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	19	3	7	4	5	6	11	2
2	5	2	2	1	0	3	2	0
3	12	0	0	0	12	0	12	0
4	15	1	0	6	8	2	12	1
5	2	0	0	2	0	2	0	0
6	8	0	8	0	0	1	7	0
7	5	0	0	5	0	2	3	0
8	17	4	5	4	4	4	12	1
9	12	0	5	6	1	2	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	24.4	65	20.57	68.242
19	25.2	43	45.97	70.895
15	32.0	47	45.26	89.687
15	24.8	35	23.77	59.027
27	27.0	27	41.49	47.452
15	19.2	25	23.73	51.747
18	26.0	31	33.67	47.867
16	27.9	98	24.49	75.261
15	18.8	27	23.87	43.202
15	25.1	98	20.57	64.829

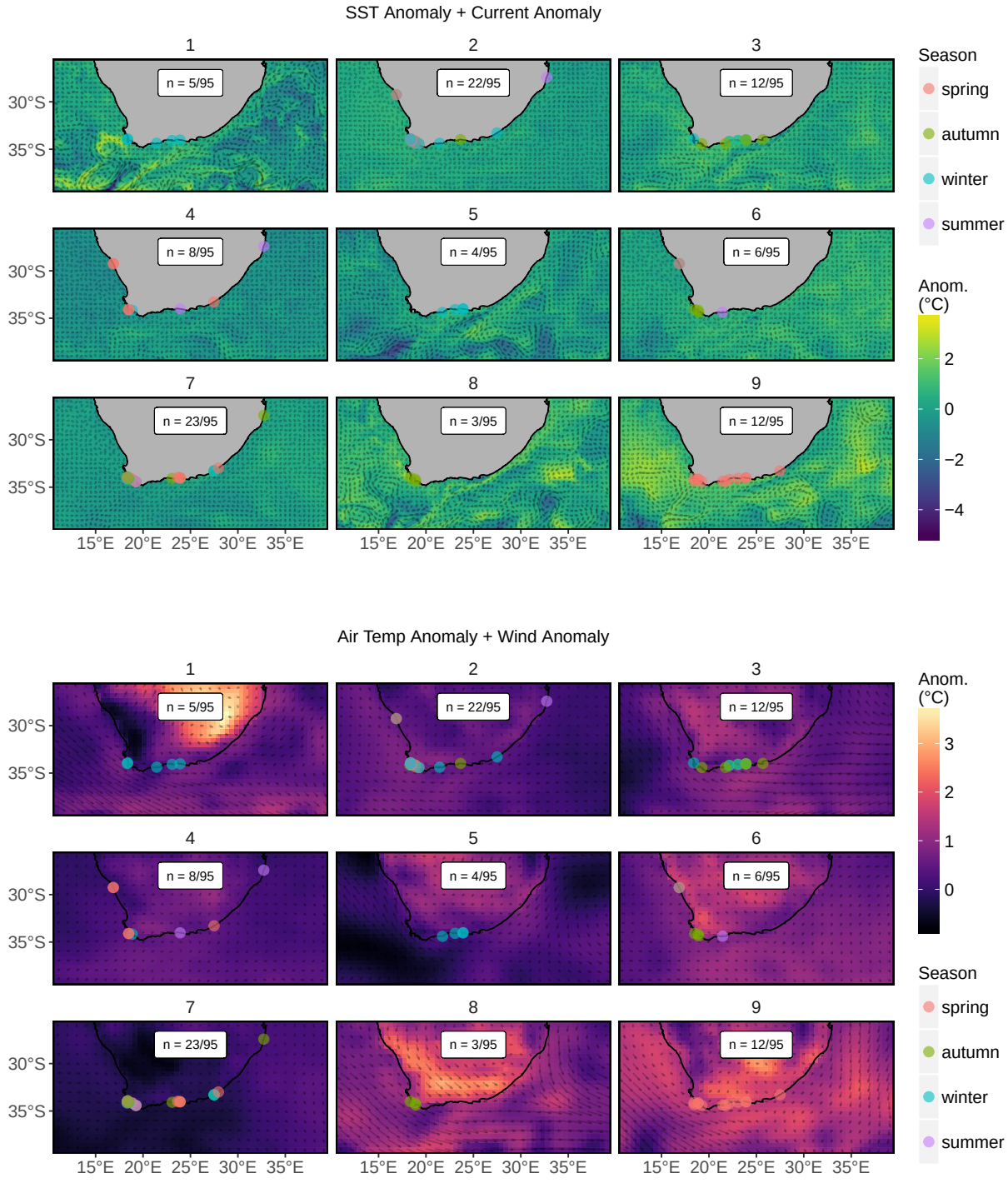

Figure 1: The results of a SOM clustering of the syoptic air-sea anaomaly data during coastal MHWs. The clusters shown here correspond to the following table and figure.

int_cum_max	int_max_min	int_max_mean	int_max_max
160.80	1.63	3.935	7.34
91.43	2.77	4.479	6.94
137.08	2.36	4.021	5.18
158.12	2.16	3.578	7.66
53.42	2.28	2.381	2.48
92.14	2.01	3.499	7.37
62.49	2.10	2.669	4.03
308.20	1.68	3.891	6.90
69.02	1.86	3.083	4.80
308.20	1.63	3.666	7.66

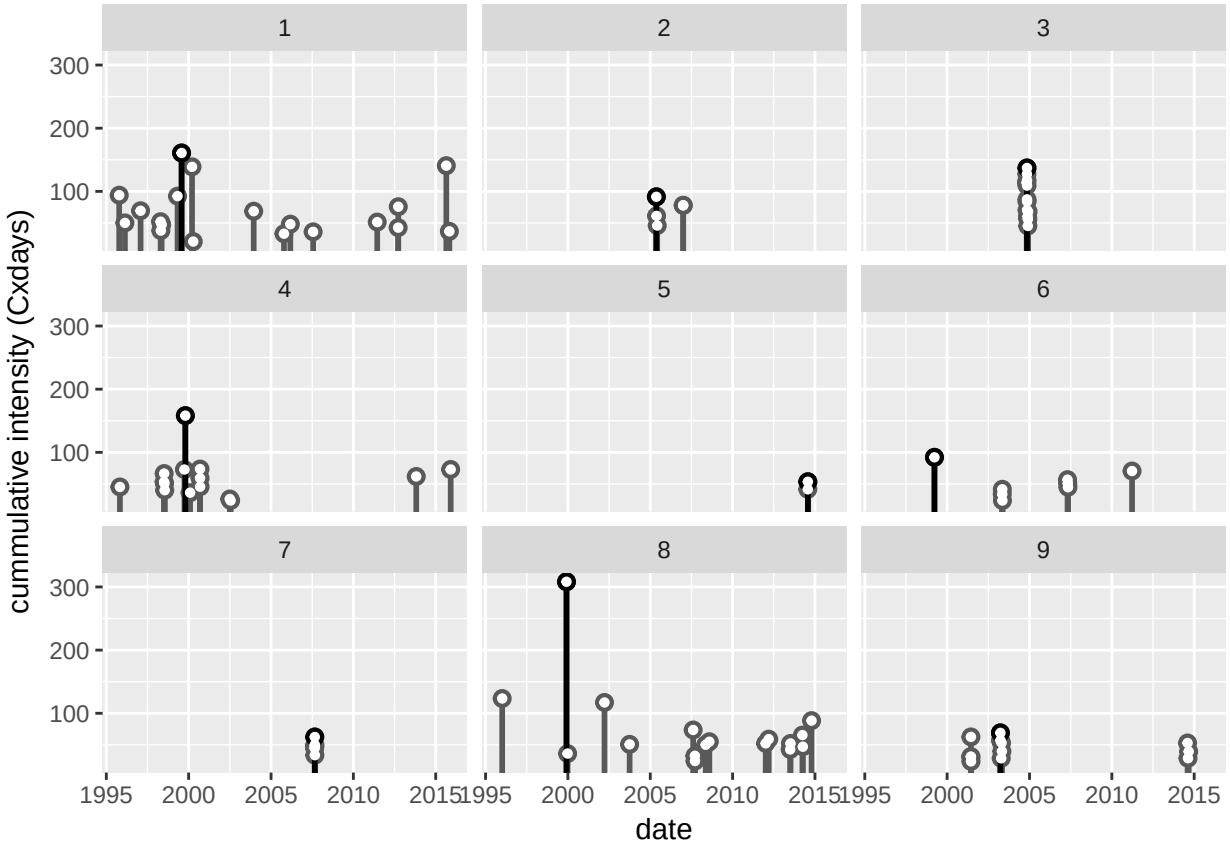

Figure 2: Lolliplots showing start date and cummulative intensity of each event by each node, as seen in Figure 1.

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.

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 a clustering, it is best to attempt to reduce it where possible. For this research that means 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 faces 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 4: 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
19	5	12	7
5	13	9	12
12	12	18	16
15	7	4	4
2	17	4	4
8	8	25	10
5	10	10	12
17	7	9	8
12	16	4	22

The results table generated from the clustering of events into different nodes shown above is not very informative because it is known that the SOM algorithm always reshuffles these data into different nodes. Due to the very high dimensions of the data, the algorithm is not able to find a single best answer. Therefore, it is more informative to further order each column from highest to lowest so as to see if the general clustering of events is similar.

Table 5: Table showing the number of events within a node scored by 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
2	5	4	4
5	7	4	4
5	7	4	7
8	8	9	8
12	10	9	10
12	12	10	12
15	13	12	12
17	16	18	16
19	17	25	22

When the data are reshuffled into descending order based on the number of events clustered into each node

we see that the results are much more similar than they first appeared. This analysis could of course be much more rigorous, but this would require the creation of several sets of figures and so it is left as is for now. The next step in this portion of the analysis then would not be the creation of figures, but rather thinking of a clever way of comparing specifically which events have been clustered together and to then create a dissimilarity index of some sort based on these results. This then would allow for the comparison to be scaled up so it could be replicated 1,000 times to be more thorough.

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 finds in the results. 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.

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. Therefore it is necessary to run all clustering techniques on air-sea values combined, as well as separately.

Do normal days cluster

The idea here is to include the 366 daily synoptic climatology values in with the synoptic MHW values to see if they cluster differently.

Clustering techniques that need to be investigated

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

Dendrograms

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 x number of times until a best fit is found. The data vectors are classified into the cluster centroid to which they are closest.

Ordiplots

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.

SOM nodes

References