An Analysis of Lower Vancouver Island Temperature Patterns

Austin Beauchamp, University of Victoria
Department of Physics & Astronomy
Submitted December 14th, 2019

Contents

1 Introduction **1**

2 Minute Data **1**
2.1 Minute Analysis . . . 2
2.2 Regression analysis . . . 3
2.3 Station Cross-Covariance . . . 9
2.4 Fourier Analysis . . . 9

3 Hour Data **14**
3.1 2D Interpolation . . . 15
3.2 EOF Analysis . . . 21

1 Introduction

This project aims to analyze temperature data from the *School-Based Weather Station Network* [1], which is a series of weather stations situated in the Greater Victoria area on lower Vancouver Island, Canada. Temperatures have been recorded at two sampling rates: 1/60 Hz and 1/3600 Hz, or one sample per minute and one sample per hour, respectively. The were seven stations recording at a sampling rate of one sample per minute, and 39 stations sampling at a rate of one sample per hour. The dataset analyzed contains recordings from 2008 to the end of 2019 for the hourly recordings, and from 2010 to the end of 2019 for the minutely recordings. The data is not always continuous; not all stations recorded data 100% of the time, however timestamps between stations are kept consistent. These non-recordings will be discussed in future sections. First, the minute resolution data is examined. The overall dataset is investigated and basic statistical analysis is conducted in time-space. The average temperature of every month over the years is analyzed and plotted with regression lines to determine if the mean temperatures are increasing. Cross-covariance's between stations is examined to explore similarities between station weather patterns. Following time-space analysis, frequency-space is explored with power spectral densities (PSDs).

2-Dimensional interpolation is also explored. As the minute data contains a much higher resolution (which is better for spectral analysis), the hourly data is used for 2D interpolation techniques. Using the interpolated data, spatial patterns between stations is examined by calculating the empirical orthogonal functions (EOFs) and plotting the strongest modes.

2 Minute Data

The seven stations that recorded temperature every minute are named: Deep Cove, Discovery Elementary, Helgesen, James Bay, Keating, the University of Victoria (UVic), and John Muir. The initial analysis will focus on the minute resolution data.

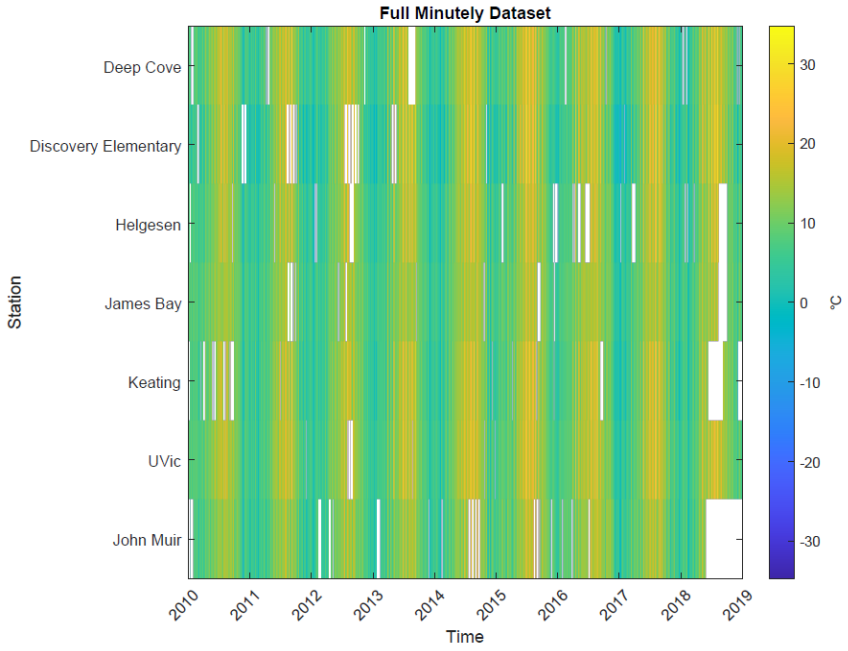

Figure 1. Full minute dataset, visualized. White values represent null recordings. Summer/winter cycles are observable as eachs station transistions from higher temperatures (yellow) to lower temperatues (blue) each year.

2.1 Minute Analysis

To first visual our dateset, the temperature from each station at each time is plotted in figure 1. Each station is labelled along the y-axis, with the time of each recording along the x-axis. Each bar represents one temperature, with white bars representing NaN's (Not a Number), which is when the station failed to record a temperature. Most stations are consistent with temperature recordings, however John Muir appears to have gone offline towards the end of 2019. Figure 2 plots the time series' more conventionally, with each temperature against time. As in figure 1, the overall temperature cycles are prominent over each year. Winter and summer cycles are seen during the lower temperatures and higher temperatures, respectively.

Table I has the basic statistics for each station. NaN's were excluded from the calculations, not using any forms of interpolation. First impressions show that the Greater Victoria area has a temperate climate, all stations hovering roughly around $10^oC \pm 6^oC$, with UVic having the highest average, and Discovery Elementary having the lowest.

	Mean $[^oC]$	Standard Deviation $[^oC]$
Deep Cove	10.91	6.2
Discovery Elementary	9.82	7.3
Helgesen	10.16	5.6
James Bay	9.99	4.5
Keating	10.41	5.9
UVic	11.11	5.5
John Muir	9.93	4.9

Table **I**. Basic statistics for full minute resolution datasets. Dropped NaN values for calculcations.

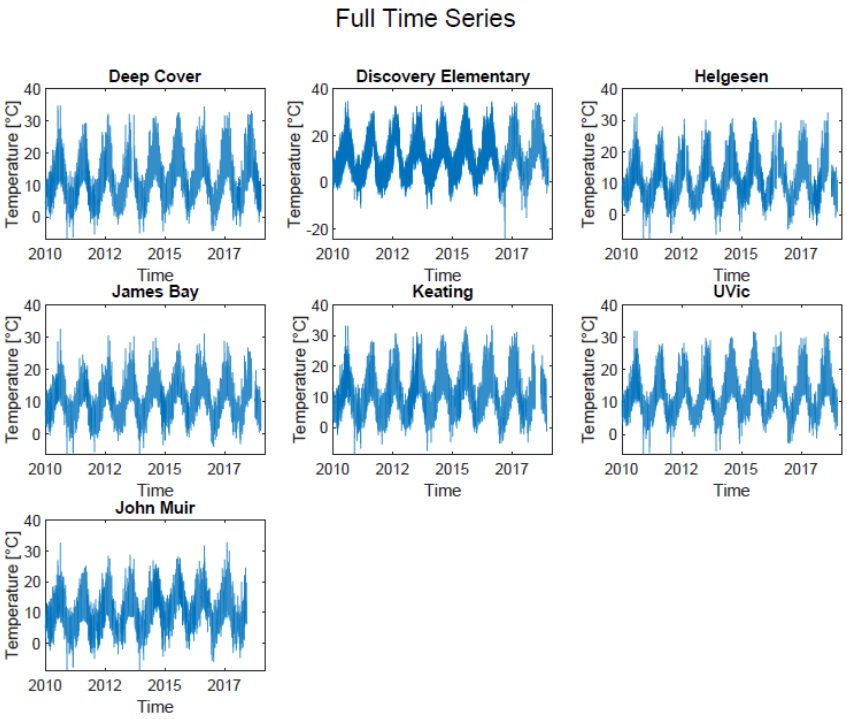

Figure 2. Each standard time series with temperatures plotted against time.

To show the temperature cycles more clearly, each time series is heavily smoothed in figure 3. The smoothing is done using a moving mean with a window size of 50000. This type of smoothing was choosing due to its quick execution and high smoothing factor. This technique demonstrates the season temperature cycles, but the high smoothing window causes a large amount of information loss. Helgesen appears to have the largest variability within the first few years, also dipping down to the lowest temperatures in most years. However, this contradicts the standard deviation in table **I**, where Helgesen has a middle deviation, whereas the smoothing graph would predict Helgesen to have the largest.

Another method of visualizing the dataset is to create approximate probability density functions by generating histograms of each sample, presented in figure 4. Each histogram is overlayed with a Gaussian curve centred at the sample mean with a sigma of the sample standard deviation. Each station's histogram approximately resembles the normal distribution, with some tending to skew slightly to the left. Consistent with table **I**, each station has similar mean temperatures and standard deviations.

2.2 Regression analysis

In modern times, a natural question to ask is whether or not the temperate is increasing due to global warming. We will attempt to tackle this problem by analyzing the average temperature of each month to determine if the temperature is increasing with 95% confidence. To start, figure 5 plots the average temperature of July across the entire dataset for each station. Stations for Deep Cove, Discovery Elementary, Helgesen, Keating, and UVic all follow very similar patterns, while James Bay and John Muir are noticeably lower in temperatures. This is not explained geographically, as James Bay and Jon Muir are distanced by about 0.3^o longitude. All stations had a peak in July of 2015, where the average temperature in 2016 dropped lower. 2019 also includes a significant increase of the July temperature, expect John Muir and Keating do not contain data for that July.

In attempt to determine whether the average monthly temperature has been increasing, the July temperatures for the UVic station are plotted in figure 6. A linear regression line is fitted to the data, along with a 95% confidence

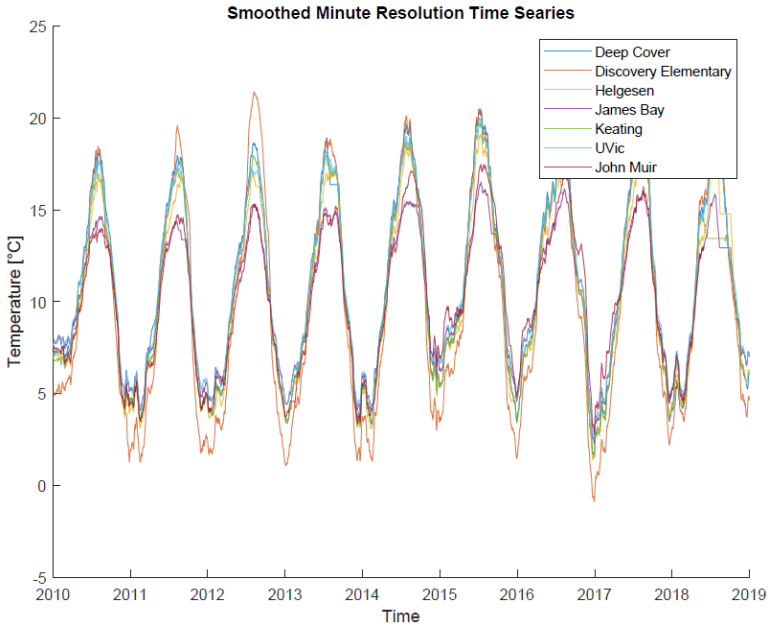

Figure 3. Heavily smoothed minute series. Smoothed using a moving mean window with size 50000.

interval. The linear fit is accomplished using a standard least squares fit. If the equation for the straight line is given by

$$y = a + bx$$

the confidence bound for the slope, δb, is given by

$$\delta b = \frac{\sigma_\varepsilon}{\sigma_x \sqrt{N-1}}$$

where σ_x is the standard deviation for the independent variable (time) and σ_ε is the error in the least squares model, given by

$$\sigma_\varepsilon = \frac{\Sigma(y_i - \widetilde{y})^2}{N-2}$$

where $\widetilde{y}$ is the value predicted by the linear fit. The confidence limits for a, represented by δa, are defined by requiring any linear fit to pass through the sample mean. The dotted line going through the middle of figure 6 represents the sample mean for all the plotted points. Because the limits of the confidence interval do not overlap with the sample mean, the average monthly temperature of July at the UVic weather station is increasing at the 95% confidence interval.

This process is repeated for every month at the UVic weather station, with the results in figure 7. Note that few months have the same conclusion are July. Only May, July, and August have been increasing in average temperatures over the last 10 years. For most months, the average temperature is not increasing at a 95% confidence level. Months other than May, July, and August have large confidence intervals due to the small amount of monthly averages and large variability. For example, September has almost zero slope on the line of best fit, and a confidence band that appears to cover the entirety of the y-axis. From this plot, the monthly mean for September is seemingly random and is not increasing nor decreasing. A much larger dataset is needed in order to draw a more concrete conclusion about the rising of average monthly temperatures.

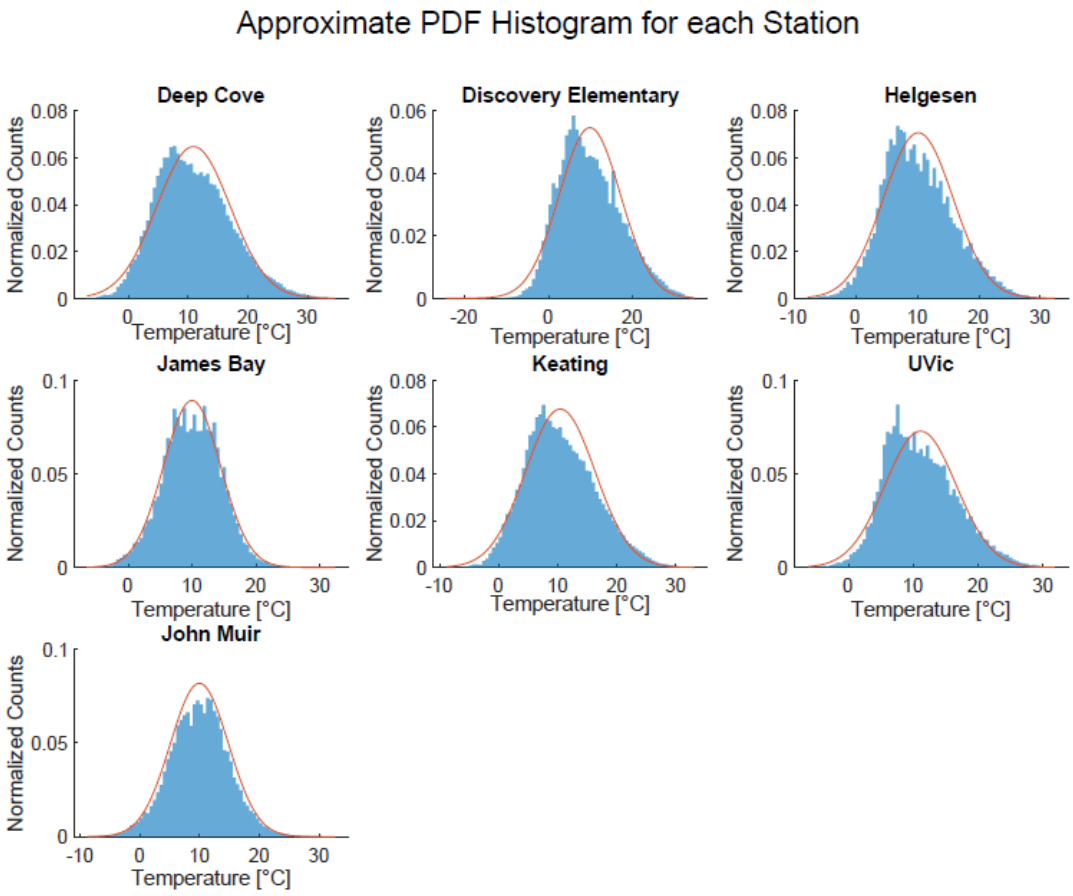

Figure 4. Histogram for each station, excluding NaN values. Overlayed is a normal distribution with centered at the sample mean and sigma of the sample standard deviation.

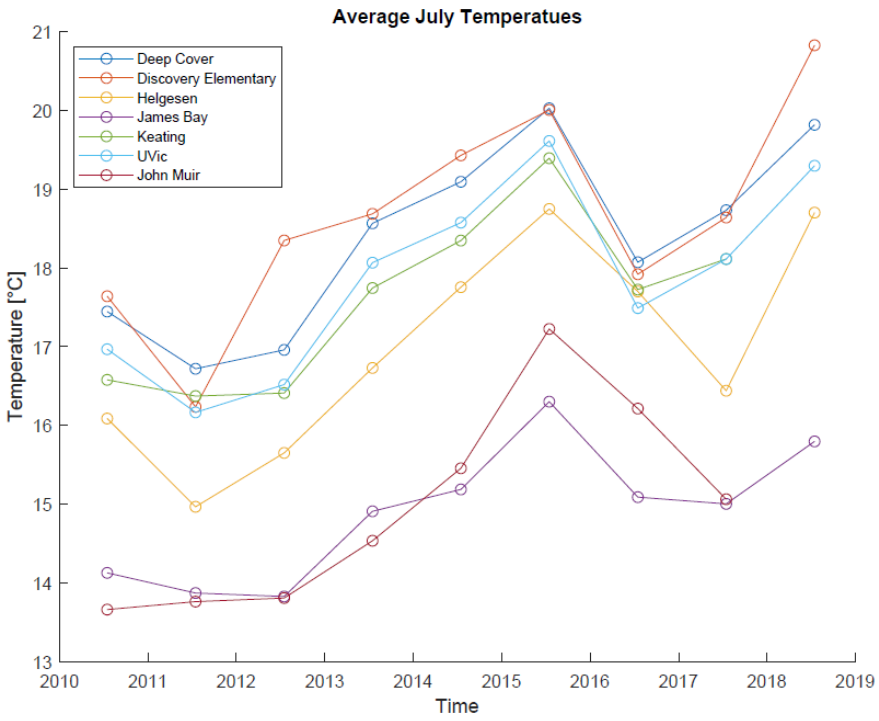

Figure 5. The average temperature for the month of July at each station over the course of the entire dataset.

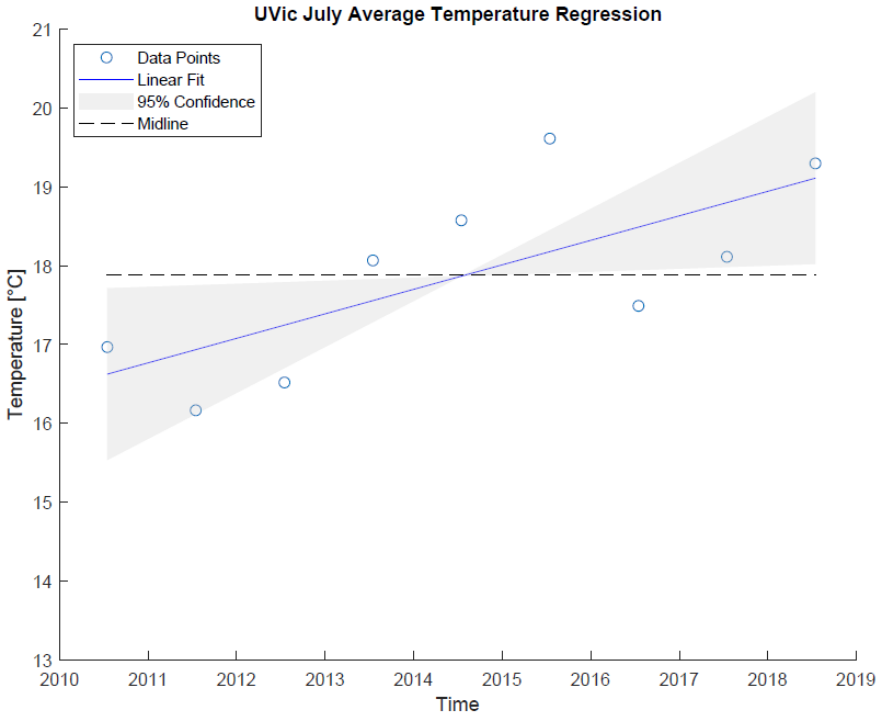

Figure 6. The average July temperature for the UVic weather station with a linear fit and 95% confidence interval.

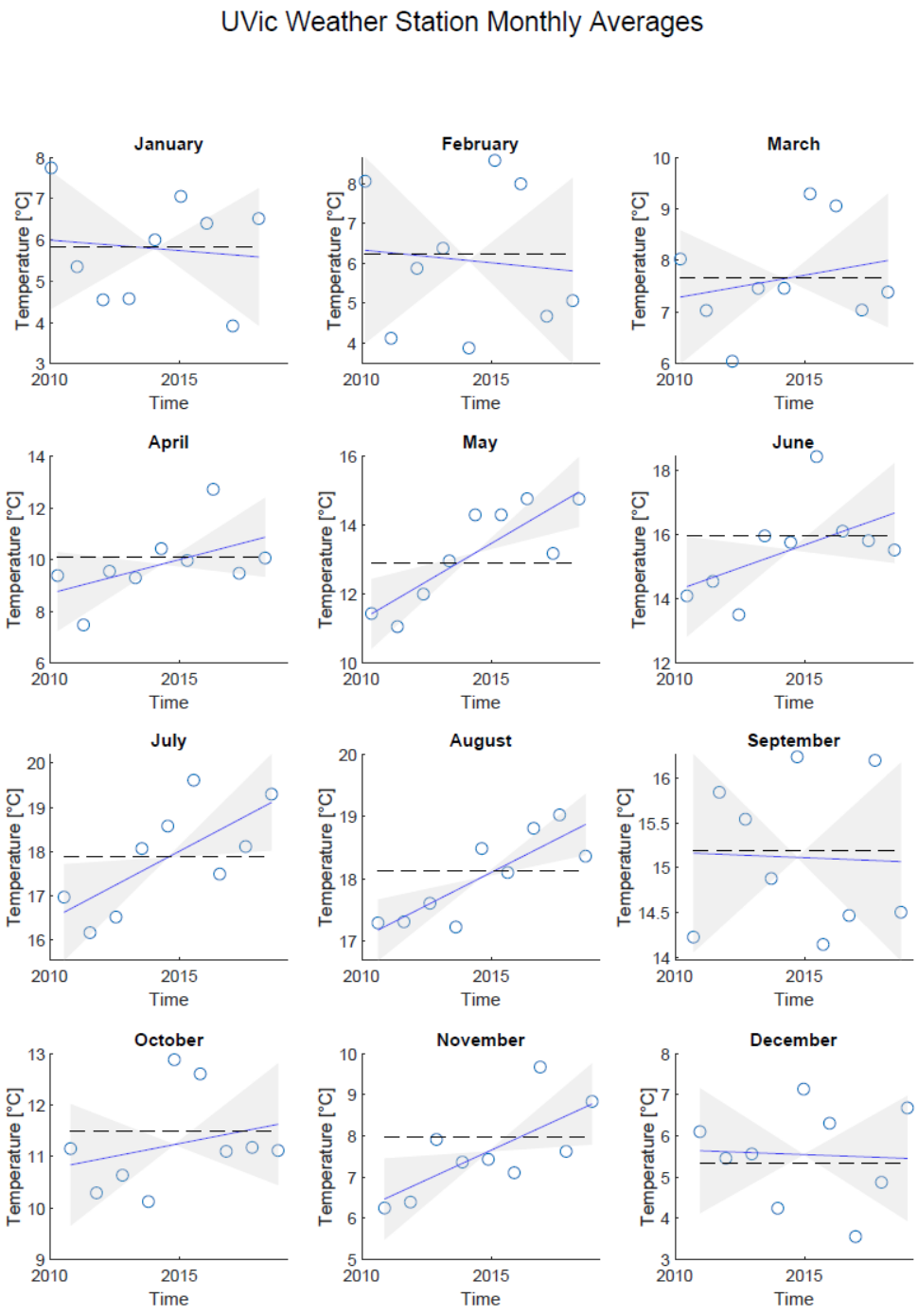

Figure 7. The average monthly temperature for the UVic weather station with a linear fit and 95% confidence interval.

2.3 Station Cross-Covariance

To compare station patterns against each other, the cross-covariance between two stations is calculated. The cross-covariance between two discrete time sequences measures the correlation between the time series when shifted (lagged) across each other at different intervals. The resulting plot shows the similarity between the two series as a function of displacement from one another. The covariance between John Muir and Helgesen is in figure 8 and the covariance for John Muir and Discovery Elementary is plotted in figure 9. In both of these figures the covariance during the winter is plotted in the first subplot, and the covariance for the summer is on the second. The max lag for the covariance function is 40 days, which is enough to allow the summer curve to drop below e^{-1}. The series are considered to not be significantly correlated when the covariance drops below e^{-1}, represented by the black dotted line on all the figures.

The most apparent difference between the summer and winter covariance plots is the magnitude of the waves. The large amplitude in the summer cycle is due to the day/night cycles having larger and more consistent variability in temperatures. The most important aspect when examining the covariance plots is not the signals amplitude, but when the signal completely falls below the dotted e^{-1} line. The covaraince signal during winter completely falls below e^{-1} after 10 days lag, whereas the summer covariance's peaks stay above e^{-1} until about 28 days. This suggests that the signals during the summer are much more correlated, potentially meaning the temperatures at John Muir, Helgesen, and Discovery Elementary, are more similar during the summer than during the winter. These station were chosen due to their proximity: John Muir and Helgesen are close, while Discovery Elementary is further up island. This is reflected in the amplitudes of the winter covariance plots, where John Muir and Helgesen in figure 8 has a smaller amplitude than John Muir and Discovery Elementary in figure 9. The smaller signal amplitude suggests a smaller variability, which is expected as close stations would have most likely be recording similar data.

2.4 Fourier Analysis

The power spectral densities for each station's time series is plotted in figures 10 and 11. Each station has the PSD plotted on the left, and its corresponding PSD in variance-preserving form on the right. The PSD is calculated using Welch's power spectral density estimate with a window size of 2^{15}, 2^{14} overlapping samples, 2^{15} frequencies, and a sampling rate of 1/60Hz. These values were chosen to get a smooth signal for the first set of peaks and lower noise at the higher frequencies. Each station has a very similar PSD, with a peak frequencies at one cycle per day (CPD), which represents each days cycle. Following peaks diminish at 2 CPD, 3 CPD, and so on until the higher frequencies become noise. The variance preserving plots convey a similar story, where majority of the power in the signal comes at one cycle per day. It is interesting to note that the highest and lowest peaks at 1CPD from the variance preserving PSDs correspond to the highest and lowest standard deviations from table **I**, given by Discovery Elementary and James Bay, respectively. This is perhaps counter-intuitive, as the station with a lower spread in temperatures might be considered to have the strongest cyclic properties.

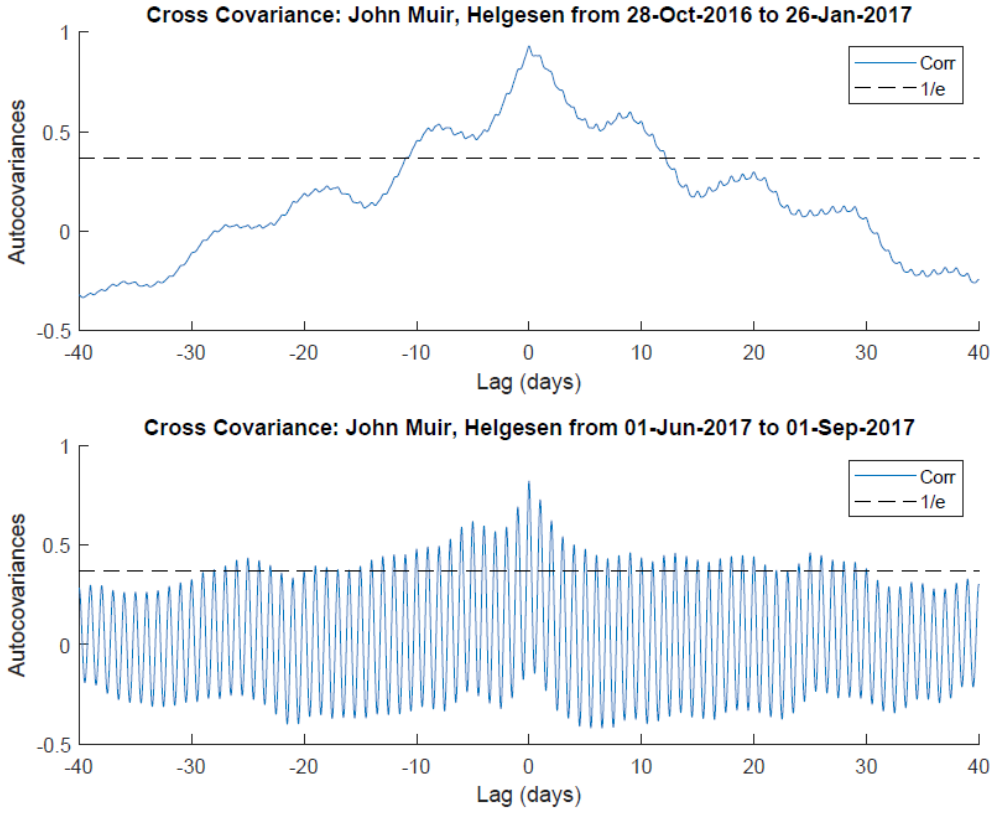

Figure 8. The cross-covariance between John Muir and Helgesen during the winter (top) and summer (bottom).

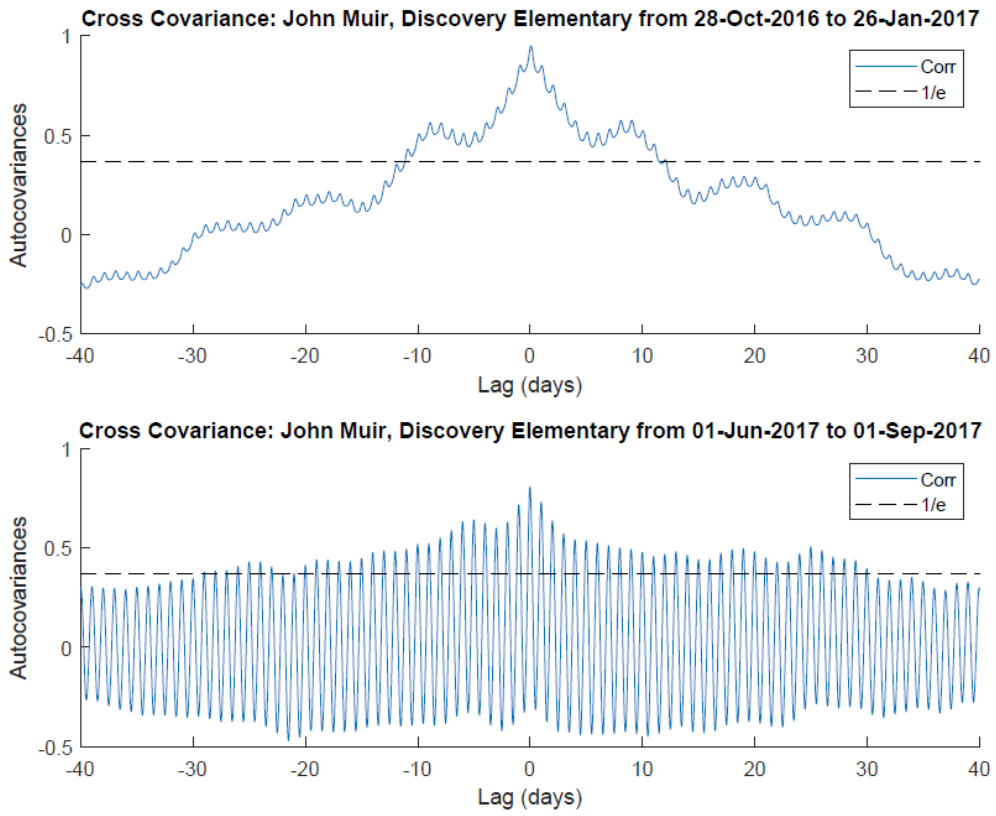

Figure 9. The cross-covariance between John Muir and Discovery Elementary during the winter (top) and summer (bottom).

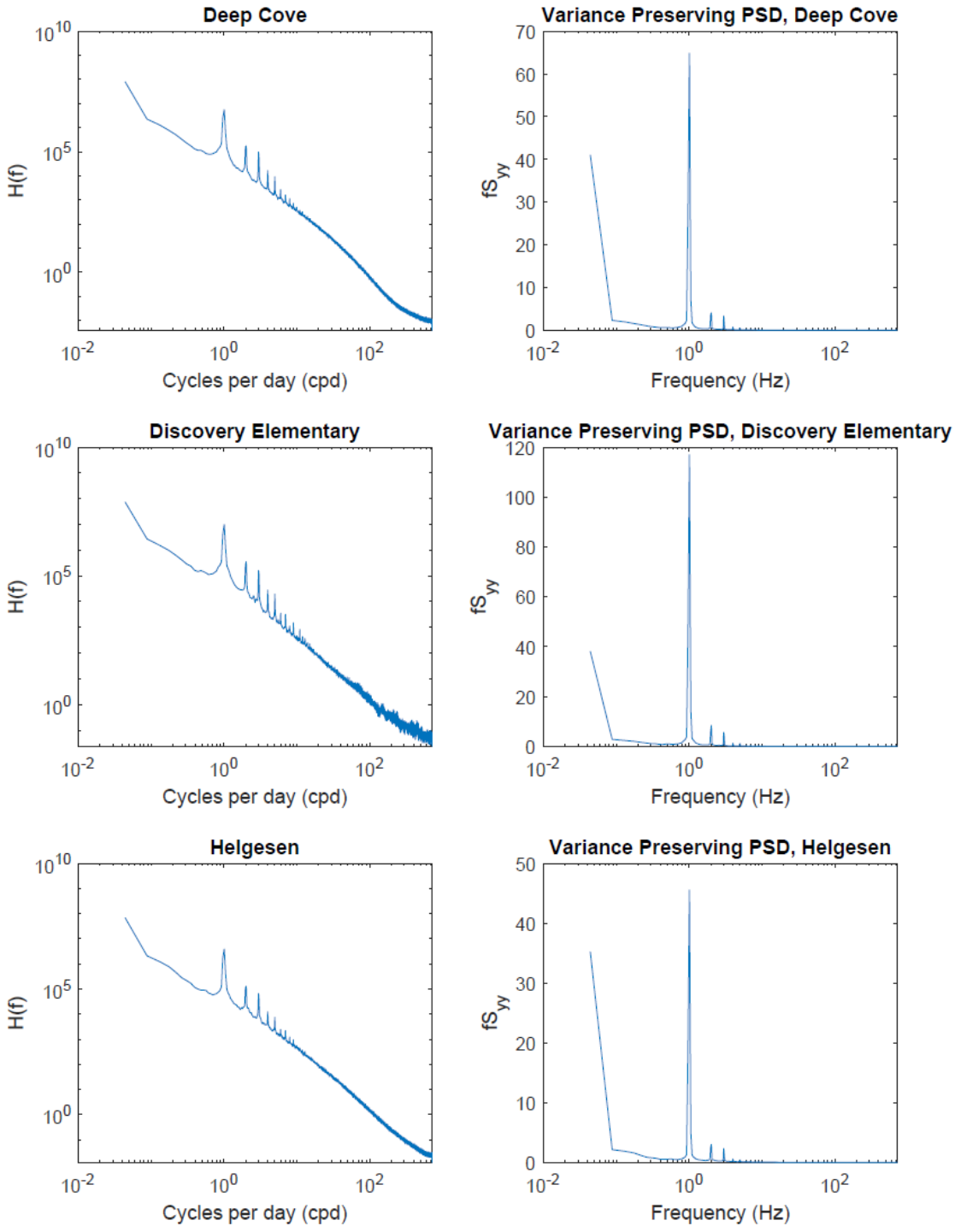

Figure 10. Power spectral densities in normal and variance preserving forms.

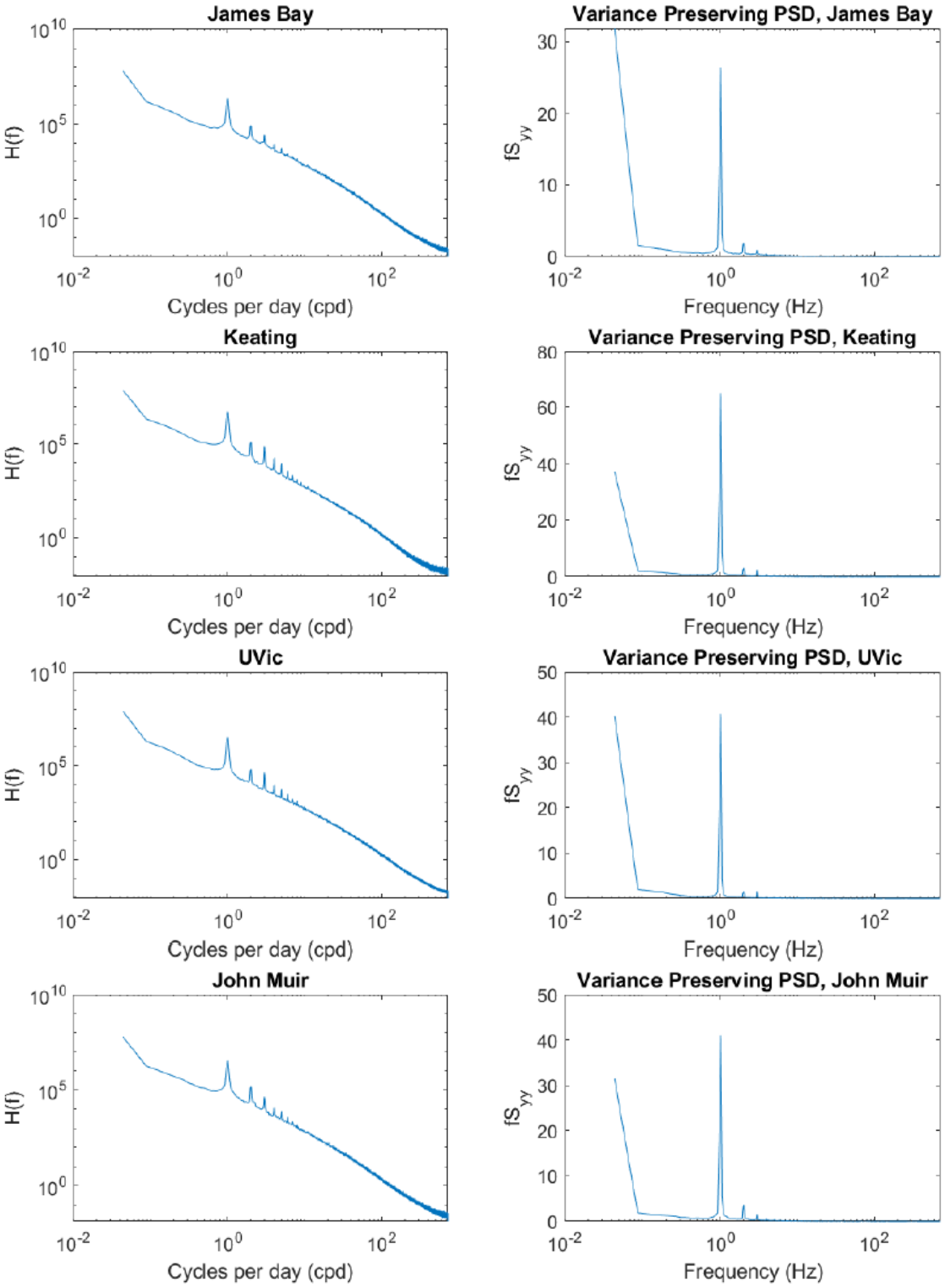

Figure 11. Power spectral densities in normal and variance preserving forms.

3 Hour Data

This section will focus entirely on the hourly dataset, mainly in regards to 2-dimensional plotting and interpolation techniques. With 39 stations in total, they will not be labelled with names. Instead, figure 12 displays the location of each station on the map, indicated by the red dots. Most stations are scattered within Victoria and the metropolitan, but some stations spread out up island and towards the west. Following the same technique for displaying the entire minute dataset, figure 13 shows the data from all 39 stations from 2008 to the end of 2019. Some stations have significantly longer periods of null-recordings, most prominently are stations 20, 29, 33, 37, and 38. These stations are dealt with in future subsections. The yearly cycle between seasons that was seen in figure 1 is also present in figure 13, with a similar range of temperatures from -35^oC to $+35^oC$.

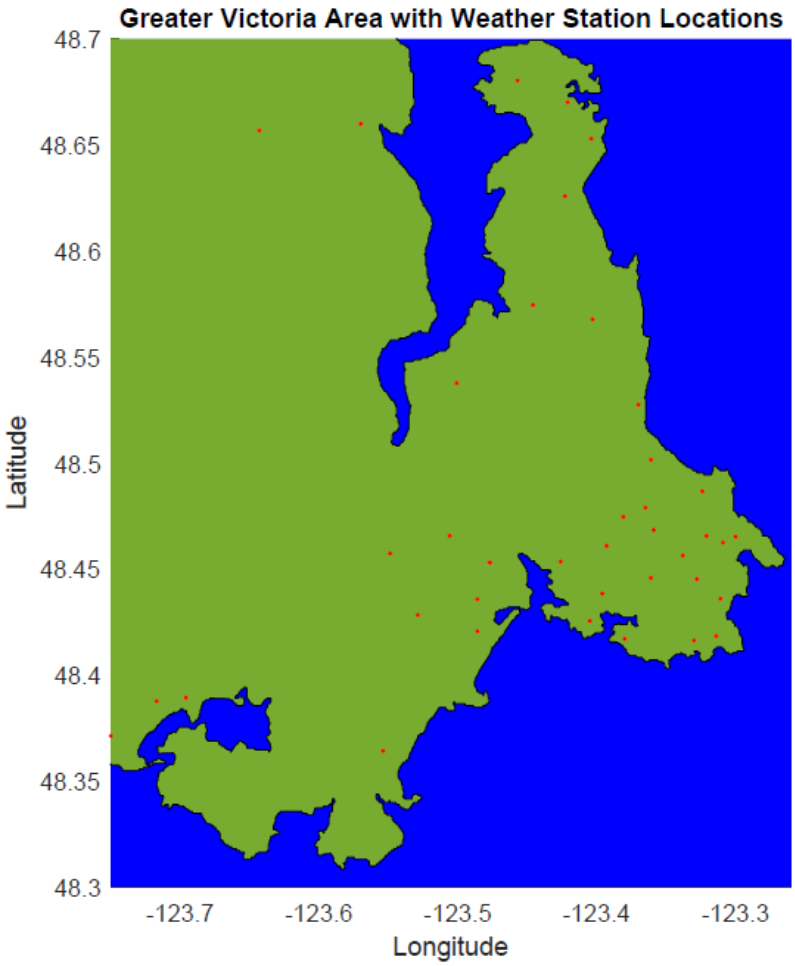

Figure 12. Map of lower Vancouver Island displaying all the weather stations that record the temperature once per hour.

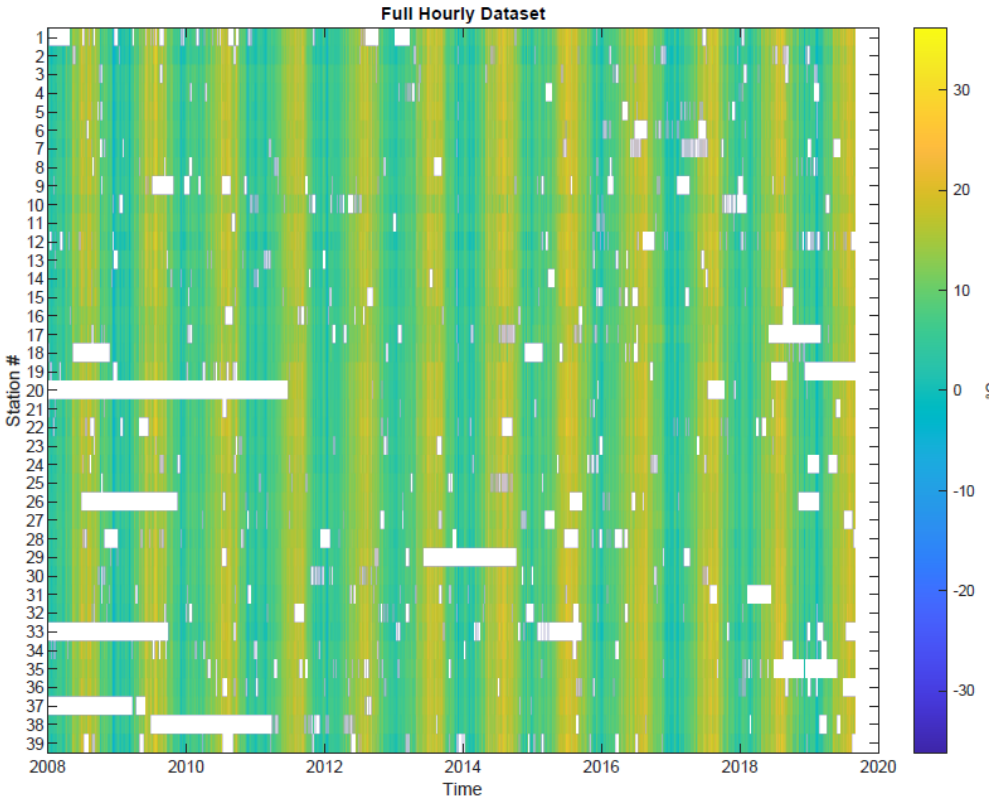

Figure 13. Full hourly dataset showing all temperature recordings from 39 different weather stations.

3.1 2D Interpolation

As stations only take discrete temperature measurements at a single location in space, the entire island will not be completely mapped out for temperatures. For the average person, they are less concerned with the exact temperature at an arbitrary station, and more-so with the general temperature of the island. This is where 2D interpolation comes into play. Like the contour plot in [1], the temperatures from all the stations are interpolated to create a general idea of what the temperature is like anywhere on the island. The four types of interpolation explored in the following section are:

- Cubic interpolation
- Nearest Neighbor interpolation
- Inverse square distance gridding
- Exponential gridding

Cubic interpolation is implemented from MATLAB's *griddata* builtin function, which describes the function as "triangulation-based cubic interpolation supporting 2-D interpolation only" [2]. Nearest neighbor interpolation simply sets each point to the same temperature as the nearest station. This gives rise to a Voronoi diagram. Nearest neighbor does not create a good approximation for the temperatures across the island, but generates from interesting patterns. Inverse square distance and exponential gridding both generate weights defined by a weighting function. For each point

being interpolated, the distance to every station is calculated. This distance is plugged into a function to generate a weight. Each station has an associated weight, which is then multiplied by its temperature. The interpolated temperature will be the addition of all the weighted values from each station. Mathematically,

$$z_i = \sum_n w_n z_n$$

where z_i is the new temperature at a point (x_i, y_i), w_n is the weight at station n with a temperature z_n, and the weights are normalized

$$\sum_n w_n = 1$$

The inverse square weights are given by

$$w_n = \frac{1}{r_n^2}$$

and the exponential weights by

$$w_n = e^{-r_n^2/2\sigma^2}$$

where σ is a smoothing constant. The distance in both equations is given by

$$r_n^2 = (x_n - x_i)^2 + (y_n - y_i)^2$$

where x_n, y_n is the location of the point being interpolated and x_i, y_i are the coordinates of each station. Figures 14, 15, 16, 17 show the four different gridding techniques at different times of the year.

MATLAB's griddata function only interpolated within the convex hull, not extrapolating any temperatures outside. The cubic interpolation sometimes had jagged temperature edges, somewhat observable in figure 16. The nearest neighbor interpolation has a Voronoi diagram overlaid, however the Voronoi diagram was not used in shading the cells. This shows that a nearest neighbor interpolation will result in the same plot as a Voronoi diagram. However, the Voronoi diagram is more difficult to shade in outer cells as their lines are unbounded, creating no clear shading area. The inverse square and exponential gridding both have much more convincing results than the first two. Inverse square interpolation stays continues over the grid, and each station has a heavy influence on areas it closely surrounds. The exponential interpolation is extremely smooth across the land, not allowing for single stations to have a large weight on nearby areas. Our of the four times chosen, the highest temperatures occurred during September in figure 17, and the lowest temperatures in February, figure 15.

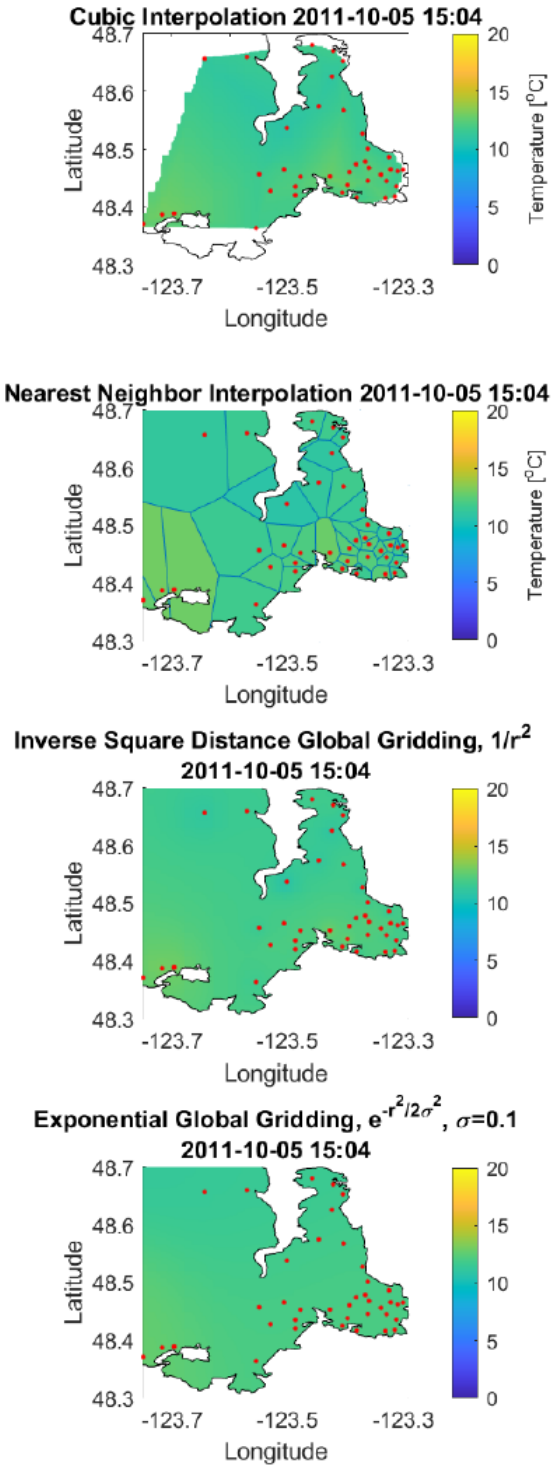

Figure 14. Winter gridded data.

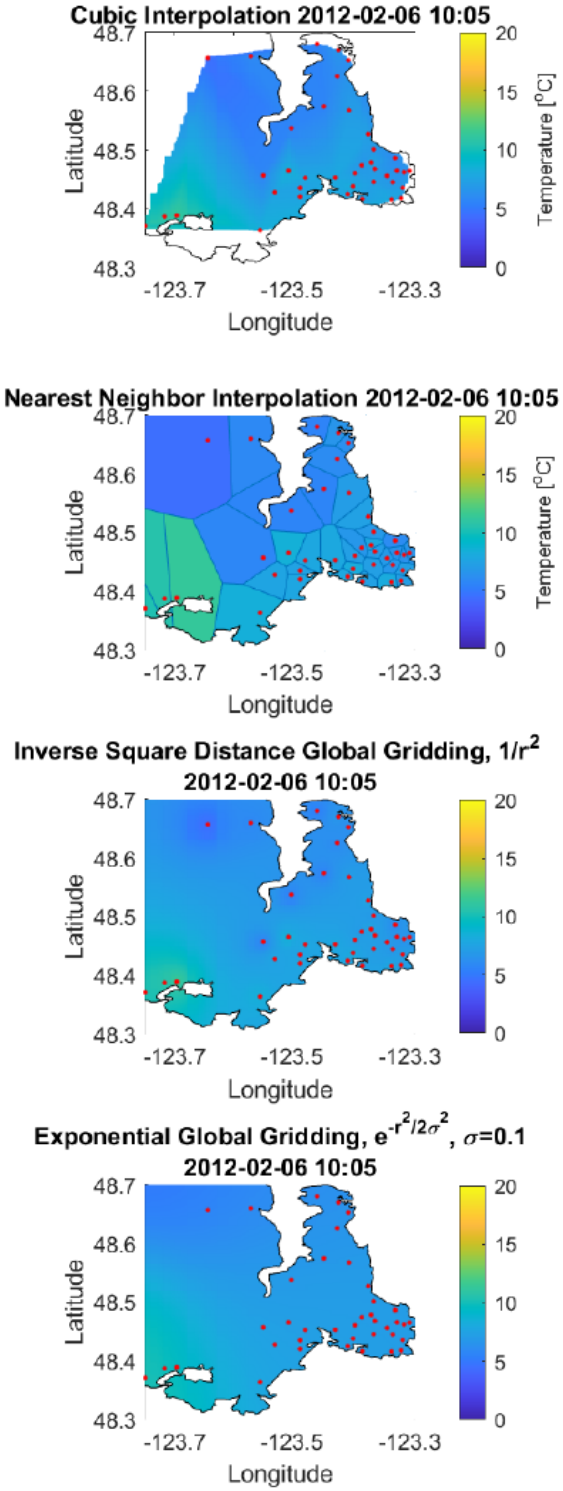

Figure 15. Spring gridded data.

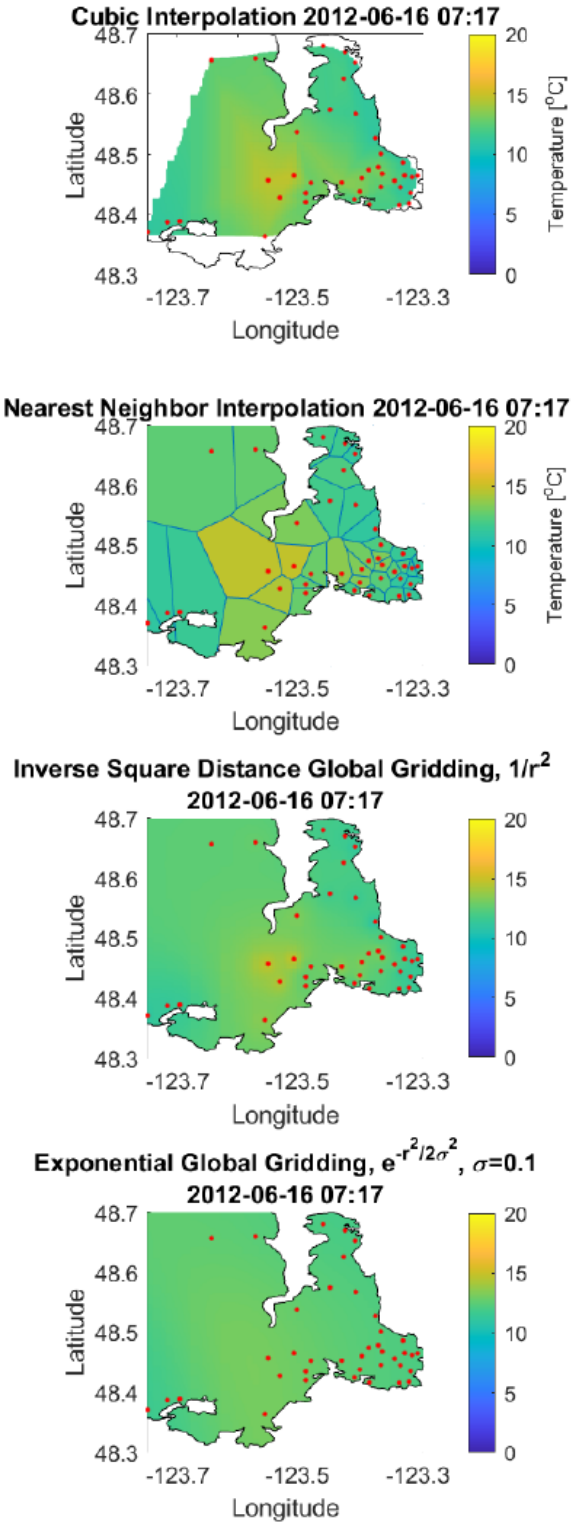

Figure 16. Summer gridded data.

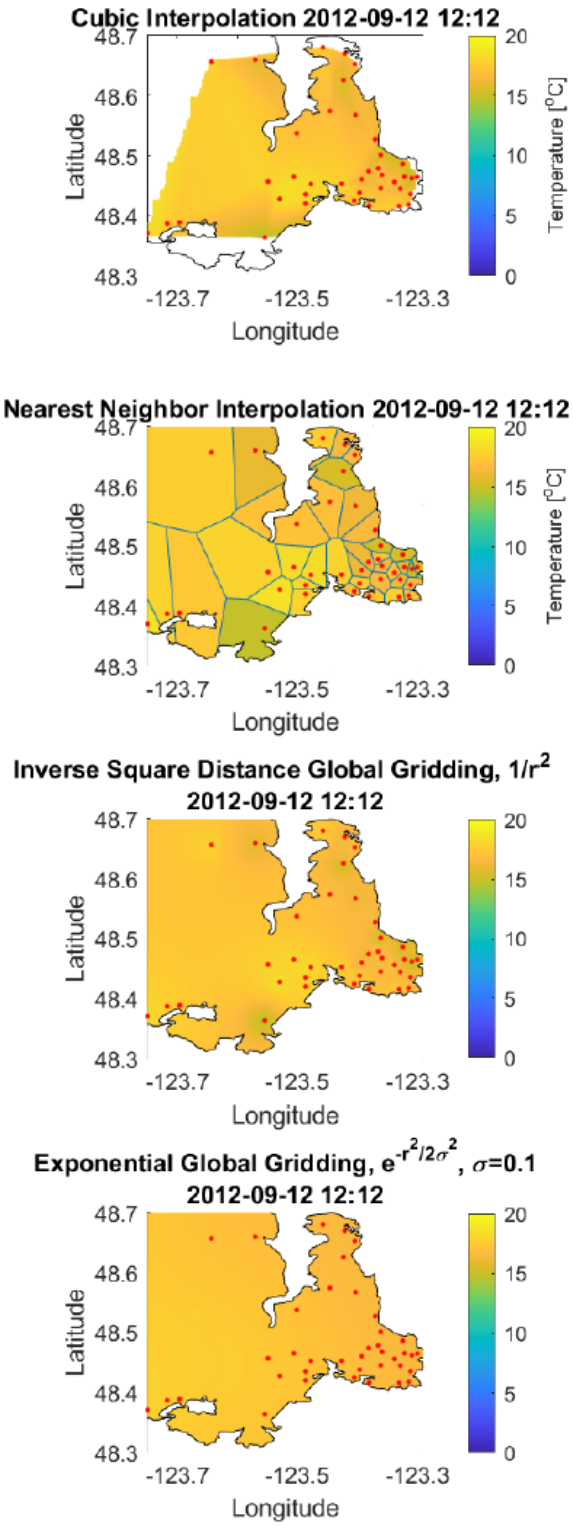

Figure 17. Fall gridded data.

3.2 EOF Analysis

With the purpose of the 2D interpolation leading to empirical orthogonal function analysis, only the inverse square interpolation is carried forward. Inverse square is chosen due to its continuity and stations' recordings having more influence in their local surroundings, generating an accurate weather map. The EOFs are interpolated using a 500x500 grid. This grid is much smaller than previous due to the large amounts of memory required. Keeping the original resolution resulted in system memory overflow. The reduced grid allows for much quicker calculations and the resolution is still more than acceptable for analysis. The 500x500 spatial grid containing the interpolated temperature information was stacked 250 times, creating a 500x500x250 matrix, where the third dimension represents the temporal domain. The temporal domains varied by one hour. The starting point for the first matrix is on October 1st, 2017. The following 250 hours brings the final matrix to about October 12th, 2017.

The three most influential modes are plotted in figures 18, 19, and 20. Each figure has the percent of variance explained by each mode included in the title. Figure 18 contains 99.4% of the variability of the dataset. The strongest correlations are all inside the main metropolitan area of Greater Victoria. Modes two and three contain significantly lower importance, with variance percentages of 0.4% and 0.2%, respectively. The main area of variability in mode 2 appears up island, along the ferry terminal and airport. Mode 3's variability exists everywhere else, towards the middle of the island. The principal components are plotted in figure 21. As expected, mode 1's principle component amplitude is much higher than modes 2 and 3, which are close together, showing that mode 1 contributed most significantly to the overall temperature variations.

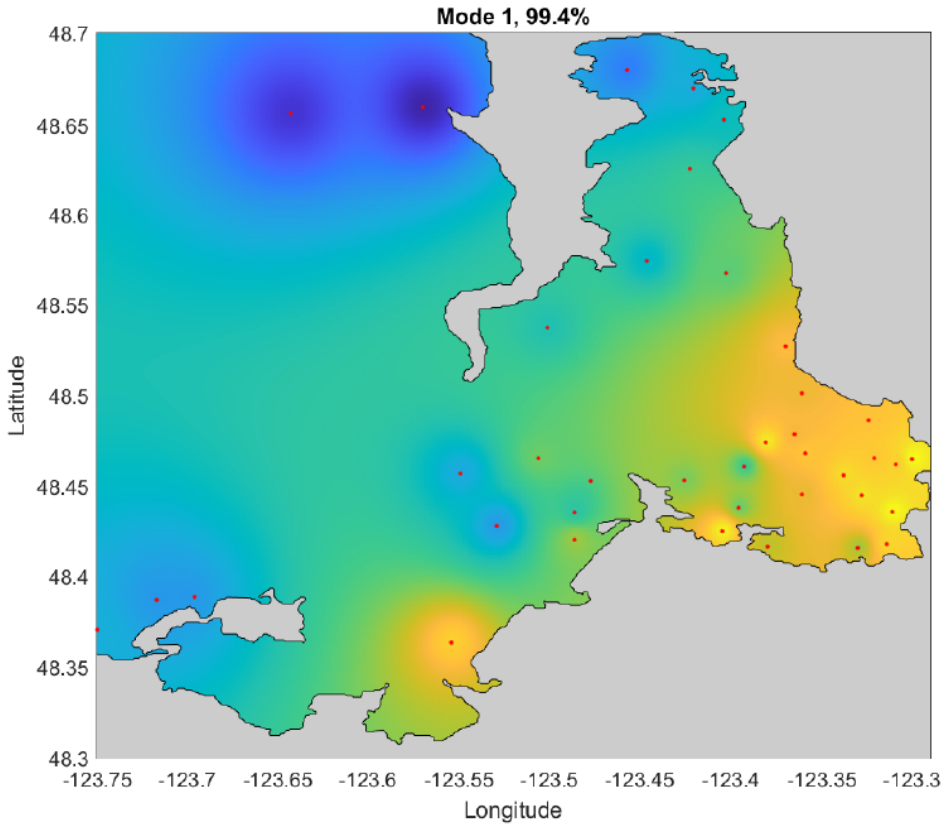

Figure 18. The most important mode from the EOF analysis. Mode 1 accounts for a high majority of all variability.

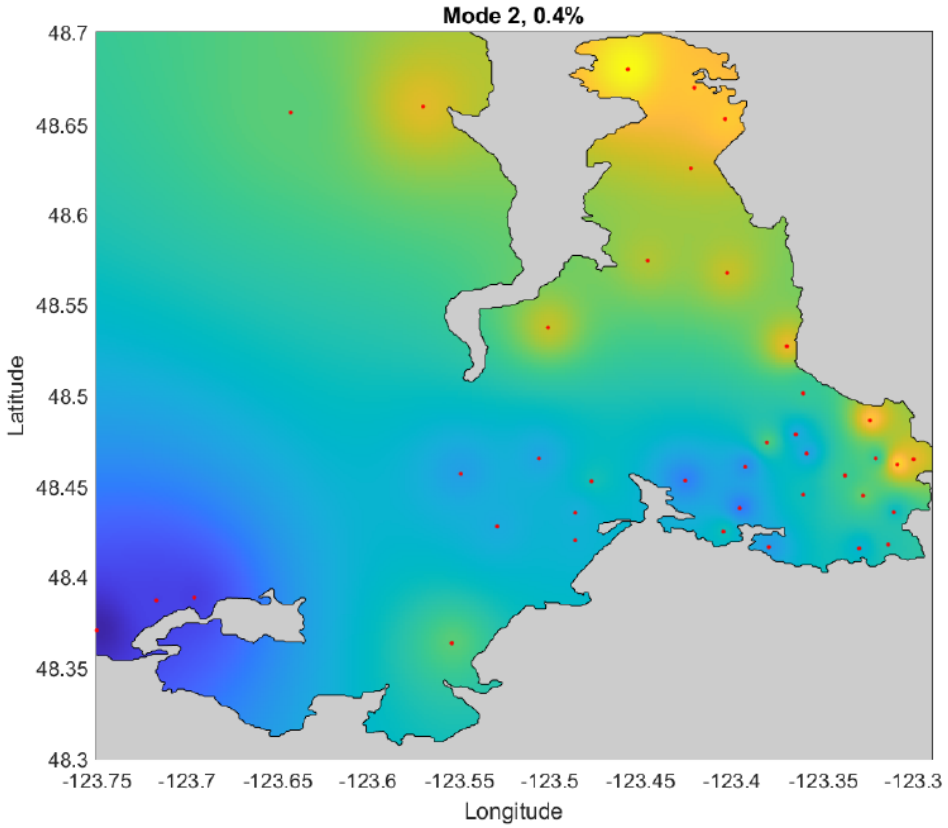

Figure 19. Second most important mode.

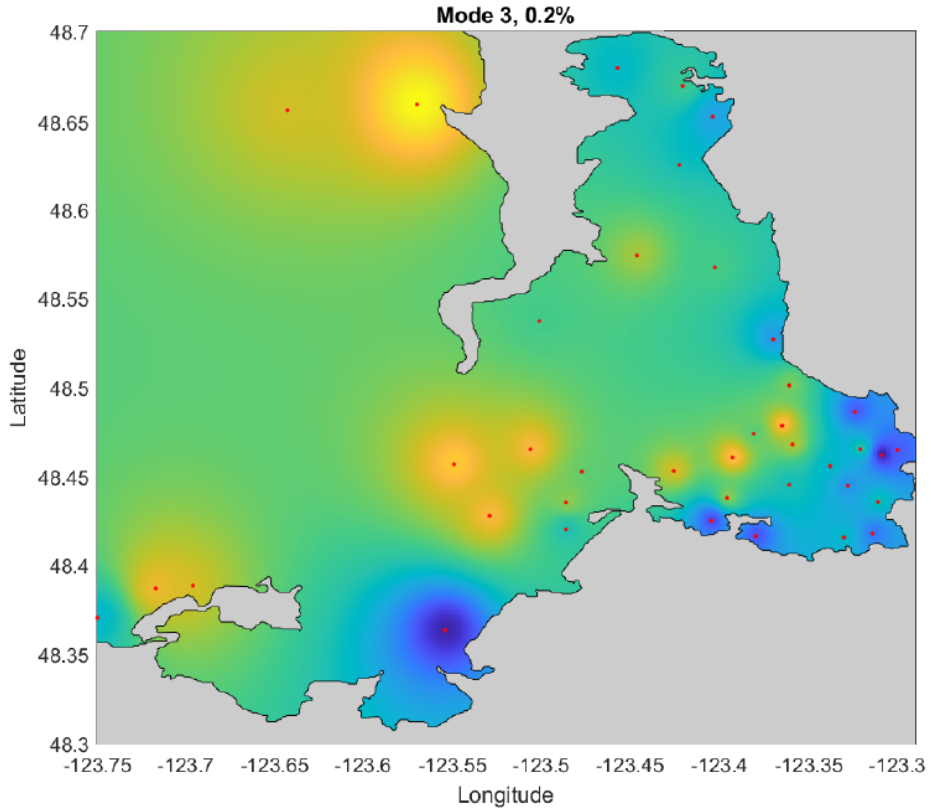

Figure 20. Third most important mode.

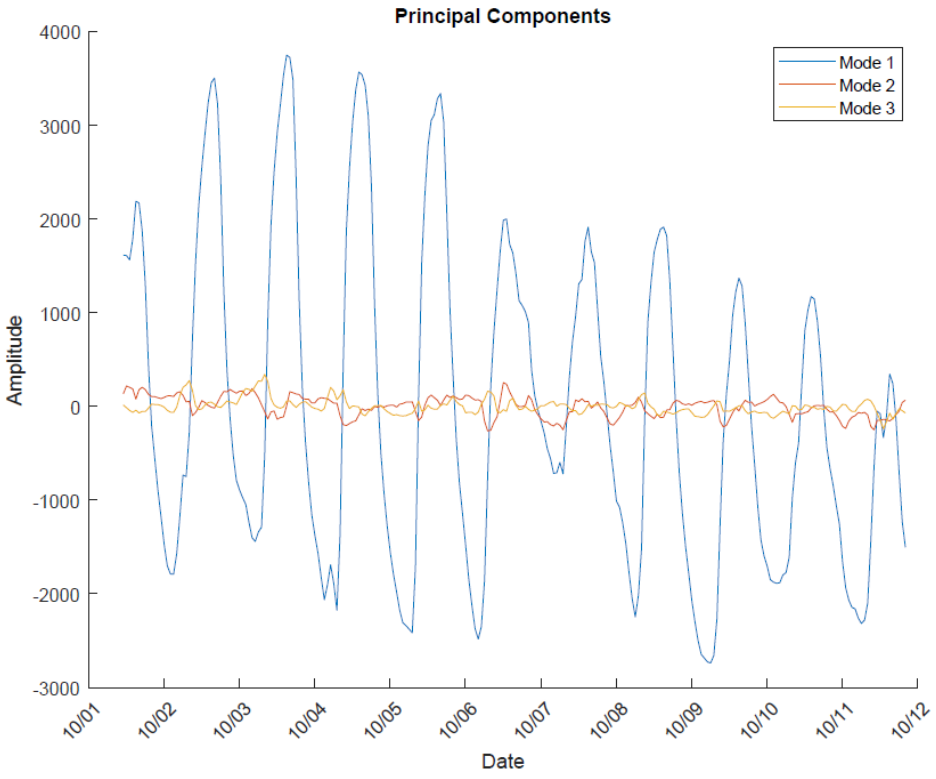

Figure 21. Principal comonents from the three most important modes..

References

[1] E. Wiebe, "School-Based Weather Station Network," *School-Based Weather Station Network - Victoria, British Columbia.* [Online]. Available: http://www.victoriaweather.ca/. [Accessed: 4-Dec-2019].

[2] "graddata" Interpolate 2-D or 3-D scattered data - MATLAB. [Online]. Available: https://www.mathworks.com/help/matlab/ref/griddata.html. [Accessed: 13-Dec-2019].