ANOMALY CONSTRUCTION IN CLIMATE DATA: ISSUES AND CHALLENGES

JAYA KAWALE*, SNIGDHANSU CHATTERJEE***, ARJUN KUMAR*, STEFAN LIESS**,
MICHAEL STEINBACH*, AND VIPIN KUMAR*

Abstract. Earth science data consists of a strong seasonality component as indicated by the cycles of repeated patterns in climate variables such as air pressure, temperature and precipitation. The seasonality forms the strongest signals in this data and in order to find other patterns, the seasonality is removed by subtracting the monthly mean values of the raw data for each month. However since the raw data like air temperature, pressure, etc. are constantly being generated with the help of satellite observations, the climate scientists usually use a moving reference base interval of some years of raw data to calculate the mean in order to generate the anomaly time series and study the changes with respect to that.

In this paper, we evaluate different measures for base computation and show how an arbitrary choice of base can skew the results and lead to a favorable outcome which might not necessarily be true. We perform a detailed study of different base selection criterion and base periods to highlight that the outcome of data mining can be sensitive to choice of the base. We present a case study of the dipole in the Sahel region to highlight the bias creeping into the results due to the choice of the base. Finally, we propose a generalized model for base selection which uses Monte-Carlo based methods to minimize the expected variance in the anomaly time-series of the underlying datasets. Our research can be instructive for climate scientists and researchers in temporal domain to enable them to choose the right base which would not bias the outcome of the results.

1. Introduction

An important component of Earth Science data is the seasonal variation in the time series. Seasons occur due to the revolution of the Earth around the Sun and the tilt of the Earth's axis. The change in seasons brings about annual changes in the climate of the Earth such as increase in temperature in the summer season and decrease in temperature in the winter season. The seasonality component is the most dominant component in the Earth science data. For example, consider the time series of monthly values of air temperature at Minneapolis from 1948-1968 as shown in Figure 1. From the figure, we see that there is a very strong annual cycle in the data. The peaks and valleys in the data correspond to the summer and winter season respectively and occur every year. The seasonal patterns even though important are generally known and hence uninteresting to study. Mostly, scientists are interested in finding non-seasonal patterns and long term variations in the data. As a result of the effect of seasonal patterns, other signals in the data like long term decadal oscillations, trends, etc. are suppressed and hence it is necessary to remove them. Climate scientists usually aim at studying deviations beyond the normal in the data.

In order to remove seasonality from the raw data, climate scientists generally remove the monthly mean value from the raw data. For example, although more than 100 years of data are available for the temperature anomaly time series at the National Climatic Data Center, only the 100 years 1901-2000 are used to calculate the annual cycle [3]. Often, climate scientists only take 30 years as a reference interval and construct anomalies with respect to that interval. There are several important results and implications derived from the anomalies constructed using a short reference base. In general, climate data has complex structures due to spatial and temporal autocorrelation.

*Department of Computer Science, University of Minnesota, kawale, arkumar, steinbac, kumar@cs.umn.edu
Department of Soil, Water and Climate, liess@umn.edu * School of Statistics. University of Minnesota chatterjee@stat.umn.edu.

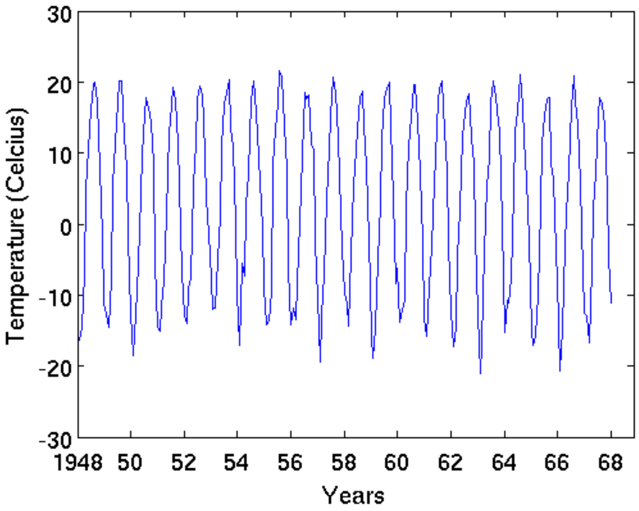

FIGURE 1. The figure shows the monthly mean air temperature at Minneapolis for a 20 year period. From the figure we can see that there is a very high annual cycle and the temperatures go up and down with the change of seasons.

The choice of the base significantly impacts the patterns that can be discovered from it and some really important climate phenomenons are computed using a fixed base. For example, teleconnections or long distance connections between two regions on the globe are represented by time series called *climate indices*. Climate indices are time series that summarize the behavior of the selected regions and are used to characterize the factors impacting the global climate. These climate indices are computed by the Climate Prediction Center [1] using a moving 30-year base period and currently they use a base period of 1981-2010. Another important set of results computed using a fixed base are incorporated in the International Panel on Climate Change (IPCC) Fourth Assessment Report on understanding climate change[13].

In this paper, we show how an arbitrary choice of base can skew the results and lead to favorable outcome which might not necessarily be true. We examine four simple criterions for base selection and empirically evaluate the differences in them. Our empirical evaluation of the different measures reveals that the z-score measure is quite different from the other measures like mean, median and jackknife. We further study the impact of using different base period to highlight that the outcome of further analysis can be sensitive to the choice of the base. We present a case study of the Sahel region to show that the dipole in precipitation in the region moves around and even disappears with the choice of a different base. Finally, we propose a generalized model for base selection which uses Monte-Carlo based sampling methods to minimize the expected variance of the underlying datasets. Our research can be especially instructive to climate scientists in helping them construct a generalized anomaly that does not create a bias in their analysis. Further, other researchers in temporal domain can also benefit from our work and it will enable them to choose a bias-free base. The main contributions of our paper are as follows:

- We present a systematic evaluation of four different measures of computing the base to construct the anomalies. Our evaluation shows that using the mean for anomaly computation might not be the right thing to do.

- We show that using a short base reference introduces a bias in the variance and show an alternative approach to take care of the bias.

- We present a case study of Sahel region to highlight that outcome of further analysis like dipole detection can be sensitive to the selection of the base.
- We propose an algorithm based on Monte Carlo sampling for automatic selection of the appropriate base which minimizes the variance across the time series. The algorithm suggests using weighted base of 55 year time period rather than 30 years for our data spanning 62 years.

The paper is organized as follows: In Section 2, we define the related work examining the issues with base construction. In Section 3, we describe the dataset used for our study. In Section 4, we examine the different measures and different time periods that can be used for anomaly construction. In Section 5, we describe our experiments evaluating the different measures and time periods. We also present a case study of the Sahel region to show the impact of the different base period in dipole analysis. In Section 6, we present a generalized approach for anomaly construction that computes a weighted anomaly using Monte Carlo sampling and also present our results based on the approach. Section 7 includes the discussions and the future road map for our work.

2. Related Work

Anomaly computation is a fundamental problem in climate science as most of the analysis of climate data relies upon computation of the anomalies as the first step. There have been some studies in the climate domain analyzing anomalies. Climate scientists mostly use a 30 year period to construct the anomalies and remove the annual cycle. Other ways to remove the annual cycle are 1) computing second moment statistics over each individual season by removing the first two harmonics of the respective time series; and 2) averaging the second moment statistics over all years. More techniques to remove the annual cycle include removing the first two or three harmonics (periods of 365.25, 182.625, and 121.75 days) e.g., [8] and [4]. Some of the less common practices involve looking at more sophisticated techniques like removing the cyclostationary empirical orthogonal function [11] or bandpass filtering, e.g. using a low-pass filter with 0.5 cycle/year [15]. More general methods are described in Wei et al. [19].

However, these procedures fail to take into account the natural interannual variability that should remain visible in the data. Therefore the procedures result in biased estimates of certain statistics [17]. In particular, lag-autocorrelations are systematically negatively biased, which indicates that uncertainty is added to climate data. Trenberth [17] shows for first order autoregressive time series that the autocorrelations computed after the annual cycle is removed become negative after just a few days lag. Consequently, the stochastic character of meteorological time series can result in less statistically significant analysis. Kumar *et al.* [12] state that the analysis of observed climate data often lacks separation of the total seasonal atmospheric variance into its external and internal components, with external components being the influence of atmospheric initial conditions, the coupled air-sea interactions, and boundary conditions other than sea surface temperatures, whereas internal components are described by the atmospheric variability over time. Removing the annual cycle should provide insight into the internal variability while leaving the external forcing intact.

Tingley *et al.* [16] discuss the impact of using a short reference interval in anomaly construction. They show that using a short reference period, the variance of the records at the time interval is reduced and inflated elsewhere. They show that the choice of the reference interval has a significant impact on the second spatial moment of the time series in the temperature data set whereas the first moment of the time-series is largely unaffected . They further use two factor ANOVA model within a Bayesian inference framework.

Despite the importance of anomalies in the further impact on the results, there is no firm consensus on how to deal with the systematic construction of anomalies and their impact on the various results.

Apart from this, the authors are not aware of any systematic study comparing the different aspects of anomaly construction in the climate data.

3. Dataset

We use the data from the NCEP/NCAR Reanalysis project provided by the NOAA / OAR/ ESRL PSD, Boulder, Colorado, USA [9]. The goal of the NCEP project is to produce a comprehensive atmospheric analysis using historical data (1948 onwards) from observations as well as other analysis like projection. As a result of these analysis, there is a complete data assimilation for every grid point on the Earth.

The NCEP/NCAR Reanalysis project has data assimilated from 1948 – present and is available for public download at [2]. We use the monthly time resolution of data and it has a grid resolution of 2.5° longitude x 2.5° latitude on the globe. We use the precipitation, air temperature and sea level pressure data for our analysis as they represent the most important climate variables. In all, we have 62 years of data (corresponding to 744 monthly values) for 10512 grid locations on the globe.

4. Different aspects of Anomaly Construction

We examine two aspects of anomaly construction: 1) the measure for anomaly construction and 2) the period used for anomaly construction in the following subsections.

4.1. Different measures for Anomaly Construction.

The central idea behind anomaly construction is to split the data into two parts: (a) data with expected behavior, and (b) anomaly data that shows the variability from the expected, which is generally used for understanding climate change phenomenon. For a given location i, its anomaly times series f_i' is constructed from the raw time series f_i by removing a base vector b_i from it as follows:

$$f_i' = f_i - b_i \tag{1}$$

A simple measure of computing the base b_i is by taking the mean of all data $(\overline{f}_i)$ present for location i. However the sample mean would not be a good measure as the Earth science data is associated with a large amount of seasonality. In order to account for this the base b_i is computed by taking a monthly mean for each month separately. It is not yet clear whether the mean is the right way to compute the base or if there is a better measure to compute the base. We examine four simple measures of base computations as follows:

- **Mean**: In this measure, the monthly mean values of the raw data are considered as the base and subtracted from the data to get the anomaly series.

- **Z-score**: Another possible way to construct the anomalies is to remove the monthly z-score values from the raw data. The z-score also accounts for the standard deviation in the monthly values.

- **Median**: This is constructed by removing the monthly median values instead of the monthly means as median can be a more robust measure when the data is skewed.

- **Jackknife**: This approach involves considering all points apart from the point itself in the computation of the mean and variance measures and it produces an unbiased estimation of variance just like Maximum aposterior Estimate (MAP).

We elaborate these measures and how they are computed in the following sub-sections.

4.1.1. *Mean.* Monthly mean computation is the most widely used method to extract the anomalies from the raw data. The mean subtraction makes the anomaly time series to have a zero mean. More formally,

$$f_i'(t,m) = f_i(t,m) - \mu_m, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \tag{2}$$

where total-start and total-end values represent the actual size of the data. In general it is known that taking mean would minimize the variance of the resulting series but it can also lead to over fitting and conclusions that might not be true. Further, instead of using the entire data for base computation, a short reference interval can be chosen. For example, if the data begins from 1900 to 2010, the base start and end years could be chosen as 1960-1990. We further discuss the issue of choosing a short base in Section.4.2.

4.1.2. *Z-score.* The z-score normalization ensures that the resulting anomaly series has mean = 0 and standard deviation = 1. As a result, z-score can be considered to be more robust than the mean but at the same time z-score based standardization can eliminate variations across different locations on Earth which might not be desirable. The z-score measure is computed as follows:

$$f_i'(t,m) = \frac{f_i(t,m) - \mu_m}{\sigma_m}, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \tag{3}$$

4.1.3. *Median.* In scenarios where data is skewed, mean can be sensitive to outliers. In such settings, median is typically considered to be more robust to outliers. As a result, we consider median as a method for base computation:

$$f_i'(t,m) = f_i(t,m) - median_m, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \tag{4}$$

4.1.4. *Jackknife estimate.* The Quenouille Tukey jackknife approach [20] is a useful nonparametric estimate of mean and variance. The basic idea behind the jackknife estimator is to systematically compute the mean estimate by leaving out one observation at a time from the sample set. Let $f_1, f_2, \ldots, f_n$ be the n points in the time series of a location x. The jackknife mean estimate is computed at point f_i by taking the mean of all points except f_i as follows:

$$Mean(f_i) = \frac{f_1 + \ldots + f_{i-1} + f_{i+1} + \ldots + f_n}{n-1} \tag{5}$$

Thus the anomalies are constructed by excluding the value at each point f_x^i. We however still use all the monthly values only to compute the jackknife estimate at each point. The variance measure using the jackknife approach turns out to be:

$$Variance = \left(\frac{n}{n-1}\right)^2 \times Variance(f_1, \ldots, f_n) \tag{6}$$

In order to see this consider $f_1, \ldots, f_n$ to be variable during a given month. Then we have to following:

$$\begin{aligned} Variance &= \frac{1}{n}\sum_i (f_i - Mean(f_i))^2 \\ &= \frac{1}{n}\sum_i (f_i - \frac{n}{n-1} \times Mean + \frac{1}{n-1} \times f_i)^2 \\ &= \frac{1}{n}\sum_i \frac{n^2}{(n-1)^2} \times (f_i - Mean)^2 \\ &= \left(\frac{n}{n-1}\right)^2 \times Variance(f_1, \ldots, f_n) \end{aligned}$$

The variance essentially turns out to be an unbiased estimate and is similar to the maximum aposterior probability (MAP) estimate of the model. MAP is similar maximum likelihood estimate (MLE) but also incorporates a prior distribution over the quantity one wants to estimate. MAP estimation can therefore be seen as a more robust form of MLE estimation. However the main problem with an approach based on jackknife is that it requires a lot of computation.

4.2. **Different Time Periods for Anomaly Construction.** As mentioned earlier, an anomaly series is constructed from the raw time series by removing a base value from it. The base value is generally considered to be the mean of the data. Since the true theoretical mean is not known, the base value is created by taking the sample mean of the data. However, most of the times a short reference interval is chosen to compute the base and changes with respect to that are studied. There is no absolute truth or guidelines available to choose the reference interval. Climate scientists generally choose the base as a moving 30 year period and study the changes with respect to that. However a moving short reference interval is problematic and can result in spurious results and conclusions. In order to highlight the problems associated with picking an arbitrary short base, we consider an example of teleconnections looking into the drought of the Sahel region in Africa in the Section 5.3.

5. Experiments and Results

5.1. **Comparison of Different Measures of Anomaly Construction.** Our first task is to empirically evaluate the differences in the four different measures described in Section 4.1. We use the precipitation data for our analysis. Using all the 62 years of data from the NCEP/NCAR website, we first construct an anomaly series for each location on the Earth using the four different measures. Further, we also construct complex networks by taking pairwise correlation between all locations on the Earth as used by several researchers like [14], [10], [5], [18] to find patterns in climate data. The nodes in the graph represent all the locations on the Earth and the edges represent pairwise correlation between the anomaly time series of all the nodes on the Earth. Our goal is to evaluate whether there are statistically significant differences between different measures to compute anomalies. In order to measure the statistically significant difference, we consider the following three criterion:

- *Mean based difference*: We compute the mean of anomaly time series using different measures and then compute the difference in mean for each pair of measure. The mean difference would be statistically significant if we can say with 95% confidence that the mean of difference is non-zero.
- *Correlation based difference*: Here we compute the correlation of every point with respect to other points on the globe using the four measures and check if the correlation values are impacted by using different measures for anomaly construction.
- *Monthly variance based difference*: Here we check if the monthly variance of the anomaly time series at each location is different for pairs of anomaly computation measure or not.

We use *t-test* to test if the difference between two measures follows a Gaussian distribution with $mean = 0$ and unknown variance. Thereby, our null hypothesis, $H0$ is that two measures lead to the same result and alternate hypothesis, Ha is that the two measures are different.

Tables 1, 2 and 3 show the number of locations where two measures lead to significant differences in the anomaly time series. Here we make an observation that z-score and median lead to significant differences from each other as well as the mean and the jackknife. Z-score based base computation yields the most significant difference as it leads to statistically significant changes in correlations and monthly variances at more than 9000 grid locations on the Earth. The z-score measure also stands out if we look at the monthly variance of each point. On the other hand, mean and jackknife seem to be similar. Median differs from the two over the mean difference based comparison. This is perhaps expected as the median and the mean values are not the same and all the other bases have zero monthly mean. Overall this result indicates that different measures used to compute base can lead to drastically different results. This result makes it intuitively clear that z-score might not be the best way to compute the base. In order to compare the mean and the median, we examine the skew in the anomaly time series after using the mean to construct the anomalies. To determine the skew, we check the *kurtosis* of the anomaly series at each location on the Earth. The kurtosis falls within the range of 2.6-3.5 for more than half of the locations on the Earth which is acceptable for a normal distribution. However, some locations have a very high skew and the kurtosis value is as high as 10. Fig 2 shows the histogram of kurtosis for all the locations and also shows the skew in the time series at a random location on the Earth. This suggests that the mean might not a good measure to compute the anomalies and median might be a better choice. However, further investigations are still needed to understand the right measure for anomaly computation.

Method	Mean	Median	z-score	Jackknife
Mean	-	692	0	0
Median	-	-	1281	674
z-score	-	-	-	0
Jackknife	-	-	-	-

TABLE 1. Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the **anomalies** at the different locations for precipitation.

Method	Mean	Median	z-score	Jackknife
Mean	-	0	5303	0
Median	-	-	5152	0
z-score	-	-	-	5303
Jackknife	-	-	-	-

TABLE 2. Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the **correlation of each location** with the different locations for precipitation.

Method	Mean	Median	z-score	Jackknife
Mean	-	0	9152	0
Median	-	-	9152	0
z-score	-	-	-	8998
Jackknife	-	-	-	-

TABLE 3. Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the **monthly variance** at the different locations for precipitation.

5.2. **Comparison of different time periods for anomaly construction.** The previous results show that for a fixed base period there exists different ways to compute the base which can lead to drastic differences in the anomaly time series. *Here we try to see if we can fix a measure (say mean) then check if varying the base period affects the anomaly time series.* In order to do this, we examine three base periods: a) first 20 years b) entire 62 years and c) last 20 years. We also experiment with base period length of 30 years but that leads to similar results so for the sake of presenting the extremes, we present results choosing 20 years as a base.

We construct the anomaly series for each location corresponding to the given base periods. We selected mean as the measure of computing the base. In order to compare the time series, we used KL-divergence criteria to see if different base periods have different effects on anomaly time series.

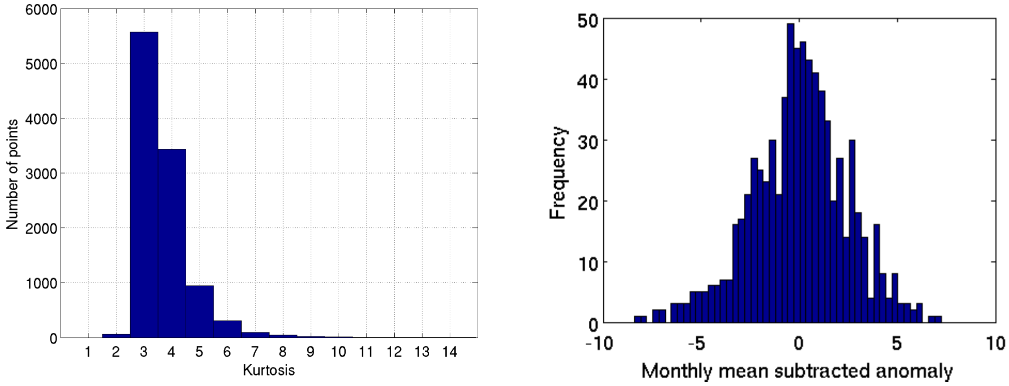

FIGURE 2. a) Kurtosis histogram b) The mean subtracted anomaly shows a skew in the data.

The KL-divergence is defined as follows:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{7}$$

KL-divergence of 0 means that the two series are exactly the same. A KL-divergence value indicates that the series are quite different. We plot the divergence value for each location on the globe in Figure 3. The white region shows that these locations are severely affected by our choice of base. In general last 20 years vs 62 years (second figure) has a light shade of gray indicating that all the locations on earth would be affected (in their anomaly series) if we make a choice between last 20 years vs all 62 years.

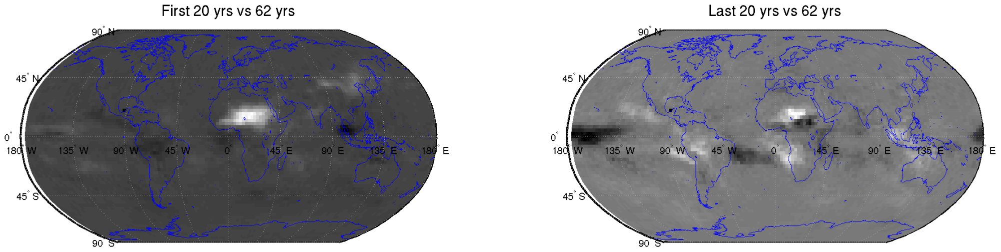

FIGURE 3. KL-divergence of the anomaly series the different bases a) first 20 years vs entire 62 years b) last 20 years vs entire 62 years and c) first 20 years vs last 20 years. The white shaded regions represent regions of maximum divergence.

Also the variance in the anomaly time series changes when we move the 20 year base period across the entire length of the time-series. Fig. 4 shows the change in variance at two random locations by picking up different 20 year base periods by varying the starting times from 1948-1988. From the figure, we see that average variance in the anomalies at different points varies using different start times for the base periods. This makes the problem complex as different regions show minimum variance in different windows of the time period.

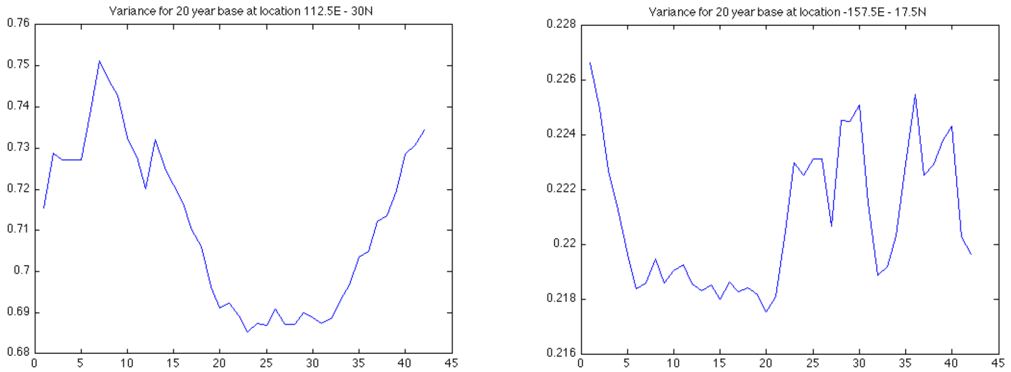

FIGURE 4. Change in variance of two random locations on the Earth choosing a 20 year reference period and moving the starting year from 1948-1988.

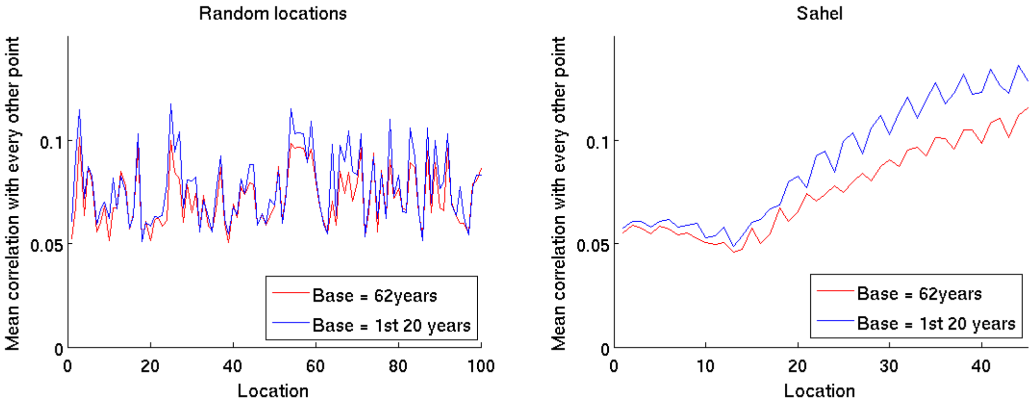

FIGURE 5. a) Mean correlation of 100 random points with all the points in the globe for precipitation using the entire 62 years as the base and only the first 20 years as the base. b) Mean correlation of 100 random points with all the points in the globe for precipitation using the entire 62 years as the base and only the first 20 years as the base. The difference in correlation (red and blue) is much more pronounced in Sahel as compared to the random locations.

In order to further analyze the impact of the choice of short reference base on the correlation of anomaly time-series, we consider two anomaly construction scenarios using the base as: a) first 20 years from 1948-1967 and b) using the entire 62 year time period from 1948-2009. We examine the changes in the mean correlations of locations with respect to every other location on the Earth using the two base period. Figure 5 shows the change in mean correlation of 100 random locations and the locations in Sahel using the two base periods to compute the anomalies. From the figures, we see that the locations in Sahel are much more impacted by the change in the base period as compared to the 100 random locations. We also find similar trends in other variables like pressure and temperature in Sahel but do not report it due to lack of space. These results underline the fact that a reference interval is crucial in the computation of the anomalies. In the next section, we show a case study on teleconnections where the actual analysis results and implications are impacted by the choice of the reference base.

5.3. **Case study of the Sahel dipole.** *Teleconnections* are long distance connections connecting the climate of two places on the Earth. One such class of teleconnections are the dipoles which consist of two regions having anomalies in the opposite direction and thus having negative correlation. The climate in Sahara and Sahel region of Africa has undergone some radical shift in the past century. The region received heavy rainfall till about 1969 until when it went into a period of severe drought for about 30 years which brought a *regime shift* in the region. The drought in the region and its environmental causes and consequences have been well studied in the past [6].

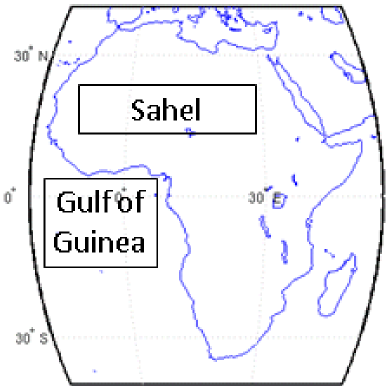

FIGURE 6. Sahel and the Gulf of Guinea in Africa.

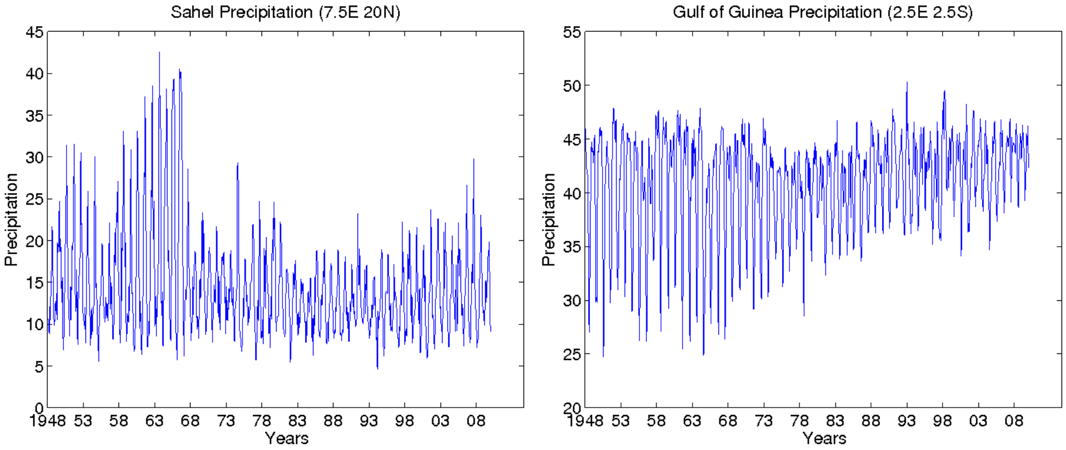

FIGURE 7. Raw precipitation time-series at Sahel and the Gulf of Guinea.

The precipitation in the region has recovered slightly but not enough to come back to the same levels as that before 1969. The severe loss of precipitation at Sahel was accompanied with a heavy increase in precipitation at the same time in the Gulf of Guinea around Africa, thus forming a dipole in precipitation [7]. The two regions Sahel and the Gulf of Guinea are marked in the Fig.6. The raw precipitation time series of the two locations in Sahel (7.5E, 20N) and the Gulf of Guinea (2.5E, 2.5S) are shown in the Fig. 7.

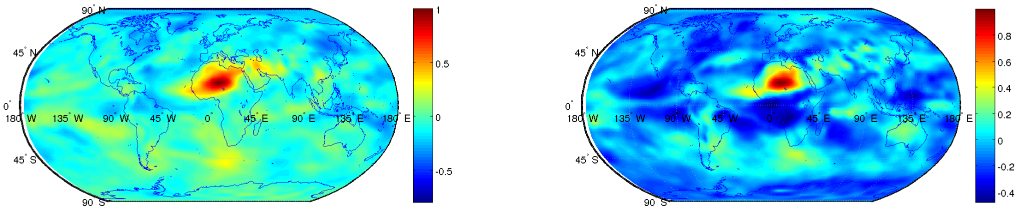

FIGURE 8. Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 1st network (1948-1967). The figure shows the presence of a dipole (positive and negative correlations as shown by red and blue regions) in the **right** picture and not the left one.

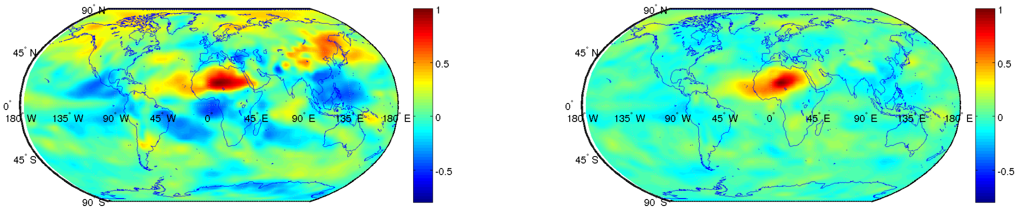

FIGURE 9. Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 2nd network (1968-1987).

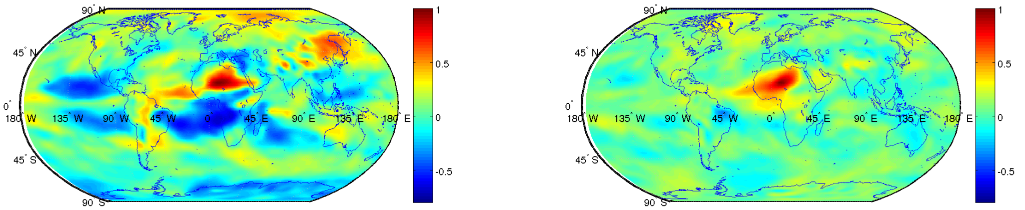

FIGURE 10. Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 3rd network (1987-2007). The figure shows the presence of a dipole (positive and negative correlations as shown by red and blue regions) in the **left** picture and not the right one.

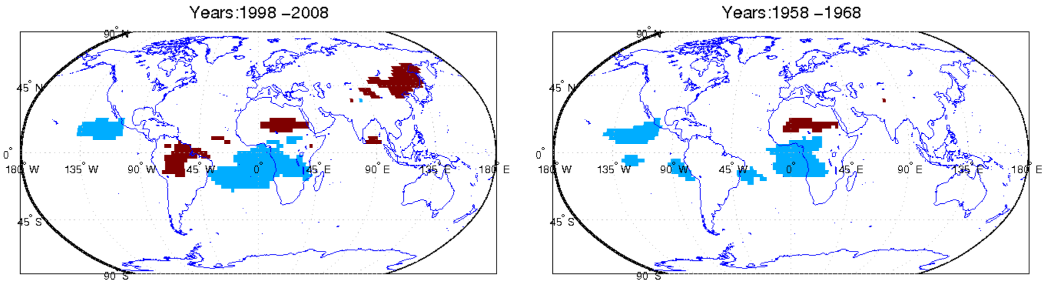

FIGURE 11. Different regions and different time periods are identified as dipoles in precipitation using the first 20 years as the base and the last 20 years as the base. (The red and blue regions represent two ends of the dipole and have negative correlation in their anomalies.)

From the figure, we can see the dramatic decrease in precipitation in Sahel and an increase in precipitation in the Gulf of Guinea around the 1970s. Now using the 62 years of NCEP precipitation data, we choose two base years, the first 20 years (1948-1967) and the last 20 years(1988-2007). We further construct three networks by taking pairwise correlation between the anomaly time series of all locations on the Earth for a 20 year time period each (1948-1967, 1968-1987 and 1988-2007). Consider the point A(7.5E, 20N) in Sahel. Let us examine the correlations of this point with all the regions on the Earth. Fig. 8, 9 and 10 show the correlation of all the points on the Earth with respect to a single point in Sahel for the three time periods 1948-1967,1968-1987 and 1988-2007 respectively. From the figures, we see that the if we choose the base period to be the first 20 years, the Sahel dipole is clearly visible (positive and negative correlations as shown by red and blue regions in the figures) in the period 1988-2007, however if we choose the last 20 years as the base period the dipole is seen in the interval 1948-1967. Further we use the dipole detection algorithm on the complete network as given in [10]. The algorithm begins by picking up the most negative edge on the Earth and grows the two ends of the negative edge into two regions such that they are negatively correlated with each other and positively correlated within each other. Using the algorithm, we see that the Sahel dipole appears in different time periods and also in different regions as also shown in the Fig.11. Thus the choice of a base period severely impacts the results and subsequently the interpretations that can be drawn from the results. Hence extra caution needs to be exercised while constructing anomalies in order to avoid spurious conclusions to be drawn from the results.

6. A Generalized Approach for Anomaly Construction

In the previous section, we saw that there is a bias introduced in the results upon considering different measures of the base and different durations. So the primary question arises, *What is the right base to choose for anomaly construction?*. In this section, we discuss our approach to handle the problem of the anomaly construction. The intuition behind our approach is to have a weighted mean to construct the anomalies and use an objective criteria to pick up the right set of weights using Monte Carlo sampling. The weighted base for anomaly construction for a location i is created as follows:

$$b_i(t, w) = \sum_{t=t0}^{t0+k} w_t * f_i(t) \quad subject\ to \sum_{t=t0}^{t0+k} w_t = 1 \tag{8}$$

where $t0$ represents the starting time period, k represents the length of the time period. We further assume that the weights w_t are the same for each year and do not depend upon the month in the year. Further, the anomalies are constructed by removing the weighted base for each month as

Algorithm 1: A Generalized approach for Anomaly construction.

```
Let $f_i(t)$ be the monthly values of raw time series of location $i$
Let, $N$ = Length of total time period.
Let, $T_{base}$ = Shortest length of reference interval.
Let, $NumSimulations$ = Number of simulations to run.
Initialize $GlobalVariance$, $OptimalWeights$ to $\infty$
repeat
  for $k \in T_{base}, ...T_N$ do
    for $t \in T_0, ...T_N$ do
      for $i \in 1....NumSimulations$ do
        Compute weight vector $w_1, w_2, ......, w_t$
        subject to the constraint $w_1 + w_2 + .... + w_t = 1$. using a Dirchlet prior.
        Compute the weighted base as $b_i(t, w) = \sum_{t=t0+1}^{t0+k} w_t * f_i(t)$
        Compute the Anomalies from the weighted base as $f_i'(t, w) = f_i(t) - b_i(t, w)$
        Compute the Variance of all the anomaly time-series across the globe.
        if $Variance < GlobalVaraince$ then
          Update the $GlobalVariance$ and $OptimalWeight$
          $GlobalMedian = Median$
          $OptimalWeight = w_1, w_2, ......, w_t$
        end if
      end for
    end for
  end for
until convergence
```

follows:

$$f_i'(t, w) = f_i(t) - b_i(t, w) \tag{9}$$

We run Monte Carlo simulations to get the right set of the weights w_t and define the objective function as minimizing the variance of the anomaly time series over time and space. By minimizing the variance, we are trying to enforce uniformity over the data. There can be some other objective functions like the median of the lowest 10% of the correlations. The intuition behind this objective function is that for computing dipoles, we need to examine the most negative correlations. Hence we want to find a weight and a base vector corresponding to our criterion for dipoles. However, we consider a general objective function that is not dependent upon the problem. The further details of the algorithm are present in Algorithm 1.

6.1. Results.

We use the precipitation data and run our Monte Carlo based simulation algorithm to get the right reference base period. Figure 12 represents the final converged weights. The other parameters of the final convergence of the algorithm are as mentioned in Table 4. Using our new weighted anomaly, we re-construct the correlation plots around the Sahel to get a sense of the dipole in the Sahel. Fig 13 shows the new results using the dipole in the Sahel using the weighted anomaly. Using a bias free base gives us confidence about the non-spuriousness of the discovered climate pattern or climactic phenomena such as a dipole. It implies that a dipole does exist in the region and that bases chosen which result in the dipole appear vanishing are not good bases. This objective function thus helps us in observing phenomena which would be more prominent if favorable bases are assumed but a bias-free base gives us a worst case scenario and more confidence in the results.

7. Discussion and Conclusions

The issue of anomaly construction is a fundamental problem in climate science as most of the analysis and results are derived after the raw data is transformed into an anomaly series. However there are no current guidelines available on anomaly construction and climate scientists usually

Table 4. Final algorithm convergence details.

Parameter	**Value**
Period	55
Starting year	1948

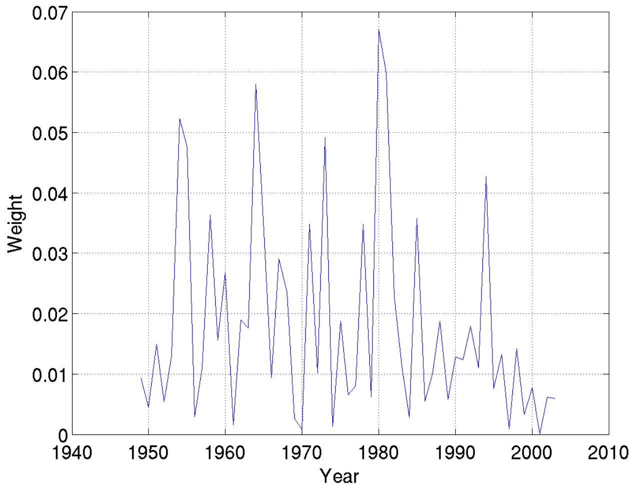

Figure 12. Final converged weight vector.

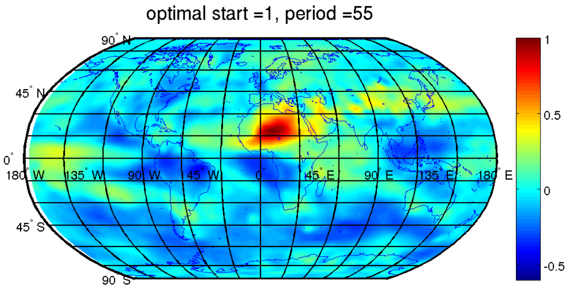

Figure 13. A correlation map as seen from the Sahel location A(7.5E, 20N).

rely upon computing a moving reference base for anomaly construction. In this paper, we examine the various issues pertaining to the construction of the anomalies. We assess the four methods of anomaly construction i.e. mean, median, z-score and jackknife. Our results show that if z-score is used as a measure for anomaly computation then the correlation values across different locations come out to be significantly different at 95% confidence interval. The mean, median and the jackknife measure do not show significant differences. However, due to the skewness in the data, the mean might not be a good measure and the median might be a good measure in such a case. However, further investigation is required to understand the right measure should be used.

We further show the bias in results introduced due to a choosing a short reference interval and show the difference in conclusions and results using a case study of the Sahel dipole. It is important to handle the bias introduced due to a short base as subsequent conclusions derived from it get affected. We further propose a generalized algorithm to handle the the issue of a bias-free base. Using our algorithm, we get the optimal base period to be 55 years. The algorithm can be modified to have different objective functions to handle different specific scenarios. As a part of our future work, we will examine different approaches to learn the weight vector as opposed to using the Monte Carlo simulations. We will also evaluate different objective measures and their impact on the base construction.

Acknowledgments

This work was supported by NSF grants IIS-0905581 and IIS-1029771. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

References

[1] Climate prediction centre, http://www.cpc.ncep.noaa.gov/.

[2] Ncep data download link, http://www.esrl.noaa.gov/psd/data/.

[3] Temperature anomaly time series, national climatic data center, noaa, http://www.ncdc.noaa.gov/ghcnm/time-series/index.php.

[4] G. Compo, G. Kiladis, and P. Webster. The horizontal and vertical structure of east asian winter monsoon pressure surges. *Quarterly Journal of the Royal Meteorological Society*, 125(553):29–54, 1999.

[5] J. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal-Special Topics*, 174(1):157–179, 2009.

[6] J. Foley, M. Coe, M. Scheffer, and G. Wang. Regime shifts in the sahara and sahel: interactions between ecological and climatic systems in northern africa. *Ecosystems*, 6(6):524–532, 2003.

[7] A. Giannini, R. Saravanan, and P. Chang. Oceanic forcing of sahel rainfall on interannual to interdecadal time scales. *Science*, 302(5647):1027, 2003.

[8] C. Jones and J. Schemm. The influence of intraseasonal variations on medium-to extended-range weather forecasts over south america. *Monthly Weather Review*, 128(2):486–494, 2000.

[9] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph. The ncep/ncar 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, 77:437–471, 1996.

[10] J. Kawale, M. Steinbach, and V. Kumar. Discovering dynamic dipoles in climate data. In *SIAM International Conference on Data mining, SDM*. SIAM, 2011.

[11] K.-Y. Kim and C. Chung. On the evolution of the annual cycle in the tropical pacific. *Journal of Climate*, 14(5):991–994, 2001.

[12] A. Kumar, B. Jha, Q. Zhang, and L. Bounoua. A new methodology for estimating the unpredictable component of seasonal atmospheric variability. *Journal of Climate*, 20(15):3888–3901, 2007.

[13] I. P. on Climate Change. *Fourth Assessment Report: Climate Change 2007: The AR4 Synthesis Report*. Geneva: IPCC, 2007.

[14] M. Steinbach, P. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 446–455. ACM, 2003.

[15] D. Thomson. Dependence of global temperatures on atmospheric co2 and solar irradiance. *Proceedings of the National Academy of Sciences of the United States of America*, 94(16):8370, 1997.

[16] M. Tingley. A bayesian anova scheme for calculating climate anomalies, submitted 2011.

[17] K. Trenberth. Some effects of finite sample size and persistence on meteorological statistics. part i: Autocorrelations. *Mon. Wea. Rev*, 112:2359–2368, 1984.

[18] A. Tsonis, K. Swanson, and P. Roebber. What do networks have to do with climate? *Bulletin of the American Meteorological Society*, 87(5):585–595, 2006.

[19] L. Wei, N. Kumar, V. Lolla, E. Keogh, S. Lonardi, and C. Ratanamahatana. Assumption-free anomaly detection in time series. In *Proceedings of the 17th international conference on Scientific and statistical database management*, pages 237–240. Lawrence Berkeley Laboratory, 2005.

[20] C. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.