Final Project For Physics 411: Analyzing Lower Vancouver Island Temperature

Mateo Branion-Calles, University of Victoria

December 15th 2021

I. Introduction

As part of this study, this project's goal was to analyze temperatures collected by the School-Based Weather Station Network. The data provided is broken into two sampling frequencies: minutely data with sampling rate of 1/60 Hz and hourly data with a sampling rate of 1/3600 Hz. The time, associated with each data set is in ordinal time and is not human readable. A general conversion from ordinal to human-readable times (YYYY/MM/DD) was required to achieve this. Both the hourly and minutely data have gaps of information where an individual station did not record any data. This has led to some stations with several months of missing data. Because of this, some stations were omitted from the final results of this project as it did not make sense to average and interpolate such a large time period. The analysis of this report includes finding the average temperature of July across all stations from the minutely data, looking at the Probability Density Function (PDF), Power Spectral Density (PSD), the Variance Preserving PSD, Spectrogram and cross-correlations. For the hourly data we will look at local gridding interpolations, the EOF of grid data, the strongest modes, and the time series of the EOF grid. This report is organized in such a way that the text is presentable and the images are readable. This was done so because many of the images are generated on subplots and require a larger font size to view the x and y ticks as well. In order to make this report even more readable the images will be submitted in a separate PDF titled 'Appendix' so that the reader has both the text and images simultaneously.

II. Interpolation of Data

Before we begin to analyze our data we must first dive into what is and isn't usable data. For our Minutely data the white spaces in Figure 1 visualize when our stations were not recording temperature. From this we can determine when reliable statics can be conducted. For example it is evident that the John Muir station was not recording temperature during most of 2018. The Minutely data is well-documented with only a few gaps of data that are easily masked and interpolated.

In order to fill the gaps of data, we applied an averaging mask to any NaN data point, effectively taking the mean temperature of all the other stations at the corresponding time of the NaN data point. Because there are instances where all stations have NaN at a specific time, a Cubic Spline Interpolation is used to fill the remaining values. As we can see in Figure 2, the white gaps have been filled with yellow, blue and green and are consistent with winter and summer cycles. No stations were removed from the minutely data. The same logic was applied to the hourly data. The hourly stations: Willows, UVicSci, Shawnignan School, Quadra, Margret Jenkins and Kelset were removed from our analysis of the hour data due to large discontinuities (Figure 4.).

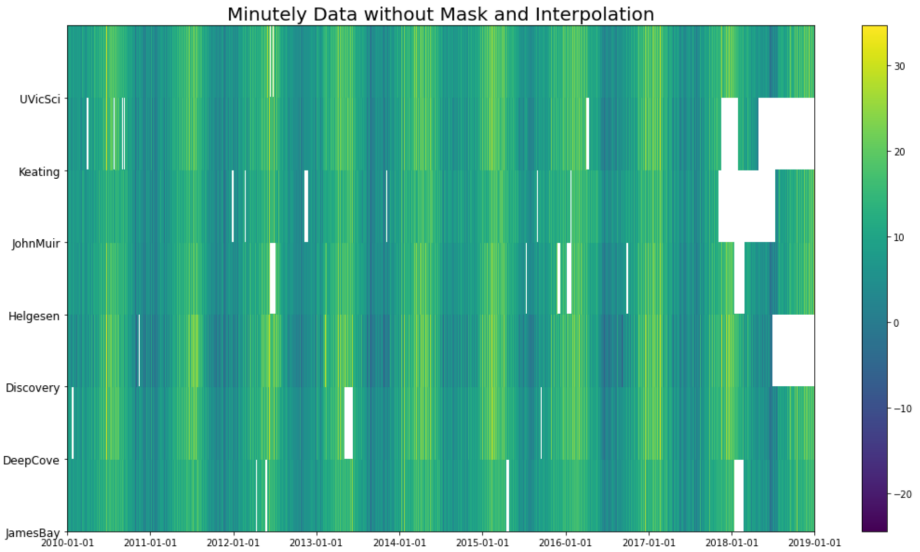

Figure 1. Unaltered Minutely Data across all 7 stations. Summer Cycles can be seen in yellow, where winter cycles can be seen in blue. White values show when a station was not collecting data

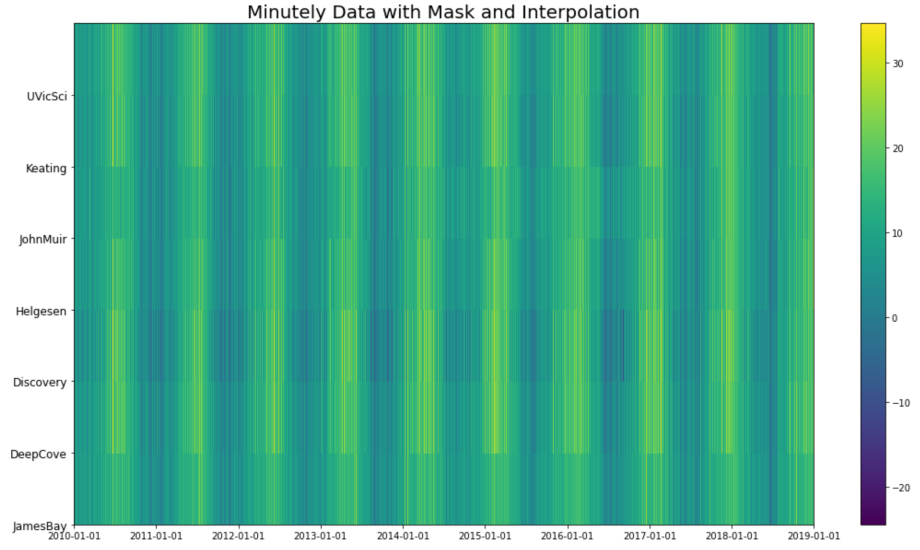

Figure 2. Masked and Interpolated Minutely Data. See Figure 1. for season cycle information

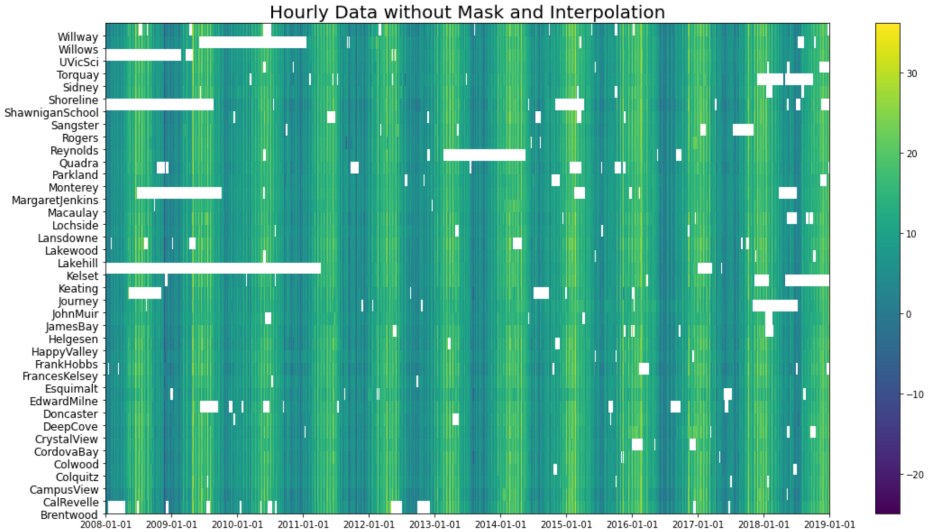

Figure 3. Unaltered Hourly Data across all 39 stations. See Figure 1. for season cycle information

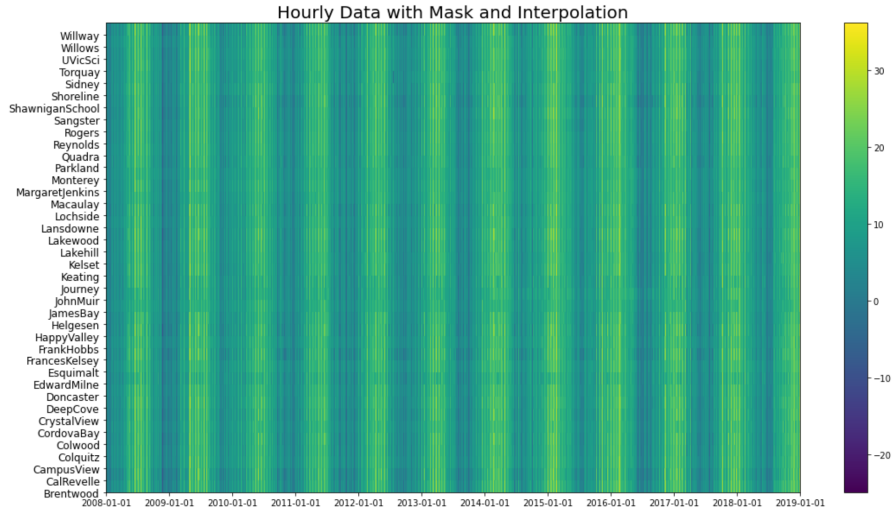

Figure 4. Masked and Interpolated Hourly Data across all 39 stations. See Figure 1. for season cycle information

III. Minute Data

First, it should be mentioned that the minute data starts from January 1st 2010 and ends August 31st 2018. Second, it should be noted that all of the statistics and minute data analysis conducted on this report was done with the interpolated minute data, thus there is not a single NaN value associated with any of the following figures or tables.

Table 1. provides us with some standard statistics such as mean and standard deviation. Most weather stations have a mean temperature of about 10 C° with an standard deviation of about 6 C°. Considering that the stations are not too far apart, it is logical that the data sets mean and standard deviations would be relatively similar. It would be interesting to have data on the northern peak of Vancouver Island to analyze the difference of average temperatures with the southern region of the island.

III.1 Time Series and Filter

Figure 5. shows the full time series of the minute data after it has been masked and interpolated. Seasonal trends can be harder to view when looking at the full time series, so a uniform filter from the scipy python pack was applied to our minute data. This process is similar to that of a moving mean window. The window size used was 50,000. Figure 6, our filtered time series, shows the seasonal changes of each station in a readable format.

III.2 PDF

Another way we can look at the time series is if we were to generate an approximate probability density function. This can be done by binning our data sets and creating a normalized histogram, which represents a normal distribution. Figure 7 shows a histogram for all 7 stations overlapped with a normalized Gaussian function. The Gaussian function makes use of each station's standard deviation and mean. As we can see both approximate PDF and the Gaussian function are centred at each station's temperature mean. The histogram visually shows us that the temperature for all stations is rarely less than 0° and did not raise above 25C° . All but James Bay and John Muir are skewed left. The reason for this is unknown.

III.3 Mean and Standard Deviation of Summer and Winter

Another important analysis to look at is how the temperature mean, stand deviation, minimum and maximum of any given month varies from year to year. If we are to look at Figure 8, we see a steady increase for all stations between the years 2010 and 2015. We see a peak in 2015 and then drop in 2016. But after including 2016, the trend oscillates. However, 2019 still has a higher temperature than 2010. Figure 9 is similar to Figure 8, but looking at the standard deviation, we can see in Figure 9 that the trend is similar to that of the mean for the same time period. Further analysis requires us to look at the difference between the maximum and minimum temperature for July. We see in Figure 10. that the difference in temperature is shrinking. If we then just look at the minimum temperature for July across the same time period we also see that the minimum temperature is increasing. The minimum temperature is typically associated with nighttime or an extremely cold day. From this, we can draw two potential conclusions: either the coldest night is getting warmer, or the coldest day is getting warmer, or possibly both. In any case July is not getting colder. Further analysis would require us to look at the standard deviation of night and day separately for July. Personal time to further analyze

this does not exist. This is a small sample size and strong connection between this and climate change cannot be determined. Perhaps if we had data for the years 1990 to 2021, we would see a stronger climb. It is very likely the heat dome from 2021 would contribute to a large standard deviation.

III.4 Power Spectral Density

A power spectral density for all seven stations was generated in order to find recurring characteristics of our time series. Figures 12 and 13 show the PSD and Variance Preserving PSD respectively. Figure 12 has a 95% confidence interval associated with it, however the error bars are short enough that they do not appear visible on the graphs. These PSDs were generated using scipys Welch's method with a sampling rate of 1/60Hz and 2^{14} segments. Every station has an extremely similar trend with a noticeable diurnal cycle likely associated with sun set and sun rise as these two events occur once daily, and would have the most drastic recurring daily change to temperature. We also see the harmonics of the PSD in figure 13. The harmonics do not represent anything physical but are just artifacts of the time series used to generate a PSD.

III.5 Cross Correlation

A Cross Correlation test was conducted between all the stations of the minute data using the entire data set. The lowest correlation value was found to be around 0.93 and the highest (excluding the obvious perfect correlation between the same stations) was around 0.98, showing us that if something affects one station there is a high probability it will affect the others.

III.6 Spectrogram

Our Spectrograms visually shows us the evolution of the PSD. A frequency of one cycle per day was used. This allows us to see the PSD as it changes over time. Not too much stands out in the spectrogram that our PSDs from figure 12 and 13 did not tell us. As per usual the cycle is governed by the diurnal cycle with a clear yellow line at 1 CPD. But one thing the spectrogram does show us is how the PSD loses a significant amount of energy during the winter seasons, evident by the Discover Station, which shows the largest change of energy during the winter.

III.7 Discovery Station

The Discovery Station had the largest standard deviation of any station and it is evident in both the PSD and the spectrogram. It is most evident in how much noise is in frequencies greater than 10 CPD and how much larger the harmonics are in the PSD variance preserving compared to the other graphs.

Station	Mean (C°)	Standard Deviation (C°)
JamesBay	10.21	4.61
DeepCove	10.95	6.17
Discovery	9.94	7.14
Helgesen	10.29	5.70
JohnMuir	10.17	5.06
Keating	10.64	5.94
UVicSci	11.18	5.54

TABLE I. Mean and Standard Deviation of the entire Minute Data

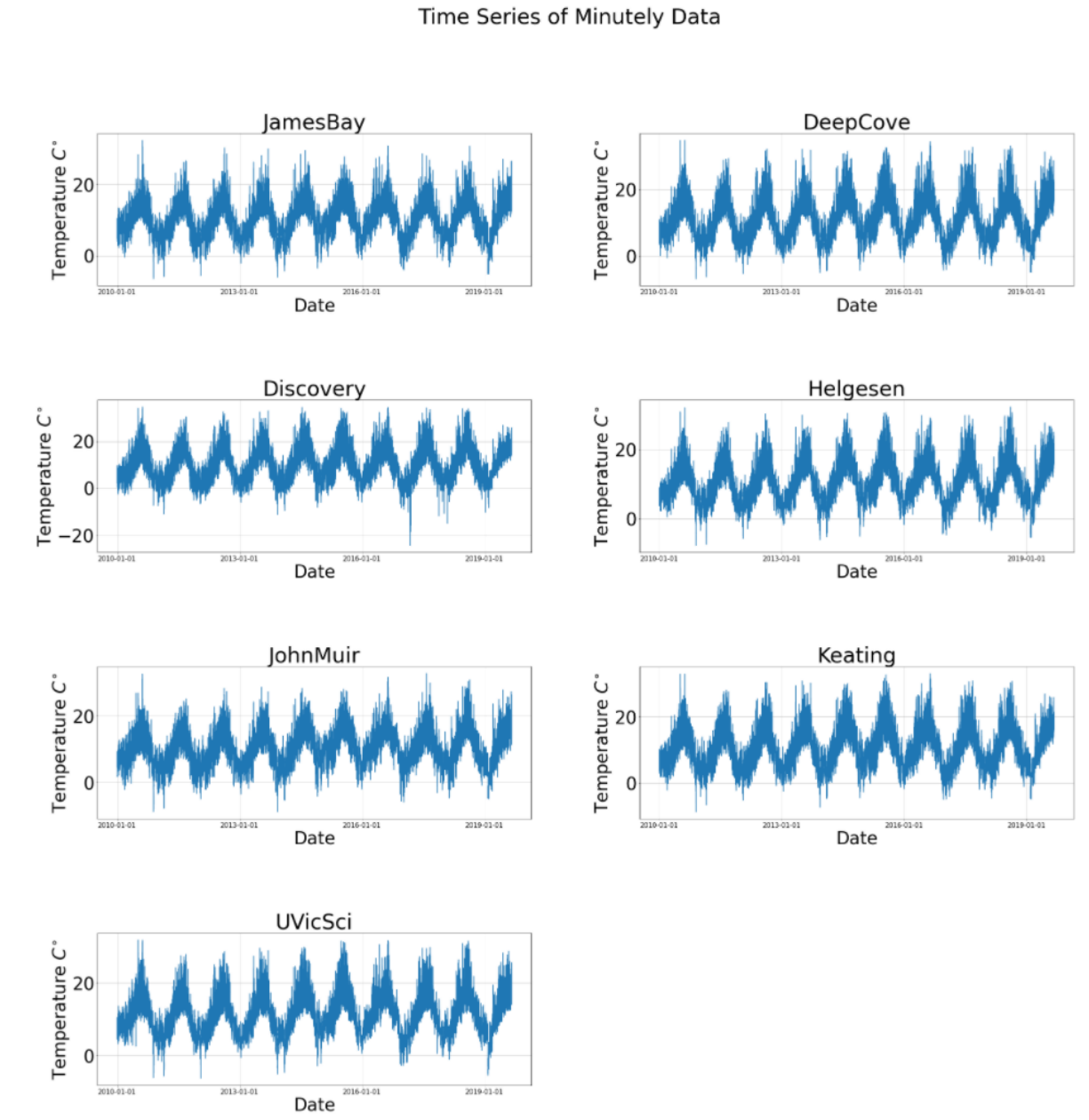

Figure 5. Full time series of all stations in the minutely data after interpolation and masking. Expected seasonal patterns

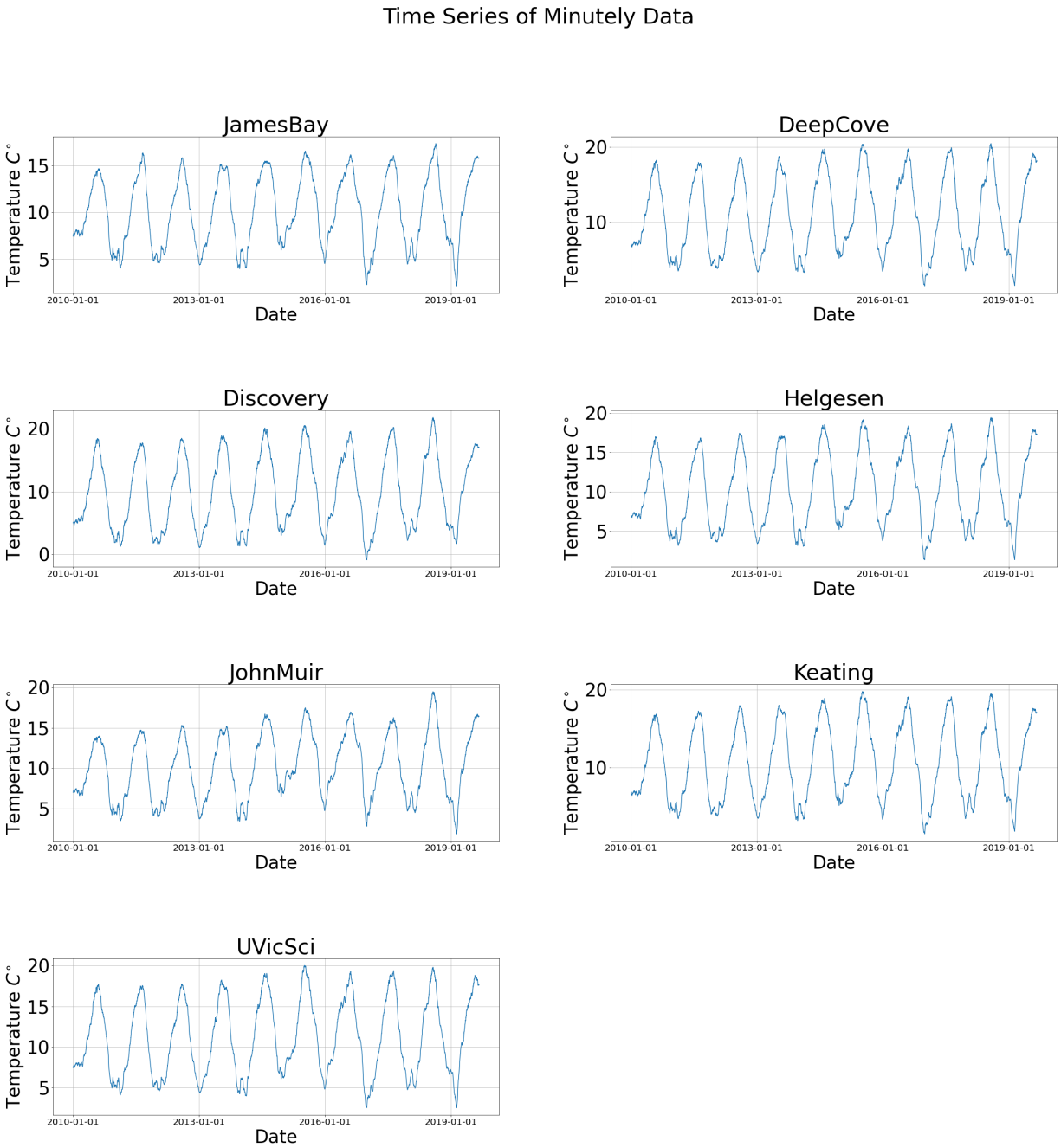

Figure 6. Uniform filtered Time Series of all stations in the minutely date. Information has been lost but the periodicity of seasons is easier to read.

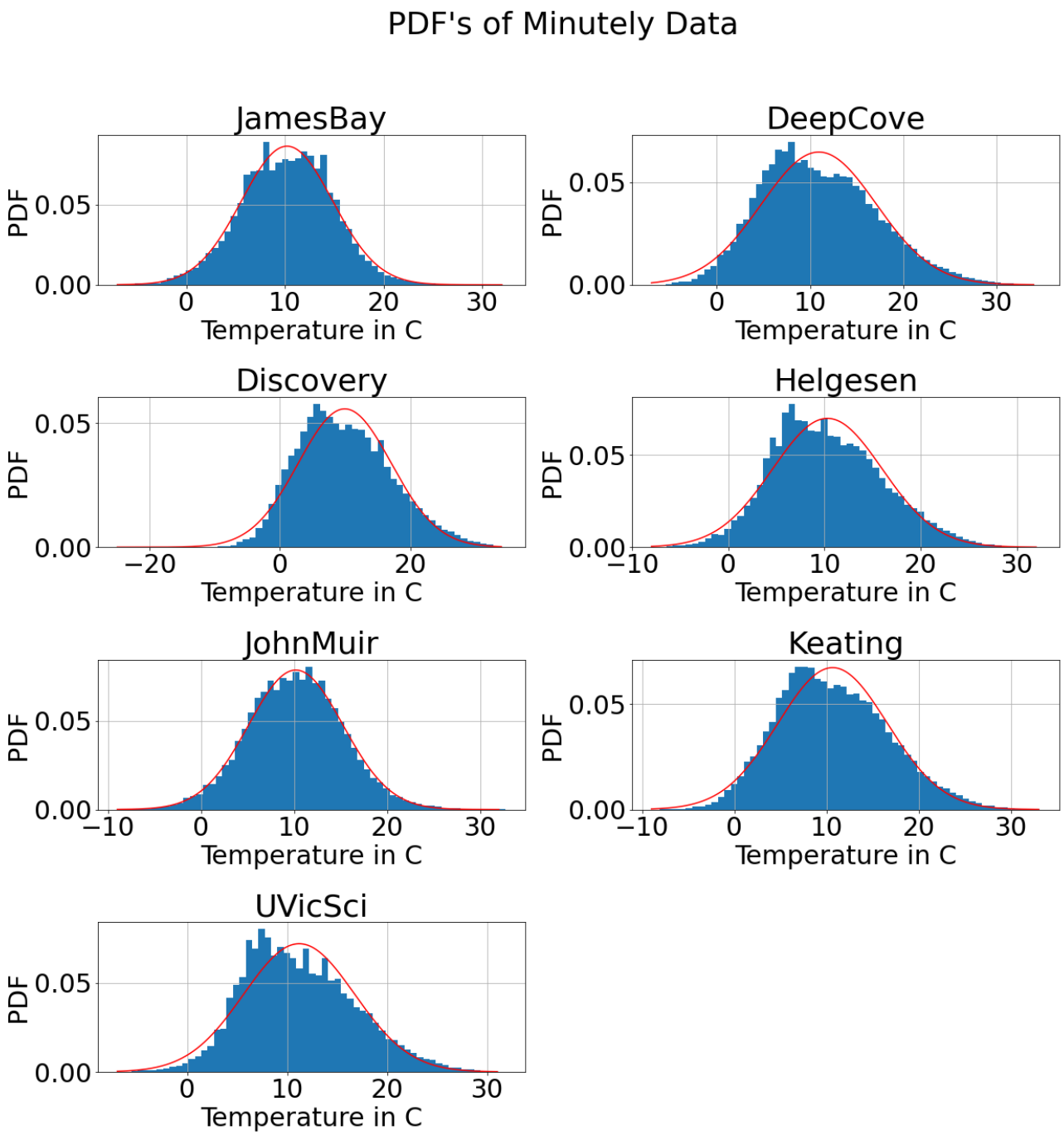

Figure 7. Approximate probability density of interpolated minute data with a normalized Gaussian fit centred at the mean with width Ω

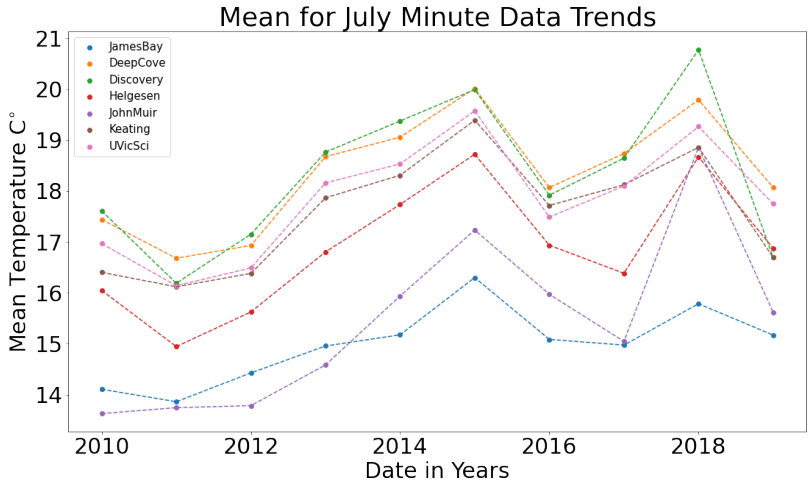

Figure 8. Mean temperature for minute data spanning from the year 2010-2019 looking only at July Data. Visible increase in mean temperature over the years

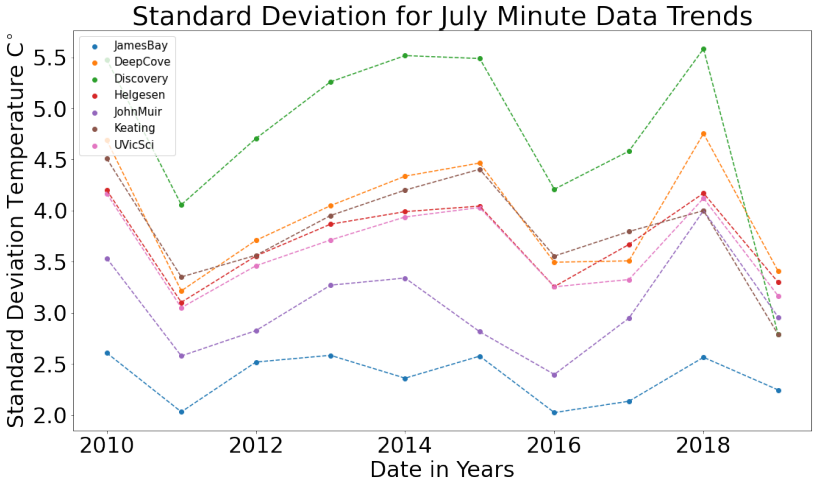

Figure 9. Standard Deviation temperature for minute data spanning from the year 2010-2019 looking only at July Data. Visible increase in standard deviation of temperature over the years

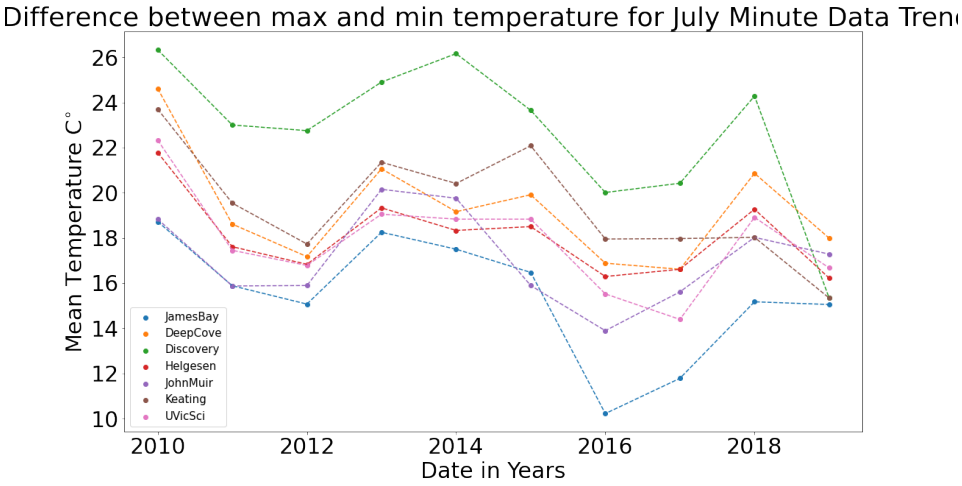

Figure 10. Difference between the maximum and minimum temperature values of July for a given station and year. Difference is visibly decreasing.

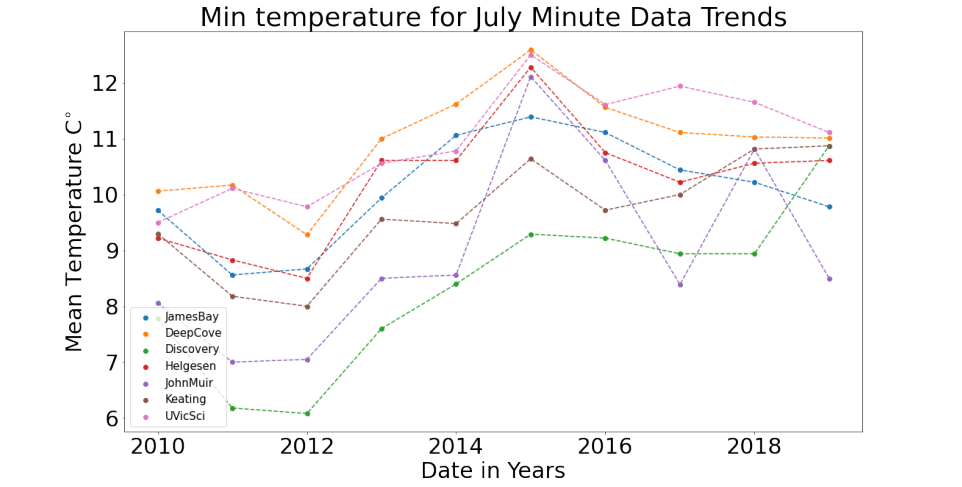

Figure 11. Minimum temperature of July for the minute Data. Visibly increasing linearly

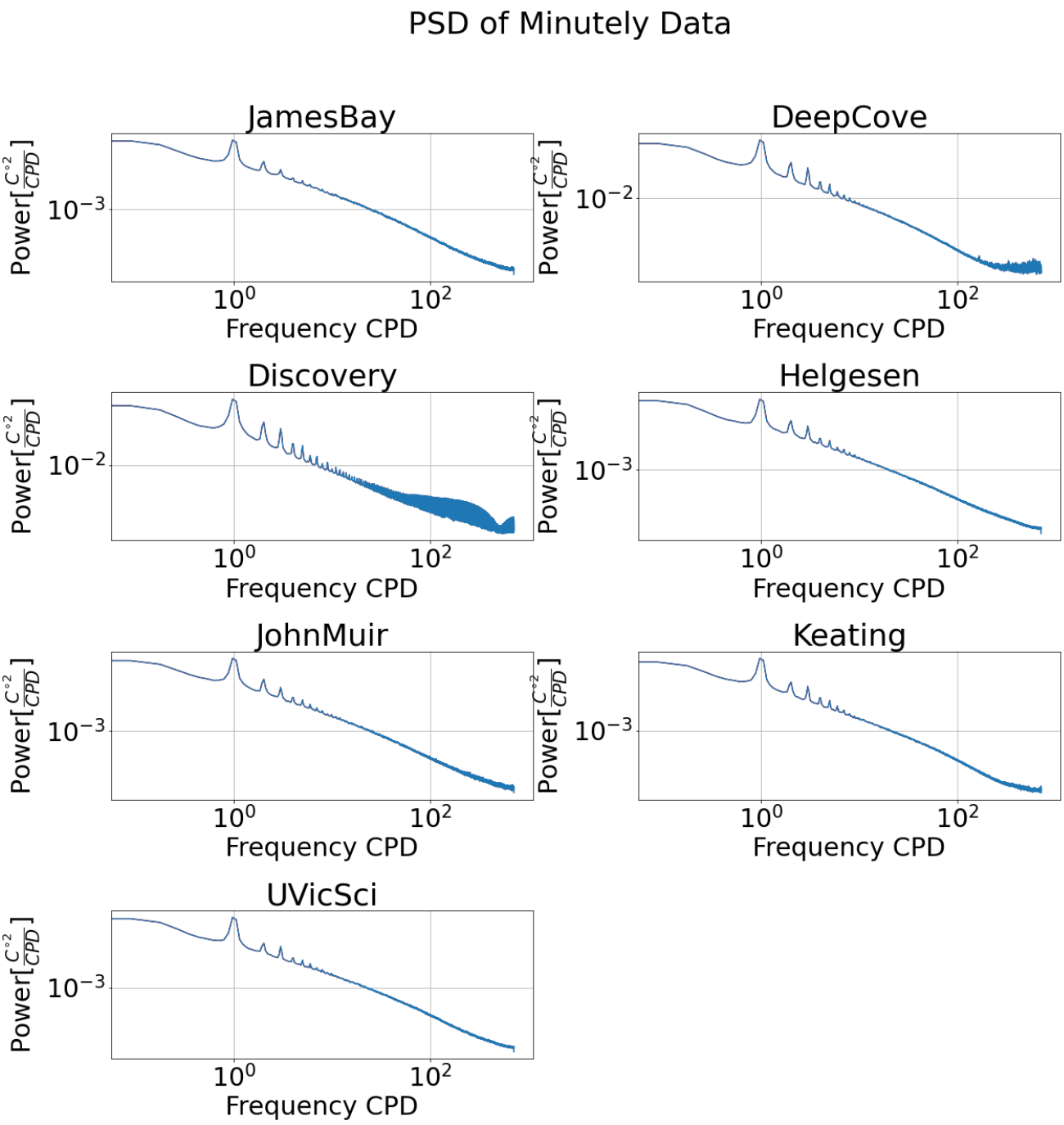

Figure 12. Power Spectral Density of all 7 stations of the minute data. 95% confidence interval is included but is not visible due to the uncertainty being so insignificant

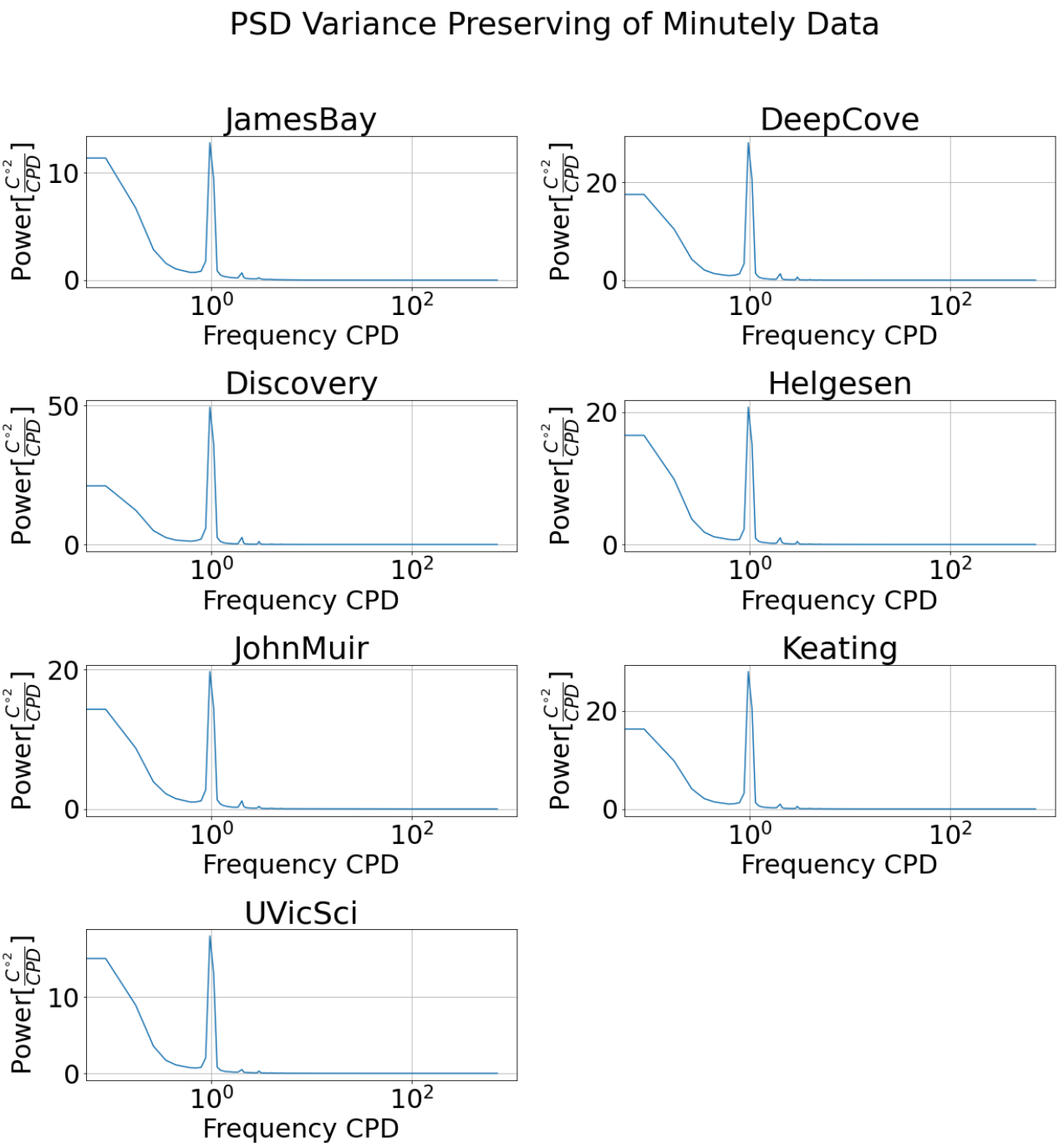

Figure 13. Variance preserving PSD of all 7 stations. Sun set and sun rise at the diurnal cycle (1CPD)

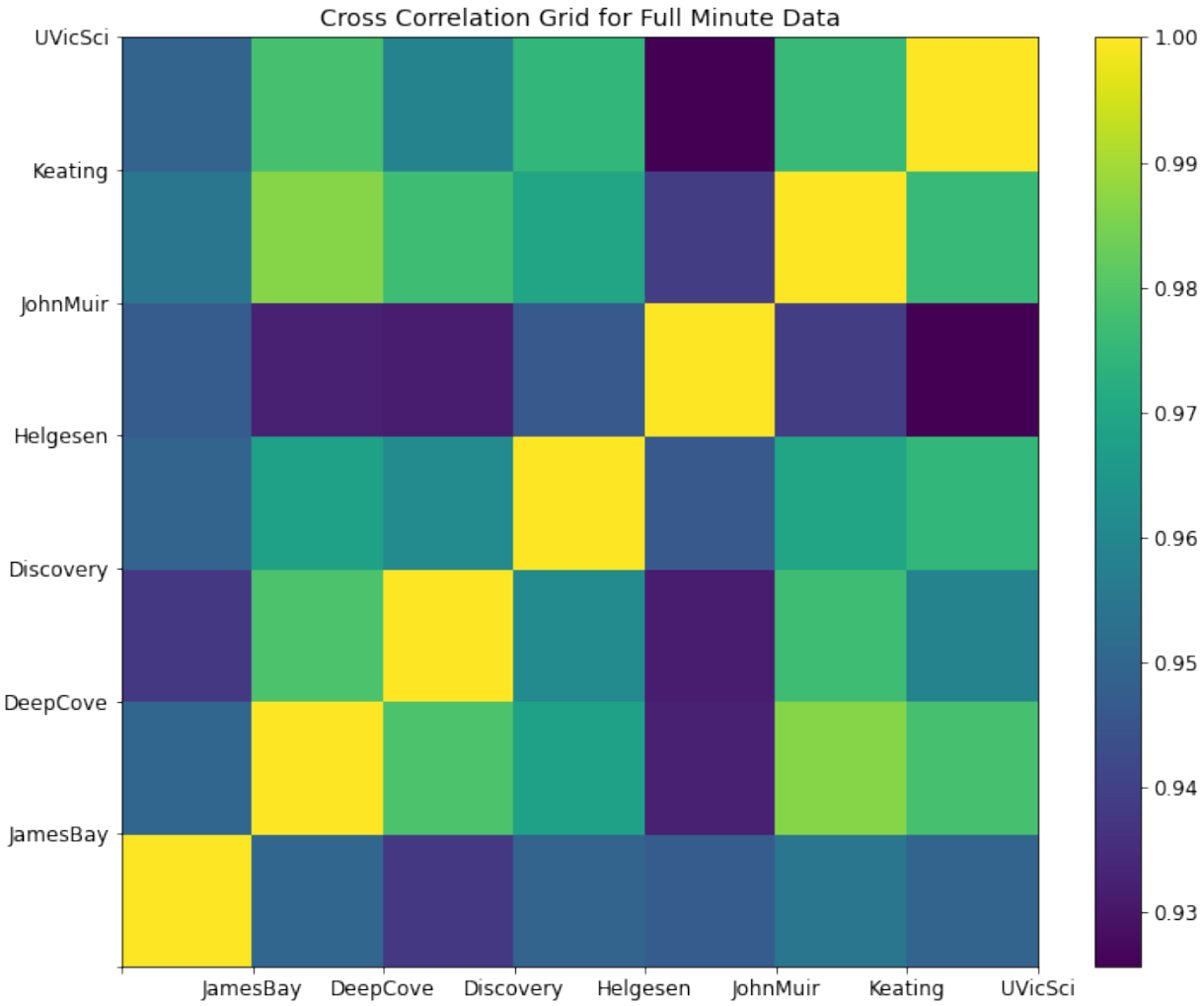

Figure 14. Cross Correlation between the minute stations across the whole dataset. The yellow diagonal is auto correlated as the correlation between the same station with the same phase will be 1.0. The Cross Correlation between every station is very significant

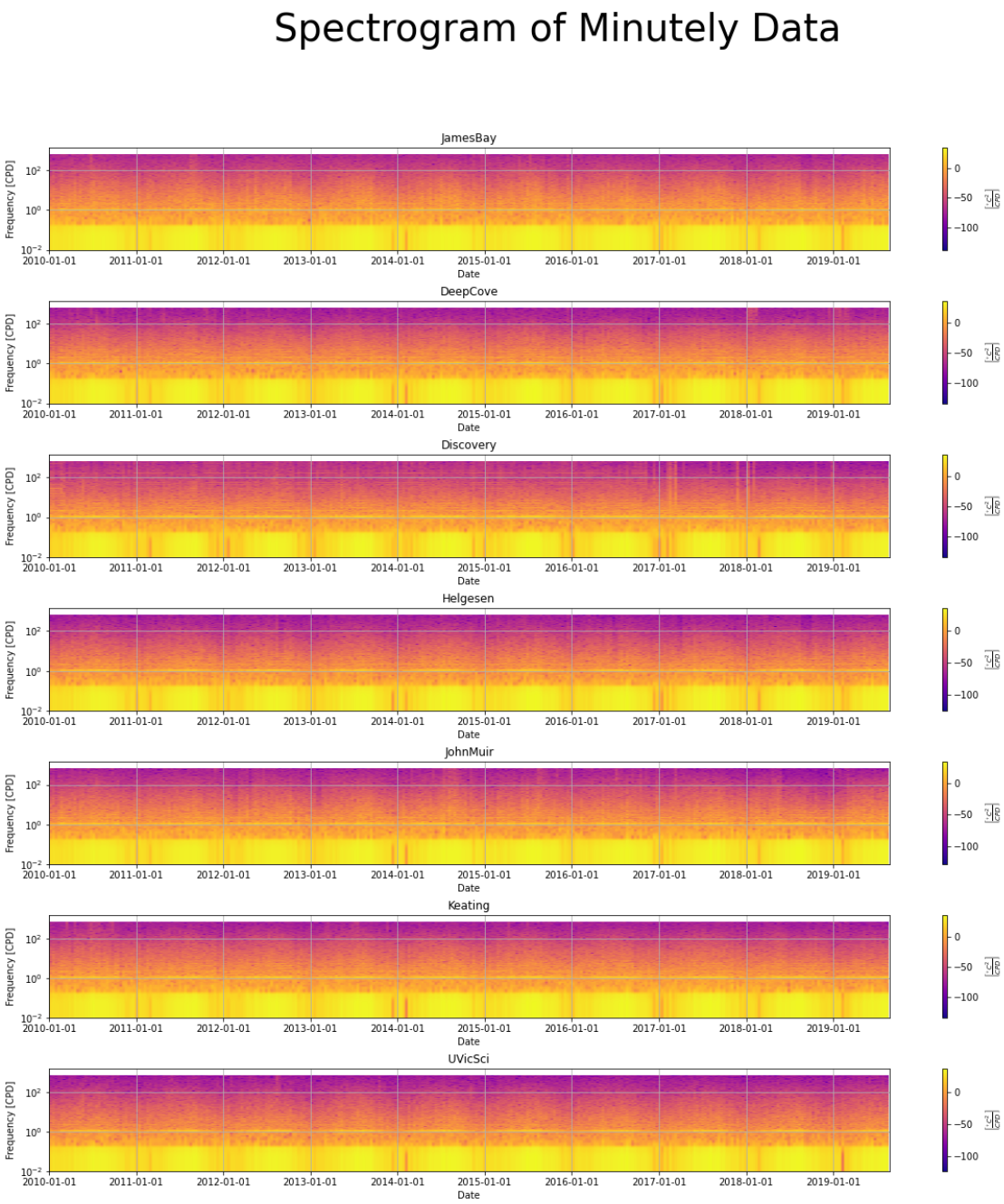

Figure 15. Spectrogram of all 7 minute stations. The sample frequency is 1/60 Hz or 1 PSD per day

IV. Hourly Data

IV.1 Local Data interpolation

The following data show we applied three different types of local interpolations to our spatial data. This was done by creating a mesh grid of the longitude and latitude. The mesh grid was created by using the corner values from the hourly data. This grid is visualized in the 2nd graph of figure 16, the nearest neighbor interpolation. It is immediately obvious that the nearest neighbour interpolation is the least accurate interpolation method of the 3; as it does not truly interpolate between points, but assigns value to a local unassigned area, rather than take into consideration multiple surrounding stations. We can see from the other graphs that the linear and cubic interpolation methods consider all surrounding stations. Like most interpolation functions, the cubic method is less efficient but more accurate than linear as it has more degrees of freedom to interpolate. We see in the Cubic and linear figure that the areas near the coastline tend to be warmer than those further inland.

IV.1 EOF: Strongest mode time series And Spatial Data

An EOF was generated using a cubic grid. This grid was generated using a Spatial grid of dimensions 400x200 latitude and longitude coordinates. Unfortunately, my computer did not hold enough memory to run the program as my kernel would crash or I would get a memory overload error when I attempted to run the program so I asked Kenny, to run my EOF data as his computer was able to handle the memory overload. Figure 17's second graph accounts for the 98.1% of the variability of the time series. The following time series represent the descending importance of variability with variabilities 0.09%, 0.06%, 0.04% respectively. Figure 19 shows the spatial variability of the strongest modes in Figure 18. From Figure 19 we can see that the variability of inland is quite high compared to that of the coastline. If time permitted and my computer had the computational power to run this code, comparing EOF plots from summer and winter would be interesting to observe.

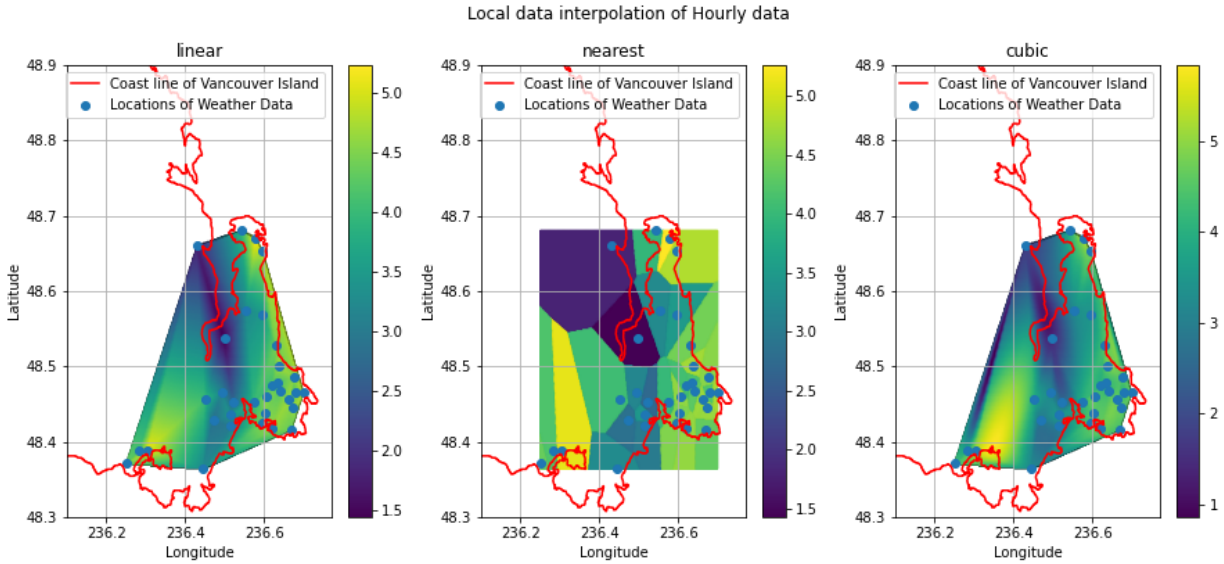

Figure 16. Local data interpolation of linear, nearest neighbor and cubic

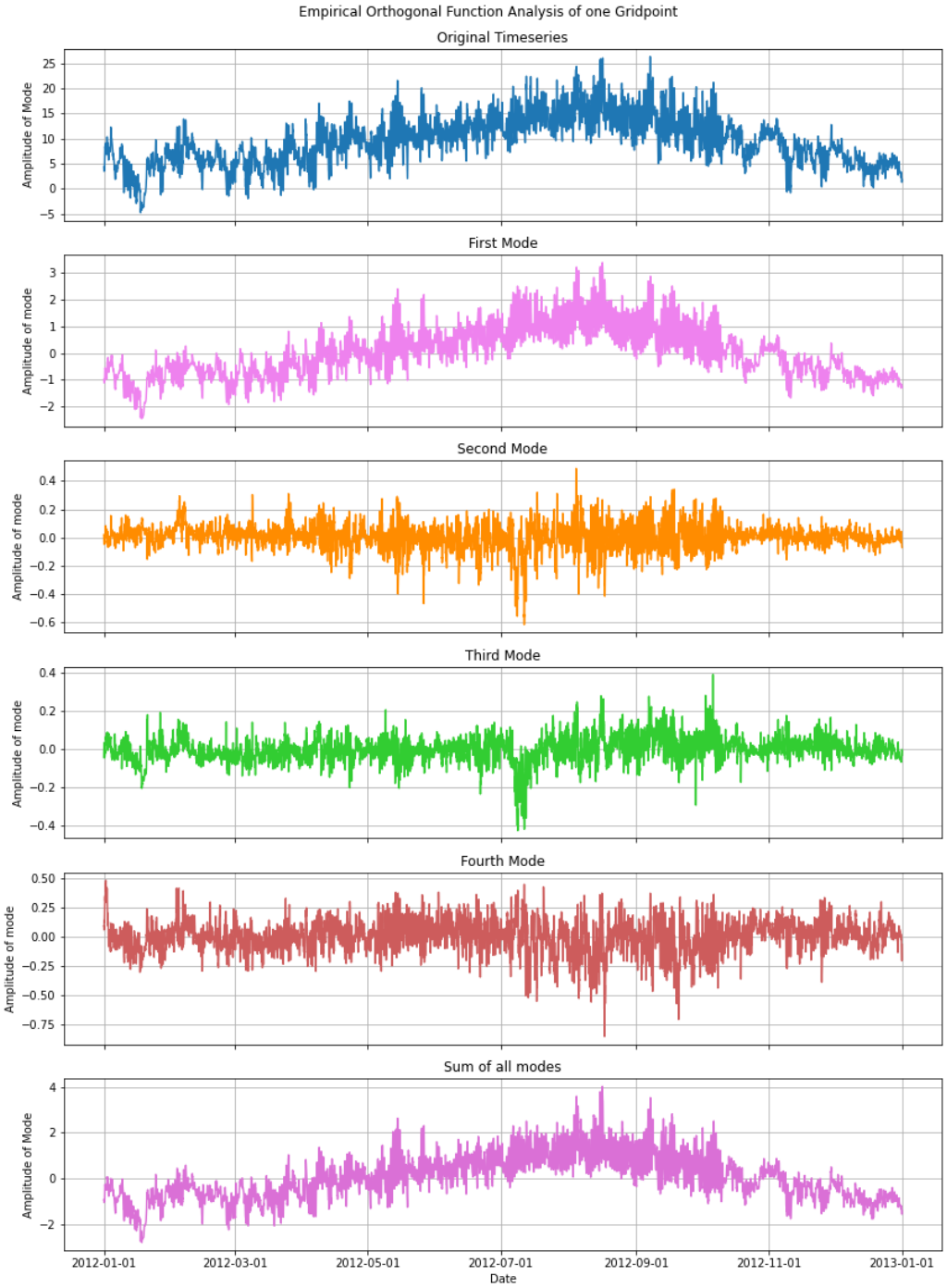

Figure 17. Strongest modes of hourly data

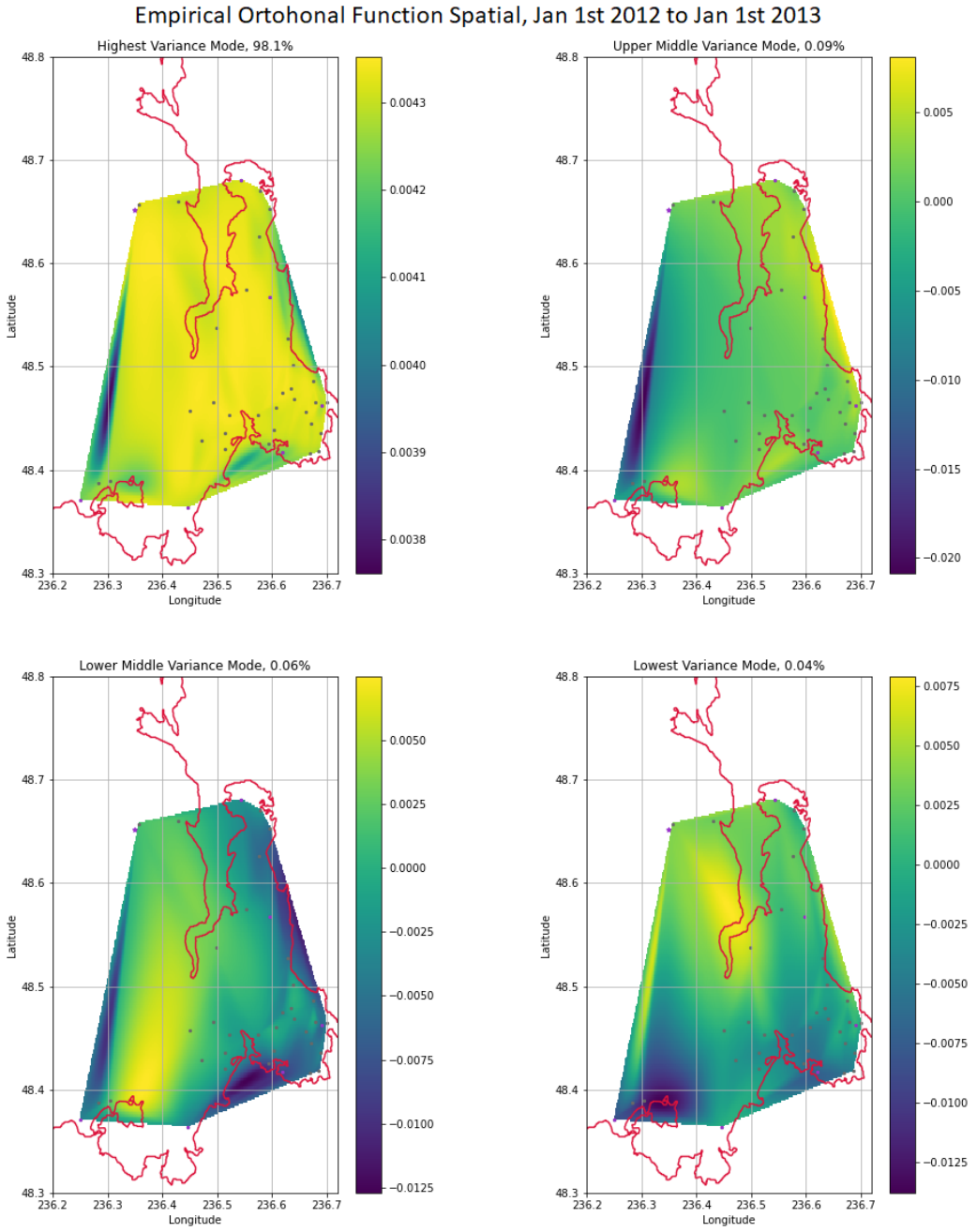

Figure 18. EOF Grid data

V. Summary

Both the minute and hourly data were masked and interpolated to have no NaN values. 6 Stations were removed from the hourly data but no stations were removed from the minutely data. The following is related to the minute data. A filter was applied to the full time series to ensure that the periodicity of yearly seasons were in fact accounted for in our minutely data. An approximate probability density function was generated using a histogram. This histogram was skewer left for 5 of the 7 stations and showed that temperatures rarely dipped below $0C^{\circ}$ nor raised over $25C^{\circ}$. As expected, our power spectral density had a diurnal cycle, likely the sun set and sun rise were affecting the temperature drastically once a day. All stations had a nearly identical PSD graph. A cross correlation was then applied and found that every station had significant correlation. Lastly for the minute data, a spectrogram was applied. As expected the spectrogram had a clear peak at 1 cycle per day. The spectrogram gave us insight into the energy loss during the winter seasons and the energy gain during the summer. The Discovery station had a larger variance than the other 6 stations and this is evident in the PSD and the spectrogram. The following will be a summary on the hourly data. Three Local data interpolations were applied to masked and interpolated hourly data. From here, we have determined 2 things. Nearest neighbour is useless in this context, as it resembles a birds eye view of farm land, and thus incredibly inaccurate. The linear and cubic methods have an extremely similar trend but cubic has higher temperatures further inland than linear and linear has warmer temperatures near the coast Our EOFs most important node had a variability of 98.1%. We can see that variability inland is quite high and that the coastline had much less variability than that of the inland.