Answer to reviewer 2

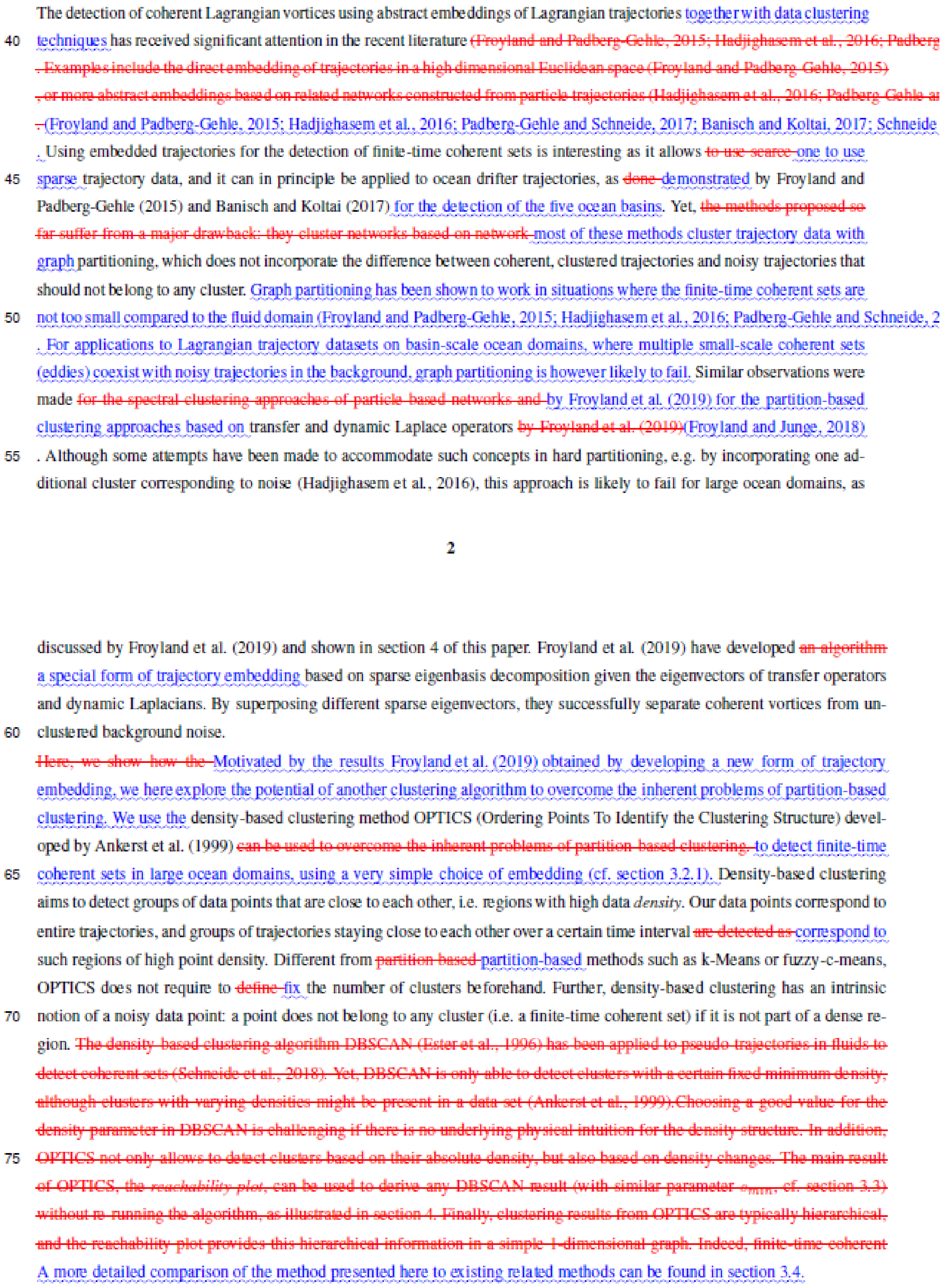
General answer:
We thank the reviewer for the critical comments, and in particular for the detailed analysis of other methods and their comparison to our approach. We agree with most points raised by the reviewer. We have made major adaptations to the formulations in the revised version, and explain the relation of our method to existing studies in more detail.

Please note:
The images in this file are excerpts of the revised version in latexdiff. Please apologize the formatting problems of latexdiff that cuts off references at line breaks. This is not the case in the revised version.

Comment 1
There are already several clustering methods in the literature for finding finite-time coherent sets, including a density-based clustering DBSCAN by Schneide-etal'18, which is a special case of the OPTICS approach in the manuscript. The idea of a hierarchy of finite-time coherent sets has been considered by Ma/Bollt'13. The paper Fr/Sa/Ro'19 develops a robust method to classify only those sets are that coherent, not fully partitioning the domain. In Fr/Sa/Ro'19, coherent sets at different spatial scales are also considered, similar to a hierarchy. Fr/Sa/Ro'19 also considers the Bickley jet and ocean eddies, with ocean eddies listed as a motivation in Fr/Sa/Ro'19 for developing a non-partitioning approach. Not limited to the work above, I would say there is some "upselling" of the novelty in the manuscript, and that prior work is occasionally omitted, mischaracterized, or overly criticized.

Answer to comment 1
Thank you for this comment. We did not intend to upsell our work, or omit, mischaracterize or overly criticize existing work. In fact, our work has been majorly motivated by the paper of Froyland et al. 2019. But we understand that the original manuscript appeared to do so, and we thank the reviewer to making this clear to us. We have made the following changes in the new version.

1. We mainly removed the discussion of other methods in the introduction and moved it to a separate section. In the introduction, we emphasize that our work is majorly inspired by Froyland et al. 2019. We are also more specific about the actual problem at hand, i.e. the detection of many small scale coherent sets in large-scale, noisy ocean flows.



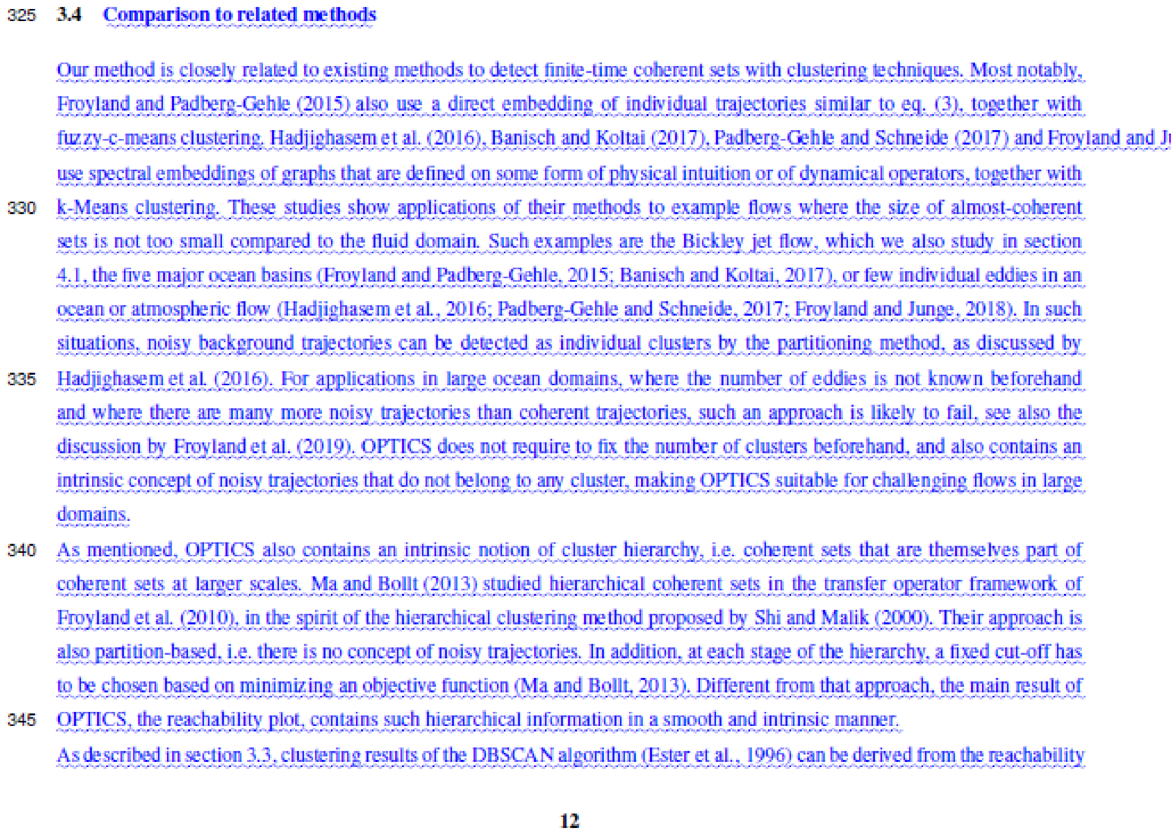
The detection of coherent Lagrangian vortices using abstract embeddings of Lagrangian trajectories together with data clustering techniques has received significant attention in the recent literature ~~(Froyland and Padberg-Gehle, 2015; Hadjighasem et al., 2016; Padberg~~ ~~. Examples include the direct embedding of trajectories in a high dimensional Euclidean space (Froyland and Padberg-Gehle, 2015)~~ ~~, or more abstract embeddings based on related networks constructed from particle trajectories (Hadjighasem et al., 2016; Padberg-Gehle ar~~ ~~.~~(Froyland and Padberg-Gehle, 2015; Hadjighasem et al., 2016; Padberg-Gehle and Schneide, 2017; Banisch and Koltai, 2017; Schneide . Using embedded trajectories for the detection of finite-time coherent sets is interesting as it allows ~~to use scarce~~ one to use sparse trajectory data, and it can in principle be applied to ocean drifter trajectories, as ~~done~~ demonstrated by Froyland and Padberg-Gehle (2015) and Banisch and Koltai (2017) for the detection of the five ocean basins. Yet, ~~the methods proposed so far suffer from a major drawback: they cluster networks based on network~~ most of these methods cluster trajectory data with graph partitioning, which does not incorporate the difference between coherent, clustered trajectories and noisy trajectories that should not belong to any cluster. Graph partitioning has been shown to work in situations where the finite-time coherent sets are not too small compared to the fluid domain (Froyland and Padberg-Gehle, 2015; Hadjighasem et al., 2016; Padberg-Gehle and Schneide, 2 . For applications to Lagrangian trajectory datasets on basin-scale ocean domains, where multiple small-scale coherent sets (eddies) coexist with noisy trajectories in the background, graph partitioning is however likely to fail. Similar observations were made ~~for the spectral clustering approaches of particle based networks and~~ by Froyland et al. (2019) for the partition-based clustering approaches based on transfer and dynamic Laplace operators ~~by Froyland et al. (2019)~~(Froyland and Junge, 2018) . Although some attempts have been made to accommodate such concepts in hard partitioning, e.g. by incorporating one additional cluster corresponding to noise (Hadjighasem et al., 2016), this approach is likely to fail for large ocean domains, as

discussed by Froyland et al. (2019) and shown in section 4 of this paper. Froyland et al. (2019) have developed ~~an algorithm~~ a special form of trajectory embedding based on sparse eigenbasis decomposition given the eigenvectors of transfer operators and dynamic Laplacians. By superposing different sparse eigenvectors, they successfully separate coherent vortices from unclustered background noise.

~~Here, we show how the~~ Motivated by the results Froyland et al. (2019) obtained by developing a new form of trajectory embedding, we here explore the potential of another clustering algorithm to overcome the inherent problems of partition-based clustering. We use the density-based clustering method OPTICS (Ordering Points To Identify the Clustering Structure) developed by Ankerst et al. (1999) ~~can be used to overcome the inherent problems of partition-based clustering.~~ to detect finite-time coherent sets in large ocean domains, using a very simple choice of embedding (cf. section 3.2.1). Density-based clustering aims to detect groups of data points that are close to each other, i.e. regions with high data *density*. Our data points correspond to entire trajectories, and groups of trajectories staying close to each other over a certain time interval ~~are detected as~~ correspond to such regions of high point density. Different from ~~partition based~~ partition-based methods such as k-Means or fuzzy-c-means, OPTICS does not require to ~~define~~ fix the number of clusters beforehand. Further, density-based clustering has an intrinsic notion of a noisy data point: a point does not belong to any cluster (i.e. a finite-time coherent set) if it is not part of a dense region. ~~The density-based clustering algorithm DBSCAN (Ester et al., 1996) has been applied to pseudo trajectories in fluids to detect coherent sets (Schneide et al., 2018). Yet, DBSCAN is only able to detect clusters with a certain fixed minimum density, although clusters with varying densities might be present in a data set (Ankerst et al., 1999).Choosing a good value for the density parameter in DBSCAN is challenging if there is no underlying physical intuition for the density structure. In addition, OPTICS not only allows to detect clusters based on their absolute density, but also based on density changes. The main result of OPTICS, the *reachability plot*, can be used to derive any DBSCAN result (with similar parameter ϵ_{min}, cf. section 3.3) without re-running the algorithm, as illustrated in section 4. Finally, clustering results from OPTICS are typically hierarchical, and the reachability plot provides this hierarchical information in a simple 1-dimensional graph. Indeed, finite-time coherent~~ A more detailed comparison of the method presented here to existing related methods can be found in section 3.4.

2. We have added an additional section to compare our method to existing approaches. There, we stress what our contribution is compared to Froyland et al. 2019: the study of an improved clustering step, instead of an improved embedding step. We also mention the downside of our method compared to Froyland and Junge (2018) and Froyland et al. (2019). Note that the hierarchical method of Ma and Bollt (2013) is powerful, but it is partition-

based and not intrinsic to the clustering algorithm, as there is a cut-off chosen at each step of the hierarchical clustering. Also, note that many of the existing methods that use k-Means did work for examples where the coherent sets are not very small compared to the fluid domain. Finally, DBSCAN has been used by Schneide et al. (2018), but not to derive explicit clustering results, and also not in the ocean context. We explain this in this new section.

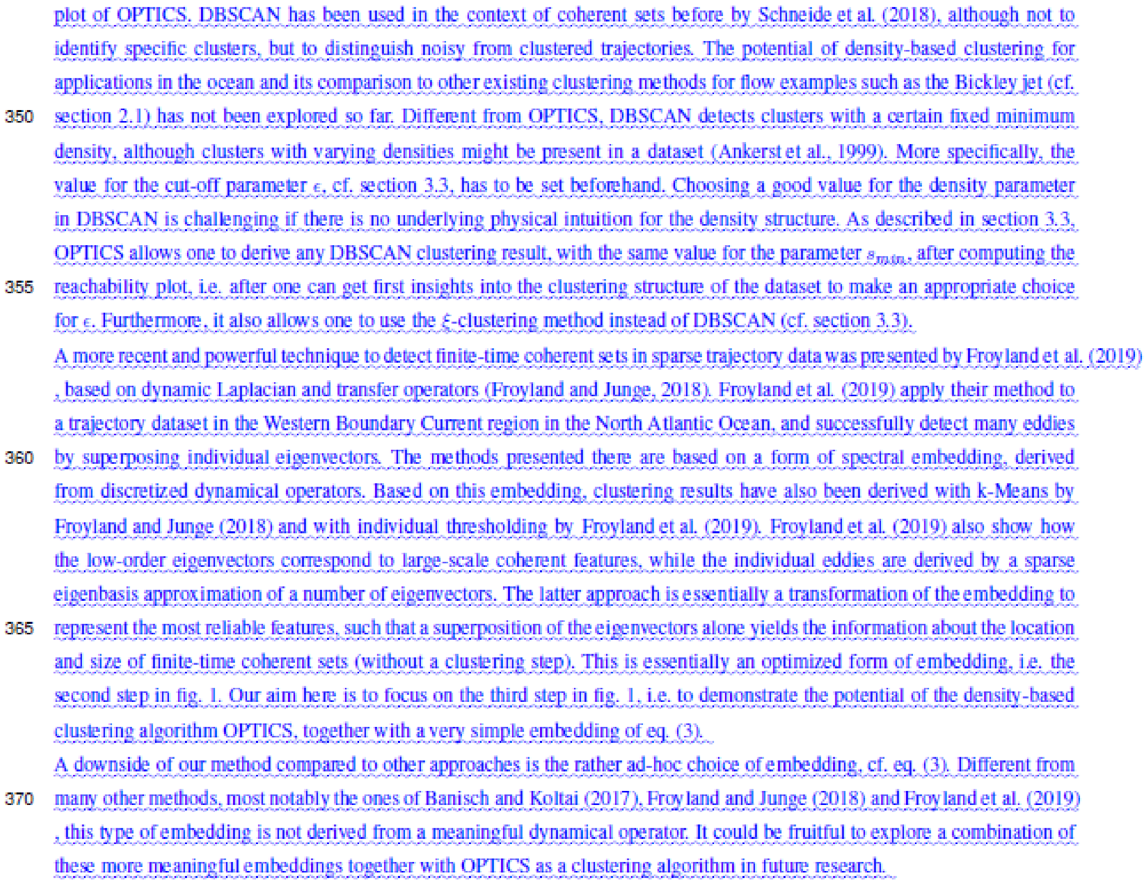


3.4 Comparison to related methods

Our method is closely related to existing methods to detect finite-time coherent sets with clustering techniques. Most notably, Froyland and Padberg-Gehle (2015) also use a direct embedding of individual trajectories similar to eq. (3), together with fuzzy-c-means clustering. Hadjighasem et al. (2016), Banisch and Koltai (2017), Padberg-Gehle and Schneide (2017) and Froyland and J use spectral embeddings of graphs that are defined on some form of physical intuition or of dynamical operators, together with k-Means clustering. These studies show applications of their methods to example flows where the size of almost-coherent sets is not too small compared to the fluid domain. Such examples are the Bickley jet flow, which we also study in section 4.1, the five major ocean basins (Froyland and Padberg-Gehle, 2015; Banisch and Koltai, 2017), or few individual eddies in an ocean or atmospheric flow (Hadjighasem et al., 2016; Padberg-Gehle and Schneide, 2017; Froyland and Junge, 2018). In such situations, noisy background trajectories can be detected as individual clusters by the partitioning method, as discussed by Hadjighasem et al. (2016). For applications in large ocean domains, where the number of eddies is not known beforehand and where there are many more noisy trajectories than coherent trajectories, such an approach is likely to fail, see also the discussion by Froyland et al. (2019). OPTICS does not require to fix the number of clusters beforehand, and also contains an intrinsic concept of noisy trajectories that do not belong to any cluster, making OPTICS suitable for challenging flows in large domains.

As mentioned, OPTICS also contains an intrinsic notion of cluster hierarchy, i.e. coherent sets that are themselves part of coherent sets at larger scales. Ma and Bollt (2013) studied hierarchical coherent sets in the transfer operator framework of Froyland et al. (2010), in the spirit of the hierarchical clustering method proposed by Shi and Malik (2000). Their approach is also partition-based, i.e. there is no concept of noisy trajectories. In addition, at each stage of the hierarchy, a fixed cut-off has to be chosen based on minimizing an objective function (Ma and Bollt, 2013). Different from that approach, the main result of OPTICS, the reachability plot, contains such hierarchical information in a smooth and intrinsic manner.

As described in section 3.3, clustering results of the DBSCAN algorithm (Ester et al., 1996) can be derived from the reachability



plot of OPTICS. DBSCAN has been used in the context of coherent sets before by Schneide et al. (2018), although not to identify specific clusters, but to distinguish noisy from clustered trajectories. The potential of density-based clustering for applications in the ocean and its comparison to other existing clustering methods for flow examples such as the Bickley jet (cf. section 2.1) has not been explored so far. Different from OPTICS, DBSCAN detects clusters with a certain fixed minimum density, although clusters with varying densities might be present in a dataset (Ankerst et al., 1999). More specifically, the value for the cut-off parameter ϵ, cf. section 3.3, has to be set beforehand. Choosing a good value for the density parameter in DBSCAN is challenging if there is no underlying physical intuition for the density structure. As described in section 3.3, OPTICS allows one to derive any DBSCAN clustering result, with the same value for the parameter s_{min}, after computing the reachability plot, i.e. after one can get first insights into the clustering structure of the dataset to make an appropriate choice for ϵ. Furthermore, it also allows one to use the ξ-clustering method instead of DBSCAN (cf. section 3.3).

A more recent and powerful technique to detect finite-time coherent sets in sparse trajectory data was presented by Froyland et al. (2019) , based on dynamic Laplacian and transfer operators (Froyland and Junge, 2018). Froyland et al. (2019) apply their method to a trajectory dataset in the Western Boundary Current region in the North Atlantic Ocean, and successfully detect many eddies by superposing individual eigenvectors. The methods presented there are based on a form of spectral embedding, derived from discretized dynamical operators. Based on this embedding, clustering results have also been derived with k-Means by Froyland and Junge (2018) and with individual thresholding by Froyland et al. (2019). Froyland et al. (2019) also show how the low-order eigenvectors correspond to large-scale coherent features, while the individual eddies are derived by a sparse eigenbasis approximation of a number of eigenvectors. The latter approach is essentially a transformation of the embedding to represent the most reliable features, such that a superposition of the eigenvectors alone yields the information about the location and size of finite-time coherent sets (without a clustering step). This is essentially an optimized form of embedding, i.e. the second step in fig. 1. Our aim here is to focus on the third step in fig. 1, i.e. to demonstrate the potential of the density-based clustering algorithm OPTICS, together with a very simple embedding of eq. (3).

A downside of our method compared to other approaches is the rather ad-hoc choice of embedding, cf. eq. (3). Different from many other methods, most notably the ones of Banisch and Koltai (2017), Froyland and Junge (2018) and Froyland et al. (2019) , this type of embedding is not derived from a meaningful dynamical operator. It could be fruitful to explore a combination of these more meaningful embeddings together with OPTICS as a clustering algorithm in future research.

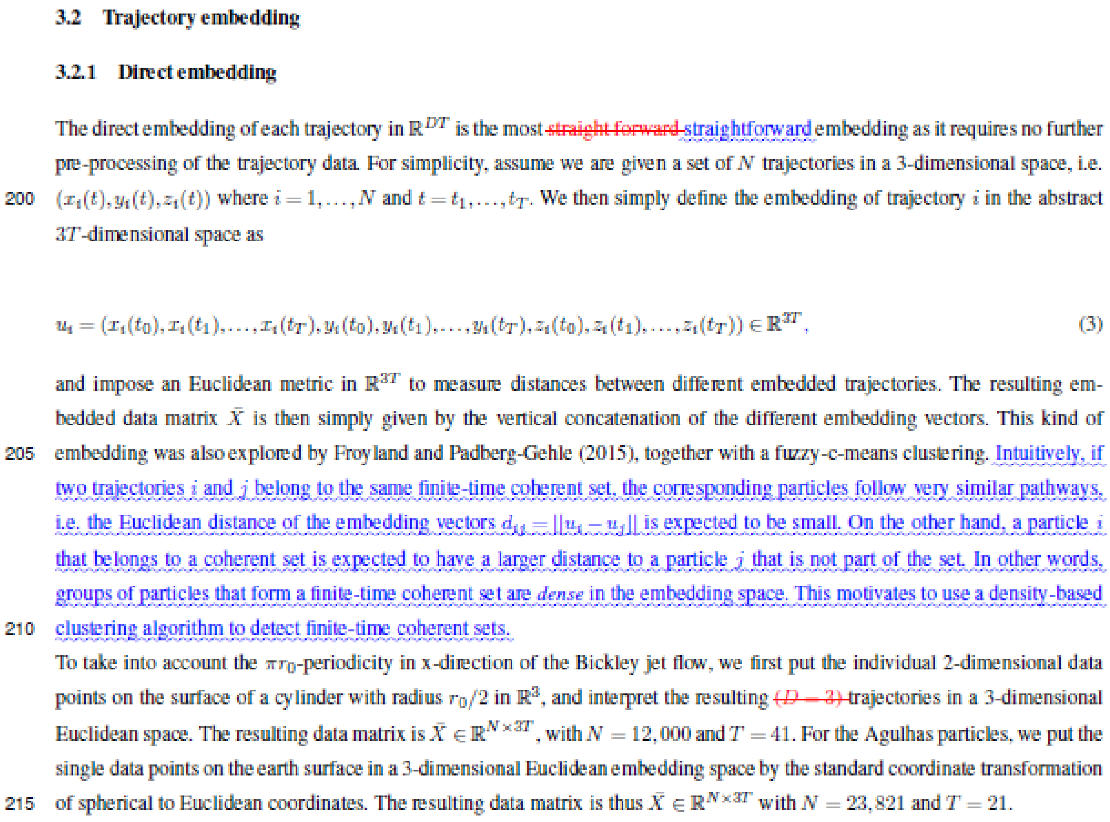
Comment 2

A positive aspect is that the (standard) "DBSCAN" and "\xi" clustering outputs of the OPTICS clustering could provide potentially useful hierarchical information, and to my knowledge this is a new way of analyzing the dynamics. Unfortunately, this is not explored much, and the authors do not provide an intuitive explanation of what the "DBSCAN" and "\xi" clustering algorithms are actually doing in their dynamical context. It would be beneficial for the authors to link the algorithms more with the dynamical inputs (trajectories) and the dynamical problem being solved. As this is the main contribution of the paper, I think this needs to be expanded much more. The reasons behind the choices of which clustering algorithm is applied to the different datasets should also be explained.

Answer to comment 2

Thank you for this comment. We were indeed lacking some form of intuition behind the two clustering methods and their application. We have made the following changes.

1. More explanation about the embedding and why the embedded trajectories create a signal in terms of data density.



3.2 Trajectory embedding

3.2.1 Direct embedding

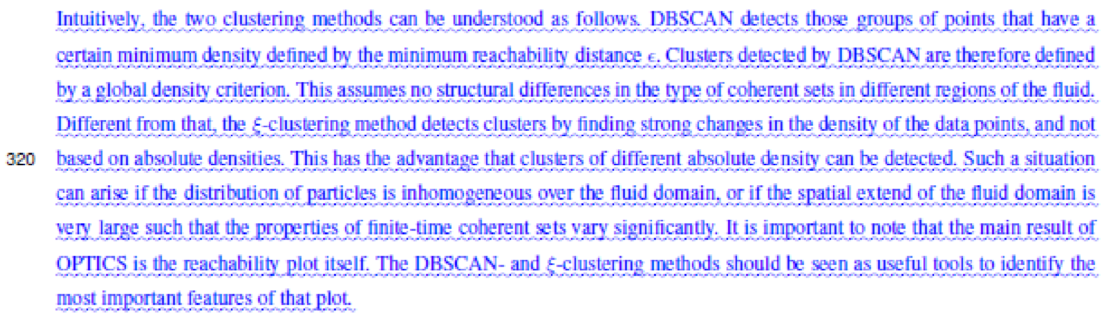
The direct embedding of each trajectory in $\mathbb{R}^{DT}$ is the most ~~straight forward~~ straightforward embedding as it requires no further pre-processing of the trajectory data. For simplicity, assume we are given a set of N trajectories in a 3-dimensional space, i.e. $(x_i(t), y_i(t), z_i(t))$ where $i = 1, \ldots, N$ and $t = t_1, \ldots, t_T$. We then simply define the embedding of trajectory i in the abstract $3T$-dimensional space as

$$u_i = (x_i(t_0), x_i(t_1), \ldots, x_i(t_T), y_i(t_0), y_i(t_1), \ldots, y_i(t_T), z_i(t_0), z_i(t_1), \ldots, z_i(t_T)) \in \mathbb{R}^{3T}, \tag{3}$$

and impose an Euclidean metric in $\mathbb{R}^{3T}$ to measure distances between different embedded trajectories. The resulting embedded data matrix $\bar{X}$ is then simply given by the vertical concatenation of the different embedding vectors. This kind of embedding was also explored by Froyland and Padberg-Gehle (2015), together with a fuzzy-c-means clustering. Intuitively, if two trajectories i and j belong to the same finite-time coherent set, the corresponding particles follow very similar pathways, i.e. the Euclidean distance of the embedding vectors $d_{ij} = ||u_i - u_j||$ is expected to be small. On the other hand, a particle i that belongs to a coherent set is expected to have a larger distance to a particle j that is not part of the set. In other words, groups of particles that form a finite-time coherent set are *dense* in the embedding space. This motivates to use a density-based clustering algorithm to detect finite-time coherent sets.

To take into account the πr_0-periodicity in x-direction of the Bickley jet flow, we first put the individual 2-dimensional data points on the surface of a cylinder with radius $r_0/2$ in $\mathbb{R}^3$, and interpret the resulting ~~$(D-3)$~~ trajectories in a 3-dimensional Euclidean space. The resulting data matrix is $\bar{X} \in \mathbb{R}^{N \times 3T}$, with $N = 12,000$ and $T = 41$. For the Agulhas particles, we put the single data points on the earth surface in a 3-dimensional Euclidean embedding space by the standard coordinate transformation of spherical to Euclidean coordinates. The resulting data matrix is thus $\bar{X} \in \mathbb{R}^{N \times 3T}$ with $N = 23,821$ and $T = 21$.

2. An intuitive explanation of the two clustering methods and their major properties.



Intuitively, the two clustering methods can be understood as follows. DBSCAN detects those groups of points that have a certain minimum density defined by the minimum reachability distance ϵ. Clusters detected by DBSCAN are therefore defined by a global density criterion. This assumes no structural differences in the type of coherent sets in different regions of the fluid. Different from that, the ξ-clustering method detects clusters by finding strong changes in the density of the data points, and not based on absolute densities. This has the advantage that clusters of different absolute density can be detected. Such a situation can arise if the distribution of particles is inhomogeneous over the fluid domain, or if the spatial extend of the fluid domain is very large such that the properties of finite-time coherent sets vary significantly. It is important to note that the main result of OPTICS is the reachability plot itself. The DBSCAN- and ξ-clustering methods should be seen as useful tools to identify the most important features of that plot.

3. We have included a DBSCAN clustering result in the main figure of the Agulhas flow example, and discuss the differences between xi and DBSCAN clustering.

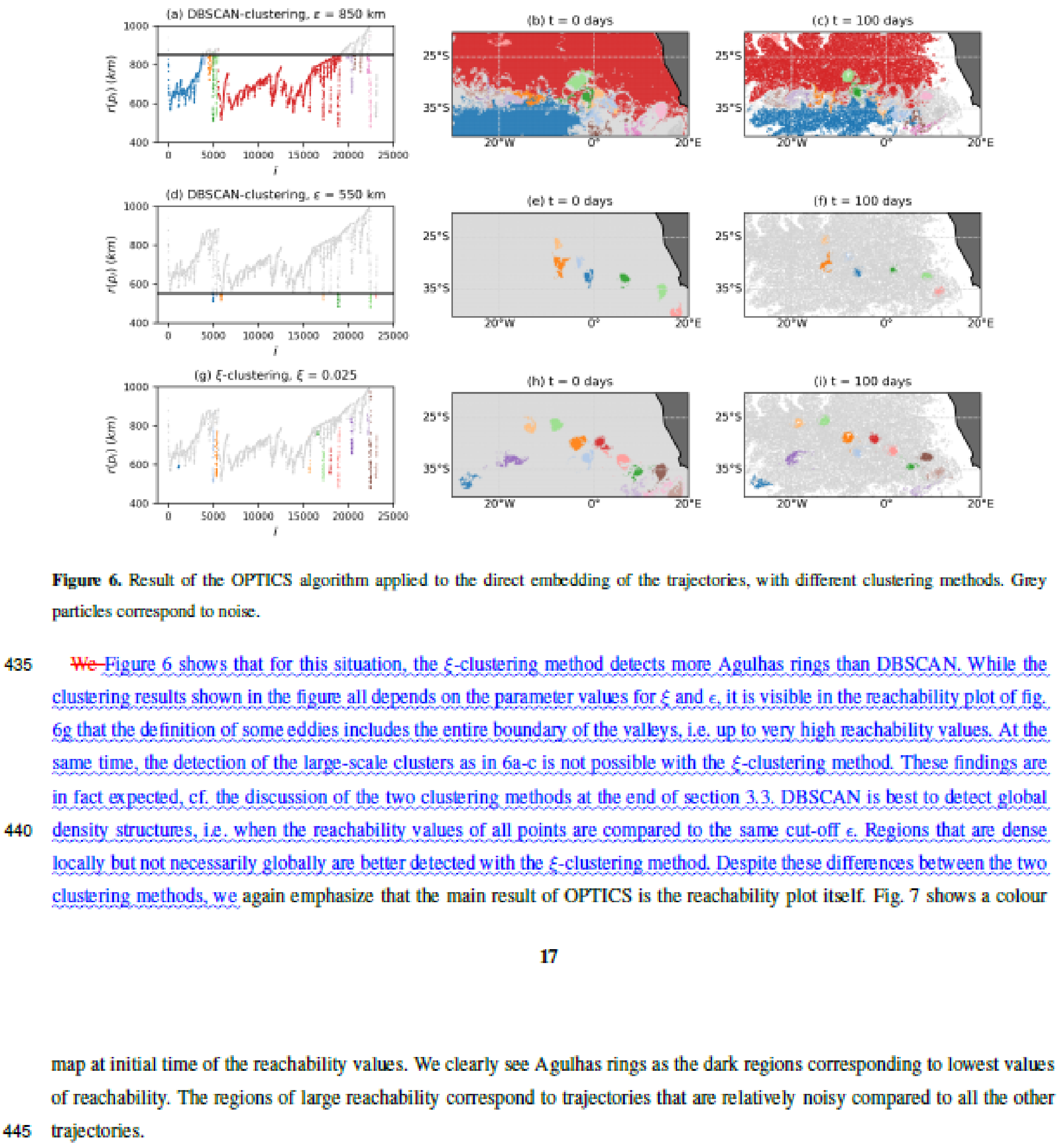

(a) DBSCAN-clustering, ε = 850 km
(b) t = 0 days
(c) t = 100 days
(d) DBSCAN-clustering, ε = 550 km
(e) t = 0 days
(f) t = 100 days
(g) ξ-clustering, ξ = 0.025
(h) t = 0 days
(i) t = 100 days
r(p_i) (km)
i
1000
800
600
400
0
5000
10000
15000
20000
25000
25°S
35°S
20°W
0°
20°E

Figure 6. Result of the OPTICS algorithm applied to the direct embedding of the trajectories, with different clustering methods. Grey particles correspond to noise.

~~We~~ Figure 6 shows that for this situation, the ξ-clustering method detects more Agulhas rings than DBSCAN. While the clustering results shown in the figure all depends on the parameter values for ξ and ϵ, it is visible in the reachability plot of fig. 6g that the definition of some eddies includes the entire boundary of the valleys, i.e. up to very high reachability values. At the same time, the detection of the large-scale clusters as in 6a-c is not possible with the ξ-clustering method. These findings are in fact expected, cf. the discussion of the two clustering methods at the end of section 3.3. DBSCAN is best to detect global density structures, i.e. when the reachability values of all points are compared to the same cut-off ϵ. Regions that are dense locally but not necessarily globally are better detected with the ξ-clustering method. Despite these differences between the two clustering methods, we again emphasize that the main result of OPTICS is the reachability plot itself. Fig. 7 shows a colour

map at initial time of the reachability values. We clearly see Agulhas rings as the dark regions corresponding to lowest values of reachability. The regions of large reachability correspond to trajectories that are relatively noisy compared to all the other trajectories.

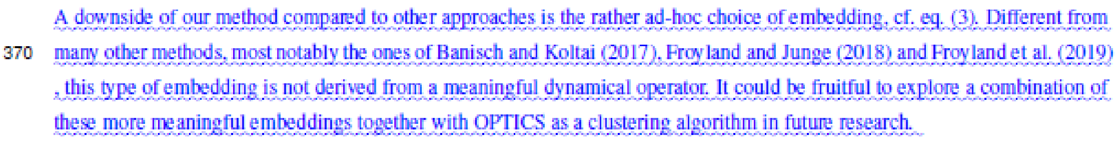
Comment 3

The (uncited) paper Froyland/Junge'18 develops a finite-element approximation of the dynamic Laplacian, which is a very accurate and robust method of finite-time coherent set extraction for low-dimensional systems of the type treated in the Wichmann manuscript. In Froyland/Junge'18 there are no free parameters, the method is unaffected by the density of the data points, and estimates are produced on the whole domain. A comparison can be made for the Bickley example in the Wichmann manuscript because the setup is identical. Wichmann et al uses a 200x60 grid of points and particle positions at times t=0, 1, 2, 3,..., 39, 40. Froyland/Junge'18 studied the same Bickley flow as in Wichmann, except that Froyland/Junge'18 used a coarser 100x30 grid of points and only particle positions at time 0 and time 40. Figure 15 in Froyland/Junge'18 shows much clearer images with fewer trajectory inputs. Thus, I think there is not a strong case for the approach in the manuscript being a better performer.

Answer to comment 3

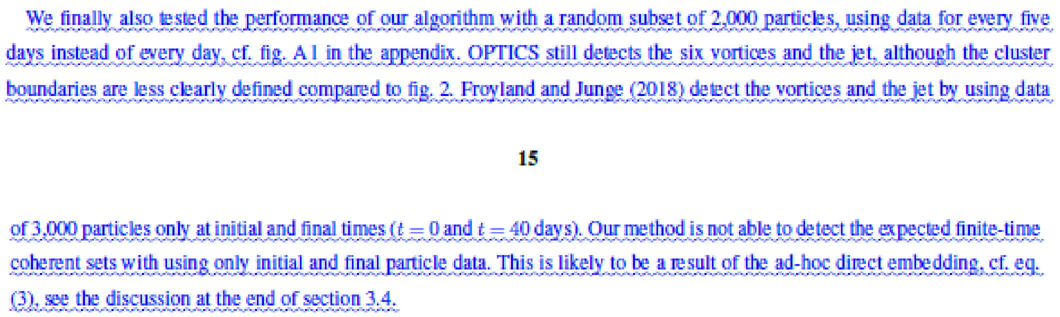
Thank you for this comment, and we apologize for not having cited that paper. Note however that the clustering results presented there are also based on k-Means clustering, and there are no free parameters only up to the choice of embedding dimension and the number of clusters. The paper also shows that the approach with k-Means works for situations where the coherent sets are not very small compared to the fluid domain, see the problems of k-Means in this context in the paper by Froyland et al. 2019. Nevertheless, the concepts presented there are powerful, as they provide a type of embedding that has a clear dynamical motivation, which is an advantage compared to our heuristic embedding. We refer to the paper at many places in the new version in different contexts:

1. End of the new section on comparison to other methods



A downside of our method compared to other approaches is the rather ad-hoc choice of embedding, cf. eq. (3). Different from many other methods, most notably the ones of Banisch and Koltai (2017), Froyland and Junge (2018) and Froyland et al. (2019) , this type of embedding is not derived from a meaningful dynamical operator. It could be fruitful to explore a combination of these more meaningful embeddings together with OPTICS as a clustering algorithm in future research.

2. We now also tested our method with the Bickley jet using less particles and less data points for each trajectory. Our method does indeed not perform as well as the method of Froyland and Junge (2018), and we want to thank the reviewer for explicitly mentioning this possible comparison.



We finally also tested the performance of our algorithm with a random subset of 2,000 particles, using data for every five days instead of every day, cf. fig. A1 in the appendix. OPTICS still detects the six vortices and the jet, although the cluster boundaries are less clearly defined compared to fig. 2. Froyland and Junge (2018) detect the vortices and the jet by using data

15

of 3,000 particles only at initial and final times ($t = 0$ and $t = 40$ days). Our method is not able to detect the expected finite-time coherent sets with using only initial and final particle data. This is likely to be a result of the ad-hoc direct embedding, cf. eq. (3), see the discussion at the end of section 3.4.

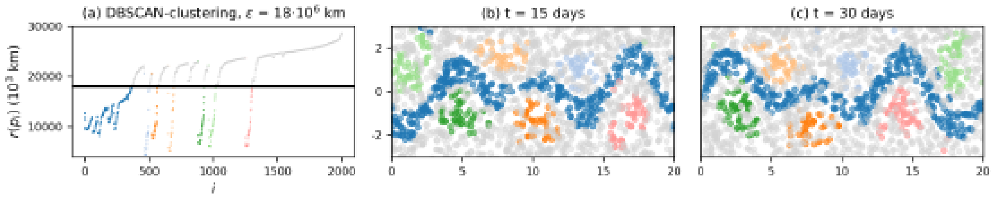
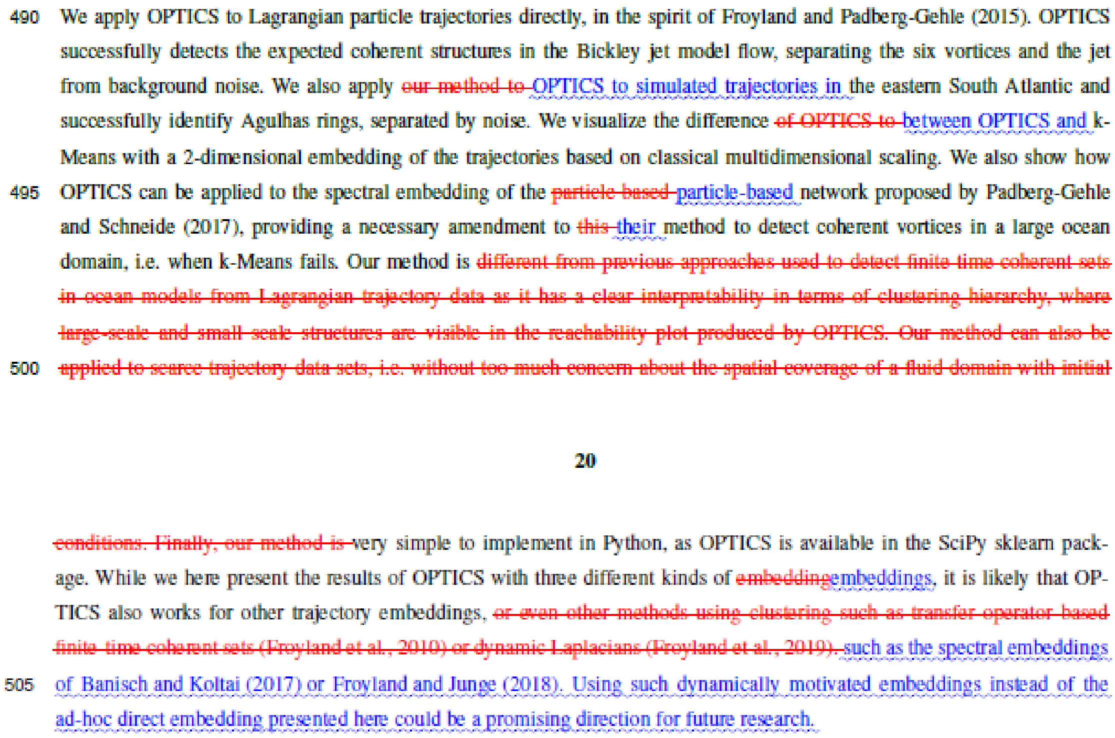


Figure A1. Result of ~~$K = 8$ k-Means clustering~~ the OPTICS algorithm for a random subset of 2,000 particles in the ~~2-dimensional embedding from classical MDS~~Bickley jet flow, ~~cf~~with particle data every 5 days instead of every day. ~~fig~~To account for the smaller number of particles, we set $s_{min} = 15$ for this case. ~~4. Axis units:~~ The six vortices and the jet are ~~in 1000 km~~still clearly visible.

3. In the conclusion, we come back to the problems of our form of embedding and mention again that a combination of the embedding of Froyland and Junge (2018) together with OPTICS could yield better results.



We apply OPTICS to Lagrangian particle trajectories directly, in the spirit of Froyland and Padberg-Gehle (2015). OPTICS successfully detects the expected coherent structures in the Bickley jet model flow, separating the six vortices and the jet from background noise. We also apply ~~our method to~~ OPTICS to simulated trajectories in the eastern South Atlantic and successfully identify Agulhas rings, separated by noise. We visualize the difference ~~of OPTICS to~~ between OPTICS and k-Means with a 2-dimensional embedding of the trajectories based on classical multidimensional scaling. We also show how OPTICS can be applied to the spectral embedding of the ~~particle based~~ particle-based network proposed by Padberg-Gehle and Schneide (2017), providing a necessary amendment to ~~this~~ their method to detect coherent vortices in a large ocean domain, i.e. when k-Means fails. Our method is ~~different from previous approaches used to detect finite-time coherent sets in ocean models from Lagrangian trajectory data as it has a clear interpretability in terms of clustering hierarchy, where large-scale and small-scale structures are visible in the reachability plot produced by OPTICS. Our method can also be applied to scarce trajectory data sets, i.e. without too much concern about the spatial coverage of a fluid domain with initial~~

20

~~conditions. Finally, our method is~~ very simple to implement in Python, as OPTICS is available in the SciPy sklearn package. While we here present the results of OPTICS with three different kinds of ~~embedding~~embeddings, it is likely that OPTICS also works for other trajectory embeddings, ~~or even other methods using clustering such as transfer operator based finite-time coherent sets (Froyland et al., 2010) or dynamic Laplacians (Froyland et al., 2019).~~ such as the spectral embeddings of Banisch and Koltai (2017) or Froyland and Junge (2018). Using such dynamically motivated embeddings instead of the ad-hoc direct embedding presented here could be a promising direction for future research.

Comment 4

The idea to not fully partition the domain has already been treated in Fr/Sa/Ro’19. Regarding the ocean eddy example in the manuscript, Fr/Sa/Ro’19 also applied the method of Froyland/Junge’18 to ocean flow and successfully extracted a greater number of eddies than Wichmann at a higher quality. On the other hand, Fr/Sa/Ro’19 used AVISO-derived trajectories rather than model output, so it could be that Wichmann is using a rougher velocity field. Wichmann also used lower trajectory density than Fr/Sa/Ro’19 by a factor of about 4; both of these items could make Wichmann’s task more difficult, compared to Fr/Sa/Ro’19.

Answer to comment 4

Thank you for pointing this out. For a detailed comparison of the both methods, it would indeed be necessary to choose exactly the same flows. Detecting a greater number of eddies in a specific ocean domain does not necessarily have an implication for the usefulness of a method. We would like to note again that the results of Froyland et al. (2019) were a major motivation for our paper, and we do not aim to compete with their method any aspects. We would rather like to show how a change of clustering algorithm, instead of a change of embedding, can also yield better results compared to partition-based clustering, see the paragraph below in the revised paper on the comparison to other methods. We believe that a combination of the embedding of Froyland and Junge 2018 together with OPTICS could be a useful extension of our method. See our answer to your comments 1 and 3 for more content relating to their method.