**Climatology anomalies data analysis for Holly Thomas**

STAT688 Statistical Consulting with Dean Billheimer

Hamza Islam Butt ([hamzabutt39@arizona.edu](mailto:hamzabutt39@arizona.edu))

Yifan Zhang ([yifanz1@arizona.edu](mailto:yifanz1@arizona.edu))

Kaiwen Liu([kaiwenliu@arizona.edu](mailto:kaiwenliu@arizona.edu))

February 10, 2023

**Executive Summary**

Holly is a climate scientist investigating extreme winter weather events. She is working with 3-dimensional spatiotemporal datasets (from a simulated climate model and observational data with anthropogenic trend, respectively) on anomalies in climatological measures such as temperature and sea level pressure. The primary goal is to assess if the distribution of anomalies corresponding to ± 40-day interval around extreme weather events is significantly different from that for all winter season days. This hypothesis needs to be tested at each geographical grid point (latitude x longitude combination).  The main recommendation is to use a one-sample t-test to test if the 40-day ( before/after) average of anomalies of extreme events is significantly different than that of all winter days. Since these tests are carried out separately for each geographical grid point, a false discovery rate (FDR) correction is also recommended to minimize false discoveries due to multiple hypothesis testing. For the observational data with anthropogenic trend, it is suggested to detrend the anomaly values using simple linear regression (or more advanced methods) before carrying out the t-tests.

**Detailed Summary**

**Background**

Holly’s research focuses on the following two datasets:

i) ‘model\_data\_no\_warming.nc’ is the simulation dataset  
ii) ‘observational\_data\_warming\_trend.nc’ is the dataset with the anthropogenic trend

In both datasets, surface temperature is the variable of interest. Holly is only interested in extreme winter weather events, so only the data from the winter season would be analyzed. Since the datasets are three dimensional (longitude, latitude and time), the datasets have been dissected into grid cells, and anomaly calculations have been performed in each individual grid cell. For each extreme event identified by the anomaly score, at each spatial location, average  
anomalies over forty days before the event and average anomalies over forty days after the event

have been pulled out to form the subsets. The main goal is to compare the  
distribution of the subsets that have been identified as the extreme winter weather anomaly events  
with the distribution of all anomaly days of the winter weather season.

**Methods**

A subset of each dataset (with and without anthropogenic trend) was provided. In each case, there were approximately 150 extreme weather events, though the exact dates weren’t available. The first step involved converting the raw climatological variables into anomalies. For the data with anthropogenic trend, time series plots were used to assess trend in the anomalies which was adjusted before proceeding. Winter days defined using the winter solstice (Dec 21 – Mar 20th) were appropriately filtered out; amongst these 150 days were randomly chosen to represent ‘extreme events’. The average anomalies for ± 40-day windows for extreme events and all winter days were computed (separately for each grid point); histograms were used to compare the distributions. Finally, one sample t-tests were carried out separately for each grid point (with and without false discovery rate correction) and their corresponding p-values presented as heatmaps. The statistical analysis was performed using R version 4.2.0; the R-code is available in an appendix at the end of this document.

**Statistical Analysis**

**Model dataset (no anthropogenic trend)**

1. Characteristics: The primary climatological variable is surface air temperature (TAS) measured in Kelvins. It is measured in three dimensions; longitude (30 points, ranging from 150.0-186.25°E), latitude (40 points ranging from 23.09-59.84°N) and time (23725 points, measured in daily increments as days since 1950-01-01. This cumulatively amounts to 65 years). TAS anomalies were calculated for every individual grid cell (total of 30 longitude x 40 latitude = 1200 grid cells) using the same method used by Cohen et al. (2018) as follows:-
   1. Calculate the average TAS for each day of the year, at every location, using the 31-year reference period 1980-01-01 – 2009-12-31 (referred to as climatological mean).
   2. The climatological mean TAS is subtracted from the raw TAS measurement at every time point. For example, at the geographical grid point 23.09°N,150°E on October 1st, 1960, the anomaly is equal to the TAS that day minus the climatological mean TAS for 1st October at 23.09°N,150°E.

The data is stored in the NetCDF format (.cd4 extension)

1. Statistical analysis:
   1. All statistical analysis was performed using R. The *netcdf4* package was used to read in the dataset, and the spatiotemporal dimension values (latitude, longitude, and time) extracted into vectors. [1]
   2. The climatological mean values for the reference period 1980-01-01 – 2009-12-31 were calculated using base R loops and netcdf4 functions. The resulting array has dimensions 30x40x23725 (lon x lat x time). [2]
   3. Winter days were defined using the winter solstice (Dec 21 – Mar 20th). For years 1950-2014, there were a total of 5760 winter days that were subset.
   4. A random sample of 150 was generated from within the above selected winter days to represent ‘extreme weather events’.
   5. The average anomaly for ‘40 days before’ or ‘40 days after’ a date is then calculated into an R array separately at each grid point location for:
      1. Extreme event days (n=150)
      2. The population of all winter days (n=5760). [3]
   6. Histograms (Fig. 1) were generated to compare how the distribution of average TAS anomalies (40 days before or 40 days after) compared between extreme events and all winter days at one grid point (23.1°N,150°E). [4]

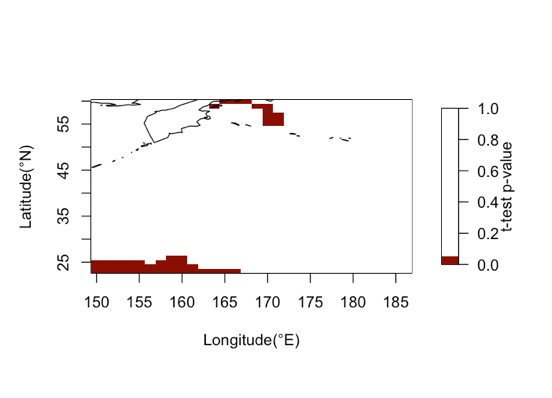
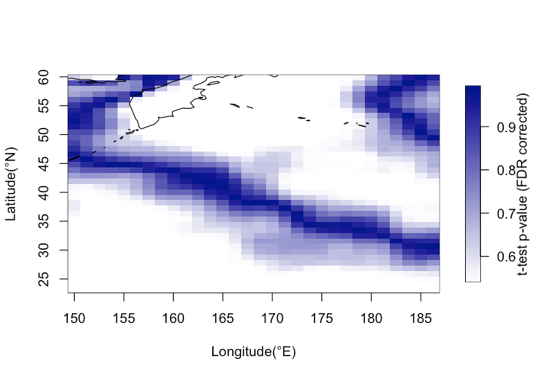
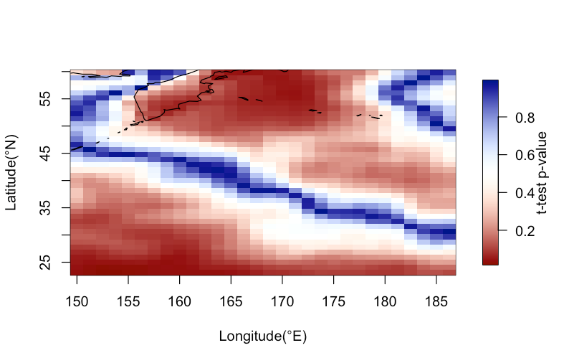
***Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated****Fig. 1: Histograms showing the comparison of average TAS anomalies between extreme events and all winter days for 40 days before(left) and 40 days after(right) the date.*

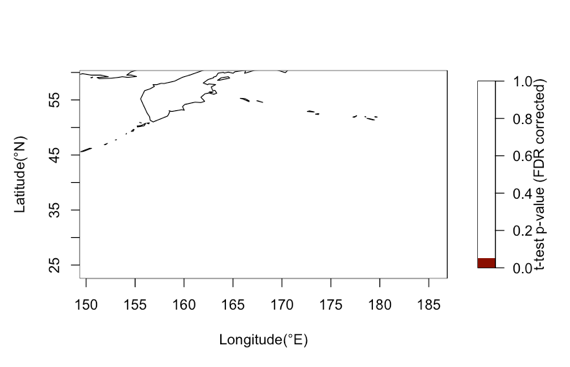
The distributions appear to be very similar, which is expected since the 150 days were randomly sampled.

* 1. One sample t-test was carried out to test the hypothesis if the mean of the extreme weather events anomalies (40 days before or 40 days after) is significantly different than that of mean of all winter days anomalies. Since the sample size of extreme events is 150, the central limit theorem is applicable. This results in a total of 30x40=1200 t-tests, whose p-values were extracted and represented as lat x lon heatmap. A stippled plot was also generated, which only highlights p-values <0.05.
  2. False Discovery Rate (FDR) was also carried out on the raw p-values in the above step separately, to account for false discoveries during multiple hypothesis testing. Heatmaps (original and stippled) (Fig. 2) were similarly generated. [5]



A

B



D

C

*Fig. 2: Heatmaps showing p-values at each of the 1200 geographical grid points from one sample t-tests comparing the average ‘40 day after’ TAS anomaly for extreme events to that for all winter days. A & B show p-values from the t-tests without FDR correction (A= original heatmap, B= stippled plot showing p-values <0.05); C& D show p-values after FDR correction (C= original heatmap, D= stippled plot showing p-values <0.05).*

It can be observed that after FDR correction, there are no significant differences in the anomaly values at any of the 1200 grid points.

**Observational study data (with anthropogenic trend)**

1. Characteristics: The primary climatological variable is 2-metre temperature (T2M) measured in Kelvins. It is measured in three dimensions; longitude (34 points, ranging from 150.38-185.38°E), latitude (36 points ranging from 23.09-59.84°N) and time (23376 points, measured in daily increments as days since 1959-01-01 (until 2022-12-31). This cumulatively amounts to 64 years).
2. Statistical analysis:
3. This dataset required some cleaning prior to usage:
4. There were ‘NA’ values corresponding to the T2M variable in the duration 2020-07-06 to 2020-07-31. Hence, only data up till the end of the previous year (i.e., 2019-12-31) was selected.
5. The observational data contains ‘29th Feb’ measurements for all leap years, which was not present in the model data. To simplify analyses, these were removed.
6. Anomaly calculation: T2M anomalies were calculated for every individual grid cell (total of 34 longitude x 36 latitude = 1224 grid cells) using the same strategy used for the model dataset. The reference period for climatological mean was the same as that used for model dataset i.e., 1980-01-01 – 2009-12-31. [6]
7. Detrending: The time series of anomalies at a single grid point (56.6°,150°E) was plotted to check for a trend. The graph shows a fluctuating and a slightly increasing trend over time. The detrend function in the R package *astsa* was used to remove the increasing trend component using simple linear regression (polynomial of order 1). The graph (Fig. 3) or the same time series after detrending shows that this trend was removed. The detrended data was used for all subsequent analyses. [7]

**A picture containing chart

Description automatically generatedChart, line chart

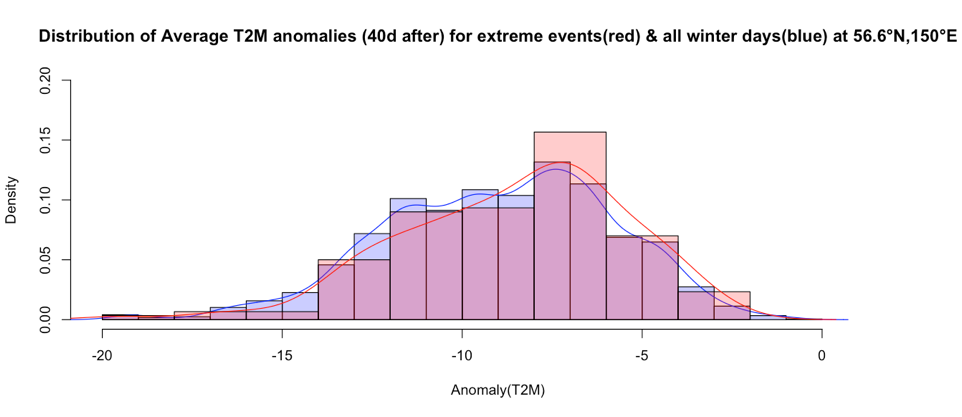
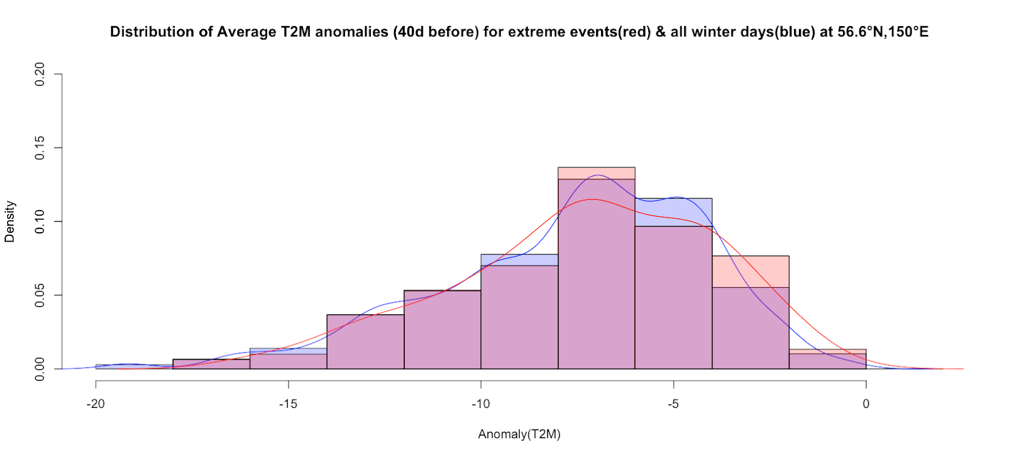
Description automatically generated**

*Fig. 3: Time series of T2M anomalies at 56.6°,150°E before (left) and after(right) detrending*

1. The subsequent steps in the analysis were the same as those used for the model data. There were 5400 winter days, from which 150 were randomly selected to represent extreme events. The histograms (for the grid point 56.6°,150°E), and all heatmaps from the analysis are given below (Fig. 4):

**Chart, histogram

Description automatically generated**

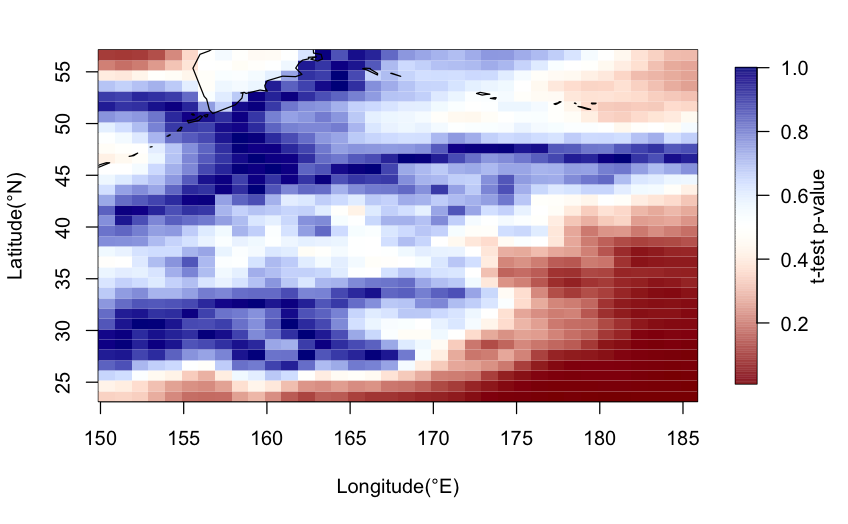
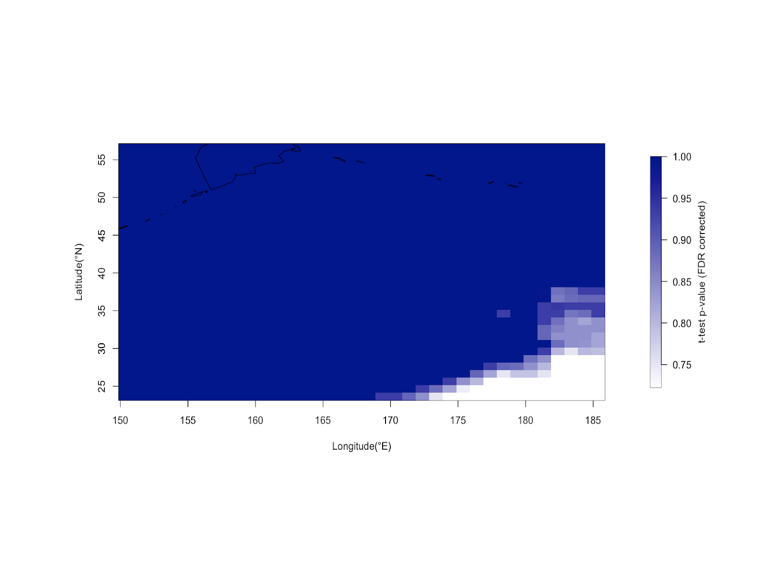
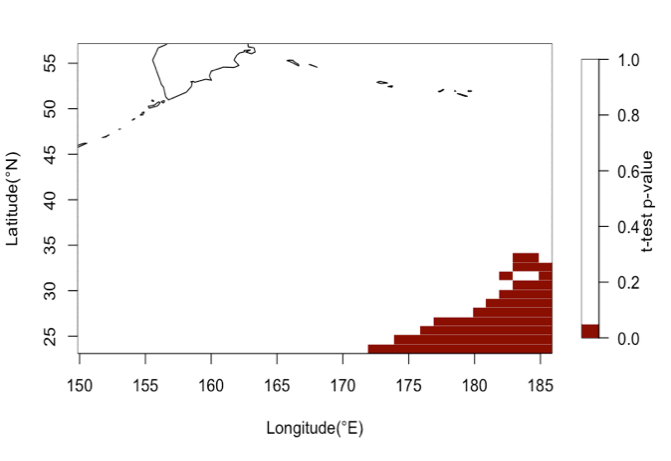
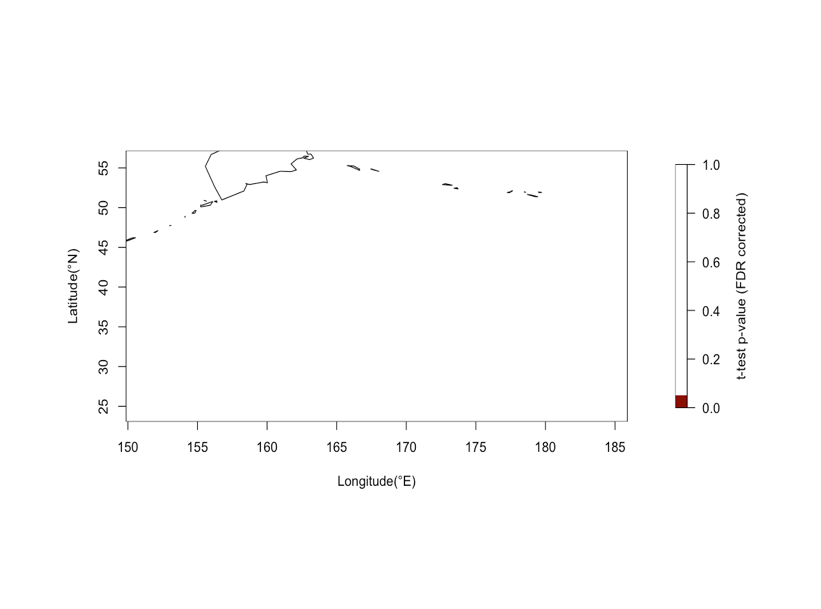


Chart, histogram

Description automatically generated

*Fig. 4: Histograms showing the comparison of average detrended T2M anomalies between extreme events and all winter days for 40 days before(top) and 40 days after(bottom) the date.*

The distributions appear to be very similar, which is expected since the 150 days were randomly sampled.



A

B

C

D

*Fig. 5: Heatmaps showing p-values at each of the 1224 geographical grid points from one sample t-tests comparing the average ‘40 day after’ detrended T2M anomaly for extreme events to that for all winter days. A & B show p-values from the t-tests without FDR correction (A= original heatmap, B= stippled plot showing p-values <0.05); C& D show p-values after FDR correction (C= original heatmap, D= stippled plot showing p-values <0.05).*

It can be observed that after FDR correction, there are no significant differences in the anomaly values at any of the 1224 grid points.

**Limitations**

One limitation of the analysis of the observational data (with anthropogenic trend) is that a simple approach involving linear regression was used to remove the trend. A more rigorous approach involving assessment of other trend patterns such as adaptive and multidecadal could be explored (Wu et al., 2007).

**References**

Cohen, J., Pfeiffer, K., & Francis, J. A. (2018). Warm Arctic episodes linked with increased frequency of extreme winter weather in the United States. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-02992-9

Wu, Z., Huang, N. E., Long, S. R., & Peng, C. K. (2007). On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(38), 14889–14894. https://doi.org/10.1073/pnas.0701020104

**Appendix**

(a)

library(ncdf4) # package for netcdf manipulation  
library(raster) # package for raster manipulation  
library(rgdal) # package for geospatial analysis  
library(ggplot2) # package for plotting  
library(fields)  
library(maps)  
library(stringr)

#Read in the data and save the spatiotemporal dimensions as vectors  
nc\_data <- nc\_open('model\_data\_no\_warming.nc')

lon <- ncvar\_get(nc\_data, "lon")

lat <- ncvar\_get(nc\_data, "lat",verbose=FALSE)

t <- ncvar\_get(nc\_data, "time")

#transform times into actual dates  
t\_trans<-as.Date(t,origin="1950-01-01",tz="MST")

(b)

#Subset the data to retain only the 1980 - 2010 data   
temps\_1980\_2010 = ncvar\_get(nc\_data,"tas",start=c(1,1,10951),count=c(-1,-1,11315))  
#Transform array to group by years and dates   
temps\_1980\_2010\_transform = array(0,dim=c(30,40,31,365))  
for(i in 1:30 ){  
 for(j in 1:40){  
 l=1  
 for(k in 1:31){  
 temps\_1980\_2010\_transform[i,j,k,]<-temps\_1980\_2010[i,j,c(l:(l+364))]  
 l=l+365  
 }  
 }  
}  
  
#Calculate the daily averages into a new array  
dailyaverages\_1980\_2010= array(0,dim=c(30,40,365))  
for(i in 1:30 ){  
 for(j in 1:40){  
 for(k in 1:365){  
 dailyaverages\_1980\_2010[i,j,k]<-mean(temps\_1980\_2010\_transform[i,j,,k])   
 }  
 }  
}  
###Calculation of anomalies for every day at every location  
  
  
#1) Generate arrays in appropriate dimensions  
  
temps\_all = ncvar\_get(nc\_data,"tas",start=c(1,1,1),count=c(-1,-1,23725))  
daily\_anomalies<-array(0,dim=c(30,40,65,365))  
temps\_all\_transform = array(0,dim=c(30,40,65,365))  
for(i in 1:30 ){  
 for(j in 1:40){  
 l=1  
 for(k in 1:65){  
 temps\_all\_transform[i,j,k,]<-temps\_all[i,j,c(l:(l+364))]  
 l=l+365  
 }  
 }  
}  
  
#2) Calculate the anomalies into the array named ‘daily\_anomalies’   
  
for(i in 1:30 ){  
 for(j in 1:40){  
 for(k in 1:65){  
 for(l in 1:365){  
 daily\_anomalies[i,j,k,l]<-temps\_all\_transform[i,j,k,l]-dailyaverages\_1980\_2010[i,j,k]  
 }  
 }  
 }  
}  
  
#3) Convert anomalies into a format called ‘daily\_anomalies\_long’ for downstream processing:  
daily\_anomalies\_long<-array(0,dim=c(30,40,23725))  
  
for(i in 1:30 ){

for(j in 1:40){  
 l=1  
 for(k in 1:65){  
 daily\_anomalies\_long[i,j,l:(l+364)]<-daily\_anomalies[i,j,k,]  
 l=l+365  
 }  
 }  
}

(c)

#Vector to extract only the winter date indices  
winter\_days<-vector()  
for(i in 1:64){  
 winter\_days\_yeari<-seq((355+(i-1)\*365),length.out = 90,by=1)  
 winter\_days<-c(winter\_days,winter\_days\_yeari)  
}  
#Array that contains averages for the +/- 40 day window of anomalies for 150 randomly selected extreme events from the winter days  
set.seed(100)  
n<-150  
indexes<-sample(winter\_days,n)  
anomalies\_array\_extreme40<-array(0,dim=c(30,40,n,2))  
  
for(i in 1:30){  
 for(j in 1:40){  
 for(k in indexes){  
 # anomalies\_array\_extreme40[i,j,which(indexes==k),1:81]<-daily\_anomalies\_long[i,j,(k-40):(k+40)]  
 anomalies\_array\_extreme40[i,j,which(indexes==k),1]<-mean(daily\_anomalies\_long[i,j,(k-40):(k-1)])  
 anomalies\_array\_extreme40[i,j,which(indexes==k),2]<-mean(daily\_anomalies\_long[i,j,(k+1):(k+40)])  
 }  
 }  
}  
  
#Array that contains the averages for the +/- 40 day window of anomalies for all winter dates (Dec21-Mar20)  
#Starts at Dec 21,1950. Ends at Mar 20, 2014  
n2<-winter\_days #The final date in dataset was removed for simplicity  
  
anomalies\_array\_pop40<-array(0,dim=c(30,40,length(n2),2))  
  
for(i in 1:30){  
 for(j in 1:40){  
 for(k in n2){  
 anomalies\_array\_pop40[i,j,which(n2==k),1]<-mean(daily\_anomalies\_long[i,j,(k-40):(k-1)])  
 anomalies\_array\_pop40[i,j,which(n2==k),2]<-mean(daily\_anomalies\_long[i,j,(k+1):(k+40)])  
 }  
 }  
}

(d)

#Example Histogram (showing a comparison of distribution of anomalies of extreme vs pop at 1 grid point)  
  
p1<-hist(anomalies\_array\_pop40[1,1,,2],probability=TRUE)  
p2<-hist(anomalies\_array\_extreme40[1,1,,2],probability=TRUE)  
  
plot(p1, col=rgb(0,0,1,1/4), xlim=c(-3,3),ylim=c(0,0.45),freq = FALSE,main="Distribution of Average TAS anomalies (40d after) for extreme events(red) & all winter days(blue) at 23.1°N,150°E",xlab="Anomaly(TAS)") # first histogram  
lines(density(anomalies\_array\_pop40[1,1,,2]), col = "blue")   
plot(p2, col=rgb(1,0,0,1/4), xlim=c(-3,3), add=T,freq=FALSE)  
lines(density(anomalies\_array\_extreme40[1,1,,2]), col = "red") # second histogram

(e)

##One sample t-tests: Separately carry out for each grid point

pval\_before40d<-pval\_after40d<-pvalfdr\_after40d<-array(0,dim=c(30,40))  
  
for(i in 1:30){  
 for( j in 1:40){  
 pval\_before40d[i,j]<-(t.test(anomalies\_array\_extreme40[i,j,,1],mu=mean(anomalies\_array\_pop40[i,j,,1])))$p.value  
 pval\_after40d[i,j]<-(t.test(anomalies\_array\_extreme40[i,j,,2],mu=mean(anomalies\_array\_pop40[i,j,,2])))$p.value  
   
 }  
}  
  
#FDR correction  
  
pvalfdr\_before40d<-matrix(p.adjust(pval\_before40d,method="fdr"),nrow=nrow(pval\_before40d),byrow=FALSE)  
  
pvalfdr\_after40d<-matrix(p.adjust(pval\_after40d,method="fdr"),nrow=nrow(pval\_after40d),byrow=FALSE)  
  
colorTable<- designer.colors(500, c( "dark red","white","dark blue"),   
 x = c(0, 50, 100) / 100)  
  
colorTable2<- designer.colors(500, c("white","dark blue"),   
 x = c(0, 100) / 100)  
  
  
#p-value spatial plots (before and after FDR correction)  
image.plot(lon,lat,pval\_after40d,legend.lab = "t-test p-value",col=colorTable,xlab="Longitude(°E)",ylab="Latitude(°N)")  
map(add=T)  
  
image.plot(lon,lat,pvalfdr\_after40d,legend.lab = "t-test p-value (FDR corrected)",col=colorTable2,xlab="Longitude(°E)",ylab="Latitude(°N)")  
map(add=T)

(f)

#Read in the data and extract vectors for spatiotemporal dimensions  
  
nc\_data <- nc\_open('observational\_data\_warming\_trend.nc')

lon <- ncvar\_get(nc\_data, "longitude")  
lat <- ncvar\_get(nc\_data, "latitude",verbose=FALSE)  
t <- ncvar\_get(nc\_data, "time")

#transform times into actual dates  
t\_trans<-as.Date(t,origin="1959-01-01",tz="MST")  
  
#transform t and t\_trans to retain all days except the last 3 years (due to NA value in t2m)  
  
t<-t[1:which(t\_trans=="2019-12-31")]  
t\_trans<-t\_trans[1:which(t\_trans=="2019-12-31")]

#Subset the data to retain only the 1980 - 2010 data   
  
#Upon inspection, this data contains all the corresponding dates for '29th Feb' for leap years. For ease of processing, these were removed  
  
t\_trans\_1980\_2010<-t\_trans[7671:18993]  
  
feb\_29ind\_1980\_2010<-which(str\_detect(t\_trans\_1980\_2010, '02-29'))  
  
#t\_trans\_ref<-t\_trans\_1980\_2010[-feb\_29ind\_1980\_2010]  
  
#The start/end positions of this period can be determined using the R command:  
#which(t\_trans=="1980-01-01")  
  
temps\_1980\_2010 = ncvar\_get(nc\_data,"t2m",start=c(1,1,7671),count=c(-1,-1,11323))  
  
#remove the Feb 29 data from the temps\_1980\_2010 dataframe  
temps\_1980\_2010<-temps\_1980\_2010[,,-feb\_29ind\_1980\_2010]  
  
#Transform array to group   
#interval<-c(seq(1,11315,by=365))  
  
temps\_1980\_2010\_transform = array(0,dim=c(36,34,31,365))  
  
for(i in 1:36 ){  
 for(j in 1:34){  
 l=1  
 for(k in 1:31){  
 temps\_1980\_2010\_transform[i,j,k,]<-temps\_1980\_2010[i,j,c(l:(l+364))]  
 l=l+365  
 }  
 }  
}  
  
#Calculate the daily averages into a new array  
dailyaverages\_1980\_2010= array(0,dim=c(36,34,365))  
  
for(i in 1:36 ){  
 for(j in 1:34){  
 for(k in 1:365){  
 dailyaverages\_1980\_2010[i,j,k]<-mean(temps\_1980\_2010\_transform[i,j,,k])   
 }  
 }  
}  
  
###Calculate the anomalies for every day at every location

#1) Generate arrays in appropriate dimensions  
  
temps\_all = ncvar\_get(nc\_data,"t2m",start=c(1,1,1),count=c(-1,-1,22280))  
feb\_29ind\_temps\_all<-which(str\_detect(t\_trans, '02-29'))  
  
temps\_all<-temps\_all[,,-feb\_29ind\_temps\_all]  
daily\_anomalies<-array(0,dim=c(36,34,61,365))  
temps\_all\_transform = array(0,dim=c(36,34,61,365))  
  
  
for(i in 1:36 ){  
 for(j in 1:34){  
 l=1  
 for(k in 1:61){  
 temps\_all\_transform[i,j,k,]<-temps\_all[i,j,c(l:(l+364))]  
 l=l+365  
 }  
 }  
}  
  
#2) Calculate the anomalies into the array named ‘daily\_anomalies’  
  
for(i in 1:36 ){  
 for(j in 1:34){  
 for(k in 1:61){  
 for(l in 1:365){  
 daily\_anomalies[i,j,k,l]<-temps\_all\_transform[i,j,k,l]-dailyaverages\_1980\_2010[i,j,k]  
 }  
 }  
 }  
}  
  
  
#3) Convert anomalies into a format called ‘daily\_anomalies\_long’ for downstream processing:  
daily\_anomalies\_long<-array(0,dim=c(36,34,22265))  
  
for(i in 1:36 ){  
 for(j in 1:34){  
 l=1  
 for(k in 1:61){  
 daily\_anomalies\_long[i,j,l:(l+364)]<-daily\_anomalies[i,j,k,]  
 l=l+365  
 }  
 }  
}

(g)

##Trend analysis and adjustment for the anomalies

t\_filt<-t[-feb\_29ind\_temps\_all]  
plot(t\_filt,daily\_anomalies\_long[1,1,],type="l",xlab="t(days since 1959-01-01)",ylab="T2M Anomaly",main="Time series of T2M anomalies at 56.6°N,150°E",col="blue")  
  
#Upon visual inspection, there appears to be a trend (slight increase in average anomalies over time)  
  
#Use a linear regression (polynomial order 1) to detrend the anomaly time series at every location. The R package astsa was used for this purpose

library(astsa)  
  
daily\_anomalies\_long\_detrended<-array(0, dim=dim(daily\_anomalies\_long))  
  
for(i in 1:36){  
 for(j in 1:34){  
 daily\_anomalies\_long\_detrended[i,j,]<-detrend(daily\_anomalies\_long[i,j,],order=1)  
 }  
}  
  
#plot the detrended time series  
  
plot(t\_filt,daily\_anomalies\_long\_detrended[1,1,],type="l",xlab="t(days since 1959-01-01)",ylab="T2M Anomaly(Detrended)",main="Time series of T2M anomalies(detrended) at 56.6°N,150°E",col="maroon")