Modeling Paleoclimactic Data Trends of Temperature

Arjun Naik

June 3, 2022

Contents

1 Introduction

1.1 Background

Paleoclimatology is the study of historical climate trends on Earth. Paleoclimactic Data is the data that is collected regarding the Earth's climate, typically spanning from anywhere between hundreds and millions of years ago. Logically, this data cannot have been physically colected and gathered by physical humans at the moment the data was measuring, so scientists find other ways to collect this data. Typically, in paleoclimatology, paleoclimactic data is collected from *natural sources,* including tree rings, ice cores, and marine phenomena.[1]

1.2 Significance

In modern times, the climate and climate is a pressing issue. Thus, in order to fully attempt to take control over the issue of climate change, it is very important to have a thorough understanding of the modern climate. To have a thorough understanding of modern climate trends, getting a good understanding of past climate trends is crucial. These past climate trends may not only include the recent past, but also the distant past. In order to do this, paleoclimatology and the understanding of paleoclimactic trends would be a useful skill and tool to employ.

1.3 Question

To be more specific, since the issue of climate change regards temperature of the Earth most directly, paleoclimactic trends regarding the average temperature of the Earth would be most useful to get a thorough understanding of. For that reason, this paper will be using only paleclimactic data regarding the temperature of the Earth. In order to get a thorough understanding of this path of data, one of the first useful steps is making a model of it. We expect to be able to model this data because in general, paleoclimactic data is known for having "repeating" aspects, which implies that modelling tools such as the use of sinusoidals will prove useful. This question adds to this idea by asking the question,

> **Is it possible to make a mathematical model of paleoclimactic data that measures the temperature of the Earth? Would this model be feasible to apply to our greater significance—gaining an understanding of modern climate trends?**

2 General Approach

In order to pursue this question with utmost efficiency, this paper will gather paleoclimactic data that measures temperature over the past thousands of years. The data that we will use will come from the National Centers for Environmental Information from NOAA.

To be more specific, the data that we will be used is from this directory.[2] In this directory, there is no file with data that models paleoclimactic data of the whole Earth's temperature. Instead, there are 698 files that model paleoclimactic data for different locations (698 locations total).

So, in order to model for the whole Earth, we will take the "averages" of the measured temperatures in the data for the individual locations. Then, we will attempt to make a mathematical model of this "averaged" dataset.

This will have the same intended purposes as if there was a single dataset that represented temperature of the whole Earths. This claim is made under the assumption that the 698 files in the directory model locations that are fairly spaced out over the whole world, such that location bias can be avoided.

[1] www.drought.gov/data-maps-tools/paleoclimatology-data

[2] www.ncei.noaa.gov/pub/data/paleo/reconstructions/climate12k/temperature/version1.0.0/Temp12k_directory_LiPD_files

3 Data

3.1 Extracting the Data

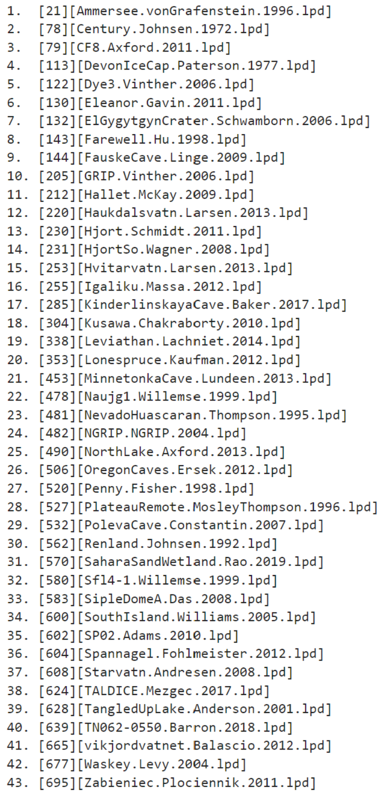
Extracting the data from this directory is a little more complex than typical data extraction, largely due to the file format. The files in the directory are in the LiPD format, which stands for Linked PaleoData, with the .lpd file extension. This data format cannot be visualized or even accessed using typical data visualization tools, such as text editors, for example. Thus, when extracting the data in python, we must employ the *LiPD Utilities* module.[3] LiPD data files are structured in a way that stores data values/measurements in something called a *timeseries*. Using simple data transformation techniques in theory should make it really easy to transform a timeseries into typical structured dataset formats. Thus, our first line of direction is to extract the timeseries from the LiPD data and transform them into dataframes, structured dataset formats.

3.2 Issues with the General Approach

We attempt this on the 698 files in the directory at one time. However, we encounter multiple stebacks with doing this.

1. The first setback is that for some of the LiPD files in the directory, a timeseries that represents temperature doesn't exist in the data, which makes the program fail immediately. The following files are the files that fall under this error:

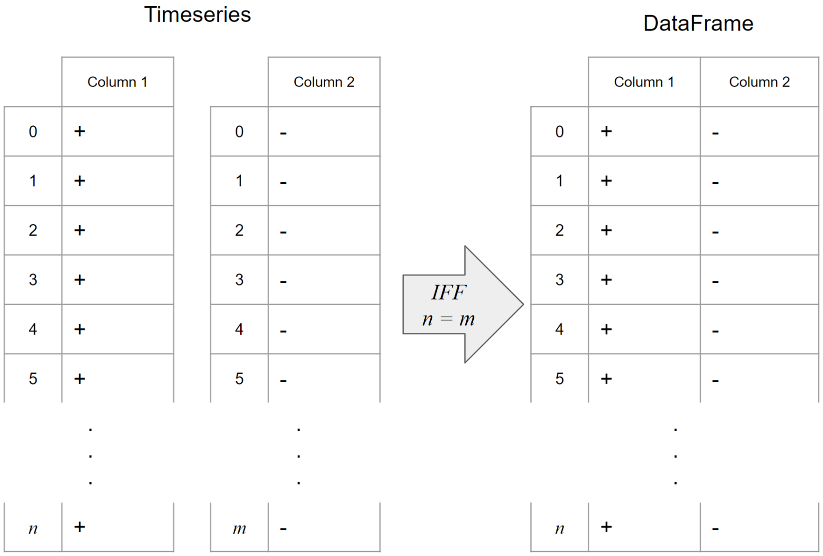
#. [Position in Directory][Filename]



```
1.  [21][Ammersee.vonGrafenstein.1996.lpd]
2.  [78][Century.Johnsen.1972.lpd]
3.  [79][CF8.Axford.2011.lpd]
4.  [113][DevonIceCap.Paterson.1977.lpd]
5.  [122][Dye3.Vinther.2006.lpd]
6.  [130][Eleanor.Gavin.2011.lpd]
7.  [132][ElGygytgynCrater.Schwamborn.2006.lpd]
8.  [143][Farewell.Hu.1998.lpd]
9.  [144][FauskeCave.Linge.2009.lpd]
10. [205][GRIP.Vinther.2006.lpd]
11. [212][Hallet.McKay.2009.lpd]
12. [220][Haukdalsvatn.Larsen.2013.lpd]
13. [230][Hjort.Schmidt.2011.lpd]
14. [231][HjortSo.Wagner.2008.lpd]
15. [253][Hvitarvatn.Larsen.2013.lpd]
16. [255][Igaliku.Massa.2012.lpd]
17. [285][KinderlinskayaCave.Baker.2017.lpd]
18. [304][Kusawa.Chakraborty.2010.lpd]
19. [338][Leviathan.Lachniet.2014.lpd]
20. [353][Lonespruce.Kaufman.2012.lpd]
21. [453][MinnetonkaCave.Lundeen.2013.lpd]
22. [478][Naujg1.Willemse.1999.lpd]
23. [481][NevadoHuascaran.Thompson.1995.lpd]
24. [482][NGRIP.NGRIP.2004.lpd]
25. [490][NorthLake.Axford.2013.lpd]
26. [506][OregonCaves.Ersek.2012.lpd]
27. [520][Penny.Fisher.1998.lpd]
28. [527][PlateauRemote.MosleyThompson.1996.lpd]
29. [532][PolevaCave.Constantin.2007.lpd]
30. [562][Renland.Johnsen.1992.lpd]
31. [570][SaharaSandWetland.Rao.2019.lpd]
32. [580][Sfl4-1.Willemse.1999.lpd]
33. [583][SipleDomeA.Das.2008.lpd]
34. [600][SouthIsland.Williams.2005.lpd]
35. [602][SP02.Adams.2010.lpd]
36. [604][Spannagel.Fohlmeister.2012.lpd]
37. [608][Starvatn.Andresen.2008.lpd]
38. [624][TALDICE.Mezgec.2017.lpd]
39. [628][TangledUpLake.Anderson.2001.lpd]
40. [639][TN062-0550.Barron.2018.lpd]
41. [665][vikjordvatnet.Balascio.2012.lpd]
42. [677][Waskey.Levy.2004.lpd]
43. [695][Zabieniec.Plociennik.2011.lpd]
```

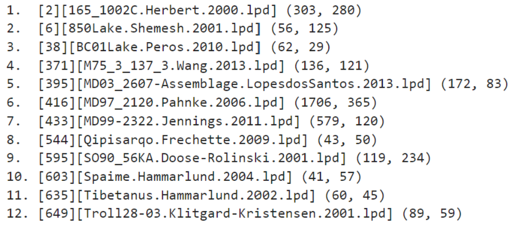
[3] https://nickmckay.github.io/LiPD-utilities/

2. The second setback is that of the files that *do* contain a timeseries that represents temperature, the timeseries for temperature conflicts with the length of the timeseries for *age*, the "x-value" that is being used to track the temperature. Because timeseries are stored in LiPD files independently of each other (to a certain degree), this does not cause any issues in the *extraction* of the timeseries itself. Rather, it causes an error in the *transformation* step of the process. The data format that we are transforming the LiPD file into are DataFrames. DataFrames are designed such that arrays (which will be the independed timeseries after conversion in this case) that go into the dataframe have the same length. The following diagram shows the difference between how independent timeseries are structured and how DataFrames are structured.



Here, the sizes of the timeseries are not reliant on each other, and can have differing values. This does not directly harm the LiPD data. However, in a DataFrame, the columns are both required to have the same length, and are determined by a single value n, rather than two—n and m. Thus, in order for timeseries to be transformed into a DataFrame, n and m must have the same value. This is the case for the majority of the files, but some still have differing sizes for age and temperature. The affected files are the following:

#. [Position in Directory][Filename] (age size, temperature size)



```
1.  [2][165_1002C.Herbert.2000.lpd] (303, 280)
2.  [6][850Lake.Shemesh.2001.lpd] (56, 125)
3.  [38][BC01Lake.Peros.2010.lpd] (62, 29)
4.  [371][M75_3_137_3.Wang.2013.lpd] (136, 121)
5.  [395][MD03_2607-Assemblage.LopesdosSantos.2013.lpd] (172, 83)
6.  [416][MD97_2120.Pahnke.2006.lpd] (1706, 365)
7.  [433][MD99-2322.Jennings.2011.lpd] (579, 120)
8.  [544][Qipisarqo.Frechette.2009.lpd] (43, 50)
9.  [595][SO90_56KA.Doose-Rolinski.2001.lpd] (119, 234)
10. [603][Spaime.Hammarlund.2004.lpd] (41, 57)
11. [635][Tibetanus.Hammarlund.2002.lpd] (60, 45)
12. [649][Troll28-03.Klitgard-Kristensen.2001.lpd] (89, 59)
```

3. Finally, the last setback is just that the process of extracting, transforming. and storing the data in the way that we want for 698 files takes up a lot of time and resources, which highly stunts the efficiency of this project.

Given these three setbacks, we now need to revise our approach.

4 Revising the Approach

In our first approach, we said that we would "average" all of the temperatures from the 698 files. However, we know this isn't possible because of the first two setbacks, and highly inefficient because of the third setback. In order to work around the first two setbacks, we can just not transform the datasets that are affected, and just do the extracting and storing steps. However, we would still need to extract, transform, and store the data from the remaining 643 files in the directory, which is too tedious and costly due to the third setback. This is not only because of the added transformation step, which takes up the most resources, but also because storing the data in the transformed DataFrame is more time-consuming and costly than just storing the timeseries.

4.1 Another Issue with the Approach

Although it is fairly straightforward to work around the first few issues, we encounter a new, more pressing issue. The way that we planned to take the averages for the temperatures in the data was that we would average each temperature across all datasets for a given set of years. However, after inspecting the data and the timeseries for *age*, we see that not only do the ages not match across datasets, but they are also not always evenly spaced, so a clever "shift" of the ages to map them onto eachother such that the averaging would work is not possible. It is also not only extremely costly to iterate through each timeseries for age and only grab the ages that exist in all of the datasets, but we also can't guarantee that there will be enough—or even *any*—data left to model and average. If there does happen to be any left, it is highly likely that it would be to scarce to get anything meaningful out of it in terms of analysis. Thus, we have to employ a completely new approach, even if it would be abstract compared to more concrete data analysis techniques.

4.2 *DEATHED Extraction And Transformation of Highly Enigmatic Data*

We develop a new approach, and call it *"DEATHED,"* which stands for:

DEATHED Extraction And Transformation of Highly Enigmatic Data[4]

From start to finish, DEATHED describes the method used on the process of going from the "enigmatic data" from the LiPD files to data suitable to model mathematically, while taking care of and working around all the described setbacks and issues.

Here are the steps and methods used in the DEATHED process:

1. Extracts the needed independent data (timeseries that include temperature) for all files (that contain temperature, in this case) in a given list of files, encluding the imporperly structured "enigmatic" data.

2. Stores the data in a way in which the structure of the data allows for "enigmatic data" (not transformation). This works around the second issue brought up.

3. Filters through the data and deterimnes which data would make the transformation step impossible, based solely on the second issue, and then removes these "illegal" data from the rest of the data.

4. Because transforming all the legal data would be too tedious, in order to work around this issue, this method randomly (or non-randomly) chooses a set number of files to perform the transformation on.

5. After choosing the files, this method independently transforms the data into structure data formats such as DataFrames for easier workability.

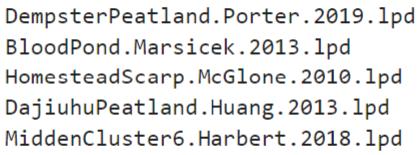
[4] Yes, "DEATHED" is indeed part of the acronym for *DEATHED!*

6. This leads to the issue brought up in section 4.1. Because we can't assume to guarantee that enough values exist in all datasets, and it will be too tedious to check, DEATHED assumes that there are *no* values that we want in the any of the datasets. Although this is almost certainly untrue, it allows us to make the data predictable. In order to do this, DEATHED choses a set of desired x-values based on the highest first value of x and lowest last value of x, and "implants" them into the data. This method of choosing the desired x-values is used because inputting values outside of the range will cause the results to be more prone to error. For each of these x-values, this method implants a missing value into the data. Here, it is worth noting that if a value in this set *does* exist, the program will keep the existing value instead of inputting a missing value. It is also worth noting that the data is sorted by age after the implantation.
7. After this step, we *can* guarantee that all the DataFrames have the desired x-values. Now, we need to fill in the missing values.
8. To fill in these missing values, DEATHED uses a technique called *Linear Interpolation,* which basically predicts what values *would exist* as measurements for the desired x-values, despite the x-values not actually existing. This technique will be explained in more detail further on in the section.
9. After the missing values are filled, this method eliminates all values in all the datasets that aren't in the list of desired x-values. This leaves us with a set of datasets that all have the same x-values.
10. This allows us to then take the average of these datasets, which outputs a single averaged DataFrame that describes the average global temperature for some amount of years, which is what this project required.

We now apply DEATHED to this project and the data it uses. Steps one through three are heavily program based in comparison to the descision making involved with those processes, so they do not warrant their own section or sections. The only thing worth noting is that from the LiPD files in the directory, we extracted and stored all the timeseries from files that contained temperature and removed the ones that didn't. Then we removed all of the "illegal" data sets.

5 Choosing Specific Data

We decide to randomly choose 5 datasets out of the 698, because if they are random, there is still no location bias, even if we can't guarantee every location being covered. The datasets we chose came from the following LiPD files.

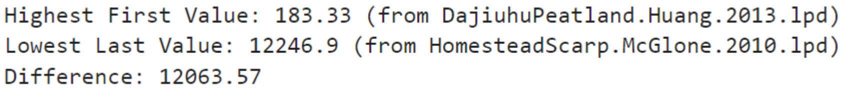


```
DempsterPeatland.Porter.2019.lpd
BloodPond.Marsicek.2013.lpd
HomesteadScarp.McGlone.2010.lpd
DajiuhuPeatland.Huang.2013.lpd
MiddenCluster6.Harbert.2018.lpd
```

The next step, transforming the data, is primarily code based, and does not require a section because it doesn't involve any significant descision making processes.

6 Determining the Desired x-values

Using DEATHED, we determine the desired x-values based on the data from the files chosen in the previous step. We get the following from the files:



```
Highest First Value: 183.33 (from DajiuhuPeatland.Huang.2013.lpd)
Lowest Last Value: 12246.9 (from HomesteadScarp.McGlone.2010.lpd)
Difference: 12063.57
```

This tells us that the desired x-values would be between 183.33 and 12246.9. For simplicity, the values we chose will be determined by the 1205 values in the following sequence:

$$190, 200, 210 \ldots 12240$$

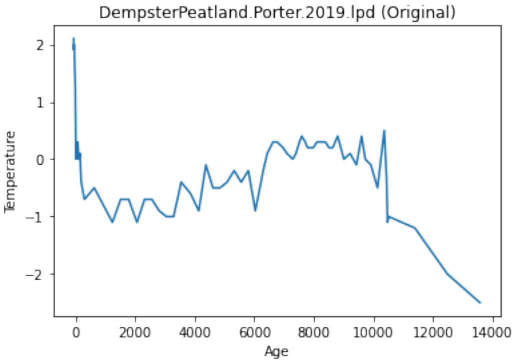
We then insert the values into the DataFrames, which is too program based to warrant more explanation.

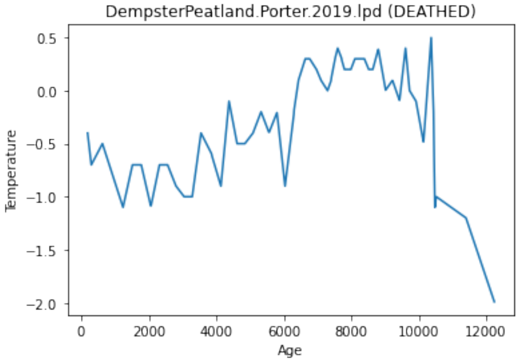
7 Data Interpolation

The previous step leaves us with DataFrames with all of our desired ages that output missing values, along wih all the real data. In order to fill in the missing values, we use a method called linear interpolation:

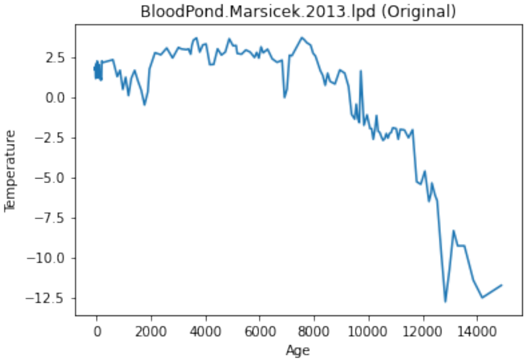
> Because the desired ages we chose were between the highest first value and the lowest last value, we can say that for each desired age, there is at least one point on both sides of the desired age that actually exists in correspondence with a real temperature. This method takes advantage of this phenomenon. For every desired value, if the age coresponds with a missing value, then the following steps are performed:
>
> 1. The two neighboring points that don't contain missing values (one on each side) are treated like they are coordinates on a Cartesian plane, with the values $(x, y) = (\text{age, temperature})$.
>
> 2. Now, a line can be drawn connecting these two points, in the format $f(x) = mx + b$.
>
> 3. Finally, using this line function, the desired value can be computed by doing $f(x)$, where x is the desired age, and $f(x)$ gives us the temperature.

We perform this step on all of the data, which fills in the missing values. Then, we eliminate the extra values for all of the datasets—the ones that aren't in the set of desired ages. The following shows graphs of the data from the five locations, before and after the implantation and interpolation.

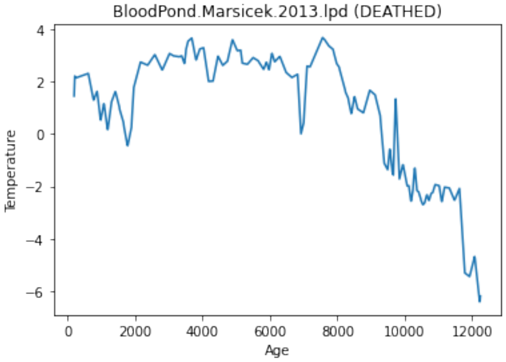




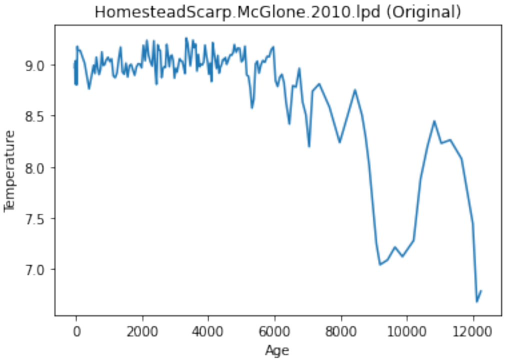
DempsterPeatland.Porter.2019.lpd (DEATHED)



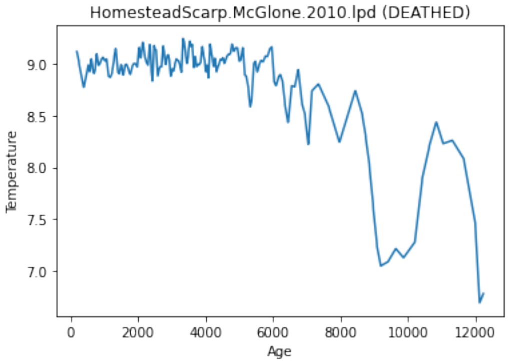
BloodPond.Marsicek.2013.lpd (Original)

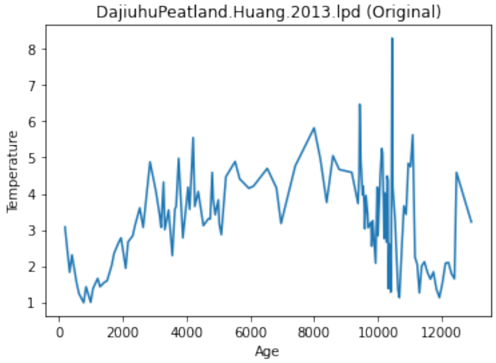


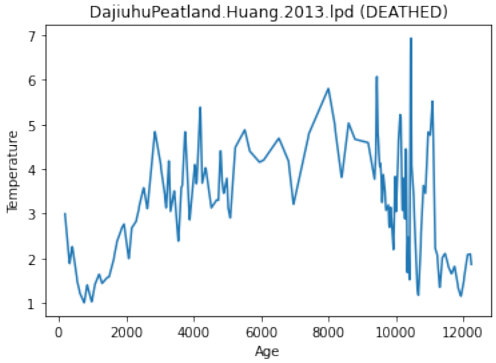
BloodPond.Marsicek.2013.lpd (DEATHED)

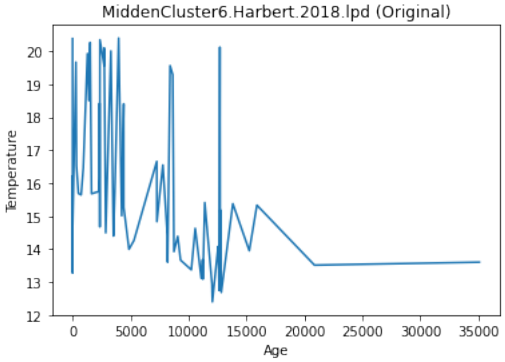


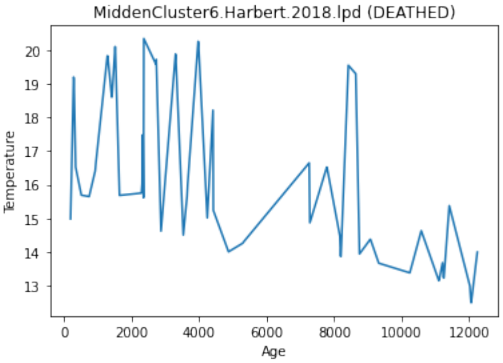
HomesteadScarp.McGlone.2010.lpd (Original)


HomesteadScarp.McGlone.2010.lpd (DEATHED)


DajiuhuPeatland.Huang.2013.lpd (Original)


DajiuhuPeatland.Huang.2013.lpd (DEATHED)

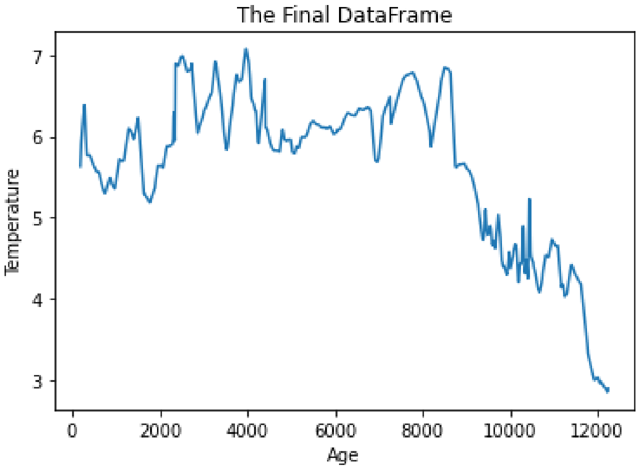

MiddenCluster6.Harbert.2018.lpd (Original)



As seen in the images, these processes don't affect the data so much. Although with the last pair of images it seems like the data changed a lot, this is because the range sharply dropped from ≈ 0–35000 to ≈ 0–12000.

8 Final Data

The previous step leaves us with 5 DataFrames that all have the same values for age. This would allow us to produce a single DataFrame that takes the average of all the temperatures across all DataFrames for all ages. We do this, and get a single DataFrame with the same ages as the individual ones, but the temperatures represent all of the frames, by taking the average values. This is the final data that we will be modelling, and it looks like the following:



9 Creating a Model

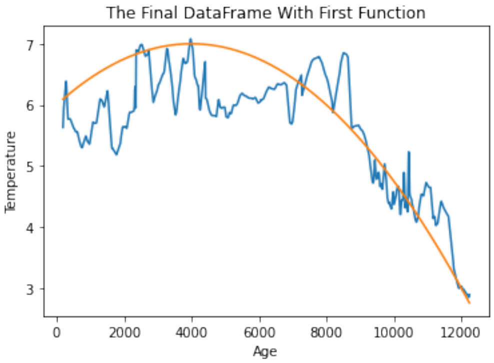
Now, we try to create a mathematical model of the data.

9.1 Quadratic

Just looking at the data visually, it seems like it has a general quadratic trend. We try the following function based on where it looks like the vertex would lie, and one of the minimum points in the data to the far right.

$$f(x) = -0.0000000625\,(x - 4000)^2 + 7$$

This produces the following graph:



9.2 Troubleshooting Specific Regions of the Model

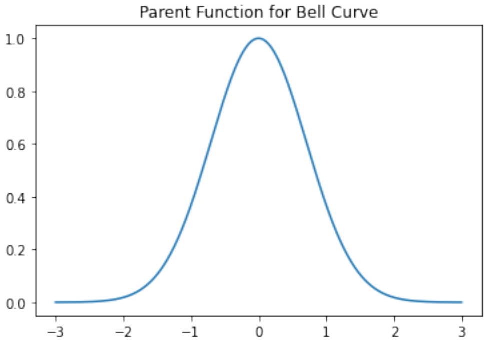
With the above model, some obvious issues lie in specific regions, even though the general trend seems to match. For example, in some places, the model is a lot higher than the data, and in other places, the model is a lot lower than the data. Thus, we need to find a way to anly adjust these parts of the model.

9.2.1 Using "Bell Curves" to Improve the Model

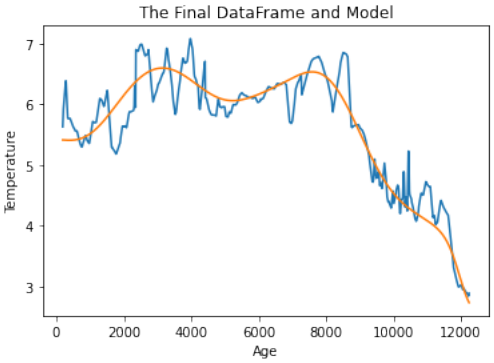
In order to do this, we use bell curves, which are determined by (transformations of) the following base equation:

$$y = e^{-(x^2)}$$

This produces the following graph:



This is extremely useful, because with the graph, only a specific portion of it is a bump, and the rest of the graph has a y-value close to 0. We can apply transformations to this graph, and add the graph to the quadratic in order to effectively adjust the height of the model in certain locations. We do this using transformations, and get the following model:



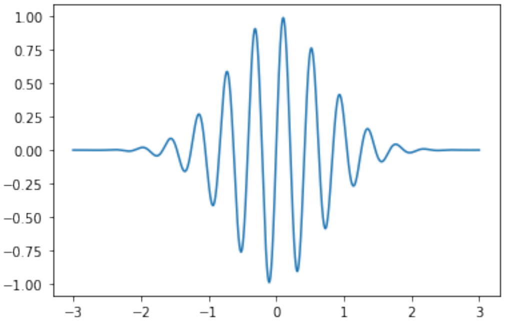
Our model as of now is determined by the following function:

$$
\begin{aligned}
f(x) = -0.0000000625\,(x-4000)^2 \\
+6.5 + 0.2^{e^{\left(-(0.0008(x-3000))^2\right)}} \\
-0.4^{e^{\left(-(0.0008(x-5000))^2\right)}} \\
+0.7^{e^{\left(-(0.0008(x-8000))^2\right)}} \\
-0.4^{e^{\left(-(0.0008(x-10000))^2\right)}} \\
-0.6^{e^{\left(-(0.0008(x-900))^2\right)}} \\
-0.7^{e^{\left(-(0.002(x-12300))^2\right)}}
\end{aligned}
$$

It is clear that this model is slightly better than the previous ones.

9.3 Sinusoidals

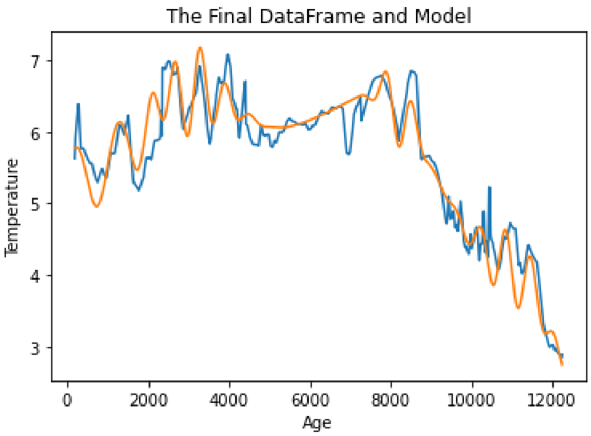
Finally, there seems to be some sinusoidal trends in the data, since it moves up and down relative to the most recent model. However, these sinusoidal trends only happen in some locations, and not others. If we add a sinusoidal to the model just like that, the trend will show up only in some locations. Thus, we need to find a way to limit the sinusoidals to specific domains. We can take advantage of the bell curve technique here. Multiplying a sinusoidal by the bell curve will only make the sinusoidal show up in the areas with the bell curve, since the rest of the graph would be multiplied by a value close to 0. For example, the following is the graph of $\sin(15x) \cdot e^{-\left(x^2\right)}$:

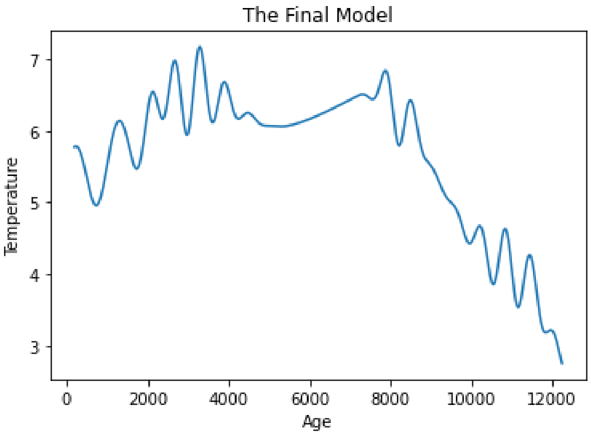


We now apply the sinusoidals to the model, multiplying them with the bell curves acordingly. Our model is now the following:

$$
\begin{gathered}
f(x) = -0.0000000625\,(x-4000)^2 \\
+6.5 + 0.2^{e^{\left(-(0.0008(x-3000))^2\right)}} \\
-0.4^{e^{\left(-(0.0008(x-5000))^2\right)}} \\
+0.7^{e^{\left(-(0.0008(x-8000))^2\right)}} \\
-0.4^{e^{\left(-(0.0008(x-10000))^2\right)}} \\
-0.6^{e^{\left(-(0.0008(x-900))^2\right)}} \\
-0.7^{e^{\left(-(0.002(x-12300))^2\right)}} \\
+1.2e^{\left(-(0.001(x-3000))^2\right)\cdot 0.5\sin(0.01x)} \\
+e^{\left(-(-0.0007(x-1000))^2\right)\cdot 0.5\sin(0.006(x+50))} \\
+0.9e^{\left(-(0.002(x-8200))^2\right)\cdot 0.6\sin(0.009x)} \\
+0.5e^{\left(-(0.001(x-11000))^2\right)\cdot \sin(0.01x)}
\end{gathered}
$$

This produces the following graph:





[5]

[5]

This matches most of the data pretty convincingly, so this will be our final model.

9.4 Statistical Analysis of Model

In order to statistically analyze the model, we use a tool called the R^2, which calculates how good a model is. It is determined by the following equation:

$$R^2 = 1 - \frac{\Sigma\left(y_i - f(x_i)\right)^2}{\Sigma\left(y_i - \mu_y\right)^2}$$

Here, the closer the R^2 value is to 1, the better the model is. Using our model and the data, we find that the R^2 of the model is **0.9010851388169703**. This is a fairly high R^2, which means that our model models the data well.

10 Broader Analysis of Model and Results

In section 9.4, we've established that the model we created is an excellent mathematical model of the data. However, we needed to make a lot of adjustments to the model in order for it to work. For example, we used the bell curves, which allowed our model to fit specific parts of the data better. Although this worked to model this specific data, this technique would not prove useful in making a model of this data in order to predict other values, since we practically designed this model around this data completely. In other words, this model has limited practical value outside the scope of this data. In addition, it is likely that many of the peaks and dips that the model fits is a result of anomalous trends, or trends that have little significance. For analyzing modern climate trends, a simpler model may have been better, such as the original quadratic that was proposed.

11 Conclusions

To answer the question,

> **Is it possible to make a mathematical model of paleoclimactic data that measures the temperature of the Earth? Would this model be feasible to apply to our greater significance—gaining an understanding of modern climate trends?**

As we have seen in section 9.4, it is possible to make a mathematical model of global temperature. However, as explained in section 10, this model is tailored to the data that this project explored, and cannot be used much outside of the scope of this data. This means that the model would hardly be useful in applying to a greater understanding of modern climate trends.

12 Further Inquiry

This paper looked at global trends of temperature. Staying in the lane of paleoclimatology, a line of further inquiry could be to look at other trends, such as percipitation. If this is not feasable in terms of paleclimactic data collection, then another line of further inquiry could be a more detailed analysis of more recent data for percipitation, temperature, or both.