Co-occurrence

Robert Schlegel

03 October 2016

Overview

This markdown document contains all of the code used to produce the co-occurrence analyses. Not all of the code is shown in the pdf output of this file however, if one were to open the .Rmd version of this file in Rstudio one could look through all of the code, which is all fully annotated. This step in the work-flow produces several files. They may all be found at the following repository https://github.com/schrob040/AHW. The naming convention is as follows: "w/x1_x2_y_z_CO.Rdata" where "w" is the file pathway, "x1" is the first dataset being used in the comparison, "x2" is the second data set being used in the comparison, "y" is the type of extreme events being compared ("hw" = heat wave, "cs" = cold-spell), "z" is the temperature metric being compared (note that the SACTN data are only measured in Tmean), "CO" is simply there to remind the user that these are co-occurrence data and ".Rdata" is the proprietary R language file type. These results are saved as .Rdata as they are muuuuch smaller than .csv files. If there is interest in being able to work with these intermediate results outside of R please let me know and I can produce .csv/ .txt/ .whatever files as requested.

```
"data/cooccurrence/SAWS_SACTN_hw_tmean_CO.Rdata"
"data/cooccurrence/SAWS_SACTN_hw_tmax_CO.Rdata"
"data/cooccurrence/SAWS_SACTN_hw_tmin_CO.Rdata"
"data/cooccurrence/SAWS_SACTN_cs_tmean_CO.Rdata"
"data/cooccurrence/SAWS_SACTN_cs_tmax_CO.Rdata"
"data/cooccurrence/SAWS_SACTN_cs_tmin_CO.Rdata"
"data/cooccurrence/SACTN_SACTN_hw_tmean_CO.Rdata"
"data/cooccurrence/SACTN_SACTN_cs_tmean_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_hw_tmean_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_hw_tmax_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_hw_tmin_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_cs_tmean_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_cs_tmax_CO.Rdata"
"data/cooccurrence/SAWS_SAWS_cs_tmin_CO.Rdata"
```

This step in the work-flow draws on results produced by the "proc/SACTN.RMarineHeatwaves.R" and "proc/SAWS.RMarineHeatwaves.R" scripts, which calculate the extreme events for both the SACTN and SAWS datasets respectively using the RMarineHeatwaves algorithm.

The output of these co-occurrence analyses is next analysed by the "proc/results.R" script to produce a more thorough understanding of any relationship of extreme events that may exist between AND/ OR within the datasets. The results from this and the "proc/results.R" scripts are visualised in the "graph/figures.R" and "graph/figures2.R" scripts. I am in the process of converting these scripts to mark down files as well.

Extreme events

As mentioned above, this step makes use of the extreme event calculation results from "proc/SACTN.RMarineHeatwaves.R" and "proc/SAWS.RMarineHeatwaves.R". What is important to note here is that the short and irregular time periods of the SACTN time series prevent a consistent analysis period from being used to establish one climatology across all time series to allow for consistent comparison. Therefore individual analysis periods

were established for each SACTN time series based on the first and last full year of data. Each of the 11 SAWS time series had their extreme events calculated 21 different time so as to match the 21 different analysis periods from the SACTN dataset. This was done for both the heat waves and cold-spells. And for the Tmean, Tmax and Tmin temperature statistics. A seperate analysis was also run on only the SAWS data using the consistent analysis period of 1981-2010 from Kruger et al. (2016) for each of the event types and temperature metrics.

Co-occurrence calculations

The extreme event results are produced in such a manner that only one step is required before co-occurrence calculations may begin. This is to split the data frames up into heat waves and cold-spells. This isn't absolutely necessary but the data frames were beginning to become too large and so I felt it best to divide them up a bit so as to make the next steps less unwieldly.

After splitting the data frames up, one must simply run the "cooccurrence()" function I wrote for just this purpose. It takes one or two minutes to compare all of the events of one dataset against those that occurred in another. Going through the comparisons pairwise, site by site. I first ran all of these co-occurrence calculations on the heat waves and cold-spells BETWEEN the SACTN and SAWS datasets. When making these comparisons I used the SAWS events that were calculated from the analysis period for the SACTN site that the SAWS events were being compared against.

I then ran the same co-occurrence calculations on heat waves and cold-spells WITHIN the datasets. When comparing SAWS events in one time series to those in another SAWS time series I used the common analysis period of 1981-2010. When comparing SACTN events in one time series to another I used the analysis period of each individual time series, meaning that most of the SACTN comparisons where not made with the same analysis periods. This issue is unfortunately not avoidable as many of the time series only have a few years of overlapping dates, which would not be a sufficient amount of time to build a convincing climatology.

It is important to note that the co-occurrence results do not show the number of events that occur within a specific range of days, but rather now show which event occurrs the most recently to the event that is being tested. The output is then the full range of statistics (including dates, distance and bearing between the sites) to allow for a more thorough analysis.

Co-occurrence Methodology

Whereas the "cooccurrence()" function appears delightfully simplistic on the surface, it actually makes use of several other special functions written to allow for parallel computing of the task at hand. If one is reading this via the .pdf version, the attendant code will not be visualised as I will describe the process in plain English. If one would like to read the code directly it may be viewed in the .Rmd version of this file. All of the functions that are described below may also be found in the "func/cooccurrence.func.R" script.

cooccurrence()

This function is designed so as to be the first and last step. It is also designed to be able to calculate co-occurrence of events between AND within different datasets. It iteratively takes a single site from a dataset and compares that against all the sites from a different dataset. If one is comparing the same data frames it removes the redundant time series so as not to produce a bunch of 100% match success results... Because that would be misleading. It then removes any extreme events that occur outside of the range of dates found within the time series being compared (NB: whilst creating this file I found that this step is not behaving itself and will need to be corrected). After shedding the events occurring outside of the appropriate range of dates, it then calculates which quantile each event belongs to. This is necessary so that later one may look at how co-occurrence proportions change when one examines only the larger events that occurred in

a time series. After doing this it then runs the "event.match()" function five times. Once for each of the quantiles that were previously calculated (i.e. 0.00, 0.25, 0.50, 0.75, 1.00). After calculating all of these results everything is munged together and saved as a .Rdata file. But read on to find out what other exciting twists and turns occur the deeper down the rabit hole the analysis goes!

event.match()

This next step on the co-occurrence journey is a small one (but is where I must move the year cropping function to ensure it works correctly). It's main purpose is to extract individual sites from the full dataset it was given, so as to allow the next function to compare only one site from the first dataset, against one site from the second dataset. As preperation for this an individual tag is created for each event.

event.latest()

This function is the work horse of the anlaysis. It's primary purpose is to find out how long the difference in days is for each event from one single site from the second dataset from one event from the site given from the first dataset. I know that is confusing, but the way the code needs to be written to allow for parallel processing prevents it from being overly clear... One way to look at it is that every layer in the stack of functions used in the anlysis is reducing the overall number of things being compared. At this layer we have only one event being compared against the events of one site. Also, as one moves down the list of functions, the focus of each functions flip flops, for some odd reason. Meaning if the first function requires an x and y variable, in the next function these will be reversed. This isn't overly important to the theory behind why I performed the analysis the way I did so I won't ramble on about it anymore. After finding the most recent (latest) event, the function then calculates the distance and bearing between the sites, as well as which coastal section of South Africa that site belongs to.

Other functions

Below are given the three other functions mentioned in the two sub-sections above. They don't perform any calculations and so I haven't written any in depth descriptions of them. They are just for convenience sake.

Co-occurrence results

The dataframes produced by this co-occurrence work-flow are rather beefy and so it is necessary to give an explanation for what all of the column headers mean.

```
load("~/AHW/data/cooccurrence/SACTN_SAWS_hw_tmean_CO.Rdata")
head(SACTN_SAWS_hw_tmean_CO, 1, addrownums = FALSE)
```

```
                      index index_start index_stop event_no duration
1 Cape Agulhas - Betty's Bay       27031      27035      162        5
  date_start  date_stop  date_peak int_mean  int_max   int_var  int_cum
1 2005-01-02 2005-01-06 2005-01-06 1.977034 2.264594 0.2646968 9.885171
  int_mean_rel_thresh int_max_rel_thresh int_var_rel_thresh
1           0.3927742          0.6906452          0.2597694
  int_cum_rel_thresh int_mean_abs int_max_abs int_var_abs int_cum_abs
1           1.963871        22.89       23.25   0.2631539      114.45
  int_mean_norm int_max_norm rate_onset rate_decline type         site
1     0.9770342     1.264594  0.1631378    0.7929106  AHW Cape Agulhas
```

```
       SACTN percentile  n ply_index percentile.idx index_start.1
1 Betty's Bay         50 23         1             0           157
  index_stop.1 event_no.1 duration.1 date_start.1 date_stop.1 date_peak.1
1          161          1          5   2005-05-18  2005-05-22  2005-05-19
  int_mean.1 int_max.1 int_var.1 int_cum.1 int_mean_rel_thresh.1
1   1.819191  2.660805 0.4755498  9.095953              0.482671
  int_max_rel_thresh.1 int_var_rel_thresh.1 int_cum_rel_thresh.1
1             1.296058            0.4590966             2.413355
  int_mean_abs.1 int_max_abs.1 int_var_abs.1 int_cum_abs.1 int_mean_norm.1
1        16.5234        17.328     0.4550948        82.617       0.8191906
  int_max_norm.1 rate_onset.1 rate_decline.1 type.1      site.1
1       1.660805     1.042706      0.4051897    MHW Betty's Bay
  percentile.1 latest                    index.2     dist     bear
1            0   -136 Cape Agulhas - Betty's Bay 113.2884 297.2109
  site_coast site.1_coast coast_index
1         sc           sc     sc - sc
```

The disgusting blob of text produced by the above code is the first line of results from "SACTN_SAWS_hw_tmean_CO.Rdata" condensed into a portrait layout. Most of the columns found in these results are the same as those described in Hobday ety al. (2015). Note that intensity is calculated as the degrees Celsius over the daily threshold (the 90th percentile) for temperatures on a given day. The explanation of each column is as follows:

- "index" = the names of the two sites being compared
- "index_start"/ "index_stop" = the row number on which the extreme event starts/ stops in the original daily temperature dataframe
- "event_no" = the number of the event as it occurred chronologically in the events for that site
- "duration" = the length of the event in days
- "date_start"/ "date_stop"/ "date_peak" = the date on which the event started/ stopped/ reached its peak intensity
- "int_mean"/ "int_max" = the mean/ max intensity of the event
- "int_var" = the variance in the daily intensity of the event
- "int_cum" = the cumulative intensity of the event, meaning the total of the intensities of each day
- "int_cum_rel_thresh" = Not sure... can't find it in Hobday et al. (2015) either
- "int_mean_abs", "int_max_abs", "int_var_abs", "int_cum_abs" = the same as the other metrics with similar names described above except given in degrees Celsius
- "int_mean_norm", "int_max_norm" = I believe this shows the mean and max intensities normalised by some metric though I'm not sure what that is...
- "rate_onset", "rate_decline" = these show the rate of onset and decline of the event though I'm not sure what exactly the units are...
- "type" = the type of the events being compared, "A*" = atmospheric, "M*" = marine, "*HW" = heat wave, "*CS" = cold-spell
- "site" = the name of one of the site's being compared
- "SACTN" = the name of the other site, this column will be removed in future versions
- "percentile" = the percentile (based on int_cum) into which the event falls compared to other events from the same site
- "n" = the number of events recorded for this site
- "ply_index" = the unique number used to compare this event to all others
- "percentile.idx" = this is the lowest percentile present in the events, this value changes when smaller events are screened out and is used for computational purposes
- index_start.1", etc. = all columns that end with"*.1" are the exact same as the other columns with the same name, except they show the statistics for the site that the first site was compared against
- "latest" = this value shows how many days the second site started after the first
- "index.2" = this should be the same as "index" and was created to double check that the functions were working correctly

- "dist" = the distance (km) between the two sites being compared
- "bear" = the bearing between the sites being compared, this should probably be standardised from 360degrees to 180degrees
- "site_coast"/" site.1_coast" = the coastal section that the first/ second site belong to, coastal section explanation may be found in Smit et al. (2013)
- "coast_index" = both coastal sections combined, for further analysis purposes

Yikes... So that hopefully won't need to go into the paper. These data are only an intermediate step on the way to the final goal after all.

Figures

And now some pretty figures after reading all of that aweful text. Now that the co-occurrence rates aren't being screened by artificial parametres, it allows for a broader image to be drawn of what the relationships may look like. The initial way I thought of doing that was with density plots.

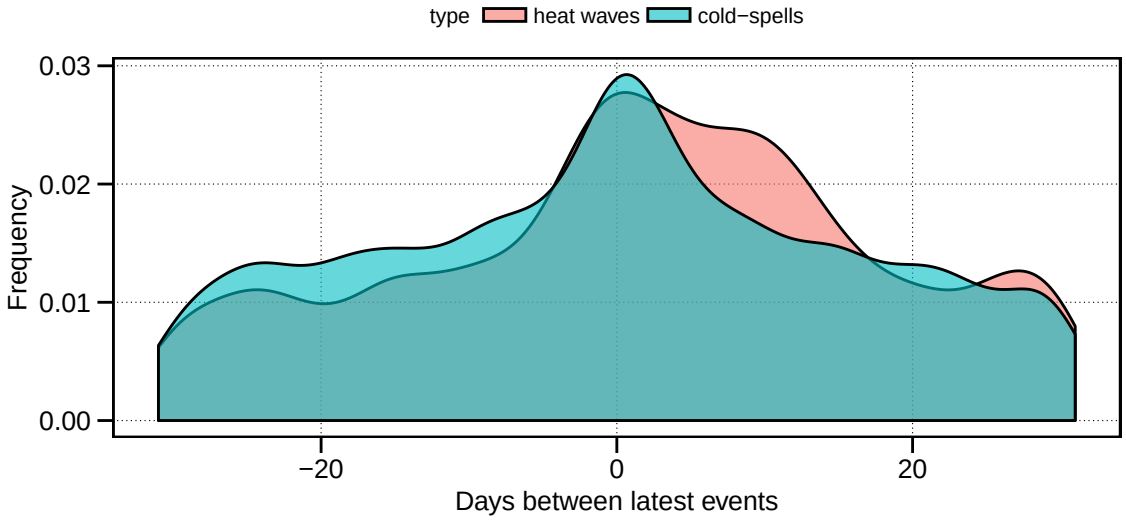

Figure 1: Density plot of days between most recently occurring events after matching all SACTN events to all SAWS events (Tmean). The co-occurrence rangs have been constrained here to only events that occurred within 31 days of one another. Positive values on the x-axis mean that the SACTN event preceeded the SAWS event.

As one may see from these figures there is a tendency towards 0 with all of the different comparisons. Most striking is that when comparing SAWS data to other SAWS data the rates of co-occurrence become very close. This is less so when comparing SACTN data to itself. And even less so when comparing SACTN to SAWS data. Please note that the y-axis is different in each figure but that the x-axis is constant. One consistent theme with all figures is that rates of co-occurrence for heat waves are gathered closer to 0 than cold-spells. We also see in Figure 1 that heat waves in the SACTN dataset more often occur before the SAWS dataset than the other way around. This may however be due to offshore atmospheric forcing that has yet to move onto shore and be recorded by weather stations, as Andries pointed out earlier.

Moving forward

There is still much I could include in this report but I think it has already become much to long so I will stop here for now. AJ and I were talking about what the next step(s) should be and it was decided that a

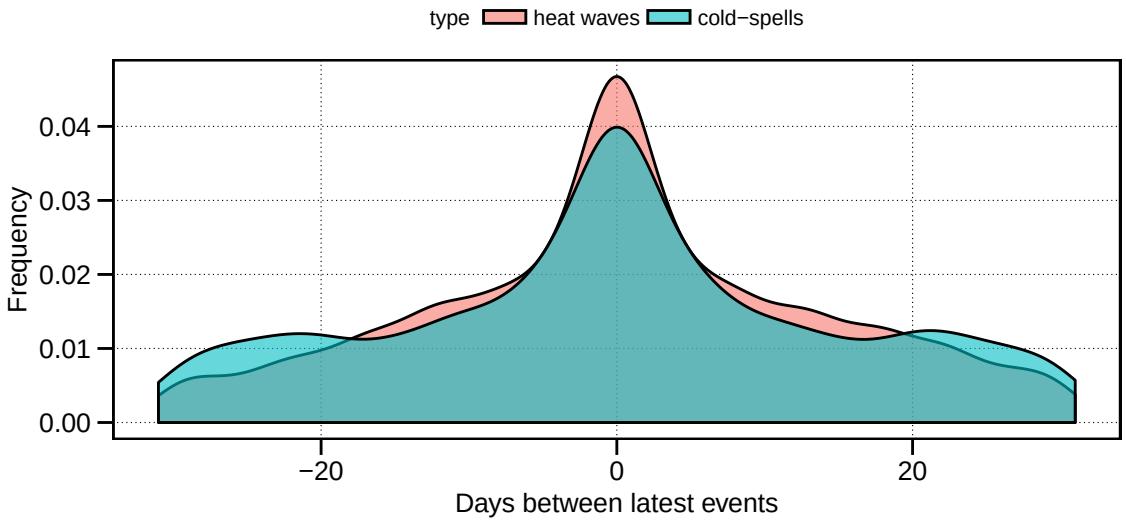

Figure 2: Density plot of days between most recently occurring events after matching all SACTN events to all SACTN events. The co-occurrence rangs have been constrained here to only events that occurred within 31 days of one another.

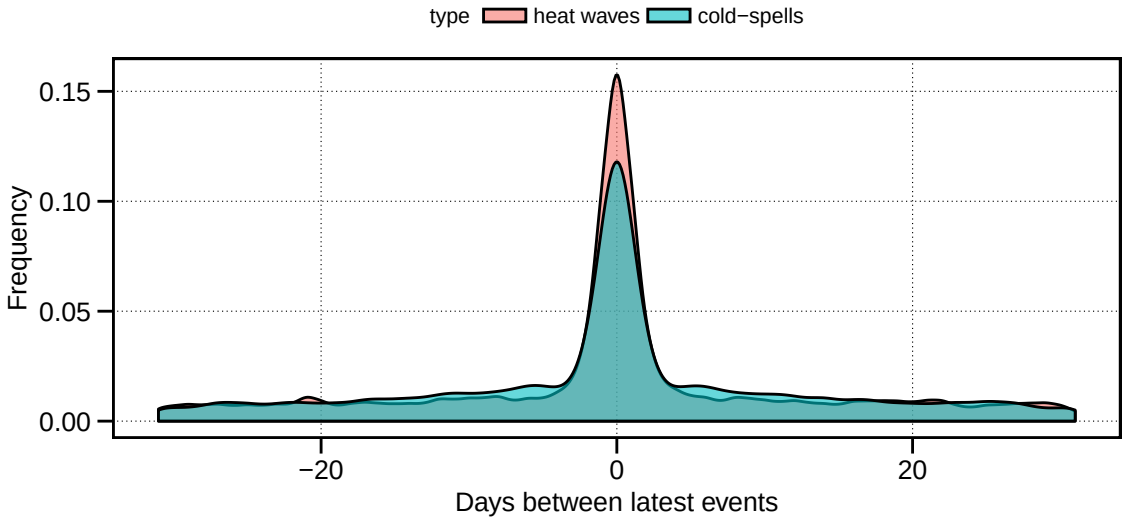

Figure 3: Density plot of days between most recently occurring events after matching all SAWS events (Tmean) to all SAWS events (Tmean). The co-occurrence rangs have been constrained here to only events that occurred within 31 days of one another.

multi-variate analysis must be performed on all of these results in order to get a better idea of which of the multitude of metrics are the most useful for any future analyses. I've got some old code in which I performed an ordination on some sites to classify them based on temperature metrics so I will go and dust that off and think about how to apply it to this. Because I have the distance and bearing between sites I can also calculated and plot which areas of the coast correlate best with which directions of travel. So that is still something that is coming down the pipe. Now that I am calculating co-occurrence a bit differently I also need to rethink how best to create a metric that shows the strength of the co-occurrence between sites. If anyone has any input/ direction they would like to provide please let me know. Otherwise I have a good idea about where to go next. And am excited to do it!