mathbf{s}(\mathbf{x_{i,j}})})_{m,n} = \sigma(\mathbf{s}(\mathbf{x}_{i,j})_m)\sigma(\mathbf{s}(\mathbf{x}_{i,j})_n)\rho(\mathbf{s}(\mathbf{x}_{i,j})_m, \mathbf{s}(\mathbf{x}_{i,j})_n)$. The derivation of Equ. (5.32) and (5.33) follows the standard linear transformation rule for mean vectors and covariance matrices. For further details, the reader is referred to, e.g., [Fel08].

The derivation of the gradient mean vector and the covariance is also possible for *irregular grid structures*. An irregular grid structure $\mathbb{I}_m = \{\mathbf{x}_i : 1 \leq i \leq m\}$ is assumed, where each grid point is connected to a set of neighboring grid points $N(\mathbf{x}_i)$ via edges. The data uncertainty at every point is modeled in exactly the same way as described for the regular case. For a given point $\mathbf{x}_i$, the derivatives along the edges in the random field are approximated by one-sided differences, and they can also be expressed by the projection of the (unknown) gradient at $\mathbf{x}_i$ onto the edges

$$\nabla Y(\mathbf{x}_i)^\top \frac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|} = \frac{Y(\mathbf{x}_j) - Y(\mathbf{x}_i)}{\|\mathbf{x}_j - \mathbf{x}_i\|}, \quad \mathbf{x}_j \in N(\mathbf{x}_i). \tag{5.34}$$

For a vertex with n neighbors this can be written as a linear system

$$\mathbf{R}\nabla Y(\mathbf{x}_i) = \mathbf{b}, \tag{5.35}$$

$$\mathbf{R} = [\mathbf{r}_1|\mathbf{r}_2|...|\mathbf{r}_n]^\top, \quad \mathbf{r}_j = \frac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|}, \tag{5.36}$$

$$b_j = \frac{Y(\mathbf{x}_j) - Y(\mathbf{x}_i)}{\|\mathbf{x}_j - \mathbf{x}_i\|}. \tag{5.37}$$

Because in a d-dimensional grid the gradient has d components, d neighbors are required to solve for the gradient. However, as in a d-dimensional grid, every grid point has typically more than d neighbors, so that the system is over-determined, i.e., it is not guaranteed that a gradient exists that solves the equations for all neighbors. Thus, the system is solved using the least squares approach and the respective normal equation

$$\mathbf{R}^\top\mathbf{R}\nabla Y(\mathbf{x}_i) = \mathbf{R}^\top\mathbf{b}, \tag{5.38}$$

$$\nabla Y(\mathbf{x}_i) = \left(\mathbf{R}^\top\mathbf{R}\right)^{-1}\mathbf{R}^\top\mathbf{b} = \tilde{\mathbf{A}}\mathbf{b}. \tag{5.39}$$

As the random variables at the grid points appear linearly in $\mathbf{b}$, and $\tilde{\mathbf{A}}$ is a linear operator, the mean vector and the covariance matrix for $\nabla Y(\mathbf{x}_i)$ can be obtained in more or less the same way as described in Equ. (5.32) and (5.33).

Since the statistical parameters describing the gradient's variation indicate the spread in both magnitude and orientation, in the following, these two properties are separated, and then, analytic expressions of the confidence intervals for each of them are derived. These intervals are then used to visualize the gradient uncertainty.

Uncertainty in Derivative

Because the derivative at a spatial point $\mathbf{x}_{i,j}$ in a 2D scalar field is dependent on direction, first, a suitable direction $\mathbf{v}(\mathbf{x}_{i,j})$ has to be selected into which to estimate the *uncertainty of the derivative*. Since the following approaches concentrate on the uncertainty analysis of features in the *mean data set*, without loss of generality, the mean gradient direction $\mathbf{v}(\mathbf{x}_{i,j}) = \frac{\mu_\nabla(\mathbf{x}_{i,j})}{\|\mu_\nabla(\mathbf{x}_{i,j})\|}$ is chosen. This choice is particularly beneficial for the cases where the data uncertainty can be modeled via multi-variate distributions for which the derivative is most likely maximum in the mean gradient direction, such as Gaussian distributions. For arbitrary distributions, however, other directions might better serve the purpose of uncertainty analysis. For instance, if a multi-modal distribution of the gradient direction can be assumed, the preferential directions are those corresponding to the observed peaks.

The uncertainty of the derivative in the mean gradient direction can itself be modeled by a scalar random variable, by projecting $\nabla Y(\mathbf{x}_{i,j})$ onto the mean gradient direction

$$D(\mathbf{x}_{i,j}) := \mathbf{v}(\mathbf{x}_{i,j})^\top \nabla Y(\mathbf{x}_{i,j}). \tag{5.40}$$

$D(\mathbf{x}_{i,j})$ describes the random variation of the directional derivative along the mean gradient direction. Because D is obtained by applying a linear operator to ∇Y, the mean and standard deviation of D are given by

$$\mu_D(\mathbf{x}_{i,j}) = \mathbf{v}(\mathbf{x}_{i,j})^\top \mu_\nabla(\mathbf{x}_{i,j}) = \|\mu_\nabla(\mathbf{x}_{i,j})\|, \tag{5.41}$$

$$\sigma_D(\mathbf{x}_{i,j}) = \sqrt{\mathbf{v}(\mathbf{x}_{i,j})^\top \Sigma_\nabla \mathbf{v}(\mathbf{x}_{i,j})}. \tag{5.42}$$

Thus, the mean derivative in the mean gradient direction is the magnitude of the mean gradient. The standard deviation indicates the variability of the derivative and serves as an uncertainty indicator. This uncertainty is subsequently called the *derivative uncertainty*. The confidence interval of the derivative uncertainty is defined as $[\mu_D - \sigma_D, \mu_D + \sigma_D]$.

Uncertainty in Orientation

In the following, confidence intervals for the uncertainty in gradient orientation are derived, i.e., intervals describing the spread of the angular variation around the mean gradient direction. The goal is to project the gradient uncertainty — given by the covariance matrix Σ_∇ — onto the unit circle, in order to obtain an angular standard deviation θ_σ around the angle θ_∇ between the mean gradient direction μ_∇ and the x-axis. This is illustrated in Fig. 5.24, where Σ_∇ is represented by the corresponding covariance ellipse.

To project the gradient uncertainty onto the unit circle, the covariance information is transformed to polar coordinates, and then the gradient magnitude is separated from the gradient direction. However, unlike the projection in Equ. 5.41, this transformation is not linear and, thus, no closed-form formulae can be derived. This problem is approached by employing an auxiliary 2D probability density function parameterized by the given gradient mean μ_∇ and covariance matrix Σ_∇. This function can then be transformed to polar coordinates. In polar coordinates, the gradient magnitude can be separated from the orientation, so that the marginal probability of the angular variations around the mean vector can be computed by integrating over the magnitude, and then used to derive confidence intervals.

A Gaussian distribution is chosen as probability density function, because it can be easily parameterized by μ_∇ and Σ_∇ to yield a 2D random variable with exactly the prescribed stochastic properties. Furthermore, it is integrable and can be transformed to polar coordinates analytically.

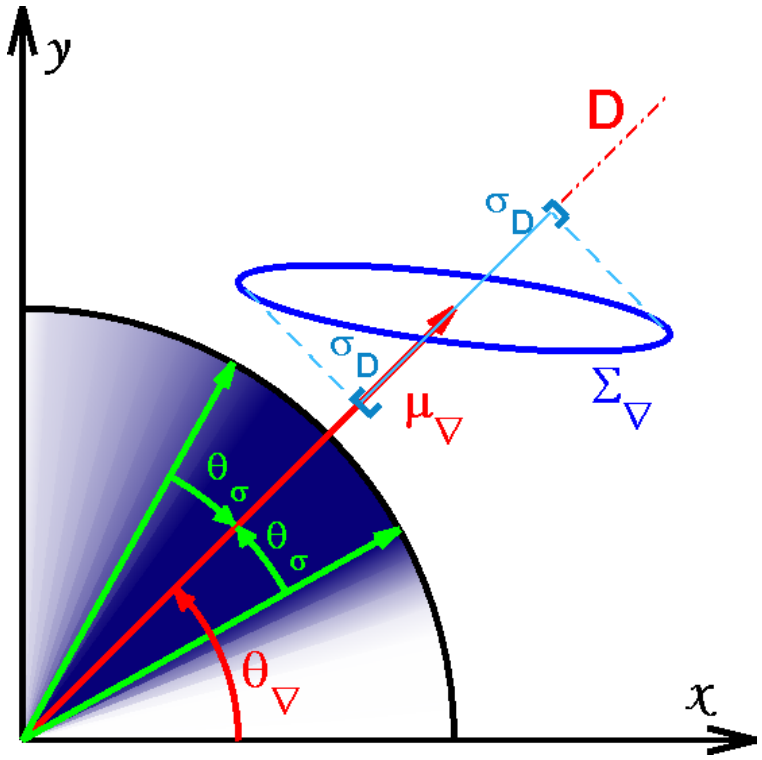

Figure 5.24: Red: Mean gradient μ_∇ and angle θ_∇ it takes with the x-axis. Blue: Covariance ellipse and standard deviation of derivative along the mean gradient direction. Green: Angular confidence interval given by the two equal angles θ_σ. Probability density $p_\leftrightarrow$ of the gradient orientation in the first quadrant mapped to opacity (dark blue).

Notably, the selected density function only serves the purpose of transferring the covariance ellipse characterizing the gradient uncertainty to the unit circle, and bears no relevance to the actual probability distribution of the gradient. Since confidence intervals are computed instead of probability density functions for the gradient orientation, any other distribution characterized by the gradient mean μ_∇ and covariance matrix Σ_∇ could have been assumed. Nonetheless, the Gaussian distribution has been chosen due to the mentioned benefits.

For a given vector $\mathbf{g}$, the bi-variate probability density function of ∇Y is defined as

$$p_\nabla(\mathbf{g}) = \frac{1}{2\pi\sqrt{\det \Sigma_\nabla}} \exp\left(-0.5(\mathbf{g}-\mu_\nabla)^\top \Sigma_\nabla^{-1}(\mathbf{g}-\mu_\nabla)\right), \tag{5.43}$$

the equation representing both the uncertainty in direction and the derivative uncertainty. To isolate the uncertainty in direction, a coordinate transformation from Cartesian to polar coordinates

is performed:

$$p_\nabla(\theta, r) = \frac{r \exp\left(E(\theta, r)\right)}{2\pi\sqrt{\det \Sigma_\nabla}}, \quad \theta \in [0, 2\pi], \quad r \in [0, \infty[, \tag{5.44}$$
$$E(\theta, r) = \left(-\frac{1}{2}\mathbf{h}(\theta, r)^\top \Sigma_\nabla^{-1} \mathbf{h}(\theta, r)\right),$$
$$\mathbf{h}(\theta, r) = r \begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix} - \mu_\nabla.$$

For every angle θ, a probability density value can be obtained by integrating the bi-variate gradient distribution over all possible magnitudes, i.e., by evaluating the θ-marginal

$$p_\nabla^\theta(\theta) = \int_0^\infty p_\nabla(\theta, r)\,\mathrm{d}r, \quad \theta \in [0, 2\pi]. \tag{5.45}$$

To derive the formula for the distribution of the gradient direction, first, the θ-marginal is expressed by the integral of the product of a first order polynomial with an arbitrary Gaussian function

$$p_\nabla^\theta(\theta) = C \int_0^\infty r \exp\Big(-(ar^2 + 2br + c)\Big)\,\mathrm{d}r, \tag{5.46}$$

with variables

$$C = \frac{1}{2\pi\sqrt{\det \Sigma_\nabla}}, \quad a = H(\cos\theta, \sin\theta), \quad b = t\sqrt{a}, \quad \text{and} \quad c = H(\mu_{\nabla x}, \mu_{\nabla y}), \tag{5.47}$$

and the auxiliary terms

$$H(x, y) = \frac{1}{2(1-\rho_\nabla^2)}\left(\frac{x^2}{\sigma_{\nabla x}^2} + \frac{y^2}{\sigma_{\nabla y}^2} - 2\rho\frac{xy}{\sigma_{\nabla x}\sigma_{\nabla y}}\right), \tag{5.48}$$
$$t = \frac{-\mu_{\nabla x}\sigma_{\nabla y}^2 \cos\theta - \mu_{\nabla y}\sigma_{\nabla x}^2 \sin\theta}{2\det\Sigma_\nabla\sqrt{H(\cos\theta, \sin\theta)}} \tag{5.49}$$
$$+ \frac{\rho_\nabla\sigma_{\nabla x}\sigma_{\nabla y}(\mu_{\nabla x}\sin\theta + \mu_{\nabla y}\cos\theta)}{2\det\Sigma_\nabla\sqrt{H(\cos\theta, \sin\theta)}}. \tag{5.50}$$

It follows that

$$p_\nabla^\theta(\theta) = C \int_0^\infty r \exp\Big(-(ar^2 + 2br + c)\Big)\,\mathrm{d}r = \tag{5.51}$$
$$= C \exp\left(\frac{b^2}{a} - c\right) \int_0^\infty r \exp\left(-\left(\sqrt{a}r + \frac{b}{\sqrt{a}}\right)^2\right)\mathrm{d}r, \tag{5.52}$$

where the change of variables $\sqrt{a}r + b/\sqrt{a} = u$ gives

$$p^{\theta}_{\nabla}(\theta) = \frac{C}{a}\exp\left(\frac{b^2}{a} - c\right)\int_{\frac{b}{\sqrt{a}}}^{\infty}\left(u - \frac{b}{\sqrt{a}}\right)\exp\left(-u^2\right)\mathrm{d}u = \tag{5.53}$$

$$= \frac{C}{2a}\exp\left(\frac{b^2}{a} - c\right)\left(\exp\left(-\frac{b^2}{a}\right) - \frac{2b}{\sqrt{a}}\int_{\frac{b}{\sqrt{a}}}^{\infty}\exp\left(-u^2\right)\mathrm{d}u\right) = \tag{5.54}$$

$$= \frac{C\exp\left(-c\right)}{2a}\left(1 - \sqrt{\pi}\frac{b}{\sqrt{a}}\exp\left(\frac{b^2}{a}\right)\left(1 - \mathrm{erf}\left(\frac{b}{\sqrt{a}}\right)\right)\right). \tag{5.55}$$

This can then be re-written to yield

$$p^{\theta}_{\nabla}(\theta) = \frac{\exp\left(-H(\mu_{\nabla x}, \mu_{\nabla y})\right)\left(1 - \sqrt{\pi}\, t\, \exp\left(t^2\right)\left(1 - \mathrm{erf}(t)\right)\right)}{4\pi\sqrt{\det\Sigma_{\nabla}}\, H(\cos\theta, \sin\theta)}, \tag{5.56}$$

with the auxiliary terms

$$H(x, y) = \frac{1}{2(1 - \rho^2_{\nabla})}\left(\frac{x^2}{\sigma^2_{\nabla x}} + \frac{y^2}{\sigma^2_{\nabla y}} - 2\rho\frac{xy}{\sigma_{\nabla x}\sigma_{\nabla y}}\right), \tag{5.57}$$

$$t = \frac{-\mu_{\nabla x}\sigma^2_{\nabla y}\cos\theta - \mu_{\nabla y}\sigma^2_{\nabla x}\sin\theta}{2\det\Sigma_{\nabla}\sqrt{H(\cos\theta, \sin\theta)}} + \frac{\rho_{\nabla}\sigma_{\nabla x}\sigma_{\nabla y}(\mu_{\nabla x}\sin\theta + \mu_{\nabla y}\cos\theta)}{2\det\Sigma_{\nabla}\sqrt{H(\cos\theta, \sin\theta)}}. \tag{5.58}$$

Here, $\mu_{\nabla x}$ and $\mu_{\nabla y}$ are the components of the mean gradient vector μ_{∇}, $\sigma_{\nabla x}$ and $\sigma_{\nabla y}$ are the standard deviations of the gradient components (square roots of the diagonal entries of Σ_{∇}), and $\rho_{\nabla} = \Sigma_{\nabla xy}/(\sigma_{\nabla x}\sigma_{\nabla y})$ is the corresponding correlation value. These values can be computed at every grid point (cf. Equ. (5.32) and (5.33)).

Equ. (5.56) expresses the probability density function of the gradient direction. But, since one is interested in using the probability density for assessing the stability of the orientation of geometric structures in the data, e.g., the orientation of an iso-contour at a certain location, the probability of occurrence of angle θ should include the probability of occurrence of $\theta + \pi$. Thus, the probability density function of the gradient orientation is used instead of the gradient direction, and the respective interval of length 2π, centered at the mean gradient direction θ_{∇} is considered. The probability density function

$$p_{\leftrightarrow}(\theta) = p^{\theta}_{\nabla}(\theta) + p^{\theta}_{\nabla}(\theta + \pi), \quad \theta \in [\theta_{\nabla} - \pi/2, \theta_{\nabla} + \pi/2], \tag{5.59}$$

serves this purpose, because it adds the two probability densities of the positive and negative gradient direction around the mean direction.

The one-dimensional continuous probability density function $p_{\leftrightarrow}(\theta)$ can now be parameterized by

mean θ_∇ and standard deviation θ_σ, where θ_σ is computed as

$$\theta_\sigma = \sqrt{\int_{\theta_\nabla - \pi/2}^{\theta_\nabla + \pi/2} (\theta - \theta_\nabla)^2 p_\leftrightarrow(\theta)\,\mathrm{d}\theta}. \tag{5.60}$$

The standard deviation is used to define the symmetric confidence interval for the orientation uncertainty, centered at the gradient mean direction and with an arc length equal to twice the standard deviation (cf. also Fig. 5.24, where, unlike the probability density, the derived confidence interval is symmetric around the mean).

For a low orientation uncertainty of the gradient, θ_σ approaches zero. The angular standard deviation reaches its maximum at $\theta_\sigma = \pi/2$. For a high orientation uncertainty, however, the distribution $p_\leftrightarrow$ tends to become uniform and the normalized standard deviation $\theta_\sigma^N := \frac{2}{\pi}\theta_\sigma$ approaches the standard deviation of the uniform distribution, $\sigma_u = 1/(\sqrt{3}) \approx 0.58$, values higher than σ_u being possible only for *multi-modal* probability density functions of the gradient orientation. The parameter θ_σ^N acts as a normalized symmetric indicator of the spread in gradient orientation.

5.3.3 Visualization of Derivative Uncertainty

The goal is to provide a visualization that allows to analyze simultaneously the shape of iso-contours in the mean values (which gives important contextual information regarding the gradient orientation), the mean gradient magnitude μ_D, and the *absolute* derivative uncertainty σ_D at every point in the domain. In addition, a visualization is presented, showing the *significance* of the absolute uncertainty, which is indicated by the relative width of a confidence interval with respect to the mean value. Only by analyzing this *relative uncertainty*, can the stability of the slope along the gradient direction be revealed. For instance, if $\mu_D - \sigma_D$ is close to zero or negative, a gradient is likely to disappear or even to invert.

To allow classifying points according to the relative uncertainty, the measure

$$\epsilon(\mathbf{x}_{i,j}) := 1 - \frac{\mu_D(\mathbf{x}_{i,j})}{\mu_D(\mathbf{x}_{i,j}) + \sigma_D(\mathbf{x}_{i,j})} \in [0,1] \tag{5.61}$$

is defined for this uncertainty. For the special case $\mu_D(\mathbf{x}_{i,j}) = \sigma_D(\mathbf{x}_{i,j}) = 0$, $\epsilon(\mathbf{x}_{i,j}) = 1$ is set. Values of the relative uncertainty approach 0.5 wherever the absolute uncertainty becomes equal to the mean gradient magnitude, and range within $[0.5, 1]$ for the critical case $\sigma_D > \mu_D$.

To convey the basic shapes of the iso-contours in the mean scalar field, as well as the spatial differences in the mean derivative, patterns are used that are oriented according to the iso-contours and vary with a frequency indicating the derivative strength. First, the range of mean values is partitioned into a number of N equally spaced intervals, every interval having a width of $(\mu_{\max} - \mu_{\min})/N$, where $\mu_{\min}$ and $\mu_{\max}$ are the smallest and largest mean value, respectively. Then, each

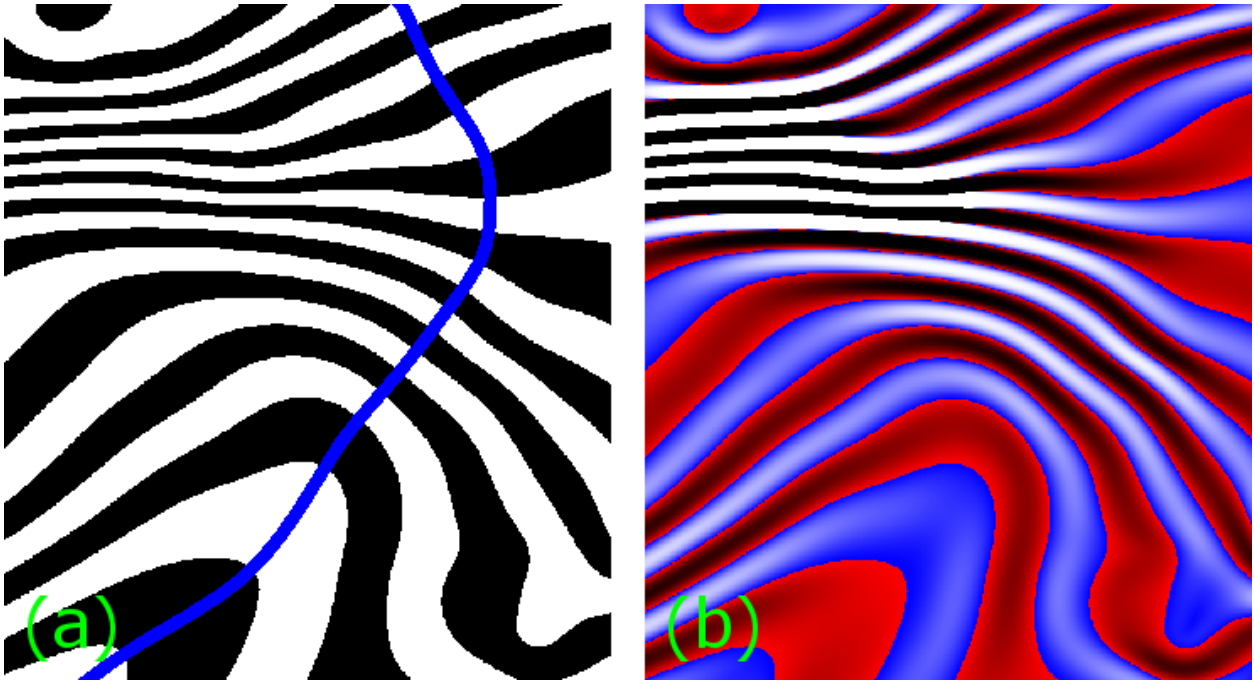

Figure 5.25: (a) The data range of the scalar field is partitioned into equally-spaced intervals, colored in black and white. The thickness of the resulting bands corresponds to the local gradient magnitude. The blue normal curve is everywhere tangential to the gradient field. (b) The gradient magnitude for the same 2D scalar field is now also encoded into color diffusion (high and low diffusion in regions with low and high gradient magnitudes, red and blue diffuse into the black and white bands, respectively).

interval is assigned either 0 or 1, in alternating order. The mapping of the data value at a grid point $\mathbf{x}_{i,j}$ is performed via

$$\kappa(\mathbf{x}_{i,j}) := \left\lfloor \frac{\mu(\mathbf{x}_{i,j})N}{\mu_{\max} - \mu_{\min}} \right\rfloor \bmod 2, \tag{5.62}$$

where $\mu(\mathbf{x}_{i,j})$ is the mean value at the point. Fig. 5.25 (a) illustrates a mapping of the κ values to black and white. In regions where the derivative in mean gradient direction is low, the white and black bands are stretched, narrowing with increasing derivative. The principal shape of the iso-contours reflects in the shape of the bands. The resulting image is called the *contour map*.

It is clear that the shape of the iso-contours and the gradient magnitudes can also be visualized directly by color-coding the initial mean data values. However, by using the proposed banding structure, the use of colors can be left for the visualization of the derivative uncertainty. Furthermore, the transition regions between the different bands provide an additional visual clue that can be used effectively to encode the uncertainty.

Absolute Uncertainty

To visually encode the confidence intervals $[\mu_D - \sigma_D, \mu_D + \sigma_D]$, a diffusion-like process is proposed that transports a *diffusion color* across the bands, and blends this color with the color of the bands. The initial color of the bands is called the *background color*. For the sake of clarity, for now it is assume that a band's background color is either black or white, and that the same diffusion color is used for all bands with the same background color. Different diffusion colors are used for the black and white bands. The diffusion is simulated in such a way that the diffusion color is injected at a band's boundary and transported into its interior. Thus, over time a band's background color mixes increasingly more with the diffusion color, taking on this color in the limit.

The confidence interval at every domain point is encoded by mapping the interval's width to the *diffusion time*, which results in a spatially varying diffusion strength. To illustrate this effect and to exemplify the dependency of the diffusion strength from the encoded values, in Fig. 5.25 (b) the gradient magnitude — also shown in the underlying contour map — is mapped to the diffusion time. The diffusion color was set to red and blue, respectively, for bands having black and white background colors. Later in this section, it will be shown how this technique can be extended to also encode the relative uncertainty, i.e., the spatial differences in the significance of uncertainty, by making the background and diffusion colors dependent on the relative uncertainty.

At each spatial point, a color diffusion is simulated that takes place along the *normal curve*, i.e., the curve passing through the point and oriented along the gradient direction. A normal curve crosses the boundary between two bands orthogonally (cf. blue curve in Fig. 5.25 (a)).

The diffusion process changes the color at a point from its initial background color to a mixture of the background and the diffusion color. The *diffusion value*, i.e., the mixing ratio of the two colors, is determined by simulating how much of the diffusion color has been transferred to the point along the normal curve, starting at the boundary of the band containing the point. The time over which the diffusion is simulated is made dependent on the absolute derivative uncertainty, resulting in stronger diffusion in more uncertain regions. The diffusion value ranges from 0 (full diffusion and dominating diffusion color) to 1 (no diffusion and dominating background color).

The color $c(\mathbf{x}_{i,j})$ at a point is determined by first computing the diffusion value $d(\mathbf{x}_{i,j})$ at the point, and then using this value for blending the diffusion color c_d with the background color c_b:

$$c(\mathbf{x}_{i,j}) = (1 - d(\mathbf{x}_{i,j}))c_d(\mathbf{x}_{i,j}) + d(\mathbf{x}_{i,j})c_b(\mathbf{x}_{i,j}). \tag{5.63}$$

Instead of simulating a physical diffusion process at run-time, a 2D diffusion texture T is pre-computed, and the actual diffusion value is looked up at a point and for a certain diffusion degree in this texture. The texture T is parameterized over the relative position u of a point between the

two boundary lines of the band containing this point (from 0 to 1), and the degree of diffusion v (from 0 to 1).

T is obtained in two steps: First, an intermediate 2D texture τ is created, parameterized over the diffusion time v_τ from 0 to a selected maximum $v_{\tau_{max}}$, and over the relative point position ($u_\tau = u$). The values of τ at the texture coordinates (u_τ, v_τ) are computed by solving the following classical 1D heat equation with given boundary values and initial conditions

$$\frac{\partial \tau}{\partial v_\tau}(u_\tau, v_\tau) = \frac{\partial^2 \tau}{\partial u_\tau^2}(u_\tau, v_\tau), \quad u_\tau \in [0,1], \quad v_\tau \in [0, v_{\tau_{max}}], \tag{5.64}$$

$$\tau(0, v_\tau) = \tau(1, v_\tau) = 0, \quad v_\tau \in [0, v_{\tau_{max}}], \tag{5.65}$$

$$\tau(u_\tau, 0) = 1, \quad u_\tau \in]0,1[. \tag{5.66}$$

The heat equation models the diffusion in the 1D domain $u_\tau \in [0,1]$ over the diffusion time v_τ. The initial condition in Equ. (5.66) models the situation where no diffusion takes place and only the background color is visible. The boundary condition in Equ. (5.65) enforces the diffusion from both boundaries of a band towards its interior.

The heat equation can be solved using a Fourier approach, giving the solution

$$\tau(u_\tau, v_\tau) = \sum_{k=1}^{\infty} \frac{2}{k\pi} \left(1 - \cos(k\pi)\right) \sin(k\pi u_\tau) \exp(-k^2 \pi^2 v_\tau). \tag{5.67}$$

These are the diffusion values stored in T. Since the summands become ever smaller for increasing k, the summation can be terminated after a sufficiently large number of terms have been added. In practice, the number of summands for an error below, e.g., single precision, is dependent on the resolution of the texture T.

The next step performs a parameter transformation, so that the v texture coordinates are $\in [0,1]$ instead of $\in [0, v_{\tau\,\max}]$. This is achieved via the mapping

$$\lambda(v_\tau) = \int_0^1 \tau(u_\tau, v_\tau)\, \mathrm{d}u_\tau, \tag{5.68}$$

which is strictly monotonic from 1 to 0 for $v_\tau \in [0, v_{\tau\,\max}]$. Because one needs to map from $v \in [0,1]$ to $v_\tau \in [0, v_{\tau\,\max}]$, the inverse function λ^{-1} is needed. While this does not have a closed form, it can be obtained using a back-mapping strategy between the domain $[0, v_{\tau\,\max}]$ and its image $\lambda([0, v_{\tau\,\max}])$. The values in the diffusion texture T (see Fig. 5.26) are then computed as

$$T(u, v) = \tau(u, \lambda^{-1}(v)), \quad u \in [0,1], \quad v \in [0,1]. \tag{5.69}$$

To finally make the texture coordinate u depend on the relative position of a point between the

Figure 5.26: The diffusion texture T. Diffusion values from 0 to 1 are mapped to a grayscale color map. The x-axis represents the relative position u of a point between two band boundaries. The y-axis represents the degree of diffusion v.

band boundaries, and v on the absolute derivative uncertainty, the following linear mappings to $[0, 1]$ are performed:

$$u(\mathbf{x}_{i,j}) = \frac{\mu(\mathbf{x}_{i,j})N}{\mu_{\max} - \mu_{\min}} - \left\lfloor \frac{\mu(\mathbf{x}_{i,j})N}{\mu_{\max} - \mu_{\min}} \right\rfloor, \tag{5.70}$$

$$v(\mathbf{x}_{i,j}) = \frac{\sigma_D(\mathbf{x}_{i,j})}{\sigma_D^{\max}}. \tag{5.71}$$

The color at a point $\mathbf{x}_{i,j}$ is then computed by using $T(u, v)$ as the diffusion value $d(\mathbf{x}_{i,j})$ in Equ. (5.63).

Relative Uncertainty

The following paragraphs introduce the use of different background *and* diffusion colors to simultaneously encode the absolute and relative derivative uncertainty into the contour map. Therefore, a color map is selected to encode $\epsilon \in [0, 1]$ (cf. Equ. (5.61)) into colors c_r. Even though this color map can be arbitrary in general, it must adhere to some requirements in the current scenario: Firstly, different hues are expected to allow separating different "key-regions", such as the regions around 0, 0.5, and 1, as well as the regions in between. Thus, one seeks for a color map with at least three major identifiable key colors. A two-color or gray-scale color table cannot achieve this with an acceptable color contrast. Secondly, the color map must contain only colors with maximum saturation and intensity, and it should not contain white or black. This is because the absolute uncertainty is encoded into the saturation/intensity of colors in HSV color space, similar to [Hen03], which is performed by blending ϵ (encoded into hue) with either black or white in

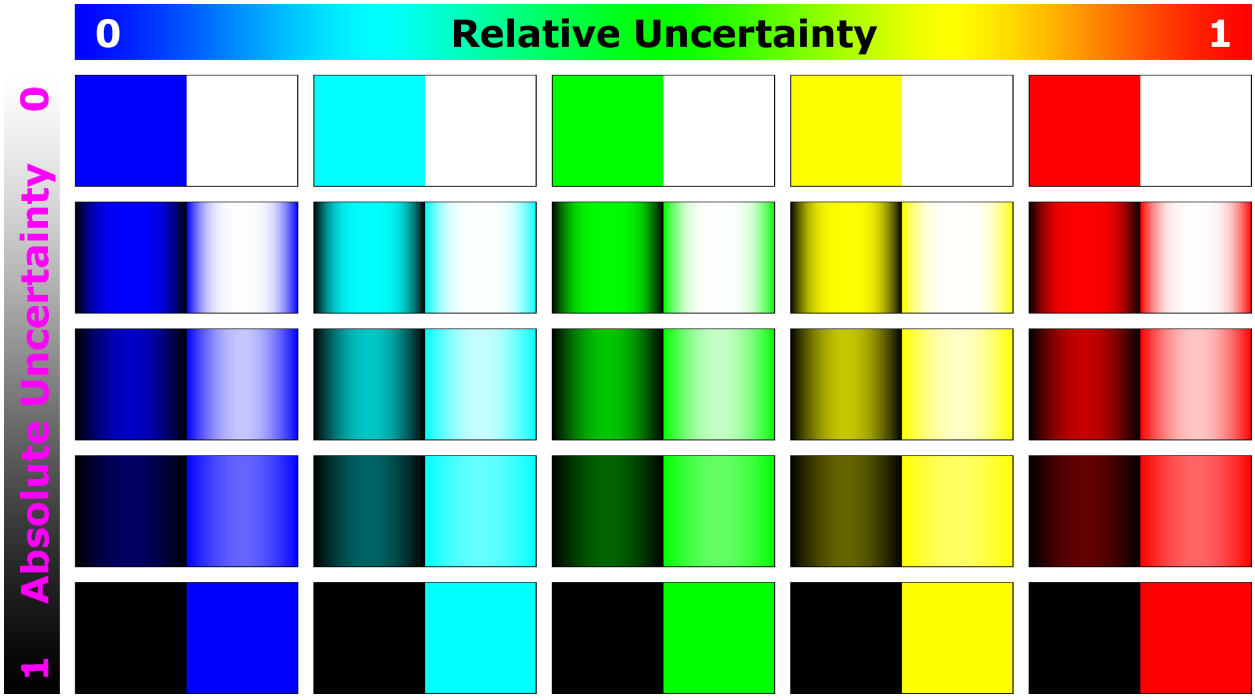

Figure 5.27: 25 cutouts of two neighboring contour bands are illustrate for different combination of absolute (vertical axis) and relative (horizontal axis) derivative uncertainty values.

the diffusion process. Consequently, color maps which let the color go over white or black (e.g., [Mor09, BPC*10, KVUS*05]) cannot be used, because they cannot always distinguish between the absolute and relative uncertainty.

In this work, the rainbow color map [blue → cyan → green → yellow → red] (cf. Fig. 5.27 color map in uppermost row) is utilized as an example for encoding the relative uncertainty, though this may not be an optimal design. The hue range is used from blue to red over green (instead of magenta), to increase the color contrast and allow a better visual differentiation of the key-regions.

In the presented model, different background and diffusion colors are used for the black and white bands in Fig. 5.25 (a). The RGB background colors for the bands in Equ. (5.63), initially white and black, respectively, is set to $c_b^1 = (1,1,1)$ and $c_b^2 = c_r$. The diffusion colors are set to $c_d^1 = c_r$ and $c_d^2 = (0,0,0)$, respectively. Consequently, in regions where the absolute uncertainty is low, only the background colors are visible, color-coding the strength of the relative uncertainty. With increasing absolute uncertainty, black diffuses into the colored bands and c_r diffuses into the white bands. Where the absolute uncertainty is high, the white bands are completely covered by c_r and the other bands are covered by black.

Fig. 5.27 illustrates the proposed diffusion scheme through 25 cutouts of two neighboring contour bands for different combinations of absolute and relative uncertainty values. The horizontal axis represents relative uncertainty values $\epsilon \in [0,1]$, mapped to the chosen rainbow color map. Along

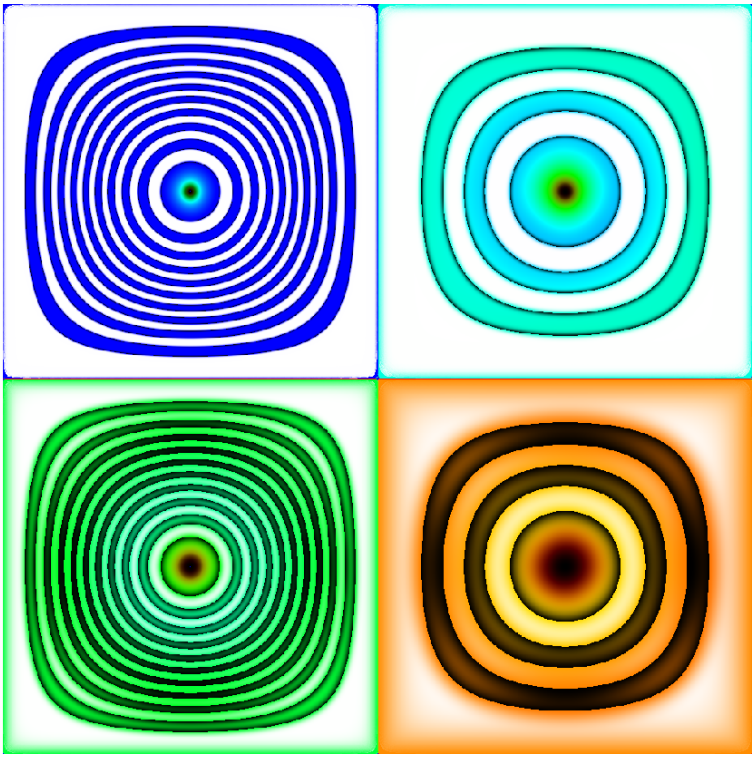

Figure 5.28: Visualization of the uncertainty in gradient derivative in an ensemble data set comprising four regions with different means, absolute, and relative uncertainties.

the vertical axis, the normalized absolute uncertainty value (cf. Equ. (5.71)) increases linearly from zero to one. For each combination, the respective color diffusion in two neighboring bands is shown. It can be noticed that combining the diffusion scheme and the color mapping allows the simultaneous visualization of both uncertainty entities. Furthermore, due to the cross-over diffusion, the relative uncertainty color is visible independently of the degree of absolute uncertainty.

In Fig. 5.28, the proposed diffusion scheme was applied to an ensemble data set comprising four well-separated regions with different absolute and relative uncertainties in gradient magnitude. A more detailed description of the process used to generate this data is given at the beginning of the results section.

As indicated by the underlying contour maps, the mean values in all four regions show a radially symmetric change in value with different slopes towards the center of the respective 2D sub-domain. In the upper examples, almost no diffusion takes place and only white and colored bands are visible. This indicates low absolute derivative uncertainty. In the lower examples, the colored and white bands turn into black and color, respectively, due to significant diffusion. Here, the gradients are affected by strong absolute uncertainty. In the upper left example, the blue color indicates very low relative uncertainty. The cyan color in the upper right example indicates a ratio of approximately $3:1$ between the mean derivative magnitude μ_D and the derivative uncertainty. In the lower left

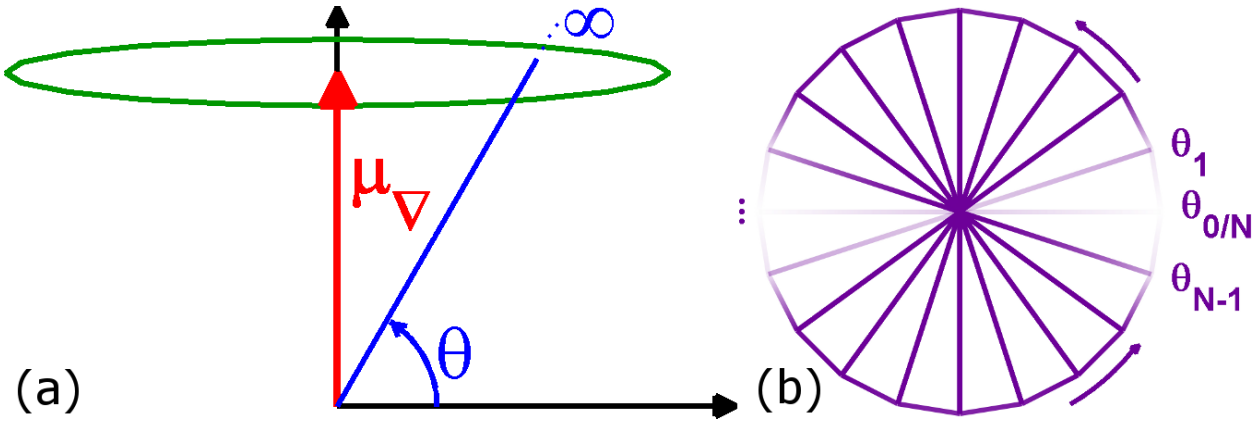

Figure 5.29: (a) 2D domain for Gaussian probability density of a vector quantity is shown. The red arrow and green covariance ellipse indicate the mean vector and confidence region, respectively. The gradient direction probability $p^{\theta}_{\nabla}(\theta)$ for an angle θ is obtained by integrating the 2D probability density function along the blue line from 0 to ∞. (b) Triangle-based glyph representation is shown. Each off-center vertex at angle θ_i is assigned a transparency according to the probability density function $p_{\leftrightarrow}$ of the angular orientation. The color is constant for all vertices and determined by the standard deviation of $p_{\leftrightarrow}$.

example, the green color ($\epsilon \approx 0.5$) shows that σ_D is as large as μ_D. Therefore, the strong mean derivative indicated by the band structure is highly uncertain. The orange color in the lower right example represents a ratio of approximately $1:7$ between μ_D and σ_D.

5.3.4 Visualization of Orientation Uncertainty

To visualize the uncertainty in orientation, the standard deviation θ_σ (cf. Equ. (5.60)) around the mean gradient direction θ_∇ is encoded into a circular glyph representation. Each glyph is represented by a triangle fan, comprising a number of equally-sized triangles that are arranged around the glyph center-point (cf. Fig. 5.29). The gradient uncertainty parameters μ_∇ and Σ_∇, which are assumed to be constant for the entire glyph, are computed at the data point at which the glyph is centered. For each triangle, color and transparency values are assigned to the two off-center vertices, while the center vertex is assigned the mean color and transparency of these vertices. Starting with $\theta_0 = 0$ at the east vertex, each off-center vertex i is assigned an angle $\theta_i = 2\pi(i/N)$ in counterclockwise order, where N is the number of off-center vertices. This ordering is illustrated in Fig. 5.29 (b).

All vertices of a glyph have the same color, which encodes the degree of uncertainty. The color of a glyph is obtained by mapping the normalized standard deviation at the glyph center point, $\theta^N_\sigma = 2\theta_\sigma/\pi$, to color, via a pre-defined color map, as shown in Fig. 5.30. The color of the glyphs goes from green over blue to magenta, as the uncertainty in the gradient orientation varies from low

Figure 5.30: Color map for normalized orientation standard deviation θ_σ^N.

to high. The highest uncertainty, corresponding to a continuous uniform orientation probability density, has a normalized standard deviation of 0.58. Normalized standard deviations higher than 0.58 occur for multi-modal probability densities, the color of such glyphs tending to red.

To allow the user to distinguish between points at which the orientation distribution is more or less uniform, the transparency of the glyphs is modified accordingly. Thus, the transparency value for each off-center vertex is either one or close to zero, depending on whether the angle of the vertex falls within the angular confidence interval around the mean angle θ_∇ or not. The opacity for points outside the angular confidence interval is set to a value slightly higher than zero, so that the individual glyphs are well-separated.

For the particular case of a Gaussian distribution of the scalar fields, where the gradients are also Gaussian distributed, the visualization method also includes the orientation probability density (cf. Equ. (5.59)) for the transparency, the color of the glyph remaining unchanged. An orientation probability density value is computed at every off-center vertex by evaluating Equ. (5.59) at the assigned θ_i for $\theta_\nabla - \pi/2 \leq \theta_i \leq \theta_\nabla + \pi/2$ and, for angles θ_i outside this range, at those angles that lie within the range and make an angle of π with the vertices θ_i. To show the individual distributions per glyph, the transparency of a vertex can be set directly to its corresponding orientation probability density value, normalized by the mode of the local probability density:

$$\alpha(\theta) = \frac{p_\leftrightarrow(\theta)}{\max_{\theta \in [\theta_\nabla - \pi/2, \theta_\nabla + \pi/2]} p_\leftrightarrow(\theta)}. \tag{5.72}$$

The latter assignment of probability density values to transparency is more revealing than the general one of confidence intervals. As illustrated in Fig. 5.31, however, the difference is typically insignificant, except in the case of (almost) uniform distributions: While glyphs having the transparency display orientation probability density values become fully opaque, this does not hold true for glyphs where the transparency shows confidence intervals. In the latter case, uni-modal distributions cannot have a standard deviation greater than 0.58 — the standard deviation for uniform distributions, corresponding to maximum uncertainty, which means that only about 60% of the glyph is opaque, due to the confidence interval. Asymmetries around the mean value are also not shown by the confidence intervals technique. However, the latter method does emphasize the multi-modal distributions better, as such glyphs stand out as always colored within the magenta-red range, irrespective of how narrow or large the spreads around their modes are. Nevertheless,

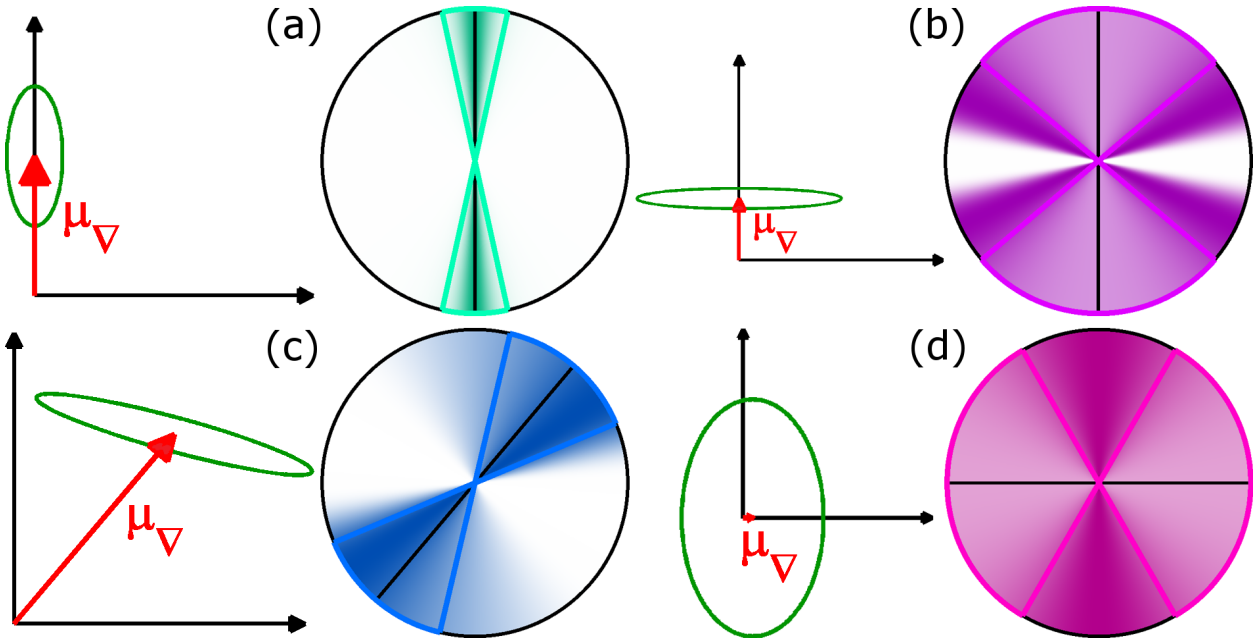

Figure 5.31: 2D gradient distributions and corresponding circular glyphs. A black line within the glyph encodes the mean gradient orientation, while lines around the mean gradient orientation, in the same color as the glyphs, but brighter, show the angular confidence intervals. (a) High gradient magnitude and low orientation uncertainty. (b) Multi-modal probability distribution. Two orientations are equally likely, the mean orientation is very unlikely and therefore unstable. (c) Asymmetric spread of the orientation. (d) Almost zero mean gradient, but strong likelihood of a vertical gradient orientation.

both glyph representation methods benefit similarly of varying transparencies, which causes a clear color contrast between orientations with high and low likelihoods. Note that the confidence intervals representation cannot be applied as is when the mean gradient is zero, because there is no specific direction defined around which to compute the confidence interval. This case is nonetheless infrequent, especially since glyphs are placed at discrete locations.

Fig. 5.31 shows four examples of bi-variate gradient distributions and the corresponding uncertainty glyphs. The mean gradient μ_∇ is shown as a red vector, the green ellipse, related to Σ_∇, indicates the 2D confidence area around the mean vector, and the black line within the glyph represents the mean orientation of the vector. The glyphs are colored according to the color map for the confidence intervals, while the transparency is set to the probability density function of the gradient orientation. Confidence intervals are shown by the thick lines around the mean orientations and have the same color as the glyphs, but brighter.

In (a), the orientation uncertainty is low, because the gradient variation mainly alters the magnitude. In (b), there is a significant uncertainty in orientation. The color and transparency mapping further indicate a multi-modal orientation distribution, i.e., two significant different orientations

are equally likely. Note that the orientation of the mean vector has a low likelihood and is, therefore, unstable. The multi-modality cannot be shown just via the transparency for the confidence intervals. In (c), the zero correlation between the gradient's components in (a) and (b) was changed to a negative value. This resulted in an asymmetric orientation distribution with respect to the mean vector, which is also not displayed by the confidence intervals. In (d), the mean gradient is on the x-axis, close to the zero vector. Although the glyph indicates a strong uncertainty, it becomes apparent that the gradient is most likely directed vertically. The confidence intervals, however, are drawn around the horizontal mean gradient orientation and miss thus the likely vertical orientation of the gradient. It becomes apparent that, while confidence intervals do not correspond entirely to the orientation probability density functions, they nonetheless give a good approximation of the orientation variability of the gradient in most cases.

Because the user is interested in the orientation stability of the gradients, the mean data set is displayed as a contour representation. A user-specified number of iso-lines is displayed for iso-values equidistantly positioned within the data range.

A glyph-based visualization, such as the one in Fig. 5.34 (b), facilitates a rapid understanding of the way uncertainty affects the orientation of the gradients, especially in the regions with lower uncertainty, where the glyphs become less opaque and the geometry of the iso-contours can be visualized concurrently with its uncertainty. This happens irrespective of the placement of the glyphs, because the pattern of the glyphs is typically orthogonal to the iso-contours of the scalar field, making the contours clearly discernible. An alternative would be to consider the orientation uncertainty of the tangent line, rather than that of the gradient, but this would not only produce no new information, it would also lower the contrast in coverage of the iso-contours between low and high uncertainty regions.

5.3.5 Results and Discussion

First, the proposed techniques are applied to a synthetic 2D scalar ensemble, which was designed specifically to demonstrate the additional insights these techniques can provide (cf. Fig. 5.32). A 2D domain was divided into 2×2 quadratic cells, and each cell was triangulated (cf. transparent grid lines in (a)). Random scalar values at the corner and center vertices were modeled via a multi-variate Gaussian random variable $\mathbf{H}$. The first component, H_1, with mean values $\mu(H_1) = 0$, models cell corners values, while four other components H_2, H_3, H_4, and H_5, with means $\mu(H_2) = 1$, $\mu(H_3) = 1$, $\mu(H_4) = 0.3$, and $\mu(H_5) = 0.3$, model the values at the center vertices in the upper and lower cells, respectively. $\mathcal{C}^1$-interpolation was performed between the values at the vertices inside each cell and triangle. Fig. 5.32 (a) shows the mean data values, linearly mapped to color via the color map indicated below the images. As shown in (b), all random variables have a constant standard deviation $\sigma(H_i) = 1$. A strongly positive correlation was modeled between the random

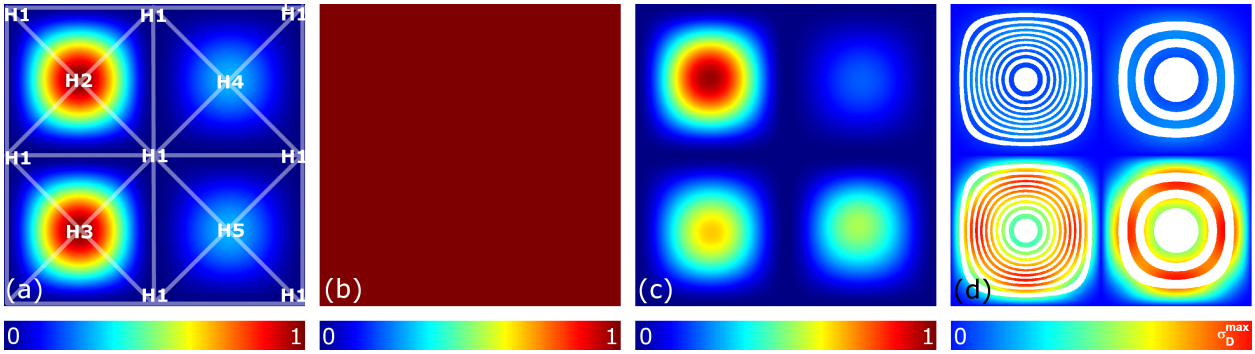

Figure 5.32: (a) Mean values in a synthetic 2D ensemble data set are shown. Data values at vertices of a triangular grid are generated via a multi-variate Gaussian distributed random variable with components H_i. Barycentric interpolation is used in between. (b) The standard deviation of each variable is constant over the entire domain, but different correlations were enforced between the variables H_i. (c) One particular ensemble member (realization) shows vastly different data distributions — and gradient variations thereof — in the four square regions. (d) Color encoding of the absolute derivative uncertainty—mapped from $[0, \sigma_D^{\max}]$ to the given color map — in the contour map.

variables H_1 and H_2, and between H_1 and H_4. All other pairs are uncorrelated (zero pairwise correlation). An ensemble of realizations for $\mathbf{H}$ was generated at the corner and center vertices via a multi-variate Gaussian random number generator.

One realization (ensemble member) of $\mathbf{H}$ is shown in (c). The strength of the derivative into the mean gradient direction can vary significantly in the lower cells, due to the smaller correlation present there. For instance, there is a significant decrease and increase, respectively, in the left and the right cell. Note that the possible variations of the derivative cannot be observed by looking only at the data mean values and standard deviations in (a) and (b).

Fig. 5.32 (d) shows the absolute derivative uncertainty σ_D. The black points in the underlying contour map were colored based on σ_D, by mapping the interval $[0, \sigma_D^{\max}]$ to color via the color map displayed below the image. Although the visualization helps to differentiate regions with low and high absolute derivative uncertainty, the significance of this uncertainty with respect to the mean values cannot be judged. As shown in Fig. 5.28, the presented novel approach can help reveal such information effectively, by using diffusion and color to encode the absolute and relative uncertainty in one single representation.

In the next example, the introduced novel techniques are demonstrated in a geophysics ensemble data set, comprising material variations at a certain depth in the earth's crust. The data set was obtained using seismic tomography and contains an ensemble of relative velocity values for shear waves, originating from earthquake source locations. For details on how the data was acquired and the information it contains, the reader is referred to [MCD*12]. Fig. 5.33 (a) and (b) show the mean value and standard deviation at each point in the covered 2D domain, with values

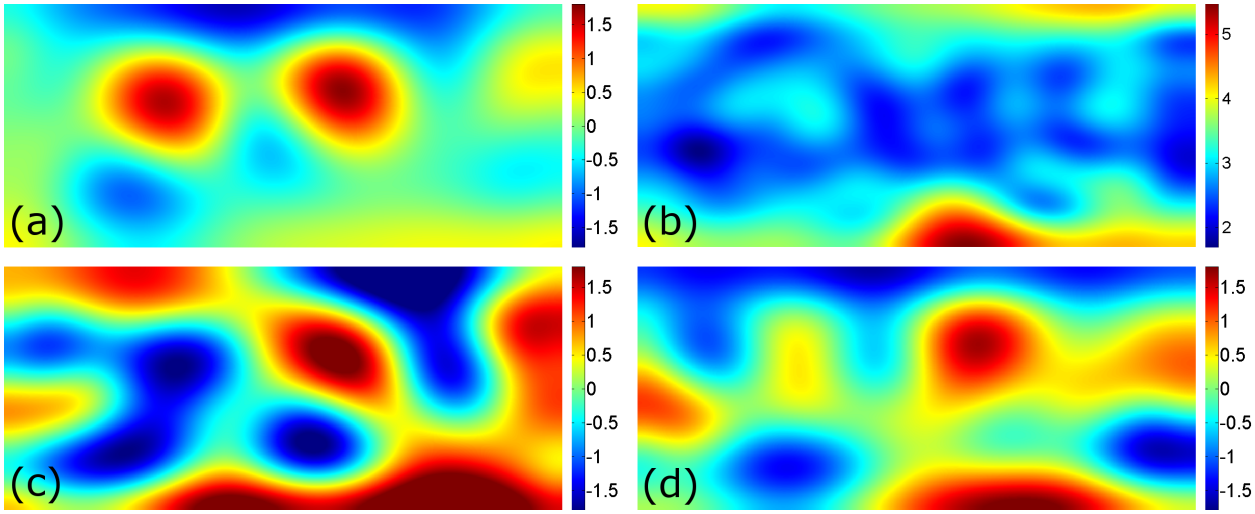

Figure 5.33: Mean values (a) and standard deviations (b) in an ensemble of 2D seismic tomography shear wave velocity variations. Two circular features with very similar occurrences can be distinguished. Two members (realizations) of the ensemble are shown in (c) and (d).

mapped linearly from blue (negative) to red (positive). Two important circular features can be observed in the mean values, but the standard deviations in both regions are almost constant. An uncertainty analysis using only mean values and standard deviations cannot thus reveal any significant differences between the two features.

Color diffusion was applied to the data set in Fig. 5.34 (a) to visualize the derivative uncertainty in the geophysics ensemble data set. The diffusion of black indicates high absolute uncertainty, which can be perceived especially in region (1). The coloring, however, indicates a high variation of the derivative in mean gradient direction across the left circular feature: the green color represents a relative uncertainty of ~ 0.5 ($\mu_D \approx \sigma_D$). The derivatives in this region have thus to be assumed unstable. This conclusion is supported by the visualization of two ensemble members in Fig. 5.33 (c) and (d). The realization in (c) shows strongly positive gradients towards the center in the right feature, but negative gradients in the left feature, i.e., the slope of the left feature is more or less inverted. The blue color of the right circular feature in 5.34 (a) indicates a relative uncertainty value of $\epsilon \approx 0$ and, consequently, significantly more stable derivatives towards the feature's center.

Fig. 5.34 (b), on the other hand, illustrates a quite stable gradient orientation across both circular features. Thus, even though both features have a circular structure, the gradient magnitude can only be assumed stable for the right feature. The values in the left circular region may decrease towards the feature center.

From an application point of view, the visualization helps to identify regions where significant material anomalies are present with respect to the surrounding structures. In the presented example,

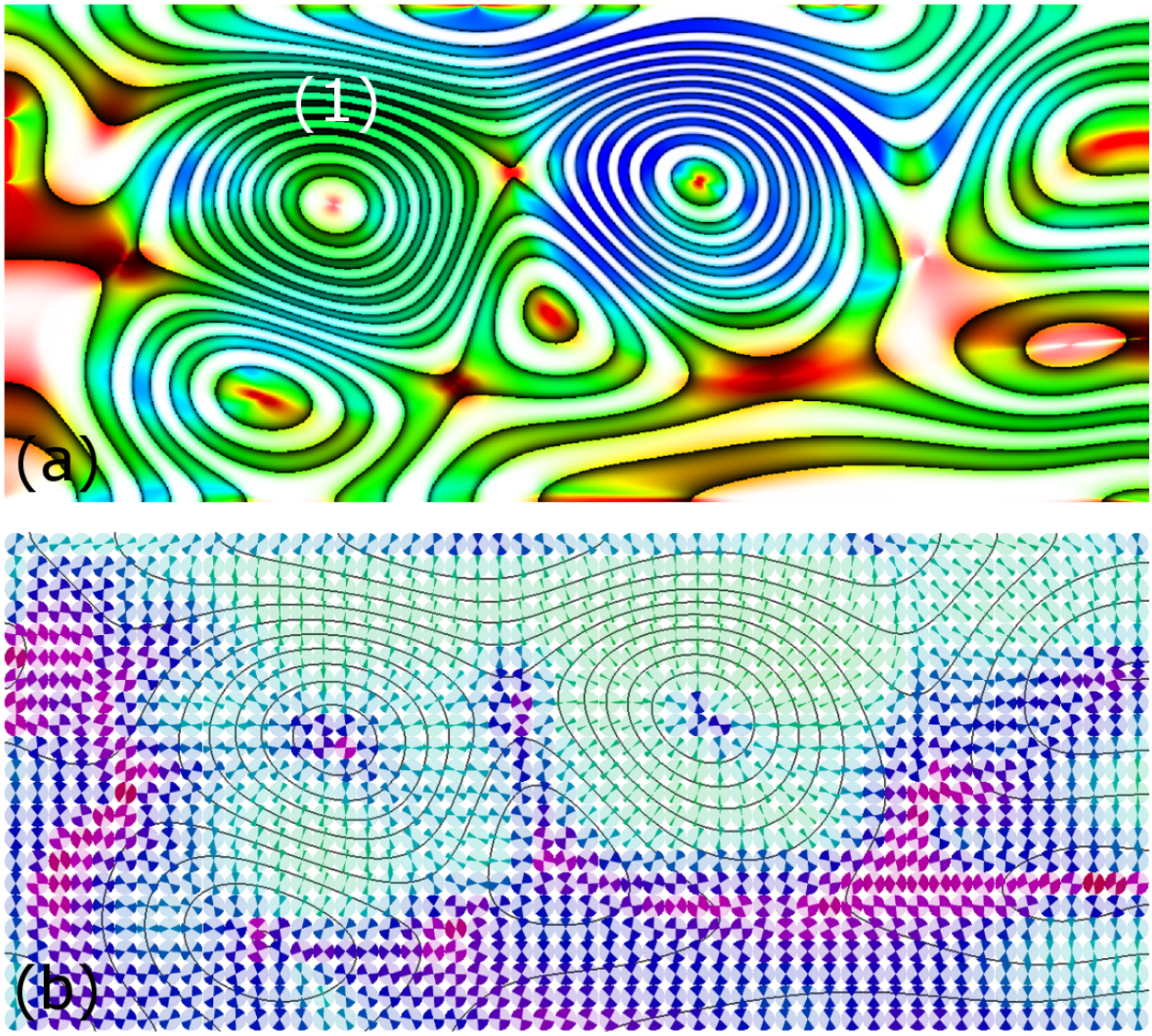

Figure 5.34: Visualization of the gradient uncertainty in the ensemble shown in Fig. 5.33. a) The two circular features can be classified as stable (right feature) and unstable (left feature) with respect to their derivative uncertainty, i.e., strong relative uncertainty (green color for $\mu_D \approx \sigma_D$) in the left region. b) Glyph-based visualization conveys high stability of the gradient orientation for the left and right features. In contrast, the small circular feature in the bottom left part is significantly affected by orientation uncertainty, indicating a very likely change in the orientation of iso-contours passing through this region.

the data contains relative velocity values for earthquake shear waves, which are characteristic for certain material structures in the earth's crust. The user is interested to identify local and global maxima and minima of the relative velocity values, because they can serve as indicators for material anomalies (e.g., minerals, sediments, ores, oil, etc.) in the earth. Since seismic tomography data

sets are always affected by uncertainty, the user is interested in analyzing the respective critical areas (e.g., maxima, minima, strong gradients, etc.) with respect to their stability. As demonstrated, the techniques proposed in this work can effectively enable such an analysis.

In a third example, uncertainty visualization is utilized to identify stable features in the mean values of an ensemble of temperature fields. The ensemble was simulated by the European Center for Medium-Range Weather Forecast (ECMWF) for two different forecast periods and pressure levels above Europe. Mean temperature values, as well as standard deviations and correlations, were computed from the ensemble members.

The mean temperature is shown in Fig. 5.35 (a), with values linearly mapped from blue to red (cold-to-warm mapping). In (b), the standard deviation is shown in temperature units, linearly mapped from blue (low) to red (high). The diffusion coloring and the glyph-based visualization of the orientation uncertainty are shown in (c) and (d).

A first observation is that the strong temperature incline in region (1) appears highly stable. This is indicated by the high-frequent bluish band structure in (c), i.e., strong gradient magnitude and $\epsilon \approx 0$, as well as by the green glyph pattern in (d). Thus, the temperature difference between the blue and cyan regions in (a) is well-resolved, affected by a relatively low uncertainty and significantly different from zero. Interestingly, the standard deviation is relatively strong in this region (cf. (b)). Consequently, the standard deviation does not necessarily allow drawing any conclusion on the variability of the derivative strength, which is also affected by correlation effects. By analyzing in Fig. 5.35 (c) the transition zone between cyan and red, as shown in Fig. 5.35 (a), only the derivatives in (2) and (4) are observed to be stable. Although region (3) exhibits a similar mean derivative, the strong presence of green ($\epsilon \approx 0.5$) indicates a high probability that the derivative becomes zero. Reliable assumptions on the temperature difference can only be made for those regions where strong derivatives are present, i.e., regions (1), (2), and (4).

Uncertainty glyphs colored according to the circular variance are shown in (d). The orientation of the gradients (and therefore the iso-lines) is stable in regions (1), (2), and (4). Although in region (3) a similar gradient strength and a clear orientation can be perceived in the mean data (a), the gradient orientation is affected by higher uncertainty and the structure of the transition zone is not well-resolved.

The fourth example presents another ECMWF temperature forecast ensemble, which was simulated over a shorter forecast period and at a different air pressure level. Fig. 5.36 (a) shows the mean data. The uncertainty visualizations in (c) and (d) reveal a prominent structure in region (1), which is stable with respect to the derivative strength and gradient orientation, and which cannot be easily identified using only the mean and standard deviation. Another interesting region is the transition zone between blue and red in (2), where the standard deviation in (b) shows a significant uncertainty in the data values. This uncertainty strongly affects the uncertainty in the derivatives,

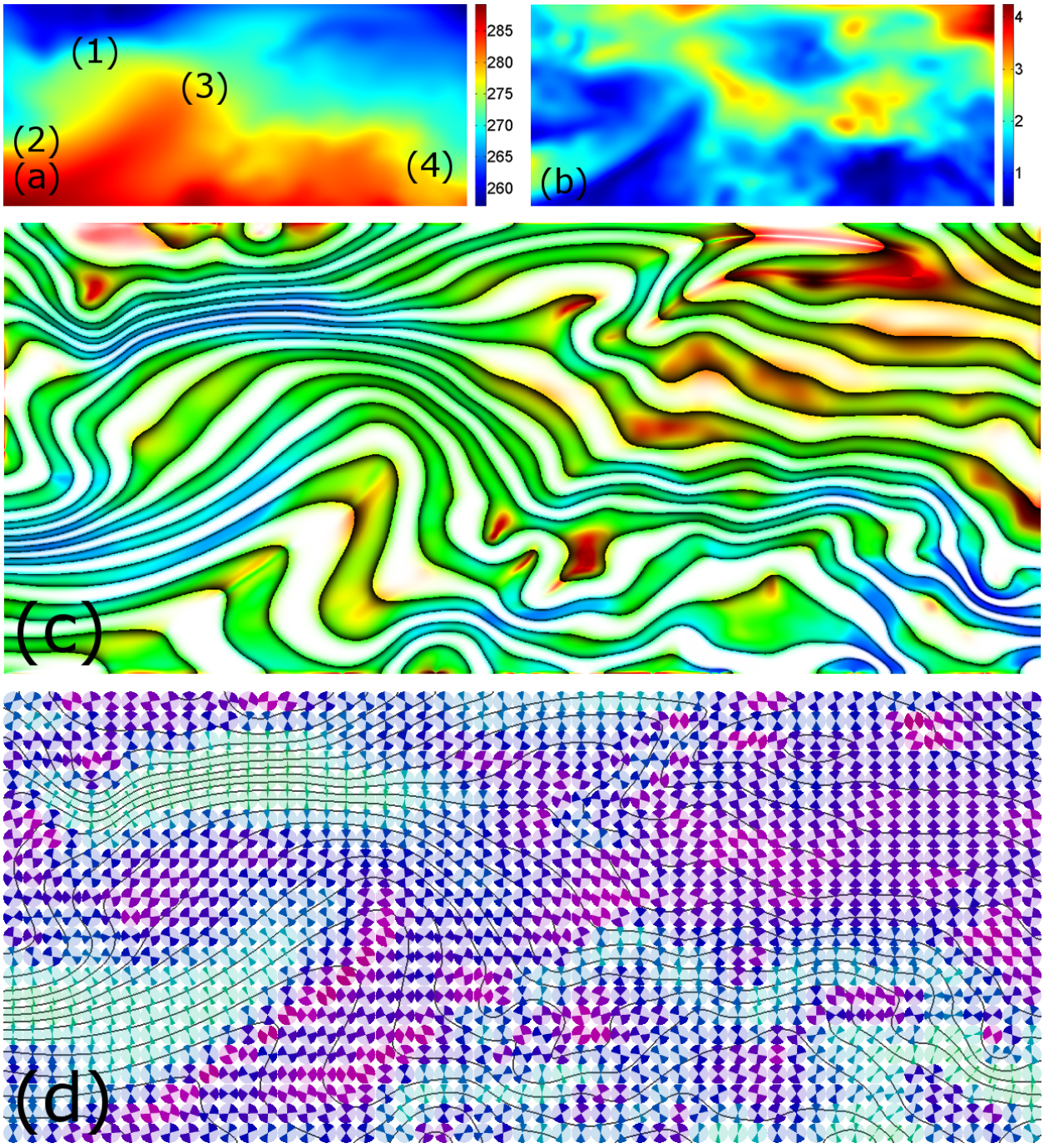

Figure 5.35: Mean values (a) and standard deviations (b) of a temperature ensemble forecast. Diffusion coloring is shown in (c). Strong mean gradients with low derivative uncertainty are identified in regions (1), (2), and (4). Green color reveals a strong relative derivative uncertainty in (3). The visualization of the uncertainty in gradient orientation (d) reveals stable orientations of iso-contours in (1), (2), and (4).

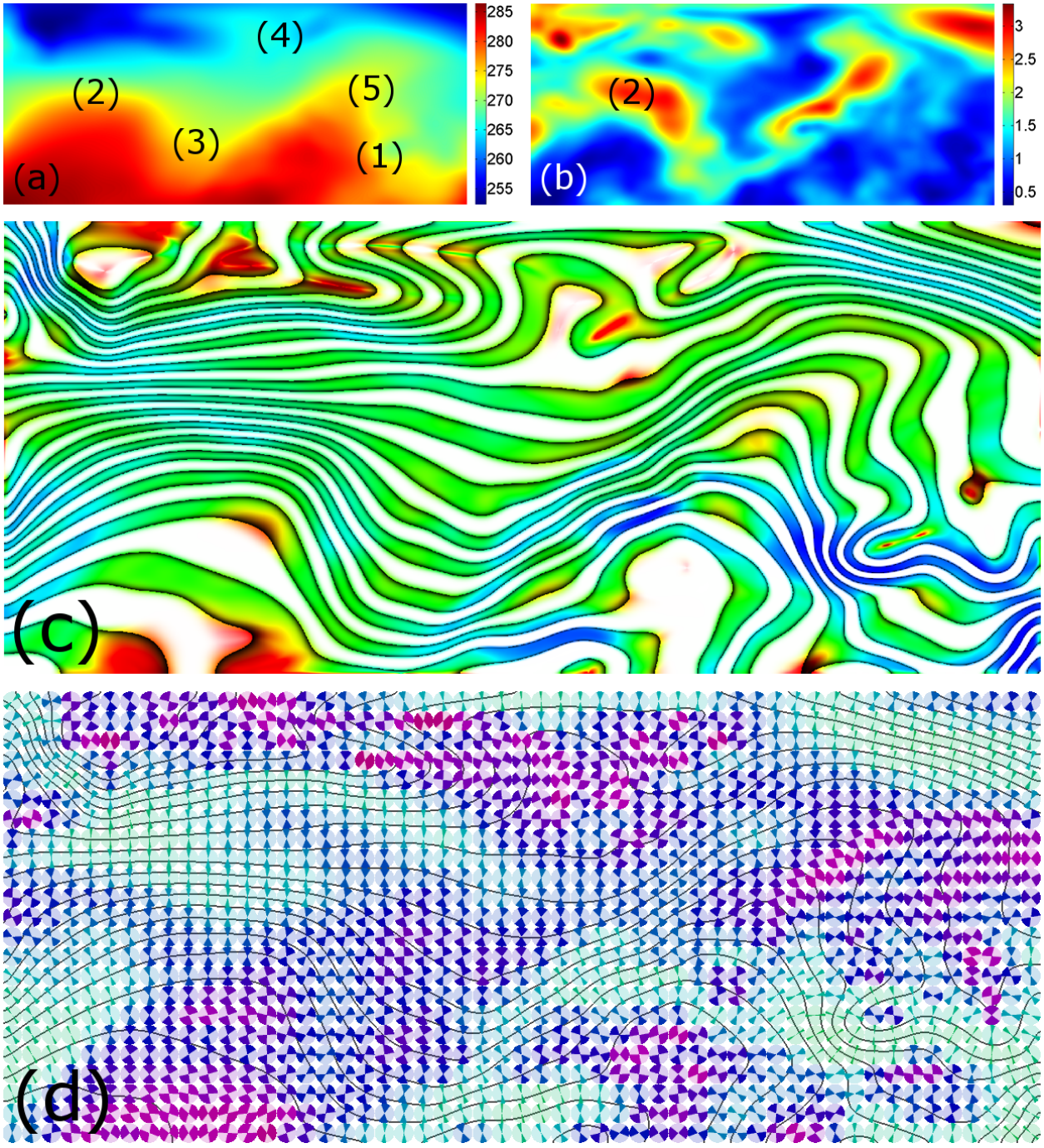

Figure 5.36: Visualizations of mean values (a), standard deviations (b), derivative uncertainty (c), and orientation uncertainty (d) for an ensemble of temperature forecasts. A prominent structure with strong mean gradients and low uncertainty can be identified, for example, in (1). In (2), a strong standard deviation and, therefore, uncertainty in derivative is shown by the green relative uncertainty color in (c). The green narrow glyphs in (d) indicate, nonetheless, that the derivative uncertainty does not affect the gradient orientation significantly. The orientation of the iso-lines is stable in (2), while in (3), (4), and (5), a significant variability in orientation and structure is emphasized.

the confidence region around the strong mean derivative almost going down to zero and resulting in a cyan to green color ($\epsilon \approx 0.4$). Interestingly, the uncertainty in gradient orientation is nearly unaffected in region (2). The orientation of the transition zone is, despite the quite high derivative uncertainty, stable. For the structures in regions (3), (4), and (5), the glyphs indicate a strong orientation variability. Moreover, the presence of color green indicates a high relative derivative uncertainty in region (3) ($\epsilon \approx 0.5$ and $\sigma_D \approx \mu_D$).

In particular in the context of weather forecasts, the proposed visualizations can help to identify regions where significant *and* stable climate changes are present. The resulting benefits are two-fold: Firstly, since weather forecast centers usually combine simulation results from different sites into a single forecast, the presented methods can help determine in which region a particular site is most reliable. This information can then be used for a region-specific selection of sites. Secondly, the uncertainty information can be used to re-parameterize a numerical forecast simulation, for instance, to reduce or even eliminate the observed instabilities in the predicted features. Such an analysis strives for the use of uncertainty for sensitivity analysis, and, herewith, the tuning of simulation technologies towards more reliable and stable outputs.

5.3.6 Conclusion

Two methods for illustrating the variability of gradients in 2D uncertain scalar fields were presented: A color diffusion model for visualizing the absolute and relative derivative uncertainty, and a glyph-based approach for the uncertainty in gradient orientation. Uncertainty parameters for gradients in uncertain discrete scalar fields have been derived, and then the mathematical basis has been introduce to derive confidence intervals for the gradient magnitude and orientation. Throughout the presented derivations, no assumption has been made on the probability distribution modeling the underlying uncertainty. The proposed methods allow a quantitative analysis of the gradient variability, which is required to assess the stability of geometric structures like iso-contours.

One future challenge will be the extension to 3D scalar fields. Although the mathematical foundations are still valid in 3D, the visualization of the resulting intervals is nonetheless difficult, because the presented approaches cannot be transferred straightforwardly to 3D. For example, the color diffusion technique cannot be embedded directly into a volume rendering approach, since, due to blending and occlusions, the band pattern cannot be perceived anymore. The glyph-based approach for visualizing the orientation uncertainty could introduce several problems in 3D: 3D glyphs have a spatial extent and occlude each other, limiting the number of glyphs that can be shown concurrently. This problem could be solved by showing glyphs only at certain spatial regions, but an interactive guidance functionality would then be needed to emphasize the interesting regions. Furthermore, the entire probability distribution encoded in one glyph cannot be seen from a specific viewpoint, requiring a view-independent mapping of the entire spherical distribution.

6 Conclusion and Future Work

This thesis introduced a number of techniques for the visual analysis of the positional and structural variability of features like iso-contours and iso-surfaces in multi-dimensional uncertain data sets. A novel approach has been presented for computing reliable probabilities of position and occurrence for iso-surface crossings in uncertain Gaussian distributed 3D scalar fields. An efficient update-scheme allows integrating the proposed algorithm into front-to-back ray-casting. Furthermore, it has been shown how to derive probability distributions for iso-contours in scalar 2D ensemble fields, and how to visually convey this information to allow an improved uncertainty analysis. Instead of assuming a parametric Gaussian uncertainty model, non-parametric spatial distributions have been derived, by computing statistical summaries and generating continuous distribution functions thereof. In addition, various techniques have been proposed for analyzing the structural variability of features in uncertain data sets, by incorporating information on correlation data or varying gradients into the uncertainty visualization approach.

Apart from diverse novel visual effects, this thesis focused primarily on the development of sophisticated mathematical models and algorithms for deriving specific characteristic parameters or indicators from uncertain scalar data sets. The proposed derivations make two prominent contributions: First, they allow a more reliable computation of uncertainty measures and an improved interpretation of derived uncertainty visualizations. Several introduced techniques significantly improve traditional approaches on visualizing uncertain data. For instance, the integration of a correlation-based uncertainty model allows the computation of iso-surface crossing probabilities independently of the sampling rate in a 3D ray-casting approach. Secondly, the introduction of certain mathematical concepts allows a significant increase in efficiency of the respective uncertainty visualization approaches. For example, the development of a sophisticated front-to-back probability update scheme based on a distance dependent correlation model allows interactive frame rates during the ray-casting process in uncertainty-aware iso-surface extraction. Compared to traditional methods, no costly pre-computations are necessary and uncertainty measures can be derived and further processed in real-time.

The examination of visual methods for analyzing stochastic dependences, correlation structures and the underlying structural uncertainty can be regarded as one of the main contributions of this thesis. For the first time, the importance of integrating information on correlation or stochastic dependence structures in visualization techniques for uncertain scalar data sets has been introduced to the uncertainty visualization community. Furthermore, several specific algorithms have been proposed for visualizing local and global correlation data, as well as for revealing the variability of higher-order structural indicators like gradients in multi-dimensional scalar data sets. Apart from increasing the reliability and efficiency of traditional uncertainty visualization approaches, the integration of information on stochastic dependences and relative uncertainties can help gain additional insight in uncertain data. This is often not possible with traditional approaches using only simple indicators like mean and standard deviation values.

The advantages of the proposed novel techniques have been demonstrated for several synthetic and real-world data sets. It has been shown how they can be utilized for analyzing structural and topological variations in geophysical scenarios. Furthermore, the proposed approaches can help analyze the reliability of the occurrence of prominent material structures in seismic tomography data sets. In addition, it has been demonstrated how the novel techniques can be used for assessing the reliability of spatial temperature structures in uncertain atmospheric forecast ensemble data sets.

Although the presented novel approaches address several prominent challenges in uncertainty visualization, they can still serve as a basis for future research in various directions. In the following, several areas of possible future work are pointed out:

- **Probability Distributions.** Several uncertainty visualization techniques for scalar data sets assume a Gaussian distribution characterizing the underlying uncertainty. This thesis introduced several algorithms for the feature analysis in both Gaussian and arbitrary distributed data sets. Future work could address the challenge of extending these approaches towards a framework where the distribution assumption can be interactively specified by the user, resulting in different uncertainty visualizations. A respective flexible approach would allow to interactively explore the impact of the probability distribution assumption on the resulting visualization. Furthermore, future work could focus on the direct integration of distribution probability values into the resulting visualization, instead of displaying less accurate and more abstract uncertainty indicators like standard deviations and confidence intervals.
- **Integrated Visualization.** One critical challenge in uncertainty visualization is the combination of different numerical entities like mean data, uncertainty indicators, correlation information, etc. in one integrated visualization. This is complicated, as usually only a very limited number of visual channels (e.g., color, opacity, glyphs, etc.) are available. This thesis presented separate approaches for different uncertainty visualization directions like iso-surface variability or correlation structures. For instance, future work could address the challenge

of combining a probabilistic iso-surface extraction method with the display of global or local correlation structures in 3D. This could significantly improve the overall interpretation of the uncertainty information.

- **Dimensionality.** The global correlation clustering algorithm, as well as the gradient uncertainty visualization both address only 2D scalar data sets. Future work could extend such algorithms to 3D data sets. This is in particular challenging as, for instance, highly optimized algorithms for large amounts of data (correlation clustering) have to be found or strong cluttering effects due to the use of 3D glyphs (gradient uncertainty) have to be approached. Nonetheless, the visualization of global correlation structures or gradients in uncertain 3D scalar data sets would be highly beneficial for the reliability analysis of prominent 3-dimensional features.

- **Feature Analysis.** This thesis presented several approaches for analyzing the variability of iso-contours and iso-surfaces in 2D and 3D scalar data sets. The next step would be an extension of the used mathematical concepts for examining the variability of other scalar data features like critical points, local and global maximums/minimums, saddle points, etc. with respect to their occurrence and spatial position. Therefore, further mathematical models would be necessary for describing the uncertainty of higher-order entities like gradients, Hessian data, curvature, etc. The integration of uncertainty information on differential geometrical quantities would allow an improved analysis of the topological variability in uncertain scalar data sets.

- **Stochastic Dependence.** This thesis stressed the importance of integrating information on stochastic dependences into uncertainty visualization algorithms. Therefore, correlation has been used as the main structural uncertainty indicator. However, this is only possible for Gaussian distributed data sets with linear dependences between random variables at different spatial domain points. For arbitrary distributions with even non-linear dependences, correlation cannot be used as a reliable indicator for structural uncertainty. For extending the analysis of structural uncertainty to arbitrary distributed data sets, novel mathematical models would be necessary for deriving, representing and visualizing non-linear dependences in uncertain scalar data sets.

- **Statistical Analysis.** Uncertain data is often given as ensembles of data sets. Uncertainty indicators are derived from such data sets and transformed to specific visual effects. However, most visualization approaches ignore the number of ensemble members and therefore the reliability of the derived uncertainty measures. For instance, in this context, it could be important to know if an uncertainty indicator (e.g., standard deviation) was derived from 1000 or only 10 ensemble members. This information could be beneficial for the interpretation stage and is directly linked to the field of statistical analysis. Furthermore, visual hypothesis tests are an interesting research direction, as they could help answer specific questions about the

data for a pre-defined confidence level. Such statistical tests are typical means for analyzing ensemble data sets.

- **Sensitivity Analysis.** In the introduction of this thesis, it was stated that uncertain data sets are often caused by uncertain input parameters used in a simulation process. For exploring the impact of input on output uncertainty of a simulation process, an analysis would be beneficial, clearly stating which parts of the resulting data set and the respective degree of uncertainty are affected by which input parameters. In this context, a comprehensive sensitivity analysis could help identify the cause of uncertainties. The integration of sensitivity information into traditional uncertainty visualization approaches could allow the analysis of the degree of uncertainty, its impact on the data, as well as its origin in one integrated (interactive) visualization. This could significantly improve the interpretation of the uncertain data in the context of the respective generation process and its dependency on initial conditions or parameters.

Although uncertain data sets can be analyzed using different visual techniques based on various motivations and mathematical concepts, an intuitive and even interactive visualization approach should always be the ultimate goal in order to allow an efficient, reliable and comprehensive interpretation of the data by the user. This is not only important for a better understanding of a physical problem or process, but can have a significant impact on further decision processes.

Bibliography

[ABT12] Aster R. C., Borchers B., Thurber C. H.: *Parameter estimation and inverse problems*. Academic Press, 2012.

[ACN08] Ailon N., Charikar M., Newman A.: Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM) 55*, 5 (2008), 23.

[AH07] Arriola L. M., Hyman J. M.: Being sensitive to uncertainty. *Computing in Science & Engineering 9*, 2 (2007), 10–20.

[AOB08] Allendes Osorio R., Brodlie K. W.: Contouring with uncertainty. *Theory and Practice of Computer Graphics 2008. Proceedings.* (2008), 59–66.

[AS64] Abramowitz M., Stegun I.: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover publications, 1964.

[AWH*12] Alabi O. S., Wu X., Harter J. M., Phadke M., Pinto L., Petersen H., Bass S., Keifer M., Zhong S., Healey C., et al.: Comparative visualization of ensembles using ensemble surface slicing. In *IS&T/SPIE Electronic Imaging* (2012), International Society for Optics and Photonics, pp. 82940U–82940U.

[Ayy06] Ayyub B. M.: *Uncertainty modeling and analysis in engineering and the sciences*. CRC Press, 2006.

[BAF08] Bostrom A., Anselin L., Farris J.: Visualizing seismic risk and uncertainty. *Annals of the New York Academy of Sciences 1128*, 1 (2008), 29–40.

[BBC04] Bansal N., Blum A., Chawla S.: Correlation clustering. *Machine Learning 56*, 1 (2004), 89–113.

[BCN*05] Banks J., Carson J. S., Nelson B. L., Nicol D. M., et al.: *Discrete-event system simulation*. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[BFMW12] Bürger K., Fraedrich R., Merhof D., Westermann R.: Instant visitation maps for interactive visualization of uncertain particle trajectories. In *IS&T/SPIE Electronic Imaging* (2012), International Society for Optics and Photonics, pp. 82940P–82940P.

[BFW02] Bastin L., Fisher P. F., Wood J.: Visualizing uncertainty in multi-spectral remotely sensed imagery. *Computers & Geosciences 28*, 3 (2002), 337–350.

[Bjo96] Bjorck A.: *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics, 1996.

[BKKZ04] Böhm C., Kailing K., Kröger P., Zimek A.: Computing clusters of correlation connected objects. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (2004), ACM, pp. 455–466.

[BKS04] Bordoloi U., Kao D., Shen H.: Visualization techniques for spatial probability density function data. *Data Science Journal 3*, 0 (2004), 153–162.

[BPC*10] Borgo R., Proctor K., Chen M., Janicke H., Murray T., Thornton I. M.: Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 963–972.

[BPFG11] Berger W., Piringer H., Filzmoser P., Gröller E.: Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. *Computer Graphics Forum 30*, 3 (2011), 911–920.

[Bro04] Brown R.: Animated visual vibrations as an uncertainty visualisation technique. In *Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia* (2004), ACM, pp. 84–89.

[BWE05] Botchen R. P., Weiskopf D., Ertl T.: Texture-based visualization of uncertainty in flow fields. In *Visualization, 2005. VIS 05. IEEE* (2005), IEEE, pp. 647–654.

[CBDT11] Coninx A., Bonneau G.-P., Droulez J., Thibault G.: Visualization of uncertain scalar data fields using color scales and perceptually adapted noise. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization* (2011), ACM, pp. 59–66.

[CCM09] Correa C. D., Chan Y.-H., Ma K.-L.: A framework for uncertainty-aware visual analytics. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), IEEE, pp. 51–58.

[CR00] Cedilnik A., Rheingans P.: Procedural annotation of uncertain information. In *Visualization 2000. Proceedings* (2000), IEEE, pp. 77–84.

[CWMW11] Chen C., Wang C., Ma K., Wittenberg A.: Static correlation visualization for large time-varying volume data. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE* (2011), IEEE, pp. 27–34.

[DK97] Davis T. J., Keller C. P.: Modelling and visualizing multiple spatial uncertainties. *Computers & Geosciences 23*, 4 (1997), 397–408.

[DKLP02] Djurcilov S., Kim K., Lermusiaux P., Pang A.: Visualizing scalar volumetric data with uncertainty. *Computers & Graphics 26*, 2 (2002), 239–248.

[DMJRM00] De Maesschalck R., Jouan-Rimbaud D., Massart D. L.: The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems 50*, 1 (2000), 1–18.

[DW90] Drezner Z., Wesolowsky G.: On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation 35*, 1 (1990), 101–107.

[ESG97] Ehlschlaeger C. R., Shortridge A. M., Goodchild M. F.: Visualizing spatial data uncertainty using animation. *Computers & Geosciences 23*, 4 (1997), 387–395.

[Eur] European Centre for Medium-Range Weather Forecasts: Demeter temperature ensemble. Available online at `http://www.ecmwf.int`.

[Fel08] Feller W.: *An introduction to probability theory and its applications*, vol. 2. Wiley-India, 2008.

[FKLT10] Feng D., Kwock L., Lee Y., Taylor R. M.: Matching visual saliency to confidence in plots of uncertain data. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 980–989.

[FKLTI10] Feng D., Kwock L., Lee Y., Taylor II R. M.: Linked exploratory visualizations for uncertain mr spectroscopy data. In *IS&T/SPIE Electronic Imaging* (2010), International Society for Optics and Photonics, pp. 753004–753004.

[FWA*11] Flath H. P., Wilcox L. C., Akçelik V., Hill J., van Bloemen Waanders B., Ghattas O.: Fast algorithms for bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations. *SIAM Journal on Scientific Computing 33*, 1 (2011), 407–432.

[GCN*09] Gallagher K., Charvin K., Nielsen S., Sambridge M., Stephenson J.: Markov chain monte carlo (mcmc) sampling methods to determine optimal models, model resolution and model choice for earth science problems. *Marine and Petroleum Geology 26*, 4 (2009), 525–535.

[Geo08] Georgii H.: *Stochastics: introduction to probability and statistics.* Walter De Gruyter, 2008.

[Ger] German Research Council (SFB 382): Fuel injection dataset. Available online at `http://www.volvis.org`.

[Ger98] Gershon N.: Visualization of an imperfect world. *Computer Graphics and Applications, IEEE 18*, 4 (1998), 43–45.

[GFWG10] Galbally D., Fidkowski K., Willcox K., Ghattas O.: Non-linear model reduction for uncertainty quantification in large-scale inverse problems. *International journal for numerical methods in engineering 81*, 12 (2010), 1581–1608.

[GR04] Grigoryan G., Rheingans P.: Point-based probabilistic surfaces to show surface uncertainty. *Visualization and Computer Graphics, IEEE Transactions on 10*, 5 (2004), 564–573.

[Gra96] Graves R. W.: Simulating seismic wave propagation in 3d elastic media using staggered-grid finite differences. *Bulletin of the Seismological Society of America 86*, 4 (1996), 1091–1106.

[GS05] Griethe H., Schumann H.: Visualizing uncertainty for improved decision making. In *Proceedings of the 4th International Conference on Business Informatics Research BIR 2005* (2005).

[GS06] Griethe H., Schumann H.: The visualization of uncertain data: Methods and problems. In *Proceedings of SimVis* (2006), vol. 6, pp. 143–156.

[Ham62] Hamming R.: *Numerical methods for scientists and engineers.* McGraw-Hill (New York), 1962.

[Hen03] Hengl T.: Visualisation of uncertainty using the hsi colour model: computations with colours. In *Proceedings of the 7th International Conference on GeoComputation* (2003), pp. 8–17.

[HG93] Hunter G. J., Goodchild M.: Managing uncertainty in spatial databases: Putting theory into practice. In *Papers from the annual conference-urban and regional information systems association* (1993), Urisa Urban and Regional Information Systems, pp. 15–15.

[HLNW11] Hlawatsch M., Leube P., Nowak W., Weiskopf D.: Flow radar glyphsâ€"static visualization of unsteady flow with uncertainty. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 1949–1958.

[HM96] Howard D., MacEachren A. M.: Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Systems 23*, 2 (1996), 59–77.

[Hol11] Holy T.:. `http://www.mathworks.com/matlabcentral/fileexchange/29702`, 2011.

[HT06] Hengl T., Toomanian N.: Maps are not what they seem: representing uncertainty in soil-property maps. In *Proc. Accuracy* (2006), pp. 805–813.

[Joh04] Johnson C.: Top scientific visualization research problems. *Computer graphics and applications, IEEE 24*, 4 (2004), 13–17.

[Jon03] Jones D.: Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor mri. *Magnetic Resonance in Medicine 49*, 1 (2003), 7–12.

[JPR*04] Jen D., Parente P., Robbins J., Weigle C., Taylor R., Burette A., Weinberg R., et al.: Imagesurfer: A tool for visualizing correlations between two volume scalar fields. In *Visualization, 2004. IEEE* (2004), IEEE, pp. 529–536.

[JS03] Johnson C. R., Sanderson A. R.: A next step: Visualizing errors and uncertainty. *Computer Graphics and Applications, IEEE 23*, 5 (2003), 6–10.

[KÏ0] Käufl P.: *Object-based Probabilistic Full Waveform Tomography.* Master's thesis, Geophysics, LMU München, 2010.

[KKZ09] Kriegel H., Kröger P., Zimek A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD) 3*, 1 (2009), 1.

[KLDP02] Kao D., Luo A., Dungan J. L., Pang A.: Visualizing spatially varying distribution data. In *Information Visualisation, 2002. Proceedings. Sixth International Conference on* (2002), IEEE, pp. 219–225.

[KMBR56] Kolmogorov A. N., Morrison N., Bharucha-Reid A. T.: *Foundations of the Theory of Probability.* Chelsea New York, 1956.

[KRT02] Komatitsch D., Ritsema J., Tromp J.: The spectral-element method, beowulf computing, and global seismology. *Science 298*, 5599 (2002), 1737–1742.

[KT99] Komatitsch D., Tromp J.: Introduction to the spectral element method for three-dimensional seismic wave propagation. *Geophysical Journal International 139*, 3 (1999), 806–822.

[KVUS*05] Kniss J. M., Van Uitert R., Stephens A., Li G.-S., Tasdizen T., Hansen C.: Statistically quantitative volume visualization. In *Visualization, 2005. VIS 05. IEEE* (2005), IEEE, pp. 287–294.

[KWL*04] Kindlmann G. L., Weinstein D. M., Lee A. D., Toga A. W., Thompson P. M.: Visualization of anatomic covariance tensor fields. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE* (2004), vol. 1, IEEE, pp. 1842–1845.

[KWTM03] Kindlmann G., Whitaker R., Tasdizen T., Moller T.: Curvature-based transfer functions for direct volume rendering: Methods and applications. In *Visualization, 2003. VIS 2003. IEEE* (2003), IEEE, pp. 513–520.

[Kyr03] Kyriakidis P. C.: Towards a systems approach to the visualization of spatial uncertainty. In *UCGIS Workshop: Geospatial Visualization and Knowledge Discovery Workshop* (2003).

[LFLH07] Li H., Fu C.-W., Li Y., Hanson A. J.: Visualizing large-scale uncertainty in astrophysical data. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1640–1647.

[LKP03] Luo A., Kao D., Pang A.: Visualizing spatial distribution data sets. In *Proceedings of the symposium on Data visualisation 2003* (2003), Eurographics Association, pp. 29–38.

[LLPY07] Lundstrom C., Ljung P., Persson A., Ynnerman A.: Uncertainty visualization in medical volume rendering using probabilistic animation. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1648–1655.

[LPK05] Love A. L., Pang A., Kao D. L.: Visualizing spatial multivalue data. *Computer Graphics and Applications, IEEE 25*, 3 (2005), 69–79.

[LRN88] Lee Rodgers J., Nicewander W. A.: Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 1 (1988), 59–66.

[Mah36] Mahalanobis P. C.: On the generalized distance in statistics. *Proceedings of the national institute of sciences of India 2*, 1 (1936), 49–55.

[MCD*12] Mosca I., Cobden L., Deuss A., Ritsema J., Trampert J.: Seismic and mineralogical structures of the lower mantle from probabilistic tomography. *Journal of Geophysical Research 117*, B6 (2012), B06304.

[Mor09] Moreland K.: Diverging color maps for scientific visualization. In *Advances in Visual Computing*. Springer, 2009, pp. 92–103.

[MRH*05] MacEachren A. M., Robinson A., Hopper S., Gardner S., Murray R., Gahegan M., Hetzler E.: Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science 32*, 3 (2005), 139–160.

[MS02] Mosegaard K., Sambridge M.: Monte carlo analysis of inverse problems. *Inverse Problems 18*, 3 (2002), R29.

[MT02] Mosegaard K., Tarantola A.: 16 probabilistic approach to inverse problems. *International Geophysics 81* (2002), 237–265.

[NO07] Narita M., Ohmori S.: Improving precipitation forecasts by the operational non-hydrostatic mesoscale model with the kain–fritsch convective parameterization and cloud microphysics. In *Preprints, 12th Conf. on Mesoscale Processes, Watervillle Valley, NH, Amer. Meteor. Soc* (2007), vol. 3.

[Nol08] Nolet G.: *A Breviary of Seismic Tomography: Imaging the Interior of the Earth and Sun*. Cambridge University Press, 2008.

[OB09] Osorio R. A., Brodlie K. W.: Uncertain flow visualization using lic. *7th EG UK Theory and Practice of Computer Graphics* (2009), 1–9.

[OGHT10] Otto M., Germer T., Hege H.-C., Theisel H.: Uncertain 2d vector field topology. *Computer Graphics Forum 29*, 2 (2010), 347–356.

[OGT11a] Otto M., Germer T., Theisel H.: Closed stream lines in uncertain vector fields. *Proc. Spring Conference on Computer Graphics (SCCG) 2*, 9 (2011).

[OGT11b] Otto M., Germer T., Theisel H.: Uncertain topology of 3d vector fields. *Pacific Visualization Symposium (PacificVis), 2011 IEEE* (2011), 67–74.

[OT12] Otto M., Theisel H.: Vortex analysis in uncertain vector fields. *Computer Graphics Forum 31*, 3pt2 (2012), 1035–1044.

[Pan08] Pang A.: Visualizing uncertainty in natural hazards. In *Risk Assessment, Modeling and Decision Support*. Springer, 2008, pp. 261–294.

[Par62] Parzen E.: On estimation of a probability density function and mode. *The annals of mathematical statistics 33*, 3 (1962), 1065–1076.

[PdJB07] Pebesma E. J., de Jong K., Briggs D.: Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. *International Journal of Geographical Information Science 21*, 5 (2007), 515–527.

[PH11] Pothkow K., Hege H.-C.: Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *Visualization and Computer Graphics, IEEE Transactions on 17*, 10 (2011), 1393–1406.

[PKJ08] Potter K., Krüger J., Johnson C.: Towards the visualization of multidimensional stochastic distribution data. In *Proceedings of The International Conference on Computer Graphics and Visualization (IADIS)* (2008), vol. 2008.

[PKR07] Potter K., Kniss J., Riesenfeld R.: Visual summary statistics. *Univeristy of Utah, Tech. Rep. UUCS-07-004* (2007).

[PKRJ10] Potter K., Kniss J., Riesenfeld R., Johnson C. R.: Visualizing summary statistics and uncertainty. *Computer Graphics Forum 29*, 3 (2010), 823–832.

[PMW13] Pfaffelmoser T., Mihai M., Westermann R.: Visualizing the variability of gradients in uncertain 2d scalar fields. *IEEE Transactions on Computer Graphics and Visualization* (2013).

[PPA*12] Phadke M. N., Pinto L., Alabi O., Harter J., Taylor II R. M., Wu X., Petersen H., Bass S. A., Healey C. G.: Exploring ensemble visualization. In *IS&T/SPIE Electronic Imaging* (2012), International Society for Optics and Photonics, pp. 82940B–82940B.

[PPH12] Petz C., Pöthkow K., Hege H.: Probabilistic local features in uncertain vector fields with spatial correlation. *Computer Graphics Forum 31*, 3pt2 (2012), 1045–1054.

[PRH10] Prassni J.-S., Ropinski T., Hinrichs K.: Uncertainty-aware guided volume segmentation. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1358–1365.

[PRJ12] Potter K., Rosen P., Johnson C. R.: From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*. Springer, 2012, pp. 226–249.

[PRW11] Pfaffelmoser T., Reitinger M., Westermann R.: Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. *Computer Graphics Forum 30*, 3 (2011), 951–960.

[PW12] Pfaffelmoser T., Westermann R.: Visualization of global correlation structures in uncertain 2d scalar fields. *Computer Graphics Forum 31*, 3pt2 (2012), 1025–1034.

[PW13a] Pfaffelmoser T., Westermann R.: Correlation visualization for structural uncertainty analysis. *International Journal for Uncertainty Quantification 3*, 2 (2013).

[PW13b] Pfaffelmoser T., Westermann R.: Visualizing contour distributions in 2d ensemble data. In *EuroVis Short Papers* (2013), The Eurographics Association, pp. 55–59.

[PWB*09a] Potter K., Wilson A., Bremer P.-T., Williams D., Doutriaux C., Pascucci V., Johhson C.: Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated visus-cdat systems. *Journal of Physics: Conference Series 180*, 1 (2009), 012089.

[PWB*09b] Potter K., Wilson A., Bremer P.-T., Williams D., Doutriaux C., Pascucci V., Johnson C. R.: Ensemble-vis: A framework for the statistical visualization of ensemble data. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on* (2009), IEEE, pp. 233–240.

[PWL97] Pang A. T., Wittenbrink C. M., Lodha S. K.: Approaches to uncertainty visualization. *The Visual Computer 13*, 8 (1997), 370–390.

[RLBS03] Rhodes P. J., Laramee R. S., Bergeron R. D., Sparr T. M.: Uncertainty visualization methods in isosurface rendering. In *Eurographics* (2003), vol. 2003, pp. 83–88.

[Ros56] Rosenblatt M.: Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* (1956), 832–837.

[SBSS06] Sambridge M., Beghein C., Simons F. J., Snieder R.: How do we understand and visualize uncertainty? *The Leading Edge 25*, 5 (2006), 542–546.

[SCB*04] Schmidt G. S., Chen S.-L., Bryden A. N., Livingston M. A., Rosenblum L. J., Osborn B. R.: Multidimensional visual representations for underwater environmental uncertainty. *Computer Graphics and Applications, IEEE 24*, 5 (2004), 56–65.

[Sil86] Silverman B. W.: *Density estimation for statistics and data analysis*, vol. 26. Chapman & Hall/CRC, 1986.

[SJK04] Sanderson A., Johnson C., Kirby R.: Display of vector fields using a reaction-diffusion model. In *Visualization, 2004. IEEE* (2004), IEEE, pp. 115–122.

[Ste07] Stensrud D. J.: *Parameterization schemes: keys to understanding numerical weather prediction models.* Cambridge University Press, 2007.

[STS06] Sauber N., Theisel H., Seidel H.-P.: Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *Visualization and Computer Graphics, IEEE Transactions on 12*, 5 (2006), 917–924.

[SWMW09] Sukharev J., Wang C., Ma K.-L., Wittenberg A. T.: Correlation study of time-varying multivariate climate data sets. In *Visualization Symposium, 2009. PacificVis' 09. IEEE Pacific* (2009), IEEE, pp. 161–168.

[SZD*10] Sanyal J., Zhang S., Dyer J., Mercer A., Amburn P., Moorhead R.: Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1421–1430.

[Tar05] Tarantola A.: *Inverse problem theory and methods for model parameter estimation*. siam, 2005.

[THM*05] Thomson J., Hetzler E., MacEachren A., Gahegan M., Pavel M.: A typology for visualizing uncertainty. In *Proc. SPIE* (2005), vol. 5669, pp. 146–157.

[TK94] Taylor B., Kuyatt C.: Nist technical note 1279: Guidelines for evaluating and expressing the uncertainty of nist measurements results. *National Institute of Standards and Technology, Washington DC* (1994).

[Wei74] Weir A. J.: *General integration & measure*, vol. 2. Cambridge University Press, 1974.

[Wei05] Weiler M.: *Hardware-beschleunigte Volumenvisualisierung auf adaptiven Datenstrukturen*. PhD thesis, Universitätsbibliothek der Universität Stuttgart, 2005.

[WFR*10] Waser J., Fuchs R., Ribicic H., Schindler B., Bloschl G., Groller E.: World lines. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1458–1467.

[WPL96] Wittenbrink C. M., Pang A. T., Lodha S. K.: Glyphs for visualizing uncertainty in vector fields. *Visualization and Computer Graphics, IEEE Transactions on 2*, 3 (1996), 266–279.

[WRL12] Wellmann J. F., Regenauer-Lieb K.: Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models. *Tectonophysics 526* (2012), 207–216.

[XLS10] Xu L., Lee T.-Y., Shen H.-W.: An information-theoretic framework for flow visualization. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1216–1224.

[YXK13] Yang C., Xiu D., Kirby R. M.: Visualization of covariance and cross-covariance fields. *International Journal for Uncertainty Quantification 3*, 1 (2013), 25–38.

[ZC06] Zuk T., Carpendale S.: Theoretical analysis of uncertainty visualizations. In *Electronic Imaging 2006* (2006), International Society for Optics and Photonics, pp. 606007–606007.

[ZDG*08] Zukab T., Downtonb J., Grayb D., Carpendalea S., Liangb J.: Exploration of uncertainty in bidirectional vector fields. *SPIE 6809* (2008).

[Zim08] Zimek A.: *Correlation Clustering*. PhD thesis, LMU München, 2008.

[ZT07] Zhang H., Thurber C.: Estimating the model resolution matrix for large seismic tomography problems based on lanczos bidiagonalization with partial reorthogonalization. *Geophysical Journal International 170*, 1 (2007), 337–345.

[ZWK10] Zehner B., Watanabe N., Kolditz O.: Visualization of gridded scalar data with uncertainty in geosciences. *Computers & Geosciences 36*, 10 (2010), 1268–1275.