Statistical Parameter Selection for Clustering Persistence Diagrams

Max Kontak, Jules Vidal, Julien Tierny

Statistical Parameter Selection for Clustering Persistence Diagrams

Max Kontak
Simulation and Software Technology
DLR German Aerospace Center
Köln, Germany
max.kontak@dlr.de

Jules Vidal
CNRS LIP6
Sorbonne Universite
Paris, France
jules.vidal@sorbonne-universite.fr

Julien Tierny
CNRS LIP6
Sorbonne Universite
Paris, France
julien.tierny@sorbonne-universite.fr

***Abstract*—In urgent decision making applications, ensemble simulations are an important way to determine different outcome scenarios based on currently available data. In this paper, we will analyze the output of ensemble simulations by considering so-called persistence diagrams, which are reduced representations of the original data, motivated by the extraction of topological features. Based on a recently published progressive algorithm for the clustering of persistence diagrams, we determine the optimal number of clusters, and therefore the number of significantly different outcome scenarios, by the minimization of established statistical score functions. Furthermore, we present a proof-of-concept prototype implementation of the statistical selection of the number of clusters and provide the results of an experimental study, where this implementation has been applied to real-world ensemble data sets.**

***Index Terms*—urgent decision making, ensemble simulation, topological clustering, statistical model selection**

I. Introduction

To support urgent decision making in the situation of a catastrophic event, ensemble simulations can be used to quantify uncertainties and to distinguish different possible outcome scenarios, which may require diverse steps to be taken by a crisis manager. In practice, modern numerical simulations are subject to a variety of input parameters, related to the initial conditions of the system under study, as well as the configuration of its environment. Given recent advances in hardware computational power, engineers and scientists can now densely sample the space of these input parameters, in order to identify the most plausible crisis evolution. The European project VESTEC [1] focuses on building a toolchain combining interactive supercomputing, data analysis and visualization for the purpose of urgent decision making. Through the VESTEC system, a crisis manager would be able to run an ensemble of numerical simulations and interactively explore the resulting data in order to help the decision making process. Three use cases are to be supported: mosquito-borne diseases, wildfire monitoring, and space weather forecasting.

The identification of the possible scenarios can be accomplished by finding clusters in the simulation results. For instance, for a time-varying wildfire simulation, the outputs of all ensemble simulations for each time step could be clustered to obtain a time series of clusterings, which can then be further analyzed. In that way, one can determine points in time at which significantly different simulations arise in the ensemble (*e. g.*, there is only one fire vs. the fire has split up into multiple parts), which is relevant for the decision maker, who can compare the different clusters with the behavior of the fire in reality to identify the most plausible crisis evolution.

Unfortunately, the output data sets of large-scale simulation codes are often too big to all fit in memory, which creates a need for reduced data representations. These can be provided by *topological data analysis* [11], [38]. So-called *persistence diagrams* have been used in many applications before (combustion [7], [19], [26], fluid dynamics [8], [22], material sciences [14], [20], [28], chemistry [6], [17], and astrophysics [34], [36]) to obtain reduced data representations. Recently, an efficient technique has been proposed for clustering persistence diagrams instead of the original simulation data [41]. With the application of urgent decision making in mind, the algorithm has been designed based on the classical k-means clustering algorithm to incorporate time constraints. However, the number of clusters k is still a parameter of the approach in [41].

In this work, we will investigate so-called *information criteria*, which have been developed for statistical model selection [9], [24], to determine the optimal number of clusters. We will present a proof-of-concept prototype implementation for the statistical selection of parameters in topological clustering and perform an experimental study on real-life ensemble data sets.

II. Related work

Existing techniques for ensemble visualization and analysis typically construct, for each member of the ensemble, some geometrical objects, such as level sets or streamlines, which are used to cluster the original members of the ensemble and to construct aggregated views. Several methods have been proposed, such as spaghetti plots [10] for level-set variability in weather data ensembles [31], [32], or box-plots for the variability of contours [42] and curves in general [29].

Related to our work, clustering techniques have been used to analyze the main trends in ensembles of streamlines [15] and isocontours [16]. Favelier et al. [13] introduced an

This work is partially supported by the European Commission grant H2020-FETHPC-2017 "VESTEC" (ref. 800904).

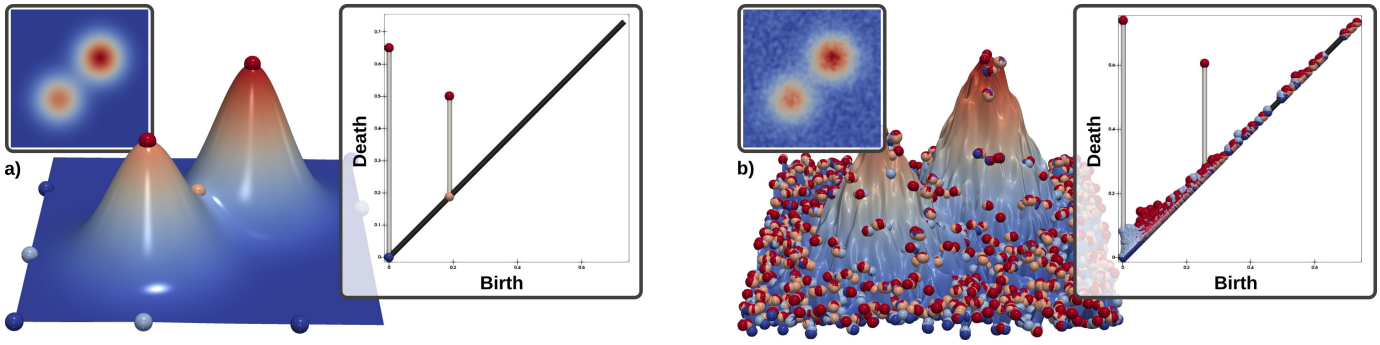


Fig. 1. The persistence diagram (right insets) reduces a data set (top left, bottom left: terrain view) to a 2D point cloud where each off-diagonal point represents a *topological feature*. In this diagram, the X and Y axes denote the *birth* and *death* of a topological feature, respectively. In these examples, points which stand out from the diagonal represent large features (the two hills, (a) and (b)), while points near the diagonal correspond to noisy features in the data.

approach, relying on spectral clustering, to analyze critical point variability in ensembles. Lacombe et al. [25] introduced an approach to cluster ensemble members based on their persistence diagrams. More relevant to our context of urgent decision making, Vidal et al. [41] introduced a method for the progressive clustering of persistence diagrams, supporting computation time constraints. However, this approach, which extends the seminal *k-means* algorithm [27], is subject to an input parameter, the number of output clusters k, which is often difficult to tune in practice.

III. Background

This section presents the technical background of our work.

A. Topological Data Analysis

Topological Data Analysis is a recent set of techniques [11], [38], which focus on structural data representations. We review in the following the main ingredients for the computation of topological signatures of data, for their comparison, and for their clustering. This section contains definitions taken from Vidal et al. [41], reproduced here for self-completeness.

a) Persistence diagrams: The input data is an ensemble of n piecewise linear (PL) scalar fields $f : \mathcal{M} \to \mathbb{R}$ defined on a PL d-manifold $\mathcal{M}$, with $d \leq 3$ in our applications. We note $f^{-1}_{-\infty}(w) = \{p \in \mathcal{M} \mid f(p) < w\}$ the *sub-level set* of f. When continuously increasing w, the topology of $f^{-1}_{-\infty}(w)$ can only change at specific locations, called the *critical points* of f. Critical points are classified according to their *index* $\mathcal{I}$: 0 for minima, 1 for 1-saddles, $d-1$ for $(d-1)$-saddles, and d for maxima.

Each topological feature of $f^{-1}_{-\infty}(w)$ (*e. g.*, connected components, independent cycles, voids) can be associated with a unique pair of critical points (c, c'), corresponding to its *birth* and *death*. Specifically, the Elder rule [11] states that critical points can be arranged according to this observation in a set of pairs, such that each critical point appears in only one pair (c, c') such that $f(c) < f(c')$ and $\mathcal{I}c = \mathcal{I}c' - 1$. Intuitively, this rule implies that if two topological features of $f^{-1}_{-\infty}(w)$ (*e. g.*, two connected components) meet at a critical point c', the *youngest* feature (*i. e.*, created last) *dies*, favoring the *oldest* one (*i. e.*, created first). Critical point pairs can be visually represented by the *persistence diagram*, noted $\mathcal{D}(f)$, which embeds each pair to a single point in the 2D plane at coordinates $\big(f(c), f(c')\big)$, which respectively correspond to the birth and death of the associated topological feature (Figure 1). The *persistence* of a pair, noted $\mathcal{P}(c, c')$, is then given by its height $f(c') - f(c)$. It describes the lifetime in the range of the corresponding topological feature.

b) Wasserstein distance between persistence diagrams: To cluster persistence diagrams, a first necessary ingredient is the notion of distance between them. Given two diagrams $\mathcal{D}(f)$ and $\mathcal{D}(g)$, a pointwise distance can be introduced in the 2D birth/death space between two points $a = (x_a, y_a) \in \mathcal{D}(f)$ and $b = (x_b, y_b) \in \mathcal{D}(g)$ by

$$d(a, b) = \big(|x_b - x_a|^2 + |y_b - y_a|^2\big)^{1/2} = \|a - b\|_2. \quad (1)$$

By convention, $d(a, b)$ is set to zero if both a and b exactly lie on the diagonal ($x_a = y_a$ and $x_b = y_b$). The *Wasserstein distance* between $\mathcal{D}(f)$ and $\mathcal{D}(g)$ can then be introduced as

$$W\big(\mathcal{D}(f), \mathcal{D}(g)\big) = \min_{\phi \in \Phi} \left(\sum_{a \in \mathcal{D}(f)} d\big(a, \phi(a)\big)^2 \right)^{1/2},$$

where Φ is the set of all possible assignments ϕ mapping each point $a \in \mathcal{D}(f)$ to a point $b \in \mathcal{D}(g)$, or to its projection onto the diagonal, $(\frac{x_a+y_a}{2}, \frac{x_a+y_a}{2})$, which denotes the removal of the corresponding feature from the assignment. The Wasserstein distance can be computed by solving an optimal assignment problem, for which efficient approximation algorithms exist [5], [23].

It can often be useful to geometrically lift the Wasserstein metric by also taking into account the geometrical layout of critical points [35]. Let (c, c') be the critical point pair corresponding to the point $a \in \mathcal{D}(f)$. Let $p^\lambda_a = \lambda c' + (1-\lambda)c \in \mathbb{R}^d$ be their linear combination with coefficient $\lambda \in [0, 1]$ in $\mathcal{M}$. Our experiments (section V) only deal with extrema, and we set λ to 0 for minima and 1 for maxima (to only consider the extremum's location). Then, the geometrically lifted pointwise distance $\widehat{d}(a, b)$ is given as $\widehat{d}(a, b) = \sqrt{(1-\alpha)d(a, b)^2 + \alpha||p^\lambda_a - p^\lambda_b||^2_2}$.

The parameter $\alpha \in [0,1]$ quantifies the importance given to the geometry of critical points and it must be tuned on a per application basis.

c) Fréchet mean of persistence diagrams: Once a distance metric is established between topological signatures, a second ingredient is needed, namely the notion of barycenter, in order to leverage typical clustering algorithms.

Let $\mathbb{D}$ be the space of persistence diagrams. The discrete *Wasserstein barycenter* of a set $\{\mathcal{D}(f_1), \mathcal{D}(f_2), \ldots, \mathcal{D}(f_n)\}$ of persistence diagrams can be introduced as the Fréchet mean of the set under the metric W. It is the diagram $\mathcal{D}^*$ that minimizes its distance to all the diagrams of the set (*i. e.*, the minimizer of the so-called Fréchet energy), that is, $\mathcal{D}^* = \arg\min_{\mathcal{D}\in\mathbb{D}} \sum_{i=1}^{n} W\big(\mathcal{D}, \mathcal{D}(f_i)\big)^2$. The computation of Wasserstein barycenters involves a computationally demanding optimization problem, for which the existence of at least one locally optimum solution has been shown by Turner et al. [40]. Efficient algorithms have been proposed to solve this optimization problem [25], including the progressive approach by Vidal et al. [41], which can return relevant approximations of Wasserstein barycenters, given some user defined time constraint $t_{\max}$, which is relevant for our urgent decision making context.

d) Topological clustering: Once barycenters between topological signatures can be computed, traditional clustering algorithms, such as the k-means [27], can be revisited to support topological data representations. Based on their efficient and progressive approach for Wasserstein barycenters, Vidal et al. [41] revisit the k-means algorithm as follows. The k-means is an iterative algorithm, where each *Clustering* iteration is composed itself of two sub-routines: *(i) Assignment* and *(ii) Update*. Initially, k cluster centroids $\mathcal{D}_j^*$ $(j = 1, \ldots, k)$ are initialized to k diagrams $\mathcal{D}(f_i)$ from the input set. Then, the *Assignment* step consists of assigning each diagram $\mathcal{D}(f_i)$ to its closest centroid $\mathcal{D}_{j(i)}^*$. This requires the computation of the Wasserstein distances W, of every diagram $\mathcal{D}(f_i)$ to all the centroids $\mathcal{D}_j^*$. Next, the *Update* step consists of updating each centroid's location by placing it at the Wasserstein barycenter of its assigned diagrams $\mathcal{D}(f_i)$. The algorithm continues these *Clustering* iterations until convergence, that is, until the assignments $i \mapsto j(i)$ between the diagrams and the k centroids do not evolve anymore. Since Wasserstein barycenters can be approximated under user-defined time constraints with Vidal's approach [41], the above algorithm also supports time constraints (see [41] for further details). Of course, a larger time constraint will, in general, result in a better clustering of the input set of persistence diagrams.

B. Statistical scores

The previously described method assumes that the number of clusters k is known *a priori*. If the number of clusters is not known in advance, so-called *information criteria* can be used to select a number of clusters *a posteriori* after the k-means algorithm has been applied for several values of k.

In our application, we will use the Akaike Information criterion (AIC, [2], [3]) and the Bayesian information criterion (BIC, [33]), which are based on the minimization of a score function of the form

$$\mathrm{IC}(k) = 2\,L(k) + p(k), \tag{2}$$

where $L(k)$ is the value of the log-likelihood function of the clustering result, when k clusters are detremined, and the term $p(k)$ penalizes the number of parameters differently for AIC and BIC. The criteria can be interpreted as a way to balance the goodness of fit (represented by the log-likelihood function) and the number of parameters: on the one hand, the goodness of fit is minimal if the number of clusters and the number of data points coincide, but the number of parameters is high in this situation. On the other hand, if the number of clusters is minimal, then the goodness of fit is, generally, very large. The minimum value of the information criterion will, consequently, be somewhere inbetween.

Under the so-called *identical spherical assumption* (see [30]), it can be shown (originally, for data from a Euclidean space) that the log-likelihood term has the form

$$L = \sum_{j=1}^{k} n_j \,\log n_j - n\,\log n - \frac{n\,d}{2}\,\log(2\,\pi\,\hat{\sigma}^2) - \frac{d}{2}\,(n-k), \tag{3}$$

where n_j is the number of diagrams mapped to the centroid $\mathcal{D}_j^*$, n is the total number of diagrams, d is the dimension of $\mathbb{D}$, and $\hat{\sigma}$ is an estimation of the in-cluster variance, for example,

$$\hat{\sigma}^2 = \frac{1}{d\,(n-k)} \sum_{i=1}^{n} W(\mathcal{D}(f_i), \mathcal{D}_{j(i)}^*)^2. \tag{4}$$

Since the dimension of $\mathbb{D}$ cannot be easily determined, we choose a value for d in our prototype implementation such that the information criteria show the expected behavior (approximately convex, being monotonically decreasing for small k and monotonically increasing for large k).

The penalty term p in (2) for the AIC is given by

$$p_{\mathrm{AIC}}(k) = 2\,k\,d,$$

whereas for the BIC it is given by

$$p_{\mathrm{BIC}}(k, N) = k\,d\,\log(N)$$

(cf. [12], Sect. 13.3), where the term $k\,d$ encodes the number of effective parameters of the statistical model, particularly, the d coordinates of the k cluster centers. Note that for a comparison of different clusterings of a fixed data set, p_{AIC} and p_{BIC} do indeed only depend on k, since both the dimension d of the underlying space as well as the number N of samples is constant.

IV. Prototype Implementation

This section details the implementation of our prototype.

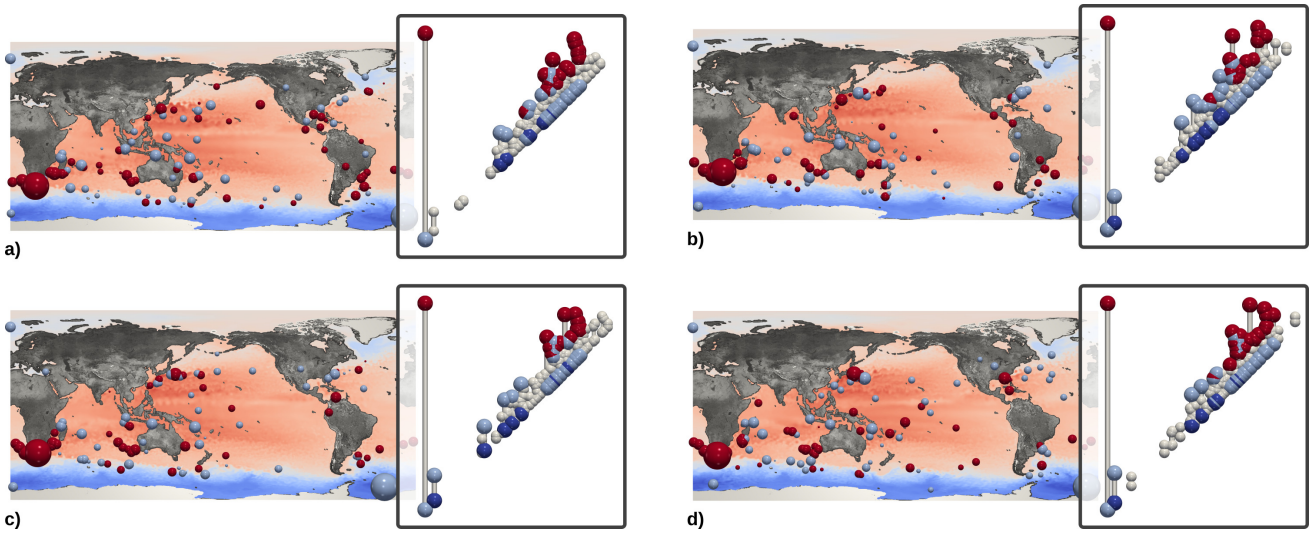


Fig. 2. Clusters automatically identified by our topological clustering approach ($t_{\max}$: 10 seconds) on the *Sea Surface Height* data-set. Left to right, top to bottom: pointwise mean of each cluster. Inset diagram: cluster centroid computed by the algorithm of Vidal et al. [41] (in the diagrams, the X and Y axes denote the *birth* and *death* of the topological features, respectively). Barycenter extrema are scaled in the domain by persistence and colored by critical index (spheres). In this example, the four clusters correspond to the four seasons.

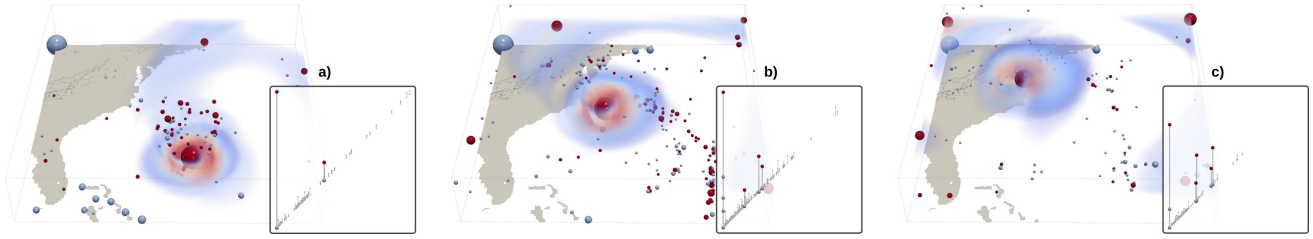


Fig. 3. Clusters automatically identified by our topological clustering approach ($t_{\max}$: 10 seconds) on the *Isabel* data-set. Left to right: pointwise mean of each cluster. Inset diagram: cluster centroid computed by the interruptible algorithm of Vidal et al. [41] (in the diagrams, the X and Y axes denote the *birth* and *death* of the topological features, respectively). Barycenter extrema are scaled in the domain by persistence and colored by critical index (spheres). In this example, the three clusters correspond to the three hurricane configurations (from left to right: formation, drift and landfall).

A. Topological clustering

For each ensemble data set, given a user time constraint $t_{\max}$, we systematically run the progressive topological clustering algorithm of Vidal et al. [41] for a range of k values (typically, 1 to 10). For this, we used the companion C++ implementation provided by Vidal et al. [41], available in the Topology ToolKit (TTK) [39]. Since the computation is independent for distinct k values, this step can be trivially parallelized (one k-clustering per process/thread).

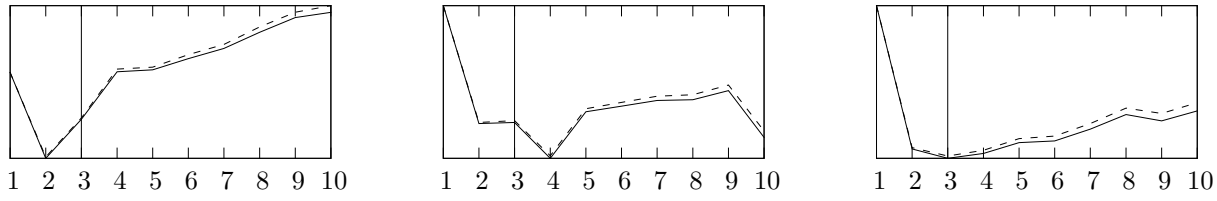
B. Statistical scores

Once the clustering has been performed for different values of k, the computation of the statistical scores (AIC and BIC) is straight-forward if the Wasserstein distances of each persistence diagram to its nearest centroid are extracted from the clustering process. Inserting these distances in (4) results in an estimation of the in-cluster variance, which can then be used in (3) to compute the value of the log-likelihood function. Combined with the computation of the penalty terms p_{AIC} and p_{BIC}, one obtains a value of the statistical score for the given clustering.

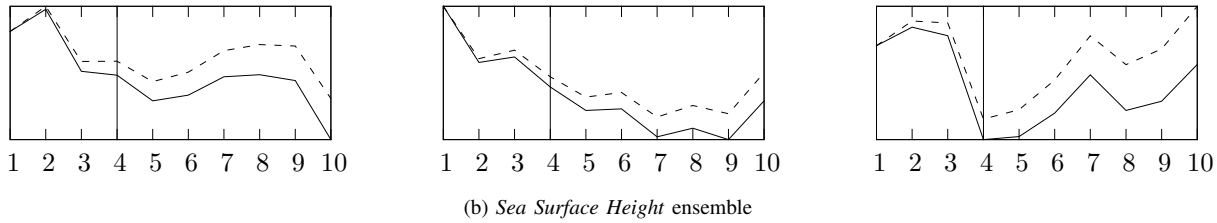
V. Results

This section presents experimental results obtained with a C++ implementation. The input persistence diagrams were computed with the algorithm by Gueunet et al. [18], which is available in the Topology ToolKit (TTK) [39].

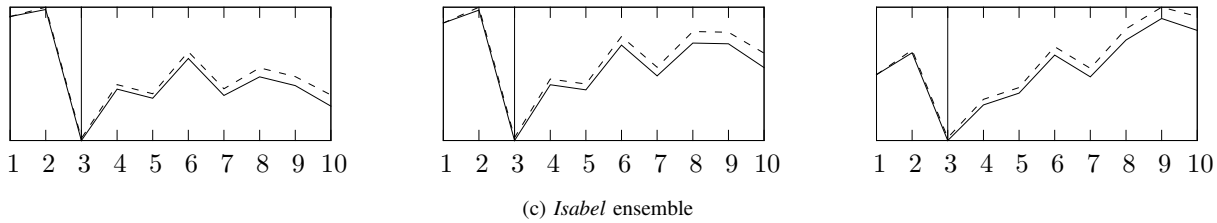
Our experiments were performed on a variety of simulated and acquired 2D and 3D ensembles, taken from Favelier et al. [13], following the experimental setup of Vidal et al. [41]. The *Gaussians* ensemble contains 100 2D synthetic noisy members, with 3 patterns of Gaussians. The *Sea Surface Height* ensemble (Figure 2) is composed of 48 observations taken in January, April, July and October 2012 (https://ecco.jpl.nasa.gov/products/all/). Here, the features of interest are the center of eddies, which can be reliably estimated with height extrema. Thus, both the diagrams involving the minima and maxima are considered and independently processed by our algorithms. Finally, the *Isabel* data set (Figure 3) is a volume ensemble of 12 members, showing key time steps (formation, drift and landfall) in the simulation of the Isabel hurricane [21]. In this example, the eyewall of the hurricane is typically characterized by high wind velocities,



(a) *Gaussians* ensemble



(b) *Sea Surface Height* ensemble



(c) *Isabel* ensemble

Fig. 4. Values of the AIC (solid line) and BIC (dashed line) for $k = 1, \ldots, 10$ for the three ensemble data sets for $t_{\max} = 1\,\mathrm{s}, 10\,\mathrm{s}, 100\,\mathrm{s}$ (left-to-right). The values have been normalized to the value for $k = 1$ for each diagram. Therefore the Y axes are not labeled. The X axes denote the number of clusters.

well captured by velocity maxima. Thus we only consider diagrams involving maxima. Unless stated otherwise, all results were obtained by considering the Wasserstein metric W based on the original pointwise metric in (1) without geometrical lifting (*i. e.*, $\alpha = 0$, subsection III-A).

In Figure 4, we have depicted the values of the statistical score functions for these three data sets for three different values of $t_{\max}$, where the number of clusters is characterized as the minimizer of the score functions. We have marked the number of clusters, determined with the most accurate clustering (that is, $t_{\max} = 100\,\mathrm{s}$) with a vertical line. We observe that for the less accurate clusterings ($t_{\max} = 1\,\mathrm{s}, 10\,\mathrm{s}$), we may obtain either a just slightly different number of clusters (Gaussians ensemble) or a number nearly doubling the optimal number of clusters (Sea Surface Height ensemble). This might seem to be a drawback with regard to the application of urgent decision making, where small values of $t_{\max}$ are desirable. However, in practice, when comparing the identified clusters with the crisis situation in reality to determine the most likely outcome, it is very helpful if the number of clusters is much lower than the number of ensemble simulations. This is still the case both for slightly different numbers of clusters and also for a twice as high number of clusters. Additionally, when determining the number of clusters in time-varying ensemble simulations, as described in the introduction, it is especially interesting if the number of clusters changes at a specific time step. We expect that these changes will also take place with the less accurate numbers of clusters. Of course, this will be analyzed in more detail in the future, when the presented method will be applied to data sets from the pilot applications used in the VESTEC project [1].

Figure 2 shows our results for the *Sea Surface Height* ensemble, where our statistical analysis estimates an optimal number of clusters of $k = 4$ and where the topological clustering [41] automatically identifies four clusters, corresponding to the four seasons: winter, spring, summer, fall (left-to-right, top-to-bottom). As shown in the insets, each season leads to a visually distinct centroid diagram.

As discussed by Vidal et al. [41], geometrical lifting is particularly important in applications where feature location bears a meaning, such as the Isabel ensemble (Figure 3). For this example, our statistical analysis estimates an optimal number of clusters of $k = 3$ and the clustering algorithm with geometrical lifting [41] automatically identifies the right clusters, corresponding to the three states of the hurricane (formation, drift and landfall).

VI. Conclusion

Motivated by urgent decision making applications, which require the clustering of ensemble simulation outputs for the determination of different crisis scenarios, we proposed a statistical technique to determine the number of clusters based on a recently published progressive clustering method for so-called persistence diagrams. We presented a proof-of-concept prototype implementation, which provided meaningful results for real-world ensemble data sets. In our upcoming research, we will incorporate the parameter selection within the clustering approach directly based on this prototype. Using Paraview Catalyst [4], [37], our approach can easily be

integrated into any simulation code. It then allows to carry out *in-situ* clustering operations on a statistically determined number of clusters at chosen iterations of the simulation, while respecting the time constraints of an urgent decision making situation. Furthermore, in the context of the European project VESTEC [1], we will apply our approach to other real-life use cases (wildfire, mosquito-borne diseases, space weather) and, especially, in an in-situ context to allow for an interaction of the decision maker with the ensemble simulations.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their thoughtful remarks and suggestions.

REFERENCES

[1] VESTEC EU Project. 2018-2021. https://vestec-project.eu.

[2] H. Akaike. Information theory and an extension of maximum likelihood principle. In B. Petrov and F. Csaki, eds., *Second International Symposium on Information Theory*, pp. 267–281, 1973.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19:716–723, 1973.

[4] U. Ayachit, A. Bauer, B. Geveci, P. O'Leary, K. Moreland, N. Fabian, and J. Mauldin. Paraview catalyst: Enabling in situ data analysis and visualization. In *Proceedings of the First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, ISAV2015, pp. 25–29. ACM, New York, NY, USA, 2015.

[5] D. P. Bertsekas. A new algorithm for the assignment problem. *Math. Program.*, 21:152–171, 1981.

[6] H. Bhatia, A. G. Gyulassy, V. Lordi, J. E. Pask, V. Pascucci, and P.-T. Bremer. TopoMS: Comprehensive topological exploration for molecular and condensed-matter systems. *J. Comput. Chem.*, 39:936–952, 2018.

[7] P. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, and J. Bell. Interactive exploration and analysis of large-scale simulations using topology-based data segmentation. *IEEE Trans. Vis. Comput. Gr.*, 17:1307–1324, 2011.

[8] F. Chen, H. Obermaier, H. Hagen, B. Hamann, J. Tierny, and V. Pascucci. Topology analysis of time-dependent multi-fluid data using the reeb graph. *Comput. Aided Geom. D.*, 30:557–566, 2013.

[9] G. Claeskens and N. Hjort. *Model selection and model averaging*. Cambridge University Press, 2008.

[10] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *The Analysis of Longitudinal Data*. Oxford University Press, 2002.

[11] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.

[12] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2016.

[13] G. Favelier, N. Faraj, B. Summa, and J. Tierny. Persistence Atlas for Critical Point Variability in Ensembles. *IEEE Trans. Vis. Comput. Gr.*, 25:1152–1162, 2018.

[14] G. Favelier, C. Gueunet, and J. Tierny. Visualizing ensembles of viscous fingers. In *IEEE SciVis Contest*, 2016.

[15] F. Ferstl, K. Bürger, and R. Westermann. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Trans. Vis. Comput. Gr.*, 22:767–776, 2016.

[16] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann. Visual analysis of spatial variability and global correlations in ensembles of iso-contours. *Comput. Graph. Forum*, 35:221–230, 2016.

[17] D. Guenther, R. Alvarez-Boto, J. Contreras-Garcia, J.-P. Piquemal, and J. Tierny. Characterizing molecular interactions in chemical systems. *IEEE Trans. Vis. Comput. Gr.*, 20:2476–2485, 2014.

[18] C. Gueunet, P. Fortin, J. Jomier, and J. Tierny. Task-based augmented contour trees with fibonacci heaps. *IEEE Trans. Parall. Distr.*, 30:1887–1905, 2019.

[19] A. Gyulassy, P. Bremer, R. Grout, H. Kolla, J. Chen, and V. Pascucci. Stability of dissipation elements: A case study in combustion. *Comput. Graph. Forum*, 33:51–60, 2014.

[20] A. Gyulassy, V. Natarajan, M. Duchaineau, V. Pascucci, E. Bringa, A. Higginbotham, and B. Hamann. Topologically clean distance fields. *IEEE Trans. Vis. Comput. Gr.*, 13:1432–1439, 2007.

[21] IEEE SciVisContest. Simulation of the Isabel hurricane. http://sciviscontest-staging.ieeevis.org/2004/data.html, 2004.

[22] J. Kasten, J. Reininghaus, I. Hotz, and H. Hege. Two-dimensional time-dependent vortex regions based on the acceleration magnitude. *IEEE Trans. Vis. Comput. Gr.*, 17:2080–2087, 2011.

[23] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. *ACM Journal of Experimental Algorithmics*, 22, 2016. Article No. 1.4.

[24] S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer, 2007.

[25] T. Lacombe, M. Cuturi, and S. Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2018.

[26] D. E. Laney, P. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Trans. Vis. Comput. Gr.*, 12:1053–1060, 2006.

[27] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28:129–137, 1982.

[28] J. Lukasczyk, G. Aldrich, M. Steptoe, G. Favelier, C. Gueunet, J. Tierny, R. Maciejewski, B. Hamann, and H. Leitte. Viscous fingering: A topological visual analytic approach. *Appl Mech. Mater.*, 869:9–19, 2017.

[29] M. Mirzargar, R. Whitaker, and R. Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Trans. Vis. Comput. Graph.*, 20:2654–2663, 2014.

[30] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734, 2000.

[31] K. Potter, A. Wilson, P. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-Vis: A framework for the statistical visualization of ensemble data. In *2009 IEEE International Conference on Data Mining Workshops*, 2009.

[32] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Trans. on Vis. Comput. Graph.*, 16:1421–1430, 2010.

[33] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 5:461–464, 1978.

[34] N. Shivashankar, P. Pranav, V. Natarajan, R. van de Weygaert, E. P. Bos, and S. Rieder. Felix: A topology based framework for visual exploration of cosmic filaments. *IEEE Trans. Vis. Comput. Gr.*, 22:1745–1759, 2016. http://vgl.serc.iisc.ernet.in/felix/index.html.

[35] M. Soler, M. Plainchault, B. Conche, and J. Tierny. Lifted Wasserstein Matcher for Fast and Robust Topology Tracking. In *IEEE Symposium on Large Data Analysis and Visualization*, 2018.

[36] T. Sousbie. The persistent cosmic web and its filamentary structure: Theory and implementations. *Mon. Not. R. Astron. Soc.*, 414:350–383, 2011. http://www2.iap.fr/users/sousbie/web/html/indexd41d.html.

[37] The Topology ToolKit usage tutorial. TTK in-situ with Catalyst. https://topology-tool-kit.github.io/catalyst.html, 2019.

[38] J. Tierny. *Topological Data Analysis for Scientific Visualization*. Springer, 2018.

[39] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE Trans. Vis. Comput. Gr.*, 24:832–842, 2017. https://topology-tool-kit.github.io/.

[40] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet Means for Distributions of Persistence Diagrams. *Discrete Comput. Geom.*, 52:44–70, 2014.

[41] J. Vidal, J. Budin, and J. Tierny. Progressive Wasserstein Barycenters of Persistence Diagrams. *IEEE Trans. Vis. Comput. Gr.*, 2019. accepted for publication.

[42] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Trans. Vis. Comput. Gr.*, 19:2713–2722, 2013.