Don't Delete Carriage Returns Above. The Title of Your Paper Goes Here Using the Paper Title Style

**Team 08**

**Dmitrii Kharkovskii /////**matric///

**Naheed Anjum Arafat /////**matric///

**Liu Dake /////**matric///

**Wu Lifu /////**matric///

**Daren Tan Zong Loong** A0111855J

Affiliations and Addresses Go Here Using the Affiliation and Address Style  
publications15@aaai.org (to delete?)

Abstract

This paper presents a real world application which relies on the Gaussian process (GP) models for probabilistic non-linear regression and prediction. Our motivational example makes use of a 1 degree by 1 degree HadSST2 dataset to predict the sea surface temperature (SST) anomaly for any valid coordinate on the map which can then be interpreted to other valuable information such as weather predictions. By providing SST anomaly values for specific coordinates, it should allow experts to plan and make better decisions based on a distribution of SST anomaly values.

1. Introduction

Sea surface temperature (SST) anomaly is a crucial oceanographic variable that are provided by continuous observations from ships, vessels and buoys. These SST anomalies are good indications of El Niño and La Niña climate cycle. For more specific locations, it can influence weather patterns across regions.

At the equator, one degree of longitude and latitude both cover about 111 kilometers, or just about 70 miles. By being able to pin-point a coordinate and produce a good measure of SST anomaly at the specific coordinate, it gives a better representation of the anomaly situation within the one degree grid. However, deploying buoys and vessels in order to cover areas within a 1 degree grid is inefficient.  
  
This brings us to our motivating application, which aims to use the Gaussian Process (GP) to provide a good distribution of SST anomaly values within the one degree grid for water-vehicles to be able to know where to move within the one degree grids.

Prediction of SST Anomalies

Our problem can be modeled by a Gaussian Process, the input is the locations (including the locations which have measurement in the HadSST2 dataset and the locations whose SST anomalies we are going to predict) and the output is the SST anomalies.

From the HadSST2 dataset, we have observation of some output, with which we can use the Bayesian model to predict the mean and uncertainty of the SST anomalies (variance) of the locations that are directly not observed. The non-parametric nature of GP model also gives us enough flexibility not to worry about whether our model can fit the data since we are giving prediction as a distribution over all such models (parameters) rather than just one model parameter.

Qualitative Advantages of GP Models

Simple Linear Regression (available in Excel) is not appropriate for modeling relationships between latitudes, longitudes and sea-surface temperature as the values of SST anomalies depends on more than one variable (requires both latitude and longitude). It is easy to comprehend from the following plots that the relationship between latitude, longitude and SST anomaly values is not linear.

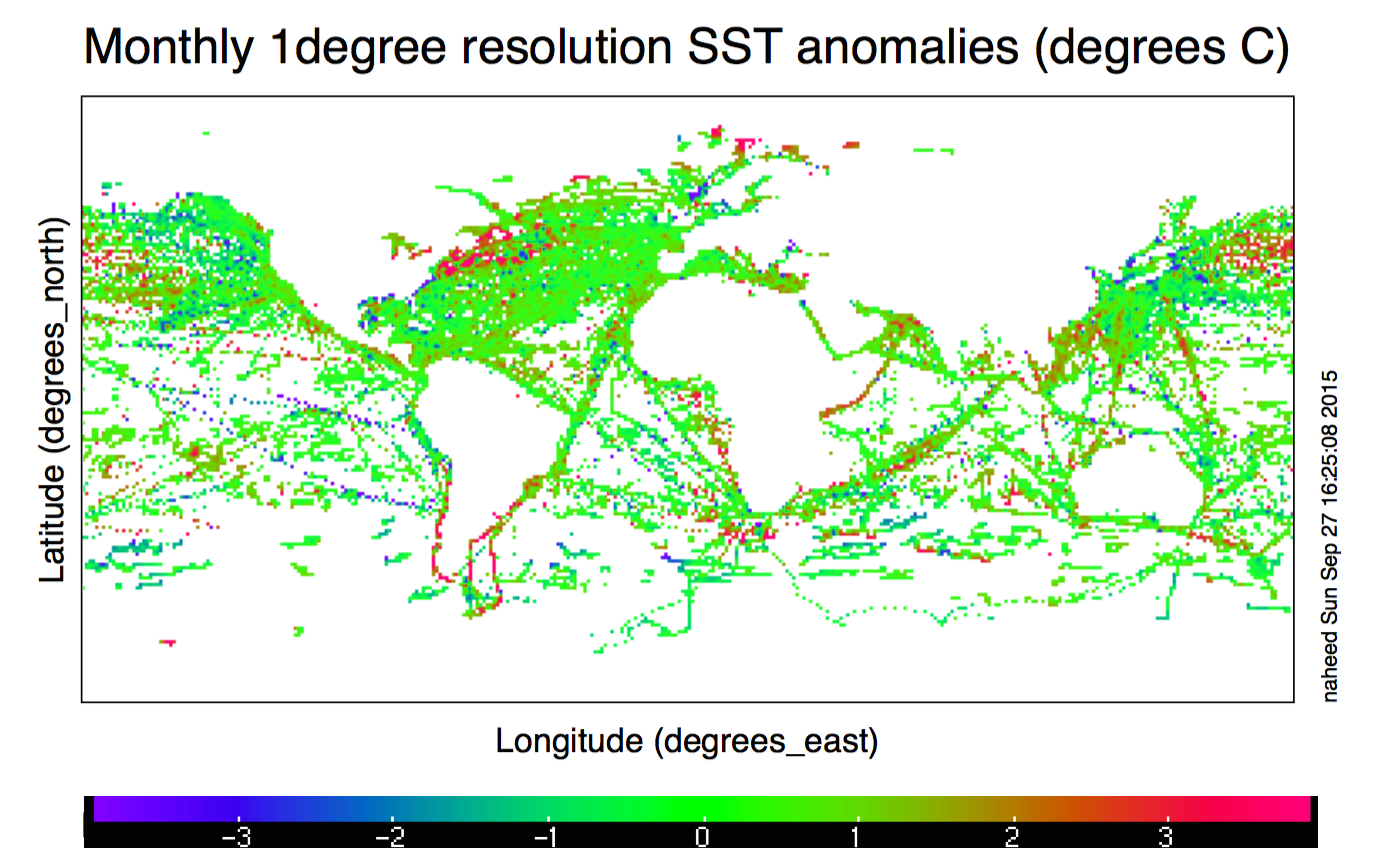


Figure 1. Monthly 1degree resolution of SST anomalies (degrees C)

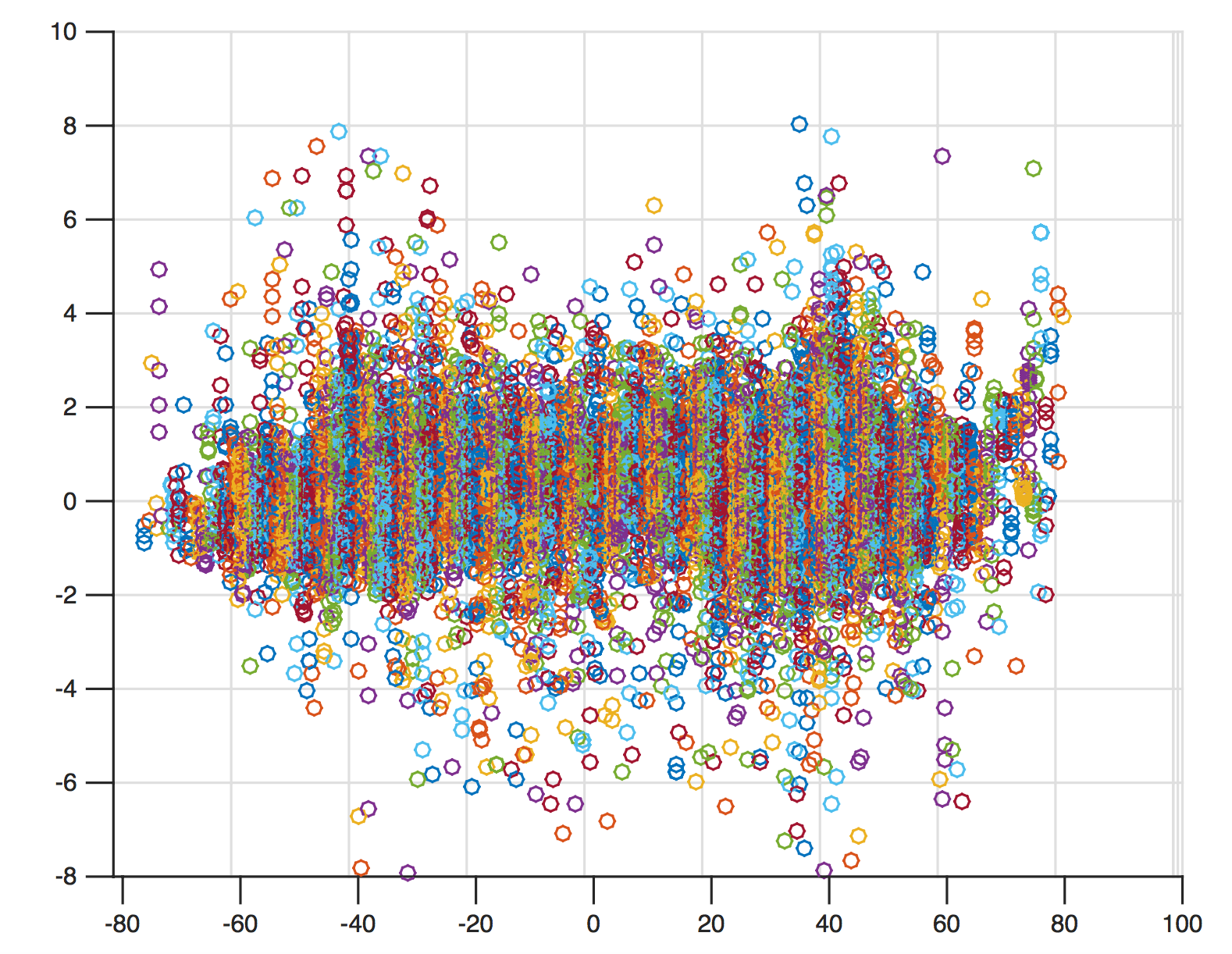


Figure 2. Plot of SST values across latitudes

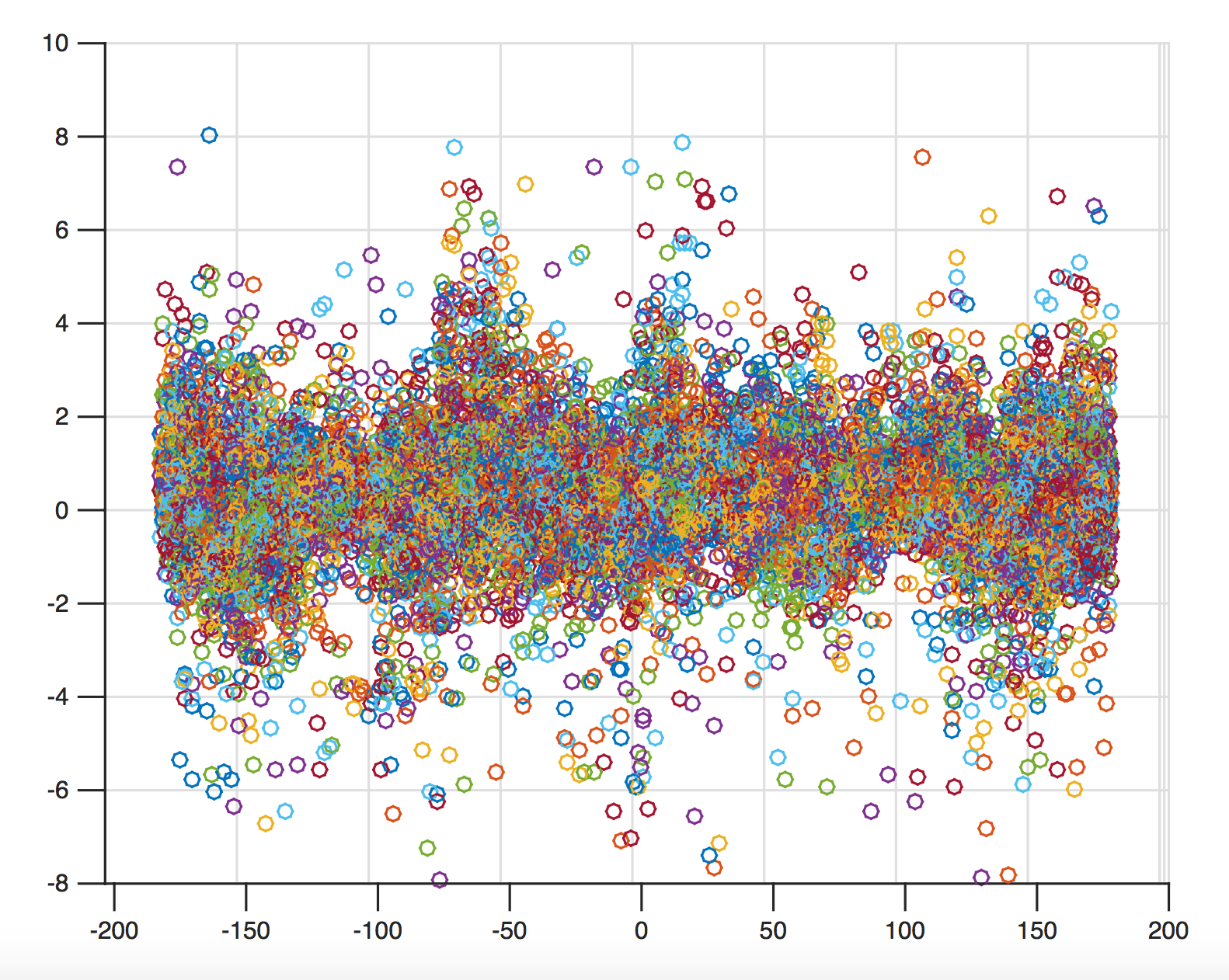


Figure 3. Plot of SST values across longitudes

From those plots, we observe that most of the regions have mean SST anomaly value (MSstA) of 0. However adjacency of blue, green and red regions (Figure 1) signifies that the relation is not linear. As a matter of fact, It is difficult to make any kind of assumptions about the relation between SST and the model parameters (i.e. whether it is linear or non-linear) as well as between the SST and the coordinates. Since the SST anomaly values varies across the coordinates in an unpredictable manner which is expected from any natural phenomena. Therefore, a non-parametric model would be a better choice for this particular application.

However, the SST anomaly values are spatially correlated, as observed from large spatially adjacent region of green dots and trails of red and blue dots (Figure 1).  Gaussian Process (GP) is particularly suitable for this kind of non-parametric regression modeling where spatial correlations among the independent variable (regressor) are known. Since we are interested to predict the MSstA for any region with known predictive uncertainty (derived from GP regression), which later can be used for planning a navigation route for a water vehicle.

**Data Requirements**

For our application, we will be relying on the 1 degree version of the HadSST2 dataset. This dataset has been refined in a way that it is time homogenous and has already passed the basic ICOADS quality checks that filter out invalid land locations or duplicates. The HadSST2 is a large data set, (approximately 65,000 data points) and is used to map a global field of SST anomalies monthly.

Our application also requires a recent dataset (i.e. using specific dataset from the same year/month) with uniformly distributed data points in order to project the SST anomalies distribution within each 1 degree grid.

**Interpretation of Data & Decision Making**

The output of our GP would be a more continuous and complete map for the ocean area. With the output of the GP, human experts would be able to see the gradient trend of ocean temperature anomaly, and gain information on how other co-related climate phenomenon (e.g. direction of ocean current, sunshine intensity at different latitude) reflected/affects ocean temperature anomaly.

Moreover, with the whole timeline of series change of data, users would be able to predict the future appearance of ocean temperature anomaly in anyplace.

**2. Technical Approach**

GP models are very popular in environmental sensing and spatial analysis because they can take into account the spatial correlation between the points of the input data which exactly fit our real world application - we assume that since the sea temperature changes continuously, the values at points, which are close to each other, should be more correlated. So using GP model will allow us to use the advantage of spatial correlation.

Another important aspect of the GP model is that it allows us to deal with noisy observations, which might also be useful. Since we do not know if the process of obtaining the HadSST2 dataset filters out noisy measurements, we should assume that the measurements are noisy.

**Additional Insights**

Since there are missing values in the dataset along with some unnecessary land co-ordinates which do not constitute our dataset, we had to trim those values which resulted in approximately 15,000 data points before running a Full-rank GP on this dataset.

**Ideas for Performance Improvements**

One proposed method of improving time efficiency is to divide the full datasets into smaller subsets that are justifiably uncorrelated (i.e. a point in pacific ocean is by no means correlated with a point in Atlantic ocean) and use a full-rank GP on each subset to get the covariance function parameters.

However, the parameters derived this way might be sub-optimal. Therefore, optimal policy for segmenting the regions into sub regions would be very crucial for the accuracy of the prediction.

For prediction, when a user requests prediction for a certain point, we will calculate which subset it belongs to and use only that covariance function to predict the value of the point. This proposed modification should not affect the results substantially given the regions are segmented optimally and provide an improved time efficiency.

**3. Experimental Evaluation**

In this section, we will demonstrate the various GP models that have been implemented to evaluate its performance in terms of accuracy, time efficiency and how each GP model performs with respect to our proposed application.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset: | Oct, 1854 | Aug, 1991 | Feb, 1999 |
| Training points: | 1854 | 12386 | 9667 |
| Testing points: | 926 | 6193 | 4833 |
| Time elapsed: | 08s39 | 55s79 | 44s76 |
| Error: | 1.222038 | 0.893325 | 1.130588 |

As mentioned above, we will be using the monthly data provided in the 1 degree version of the Hadley Centre SST data set (HadSST2)for both training and testing purposes.

**Experimental Setup**

We selected 5 dataset from HadSST2, namely October 1854, August 1991, February 1999, April 2007, and December 2012. These 5 datasets were handled with two different GP models, and one naive brute force method for comparison.

In the first section, SparseGP model was tested. 2/3 of data points in each datasets were picked out randomly and used for training. After training, we use the result to predict the rest of 1/3 data points, and compare with the ground truth values to calculate the average squared error. All five datasets were included in this section.

In the second section, we trained a full GP on each datapoint surrounded by 8 points (i.e. not NaN), then use hybrid monte carlo method to optimize parameters, and finally use the result to predict those data points surrounded by 8 other points which were not included in training, then compare to find squared error. However due to the long time of computation, we have only conducted one experiment on dataset December 2012.

The third method acts like a comparison group, where the prediction is just the average value of surrounding 8 points, and then compare with the ground truth value, to get the squared error value. Notice that in the second and third round of experiment, only those data points fully surrounded by 8 other data points were included in training, therefore could not be fairly compared with section 1, the sparseGP.

All three methods were trained and tested on same five datasets (exclude the second section), and the results are shown below.

**Model Comparisons & Analysis**

The result of SparseGP in section 1 is shown in the following table:

In the second section, there are only 3102 blocks surrounded by 8 blocks. Over these blocks, the fullGP gives a prediction error of 1.153160. The whole process spent 5-6 hours on only one data set (Dec 2012).

Finally, the brute force one gives the following result:

All the test cases in brute force method finished in a very short time.

The results above shows that, 4 out of 5 datasets predicted by brute force gives better result than SparseGP. However the limitation of brute force method is that it can only predict the point which is fully surrounded by 8 blocks. Yet, the error of directly calculating average value of adjacent blocks gives a considerable better result compared with SparseGP, except in the dataset of Oct, 1854, where only 41 points were predicted in the brute force method.

Also, from the experiment we can clearly see how Sparse GP significantly faster than full GP, where full GP computes for hours just over each points with adjacent points, without much improvement on the squared error.

**4. Conclusion**

The experiment carried out above was actually not what we expected, as the brute force method gives much better result than GP in many cases. This may caused by the randomness of the distribution of global sea surface temperature differences, which may affect GP in a misleading way. Further study need to be done in this direction.

|  |  |  |
| --- | --- | --- |
| Dataset: | Apr, 2007 | Dec, 2012 |
| Training points: | 10363 | 9909 |
| Testing points: | 5181 | 4954 |
| Time elapsed: | 48s33 | 46s84 |
| Error: | 0.886349 | 1.532515 |

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset: | Oct, 1854 | Aug, 1991 | Feb, 1999 |
| Included points: | 41 | 6642 | 3398 |
| Error: | 2.584596 | 0.492875 | 0.708249 |

|  |  |  |
| --- | --- | --- |
| Dataset: | Apr, 2007 | Dec, 2012 |
| Included points: | 3225 | 3102 |
| Error: | 0.577646 | 1.036548 |

(Can anyone add something here…I dont know what to say)

References

Lisa.stat.vt.edu, (2015). Parametric versus Semi/nonparametric Regression Models | LISA: the statistical collaboration laboratory of Virginia Tech. [online] Available at: http://www.lisa.stat.vt.edu/?q=node/7517

[Accessed 29 Sep. 2015]

Rayner, N.A., P.Brohan, D.E.Parker, C.K.Folland, J.J.Kennedy, M.Vanicek, T.Ansell and S.F.B.Tett (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set. Journal of Climate. 19(3) pp. 446-469 [Accessed 6 Oct. 2015]

**Team Contributions**

Dmitrii Kharkovskii -

Naheed Anjum Arafat -

Liu Dake -

Wu Lifu -

Daren Tan Zong Loong -

Margins (TO DELETE)

Papers **must** be printed in two-column format on 8-1⁄2 x 11 inch paper. The margins if you use 8-1/2 x 11 inch paper must be **exactly** as follows:

• Top margin: 3⁄4 of an inch

• Left margin: 3⁄4 inch

• Right margin: 3⁄4 inch

• Bottom margin: 1-1⁄4 inches

The actual image size for your paper (minus margins) is 7

x 9 inches (17.78 x 22.86 cm). Papers that deviate from these measurements will not be published.

Fonts

All fonts must be embedded for your paper to be published. No exceptions.