Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

Visualization of Global Correlation Structures in Uncertain 2D Scalar Fields

Tobias Pfaffelmoser and Rüdiger Westermann

Computer Graphics and Visualization Group, Technische Universität München, Germany

Abstract

Visualizing correlations, i.e., the tendency of uncertain data values at different spatial positions to change contrarily or according to each other, allows inferring on the possible variations of structures in the data. Visualizing global correlation structures, however, is extremely challenging, since it is not clear how the visualization of complicated long-range dependencies can be integrated into standard visualizations of spatial data. Furthermore, storing correlation information imposes a memory requirement that is quadratic in the number of spatial sample positions. This paper presents a novel approach for visualizing both positive and inverse global correlation structures in uncertain 2D scalar fields, where the uncertainty is modeled via a multivariate Gaussian distribution. We introduce a new measure for the degree of dependency of a random variable on its local and global surroundings, and we propose a spatial clustering approach based on this measure to classify regions of a particular correlation strength. The clustering performs a correlation filtering, which results in a representation that is only linear in the number of spatial sample points. Via cluster coloring the correlation information can be embedded into visualizations of other statistical quantities, such as the mean and the standard deviation. We finally propose a hierarchical cluster subdivision scheme to further allow for the simultaneous visualization of local and global correlations.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms, Viewing algorithms

1. Introduction

In a discrete spatial scalar field, the uncertainty can be modeled by a *multivariate random variable* $\mathbf{Y}$ with scalar-valued components $Y(\mathbf{x}_i)$, where each component describes the uncertainty at the point $\mathbf{x}_i$. In the following we assume that the random variables exhibit a *multivariate Gaussian distribution*, so that the uncertainty at a point $\mathbf{x}_i$ is given 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 [JS03, PWL97].

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 dependency between these values. We denote by *structural variability* the property of a structure to vary in shape due to uncertainty.

For instance, let us consider an *uncertain* 2D height field, where the variability of height values is modeled via Gaussian distributed random variables. For two such variables X and Y the mutual *stochastic dependency* 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 $\mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$, where $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ denote the *variances* of X and Y, while $\mathrm{Cov}(X,Y)$ 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 prob-

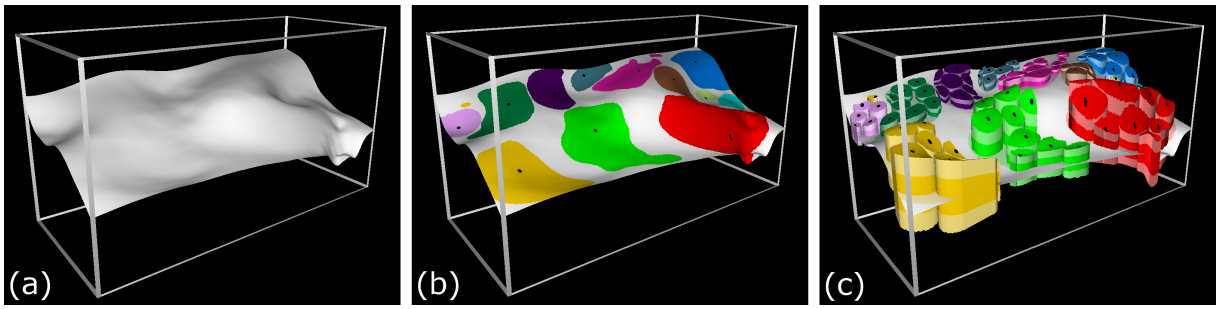

Figure 1: *(a) Mean surface in an ensemble temperature field over a 2D domain. (b) Disjoint clusters contain surface points where the uncertainty has a correlation higher than* $\rho_1 = 0.4$ *to the uncertainty at the cluster centroids (black dots). (c) Clusters are subdivided using* $\rho_2 = 0.9$ *and extruded along the third dimension according to the standard deviation at the member points.*

ability 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 our investigations.

Our Contribution. We introduce a new approach for visualizing positive and inverse correlation structures in uncertain 2D scalar fields, where the uncertainty is modeled via multivariate Gaussian distributions. Our 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, we can map the correlation information 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, we propose a novel approach for filtering the correlation information. It seeks for the most prominent spatial correlation structures and represents them as individual clusters. Therefore, we introduce *correlation neighborhoods* and use their size as a new measure for the degree of dependency 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, we further propose a subdivision scheme that breaks clusters indicating long-range dependencies 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 our visualization approach is linear in the number of spatial sample 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. For instance, Fig. 1 (a) shows a constant temperature surface above Europe. The surface visualizes the mean values in an ensemble of data sets that were simulated by the European Center for Medium-Range Weather Forecast (ECMWF). In (b), color-coded correlation clusters are visualized on the same surface. A visualization showing the subdivision of clusters into subclusters of ever higher correlation strength and the extrusion of clusters according to standard deviation is shown in (c).

The remainder of this paper is as follows: Firstly, we discuss work related to ours. Our proposed algorithm for clustering positive correlations is outlined next. This is followed by a description of the modifications that are required to allow for the construction of inverse correlation clusters. We then emphasize the combined visualization of correlation clusters and standard deviation, and we discuss the efficient rendering of the given information. We conclude the paper with an analysis of the proposed correlation visualization techniques, including some remarks on limitations and future work.

2. Related Work

The importance of uncertainty visualization has been recognized over more than a decade ago [PWL97], yet the visual indication of uncertainties in scientific data sets is still far from standard. Efforts in this area have been mainly restricted to particular domains such as geographical information systems [MRH*05], seismology [BAF08], and astrophysics [LFLH07], to name just a few. An overview and taxonomy of uncertainty visualization techniques is given in [JS03, THM*05, GS06]. The web-library at [Pot] provides a list of references to the major publications in the field.

One approach for representing uncertainty is to provide it as secondary data that is visualized in addition to the primary

data. Here, the standard deviation from a given mean value is often visualized directly via specific color and opacity mappings, animations, textures, or glyphs [WPL02, DKLP02, RLBS03, LLPY07, SZD*10]. Such approaches can provide a good indication of the local uncertainty strength.

Alternative approaches visually encode the positional variation that is caused by the uncertainty on specific features, for instance, the positional variability of surfaces in space. Techniques include the visualization of confidence surfaces [PWL97, ZWK10] and flowlines [KWTM03], surface diffusion techniques [GR04], as well as surface animations [Bro04]. The most recent approaches [PRW11, PWH11, PH10] model the uncertainty stochastically and derive probability distributions for particular stochastic events associated to isosurfaces. To the best of our knowledge, however, none of these techniques allows inferring on the possible structural surface variability.

Only 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 multifield graphs. Glyph-based visualization of local covariance structures was presented in [KWL*]. In [SWMW09], correlations in time-varying data have been investigated, and [YXK11] 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 by Bansal et al. [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].

3. Positive Correlation Clustering

In the following we assume an uncertain 2D scalar field. 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, we filter the correlation information in a pre-process so as to keep the most relevant correlation structures, but significantly reduce the memory requirements. Therefore, we first introduce a correlation-based importance measure, which we then use in a novel clustering algorithm.

3.1. Correlation Strength Model

For each vertex $\mathbf{x}_i$, we compute 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. We call

$$\eta_{\rho_1}(\mathbf{x}_i) := \{\mathbf{x}_j \in \mathbb{C} \mid \rho(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) \geq \rho_1\} \qquad (1)$$

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 dependency between the random variable $Y(\mathbf{x}_i)$ and its local and global spatial surroundings, as it counts the most prominent "correlation partners" of $\mathbf{x}_i$, independent of their position in $\mathbb{C}$.

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)

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

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 [PRW11], we propose using the size of the correlation neighborhood for a given threshold, i.e., the cardinal number, to measure the local correlation strength of each random variable.

3.2. 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 cardinal number. 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

$$\Psi_{\rho_1} := \{\mathbf{c}_0, \mathbf{c}_1, ...\} \subset \mathbb{C}, \quad (3)$$

which consists of the selected vertices $\mathbf{c}_i$ in descending order of cardinal number. 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 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. 2 (a) shows the clusters for $\rho_1 = 0.3$ in the ECMWF data set. The pink clusters show long-range bridging correlations.

Our 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 our approach, while a randomized algorithm might split up this cluster into multiple ones.

Since our 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 (cf. Fig. 8). Furthermore, in our 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 complete partitioning of the domain cannot guarantee this and lets the cluster sizes be dependent on the selection order rather than the correlation strength.

3.3. Multilevel Clustering

To enable the user to interactively analyze clusters at different correlation levels, multiple sets of clusters are precomputed for different values of ρ_1. With increasing ρ_1, in regions with low correlation strength the clusters quickly shrink and the number of clusters increases. Where clusters remain spatially extended, they indicate strongly correlated regions. Fig. 2 (b) shows these effects for the initial clusters in (a) 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.

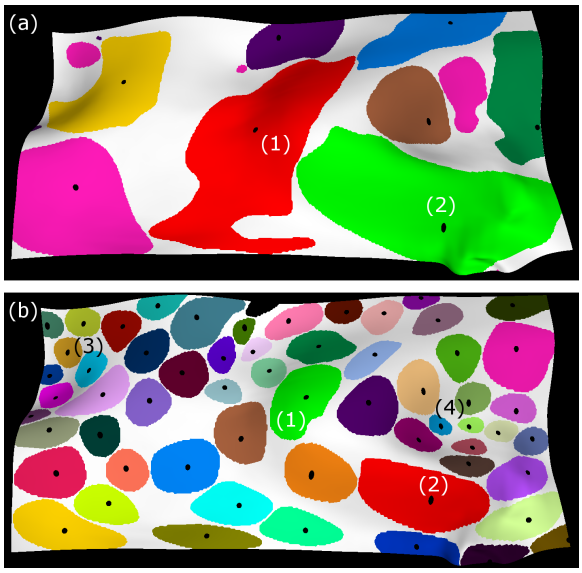

Figure 2: *(a) Positive correlation clusters for $\rho_1 = 0.3$ are visualized on the ECMWF mean surface. (b) Clustering for $\rho_1 = 0.7$ indicates strong local correlation in regions (1) and (2), and weak stochastic dependence in regions (3) and (4).*

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

3.4. Cluster Subdivision

To allow the visualization of local correlations within a global context, we introduce a subdivision scheme that splits 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

$$\Psi^{\mathbf{c}_i}_{\rho_1\rho_2} := \{\mathbf{c}_{i0}, \mathbf{c}_{i1}, ...\} \subset \eta_{\rho_1}(\mathbf{c}_i). \quad (4)$$

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

$$\eta_{\rho_1\rho_2}(\mathbf{c}_{ij}) := \{\mathbf{x}_k \in \eta_{\rho_1}(\mathbf{c}_i) \mid \rho(Y(\mathbf{c}_{ij}), Y(\mathbf{x}_k)) \geq \rho_2\} \quad (5)$$

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.

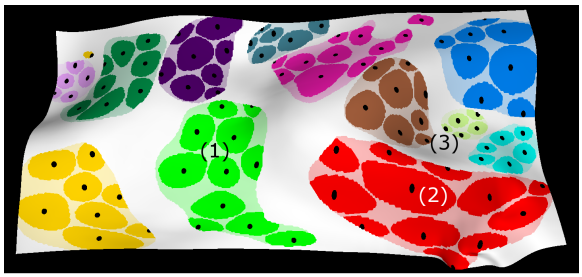

Figure 3: *Clusters for $\rho_1 = 0.4$ and corresponding sub-clusters for $\rho_2 = 0.9$ show isotropic and anisotropic correlation structures in (1) and (2), respectively. Severe cluster shrinkage indicates low correlation strength in (3).*

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. 3, were positive correlation clusters for $\rho_1 = 0.4$ (cf. Fig. 1 (b)) were subdivided using $\rho_2 = 0.9$. Compared to region (1), the sub-clusters 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.

3.5. Anisotropy Coloring

Our 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. 4 (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, we introduce a color mapping scheme that emphasizes inner-cluster anisotropy.

We first 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, we employ the GCF in Equ. (2), which controls the correlation strength by the parameter τ. We can use the known correlation and Euclidean distance in Equ. (2) to rearrange for the unknown τ:

$$\tau(\mathbf{x}_j) := -\frac{\log(\rho(Y(\mathbf{c}_i), Y(\mathbf{x}_j)))}{\left\|\mathbf{c}_i - \mathbf{x}_j\right\|^2}, \quad \mathbf{x}_j \in \eta_{\rho_1}(\mathbf{c}_i). \tag{6}$$

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, we introduce the *correlation half-value distance* (CHVD)

$$\delta(\tau) := \frac{\log(2)}{\tau}, \tag{7}$$

which indicates 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.

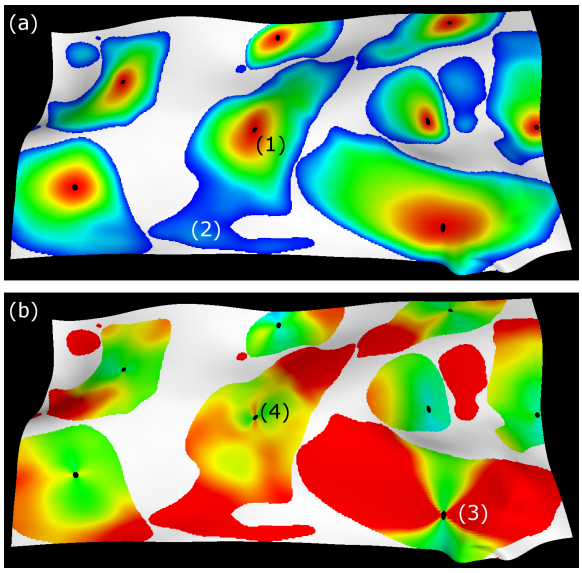

Figure 4: *(a) Correlation between cluster points and centers is mapped from $[\rho_1, 1]$ to $[blue, red]$. (b) CHVD is mapped from $[0, \delta_{max}]$ to $[blue, red]$. High correlation anisotropy around the center point is shown in (3). In (4), high anisotropy is only present close to the center point,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

In Fig. 4 (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.

4. Inverse Correlation Clustering

Linear inverse correlation between two random variables indicates that the realization of one 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 inversely correlated to each other.

As for positive correlation clustering, we define

$$\kappa_{\hat{\rho}}(\mathbf{x}_i) := \{\mathbf{x}_j \in \mathbb{C} \mid \rho(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) \leq \hat{\rho}\} \tag{8}$$

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

In inverse correlation clustering we seek for the most prominent pairs of centroids of inverse partners

$$\Phi_{\hat{\rho}} := \{(\mathbf{a}_0, \mathbf{b}_0), (\mathbf{a}_1, \mathbf{b}_1), ...\} \subset \mathbb{C} \times \mathbb{C}, \tag{9}$$

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

$$\mathbf{a}_0 := \operatorname*{arg\,max}_{\mathbf{x}_j \in \mathbb{C}} \left|\kappa_{\hat{\rho}}(\mathbf{x}_j)\right|, \tag{10}$$

$$\mathbf{b}_0 := \operatorname*{arg\,max}_{\mathbf{x}_j \in \kappa_{\hat{\rho}}(\mathbf{a}_0)} \left|\kappa_{\hat{\rho}}(\mathbf{x}_j)\right|, \tag{11}$$

$$\mathbf{a}_i := \operatorname*{arg\,max}_{\mathbf{x}_j \in \mathbb{C}, \kappa_{\hat{\rho}}(\mathbf{x}_j) \cap \bigcup_{k<i} (\kappa_{\hat{\rho}}(\mathbf{a}_k) \cup \kappa_{\hat{\rho}}(\mathbf{b}_k)) = \emptyset} \left|\kappa_{\hat{\rho}}(\mathbf{x}_j)\right|, \tag{12}$$

$$\mathbf{b}_i := \operatorname*{arg\,max}_{\mathbf{x}_j \in \kappa_{\hat{\rho}}(\mathbf{a}_i), \kappa_{\hat{\rho}}(\mathbf{x}_j) \cap \bigcup_{k<i} (\kappa_{\hat{\rho}}(\mathbf{a}_k) \cup \kappa_{\hat{\rho}}(\mathbf{b}_k)) = \emptyset} \left|\kappa_{\hat{\rho}}(\mathbf{x}_j)\right|. \tag{13}$$

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 inversely correlated to the other partner and vice versa. Clusters are computed for different negative levels $\hat{\rho}_i$.

4.1. Inverse Cluster Coloring

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). In this case, the visualization has to indicate that the sub-regions belonging to the same partner are inversely correlated to the respective other partner, but not to each other. This is achieved by hatching the clusters using different patterns. In particular, we use vertical and horizontal stripes 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, inverse clusters (in color) are visualized for $\hat{\rho} = -0.3$ in relation to the positive clusters (in grey) for $\rho_1 = 0.4$. Each color represents one inverse correlation pair, and the stripe orientation indicates the respective inverse partners within each pair.

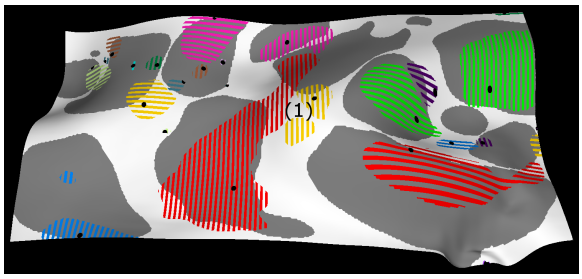

Figure 5: *Positive clusters are shown in gray, inverse clusters for $\hat{\rho} = -0.3$ are color coded. Clusters with the same color but different stripe orientation are inversely correlated.*

5. 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 we investigate uncertain scalar fields over a 2D domain and stochastically model the uncertainty via Gaussian distributions, *stochastic distance functions* (SDF) can be used to visualize the standard deviation. A SDF is defined as

$$\vartheta(v, \mathbf{x}_i) := \frac{v - \mu(\mathbf{x}_i)}{\max(\sigma(\mathbf{x}_i), \sigma_{\min})}. \tag{14}$$

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 we refer the reader to [PRW11]. 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. 6. In (a), positive correlation clusters for $\rho_1 = 0.5$, including subdivision for $\rho_2 = 0.9$, are extruded to the confidence level $\vartheta(v,\mathbf{x}_i) = 1$. The small towers in region (1) indicate significantly lower standard deviation compared to regions (2) and (3). In (b), inverse clusters for $\hat{\rho} = -0.5$ are visualized using the proposed approach. Compared to Fig. 5, some of the smaller clusters vanish, and only the most prominent clusters like the red, green, and pink clusters remain.

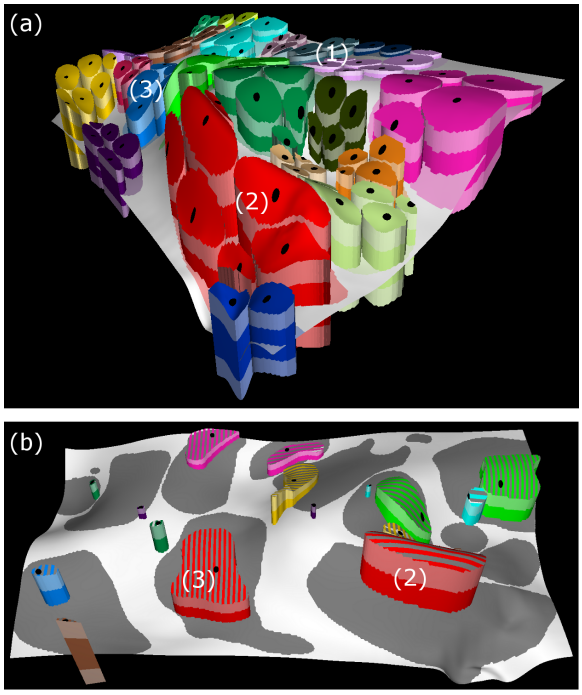

Figure 6: *(a) Positive, subdivided correlation clusters are shown for $\rho_1 = 0.5$ and $\rho_2 = 0.9$. Cluster extrusion until a selected SDF level reveals low positional variability in (1) and strong positional variabilities in (2) and (3). (b) Inverse extruded cluster partners for $\hat{\rho} = -0.5$ are shown.*

6. 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 took 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.

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 our 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.

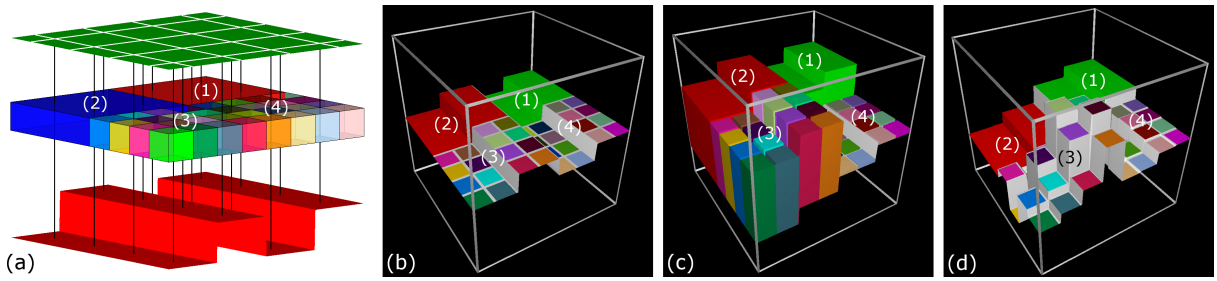

Figure 7: *(a) Geophysical setup to determine the depth of a material discontinuity (red) in the earth's interior by measuring travel times of artificial pressure waves (emitted along black lines). A material layer between emitters and discontinuity structure simulates a Gaussian error distribution in 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).*

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 isosurface 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.

7. Analysis and Discussion

We first validate the plausibility of the proposed correlation clustering approach using a simple artificial data set (see Fig. 7 (a)). We simulate a set of seismic pressure sources 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 our 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. 7 (b). It can be seen that the independent error structures shown in (a) are correctly grouped by our 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.

We now turn our attention to the ECMWF data set, and we use the proposed correlation visualization to conclude on the following peculiarities: Fig. 1 (a) and (b) show a long ridge-like 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. 2). Furthermore, cluster subdivision in Fig. 3 and anisotropy coloring in Fig. 4 (b) clearly show the alignment of the prominent correlation direction with the ridge orientation, meaning that the structural

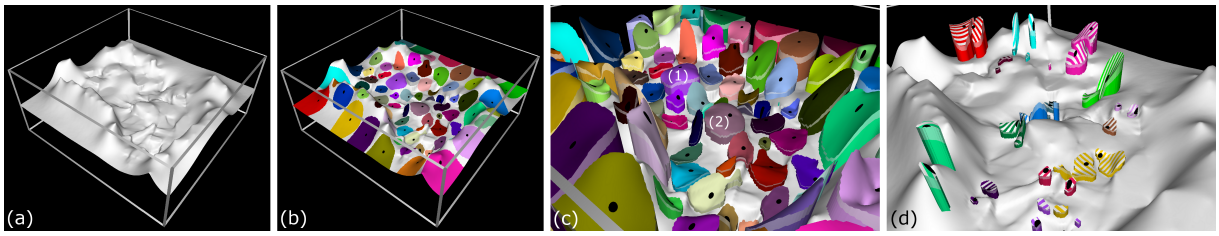

Figure 8: *(a) Mohorovičić discontinuity below Australia. (b) Strong correlation clusters close to the domain boundaries indicate strong regularizations in the simulation algorithm. (c) Close-up view reveals high and low uncertainties, respectively, at the boundaries and in the center, as well as high local correlations at (1) and (2). (d) Inverse clustering shows that strong inverse correlation takes place on a local rather than a global scale.*

variability is low along the ridge. In Fig. 2 (b), small clusters in region (4) indicate low correlation strength (see also region (3) in Fig. 3). 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 a very interesting long-range interaction between data values in the ECMWF data set. Fig. 6 (b) shows multiple inversely correlated cluster pairs. In combination with uncertainty towers, the red clusters turns out to be the most prominent. The visualization shows a strong stochastic dependency between spatial locations over long distances. For instance, a temperature decrease in region (2) is likely to cause a temperature increase in region (3), and vice versa.

In a third example we use correlation visualization to analyze the Mohorovičić discontinuity — the boundary surface between the Earth's crust and mantle — below Australia (see Fig. 8 (a)). The data was acquired using a similar geophysical setup as described in our first example. Positive correlation clusters in (b) show a rather homogeneous correlation distribution in the domain interior and high correlation strength 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 dependency 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 inversely 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 inversely 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 inversely correlated clusters are far more distant to each other and cover significantly larger regions.

8. Conclusion

Our contribution is a new approach for visualizing positive and inverse correlation structures in uncertain 2D scalar fields. We have built upon the concept of correlation neighborhoods and their cardinal numbers a novel correlation clustering algorithm. 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 dependencies in the data. The cluster representation which is build in a pre-process requires an amount of memory that is linear in the number of initial data points.

In the future we will strive to extend our approach to isosurfaces in 3D scalar fields. In general, correlation clustering as proposed can be extended to 3D, but, in this case it is not sufficient to consider only the correlations between surface points. In addition, the correlations in a 3D region enclosing the surface have to be analyzed, and, thus, more elaborate techniques are required to overcome the significantly higher computational complexity for determining correlation clusters. Furthermore, special projection or restriction schemes are required to relate 3D clusters to the structural variability of isosurfaces. Evidently, the use of correlation towers will become problematic on arbitrary surfaces due to distortions and possible penetrations of towers.

Acknowledgments

We would like to thank the ECMWF and Thomas Bodin from the RSES at ANU Canberra for providing the data sets used in this work. The work was funded by the Munich Centre for Advanced Computing (MAC) and the International Graduate School of Science and Engineering (IGSSE) at the Technische Universität München.

References

[ACN08] AILON N., CHARIKAR M., NEWMAN A.: Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM) 55*, 5 (2008), 23. 3, 4

[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. 2

[BBC04] BANSAL N., BLUM A., CHAWLA S.: Correlation clustering. *Machine Learning 56*, 1 (2004), 89–113. 3

[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. 3

[Bro04] BROWN R.: Animated visual vibrations as an uncertainty visualisation technique. In *GRAPHITE* (2004), ACM, pp. 84–89. 3

[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. 3

[DKLP02] DJURCILOV S., KIM K., LERMUSIAUX P., PANG A.: Visualizing scalar volumetric data with uncertainty. *Computers & Graphics 26*, 2 (2002), 239–248. 3

[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. 3

[GS06] GRIETHE H., SCHUMANN H.: The visualization of uncertain data: Methods and problems. In *Proceedings of SimVis 2006* (2006), pp. 143–156. 2

[Hol11] HOLY T.:. http://www.mathworks.com/matlabcentral/fileexchange/29702, 2011. 4

[JPR*04] JEN D., PARENTE P., ROBBINS J., WEIGLE C., TAYLOR II R., BURETTE A., WEINBERG R.: Imagesurfer: A tool for visualizing correlations between two volume scalar fields. 3

[JS03] JOHNSON C., SANDERSON A.: A next step: Visualizing errors and uncertainty. *Computer Graphics and Applications, IEEE 23*, 5 (2003), 6–10. 1, 2

[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. 3

[KWL*] KINDLMANN G., WEINSTEIN D., LEE A., TOGA A., THOMPSON P.: Visualization of anatomic covariance tensor fields. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1, IEEE, pp. 1842–1845. 3

[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. 3

[LFLH07] LI H., FU C., LI Y., HANSON A.: Visualizing large-scale uncertainty in astrophysical data. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1640–1647. 2

[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. 3

[MRH*05] MACEACHREN A., 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), 139–161. 2

[PH10] PÖTHKOW K., HEGE H.: Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *Visualization and Computer Graphics, IEEE Transactions on*, 99 (2010), 1–1. 3

[Pot] POTTER K.:. http://www.sci.utah.edu/~kpotter/library/uncertainVis/index.html. 2

[PRW11] PFAFFELMOSER T., REITINGER M., WESTERMANN R.: Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 951–960. 3, 7

[PWH11] PÖTHKOW K., WEBER B., HEGE H.: Probabilistic marching cubes. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 931–940. 3

[PWL97] PANG A., WITTENBRINK C., LODHA S.: Approaches to uncertainty visualization. *The Visual Computer 13*, 8 (1997), 370–390. 1, 2, 3

[RLBS03] RHODES P., LARAMEE R., BERGERON R., SPARR T.: Uncertainty visualization methods in isosurface rendering. In *Eurographics* (2003), Citeseer, pp. 83–88. 3

[STS06] SAUBER N., THEISEL H., SEIDEL H.: Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *Visualization and Computer Graphics, IEEE Transactions on 12*, 5 (2006), 917–924. 3

[SWMW09] SUKHAREV J., WANG C., MA K., WITTENBERG A.: Correlation study of time-varying multivariate climate data sets. In *Visualization Symposium, 2009. PacificVis' 09. IEEE Pacific* (2009), IEEE, pp. 161–168. 3

[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. *IEEE Transactions on Visualization and Computer Graphics* (2010). 3

[Tar05] TARANTOLA A.: *Inverse problem theory and methods for model parameter estimation*. Society for Industrial Mathematics, 2005. 3

[THM*05] THOMSON J., HETZLER E., MACEACHREN A., GAHEGAN M., PAVEL M.: A typology for visualizing uncertainty. In *Proc. SPIE* (2005), vol. 5669, Citeseer, pp. 146–157. 2

[WPL02] WITTENBRINK C., PANG A., LODHA S.: Glyphs for visualizing uncertainty in vector fields. *Visualization and Computer Graphics, IEEE Transactions on 2*, 3 (2002), 266–279. 3

[YXK11] YANG C., XIU D., KIRBY R. M.: Visualization of Covariance and Cross-covariance Fields. *International Journal for Uncertainty Quantification (to appear)* (2011). 3

[Zim08] ZIMEK A.: *Correlation Clustering*. PhD thesis, LMU München, 2008. 3

[ZWK10] ZEHNER B., WATANABE N., KOLDITZ O.: Visualization of gridded scalar data with uncertainty in geosciences. *Computers & Geosciences* (2010). 3