Comparative Visual Analysis of Transport Variability in Flow Ensembles

Mihaela Jarema
TU München
Boltzmannstrasse 3
85748, Garching bei
München, Germany
mihaela.jarema@tum.de

Johannes Kehrer
TU München
Boltzmannstrasse 3
85748, Garching bei
München, Germany
johannes.kehrer@tum.de

Rüdiger Westermann
TU München
Boltzmannstrasse 3
85748, Garching bei
München, Germany
westermann@tum.de

ABSTRACT

We propose a novel approach that enables a comparative visual exploration of the transport variability in ensembles of 2D flow fields. To reveal when and where divergences in transport occur, we first present a new approach to analyze the time-varying pairwise dissimilarities of ensemble trajectories, by using Gaussian Mixture Models (GMMs) to identify the distribution modes and the Mahalanobis distance to refine the dissimilarity measures. This enables drawing enhanced spaghetti plots, by using the color of the contour of each trajectory to encode the temporal evolution of the member, and the opacity for its representativeness relative to the ensemble behavior. To also allow a global view of the transport variability across selected sub-domains, we introduce a new graphical abstraction based on the visualization of miniaturized versions of the enhanced spaghetti plots in a small-multiples layout. To achieve this, we propose a new kind of downscaling that preserves the relevant trends in the transport behavior. We have designed a user interface comprising multiple linked views to visualize simultaneously global and local transport variations, as well as how similar the transport behavior of the ensemble members is.

Keywords
Uncertainty Visualization, Ensemble Vector Field Data, Time-varying Data.

1 INTRODUCTION

In many scientific disciplines, ensemble simulations are used to estimate the uncertainty inherent in the development of physical fields, by providing representative samples of the possible states that could evolve out of perturbed initial conditions and different models. Weather forecasting is one such example, where multiple forecasts are performed simultaneously to obtain probabilities of occurrence of specific weather events. Analyzing the temporal variability of an ensemble helps determine when and where divergences occur, and, implicitly, the locations where and the time intervals over which a simulation is more or less reliable.

Analyzing the ensemble variability is, however, not straightforward: Firstly, transport divergences occur gradually over time, due to spatial variations of the transport paths and the transport velocity along them. Analyzing the flow variability requires new concepts

to determine these divergences and, in particular, when the divergences started occurring. Secondly, visualizing the temporal variability of an ensemble – locally, at a selected domain point, or globally, to compare the transport across the domain – is challenging, since it requires new graphical abstractions to show the complex spatio-temporal ensemble evolution in an intuitive way.

Contribution: We propose a visual analytics approach to address the aforementioned challenges and explore the temporal variability of ensembles of vector fields. We provide novel means to perform a local and global visual analysis of the dissimilarities of ensemble members across the domain.

To analyze the transport deviations over time statistically, we introduce the application of the Mahalanobis distance on Gaussian Mixture Models (GMMs). Approximating the distributions of tracer particles at every grid point with mixtures of Gaussian modes enables the use of the Mahalanobis distance to assess the pairwise member dissimilarities relative to the variability allowed by each mode. To quantify the ensemble divergence over time further, we define the divergence count of an ensemble member per time step as the (normalized) number of members dissimilar to it. As members evolve, their pairwise dissimilarities and divergence counts change, revealing the time steps and locations of new (dis)similarities to others members.

The information we derive from the pairwise dissimilarities and divergence counts is then encoded graphically, to enable an effective local and global visual analysis. We propose the following novel approaches:

- **Divergence Visualization:** We enhance the spaghetti plots of the ensemble trajectories by encoding the transport evolution and divergence counts over time; this shows when and where ensemble members behave (dis)similarly, and is especially effective in cases when trajectories exhibit similar geometric shapes, but trajectory points considered at the same time instant are dispersed.

- **Small-Multiples:** We compare the transport variability across selected sub-domains via a new graphical abstraction based on miniaturized versions of enhanced spaghetti plots in a small-multiples layout. We present a new downscaling method that preserves the relevant trends in the transport behavior.

- **Similarity visualization:** We cluster ensemble members based on their flow similarity across the domain and visualize the dynamical evolution of the clusters using parallel sets.

The proposed approaches are combined into an interactive visual analytics tool, to give insight into the various aspects of the flow variability. We evaluate our approach on several meteorological ensembles and include an assessment by domain experts.

2 RELATED WORK

Uncertainty visualization, one of the top challenges in scientific visualization [1, 2], is often estimated by ensembles – representative samples of realizations of simulated phenomena, obtained by running simulations with different initial conditions and models. Such data is typically spatio-temporal, multivariate, and multivalued [3, 4], making its analysis and visualization difficult. Typical methods evaluate summary statistics and visualize these using color maps, contours, surface deformation, opacity, boxplots, or glyphs [5, 4, 6, 7].

For vector fields, Wittenbrink et al. [8] propose glyphs to show the magnitude and angular uncertainty. Lodha et al. [9] show the flow uncertainty using envelopes and animation. Pfaffelmoser et al. [6] present circular glyphs to convey the uncertainty of gradients in mean and orientation in 2D uncertain scalar fields. Jarema et al. [10] use lobular glyphs to visualize multimodal distributions for 2D directional data. Other local methods include texture mapping [11] and a reaction-diffusion model [12]. Allendes Osorio and Brodlie [13] present a modified LIC version for 2D uncertain steady vectors. For time-varying uncertain vector fields, Hlawatsch et al. [14] introduce flow radar glyphs. In the crisp case,

Hlawatsch et al. [15] downscale individual pathlines, whereas we downscale entire ensembles of trajectories. To consider the transport uncertainty, Otto et al. use particle density functions to obtain an uncertain topological segmentation for 2D [16] and 3D [17] Gaussian-distributed steady vector fields; Schneider et al. [18] analyze the transport uncertainty for unsteady vector fields. Hummel et al. [19] compare the material transport in time-varying flow ensembles by computing individual and joint vector field variances. We analyze the transport variability based on pairwise trajectory dissimilarities, rather than variances, which allows us to identify when divergences occur.

Clustering is another standard method for dealing with large and complex data [20]. Bruckner and Möller [21] propose density-based clustering to identify similar volumetric time sequences in physically-based ensemble simulations. Bordoloi et al. [22] perform realization- and distribution-based hierarchical clustering of ensemble data. Hierarchical clustering is also used to cluster trajectories, e.g., for streamlines modeling blood flow [23] or meteorological data [24]. Hollister and Pang [25] cluster streamlines using DBSCAN to derive scalar fields capturing cases when ensemble members may exhibit strong separation along trajectories, but a weak terminal separation. Although our analysis is performed on trajectories, it clusters trajectory positions and not the trajectories themselves. Thus, we are able to determine not only whether divergences occur along the trajectories, but also when they occur. Jarema et al.[10] use GMMs to cluster ensemble members based on the extent of their directional similarity locally. We extend this method to cluster ensemble members by their transport similarity, where the dissimilarity of two members is obtained by applying the Mahalanobis distance on the modes identified by means of a GMM approximation.

Fitting GMMs has been used before in uncertainty visualization. Correa et al. [26] use GMMs to model uncertainty in the visual analysis process, while Hollister and Pang [27] apply GMMs to perform probability distribution interpolation. Liu et al. [28] employ multiple Gaussian components for a volume rendering of stochastic fields. The Mahalanobis distance [29] has been used, e.g., for uncertain scalar fields, to assess the positional uncertainty of isosurfaces [30] or reveal the possible locations of critical points [31].

Other techniques for dealing with the complexity of ensemble data include coordinated multiple views, which are used to study multivariate relations via linking and brushing [3], and parallel coordinates [32] or parallel sets [33]. Nocke et al. [34], and Molchanov and Linsen [35] use coordinated multiple views for climate ensembles. Piringer et al. [36] analyze 2D function ensembles on three levels of details (member-oriented, domain-oriented, and surface plot).
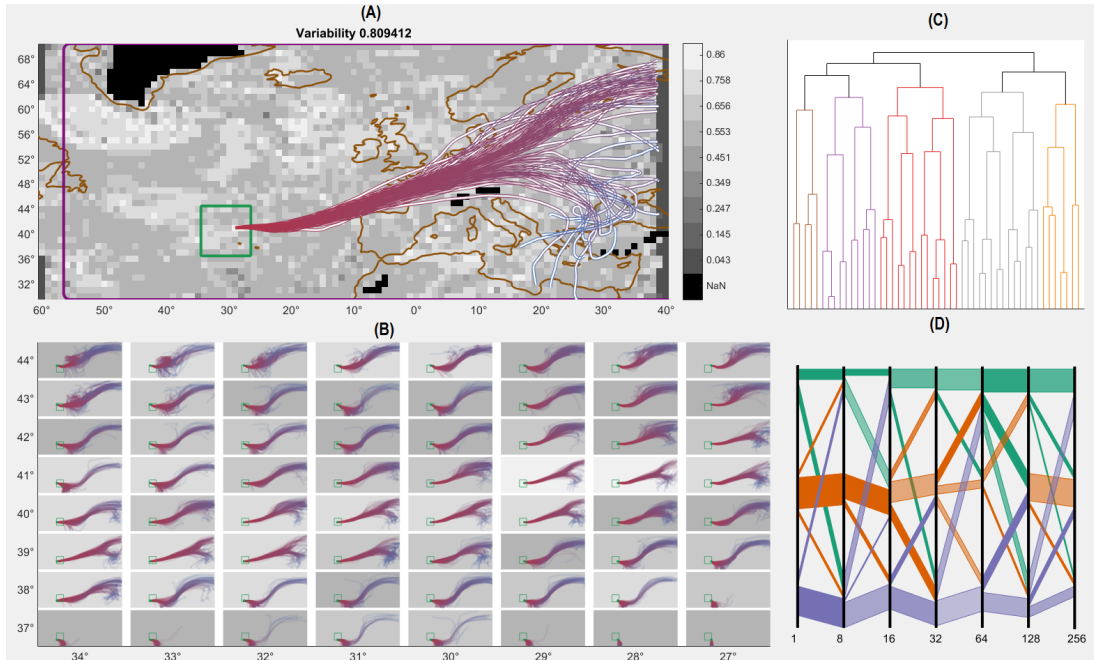
Figure 1: Multiple linked views for an ECMWF (European Centre for Medium-Range Weather Forecasts) time-varying ensemble. The spatial view (A) shows the aggregated flow variability over the domain and an enhanced spaghetti plot at a user-selected location. For the selected region marked by a green square, the detail view (B) shows downscaled spaghetti plots. View (C) shows the hierarchical clustering of the members with perturbed initial conditions. View (D) shows the variability of the clustering solution at selected time steps.

3 OVERVIEW

Our method starts with a vector field ensemble given on a 2D grid structure, and identifies at every grid point the time steps and locations of dissimilar behavior. From this, we derive measures for the temporal evolution of the transport variability, to show the variability over the domain (cf. background in Fig. 1(A)) and generate enhanced spaghetti plots at selected locations (cf. foreground in Fig. 1(A)). The enhanced plots reveal the flow variability even when particles follow geometrically similar trajectories, but with significantly different speeds, being temporally actually dispersed.

To explore and compare the flow variability at multiple locations simultaneously, concurrent spaghetti plots lead to massive clutter and occlusion. Hlawatsch et al. [15] juxtapose miniaturized trajectory images to overcome this limitation for crisp vector fields. Our ensemble visualization builds upon the concept of small-multiples [37], but includes the derived transport variability to construct small-multiples preserving the main trends of the particle trajectories (cf. Fig. 1(B)). This is achieved by downscaling trajectories based on a selection of salient time steps.

Finally, our method clusters members based on their global transport similarity, and shows these clusters via dendrograms (cf. Fig. 1(C)). The temporal evolution of clusters is encoded visually via parallel sets (cf. Fig. 1(D)), from which splitting and merging events over time can be deduced.

To quantitatively assess the flow variability, our method determines when and where members behave dissimilarly, without imposing any synthetic thresholds beyond which the behavior is considered dissimilar (cf. Sec. 4.1). This allows us to assess the spatial variability at individual time steps or over a forecast interval, and obtain the salient time steps for the small-multiples layout (cf. Sec. 4.2). We also use the method to cluster ensemble members based on their transport similarity throughout the domain (cf. Sec. 4.3).

4 TRANSPORT VARIABILITY

To find out when and where changes in the transport behavior occur, we assess members as (dis)similar based on their deviation at every time step, but without imposing any artificial thresholds. For normally distributed data, the Mahalanobis distance [29] specifies natural thresholds. Due to its use of the covariance matrix, it is scale-independent and unaffected by correlations between variables. Members are thus (dis)similar based on their deviation relative to the data variability. Ensemble data is, however, often not Gaussian distributed. We thus model the distribution of 2D particle positions (seeded at the same point) at every time step with a mixture of Gaussians, and apply the Mahalanobis distance to assess pairs of members as (dis)similar relative to the variability of the corresponding mode.

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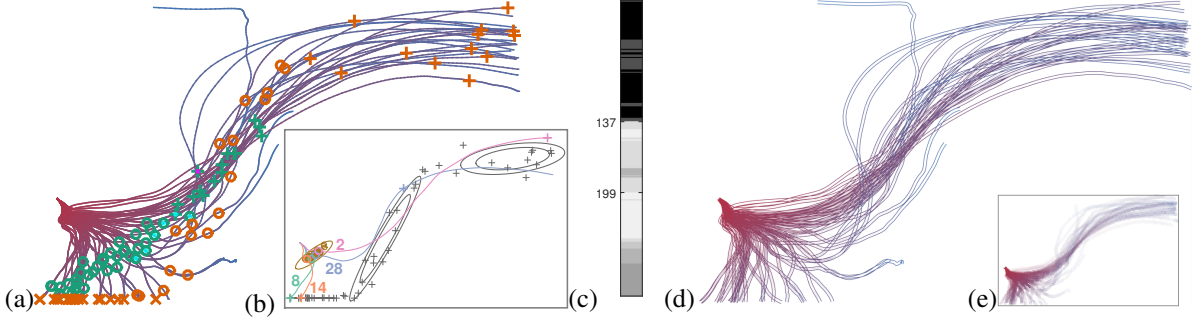
Figure 2: Pathline variability: (a) trajectories of 51 members; particle positions are shown at steps 137 and 199 with green and orange markers, respectively, and different symbols for different clusters; (b) four singled-out members: 8 (green), 14 (orange), 28 (blue), and 2 (pink), with markers for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

semble: members 2, 8, 14, and 28. Even though a bimodal distribution is fitted only at step 137 and no synthetic threshold is defined for the spread, member 8, initially an inlier, is, up to step 58, similar only to member 28 relative to the local variability (step 58 is marked by brown concentric ellipses for the Gaussian mode and circles for the particle positions). Thereafter, it also becomes and stays similar to member 14, both following the downward trend. As the deviation between members 8 and 28 increases, the two become dissimilar shortly after step 58; members 8 and 2 are dissimilar from the very beginning, despite the spread being initially very low. Actually, member 2 is dissimilar for most of the time even to 28, although they both show geometrically similar trajectories and follow the upward trend. This happens because the particles travel at different speeds (at, e.g., step 199, where the three fitted modes are shown by concentric gray ellipses and all members are marked with pluses, note that positions are modeled by different modes).

4.1 Pairwise Dissimilarity Analysis

We consider flow field ensembles – collections of n vector fields defined over the same grid structure. Our approach is designed for both stationary and time-dependent 2D vector fields, but the analysis can be extended directly to 3D. We obtain the trajectories by numerical integration, e.g., using 4-th-order Runge-Kutta methods. At every grid point, each of the n trajectories comprises m_i integration steps. To make all trajectories of equal length m, we repeat the final positions $m - m_i$ times, where $m = \max_{i=1,n}(m_i)$. Otherwise, members with similar trajectories, but of slightly different lengths $m_i > m_j$, e.g., members 8 and 14 in Fig. 2(b), would be artificially dissimilar in the $m_i - m_j$ interval. Members with trajectories of considerably different length and behavior are dissimilar as soon as their deviations are large relative to the allowed variability or the positions are modeled by different modes, e.g., members 8 and 28 in the same

figure. Moreover, the (dis)similarity of two members (i, j) is fixed after step $\max(m_i, m_j)$.

The pairwise analysis occurs in two stages, performed at every grid point and integration step of the member trajectories seeded at that grid point: In the first stage, we use GMMs to model the distribution of particle positions at each time step. We determine both the number of modes and their shapes automatically, by adapting for 2D data the procedure described in [10] for 2D directional data. Thus, unless the positions can be assumed Gaussian distributed, an Expectation Maximization (EM) algorithm fits two Gaussian modes to the positions and assigns each member to the mode that is more likely to model the observation. The process repeats until the members in each partition can be assumed Gaussian distributed (cf. Fig. 2(b) for an example of GMM partitions – shown using concentric ellipses – at several time steps). Depending on the initial conditions, EM algorithms may lead to non-repetitive solutions. To alleviate this problem, we run the GMM algorithm several times with different starting values, and use silhouettes – validation techniques for algorithms that use random initial guesses – to select the best solution from those having the most frequently met modality.

A sample of 2D observations $x_i = [X_i, Y_i]$, $i = 1, n$, is modeled with a mixture of N Gaussian modes

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

where each Gaussian mode $\mathcal{N}(\mu_j, \Sigma_j)$ has a weight α_j and is parameterized by its mean μ_j and covariance matrix Σ_j. Each observation $x_i, i = 1, n$, is modeled by Gaussian mode j with probability $p_{j,i}$, where $j = 1, N$ and $\sum_{j=1}^{N} p_{j,i} = 1$. The GMM creates a first partitioning of the members, depending on which mode is more likely to model the observations. We use *soft* clustering, where observations are dissimilar only if they belong with high probabilities to different modes. Thus, obser-
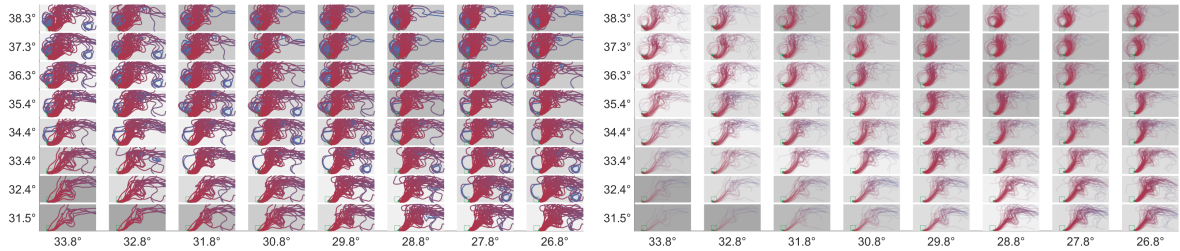
Figure 3: Small-multiples displays for ensembles are significantly more cluttered when downscaled in (left) a straightforward manner than (right) with sampling and variability encoding in opacity.

vations located on the boundary of two modes are similar to observations modeled by both modes, as shown in Fig. 2(a): a bimodal distribution is fitted for the first time at step 137 (cf. the green markers, of two types, circles and pluses); a few members (with cyan dots in the middle of the markers) are at the boundaries of both modes and thus similar to all members. However, just because only at this step there is enough proof for bimodality, this does not necessarily mean that members now modeled by different modes have started diverging at this step. Their deviation may have been large relative to the local variability already earlier, e.g., members 8 and 28, or 8 and 2 in Fig. 2(b).

Thus, in the second stage, for every pair of similar members (x_i, x_j), we compute the Mahalanobis distance relative to the corresponding covariance matrix Σ_k

$$MD(x_i, x_j) = \sqrt{(x_i - x_j)\Sigma_k^{-1}(x_i - x_j)^{\mathrm{T}}}. \quad (2)$$

Members are dissimilar if $MD(x_i, x_j) > 2.3$, which corresponds statistically to less likely deviations falling outside one confidence region (68%) for a bivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$. For boundary members, we compute the distance relative to both modes and take the higher value. Out of the six members in the previous example, only the one in the upper mode still exhibits similarities to members in the other mode. This occurs because the covariance matrix allows more variation in the vertical direction than in the horizontal one. For the same reason, the member with a magenta dot in the middle is dissimilar to all others.

4.2 Variability over the Domain

The analysis based on GMMs and the Mahalanobis distance identifies the pairwise (dis)similarities of the ensemble members at every time step. To distinguish between inliers and outliers within modes or among all members, we define the *member divergence count* (mdc$_i^t$) of ensemble member i at step t as the number of dissimilar members at that instant, normalized by the total number of members. For Gaussian distributions, positions closer to the mean have lower divergence counts than those located further. For multimodal distributions, the divergence counts generally increase for both mode inliers and outliers.

To reveal the time steps and locations with higher flow variability, we sum up and normalize the divergence counts of all members at a time step t_k to obtain a *time step divergence count* (tdc^{t_k}). The aggregated time step divergence count over a given time interval $[t_1, t_2]$ is median(tdc^{t_k})$_{t_k \in [t_1, t_2]}$. We use the median because it is a robust measure of the central tendency of the data, but other summary statistics could be used instead. Moreover, to distinguish between low and high divergence counts, we define reference values, per member (the median of the minimum divergence count at every grid point (i, j), median(min(mdc$_{i,j_k^t}$)$_{k=1,n,t=1,t_{\max}}$)), and per time step (the median of the divergence counts at the first time step, median(tdc$_{i,j}^1$)).

We also use the divergence counts to find a selection of time steps that preserves the transport behavior. Thus, we downscale trajectories without obscuring the transport trends and variability, as a straightforward downscaling would (cf. Fig. 3 (left)). For all members, we keep the salient trajectory points – the steps with changes in (dis)similarities. Other steps are sampled regularly, the sampling rate being proportional to the number of non-salient steps. While the information loss in intervals with high curvature may lead to a coarser curve approximation, the flow structure is qualitatively well-preserved (cf. Fig. 3 (right)).

The time step divergence counts for the previous example (with values in the $[0.55, 0.82]$ range) are encoded in grayscale in Fig. 2(c). While there is not enough proof for a departure from normality, the divergence counts vary in the $[0.55, 0.61]$ range. A surge to 0.76 occurs at step 137, when a bimodal distribution is fitted to the data, succeeded by a small gradual decreases as more particles follow the upward trend. Despite the geometrical similarity of the trajectories going upwards, particles move at different speeds. This leads to a second surge (0.81) at step 199, when three Gaussian modes are fitted (cf. the orange markers in Fig. 2(a)). As particles approach their final positions (and since most of them end up together, either in the bottom or the right bundle), the divergence count decreases again, although to values no lower than 0.69. During downscaling, the non-salient time steps are sampled and around 60% of the trajectory points discarded, but the flow structure and variability are retained (cf. Fig. 2(e)).

4.3 Flow-based Similarity

To analyze the transport similarity of the ensemble, we extend the work of Jarema et al. [10], who model directional data locally via GMMs and use the modes to cluster members hierarchically based on their local angular similarity over the whole domain. The binary similarity of any two members at a point depends on the angular deviation relative to local variation of their mode, and the global similarity measure is defined as the percentage of similar locations out of the total number of grid points. Here, we use the Mahalanobis distance to determine deviations in particle positions (rather than angles) that are statistically meaningful relative to the mode variability. Moreover, we perform the clustering over a whole forecast interval to cluster ensemble members based on their transport similarity (rather than their local angular similarity).

5 VISUAL ANALYSIS OF THE TRANSPORT VARIABILITY

Our user interface (cf. Fig. 1) comprises four linked views, two for the variability over the domain and another two for the variability of the clustering solution.

5.1 Flow Variability over the Domain

The spatial view (cf. Fig. 1(A)) shows the ensemble variability (here the aggregated time step divergence count for the entire forecast interval and all members) over the domain using a sequential grayscale colormap. A time slider allows an interactive visualization of the variability for single time steps or aggregated over time intervals, either for a selection of members or the entire ensemble (by using the time step divergence count). Users can select locations where to display enhanced spaghetti plots. Grayscale colormaps permit the use of a colorblind-safe red-to-blue variation for the trajectories, making them stand out against the background. Nevertheless, they suffer from simultaneous contrast effects, which hinders the correct interpretation of the displayed data [38]. To ease the variability evaluation, we inform the user about the variability at the selected location and show the global reference values.

Trajectory plots encode both the integration time and the variability (member divergence counts). Each trajectory is displayed as a curve stripe, the contour of the stripe color-coding the time and the stripe itself the variability. The color of each stripe is based on that of its contour, fading to white as the member divergence count increases relative to the local reference value (the smaller value between the local minimum and the global reference value). Thus, the stripe color for members having divergence counts close to the local reference value can hardly be distinguished from that of the contour. As the divergence count increases, the stripe color fades to white, making the trajectories of
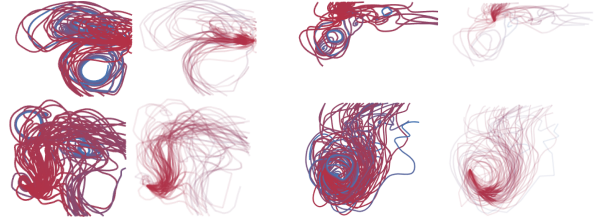

Figure 4: In a straightforward downscaling (odd plots – the first plot is the leftmost top plot) clutter and occlusion obscure the flow trends and variability; downscaling with sampling and variability encoding (even plots) conveys the main flow patterns and variability.

members with higher variability stand out less against the grayscale background than those with lower variability. The time information is nevertheless fully preserved in the color of the contours.

Fig. 2(d) shows the enhanced spaghetti plot for the previous example. The stripes begin to fade rapidly, as particles tend to spread out from an early stage. However, up to step 137, when a bimodal distribution is fitted to the data, the color of most stripes, albeit suggestive of the proximate split, has yet to fade completely. Thereafter, the stripes of several trajectories going upwards fade to white, because particles travel at considerably different velocities and member divergence counts are high. Despite the geometrical similarity of the trajectories, the fading reveals the flow variability (the increasing dispersion of the particles and number of partitions). Also, the stripes fade less for most trajectories in the lower bundle; their flow behavior is more similar, even if from their geometry alone they appear more dispersed than the trajectories in the upper bundle.

5.2 Flow Variability over Subdomains

To allow the concurrent visualization of enhanced spaghetti plots across selected subdomains with significantly less clutter and occlusion, we propose a small-multiples layout with miniaturized plots, where the downscaling preserves the salient flow trends.

The detail view (cf. Fig. 1(B)) shows a predefined number of downscaled plots, computed at regular intervals in the domain. A straightforward downscaling may obstruct the flow trends and variability (cf. Fig. 4). For example, in the first plot, the ending spiraling structures of several pathlines obscure the stable flow pattern at the beginning. Instead, we determine a selection of time steps that preserves the transport behavior (cf. Section 4.2). We also encode the integration time directly in the color of the stripes and the variability in the opacity, so that the main flow trends are clearly discernible in the miniaturized versions. Notice how the opacity use and the dense sampling of the spiraling parts (where the flow variability hardly changes anymore) in the second

plot in Fig. 4 help bring out both the flow patterns and the variability.

To enable a comparative visual analysis of the downscaled plots, all plots have the aspect ratio of the original domain and the same domain section (shown in the main spatial view by a purple rectangle). The area occupied by the detail view is displayed by a green square. To preserve context information, the green square (correspondingly scaled) is also shown in each downscaled plot, along with a green anchor ball that marks the seeding point of the trajectories. The initial domain section is the bounding box containing all spaghetti plots, but can be adjusted interactively (as was done in Fig. 1(A)).

5.3 Flow-based Similarity Visualization

The hierarchical clustering of the ensemble (based on the global flow similarity over the whole time interval) is summarized as a dendogram (cf. Fig. 1(C)). The ensemble members are shown on the horizontal axis and the clustering levels on the vertical axis. The dissimilarity levels increase as subclusters are merged, joins being represented graphically as inverted U lines. Neighboring larger subclusters are shown in different colors, to make partitions stand out. Users can select (groups) of members in the dendogram (or as text input) to visualize their spatial variability.

Insight into the dynamical evolution of the clustering solution is gained using parallel sets [33] (cf. Fig. 1(D)), with a selection of time steps along the horizontal axis and the cluster ids along the vertical one). We perform a hierarchical clustering at every selected time step. For the sake of uniformity and simplicity, we partition the members into at most three clusters, and track how the clustering solution varies over the selected time steps. We sample the time steps more densely at the beginning, e.g., in powers of two, because we noticed that the divergence counts and local clustering solutions vary more earlier in the integration (due, e.g., to changes in flow regime occurring early). The clustering variability is shown using branches connecting clusters at consecutive selected steps, the thickness of each branch depending on the corresponding number of members. Clusters are ordered so as to support id continuity from one step to another. Thus, we compute the number of common members for every pair of clusters at consecutive time steps, and assign cluster ids in decreasing order. To facilitate tracking the split and join events, bars starting from the same cluster have the same color: green, orange, and purple for the first, second, and third cluster, respectively.

This encoding reveals how the clustering solution varies from one time step to another, but not how the cluster memberships vary. To gain insight into this kind of variability, we determine, for the members in each branch, the cluster ids at the previous step and compute their

cardinalities. We then map the ratio of the largest cardinality to the number of members, to the opacity of each branch. The more opaque the branch is, the less the cluster componence has changed. To shed further light onto the clustering variability, users can select a branch to view the clustering evolution over all selected steps for the members in the selected branch (cf. Fig. 6).

6 RESULTS

We illustrate our framework on two ECMWF (European Centre for Medium-Range Weather Forecasts) ensembles of dimensions 101×41, each comprising a control run and another 50 members with perturbed initial conditions. Details on the system are available in [39].

6.1 Transport Variability Analysis

The first ensemble is a time-varying wind forecast at a pressure level of 850 hPa, initiated on October 15, 2012. The geolocated variability field (aggregated over the entire forecast interval for all members) is shown in Fig. 1(A). Locations where data was not available at this pressure level have been marked as "Not a Number (NaN)". We noticed that pathlines seeded over heterogeneous color regions (cf. Fig. 1(B)) exhibit dissimilar flow behavior, as opposed to those seeded over homogeneous areas. The detail view enables a comparative visual analysis of the pathlines, revealing the numerous changes in flow patterns and variability across the subdomain. In Fig. 1(B), for instance, trajectories seeded in the bottom line (to the left), where most pathlines go southwards, display aggregated divergence counts similar to the time step reference value (0.55). The variability increases to the right (to 0.65) as more particles go northeastwards, and decreases again further to the right, where most pathlines show no change of regime. In the next row (to the left), most particles follow a northeastward trend, but with various velocities and increasing dispersion, in spite of the geometrical similarity of the trajectories (values are around 0.7). The variability increases further (close to 0.8) in the third row, where the trajectories bifurcate geometrically as well.

The clustering solution of the ensemble is shown in Fig. 1(C) for the members with perturbed initial conditions. At first, we performed a hierarchical clustering of all members. However, because the control run was very similar to the majority of the other members, the large cluster forming around the dendogram hindered the identification of natural divisions in the hierarchical tree. Upon excluding the control run, we could detect natural groupings in the dendogram, which revealed an interesting pattern that should be further explored together with domain experts: ensemble member *i* was
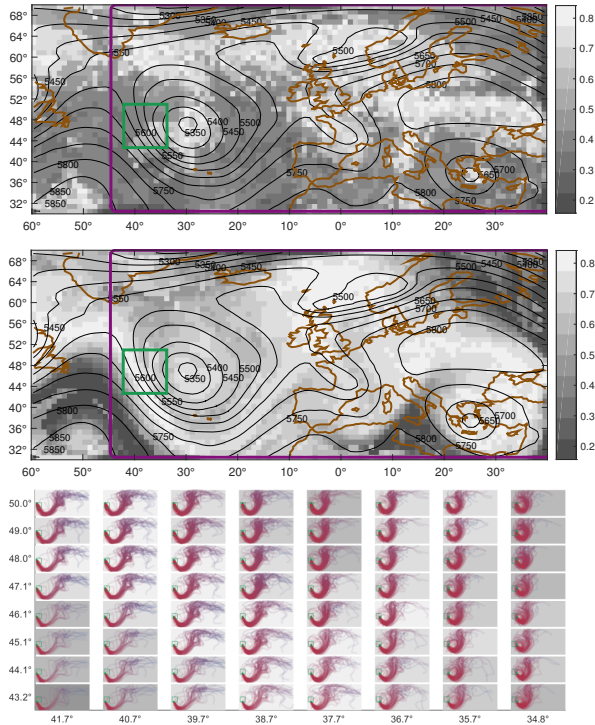
Figure 5: Spatial view for stationary ECMWF ensemble showing the variability at the (top) first time step and (middle) over the whole time interval. (Bottom) Detail view with downscaled streamline plots.

typically considerably more similar to ensemble members $i + 20$ and, if available, $i + 40$, than to other members. Notice that, in the dendogram, groupings of two or three members predominate at the low levels, larger clusters forming at distinctly higher levels.

The second ensemble is a stationary wind forecast ensemble at the same pressure level, valid on October 19, 2012. The geolocated variability for all members is displayed over the entire domain in Fig. 5, for the first time step (top) and aggregated over the whole forecast interval (middle). The isocontours of the geopotential height field of the control run are shown as black contours. Initially, regions of high variability are found mostly at the pressure centers, but their extent increases over time, as particles seeded in regions of lower variability enter regions of higher variability. This is also noticeable in the detail view (bottom), where, at all locations, most streamlines follow the southeastwards-northeastwards trend of the isocontours before they begin to spread out. In fact, the variability in this region remains relatively low for much longer than in other regions. In the end, the regions with low variability occupy much smaller extents, e.g., the low-frequency elongated region of high pressure in the bottom left corner.

The clustering solution for this ensemble considers all members. Fig. 6 (top) shows its time variability at seven selected time steps. The clustering solutions appear quite stable at the first two steps, but the variability in-
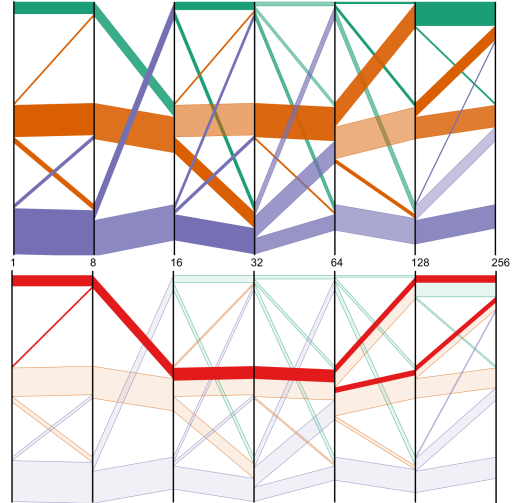

Figure 6: (Top) Clustering variability at selected time steps (cluster ids on the vertical axis, time steps on the horizontal axis). (Bottom) Membership variability for selected subcluster.

creases afterwards. The opacity of the branches gives a first insight into the cluster membership variability. For instance, the thickest of the orange branches, joining the second clusters at selected steps 16 and 32, is quite transparent, implying that the members in the branch are a mixture of those in the previous green and orange branches. The orange branch joining the second and third clusters is, however, opaque, as the componence of its members has not changed from the previous step. To gain further insight into the membership variability, users can select branches to see how the cluster memberships of their members vary at the selected steps. Fig. 6 (bottom) shows an example for the green branch joining the second cluster at step 16. Observe that, except for a temporary split towards the end, the members in the branch are always clustered together.

6.2 Implementation, Performance Analysis, and Scalability

We ran our tests on a standard desktop PC, equipped with an Intel i7-4790 quad-core processor running at 3.6 GHz and with 12 GB RAM. Fitting GMMs and computing the pairwise (dis)similarities using the Mahalanobis distance can be performed in parallel at every time step and every grid point. The operations can thus be parallelized in a straightforward way on the GPU, resulting in computation times of under five seconds for all datasets. As shown in the accompanying video, we are able to handle the comparative analysis session at interactive frame rates.

Our approach is scalable for ensembles larger than those used in our examples. Although the readability of the dendogram may suffer if the number of members is larger than a certain limit, 51 ensemble members, like in our case, is a typical number for meteorological

ensembles. Similarly, the readability of the parallel coordinates depends on the number of selected time steps and may require focus and context techniques, such as zooming and panning, which are part of our future work. An increase in the grid dimensions affects neither the two views for the clustering variability, nor the detail view, since the number of downscaled plots displayed is fixed, but only the spatial view, where the size of the pixel of each grid point decreases.

7 CONCLUSION AND EVALUATION

In this paper, we proposed a novel framework that enables an interactive comparative visual analysis of the transport variability of ensembles of 2D vector fields. We showed that our approach is able to determine the pairwise (dis)similarities of the ensemble members at every time step, without imposing artificial thresholds for the deviations. To achieve this, we computed the pairwise dissimilarities using the Mahalanobis distance on the Gaussian components identified by a GMM algorithm. Based on this fine-grained analysis, we proposed means to convey the variability of the spatio-temporal evolution of an ensemble and its clustering solution.

In developing our techniques, we collaborated closely with domain experts from meteorology, and provide herewith a summary of their informal feedback. The experts found the proposed methods useful in gaining a fast insight into the flow predictability over the domain and the time interval over which the forecasts can be trusted. They also described the techniques as useful in determining regions of different qualitative flow and highly appreciated our small-multiples approach that enabled them to observe the flow behavior at several locations concurrently. Nevertheless, they were initially puzzled about certain neighboring locations where the flow looked similar, but the variability values were quite different. Detailed inspections of such cases revealed the dissimilarities in flow behavior that had lead to the different variability values, although occasionally the differences were caused by a suboptimal partitioning of the GMM algorithm. The experts were also surprised by the spread of particle positions along geometrically similar trajectories and commended the enhanced spaghetti plots for bringing out the variability in the flow. Regarding the clustering solution, they were interested to further investigate the reason for the pattern present in the clustering solution for the time-varying data.

In the future, we plan to enrich the possibilities of exploring the variability of the clustering solution and the dynamics of the ensemble. We also intend to extend our analysis to 3D vector data. While the mathematical extension is straightforward, novel graphical abstractions are necessary to reveal the ensemble variability that reduce the clutter and occlusion problems inherent to 3D.

9 REFERENCES

[1] G.-P. Bonneau, H.-C. Hege, C. Johnson, M. Oliveira, K. Potter, P. Rheingans, and T. Schultz, "Overview and state-of-the-art of uncertainty visualization," in *Scientific Visualization*, pp. 3–27, Springer, 2014.

[2] K. Potter, P. Rosen, and C. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," in *Uncertainty Quantification in Scientific Computing*, pp. 226–249, 2012.

[3] J. Kehrer and H. Hauser, "Visualization and visual analysis of multi-faceted scientific data: A survey," *IEEE TVCG*, vol. 19, no. 3, pp. 495–513, 2013.

[4] A. Love, A. Pang, and D. Kao, "Visualizing spatial multivalue data," *IEEE Comput. Graph. Appl.*, vol. 25, no. 3, pp. 69–79, 2005.

[5] K. Potter *et al.*, "Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated ViSUS-CDAT systems," *J. Phys.: Conf. Ser.*, vol. 180, p. 012089, 2009.

[6] T. Pfaffelmoser, M. Mihai, and R. Westermann, "Visualizing the variability of gradients in uncertain 2D scalar fields," *IEEE TVCG*, vol. 19, pp. 1948–1961, Nov 2013.

[7] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson, "Visualizing summary statistics and uncertainty," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 823–832, 2010.

[8] C. Wittenbrink, A. Pang, and S. Lodha, "Glyphs for visualizing uncertainty in vector fields," *IEEE TVCG*, vol. 2, no. 3, pp. 266–279, 1996.

[9] S. K. Lodha, A. Pang, R. E. Sheehan, and C. M. Wittenbrink, "UFLOW: Visualizing uncertainty in fluid flow," in *Proc. IEEE Visualization*, pp. 249–254, 1996.

[10] M. Jarema, I. Demir, J. Kehrer, and R. Westermann, "Comparative visual analysis of vector field ensembles," in *Proc. IEEE VAST*, pp. 81–88, 2015.

[11] R. P. Botchen, D. Weiskopf, and T. Ertl, "Texture-based visualization of uncertainty in flow fields," in *Proc. IEEE Visualization*, pp. 647–654, IEEE, 2005.

[12] A. R. Sanderson, C. R. Johnson, and R. M. Kirby, "Display of vector fields using a reaction-diffusion model," in *Proc. IEEE Visualization*, pp. 115–122, 2004.

[13] R. Allendes Osorio and K. W. Brodlie, "Uncertain flow visualization using LIC," *Theory and Practice of Computer Graphics, Eurographics UK Chapter Proceedings*, pp. 215–222, 2009.

[14] M. Hlawatsch, P. Leube, W. Nowak, and D. Weiskopf, "Flow radar glyphs & static visualization of unsteady flow with uncertainty," *IEEE TVCG*, vol. 17, no. 12, pp. 1949–1958, 2011.

[15] M. Hlawatsch, F. Sadlo, H. Jang, and D. Weiskopf, "Pathline glyphs," *Comput. Graph. Forum*, vol. 33, pp. 497–506, May 2014.

[16] M. Otto, T. Germer, H.-C. Hege, and H. Theisel, "Uncertain 2D vector field topology," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 347–356, 2010.

[17] M. Otto, T. Germer, and H. Theisel, "Uncertain topology of 3D vector fields," in *PacificVis*, pp. 67–74, 2011.

[18] D. Schneider, J. Fuhrmann, W. Reich, and G. Scheuermann, "A variance based FTLE-like method for unsteady uncertain vector fields," in *Topological Methods in Data Analysis and Visualization II*, pp. 255–268, Springer, 2012.

[19] M. Hummel, H. Obermaier, C. Garth, and K. Joy, "Comparative visual analysis of Lagrangian transport in CFD ensembles," *IEEE TVCG*, vol. 19, no. 12, pp. 2743–2752, 2013.

[20] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[21] S. Bruckner and T. Möller, "Result-driven exploration of simulation parameter spaces for visual effects design," *IEEE TVCG*, vol. 16, no. 6, pp. 1468–1476, 2010.

[22] U. Bordoloi, D. Kao, and H.-W. Shen, "Visualization techniques for spatial probability density function data," *Data Sci. J.*, vol. 3, pp. 153–162, 2004.

[23] S. Oeltze, D. J. Lehmann, A. Kuhn, G. Janiga, H. Theisel, and B. Preim, "Blood flow clustering and applications invirtual stenting of intracranial aneurysms," *IEEE TVCG*, vol. 20, no. 5, pp. 686–701, 2014.

[24] F. Ferstl, K. Bürger, and R. Westermann, "Streamline variability plots for characterizing the uncertainty in vector field ensembles," *IEEE TVCG*,

[25] B. E. Hollister and A. Pang, "Visual Analysis of Transport Similarity in 2D CFD Ensembles," in *IS&T Electronic Imaging Conference on Visualization and Data Analysis*, pp. 1–11, 2016.

[26] C. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Proc. IEEE VAST*, pp. 51–58, 2009.

[27] B. E. Hollister and A. Pang, "Interpolation of non-Gaussian probability distributions for ensemble visualization," tech. rep., Jack Baskin School of Engineering, UC Santa Cruz, 2013.

[28] S. Liu, J. Levine, P. Bremer, and V. Pascucci, "Gaussian mixture model based volume visualization," in *Proc. LDAV*, pp. 73–77, Oct 2012.

[29] P. C. Mahalanobis, "On the generalised distance in statistics," *National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[30] T. Pfaffelmoser, M. Reitinger, and R. Westermann, "Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields," in *Comput. Graph. Forum*, vol. 30, pp. 951–960, 2011.

[31] M. Mihai and R. Westermann, "Visualizing the stability of critical points in uncertain scalar fields," *Computers& Graphics*, vol. 41, pp. 13 – 25, 2014.

[32] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates," *STAR Proceedings of Eurographics*, pp. 95–116, 2013.

[33] F. Bendix, R. Kosara, and H. Hauser, "Parallel sets: visual analysis of categorical data," in *IEEE Symposium on Information Visualization*, pp. 133–140, 2005.

[34] T. Nocke, M. Flechsig, and U. Böhm, "Visual exploration and evaluation of climate-related simulation data," in *Winter Simulation Conference*, pp. 703–711, 2007.

[35] V. Molchanov and L. Linsen, "Visual exploration of patterns in multi-run time-varying multi-field simulation data using projected views," in *Proc. WSCG*, vol. 21, pp. 49–58, 2014.

[36] H. Piringer, S. Pajer, W. Berger, and H. Teichmann, "Comparative visual analysis of 2D function ensembles," *Comput. Graph. Forum*, vol. 31, pp. 1195–1204, 2012.

[37] E. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[38] C. Ware, *Information visualization: Perception for design*. Elsevier, 3rd ed., 2013.

[39] M. Leutbecher and T. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008.

vol. 22, no. 1, pp. 767–776, 2016.