Cluster Results

Concepts that need to be investigated

Much has been done thus far in regards to the research, and a clear picture of what is further required now exists. The work with SOMs is likely sound, but does require that a few more variables be tested. Furthermore, it is not yet certain that SOMs will be the best clustering technique. To that end, K-means clustering and hierarchical clustering have also been identified as alternative techniques. This now gives rise to the possibility that two papers could emerge from this work. One on the resultant clustering of synoptic air-sea states during coastal MHWs, and another paper that discusses the strengths of these various clustering techniques.

The metrics for each MHW in each cluster

This requires that once the different events have been clustered, regardless of the technique used, or the variables controlled for within (see below), a summary of the event metrics must also be provided. These then will allow for the second more meaningful round of the interpretation of the results.

Table 1: The possible metrics that may be of interest for summarising the events clustered into each node. Node numbers given here correspond to Figure 1 and 2. (continued below)

node	count	summer	autumn	winter	spring	west	south	east
1	14	1	8	5	0	0	12	2
2	11	2	1	6	2	4	6	1
3	7	1	5	0	1	1	6	0
4	11	1	1	4	5	3	8	0
5	3	0	2	1	0	1	2	0
6	12	0	0	0	12	0	12	0
7	19	4	4	4	7	7	11	1
8	4	1	0	2	1	2	2	0
9	14	0	6	6	2	4	10	0
NA	95	10	27	28	30	22	69	4

Table 2: Table continues below

duration_min	duration_mean	duration_max	int_cum_min	int_cum_mean
15	23.6	40	20.57	50.051
15	28.5	98	23.87	72.412
15	17.4	21	28.74	55.614
16	28.2	61	24.49	73.968
18	34.0	65	65.53	92.234
15	32.0	47	45.26	89.687
15	21.9	43	26.14	60.822
15	16.5	18	23.77	42.931
15	24.3	48	32.97	55.592
15	25.1	98	20.57	64.829

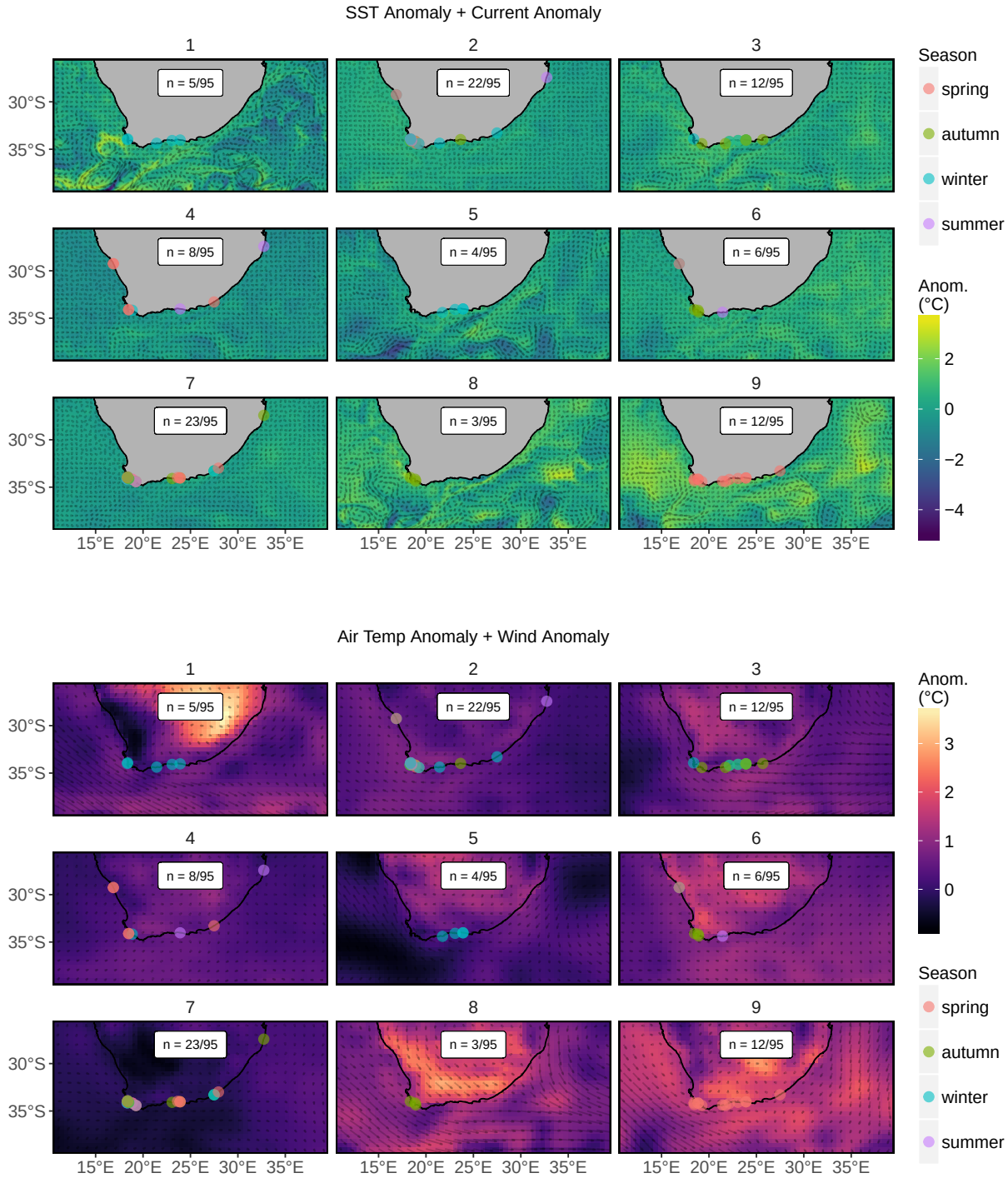

Figure 1: The results of a SOM clustering of the syoptic air-sea anaomaly data during coastal MHWs. The clusters shown here correspond to the following table and figure.

int_cum_max	int_max_min	int_max_mean	int_max_max
138.89	1.63	3.019	5.35
308.20	1.86	3.632	7.66
92.14	2.77	4.312	7.37
160.80	1.68	3.697	6.85
140.73	4.12	4.477	4.97
137.08	2.36	4.021	5.18
94.00	2.13	3.904	7.34
69.35	2.44	3.389	5.29
117.17	2.10	3.275	6.90
308.20	1.63	3.666	7.66

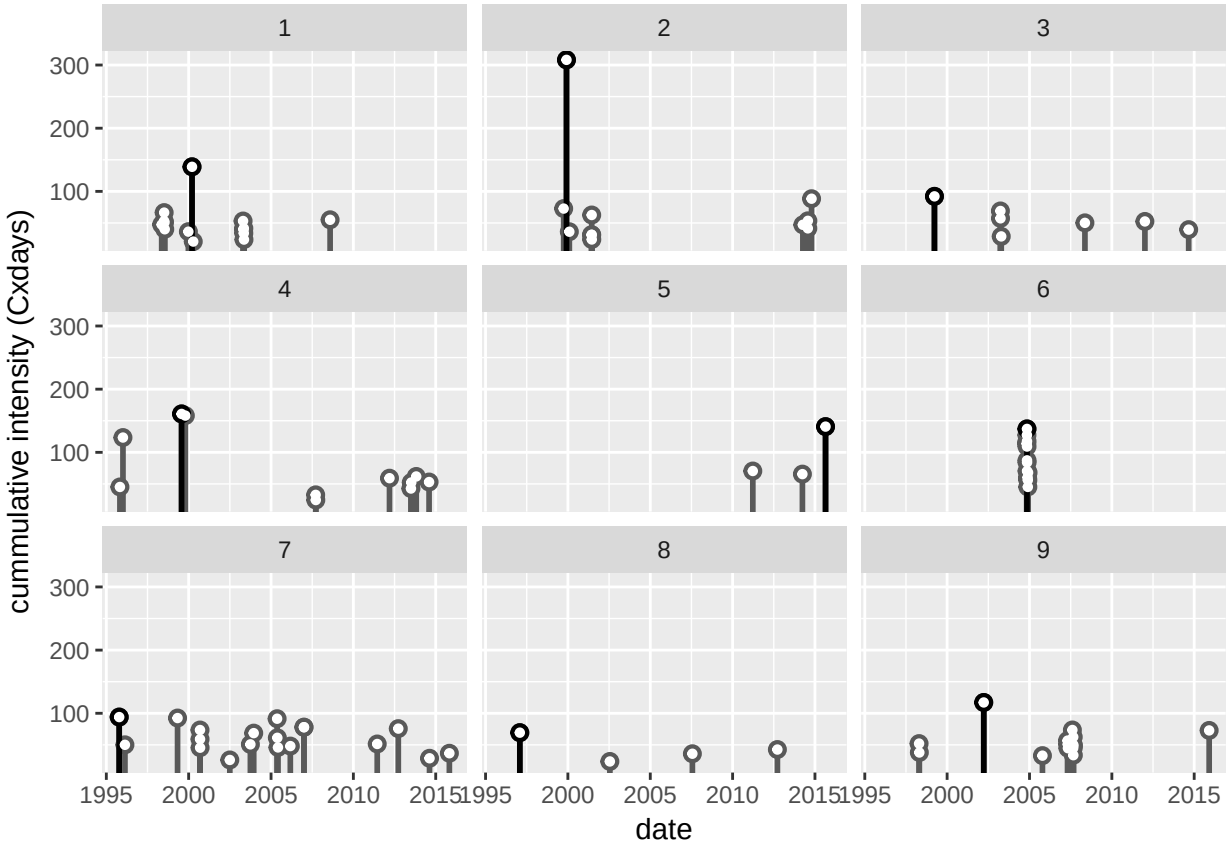

Figure 2: Lolliplots showing start date and cummulative intensity of each event by each node, as seen in Figure 1.

Figure 2 most clearly demonstrates that the SOM nodes in Figure 1 generally consist of either several events that occurred at the same time, or a blend of many disparate events. The nodes that contain more events are therefore much more 'neutral', meaning there is very little visible by way of air-sea anomalies. This is as I had feared and is currently my largest criticism of this technique. If one were to add more SOM nodes (a bad idea given that there are only 95 data vectors to be clustered), the new nodes will only allow some of the events within the larger clusters to break away and form their own node. The complexity of the air-sea states is such that any consistent patterns that may occur are obfuscated by everything else happening in the study area. Therefore it is necessary to reduce the dimensions of the input data by drawing a more narrow box around the study area to investigate the effect this may have. But first, the BRAN data need to be appropriately rounded down to a resolution of 0.5 degrees in order to match the ERA-Interim data.

Comparing large replication sets

Because the SOM nodes are shuffled during each run, comparing the results directly becomes difficult. We can however iterate the SOM analysis many times and save a vector of results for each event showing into which node it was cast for each run. This then creates a unique information vector for each event that can then be used to further cluster the events into 'mean' clusters. This helps to address the issue of SOM node 'drift'. The following two tables show the results of a SOM run on the data 10 times.

Table 4: Table showing the number of events within the same node over 10 runs on the same data. Note how variable the node assignments may be.

1r	2r	3r	4r	5r	6r	7r	8r	9r	10r
16	7	12	16	6	8	7	12	5	6
10	12	8	8	7	6	8	8	12	6
18	13	12	12	14	12	12	17	7	17
10	4	13	14	4	9	10	4	8	7
9	3	20	10	6	26	14	4	22	3
8	17	7	9	19	9	9	17	11	18
6	12	7	9	12	7	8	9	9	12
6	8	7	11	8	5	13	9	9	9
12	19	9	6	19	13	14	15	12	17

Table 5: Table showing the number of events within a node scored in descending order. The different columns show each run. Note that the results are generally consistent, but not completely.

1r	2r	3r	4r	5r	6r	7r	8r	9r	10r
6	3	7	6	4	5	7	4	5	3
6	4	7	8	6	6	8	4	7	6
8	7	7	9	6	7	8	8	8	6
9	8	8	9	7	8	9	9	9	7
10	12	9	10	8	9	10	9	9	9
10	12	12	11	12	9	12	12	11	12
12	13	12	12	14	12	13	15	12	17
16	17	13	14	19	13	14	17	12	17
18	19	20	16	19	26	14	17	22	18

Now let's up the anti a bit and run the SOM 100 times. Because the resultant data frame will become unwieldy, we will create means across the number of clustered events as coerced into descending order. These then will be compared to the mean for the 1 run and 10 run data frames.

Table 6: Table showing the number of events within a node scored in descending order. The different columns show the mean of 1, 10, and 100 SOM runs. Note that the results are remarkably similar.

1r	10r	100r
6	5	5
6	6	6
8	7	8
9	8	8

1r	10r	100r
10	10	10
10	11	11
12	13	13
16	15	15
18	19	19

As we may see from the table above, the 'wobble' present in the SOM analysis is not great. I considered running the analysis 1,000 times but this would take several hours on my computer and I think that the consistency shown between 1, 10, and 100 runs is sufficient to put this question to rest.

RI vs PCI

Even though the 'wobble' in the SOM algorithm has been shown to be small, it would be better if it was non-existent. This would allow for better comparison of the variables below as well as complete reproducibility by anyone interested in doing the work themselves. Previously with this work SOMs were being run with random initialisation (RI). In order to ensure consistent results it is necessary to use principal component initialisation (PCI). The `kohonnen` package that has been used thus far for SOMs does not have this capability so it is necessary to use a different package, `SOMbrero`. This package however uses `princomp` when calculating PCI, which doesn't work when one has more columns than rows. So yet another self-organising map implementation `yasomi` must be used as this allows one to chose `prcomp` for PCI, which works with our wide dataframe.

Table 7: Table showing the number of events within the same node over 10 runs on the same data using PCI. Note how events are always placed in the same nodes.

1r	2r	3r	4r	5r	6r	7r	8r	9r	10r
14	14	14	14	14	14	14	14	14	14
11	11	11	11	11	11	11	11	11	11
7	7	7	7	7	7	7	7	7	7
11	11	11	11	11	11	11	11	11	11
3	3	3	3	3	3	3	3	3	3
12	12	12	12	12	12	12	12	12	12
19	19	19	19	19	19	19	19	19	19
4	4	4	4	4	4	4	4	4	4
14	14	14	14	14	14	14	14	14	14

As we may see in the table above, using principal component initialisation, instead of random initialisation prevents any of the 'wobble' of events between nodes. For this reason it is preferable to use PCI instead of RI moving forward. Because all SOMs will now be initialised with PCI, it will not be included in the file names. Rather it will be assumed to be the standard.

Effect of pixel resolution on clustering

The more dimensions/ variables one introduces to a cluster analysis, the more stress will exist in the results. As large stress values are generally considered to be a negative result in clustering, it is best to attempt to reduce it where possible. One way of doing that for this research is by reducing the pixel resolution of the reanalysis products. There are two reasons that this cannot simply be done out of hand. The first is that the reduction in resolution may affect the clustering of the events. So this must be documented. The other

problem this presents is that the reduction of pixel resolution would require that any results produced be shown at this same reduced resolution. And because the goal is to show meso-scale forcing on the coast, higher pixel resolutions would be preferable. Regardless, the ERA-Interim data are at a resolution of 0.5 degrees, which requires that the BRAN data be reduced to this same resolution for appropriate cluster comparison. Beyond this initial required reduction in resolution, the question then is what effect does the further rounding of the data produce? Here we look at three resolutions: 0.5, 1.0 and 2.0 degree lon/ lat.

Table 8: Table showing the number of events clustered into which of the 9 SOM nodes. The different columns show the effect that reducing the resolution of the data has on the clustering.

res_all	res_0.5	res_1.0	res_2.0
23	14	15	13
9	11	11	12
10	7	8	9
10	11	11	9
10	3	2	2
12	12	12	12
5	19	15	17
9	4	6	8
7	14	15	13

The results table generated from the clustering of events into different nodes shown above shows that when the resolution of the BRAN data are not rounded down to that of the ERA-Interim data they have a pronounced effect on the clustering of the events. Specifically more events are lumped together in node 1, the "other" node. By reducing the resolution of BRAN to match ERA, the clustering of events becomes more even across the nodes. As we reduce the resolution further there is some shifting of a few events but the overall pattern remains. It is also helpful to see the nodes in descending order of the number of events in each nodes as seen below.

Table 9: Table showing the number of events within a node scored in descending order. The different columns show the effect that reducing the resolution of the data has on the clustering.

res_all	res_0.5	res_1.0	res_2.0
5	3	2	2
7	4	6	8
9	7	8	9
9	11	11	9
10	11	11	12
10	12	12	12
10	14	15	13
12	14	15	13
23	19	15	17

When the data are reshuffled into descending order based on the number of events clustered into each node we see that the results are very similar for resolutions coarser than 0.5. This shows us that it is not necessary to consider the use of coarser resolutions as 0.5 is most appropriate.

Effect of lat/ lon extent on clustering

With more traditional cluster analyses, the values being compared would have far fewer dimensions. In this regard one would endeavour to only include variables that seem relevant to the question being asked. For example, if clustering different rock pools by the species found within them, one would likely create better results by not including any anomalous species found in the results (like a cow fish). In regards to this work, it is best to include only the pixels that are likely relevant to the meso-scale features that may be impacting the coast. More specifically, cutting out the Agulhas retroflection above the Southern Ocean will prevent any behaviour there from affecting the clustering of events that are occurring along the coastline of South Africa.

That all being said, it may indeed be relevant, at least in the sense of potential teleconnections, to include the Agulhas retroflections over the Southern Ocean. I have decided to trim the total study area by 1 degree on the East, South and West extents, rather than just the South extent. The northern border of the study area is left unchanged as this would begin to remove some of the coastal stations. The effect of trimming the study area 1 degree at a time is presented below. A SOM is run on each trimmed dataframe. A total of 5 degrees are iteratively trimmed.

Table 10: Table showing the number of events within each node scored in descending order. The different columns show the differing amounts of spatial reduction in the extent of the study area. Note that the results remain similar when 1 to 4 degrees of pixels are trimmed, but there appears to be a difference in the results when the full study range is used and when 5 degrees of lon & lat have been removed from the East, South, and West edges of the study area.

trim_0	trim_1	trim_2	trim_3	trim_4	trim_5
3	3	2	2	2	5
4	7	8	9	9	9
7	8	9	10	9	9
11	9	10	11	10	10
11	11	11	12	11	12
12	12	12	12	12	12
14	14	12	12	13	12
14	15	15	12	13	13
19	16	16	15	16	13

The table above shows that removing the first degree of lon/ lat around the study area does have an effect on the SOM clustering. Leaving the study area 'as is' is shown in the column labeled 'trim_0'. These are the results shown in the previous tables and figures. Removing 1 to 4 degrees of lat/ lon has all have similar effects and appear to spread the clustering of the events a bit more than using the full study area. Trimming 5 degrees of lat/ lon smooths the clustering noticeably more. The difference between these different extents is not large, but I think it does show that the edges of the study area are having an effect on the clustering. And that reducing the area does allow for a more even clustering of the events into nodes. It may be best to use a study area with 1 degrees trimmed from the East, South, and West edges. But for now I shall continue to use the full study area.

Effect of running air and sea variables separately

It may be that air and sea values work in tandem with one another to force MHWs, but it is more likely that they do not. Except for perhaps VERY extreme situations (e.g. once per decade). Therefore it is necessary to run all clustering techniques on air-sea values combined, as well as separately, in order to quantify the

potential effect they have on clustering. Up until this point all variables have been run together, here we pull them apart to see how the results may differ. We do this by running all air or sea variables together and then by running air or sea temperature and wind/ current vectors separately.

Table 11: Table showing the number of events within a node scored in descending order. The different columns show the SOM results based on subsets of the data. The first column 'all_all', shows the previous results on the full, unsubsetted, untrimmed, etc. data. Note that the full air data (i.e. temp, U and V) is the most similar to the un-subsetted results.

all_all	air_all	air_temp	air_uv	sea_all	sea_temp	sea_uv
3	1	7	5	5	7	2
4	6	8	6	7	8	7
7	8	9	7	8	8	8
11	9	10	7	8	9	8
11	12	10	9	9	11	9
12	13	11	13	10	11	12
14	14	12	14	12	11	13
14	14	13	15	13	12	15
19	18	15	19	23	18	21

As we see in the table above, separating the variables from one another has a large effect on the clustering of the data. Unexpectedly, the air data appear to produce results the most similar to the results generated by running the SOM analysis on the full data set. This suggests that the air data (UV more than temperature) are driving the SOM clusters, and not the sea data. My general perception of all of the tests and results thus far has been that it is the sea, and not the air, that is usually driving events. Which makes this result a bit trickier to deal with. The importance of air data on clustering the data must be acknowledged, but I also think that this provides a clear argument against clustering the data using all of the variables at once because we generally want to see what the effect the ocean temperature is having on the coastal MHWs. One reason why air temperature and winds may be having a stronger effect on the clustering is that they are much more even, allowing for easier clustering of the results. This issue may require further analysis but I will leave it as is for now.

Do normal days cluster apart from events?

The idea here is to include the 366 daily synoptic climatology values in with the synoptic MHW values to see if they cluster differently. From previous SOM and K-means cluster analyses on these daily climatology values it was determined that 2 - 4 nodes/ clusters was optimal. Anything more than that was over fitting. I attribute this to the much more even nature of the climatological data. Therefore I hypothesised that if one were to cluster the daily clims with the event data they would be placed into different clusters. There are 95 events and 366 daily clims, so this should allow for a pattern to be seen if one does exist. Up until this point SOMs have been the primary tool of investigation into clustering, but for this last step I opt to use HCA to produce a dendrogram of the results. I do this because it provides a better visual index of the difference between the points. It also allows for a nice visual representation of the two different classes of data being compared (i.e. event data and daily clim data). Additionally, an MDS is run to allow for another dimension of spatial difference to be inferred.

Before running HCA and any other cluster or ordination techniques on these data we want to see how many clusters would be reasonable to use. The figure above shows that 4 to 6 is a good choice. With that known, we now run HCA, create a dendrogram and overlay some clusters.

The dendrogram from the HCA very clearly shows that the event data separate out from the daily clim data

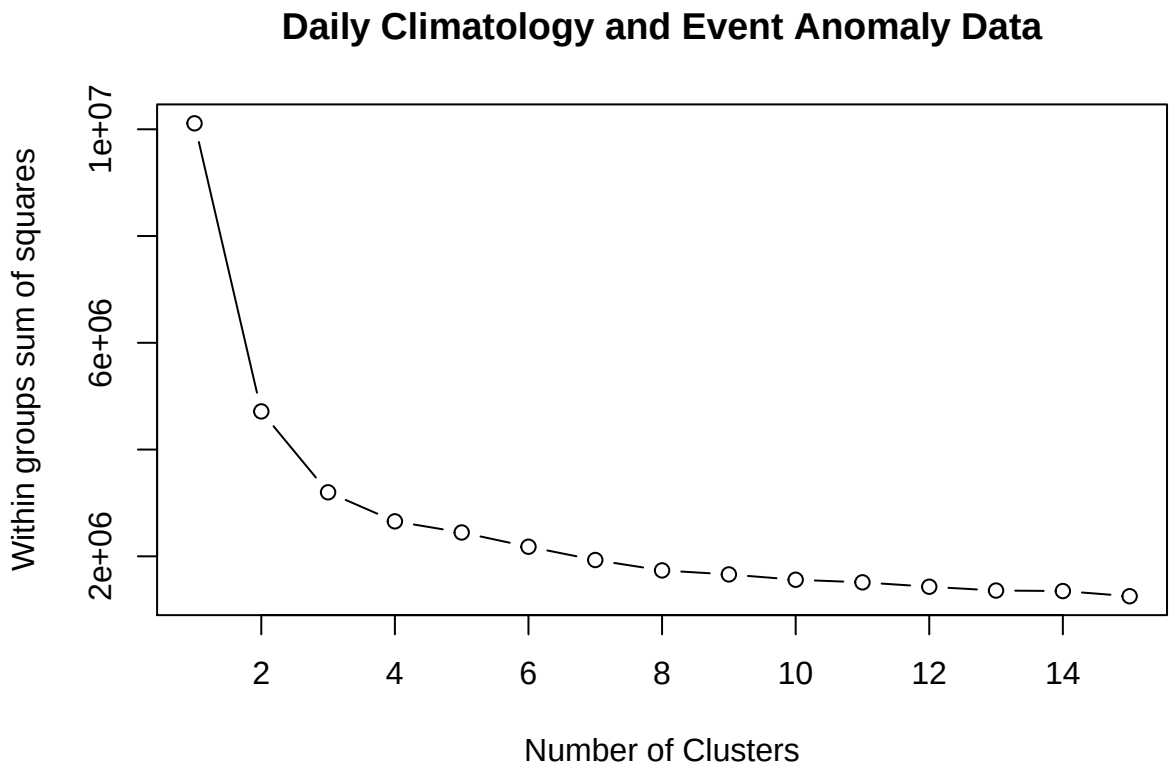

Figure 3: Plot showing the decrease in WGSS as more clusters are used for the results of an HCA on the anomaly values for synoptic air-sea states during events and daily climatologies.

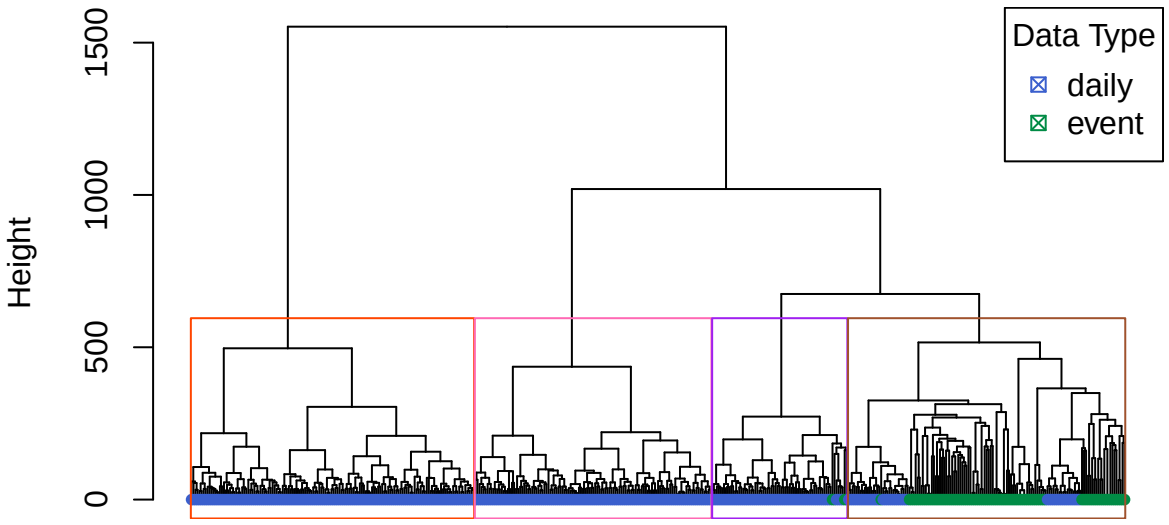

Figure 4: Dendrogram showing the results of an HCA on the anomaly values for synoptic air-sea states during events and daily climatologies. The daily clims and event data are shown with different colours. The dates or event names are included but are not legible.

almost perfectly. Besides the overlayed clusters, one may also see that the final branch on which each event sits is much longer than the daily clims. This means that not only are the events clustered apart from the daily clim data, but also that the individual events are also much more dissimilar from any of the other data points than the daily clims. But let's not stop there. The dendrogram makes a very convincing case for the dissimilarity between event and daily clim data, but let's add another dimension by creating an ordiplot via MDS.

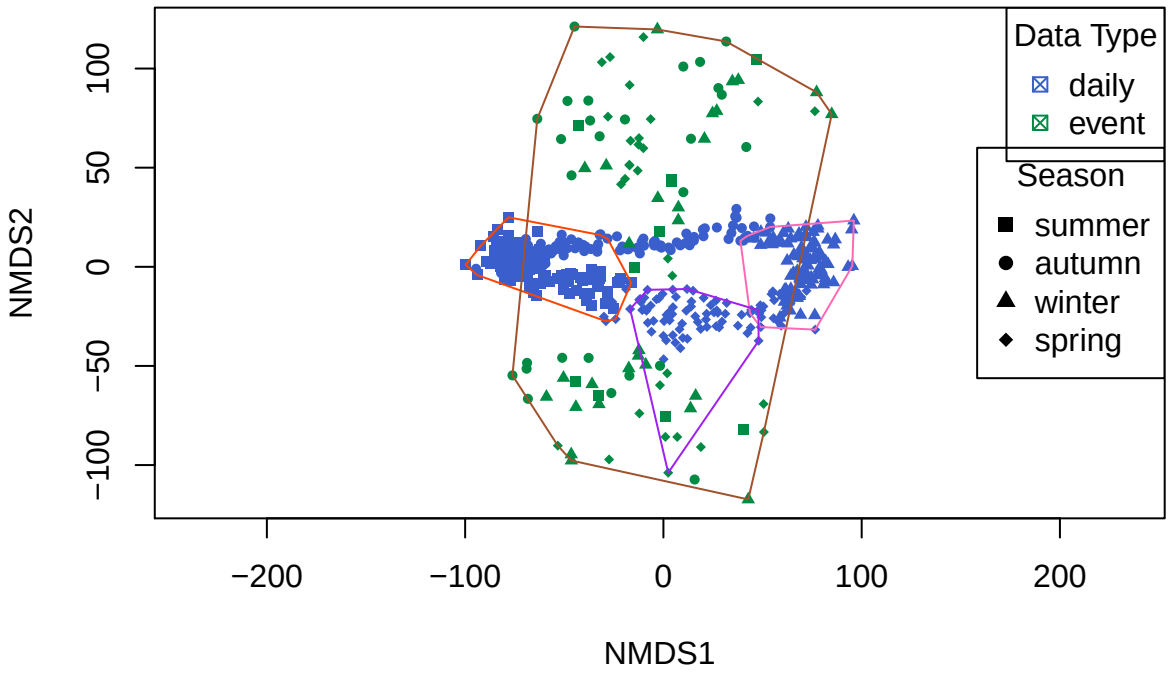

Figure 5: Ordiplot showing the results of an MDS on the anomaly values for synoptic air-sea states during events and daily climatologies. The daily clims and event data are shown with different colours. The dates or event names have not been included but are available. The clusters from the dendrogram are shown here with corresponding colours.

I find these results very exciting. I think this ordiplot shows very clearly that the synoptic air sea states during the 366 daily climatologies are different from almost all of the synoptic air-sea states during coastal MHWs. As one may see from the flat ellipse of blue squares (the daily clim points), the variance represented in the x axis is seasonality. Indeed, if the dates are included in the figure above they are in a contiguous state. With January 1st in the top left edge of the ellipse of blue squares and the dates then move clockwise. So May is roughly in the middle of the top of the ellipse and October in the middle on the bottom. The synoptic states during events appear to be controlled by the variance represented by the y axis. This must be some sort of variance that is aseasonal. Likely the anomalous characteristics of air and or sea that occur during the events. This will require further investigation but I think it will prove to be a very strong result. Even if it isn't central to the question of what are the air-sea states during extreme events, it certainly helps to show that whatever those states may be, they are different from the common air-sea states. Also worth noting is that the daily climatologies for summer and winter do not cluster at all with any of the events. They are almost all clustered with autumn, and a few with spring days.

This dot plots shows the seasonality of the clustering in a chronological order. The colours of the dots show if they are daily climatologies or event data. The x axis shows the date or event name, but there are to many to read. The take away message from this is to see how the clustering very clearly progresses throughout the year in a very even fashion. With cluster 1 representing summer, 2 shows Autumn, 3 winter, and 4 is the spring days. Beyond that, we see that almost all of the event data is clustered in with the Autumn data. This means that conditions during autumn most closely resemble the air-sea state during an extreme event. This could be taken to mean that ecosystems are naturally at more risk during this time of year. Or, perhaps,

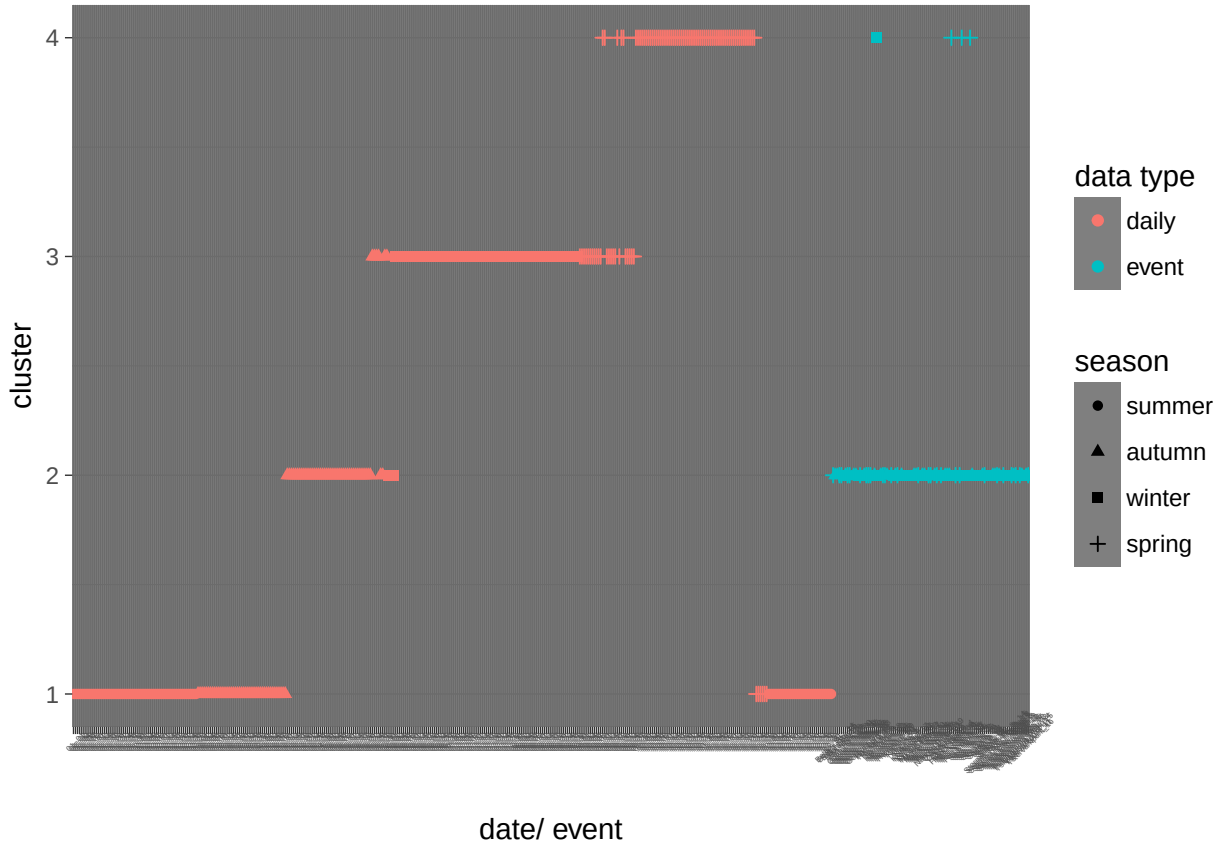

Figure 6: Dotplot showing the clustering of the anomaly values for synoptic air-sea states during events and daily climatologies as shown in the dendrogram and ordiplot. The daily clims and event data are shown with different colours. The dates or event names are shown on the x axis but are illegible.

due to the consistency of the seasonality, that species would be more prepared for these conditions during this time of year and therefore less susceptible. One would need to do more research to say. But that isn't the focus of this work anyway. When this figure is taken in conjunction with the MDS plot above we are able to say that the event data is not most like the autumn daily clims, but rather they are the least dissimilar to these data. Because they are very different from the daily clims.

I understand that this may not look as clear to the reader as it does to me, so please let me know in what ways I am failing to communicate the patterns I see in these data so I can better delve deeper into them in order to paint a clear picture for the publication.

Hierarchical clustering

HCA differs from the other two techniques outlined below in that it does not cluster the data simultaneously, based on the least stress that can be found between data vectors. Rather it iteratively divides (or combines) data vectors as the algorithm moves down (or up) a classification tree. Always looking for the point at which clusters of data may be split (or combined). This method may benefit this research due to this one dimensional approach to clustering. It will display the patterns in a more linear way. This was already run above and it did indeed provide a strong argument for how normal days and MHW days cluster out from one another. Below we will look at HCA applied only to event days.

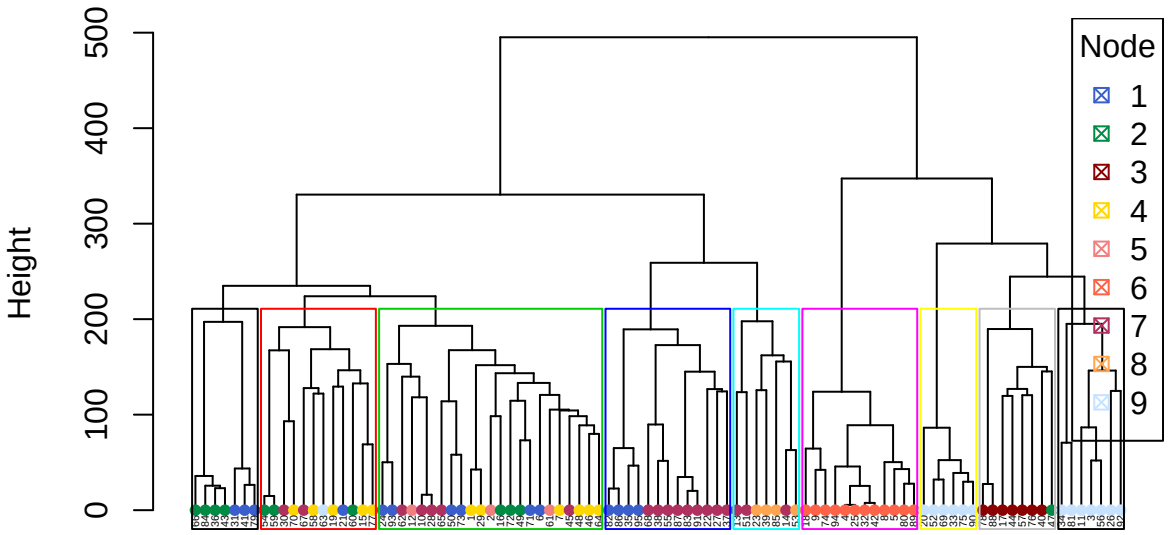

Figure 7: Dendrogram showing the results of an HCA on the anomaly values for synoptic air-sea states during events. The SOM nodes as calculated previously are shown in colour.

K-means clustering

The simplest method of clustering, and for that reason still one of the best. This is a basic algorithm that takes all data vectors and positions them in a 2D space. It then picks K points and sees, given the best possible fit of all dimensions being used, which data vectors are closest to which of the K points. This process is then repeated n times until a best fit is found. The data vectors are classified into the cluster centroid to which they are closest. This method makes two critically flawed assumptions for our purposes. The first is that it assumes the data will be distributed around the centroids in a spherical manner, and the second is that K-means attempts to find equitable samples in each node. Also due to the width of the dataframe being used, there are no built in R functions that I have found that allow one to plot the results 'out of the box'. I

therefore run an PCA on the data in order to extract the two largest principal components and plot these via an ordiplot with the K-means results overlayed.

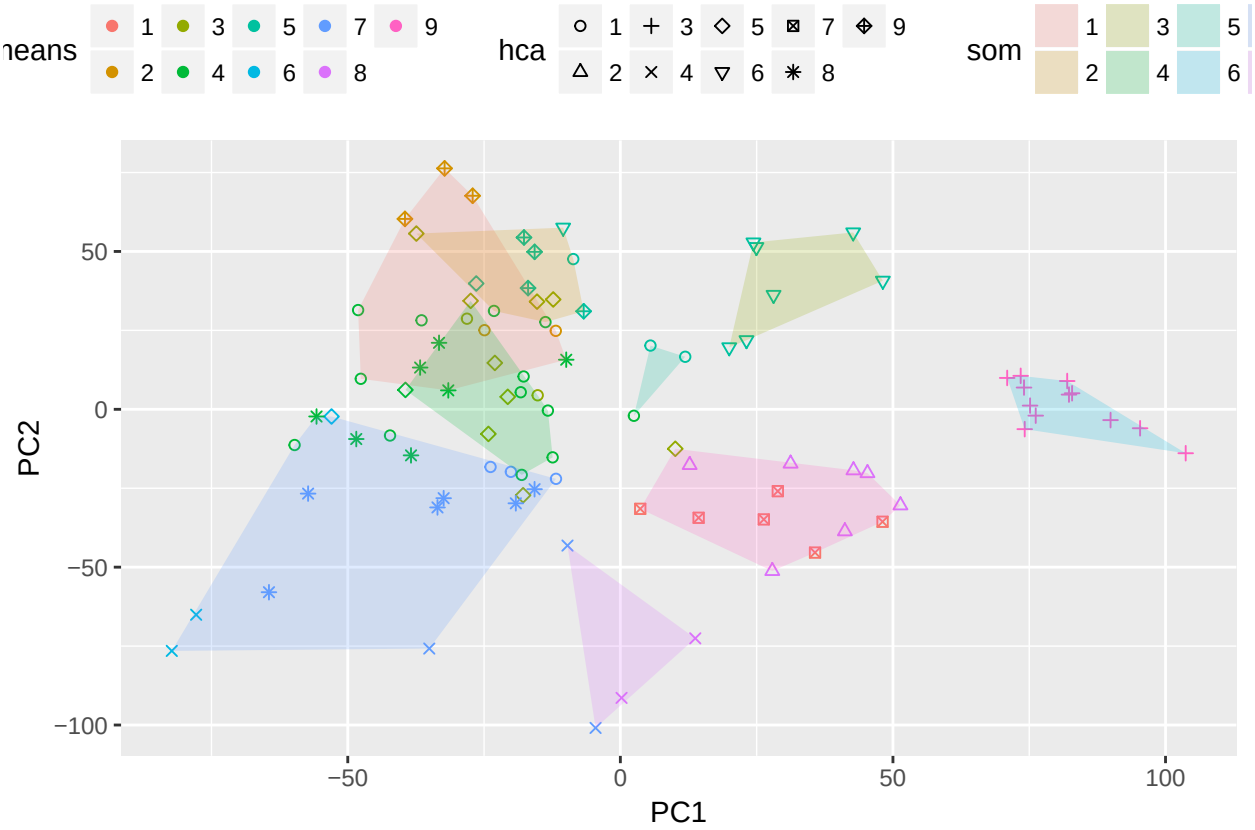

Figure 8: Ordiplot showing the results of K-means clustering on the event data. Axes determined via PCA. The colour of the points shows which K-means cluster the event has been placed in, whereas the colour of the poolygons shows in which cluster the events have been placed via HCA. Lastly, the shape of the point shows the results of the SOM clustering.

Jikes! There is a lot going on in that figure. One could spend hours going over this. After several minutes of study it appears that the HCA and K-means methods are more similar to one another than the SOM. That being said, it seems that the PCA may agree best with the SOM results. And I'm not sure what to think of that. Either way, these are very interesting results and even though there is a big clomp in the center, this is to be expected and from the ordination work I've done in the past these results are much more clear than most things I've seen. Very encouraging. If I were to point out only one thing on this figure it would be that the clomp of 12 events that have always been clustered together, no matter what method is employed, stand out very clearly on the far right. Wonderful!

SOMs

The originally proposed technique and perhaps, once this dust settles, the reigning champion. The SOM technique is apart from the previous two methods in that it accounts for the gradient that exists between the nodes it clusters the given data into. Meaning that the positions of the nodes in 2D space is relevant, unlike HCA and K-means. It is unnecessary to revisit the SOM clustering here as the majority of this document has already shown those results.

MDS

Multi-dimensional scaling provides another possible layer of interpretation of these data, even though it is not in itself a clustering or ordination technique. By highlighting which pixels on the map belong to which meso-scale properties (e.g. Agulhas, Benguela, Agulhas retroflection) it is then possible to 'environmentally fit' the effect of these pixels, and therefore meso-scale features, on top of the ordiplots generated by MDS. This however presents a large undertaking and I think it begins to move outside of the scope of this proposed research project. It would however be an excellent next step in the process. The existing categorical variables of coast and season can however be readily fit to the data and this is shown below.

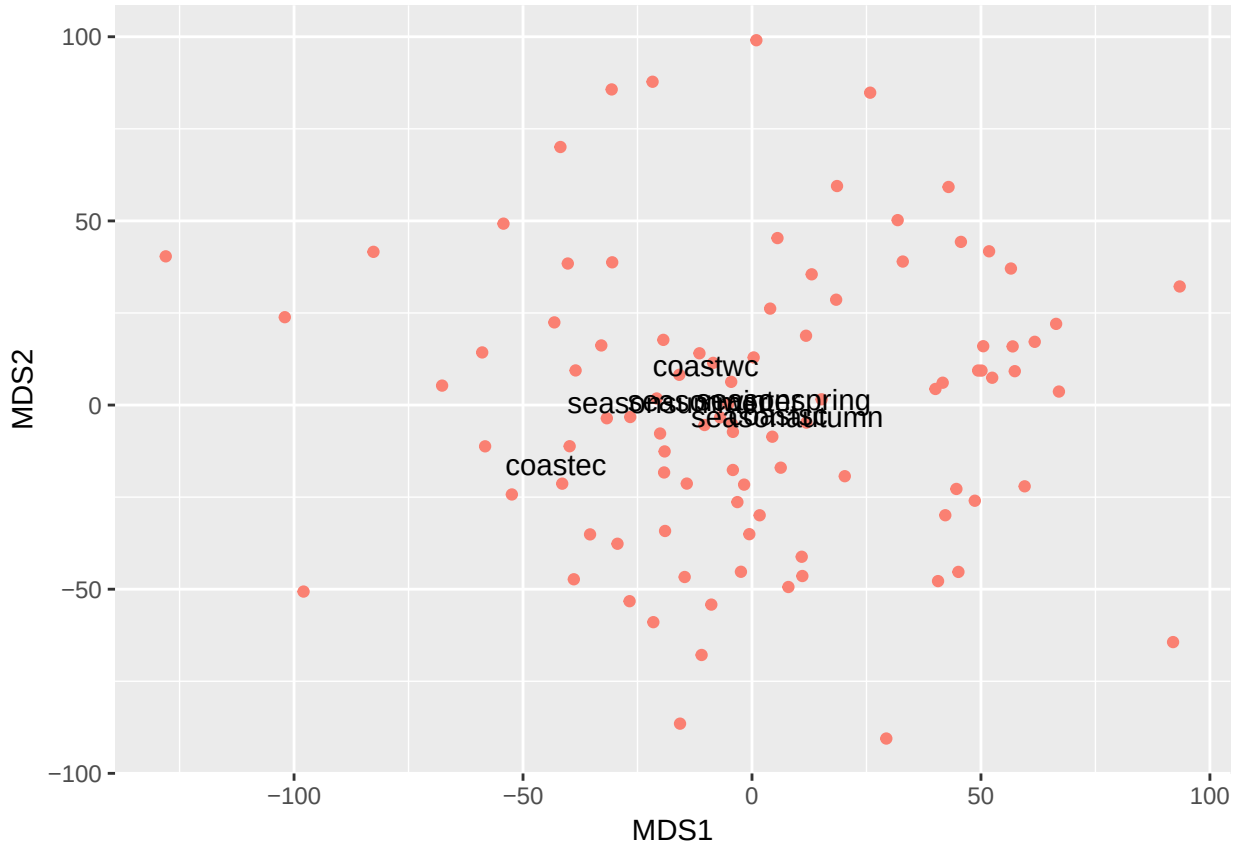

Figure 9: Ordiplot showing the MDS results on the event data. The two environmental variables of season and coast have been fit to the data. Vectors are not available as these are discrete variables.

I intentionally omitted all of the other clustering results from this figure so as to allow for the effect of the environmental variables to stand out on their own. Notably, they do not. There is very little environmental effect from these two categories. The east coast events stand out a bit from the rest, which is worth mentioning. This is not surprising, and only serves to support the hypothesis that the east coast, controlled by the (relatively) consistent Agulhas current is different from the south and west coasts. If we look at the significance of the fit of these variables we see that the fit of the coastal variables is significant at $p = 0.044$ but that season is not at $p = 0.509$.

ANOSIM

Before we look at all of the results laid out next to each other, let's have a peak at the analysis of similarity for the clusterings of the three methods.

All three clustering techniques produce significantly different clusters when 9 are used at $p = 0.001$. This is good as it means I don't have to redo everything to satisfy this requirement... From a non-lazy point of view this is also good because it shows that these techniques are effectively finding differences in the data and partitioning them accordingly. Rather surprisingly (or perhaps not) HCA is by far the winner with the largest R (this represents cluster dissimilarity, not a correlation value, confusing yes) = 0.656. K-means is in second at R = 0.205 with SOM taking up third at R = 0.14. This is perhaps not surprising in that these results represent the gradient in selectivity that these three techniques employ when clustering/ ordinating data.

Comparison

Now that the clustering/ ordination of all of the different techniques has been run, we may plot all of them side by side. There are of course a near limitless number of ways in which this may be done, but I've chosen to use lolliplots. Because I like them. I also think they do a good job of conveying the many dimensions of the data well.

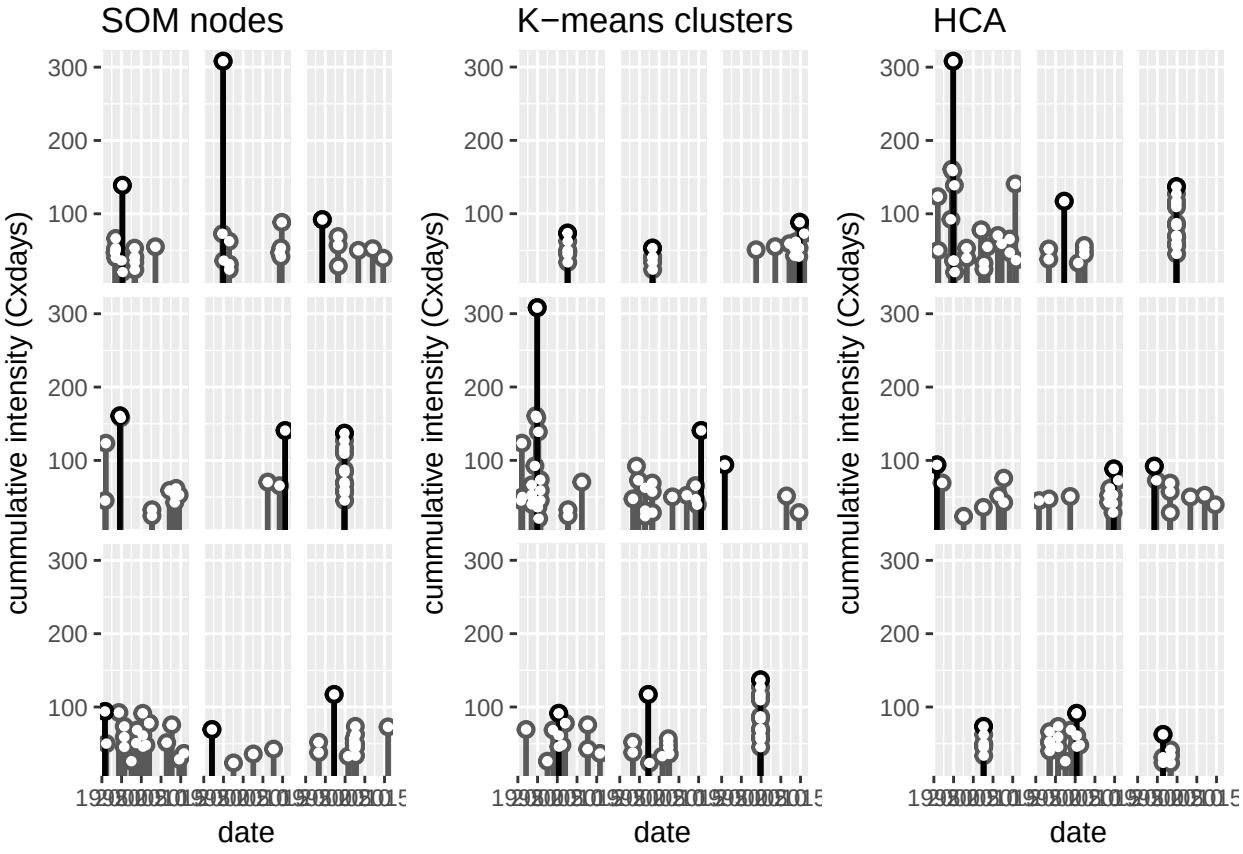

Figure 10: Lolliplots for all three methods shown together in three panels. Each facet shows the events clustered into each grouping.

This will take some time to unpack. It is also still necessary to visualise these clusters in order to see more clearly what it is that the air-sea state is up to that had the different methods setting their minds on how to cluster the events as they have. But for now these lolliplots allow for a cursory examination of the clustering in a different manner than the figure preceding this one. It is important to note that while the orientation of the nodes in the SOM results on the top panel are relevant, the cluster/ ordination layout of the 9 facets in the HCA and K-means results are not. This makes it a bit difficult to compare the results as it is not a direct task. Node/ cluster labels have been omitted for a cleaner visualisation.

References