Preparing data for extreme events and climate change analysis using MATLAB

Alexandre Martinez, University of California, Irvine

Teaching Computation in the Sciences Using MATLAB

Introduction

Because of the recent increase in the frequency of extreme weather events widely attributed to climate change, there is a concern in the scientific community about the intensity and duration of hydrological disaster such as drought. Drought is typically defined using indices based on precipitation (Such as the Standard Precipitation Index, SPI), sometimes combined with other variables (temperature, soil moisture, etc.). Therefore precipitation is the main component. The intensity of drought, also known as drought severity, is defined as the cumulative (or average) deviation of SPI (Standard Precipitation Index) below a threshold level, while the drought duration is defined as the time period over which drought occurs (J. W. Kwak, 2012).

One issue is to have a consistent long record of data. We are now looking into these issues.

Learning objectives

The principal obective of this activity is to be critical when using data and be able to quickly indentify a dataset with low consistency.

Data and method

We are using CRU (Climate Research Unit) precipitation dataset, which is a monthly gridded (0.5 degree) product. For technical reasons (processing time) we have clipped the data to North America only. We are now exploring the datasets to see if it is suitable for climate change analysis. Climate change refers to long period of time and drought is an extreme event, which means that it doesn't happen every year. Therefore, if we want to study a 100-years return period, we need at least 100 years of data. The CRU datasets starts in 1901 so should be suitable.

Data source: https://crudata.uea.ac.uk/cru/data/hrg/

Instruction

Data import

First, we need to load the data and perform a quick inspection

```
load([pwd '\dataFolder\prcp.mat']);     % Load the dataset
size(data)
```

```
ans = 1×3
```

```
   240   181   1272
```

We have a matrix of 240x181x1272, which means we have 1272 months or 106 years of record. This looks right.

Data inspection

Let's now display the cumulative precipitation.

```
load coast;                              % Load the coastline for mapping
sLat        = -30:0.5:60;
sLon        = -180:0.5:-60.5;
clab        = 'Precipitation [mm/mo]';
figure('Name', 'Monthly precipitation');
axesm('MapProjection', 'miller','GLineWidth', 1.0, 'MeridianLabel', 'on','MLabelParallel'...
, 'south','ParallelLabel', 'on','MLineLocation', 60, 'PLineLocation', 30,...
'MapLonLimit',[-180 -60],'MapLatLimit',[0 59.5],'FFaceColor',[.75 .75 .75]);
pcolorm(sLat, sLon, squeeze(nansum(data,3))');
plotm(lat, long, 'k');
colormap(flipud(jet));
framem on; gridm on; tightmap; title('Cumulative precipitation');
c               = colorbar;
c.Label.String  = clab;
```

Question: Why the ocean and most of USA are in red colors?

Answer: Ocean is set-up as 0 values (not -99 or NaN as in other dataset). We also notice most of the map is red, probably because of few pixel have extremely high values, so the colormap put most of the pixel toward the 0 values.

Activity: Write a function to plot the distribution of the cumulative precipitaiton and use it to remap the previous dataset.

Solution: Here is one function

```
cumulPrcp = nansum(data,3);
figure()
histogram(reshape(cumulPrcp,1,240*181));
```

We don't see anything, let's change the bins of the histogram and remove the ocean's values.

```
cumulPrcp(cumulPrcp==0) = NaN;
figure()
histogram(reshape(cumulPrcp,1,240*181),[0:10^6:10^7]);
```

Conclusion: We can see now that most of the data falls below 4x10^6. The other data are either an extreme local phenomena or some errors. Let's change our color scale and plot the map again.

```
figure('Name', 'Monthly precipitation');
axesm('MapProjection', 'miller','GLineWidth', 1.0, 'MeridianLabel', 'on','MLabelParallel'...
, 'south','ParallelLabel', 'on','MLineLocation', 60, 'PLineLocation', 30,...
'MapLonLimit',[-180 -60],'MapLatLimit',[0 59.5],'FFaceColor',[.75 .75 .75]);
pcolorm(sLat, sLon, cumulPrcp');
plotm(lat, long, 'k');
colormap(flipud(jet));
framem on; gridm on; tightmap; title('Cumulative precipitation');
c               = colorbar;
c.Label.String  = clab;
caxis([0 4*10^6])
```

The map looks nicer and we expect Central America to be wetter than USA, and the Gulf of Mexico to be the wettest part. Let's now plot the time series of global preciptiation.

Let's now plot the time evolution.

```
figure()
plot(nansum(nansum(data,2),3));
title('Global precipitation');
xlabel('Year');
```

Question: By looking at this time serie, what can you say about the dataset and why do you think it is like this? Develop some hypothesis and write a code to test them.

Solutions:

What we immediately notice is that the global precipitation is increasing with time, by a factor 10. Is it reasonable? **NO**. Do we get 10 times more rain than before? **NO**.

Students should remember that they are dealing with measured precipitation, if we measure less precipitation in the past, it is not because it was raining less, but because there was probably less rain gages, less observations.

Students should develop some strategies to understand what is going on, here are a few possibilities:

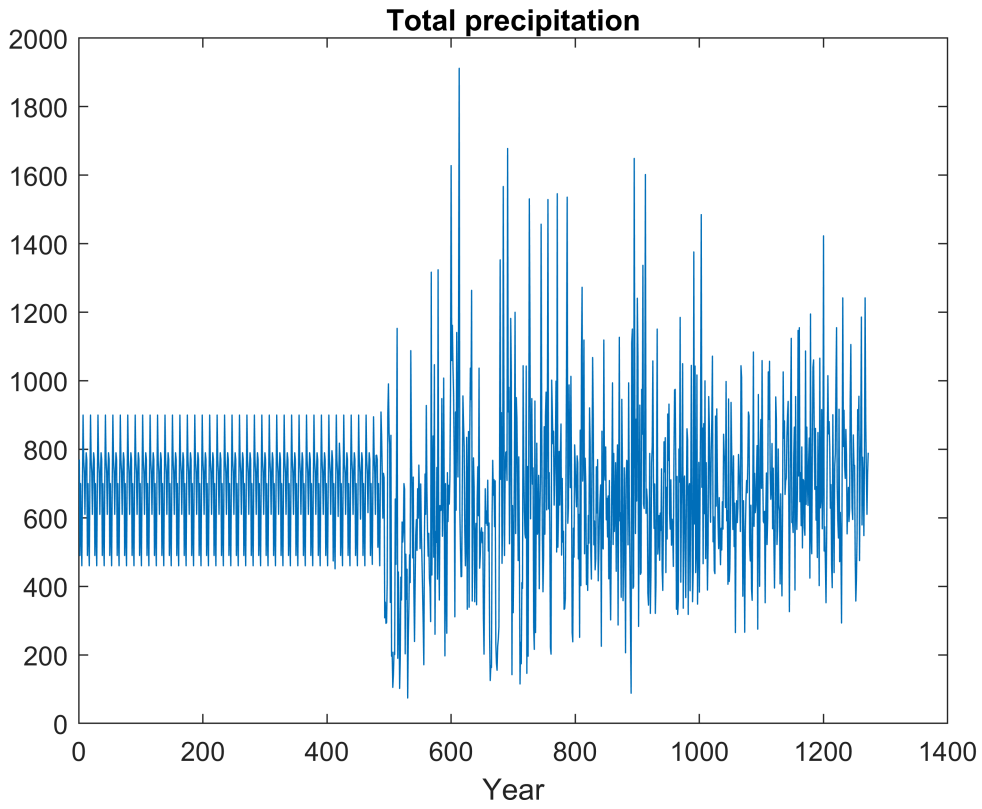
- Precipitation set to another number smaller than the actual value (NaN, 0, etc) when no data.
- Some pixels are missing data because there were no measurement made at the time (no rain gages installed).

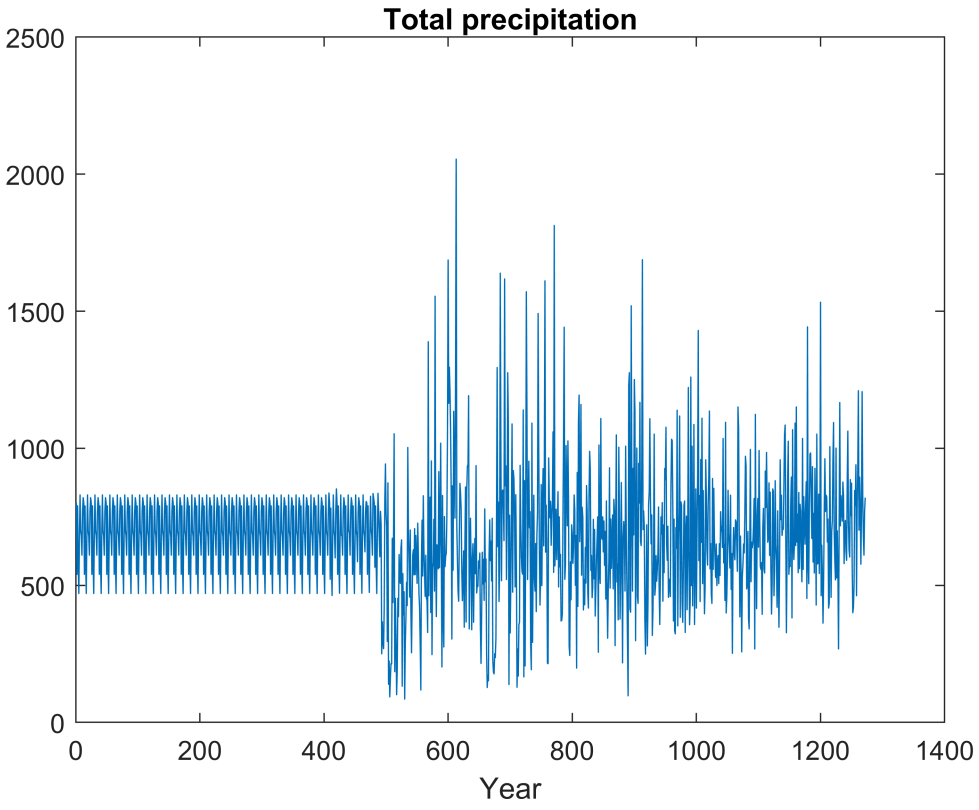
- We can select some pixel randomly and check them, or thing about area that are difficul to access (desert, mountains) or less populated (so less rain gages).

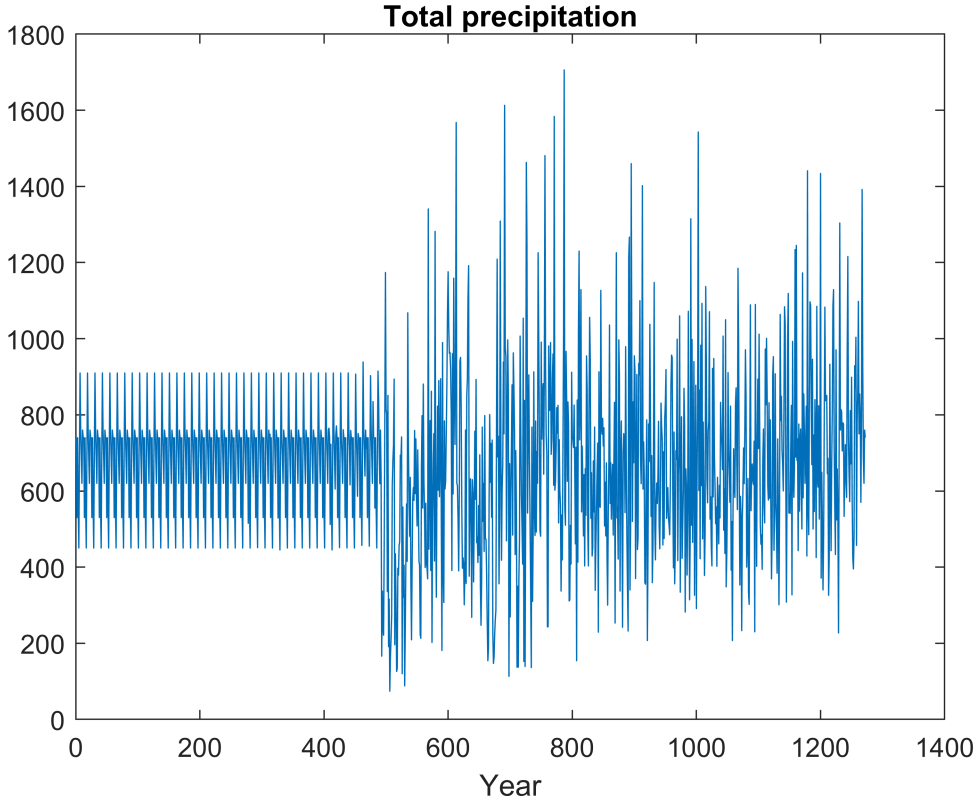
Data inspection - Random pixel selection

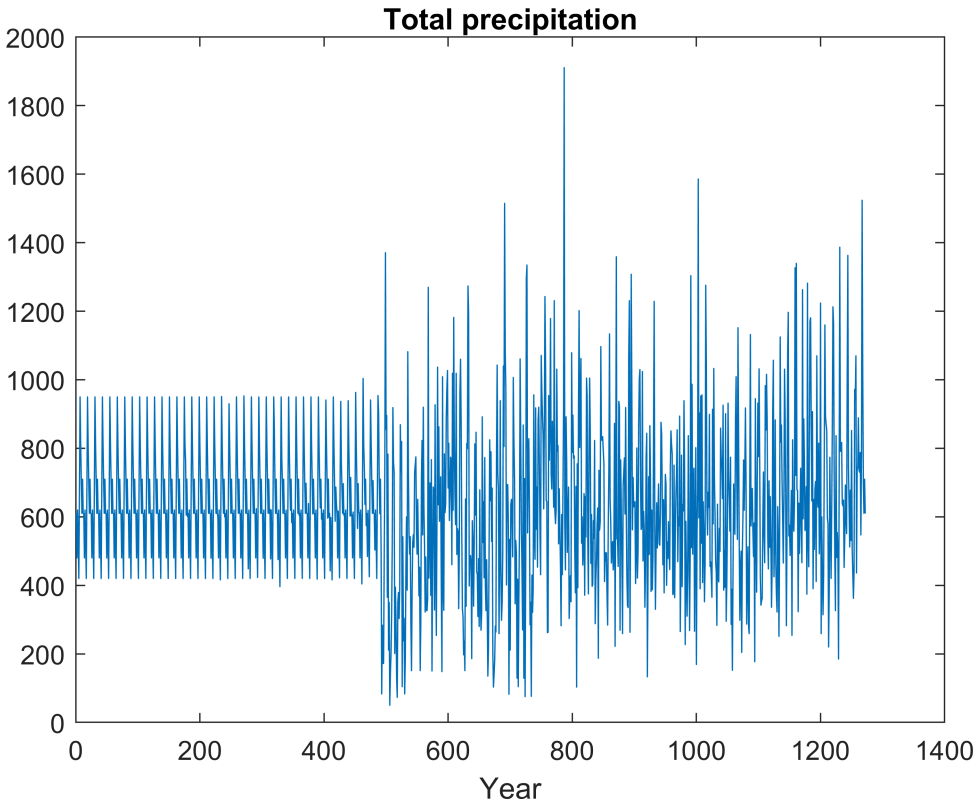
Let's look at some random pixels that we can suspect to have missing early data (let say in Grand Canyon).

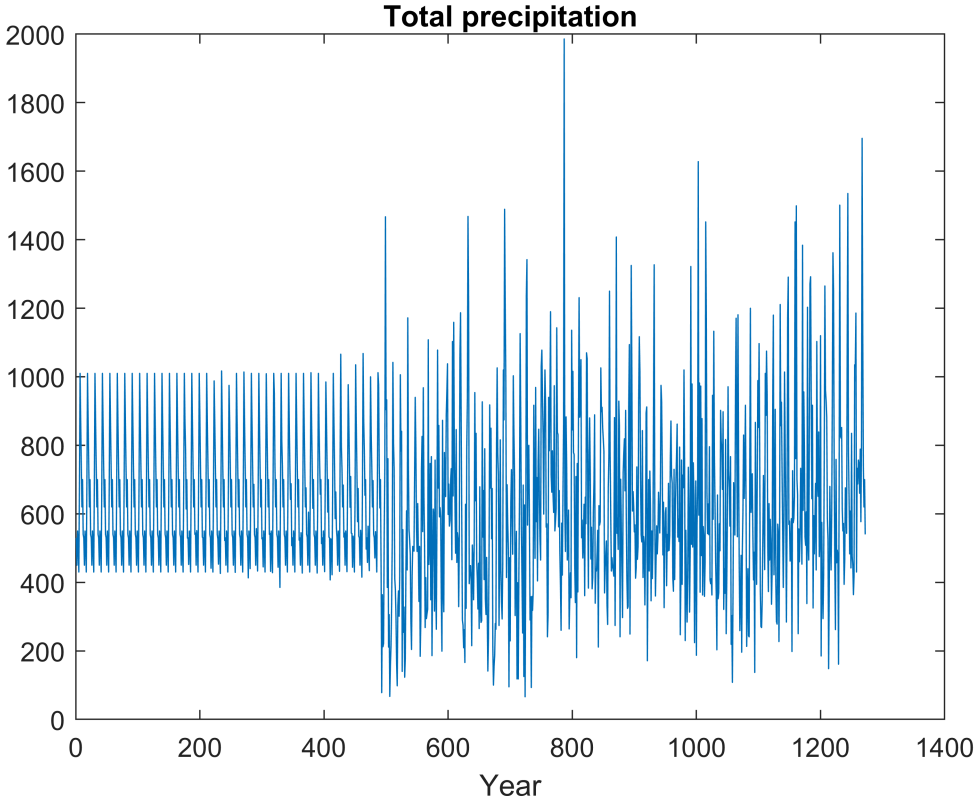
```
for i=1:20
    figure()
    plot(squeeze(data(240-i,180-i,:)));
    title('Total precipitation');
    xlabel('Year');
end
```

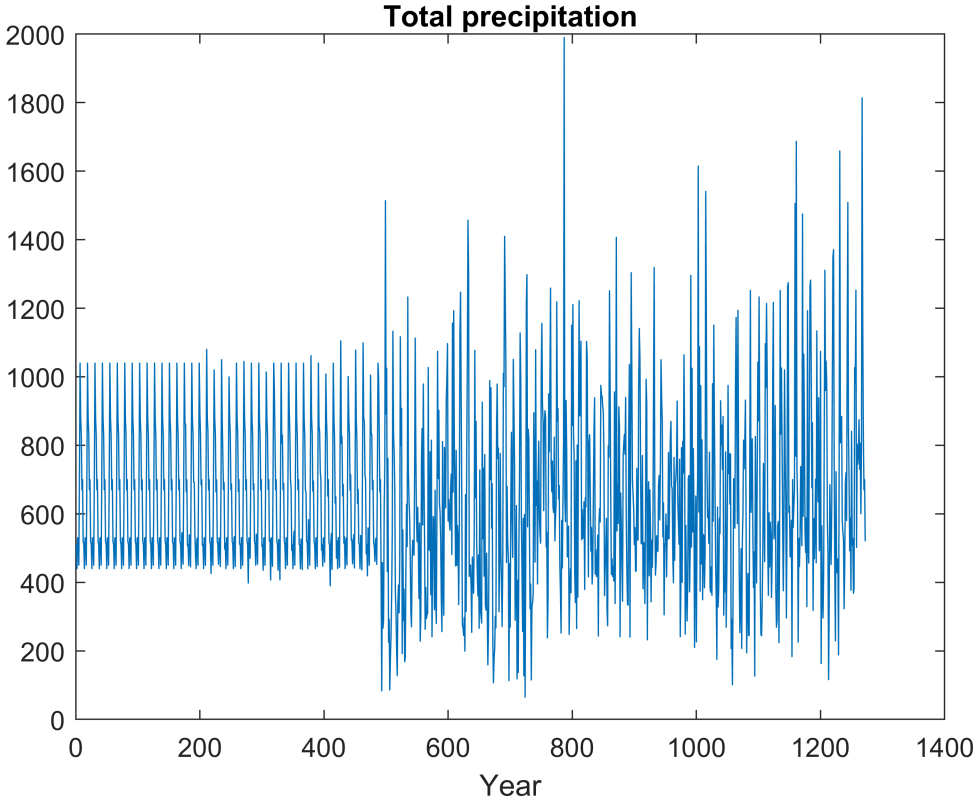


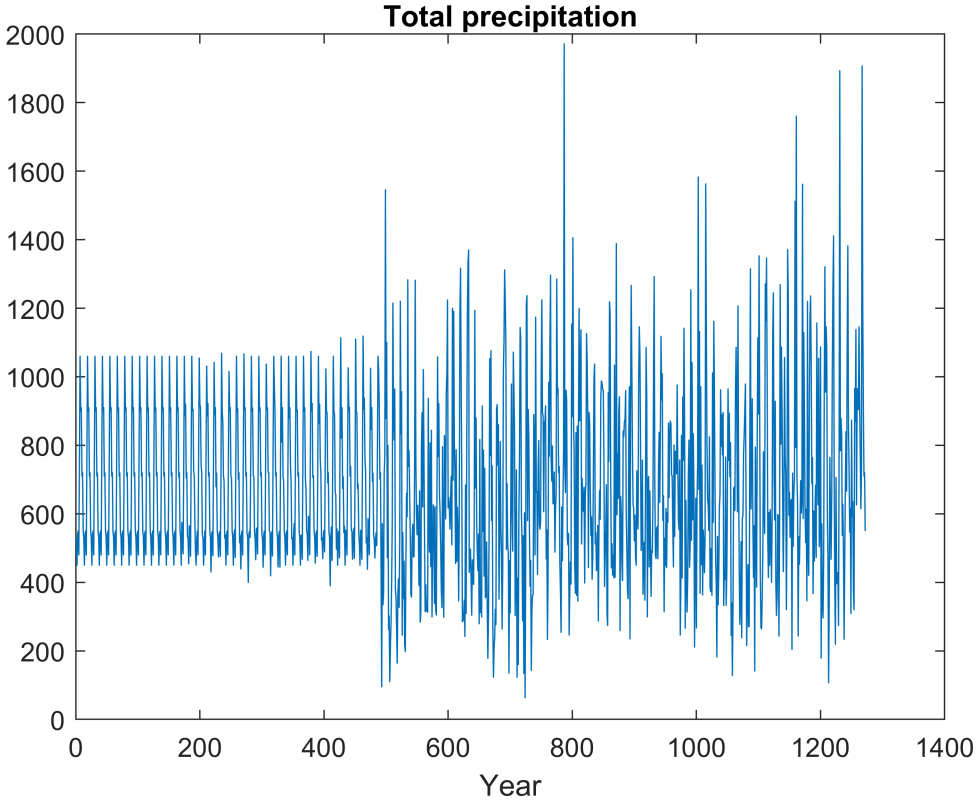

Total precipitation
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year

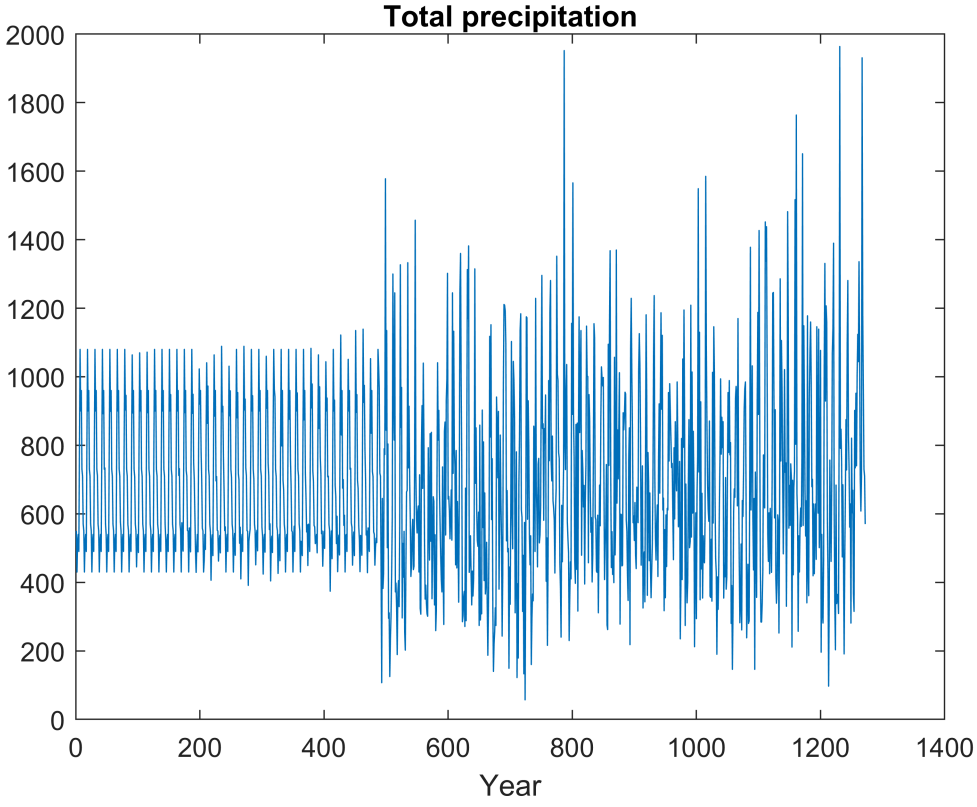

Total precipitation
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

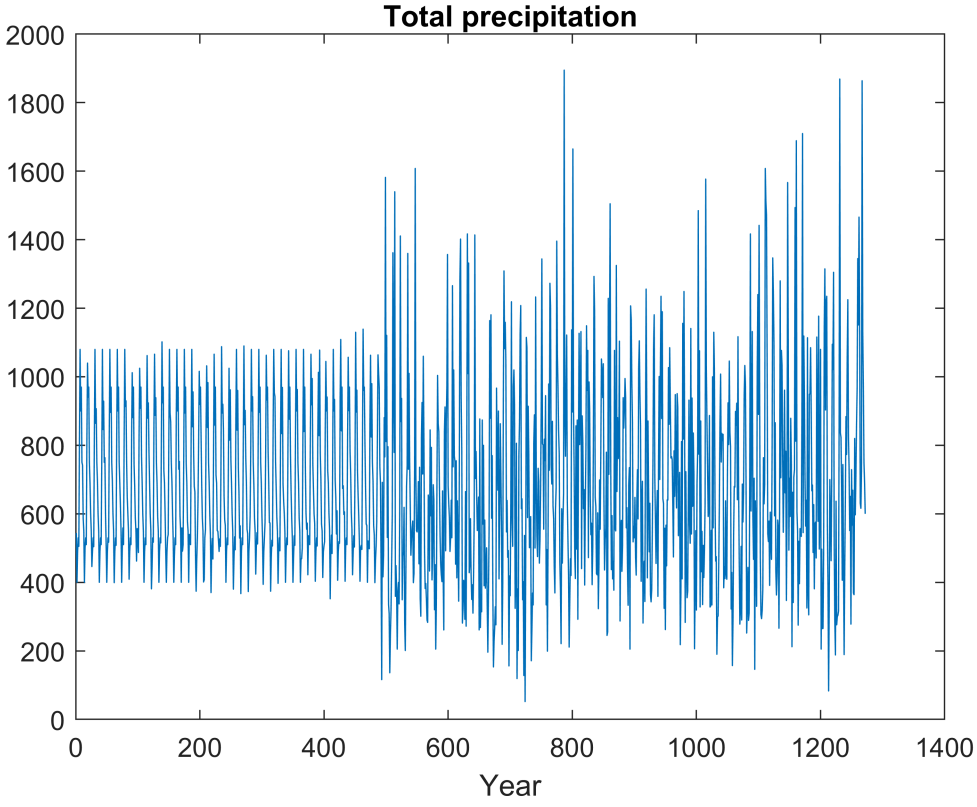

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

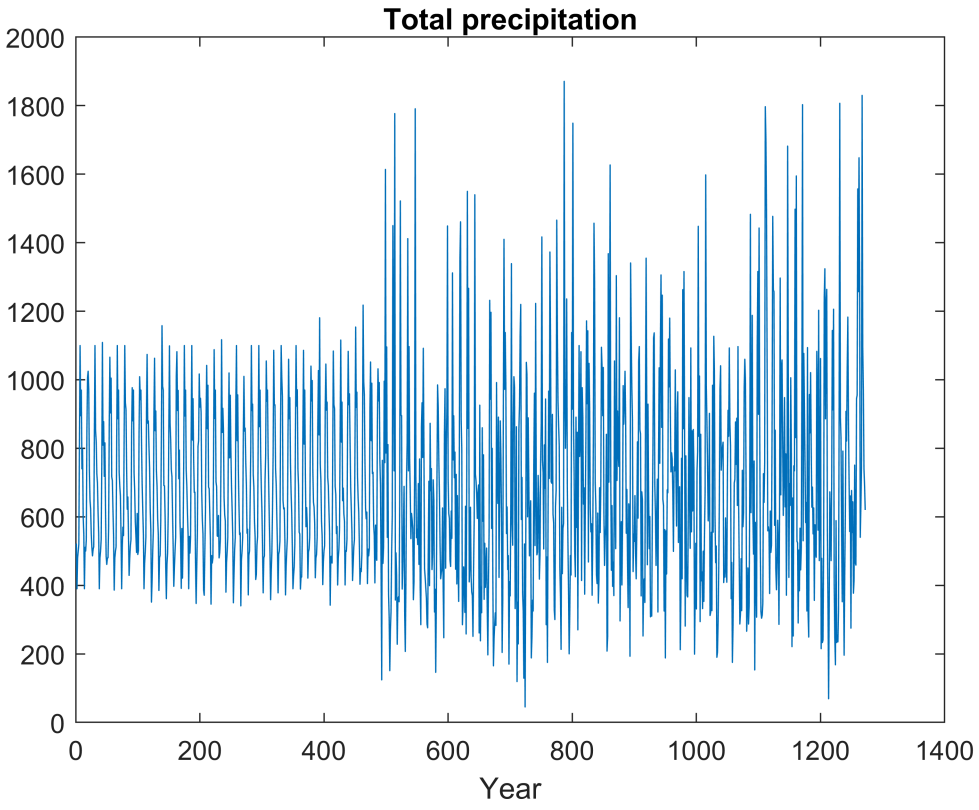

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

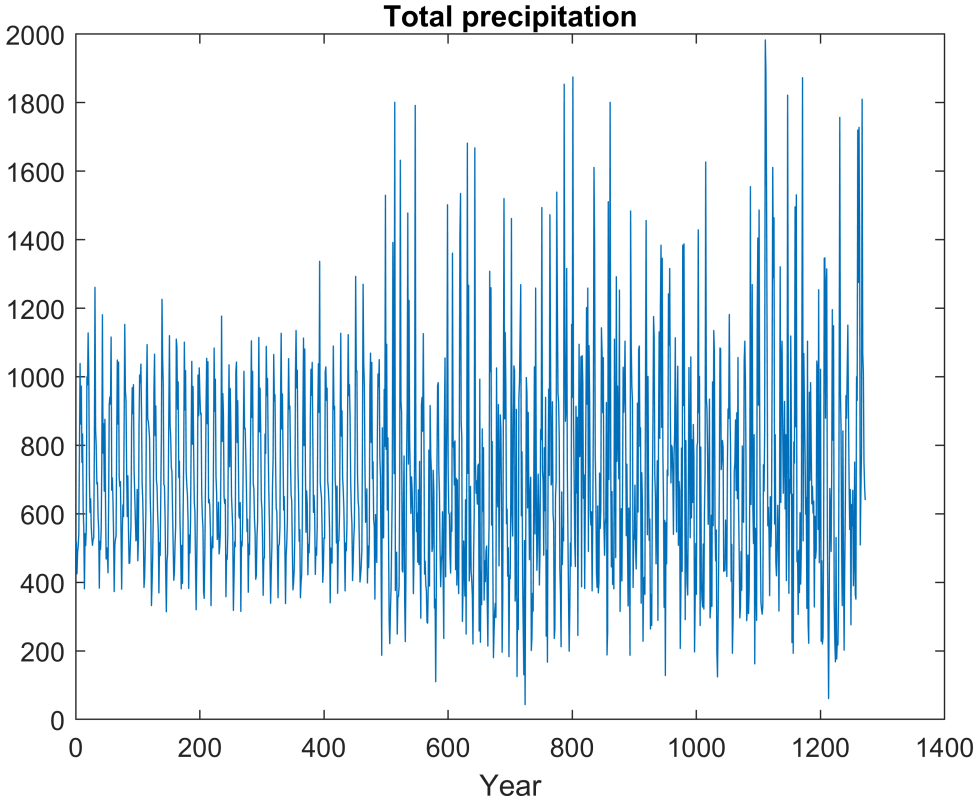

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

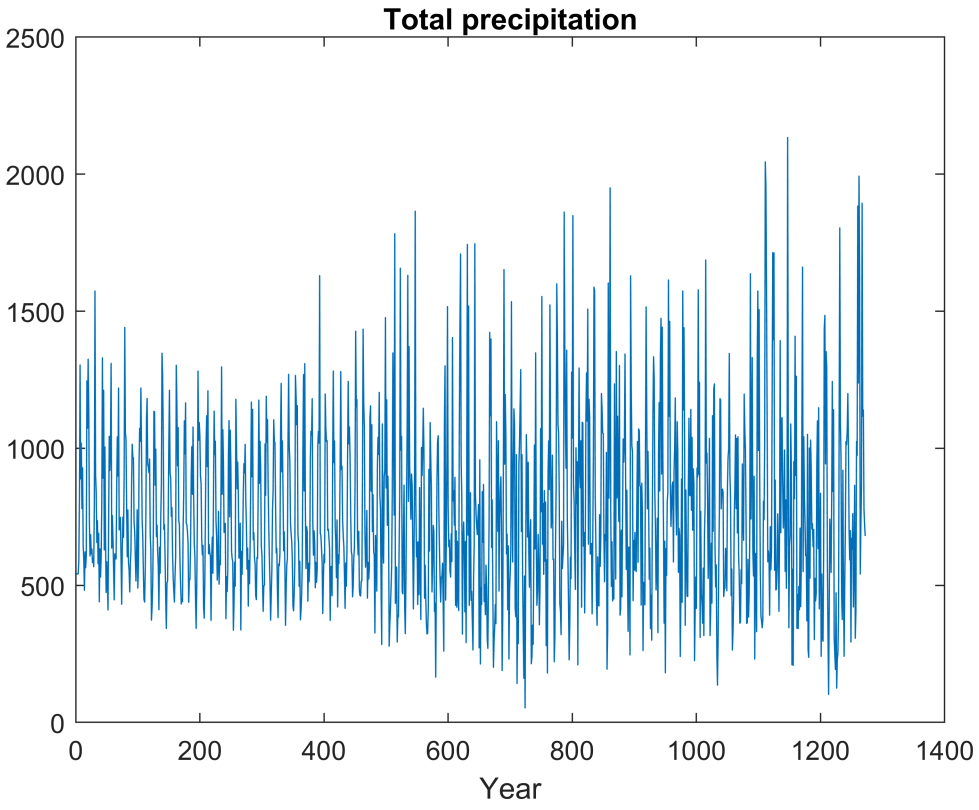

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

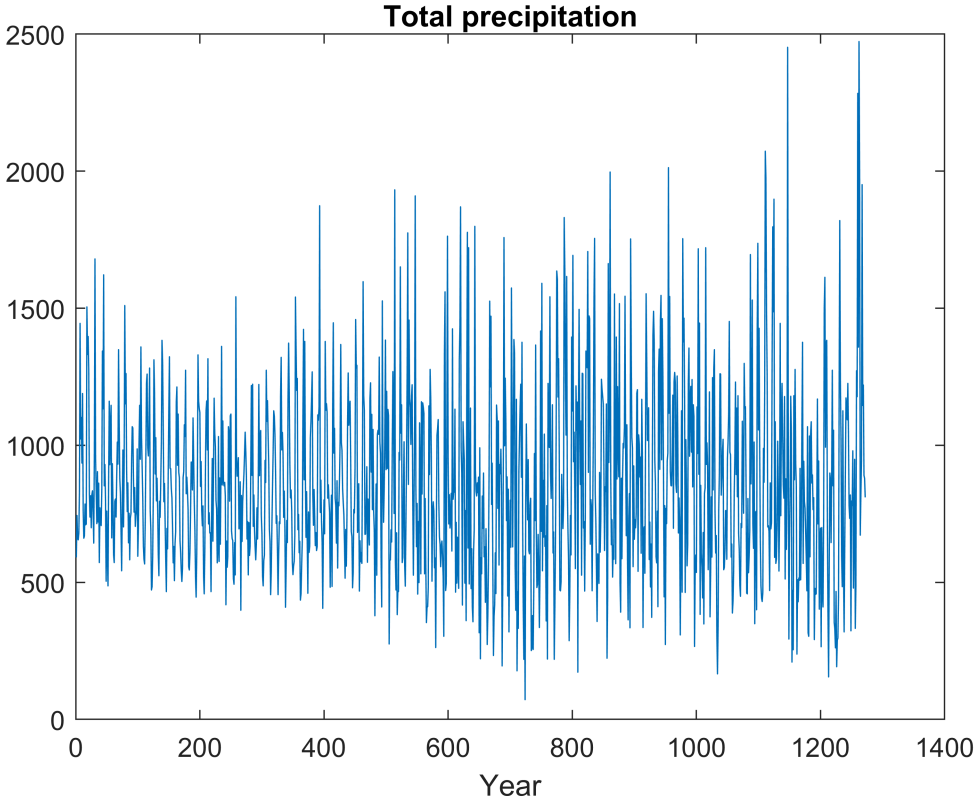

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

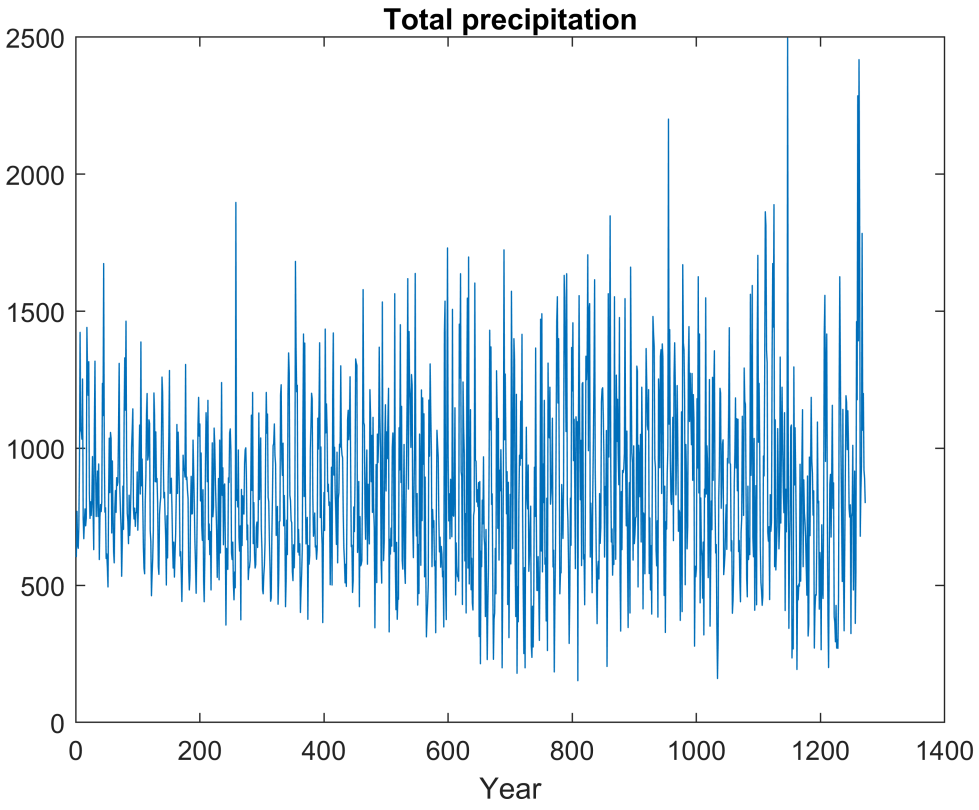

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

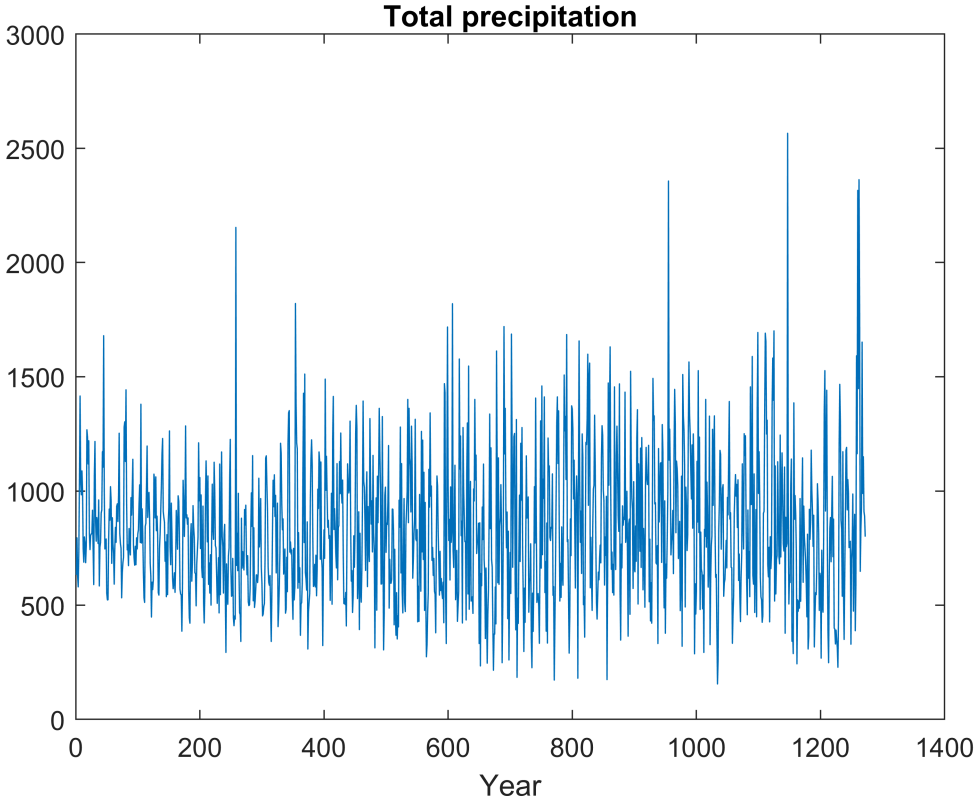

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

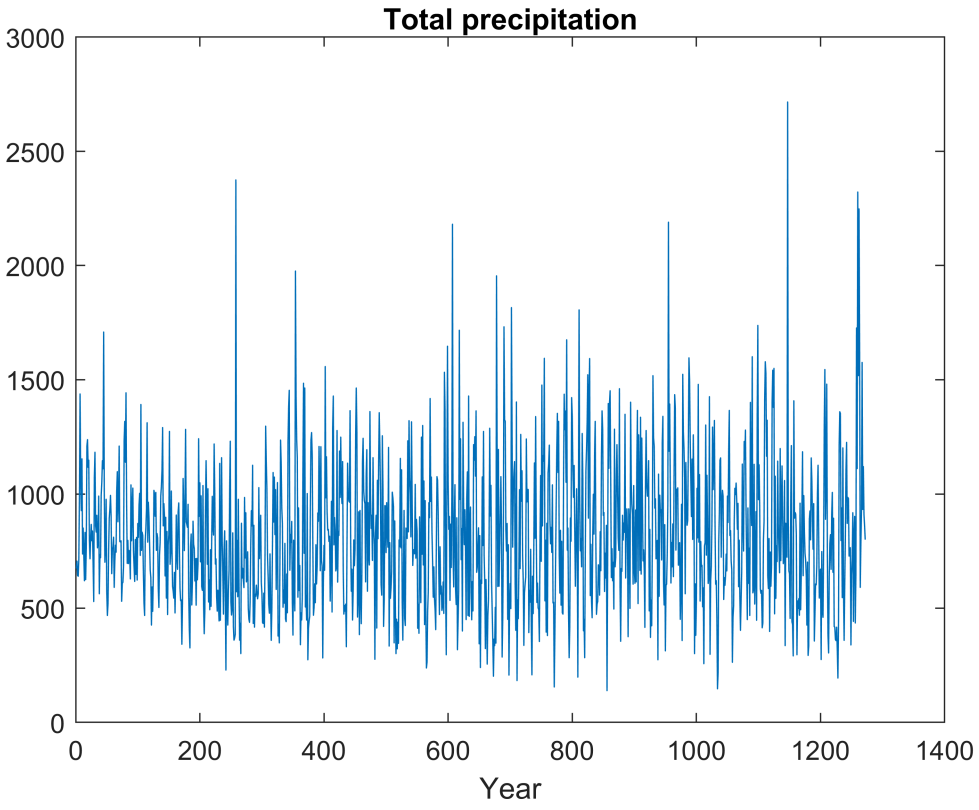

Total precipitation
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
Year

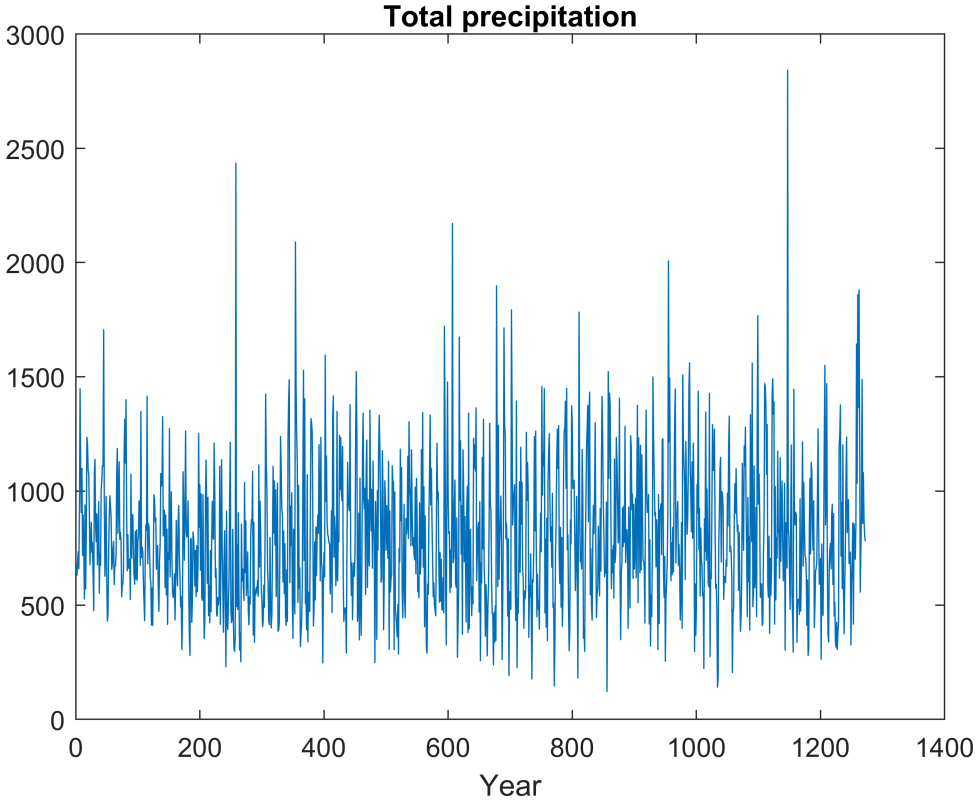

Total precipitation
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year

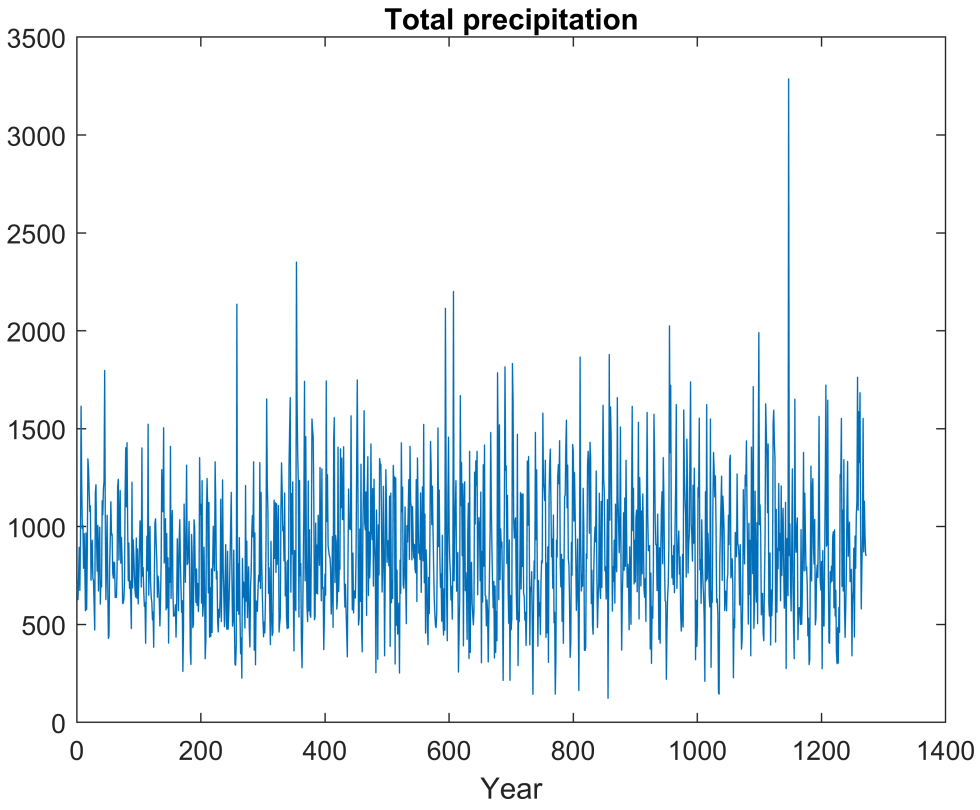

Total precipitation
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year


Total precipitation
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year


Total precipitation
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year


Total precipitation
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year


Total precipitation
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Year



Question: How many categories of figure do you see and why are they like this?

Answer: We should notice three types of pixels: only 0 values, some values with strong consistent patterns then fluctuations, some values without any patterns.

The student should understand that pixels with no data are not 0, otherwise we would have time series with 0 values and then actual precipitation values. In the figures, we can see some patterns, with periodicity of 12 months meaning that the same value is used for each January, February, March, etc... It means that the value used for each month is uniform and different for each pixel, i.e. it has to be climatology data.

Filtering the pixels

Let's filter the pixels with climatology data. One way to do it is to use a function looking at recurrent patterns (climatology) for each pixel.

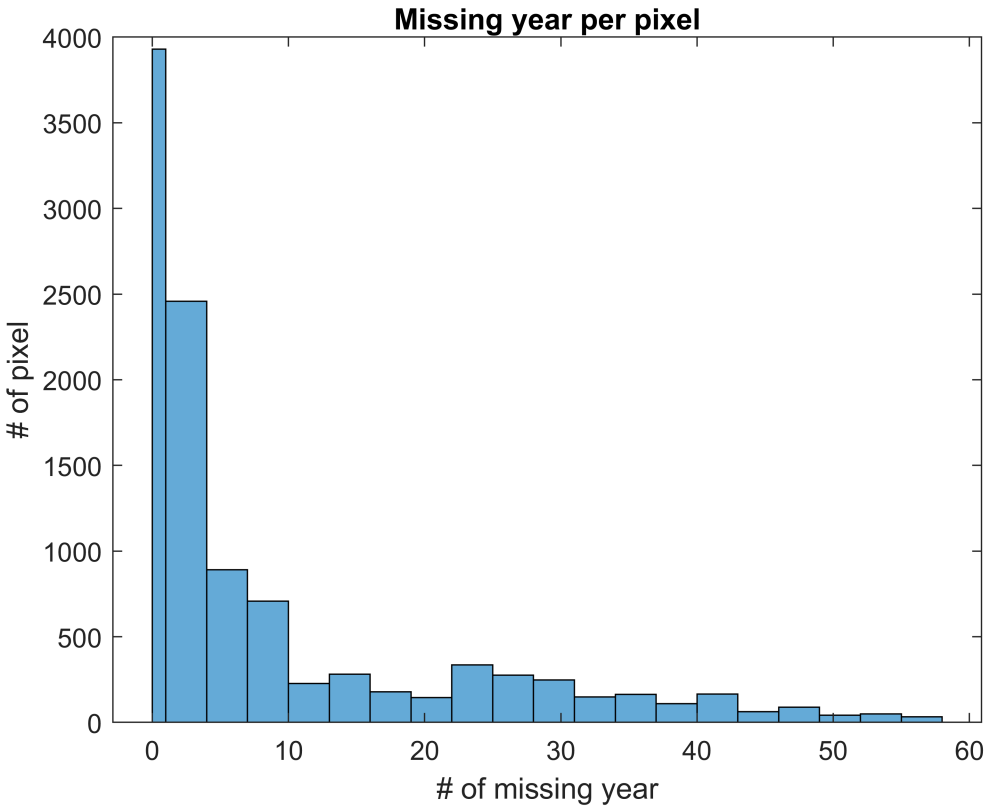
```
%Create a mask of useful pixels (with 100 years or 30 years of records)
[mask, maskHalf] = f_IO_maskNaN(data,'dataFolder');
```

```
Task started: Creating a mask of useful pixels
     The task has been completed in 1.071 sec
```

Let's check how much data we are missing

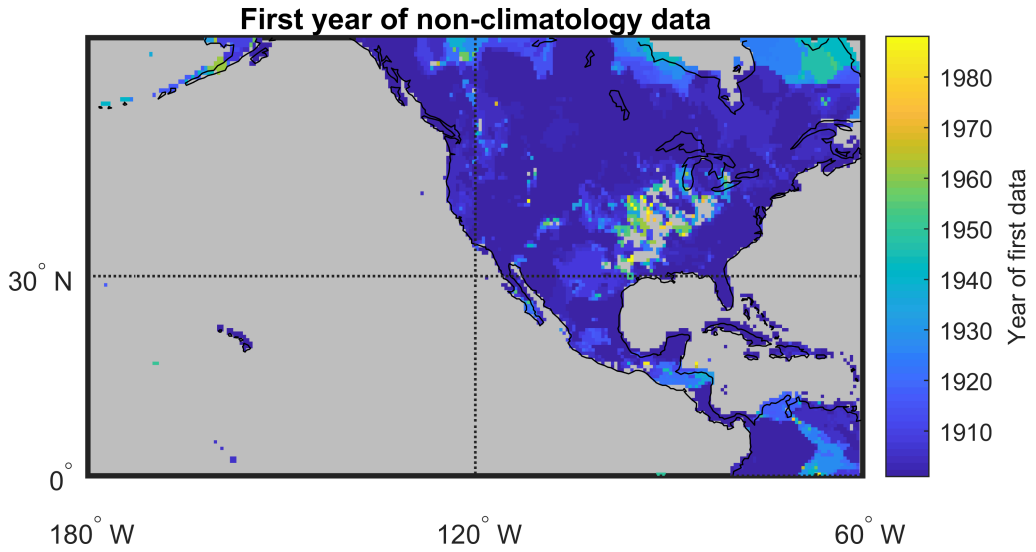
```
% Plot some statistics about the missing pixels
```

```
figure();
histogram(maskHalf(:,3),[0 1:3:60]);
title('Missing year per pixel');xlabel('# of missing year');ylabel('# of pixel');
```



And where is it located.

```
maskHalfMap = nan(size(data,1),size(data,2));
for ii=1:size(maskHalf,1)%create a matrix of strting year
maskHalfMap(maskHalf(ii,1),maskHalf(ii,2)) = maskHalf(ii,3)+1901;
end
plotGlobalMap(maskHalfMap,'startingYear','First year of non-climatology data');
```



What we now see is that a lot of data are missing in higher latitude and Central USA (starting data is about 1940, which leaves only 60 years of data).

Questions

1. What would have happened if we use this dataset to calculate return levels of extreme events?
2. To what extend can we use extreme event analysis theories on this dataset?

Elements of answers

1. If we work on a pixel with climatology data and don't remove it, we are establishing a long period where there is no extreme events (all data are climatology), therefore any extreme event in the full dataset will be considered as much rare as it actually is, so we will be overestimating extremes.
2. Some datasets have records from 1940 only (in high lattitude), therefore with only 60 years of data, we can't study centenial events at all but are limited to 50 years event or so.
3. For more information on using half dataset, see https://serc.carleton.edu/teaching_computation/workshop_2018/activities/210225.html

Conclusion

The main conclusion is that the data are not complete. Even if we have data non NaN and non 0 data for every month and every pixel, some of them are not actual data but climatology data (long term average). Therefore not checking the dataset prior to any extreme event analysis could lead to significant errors and over-estimation of return periods. It is therefore critical to always assess the quality of the data used and their

meanings. Getting an answer without error message (typically induced by NaN values) is not synonym of not having fundamental errors in the analysis performed.

References

AghaKouchak, A., et al. "Remote sensing of drought: Progress, challenges and opportunities." *Reviews of Geophysics* 53.2 (2015): 452-480.

Martinez, A. "Climate change and extreme values analysis", *Teaching Computation in the Sciences Using MATLAB Exemplary Teaching Activities collection* (2018). https://serc.carleton.edu/teaching_computation/workshop_2019/activities/231095.html