Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Visual Analysis of Spatial Variability and Global Correlations in Ensembles of Iso-Contours

F. Ferstl, M. Kanzler, M. Rautenhaus and R. Westermann

Computer Graphics and Visualization Group, Technische Universität München, Germany

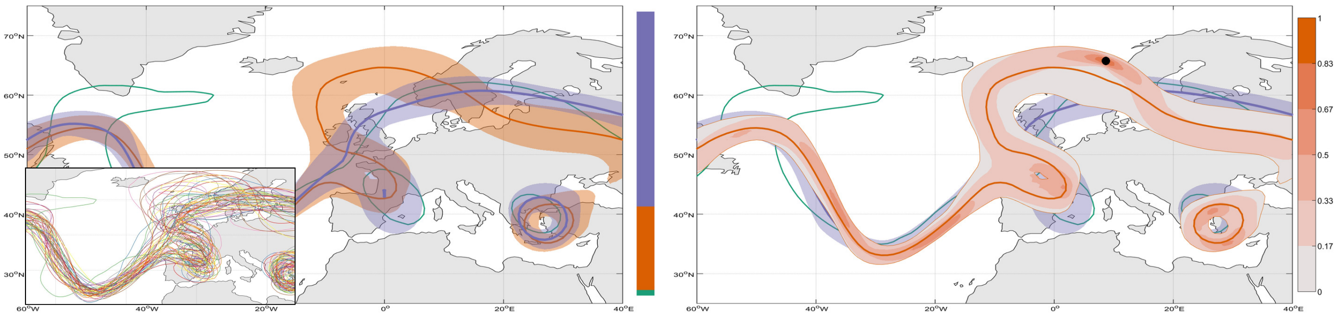


Figure 1: *From a set of iso-contours (inlay), our method computes an abstract visualization of the uncertainty that is carried by this set (left). The visualization is further refined by showing correlations, i.e. probabilities for the joint occurrence of iso-contours in a selected region (black circle) and other locations along the main trends (color coded in the right image).*

Abstract

For an ensemble of iso-contours in multi-dimensional scalar fields, we present new methods to a) visualize their dominant spatial patterns of variability, and b) to compute the conditional probability of the occurrence of a contour at one location given the occurrence at some other location. We first show how to derive a statistical model describing the contour variability, by representing the contours implicitly via signed distance functions and clustering similar functions in a reduced order space. We show that the spatial patterns of the ensemble can then be derived by analytically transforming the boundaries of a confidence interval computed from each cluster into the spatial domain. Furthermore, we introduce a mathematical basis for computing correlations between the occurrences of iso-contours at different locations. We show that the computation of these correlations can be posed in the reduced order space as an integration problem over a region bounded by four hyper-planes. To visualize the derived statistical properties we employ a variant of variability plots for streamlines, now including the color coding of probabilities of joint contour occurrences. We demonstrate the use of the proposed techniques for ensemble exploration in a number of 2D and 3D examples, using artificial and meteorological data sets.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

To analyze the *uncertainty* that is represented by an ensemble of iso-contours, the variability of the ensemble members needs to be characterized, and the major trends and outliers in the shape and spatial location of the contours need to be determined. A popular approach to visually analyze an ensemble of iso-contours is a so-called spaghetti plot, which shows the contours simultaneously in a single image.

The inset in Fig. 1(left) shows a spaghetti plot of iso-contours—distinguished by color—in a 2D geopotential height field. Since iso-contours can have very complicated shape and topology, spaghetti plots produce visual clutter when many contours overlap, and major trends, outliers, and statistical properties of the contour distribution cannot easily be conveyed. Thus, it is difficult to draw conclusions from a spaghetti plot on the different topological and structural characteristics of the contour set. For iso-contours in 3D fields, this problem becomes even more severe.

To overcome the limitations of spaghetti plots, Whitaker et al. [WMK13] and Mirzargar et al. [MWK14] have introduced visual abstractions for curve-like features based on the concept of statistical data depth, which enables the computation of a global ranking of such features and the use of this ranking to derive probabilistic indicators of the feature occurrence around the median. Ferstl et al. [FBW16] proposed to model statistically the distribution of streamlines, and to derive clusters indicating the major trends represented by an ensemble. By visualizing these clusters instead of the entire set of streamlines, the main trends can be conveyed effectively to the user.

In this work we show how to extend the method by Ferstl et al. for iso-contours in multi-dimensional scalar fields. This is challenging, because it first requires transforming the contours into a representation from which their similarity can be inferred. In contrast to ensembles of streamlines, which start at the same seeding position and can be computed using the same path-parametrization, such a transformation is not immediately given for iso-contours. In principle, geometry-based approaches for curve clustering [OLK*14] can be employed, yet they have difficulties when the contours exhibit different topologies and no consistent parameterization is given, which is common in real-world applications.

In addition to analyzing the major trends given by an ensemble of iso-contours, it is also interesting to investigate the relationship between the occurrence of a contour at different locations. In particular answering the question about the probability that a contour going through a specific location is also going through some other location can help to further explore the distribution of the iso-contours within each cluster. Especially when dealing with iso-contours or pathlines in time-varying fields, the analysis of such joint occurrences enables the exploration of interesting interrelations between the data properties at different locations and simulation times, and it can even help making predictions on the possible occurrences of features. To the best of our knowledge, a visual analysis approach for this kind of dependencies has not been proposed so far.

In this work, we address the aforementioned problems and propose new methods to explore the distribution of iso-contours in multi-dimensional scalar fields. Our specific contributions are:

- Building upon a distance field transformation of iso-contours [TN14, BM10] to obtain a consistent contour parametrization, we transform the contour ensemble into a reduced order space. According to [FBW16] we extract the major trends via clustering in the reduced order space, yet instead of random sampling fitted multivariate normal distributions we show how to derive analytically the corresponding visual abstractions in the spatial domain.
- We provide a mathematical formulation for the problem of computing the joint occurrences of contours at different locations via integration in the reduced order space over a region bounded by hyper-planes.

For visualizing the main trends in the iso-contour distribution we use variability plots as proposed by Ferstl et al. [FBW16]. In addition, we provide an interactive picking mechanism and graphically highlight joint occurrences with the picked region using color coding. To demonstrate the kind of information our approach can convey, we have performed a number of experiments using synthetic and real-world data sets from meteorology.

2. Related Work

Our technique belongs to the broader class of uncertainty visualization techniques. It falls into the category of ensemble visualization and uses curve clustering for computing confidence bands for sets of iso-contours.

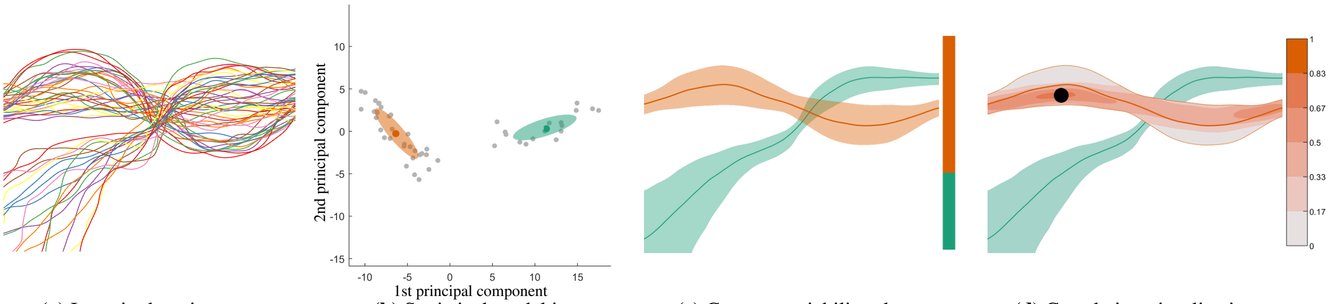
The importance of uncertainty visualization has been recognized over more than a decade ago [PWL97], and in a number of works since then, overviews and taxonomies of uncertainty visualization techniques have been given, for instance, by Johnson and Sanderson [JS03]. For the most recent survey on the topic let us refer to the book by Bonneau et al. [BHJ*14]. In ensemble visualization, it is assumed that the uncertainty is represented by a set of possible data occurrences, rather than a stochastic uncertainty model. Obermaier and Joy [OJ14] classify ensemble visualization techniques into feature-based and location-based approaches.

Related to our work are uncertainty visualization techniques which encode visually 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], surface diffusion techniques [GR04] as well as surface animations [Bro04]. Some more recent approaches [PRW11, PH10, PWH11] model the uncertainty stochastically and derive probability distributions for particular stochastic events associated to iso-surfaces.

The visualization of feature variability in ensemble fields is often performed via spaghetti plots of selected contour lines or threshold probabilities of 2D fields such as surface wind speed [PWB*09, Wil11]. Glyphs and confidence intervals were introduced to emphasize the Euclidean spread of contour and curve ensembles [SZD*10, MWK14, FBW16] and, specifically, contour boxplots [WMK13] have been applied to weather forecast data [QM16]. Locations in 3D flow ensembles are characterized based on the divergence of transport patterns [HOGJ13].

The analysis of mutual data dependencies has mostly been addressed via the visualization of data correlations. For instance, via color mapping and slicing [JPR*], by means of correlation fields and multi-field graphs [STS06], via glyph-based visualizations of local covariance structures [KWL*], or by correlation clustering to group objects with similar pair-wise correlations [BBC04, PW12].

Our work is also related to clustering approaches for parametrized curves in 2D and 3D space, which are most often based on geometric similarity measures. For an overview of similarity measures using geometric distances between curves let us refer to the comparative study by Zhang et al. [ZHT06] and, alternatively, for an overview on probabilistic model-based curve clustering algorithms let us refer to the work by Gaffney [Gaf04]. Agglomerative Hierarchical Clustering (AHC) with different cluster proximity measures has been shown for curves by McLoughlin



(a) Input in domain space **(b)** Statistical model in PCA-space $\mathbb{R}^r$ ($\approx$ SDF-space $\mathbb{R}^M$) **(c)** Contour variability plot **(d)** Correlation visualization

Figure 2: *Overview of our method:* ***(a)*** *Input set of iso-contours.* ***(b)*** *Statistical model of the corresponding signed distance functions in the principal component basis, represented by confidence ellipses and geometric cluster medians.* ***(c)*** *Contour variability plot, obtained by transforming ellipses and medians back to domain space (with cluster strengths indicated by a bar plot).* ***(d)*** *Visualization of global correlations in the orange cluster by picking a circular region (black) and color coding the conditional probability of contours going through other regions, given that they go through this picked region.*

et al. [MJL*13]. The overview article by Oeltze et al. [OLK*14] evaluates different clustering approaches for streamlines using geometry-based similarity measures. Thomas et al. [TN14] cluster iso-contours in a rotation and scale invariant feature space to find (self-) symmetries in iso-surfaces. To find similarities between different iso-contours of the same scalar field, Carr et al. [CBB06] use histograms and iso-surface statistics whereas Bruckner and Möller [BM10] use distance functions and information-theoretic measures.

An alternative to geometry-based curve clustering approaches are dimension-reduction techniques, where the initial curves are represented in some low-dimensional subspace. For shape analysis, Rathi et al. [RDT06] use kernel PCA on distance functions for shape analysis, and Leventon et al. [LGF00] employ the power of PCA on distance functions and use a multivariate normal distribution as a statistical shape model. Fofonov et al. [FML15] use distance functions and dimension reduction to visualize a set of contour time series as 2D or 3D trajectories.

3. Overview

Given an ensemble of 2D or 3D scalar fields $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^M$ (defined on the same grid comprised of M vertices), we consider the iso-contours which are extracted from each ensemble member $\mathbf{s}_i$ for the same iso-value v. Note that we are not considering value uncertainty in this work (i.e., the uncertainty in the choice of the iso-value itself) and therefore do not consider the possibility that different ensemble members can contain similar iso-contours with different iso-values. A comparative analysis of iso-contours is difficult, because they often consist of many different, disconnected parts and can vastly differ in length, i.e., they vastly differ in the number of degrees of freedom. To account for this, we fix this number to the number of grid vertices M, by representing the contours implicitly via signed distance functions (SDF). For each contour i of scalar field $\mathbf{s}_i$ we compute a corresponding SDF $\mathbf{d}_i \in \mathbb{R}^M$, which specifies at each grid point of $\mathbf{s}_i$ the signed distance to the closest point on the contour (with the sign indicating on which side of the iso-contour the point is located). While, in principle, each $\mathbf{d}_i$ is a grid of signed distance values, it can also be interpreted as a vector with M entries and, hence, as a point in the Euclidean space $\mathbb{R}^M$ which directly corresponds to contour i. We will subsequently refer to $\mathbb{R}^M$ as *SDF-space*.

By using the SDF representations, we can build upon the concept proposed by Ferstl et al. [FBW16] for generating so-called streamline variability plots, i.e., lobular visual abstractions for conveying the main trends in an ensemble of particle trajectories. An overview of the corresponding steps we perform in the current work for ensembles of iso-contours is shown in Fig. 2. We call the resulting visual abstractions *bands*, which can be generated in 2D and 3D using equivalent operations.

From the SDF representation of the input contours (Fig. 2a), we compute a statistical model of the contour distribution. This is required to obtain a mechanism for generating "continuous" distributions in the spatial domain. To this end, we perform a principal component analysis (PCA) of the M-dimensional point set $\mathbf{d}_1, \ldots, \mathbf{d}_n$. This is depicted in Fig. 2b, where the gray points are the points $\mathbf{d}_i$ in the resulting principal component basis (for illustration purposes the depiction is restricted to the projection to the first two components), and each of those points corresponds to one contour from Fig. 2a. Subsequently, we determine a clustering of the transformed points in the so-called *PCA-space*. By fitting a multivariate normal distribution (MVN) to each cluster, clusters can be represented by their geometric median and a confidence ellipse with a user-defined size in units of standard deviation (Fig. 2b). E.g., the points in Fig. 2b are grouped into two clusters, and each cluster is represented by a confidence ellipse with a size of one standard deviation (colored ellipses) and its geometric median (colored dots).

Ferstl et al. performed a dense sampling of the ellipses to generate new streamline realizations following the main trends in the data, and these realizations were transformed into the spatial domain to form the variability lobes. We show in our work that for sets of iso-contours, represented via distance functions, the shape of the corresponding bands is fully determined by a box-shaped region

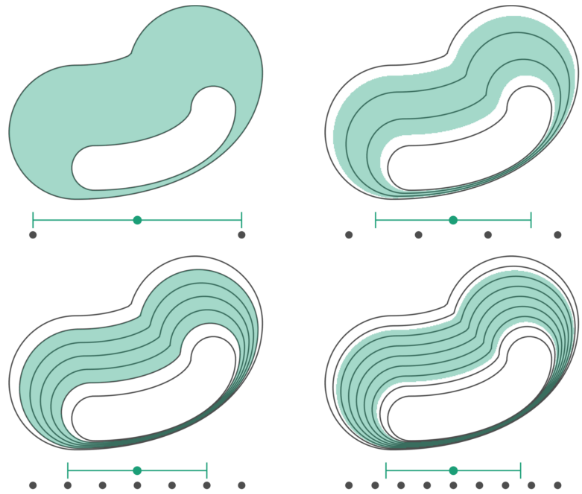


Figure 3: *Standard deviation of distance functions: Four different ensembles of equally spaced iso-contours (black lines) and corresponding confidence bands (colored green) are shown. Bands indicate the regions within one standard deviation from the mean contour. Below each image is a corresponding 1D plot of equally spaced points and an error bar showing mean and standard deviation.*

in PCA-space. From this observation we derive a method to analytically transform the confidence ellipses in PCA-space into their corresponding representations in the spatial domain. This transformation yields SDFs from which a *contour variability plot* can be derived (Fig. 2c). Conceptually, the visualized bands correspond to the boundaries of confidence ellipses in PCA-space and illustrate the standard deviations of the per-cluster sets of contours, which is illustrated in Fig. 3.

Since box-shaped regions can be used in PCA-space instead of oriented ellipses to generate exactly the same bands, some information that is given by the ellipses' orientation is lost. The orientation is determined by the off-diagonal entries of the covariance matrix of the derived clusters, and these entries indicate the pairwise correlations in the clustered data. It is important to note that neglecting this information does not affect the shapes of the bands generated, yet the missing information can be revealed by another method allowing us to perform a refined cluster analysis including joint probabilities for the occurrence of iso-contours at different locations. We compute these probabilities by performing partial set operations in PCA-space, which take into account the orientation of the confidence ellipses. The derived information about the inherent dependencies between different locations are finally used to indicate the clusters' sub-structures (Fig. 2d).

4. Contour Variability Plots

In the following we assume that a PCA of the SDFs $\mathbf{d}_i$ has been computed, and that the full set of $r := n-1$ principal components (PCs) is used (the latter assumption will be relaxed later on). Thus, each $\mathbf{d}_i$ is represented exactly by a corresponding $\mathbf{c}_i \in \mathbb{R}^r$:

$$\mathbf{d}_i = R(\mathbf{c}_i) := \bar{\mathbf{d}} + U\mathbf{c}_i \quad (i = 1,\ldots,n)\,. \tag{1}$$

Here, the vector $\bar{\mathbf{d}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i$ denotes the mean SDF, and the columns of $U \in \mathbb{R}^{M\times r}$ are the PCs, which form an orthonormal basis of a subspace of the SDF-space. Let us also introduce the function $R(\mathbf{c}_i) : \mathbb{R}^r \to \mathbb{R}^M$ to denote the corresponding reconstruction operation from PCA-space to SDF-space. R can be applied to arbitrary points $\mathbf{c} \in \mathbb{R}^r$, but then the resulting reconstructions $R(\mathbf{c})$ are, in general, not valid distance functions. However, they have similar properties like smoothness and monotonicity, and their zero-contours can be interpreted as artificial iso-contours.

In addition to the PCA, we assume that the $\mathbf{c}_i$ have been clustered, and for each cluster k an approximating MVN distribution has been computed. Each MVN distribution is parameterized by a covariance matrix $\Sigma_k \in \mathbb{R}^{r\times r}$ and mean value $\mu_k \in \mathbb{R}^r$. Since the number of points in a cluster is relatively small compared to the dimension r, usually the covariance matrix does not have full rank. However, we formally assume that Σ_k^{-1} exists, which does not pose a problem because it is not required in our final computation scheme.

To compute the standard deviation of a set of iso-contours, we represent the MVN distribution of each cluster in PCA-space using a single confidence ellipse. From this ellipse, the band representing the standard deviation of the corresponding set of iso-contours can be computed analytically, not requiring an exhaustive sampling and line rasterization process as for streamlines [FBW16].

4.1. Band Computation

The boundary $\mathcal{E}$ of a confidence ellipse for any MVN distribution $\mathcal{N}(\mathbf{x},\mu,\Sigma)$ with covariance matrix Σ and mean μ to a given number of standard deviations α can be defined as the set of points $\mathbf{x}$ with Mahalanobis distance α:

$$\mathcal{E}(\Sigma,\mu,\alpha) = \left\{\mathbf{x} \;\middle|\; (\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) = \alpha^2\right\}.$$

If we now consider the boundary of the confidence ellipse $\mathcal{E}_k := \mathcal{E}(\Sigma_k,\mu_k,\alpha) \subset \mathbb{R}^r$ representing a cluster k in PCA-space, a set of SDFs $R(\mathcal{E}_k) \subset \mathbb{R}^M$ are obtained when this ellipse is transformed to SDF-space. The corresponding band is defined as the set of points in domain space which is covered by at least one zero-contour of all SDFs in $R(\mathcal{E}_k)$.

In general, the extraction of a band from a (finite or infinite) set of distance functions $\mathcal{D} \subset \mathbb{R}^M$ can be realized analytically by determining a scalar function $L(\mathcal{D}) \in \mathbb{R}^M$ as

$$L(\mathcal{D}) = \operatorname{cmin}\left\{-\operatorname*{cmin}_{\mathbf{d}\in\mathcal{D}}\{\mathbf{d}\},\ \operatorname*{cmax}_{\mathbf{d}\in\mathcal{D}}\{\mathbf{d}\}\right\}, \tag{2}$$

where cmin and cmax denote component-wise minimum and maximum operators. By construction, $L(\mathcal{D})$ is positive inside the band, negative outside the band, and its zero-contour corresponds to the band's boundary. This is illustrated in Fig. 4 for a 1D example.

To evaluate Eq. (2) for a set $\mathcal{D}$ of SDFs we require the minimum and maximum values in each of the M dimensions ($\hat{=}$ at each grid point). This means that each band function $L(\mathcal{D})$ is fully defined by the axis-aligned bounding box of $\mathcal{D}$ in SDF-space and, hence, a

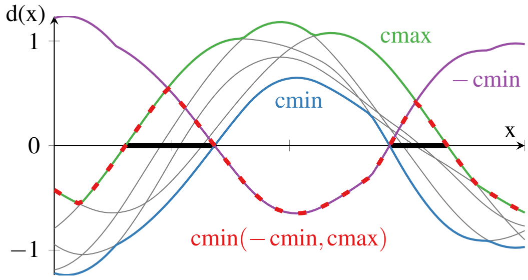


Figure 4: *Illustration of $L(\mathcal{D})$ (Eq. (2)). For a set of 1D SDF-like functions $\mathcal{D} = \{d_1, ..., d_5\}$ (gray lines), the "band" covered by their zero-contours (thick, black lines) is the region where $L(\mathcal{D}) = \mathrm{cmin}\{-\mathrm{cmin}_i\, d_i,\ \mathrm{cmax}_i\, d_i\} \geq 0$.*

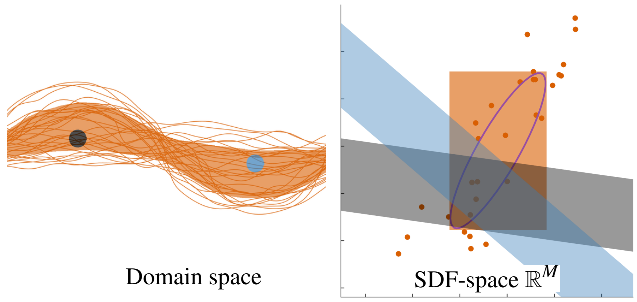


Figure 5: *Domain and SDF space: Each contour in an ensemble (orange lines) corresponds to a point in SDF-space (orange points). A band of standard deviation corresponds to an axis-aligned rectangular region in SDF-space (orange regions), which is determined by the confidence ellipse that is fitted to the points in SDF-space (purple line). The contours going through the black and blue circle, respectively, form slabs in SDF-space. From their intersection, a joint occurrence in the two circles can be computed.*

band in domain space corresponds to an axis-aligned box in SDF-space (see Fig. 5).

Because R is an orthonormal transformation, $R(\mathcal{E}_k)$ in SDF-space is also a confidence ellipse. Formally, this ellipse can also be determined directly in SDF-space by fitting a MVN distribution with $\Sigma_k^M \in \mathbb{R}^{M\times M}$ and $\mu_k^M \in \mathbb{R}^M$ to the SDFs $\mathbf{d_i}$ of cluster k. This gives the M-dimensional confidence ellipse $\mathcal{E}_k^M := \mathcal{E}(\Sigma_k^M, \mu_k^M, \alpha)$, which is equal to $R(\mathcal{E}_k)$. Consequently, the corresponding band functions $L(R(\mathcal{E}_k))$ and $L(\mathcal{E}_k^M)$ are equal, too.

In practice, on the other hand, the calculation of Σ_k^M is not feasible due to its large size. However, as shown before, $L(\mathcal{E}_k^M)$ is fully defined by the axis-aligned bounding box of $\mathcal{E}_k^M$ in $\mathbb{R}^M$ and, in turn, the bounding box of a confidence ellipse is fully defined by its mean and the diagonal of its covariance matrix. Hence, for the computation of $L(\mathcal{E}_k^M) = L(R(\mathcal{E}_k))$, we can simplify Eq. (2) to

$$L(\mathcal{E}_k^M) = \mathrm{cmin}\Big\{ -\Big(\mu_k^M - \alpha\sqrt{\mathrm{diag}(\Sigma_k^M)}\Big),\; \mu_k^M + \alpha\sqrt{\mathrm{diag}(\Sigma_k^M)}\Big\}. \tag{3}$$

This means that instead of transforming a confidence ellipse from PCA-space to domain space, we can construct a band locally (directly from all $\mathbf{d}_i$ in the corresponding cluster). The band is the region enclosed by the zero-contours of two fields which are computed point-wise as "*mean* $\pm\,\alpha \cdot$ *standard deviation*".

5. Analysis of Global Correlations

In the previous chapter we have shown the construction of the confidence bands in domain space from axis aligned bounding boxes in SDF-space. We now demonstrate that by considering in addition the orientation and extent of the confidence ellipses in PCA-space, we can derive information concerning the probabilities of joint occurrences of iso-contours. Underlying this insight is the fact that the ellipses encode global correlations between the different SDFs $\mathbf{d}_i$, giving rise to our intended contour analysis.

By using the MVN distribution of a single cluster, we investigate the event of iso-contours going through a given location in the domain. Therefore, let us assume that $\mathbf{y}$ is the position of a grid point with index $j \in \{1, ..., M\}$ in the 2D or 3D domain. Then, all SDFs $\mathbf{d} \in \mathbb{R}^M$ corresponding to contours through $\mathbf{y}$ have to be zero at that grid point j, i.e., $(\mathbf{d})_j = 0$. If $\mathbf{y}$ does not fall exactly onto a grid point, we can use linear interpolation. Let $\mathbf{w}_y \in [0;1]^M$ denote the linear interpolation weights for point $\mathbf{y}$, where $\mathbf{w}_y$ has at most four or eight non-zero entries in 2D or 3D, respectively, and the weights sum up to one. Then, all contours through $\mathbf{y}$ have to satisfy $\mathbf{d}^T\mathbf{w}_y = 0$, meaning that the set of contours through $\mathbf{y}$ corresponds to a hyper-plane in SDF-space (with normal $\mathbf{w}_y$).

To compute the probability of the occurrence of an event, we have to extend the hyper-plane to a volumetric region. This is because the probability that a contour is going exactly through a given point is zero. A first idea would be to vary $\mathbf{y}$, but this would change the normal of the corresponding hyper-plane. Instead, we use the fact that the SDFs store distances to the contours and consider the event of crossing a circular region rather than a point location.

Let $I(\mathbf{y}, s)$ denote the event of an iso-contour coming closer to the point $\mathbf{y}$ than a prescribed radius s. Then the following set of SDFs corresponds to $I(\mathbf{y}, s)$:

$$\{\mathbf{d} \in \mathbb{R}^M \mid |\mathbf{d}^T\mathbf{w}_y| \leq s\}$$

This is a subset of $\mathbb{R}^M$ that is enclosed between two parallel hyper-planes, and we will refer to such a subset as *slab* in the following. A slab can be "restricted" from SDF-space to PCA-space by substituting $\mathbf{d} = R(\mathbf{c}) = \bar{\mathbf{d}} + U\mathbf{c}$ according to Eq. (1), resulting in a r-dimensional slab denoted by $\mathcal{S}(\mathbf{y}, s)$:

$$\mathcal{S}(\mathbf{y}, s) = \{\mathbf{c} \in \mathbb{R}^r \mid |\mathbf{c}^T(U^T\mathbf{w}_y) + \bar{\mathbf{d}}^T\mathbf{w}_y| \leq s\}$$

For a set of contours, represented by an MVN distribution $\mathcal{N}(\mathbf{c}, \mu_k, \Sigma_k)$ of a cluster k in PCA-space, we can now calculate the probability of $I(\mathbf{y}, s)$ by integrating over the corresponding slab:

$$P\big(I(\mathbf{y}, s)\big) = \int_{\mathcal{S}(\mathbf{y},s)} \mathcal{N}(\mathbf{c}, \mu_k, \Sigma_k)\, d\mathbf{c}.$$

Furthermore, to investigate correlations between different locations in the domain, i.e., to determine probabilities of joint occur-

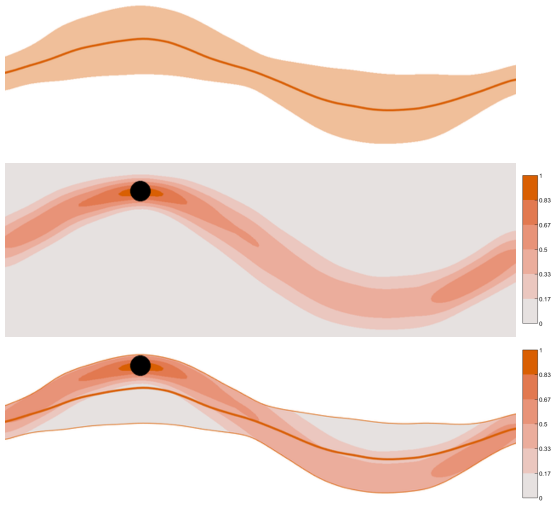


Figure 6: *Coloring of bands for the contour cluster in Fig. 5. Top: A unicolored variability plot with band and median. Middle: Color coding of the probability function $f(\mathbf{z})$ at every domain point, and at points in the band (bottom).*

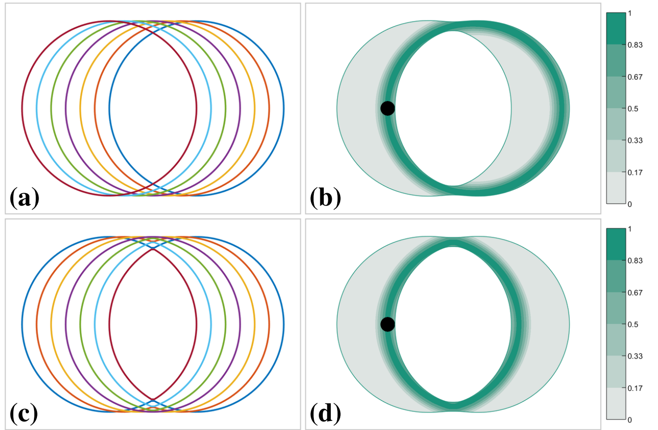


Figure 7: *Color coding of conditional probabilities.* ***(a)*** *Spaghetti plots comprised of shifted circles of the same size, and* **(c)** *the same circles cut off and mirrored at the center vertical axis.* ***(b+d)*** *The bands corresponding to one standard deviation $\alpha = 1$ (bounded by green lines). Only when coloring the bands according to $f(\mathbf{z})$ in a region around point $\mathbf{y}$ with radius t (black circle), the contours' structures can be revealed.*

rence of contours at these locations, we can consider a second circular region at some other point location $\mathbf{z}$ with radius t. To calculate the joint probability of the event $I(\mathbf{y},s) \wedge I(\mathbf{z},t)$, which indicates that a contour comes within a radius s of $\mathbf{y}$ *and* within a radius t of $\mathbf{z}$, we have to integrate over the union of the two corresponding slabs (cf. Fig. 5):

$$P\big(I(\mathbf{y},s) \wedge I(\mathbf{z},t)\big) = \int_{\mathcal{S}(\mathbf{y},s)\cap\mathcal{S}(\mathbf{z},t)} \mathcal{N}(\mathbf{c},\mu_k,\Sigma_k)\, d\mathbf{c} \tag{4}$$

This r-dimensional integral is difficult to compute directly, yet it can be transformed analytically into the following integral over a rectangular region of a 2D MVN distribution:

$$\begin{aligned} &\int_{\beta(\mathbf{w}_y)-s}^{\beta(\mathbf{w}_y)+s} \int_{\beta(\mathbf{w}_z)-t}^{\beta(\mathbf{w}_z)+t} \mathcal{N}\big((x_1,x_2)^T, \mathbf{0}, \Sigma_k'\big)\, dx_2\, dx_1 \\ &\text{with } \beta(\mathbf{w}) = -\mu_k^T(U^T\mathbf{w}) - \bar{\mathbf{d}}^T\mathbf{w} \\ &\text{and } \Sigma_k' = \begin{pmatrix} \mathbf{w}_y^T U\Sigma_k U^T \mathbf{w}_y & \mathbf{w}_z^T U\Sigma_k U^T \mathbf{w}_y \\ \mathbf{w}_z^T U\Sigma_k U^T \mathbf{w}_y & \mathbf{w}_z^T U\Sigma_k U^T \mathbf{w}_z \end{pmatrix}. \end{aligned} \tag{5}$$

For a proof of this statement let us refer to the supplementary material, yet what can be seen immediately is that a change of variables can be performed such that the normals of the two r-dimensional slabs are aligned with the first two coordinate axes. This allows collapsing all remaining dimensions of the MVN distribution, leaving a 2D problem. The computation of 2D rectangular normal probabilities like this is a standard problem in computational statistics and can be performed, e.g., using the method by Genz et al. [Gen04], which is included in a number of statistics libraries. To visualize the result of Eq. (4+5), we choose locations $\mathbf{y}$ and $\mathbf{z}$ and prescribe the corresponding radii s and t. The user selects $\mathbf{y}$ and lets compute the joint probabilities for all other regions. For simplicity, we always use equal radii in this work, i.e., $s = t$. Additionally, to obtain probability values in the range $[0;1]$, we normalize the computed values with $P\big(I(\mathbf{y},t)\big)$. This yields the following function $f(\mathbf{z})$, which is defined over the entire domain and describes the conditional probability that contours come within radius t of $\mathbf{z}$, given that they also come within radius t of $\mathbf{y}$:

$$f(\mathbf{z}) = \frac{P\big(I(\mathbf{y},t) \wedge I(\mathbf{z},t)\big)}{P\big(I(\mathbf{y},t)\big)} \tag{6}$$

We optionally use $f(\mathbf{z})$ to color-code the bands in our visualizations based on a user defined circle $(\mathbf{y},t)$, which is shown in Fig. 6. In this way, we can highlight regions which are likely to contain contours that run through the selected region and, therefore, we can reveal correlations between different regions in the domain which remain hidden when the bands are visualized in a single color. An example for such a scenario is shown in Fig. 7. Note that computing the normalization factor $P\big(I(\mathbf{y},t)\big)$ requires an additional integration over a single slab. This can be achieved either by reducing the integration to a 1D problem (analogous to the reduction of Eq. (4) to 2D) or by simply making one of the slabs in Eq. (4+5) infinitely large.

6. Contour Variability Plots

In this section we give additional details concerning the PCA computation and the clustering process in PCA-space. Furthermore, we describe how to compose and generate the final contour variability plots.

PCA: In practice, we do not use the full set of $(n-1)$ PCs, but we reduce the PCA-space to a dimension $r \leq n-1$. To determine r, we use the common *explained variance* criterion: Let σ_j denote the variance of the PCA coefficients along the j-th PC. Then the total amount of explained variance by r PCs is defined as: $\mathrm{ex}(r) = \sum_{j=1}^{r} \sigma_j / \sum_{j=1}^{n-1} \sigma_j$. Based on a threshold $\tau \in [0,1]$, we choose the smallest r which satisfies $\mathrm{ex}(r) \geq \tau$. Since the input iso-contours often contain many small details and the dimensionality of the PCA-space only has a minor impact on the performance of our implementation, we usually only drop the very insignificant PCs. I.e., we use $\tau = 0.999$ in all our experiments, which, for the meteorological ensembles used in our work, leads to roughly 35-45 required PCs out of 51.

Clustering: To separate the major trends in a contour ensemble, the resulting PCA coefficients $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^r$ are clustered. We do not fit multiple MVN distributions directly to the data using the commonly applied Expectation Maximization algorithm, because it leads to numerical problems due to the high dimensionality of the PCA-space and does not ensure a good spatial separation of the clusters in this space. Instead, we use agglomerative hierarchical clustering (AHC) with average linking, which is initialized with the Euclidean distances in PCA-space as distance measure. Furthermore, we perform an automated guess for the optimal number of clusters using the L-method [SC04]. This initial guess can by changed by the user, causing clusters to split and merge recursively according to the AHC hierarchy. For a deeper discussion of these choices for clustering we refer the reader to [FBW16].

Median Contour: For each cluster of input contours, we draw an artificial median contour (which does not exist in the original set of contours) as a single representative for that cluster. We compute the *geometric median* $\hat{\mathbf{c}}_k$ of the corresponding points in PCA-space and transform it back to the initial domain by extracting the zero-contour of $R(\hat{\mathbf{c}}_k)$. We draw the median contour with a line width corresponding to the relative size of the cluster, and each plot is augmented with an additional **bar plot** to show these sizes exactly.

Bands: For each cluster, a band corresponding to a user-defined size α in units of standard deviation is drawn (we use $\alpha = 1$ in all our examples). Bands are computed according to Eq. (3) and visualized as filled, transparent polygons in 2D, and using iso-surface raycasting in 3D. Since thin bands can often not be represented by a single scalar field, we use two intermediate fields and delay the outermost *cmin* operation in Eq. (3) until the values have been interpolated from these intermediate fields.

Outliers are detected by our hierarchical clustering as clusters of cardinality one. The contour in an outlier cluster is drawn directly and no band is shown (cf. Figs. 1, 9g+h, 10c+d). To this end, we do not explicitly search for outliers.

Correlation Color Coding: The user can click on a lobe in order to pick a circle of interest (always shown in black), defining the circle $(\mathbf{y}, t)$. The picked lobe is then visualized on top of all others and color coded with $f(\mathbf{z})$ according to Eq. (6), using the same radius t for both involved circles. The values that are shown are therefore actual probability values (and not a probability density), indicating at every point $\mathbf{z}$ the conditional probability that a line through the black circle is also going through a circle of the same size centered at $\mathbf{z}$.

7. Results

This section presents meteorological examples of 2D and 3D contour variability plots and of 2D correlation visualizations, using weather forecast data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (ENS).

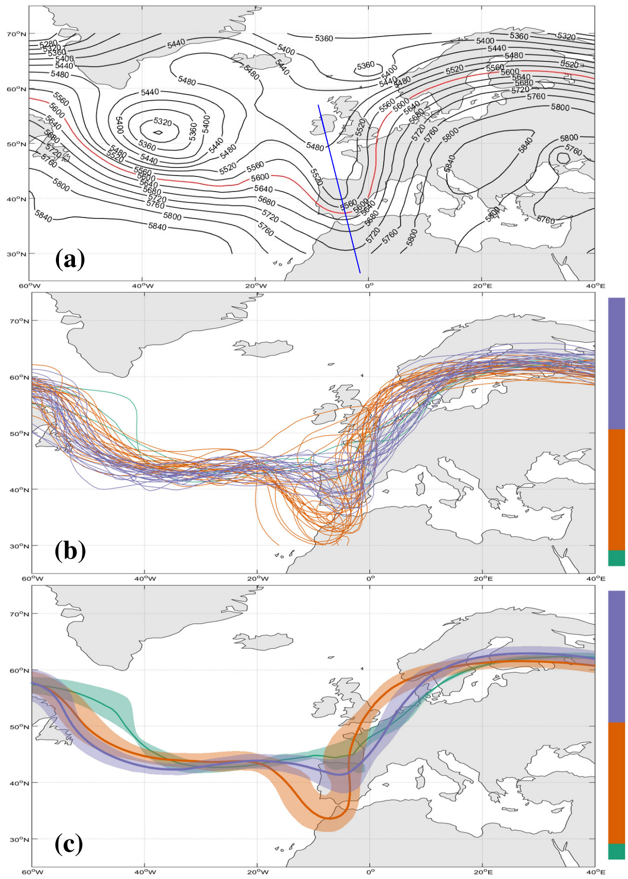


Figure 8: *Spatial variability of an ensemble of 500 hPa geopotential height contour lines. The ECMWF ENS forecast from 00:00 UTC, 15 October 2012, valid at 00:00 UTC, 20 October 2012, is shown.* ***(a)*** *Geopotential height contours (m) of the ensemble control forecast (5600 m highlighted in red). Note the distinct trough over Spain (blue axis).* ***(b)*** *Spaghetti plot of the 5600 m contour lines of all 51 ensemble members, colored by cluster membership.* ***(c)*** *Contour variability plot.*

The forecasts include 50 perturbed members and a control run. For details, we refer to e.g. [LP08]. For the present study, we use the forecast from 00:00 UTC 15 October 2012 that has been used previously by Rautenhaus et al. [RKSW15]. The example region encompasses $101 \times 41 \times 62$ grid points. For the 2D examples, the data were interpolated to vertical levels of constant pressure.

For 2D interactive demonstrations, our implementation uses the MATLAB implementations of PCA and AHC. The publicly available FORTRAN implementation of the method by Genz et al. [Gen04] was used for the calculation of the 2D rectangular, normal probabilities. A custom implementation was used for computing signed distance transforms in which, firstly, we compute the unsigned distance field to an iso-contour of a given scalar field using fast marching [Set99] and, secondly, determine the sign of the resulting distance value at every grid point by comparing its scalar value against the specified iso-value. Since the conditional proba-

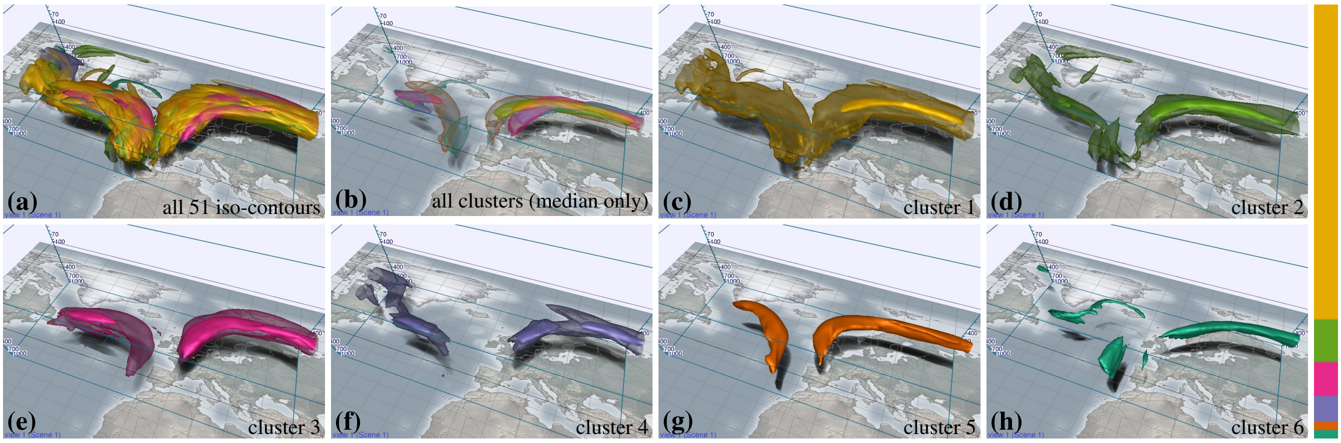


Figure 9: *Contour variability plot for an ensemble of 3D wind speed contours representing the jet stream. 50 ms^{-1} iso-surfaces of the ECMWF ENS forecast from 00:00 UTC, 15 October 2012, valid at 18:00 UTC, 19 October 2012 are shown. For comparison, the same scene as in Fig. 6 in [RKSW15] has been chosen. The images show **(a)** a "3D spaghetti" plot of all 51 iso-surfaces (colored by cluster membership), **(b)** the median contours of six clusters identified by the algorithm, and **(c-h)** the median contour (opaque) and the outer boundary of the "band" (transparent) for the six clusters. Note that clusters 5 and 6 contain one outlier each.*

bilities $f(\mathbf{z})$ typically are smooth, we compute them in a texture using half the grid spacing of the original grid.

All presented results were generated on a standard desktop PC (Intel Xeon X5675 processor with 6 $\times$ 3.0 GHz, 8 GB RAM and an NVIDIA Geforce GTX 680). On this machine, the generation of a contour variability plot for an ensemble of 2D forecast fields (101×41 grid) takes on average 270 ms. The timing includes 66 ms for the distance transformation, 31 ms for the PCA of the distance functions, and 31 ms for hierarchical clustering. The time increases by 80 ms if a correlation visualization for a single cluster is computed. For the 3D example, the contour variability plot (i.e., the SDFs for median contours and the outer boundaries of the "bands") was pre-computed and rendered via iso-surface raycasting in the open-source visualization tool Met.3D [RKSW15]. The pre-computation required 11.8 s, including 8.1 s for the distance transform, 2.6 s for the PCA and 40 ms for hierarchical clustering.

7.1. Contour variability plots

To demonstrate the proposed contour variability plots, we discuss two meteorological examples. Figure 8 shows 2D contours of the forecast geopotential height field at 500 hPa. Spaghetti plots of such fields are frequently encountered in meteorology (e.g., [Wil11]). The geopotential height field of the control forecast (Fig. 8a) shows a distinct trough extending from Ireland over Spain (blue line). We focus on the 5600 m contour. The spaghetti plot of all members' 5600 m contours (Fig. 8b) reveals variability in the line ensemble, in particular in the trough region. Without cluster coloring, it is impossible to discern differences and trends (not shown). The colouring in Fig. 8b increases the information content; the variability plot in Fig. 8c clearly shows the possible scenarios. The purple and orange clusters contain the majority of the members. The bands of both clusters are largely similar, except for the region around and east of the trough. The lines in the orange cluster represent a much stronger trough. In contrast, contours in the third cluster (green) show notable differences both south of Greenland and in the trough region. We note that the clustering algorithm takes into account the entire domain. If the user's interest is centred on a particular feature in a particular region, it may be useful to let the user select a subregion.

In Fig. 9, 3D contour clustering is applied to an example from Rautenhaus et al. [RKSW15]. 50 ms^{-1} wind speed iso-surfaces are used as representative features of the jet stream, strong winds (with core wind speeds often exceeding 50 ms^{-1}) within narrow bands typically encountered around 10 km height (e.g., [Ahr08]). Rautenhaus et al. [RKSW15] (their Fig. 6) show iso-surfaces of selected single members and of ensemble statistical quantities, rendered by the Met.3D system. To discern different scenarios of jet feature characteristics, the user of Met.3D can animate over the ensemble members, i.e., cycle through the display of single members. Our clustering method largely enhances this simple approach. Fig. 9a shows the iso-surfaces of all members in a "3D spaghetti" plot. In 3D, differences between the members are even more difficult to discern than in the 2D case; there is little more information in the plot that in the plot of the ensemble maximum in Fig. 6f of [RKSW15]. In 3D, plots of all cluster medians and their bands (as in Fig. 8c) also suffer from cluttering. We hence visualize a composite of the cluster medians only to show the major scenarios of the jet stream feature (Fig. 9b), as well as individual medians and bands of six clusters that our method has detected (ordered from the largest to the smallest cluster in Fig. 9c-h). While with this approach, the user still needs to animate over the clusters to get information on possible jet characteristics, the number of animation steps is reduced from 51 members to 6 clusters.

7.2. Correlation Visualization

Examples of plots with the proposed visualization of global correlations are depicted in Fig. 10. Fig. 10a shows correlations of iso-contours passing through a user-positioned circle within the purple

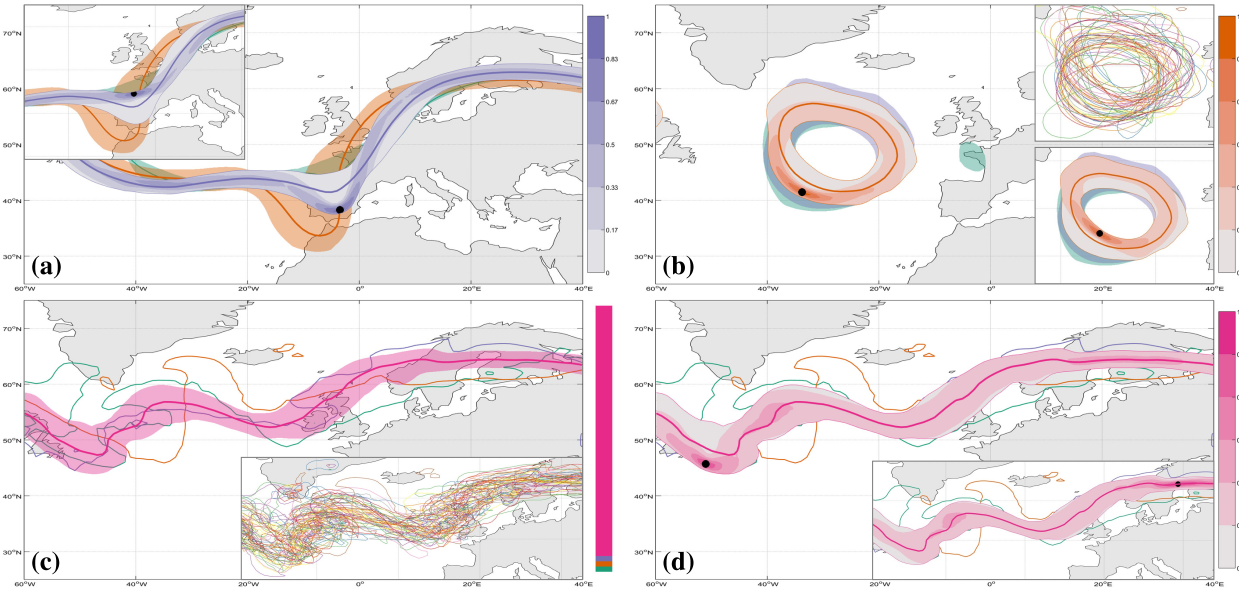


Figure 10: *Spatial correlations. **(a)** Correlations in the purple cluster of Fig. 8c. **(b)** Correlations of a cluster of 680 m contour lines of geopotential height at 925 hPa show a behavior similar to the idealized case in Fig. 7b (forecast from 00:00 UTC, 17 October 2012, valid at 12:00 UTC, 21 October 2012). **(c)** Contour lines of 0 °C of the temperature field at 850 hPa are clustered into a single cluster and three outliers (same forecast as in Fig. 8). **(d)** Except for the westernmost part of the analyzed region, these lines are almost entirely uncorrelated.*

cluster of the example in Fig. 8c. Over western Europe, strong correlations are visible; iso-contours that describe a more pronounced trough (circle positioned over southern Spain) tend to proceed on the western edge of the cluster band over the North Sea (i.e., deeper troughs are more narrow). In other regions, however, the position of the lines is largely uncorrelated to the strength of the trough. Closer to the surface (925 hPa and 36 hours later; Fig. 10b), the low-pressure system visible south of Greenland in Fig. 8a can be well identified in the geopotential height field. Here, contour lines in the selected cluster resemble a situation similar to the one in Fig. 7a+b. We conclude that the system has similar size in the individual ensemble members within the cluster, but is shifted in space.

Our third example (Fig. 10c+d) shows 0 °C temperature iso-contours at 850 hPa. Here, the contour lines contain many high frequency details; in the spaghetti plot, structures are impossible to discern (Fig. 10c). Our clustering methods yields one large cluster and three outliers. Positioning the circle in the cluster band south of Greenland (Fig. 10d) reveals correlations in the westernmost part of the region, where the low pressure system pushes cold air from the northwest towards the southeast. Hence, if air below 0 °C has been pushed as far south as the circle location at 50 °W, the cold air has also been pushed to the southern side of the variability band further west. East of the circle, however, the contour lines passing through the circle are distributed almost uniformly within the band, indicating that the deviations of the contours from the cluster median are of random nature and not globally correlated. In this case, we cannot make any conclusion, e.g., about the 0 °C boundary being located further south over Great Britain if it is located further south at 50 °W. The uniform distribution is confirmed by placing the circle at different locations (one of which is shown in the inset).

8. Conclusion and Future Work

In this work we have shown that the use of distance functions enables us to generate contour variability plots which allow for a characterization of the uncertainty that is represented by an ensemble of iso-contours. In contrast to spaghetti plots, the visual abstraction that is provided by our method effectively conveys the major trends and outliers in a given set of iso-contours. While our method is naturally limited in 3D due to overlaps and occlusion problems, our clustering and visual abstraction nevertheless provide an improvement over having to animate through single ensemble members. In addition, the proposed color coding of confidence bands can assist a user in detecting global correlations between occurrences of contours at different locations. It enables the interactive exploration of interior structures in clusters. Thus, it can help to improve an initial clustering and to decide whether a local or a global analysis of a set of contours is appropriate. In the future, we plan to further investigate the potential of our approach to analyze the time evolution of dynamical phenomena and to use our clustering to determine the occurrence of similar features yet at different locations and times. Moreover, we will address the current limitation of our method in 3D, that despite clustering and visual abstraction, occlusions and visual clutter thereof are introduced.

Acknowledgements

Access to ECMWF prediction data has been kindly provided in the context of the ECMWF special project “Support Tool for HALO Missions”. This work was supported by the European Union under the ERC Advanced Grant 291372—SaferVis: Uncertainty Visualization for Reliable Data Discovery.

References

[Ahr08] AHRENS C. D.: *Meteorology Today*, 9 ed. Brooks Cole, 2008. 8

[BBC04] BANSAL N., BLUM A., CHAWLA S.: Correlation clustering. *Machine Learning 56*, 1 (2004), 89–113. 2

[BHJ*14] BONNEAU G.-P., HEGE H.-C., JOHNSON C., OLIVEIRA M., POTTER K., RHEINGANS P., SCHULTZ T.: Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*. Springer, 2014, pp. 3–27. 2

[BM10] BRUCKNER S., MÖLLER T.: Isosurface similarity maps. *Computer Graphics Forum 29*, 3 (2010), 773–782. 2, 3

[Bro04] BROWN R.: Animated visual vibrations as an uncertainty visualisation technique. In *GRAPHITE* (2004), pp. 84–89. 2

[CBB06] CARR H., BRIAN D., BRIAN D.: On histograms and isosurface statistics. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 1259–1266. 3

[FBW16] FERSTL F., BÜRGER K., WESTERMANN R.: Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 767–776. 2, 3, 4, 7

[FML15] FOFONOV A., MOLCHANOV V., LINSEN L.: Visual analysis of multi-run spatio-temporal simulations using isocontour similarity for projected views. *IEEE Transactions on Visualization and Computer Graphics, to appear* (2015). 3

[Gaf04] GAFFNEY S.: *Probabilistic Curve-Aligned Clustering and Prediction with Mixture Models*. Ph.d. dissertation, University of California, Irvine, 2004. 2

[Gen04] GENZ A.: Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing 14*, 3 (2004), 251–260. 6, 7

[GR04] GRIGORYAN G., RHEINGANS P.: Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics 10*, 5 (2004), 564–573. 2

[HOGJ13] HUMMEL M., OBERMAIER H., GARTH C., JOY K.: Comparative visual analysis of Lagrangian transport in CFD ensembles. *IEEE Trans. on Visualization and Comp. Graph. 19*, 12 (2013), 2743–2752. 2

[JPR*] JEN D., PARENTE P., ROBBINS J., WEIGLE C., TAYLOR R., BURETTE A., WEINBERG R.: Imagesurfer: A tool for visualizing correlations between two volume scalar fields. In *IEEE Visualization, 2004*, pp. 529–536. 2

[JS03] JOHNSON C., SANDERSON A.: A next step: Visualizing errors and uncertainty. *IEEE Comp. Graph. and Appl. 23*, 5 (2003), 6–10. 2

[KWL*] KINDLMANN G., WEINSTEIN D., LEE A., TOGA A., THOMPSON P.: Visualization of anatomic covariance tensor fields. In *Proc. IEEE Eng. in Medicine and Biology Society, 2004*, pp. 1842–1845. 2

[LGF00] LEVENTON M., GRIMSON W., FAUGERAS O.: Statistical shape influence in geodesic active contours. In *Proc. IEEE Computer Vision and Pattern Recognition* (2000), pp. 316–323. 3

[LP08] LEUTBECHER M., PALMER T.: Ensemble forecasting. *Journal of Computational Physics 227*, 7 (2008), 3515–3539. 7

[MJL*13] MCLOUGHLIN T., JONES M. W., LARAMEE R. S., MALKI R., MASTERS I., HANSEN C. D.: Similarity measures for enhancing interactive streamline seeding. *IEEE Transactions on Visualization and Computer Graphics 19*, 8 (2013), 1342–1353. 3

[MWK14] MIRZARGAR M., WHITAKER R., KIRBY R. M.: Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Trans. on Visualization and Computer Graphics 20*, 12 (2014), 2654–2663. 2

[OJ14] OBERMAIER H., JOY K.: Future challenges for ensemble visualization. *IEEE Comp. Graphics and Applications 34* (2014), 8–11. 2

[OLK*14] OELTZE S., LEHMANN D. J., KUHN A., JANIGA G., THEISEL H., PREIM B.: Blood flow clustering and applications in virtual stenting of intracranial aneurysms. *IEEE Transactions on Visualization and Computer Graphics 20(5)* (2014), 686–701. 2, 3

[PH10] PÖTHKOW K., HEGE H.: Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 99 (2010), 1–1. 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, pp. 951–960. 2

[PW12] PFAFFELMOSER T., WESTERMANN R.: Visualization of global correlation structures in uncertain 2D scalar fields. *Computer Graphics Forum 31* (2012), 1025–1034. 2

[PWB*09] 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 *Proc. IEEE Int. Conf. on Data Mining Workshops* (2009), pp. 233–240. 2

[PWH11] PÖTHKOW K., WEBER B., HEGE H.-C.: Probabilistic marching cubes. *Computer Graphics Forum 30* (2011), 931–940. 2

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

[QM16] QUINAN P. S., MEYER M.: Visually comparing weather features in forecasts. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 389–398. 2

[RDT06] RATHI Y., DAMBREVILLE S., TANNENBAUM A.: Statistical shape analysis using kernel PCA. *Proc. SPIE 6064* (2006), 60641B. 3

[RKSW15] RAUTENHAUS M., KERN M., SCHÄFLER A., WESTERMANN R.: Three-dimensional visualization of ensemble weather forecasts – Part 1: The visualization tool Met.3D (version 1.0). *Geoscientific Model Development 8*, 7 (2015), 2329–2353. 7, 8

[SC04] SALVADOR S., CHAN P.: Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence* (2004), pp. 576–584. 7

[Set99] SETHIAN J. A.: *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge Univ. Press, 1999. 7

[STS06] SAUBER N., THEISEL H., SEIDEL H.: Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Trans. on Visualization and Comp. Graphics 12*, 5 (2006), 917–924. 2

[SZD*10] SANYAL J., ZHANG S., DYER J., MERCER A., AMBURN P., MOORHEAD R. J.: Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics 16* (2010), 1421–1430. 2

[TN14] THOMAS D., NATARAJAN V.: Multiscale symmetry detection in scalar fields by clustering contours. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2427–2436. 2, 3

[Wil11] WILKS D. S.: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 2011. 2, 8

[WMK13] WHITAKER R. T., MIRZARGAR M., KIRBY R. M.: Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2713–2722. 2

[ZHT06] ZHANG Z., HUANG K., TAN T.: Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Proc. 18th Int. Conf. on Pattern Recognition* (2006), pp. 1135–1138. 2