Amirtha Varshini.R.L., **EPSS137**, College of engineering,Guindy(Anna University)
S2, Kumaar\'s Kurunji, Plot 39Andal Nagar 2nd Main Road, Adambakkam,Chennai 600088
Guide:
Dr.Raaj Ramsankaran, IIT Bombay, Mumbai
Student Mentor:
Swathy Sunder, IIT Bombay, Mumbai

Abstract

Sea surface temperature (SST) is an essential climate variable for understanding and studying the climate system and quantifying ongoing climatic change like sea level change, tropical cyclones, the Asian summer monsoon, etc. Satellite instruments measure sea surface temperature by checking how much energy is emitted by the ocean at the thermal bands. Over the past 50 years, various agencies have launched plenty of satellites on board and scientists have come up with different methods (viz. Physical and data driven) to measure SST. However, most of the SST products are of low or medium resolution when it comes to coastal applications. The Landsat 8 Thermal Infrared (TIR) sensor gives an opportunity to get high resolution SST data of coastal regions. However, like any other infrared sensor, it suffers huge data loss due to cloud cover especially in the tropical region. From the literature, it was found that Sunder et al.,2020 have proved that Machine learning techniques are effective in the estimation of cloud free SST from MODIS data. Hence, in this study, a similar approach is proposed for retrieving SST from Landsat 8 Thermal Infrared Remote Sensing (TIRS) data. Sunder et al.,2020 demonstrated that variables such as the bright temperature at 11 and 12 µm, latitude, longitude and Julian day can be used for the estimation of SST. The study was carried out for the period of 2013-2015 for study area - Bay of Bengal (BoB). The Landsat-8 dataset was retrieved and processed with the help of google earth engine. SST was generated with ML techniques by training the dataset based insitu and the accuracy of results was evaluated by in-situ data collected from Centre ERS d'Archivage et de Traitement (CERSAT)—French ERS Processing and Archiving Facility (CERSAT, 2018). The resulting products were compared with the MODIS SVR SST values estimated by Sunder et al., 2020. The SVR algorithm performed satisfactorily during training and testing. However the results can be improved further if the datasets were trained for a longer time period with a wider range of reference datasets.

Table of Contents

Abbreviations

SST	Sea Surface Temperature
GEE	Google Earth Engine
BoB	Bay of Bengal
SVM	Support Vector Machine
SVR	Support vector Regression
ML	Machine Learning
BT	Brightness Temperature
GC	Google colab
CSV	Comma Separated values file
DN	Digital Number
B10	Band 10 - (10.60 - 11.19 µm), Thermal infrared 1
B11	Band 11 - (11.50 - 12.51 µm),Thermal infrared 2
RMSE	Root mean squared error
MAE	Mean absolute error

1.INTRODUCTION

1.1Background/Rationale

Sea Surface Temperature (SST) is considered as one of the fundamental geophysical variables, used to define the physical environment and the variability of aquatic ecosystems. It is an essential variable in the modeling of oceanography, marine weather, etc. and it is a crucial variable to assess the effects of global warming on the upper layer of the ocean, which is an indicator of the health of coastal ecosystems. Some measurements are made using marine instruments, but nowadays satellites are used to extract global SST data. High-resolution sea surface temperature (SST) estimates are greatly dependent on satellite-based infrared radiometers, which are proven to be highly accurate in the past decades.

It has been proved that Surface warming in the vast ocean regions is accompanied by reductions in productivity. The relationship between temperature and nutrient concentration of the surface water is inversely proportional to each other depending upon the location. Chlorophyll concentration assessment can be done only with the help of SST. Analysis of the relationship between sea surface temperature (SST) and chlorophyll-a (chl-a) increases our understanding of the productivity of the ocean. Satellite images provide proof about important information on oceanographic conditions, transformations and simultaneously help marine environmental monitoring and assessment.

Also the knowledge of fisheries directly depends on SST. Millions of people live along the Indian coastline which spans over 8100 km. The people are depending on fishing for their livelihood. In the vast and turmoil Indian ocean, locating and catching fish is always a challenging task for those people. Often, the search for fishes ends up being long and unproductive leading to rise in the price of the fishes and low profits. Potential and timely information about the location of major fish populations, would be of greater help to the fisherman community to have a better profit for their efforts. For that, the knowledge about SST is very much essential. SST is directly linked to the fish reproduction and migration. The usage of remote sensing data and its applications can be very useful for the management of coastal oceans and to devise the methods to use the satellite data to harvest food from sea. At this juncture, the scientists from marine sciences, remote sensing and fishery science collaborated to develop a technique that can use the remotely sensed sea surface temperature (SST) to identify the locations of fish aggregation. Increase in mean Sea surface Temperature over years has shown a greater impact on fish growth in the ocean. This was identified by SST dataset over the area.

So estimating SST is now a responsibility of the scientific community as accurate as possible. The knowledge of global SST distribution and temporal variation is a key input to forecasting and prediction systems. SST fields constrain upper-ocean circulation and thermal structure on daily, seasonal, decadal and climatic timescales, for the exchange of energy between the ocean and atmosphere in coupled ocean-atmosphere models, and as boundary conditions for the ocean, weather and seasonal forecasting models. Other applications include maritime safety, military operations, ecosystem assessment, supporting fisheries and tourism, transport and energy, human health, food security, and environmental policy.

Using brightness temperatures (BT) generated by the infrared sensors of the satellites, the SST can be retrieved. Most of the near-polar orbiting satellites have shown the accuracy of a root-mean-square error (RMSE) of less than 1 K for the global ocean as well as the regional seas. Most of these studies have focused on satellite imagery with medium or low spatial resolution of a few kilometers, such as 1 km in the case of NOAA/AVHRR or 4 km in the case of geostationary satellites. However, the presence of clouds is a big obstacle. This problem is more prominent across tropical regions such as Bay of Bengal (BoB), restricting the availability of cloud-free SST data for ocean applications.

Most of the operational products for estimating cloud- free SST are based on optimum interpolation (OI) approach. Also these data driven - model approaches need information about the decorrelation scales and co-variance functions. As of now, most of the global daily SST products are typically grid resolution between 0.05° ×0.05° and 0.25° ×0.25° or 5 to 25 km range. Due to spatial and temporal averaging applied for interpolation (Reynolds and 100 Chelton, 2010),the resolution of the products tend to be much coarser than the grid resolution. To the best of the authors' knowledge, till date only four attempts (Chao et al. 2009, Buongiorno

Nardelli et al. 2013, Chin et al. 2017 and Sunder et al.,2020) have been made to provide cloud-free SST products at 0.01 × 0.01 grid resolutions.

Considering major and complex problems faced in Remote sensing and geophysical variables, Machine learning techniques have proved to give convenient and easy solutions especially for remote sensing data. Machine Learning comprises a number of techniques such as Artificial Neural Networks (ANN), Support vector machines/support vector regression (SVM/SVR), decision trees, self organising map, ensemble methods such as random forests, neuro-fuzzy, genetic algorithm and multivariate adaptive regression splines. It was concluded that the SVR, a ML technique which is an effective technique for generating high resolution SST estimates compared to other ML techniques(Sunder et al.,2020).

Till date, the efforts to develop daily high-resolution cloud-free SST products mainly focused on fusion of multiple satellite and in situ data products, which involves complex computations and they are computationally expensive as well as At the same time, machine learning algorithms are useful in the estimation of various geophysical variables even during sparse data conditions, but its capabilities were still not assessed for estimation of cloud-free SST.

Therefore, to address the purpose, this study aims to estimate high resolution cloud-free SST from Landsat 8 with the help of GEE and ML technique, Support Vector Regression.

1.2 Statement of the Problem

Major changes on the ocean are identified by variations and fluctuations of Sea Surface Temperature. SST measurements benefit a wide spectrum of operational applications, including climate and seasonal monitoring/forecasting, military defense operations, validation of atmospheric models, sea turtle tracking, evaluation of coral bleaching, tourism, and commercial fisheries management. SST being vital for these many events, cloud-free SST with high spatial resolution is sparsely available. Most of the institutions (like NASA Ocean biology processing groups) use SST data estimated from non-linear sea surface temperature algorithms which use infrared(IR) bands to estimate the SST. But they undergo a huge loss of data considering places with the heavy cloud cover in tropical regions. Coastal applications which depend on SST might face problems during monsoon. Later microwave(MW) sensors were the solution for cloud free data since they penetrate through clouds. But they came with very high error due to various reasons like atmospheric absorption, large footprint,etc. Recently various combinations of sensors like IR-IR combination, MW-MW combination, MW-IR combination were tried to estimate cloud free SST. But they were time-consuming, demanded complex approaches and various assumptions. So, till date Scientific community has been trying to estimate cloud-free SST using various sensor combinations, fusion of in situ data and satellite images, and various complex approaches for the same.

Hence, the present study focuses on the estimation of cloud-free SST from an ultra high resolution sensor viz. Landsat 8.

1.3 Objectives of the Research

Landsat 8 can provide high resolution data but they are blocked by cloud cover unlike microwave data. But cloud- free SST from Landsat 8 can be estimated with the help of advanced Machine learning techniques such as SVR despite the presence of clouds. They just need in situ data for training models.

1.3.1 Overall Objective

The main objective of this research is the estimation of high resolution cloud-free sea surface temperature (SST) on a daily basis using machine learning technique from Landsat 8 data.

1.3.2 Specific objectives

- Extract the brightness temperature values of Landsat 8 data using google earth engine.
- Develop a machine learning model based on the algorithm for the estimation of cloud free SST.

- Intercompare the testing results with MODIS SVR SST,(Sunder et al.,2020).

1.4 Scope

The scope of this project is to evaluate the high resolution cloud free SST from the Landsat 8 Thermal Infrared Remote Sensing (TIRS) data. First, retrieve the DN from the landsat data using GEE. Second, calculating the bright temperature based on the radiation transfer equation and Planck's law. Finally, evaluate the cloud-free SST with SVR technique using Weka. The accuracy of inversion results is evaluated by in-situ SST and the SST generated by Sunder et al.,2020.

2 LITERATURE REVIEW

2.1 Information

The existing theories about Sea surface temperature are that the SST product provides sea surface temperature at 1- km (Level 2) and 4.6 km, 36 km, and 1° (Level 3) resolutions over the global oceans. Derived from radiance measurements collected by the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard NASA's Terra and Aqua satellites, the sea surface temperature (MODIS) is an estimate of the warmth of the ocean's "skin" (top millimeter). The algorithm uses multiple atmospheric window techniques to estimate the atmospheric parameters that are required to compensate for absorption and scattering of energy radiated and reflected by the ocean. The bulk temperature of the near-surface ocean is the temperature of the upper 10-20 cm to 1 meter as measured by conventional thermometers on buoys and ships. Extensive analysis has been done of satellite and *in situ* data to enable the algorithm to estimate bulk temperature as well as skin temperature.

Sea surface temperature determination is based on MODIS-calibrated mid- and far-infrared (IR) radiances (Bands 20, 22, 23, 31, and 32 from MODIS), using an algorithm that exploits the differences in atmospheric transmissivity in the different IR bands to enable highly accurate estimation of the atmospheric effects. A set of spatial and temporal homogeneity tests is applied to validate the quality of the cloud-free dataset. Sun glint is a significant source of error in the mid-wave IR bands. Ocean warming is a hallmark of the climate pattern El Niño, which changes rainfall patterns around the globe, causing heavy rainfall in the southern United States and severe drought in Australia, Indonesia, and southern Asia. On a smaller scale, ocean temperatures influence the development of tropical cyclones (hurricanes and typhoons), which draw energy from warm ocean waters to form and intensify.For all of these reasons scientists monitor the sea's surface temperature.

So most of these institutions have been using nonlinear algorithms to generate SST. But the major problem faced by them is the cloud cover, which can be so erroneous considering tropical regions. Coastal application suffers huge loss and deviation areas with heavy cloud cover(blockage or gap in dataset). The scientific community has been trying various combinations of sensors and complex approaches to estimate cloud-free SST. It has been proven that Machine learning techniques could provide an easier solution around complex problems faced with respect to Remote sensing of geophysical variables(Wang and Deng, 2017; Lary et al., 2016).

Regarding the various machine learning techniques, it was proved that the SVR algorithm has been working well for the Indian regions like BoB in terms of error values(Sunder et al.,2020).

SST has not been generated using Landsat 8 dataset. This research aims to generate SST from Landsat 8 Image Collection.

2.2 Summary

Sea surface temperature is highly essential for Ocean Biology and study of various scientific phenomena. Landsat 8 has a very high resolution. So this research uses google earth engine to generate Brightness temperature and use ML - SVM to generate High- Resolution cloud-free SST from Landsat 8.

3 METHODOLOGY

3.1 Concepts

3.1.1 Study Area:

The study areas selected for this research work is an ocean area situated in Bay of Bengal. Study area is the south-western part along the Indian Coastline. This area is within the tropical latitude and longitudinal extent of 12 ° to 22° N and longitudes of 82° to 95° E. BoB receives runoff from major rivers such as Ganga and Brahmaputra into the northern bay. Therefore, the surface layer in the BoB is fresh; resulting in a higher salinity. As a consequence, typical profiles of temperature and salinity in the basins differ considerably.

Since this region is also known for its heavy cloud cover, with high resolution Landsat 8 images and SVM technique, high resolution cloud free SST can be generated.

3.1.2 Datasets:

The study involves use of satellite and in-situ data of three years from January 2013 to December 2015. Here, the Landsat - 8 satellite data was selected because they are available on daily scale at grid resolution. This satellite has a 16-day repeat cycle with an equatorial crossing time: 10:00 a.m. +/- 15 minutes. Landsat 8 acquires about 740 scenes a day on the Worldwide Reference System-2 (WRS-2) path/row system, with a swath overlap (or sidelap) varying from 7 percent at the equator to a maximum of approximately 85 percent at extreme latitudes. Data products created from over 1.8 million Landsat 8 OLI/TIRS scenes are available in GEE. In our study, Landsat 8 TIRS data are used to retrieve the offshore SST, with a spatial resolution of 30 m, 15 viewing angle, and 16-day revisit period. It has two thermal infrared channels with a 10–12 μm wavelength. It has high resolution imagery and required band imagery.Therefore, the Landsat 8 satellite is highly efficient for this research.

3.1.3 In-situ

In situ data is used in this research to train the Machine learning model and for comparison purposes. So it has a greater importance in this research. The in-situ SST data used in this study were obtained from the Centre ERS d'Archivage et de Traitement (CERSAT)—French ERS Processing and Archiving Facility (CERSAT, 2018). CERSAT collects surface-level in-situ SST data from Coriolis data center. Distribution of the data points used in this study is for the BoB region. Data collected during the years 2013-2014 were used for training(1174 data points) and the data collected during 2015 were used for independent testing(582 data points). In situ data sets which have been collected within ±5hrs of the satellite overpass were used. Detailed properties of the in-situ data used in this study are mentioned in Table.1. This will be used for training the dataset to obtain cloud-free SST and also as reference for checking the accuracy of the SST derived from Landsat 8.

Table 1 : Statistical Characteristics of In-Situ

Statistics	**Training set**	**Test set**
Minimum(℃)	*23.8*	*11.539*
Maximum(℃)	*32*	*29.264*
Mean(℃)	*28.032*	*25.203*
StdDev(℃)	*2.515*	*2.988*

3.1.4 Platform

Google Earth Engine (GEE) is a cloud-based platform that enables large-scale scientific analysis and visualization of geospatial data sets. Google Earth Engine is a platform to combine a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities. The data repository is a collection of over 40 years of satellite imagery for the whole world. The cloud computing power of GEE enables the processing of Big data, combined with other vector data, within the cloud environment and removes the need to store, process, and analyze the large volumes of satellite data. Overall, GEE has opened a new big data paradigm for storage and analysis of remotely sensed data at a scale that was not feasible using desktop processing machines. One of the merits of GEE is the Landsat Image Collection and the programming interface that allows users to create and run custom algorithms. The whole million set of images can be easily narrowed down to the region of interest. This has the facility to import dataset like co -ordinates from in-situ data. It also can export the data to drive as a table(CSV), GeoTiff, and KML files.

WEKA is a data mining system that implements data mining algorithms. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset and results can be stored in various formats.

3.1.5 Machine Learning Algorithm

ML has proven to be of greater importance in classifying, analysing remotely sensed data and other geospatial analysis. The satellite data used for this research is a temporally varying dataset taken in extremely different locations.

The types of the ML algorithms commonly used are artificial neural networks (ANN), support vector machines (SVM), self-organizing map (SOM), decision trees (DT)(e.g., Shahin et al., 2001, Shahin and Jaksa, 2005, Das and Basudhar, 2008, Samui, 2008a, Samui, 2008b, Samui, 2012, Azamathulla and Wu, 2011, Azamathulla et al., 2011, Azamathulla et al., 2012, Garg et al., 2014a, Garg et al., 2014b, Garg et al., 2014c). In this study support vector machines (SVM) are trained with an imbalanced training set. SVM is a supervised machine learning algorithm which can be used for classification or regression problems. A training set is a subset to train a model. A test set is a subset to test the trained model.

Generally the regression based on support vector machines (SVRs) can be explained as follows:

SVM estimates by minimising an upper bound on the probability that the estimation error may be above a given threshold (Moser et al., 2009). For a set of training samples $((x_1, y_1), (x_2, y_2),.... (x_h, y_h)..., (x_N, y_N))$of sample size N(where x_h is the h^{th}feature vector corresponding to the reference measurement $y_h: h = 1, 2, 3,.... N$),the resulting approximation can be expressed as a linear combination of suitable kernel functions centred on a subset of training samples as given by Eqn.(3) (Moser et al., 2009; Vapnik, 1998). A kernel is a similarity function which is fed into a machine learning algorithm. But in the case of similarity functions, a kernel function can be defined which internally computes the similarity between images, and then feeds into the learning algorithm along with the images and label data.

$$f(x) = \sum_{h\epsilon S} \beta^*_h(x_h, x|\gamma)+ b^*$$

Where $\beta^*_1,\beta^*_2\ldots\beta^*_N$is the weight coefficient of the linear combination K(,.|γ) is a kernel function in general, by a vector γof r real valued parameter ($\gamma\varepsilon R^r$), b^* is a bias term, and $S = \{h: \beta^*_h \neq 0\}$.If $h\epsilon S$,i.e., $\beta^*_h \neq 0$,the training sample $x_h$is named ‘Support vector’(Moser et al ., Vapnik,1998).

In this research, Weka software by Witten has been used to develop ML based SST algorithms. With in, SVM classifier, It is found that the SVM model based on Pearson VII kernel function (PUK) shows the same applicability, suitability, performance in prediction of yarn tenacity as against SVM based RBF kernel.(Khalid Aa Abakar, Chongwen Yu). The general form of the pearson VII function for curve fitting purposes is given by the following relationship:

$$f(x) \ = \ H/[1 \ + \ 2(x - x_0)\sqrt{2^{(1/\omega)} - 1/\sigma)} \ ^2] \ ^{\omega}$$

Where H is the peak height at the centre x_0 of the peak and x represents the independent variable. The parameters ω *and* σcontrol the half width (also named Pearson width) and the tailing factor of the peak.

3.2 Methods

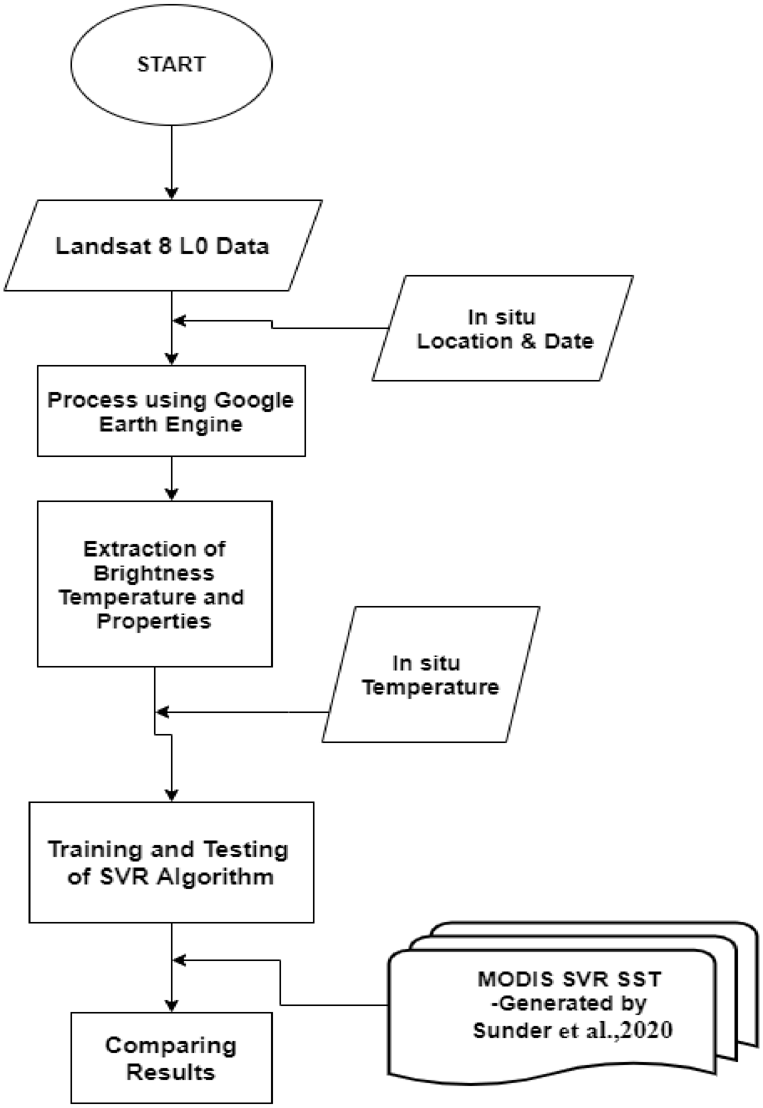

START
Landsat 8 L0 Data
In situ Location & Date
Process using Google Earth Engine
Extraction of Brightness Temperature and Properties
In situ Temperature
Training and Testing of SVR Algorithm
MODIS SVR SST -Generated by Sunder et al.,2020
Comparing Results

Figure 1. Flow chart of methodology

Figure 1 demonstrates the overall flow of events in the study. The landsat 8 images can be easily accessed using Google earth engine.Then in GEE itself a code was developed to extract band 10,11 values and generate brightness temperature with the help of latitude and longitude. It has been proved that latitude and longitude have a direct effect of data - driven approach of SST estimation(Alavi et al.,2016 and Picart et al., 2018). Also the date is converted into Julian days to account the seasonal characteristics. In situ data and Brightness temperature table was merged along with the respective coordinates and Julian day for the training and testing purposes.

3.2.1 Brightness temperature Extraction - GEE

The brightness temperature is calculated based on the radiance of the thermal infrared bands. In this research, using GEE, the landsat 8 collection is imported and filtered using various factors such as Date, Location and band specification. The Features and geospatial properties are extracted from in situ data for each year and exported as a CSV file. The total number of landsat image collection for a single year usually goes from 1 to 10 million. First landsat image collection is filtered based on bands(B10, B11) and region of Interest. Then based on each date from the in situ, the image collection is further filtered. Then based on location, for each image band values and properties are extracted and collected together. The properties include 10, 11 band values, location coordinates and time of extraction. Then they are exported as CSV files. Due to the limitation of Google earth engine, the in situ is split into sets of 1000 locations. Their respective CSV is merged together as a single CSV. Finally, CSV files are categorized based on years. The CSV file is then imported in Google collab where the extracted DN value is converted to Brightness temperature using the following formulas:

$$L_{\lambda} = ML * Q_{cal} + A_{L}$$

Where :-

$L_{\lambda} - Spectral\,radiance\ (W/(m^{2} *\ sr\ *\ \mu m)\ )$

$M_{L} - Radiance\ multiplicative\ scaling\ factor\ for\ the\ band$

(RADIANCE_MULT_BAND_n from the metadata)

$A_L - Radiance\ additive\ scaling\ factor\ for\ the\ band$

(RADIANCE_MULT_BAND_n from the metadata)

$Q_{cal} - Level\ 1\ pixel\ value\ in\ DN$

TIRS data can also be converted from spectral radiance (as described above) to brightness temperature, which is the effective temperature viewed by the satellite under an assumption of unity emissivity. The conversion formula is as follows:

$$T = \frac{K_2}{ln(\frac{K_1}{L_\lambda}+1)}$$

Where:

$T - Top\ of\ atmosphere\ brightness\ temperature\ (K)$

$L_\lambda - Spectral\ radiance\ (W/(m^2 * sr * \mu m))$

$K_1 - Band - specific\ thermal\ conversion\ constant\ from\ the\ metadata$

(K1_CONSTANT_BAND_x, where x is the thermal band number)

$K_2 - Band - specific\ thermal\ conversion\ constant\ from\ the\ metadata$

(K2_CONSTANT_BAND_x, where x is the thermal band number)

Finally, the brightness temperature at

Band 10 - (10.60 - 11.19 µm) Thermal infrared 1, resampled from 100m to 30m

Band 11 - (11.50 - 12.51 µm) Thermal infrared 2, resampled from 100m to 30m

are generated.

3.2.2 SVM on the Brightness temperature dataset

The Brightness temperature is evaluated using the formula above.The brightness temperature is obtained in kelvin. It is then converted to ℃ by a subtraction of -273.15 from the brightness temperature. Now the In - situ data and GEE generated temperature table is merged based on common latitude, longitude and minimum time difference for training and testing purposes.

Cross validation technique is used on the dataset. So basically, Cross-validation is a technique in which is used to train a model using the subset of the data-set and then evaluate using the rest of the subset of the data-set. Usually for this method, training is on the 70% of the given data-set and the rest 30% is used for the testing purpose. The 2013 and 2014 dataset is used as a training set. The dataset for the remaining years is categorized as the testing set. Generally in this method, the data-set is split into k number of subsets(known as folds) then the training is performed on all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, for each time a different subset reserved for testing purposes iterate k times. The fold is kept and used a default value as 10 and tried. The more fold, lesser the error.

Now the dataset is put in the Preprocessing of Weka for classification. SMOreg is fixed as the classifier that implements the support vector machine for regression. Using the SMOreg, different kernels are tried. PUK is found as the best kernel. Using cross validation, keeping folds as 100, the training is performed on the training set. The dataset is trained. The model is saved. Then the trained model is applied on the test set. Then the results were visualized and the errors were analysed.

4 RESULTS AND DISCUSSION

The SVR algorithm was evaluated on the data set. The algorithm was trained and tested using In situ data and Brightness temperature developed using Google earth engine. Figure 2 and figure 3 shows the comparison in situ sst and SVR sst generated using this SVR algorithm. R^2,Root mean square error (RMSE) and mean absolute error(MAE) values were obtained from the result. Basically, MAE measures the average magnitude of the errors in the set of predictions. And RMSE is a measure of how spread out these residuals(prediction error) are. The R-squared value, denoted by R^2, is the square of the correlation. It measures the proportion of variation in the dependent variable that can be attributed to the independent variable. Therefore, a comparison of R^2 , RMSE and MAE(error values) can be used to determine whether the forecast contains large but infrequent errors. Larger the difference between RMSE and MAE, the more inconsistent the error size is. Finally the results were inter-compared with MODIS SVR SST generated by Sunder et al.,2020.

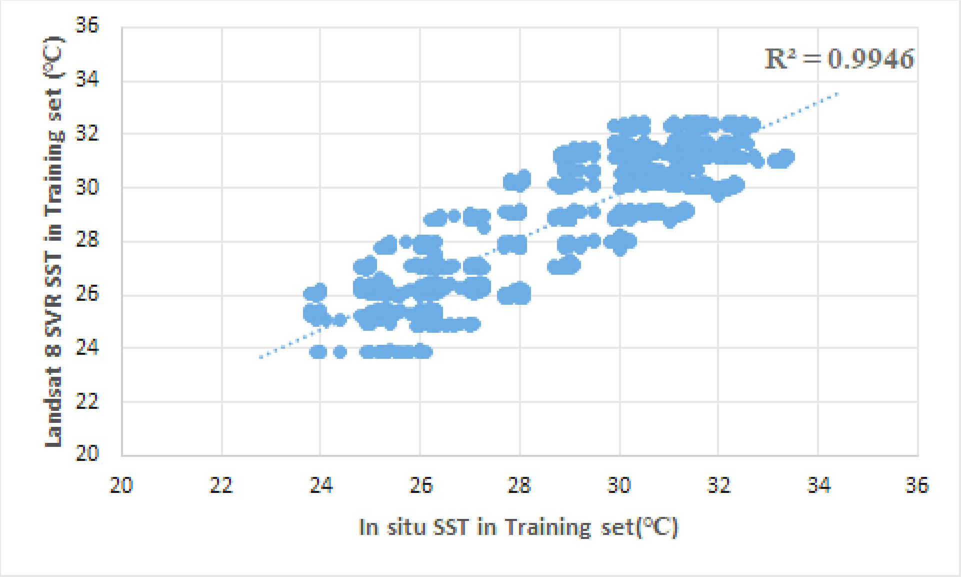


Figure.2: Representation of In situ and SVR SST in Training set in Bay of Bengal

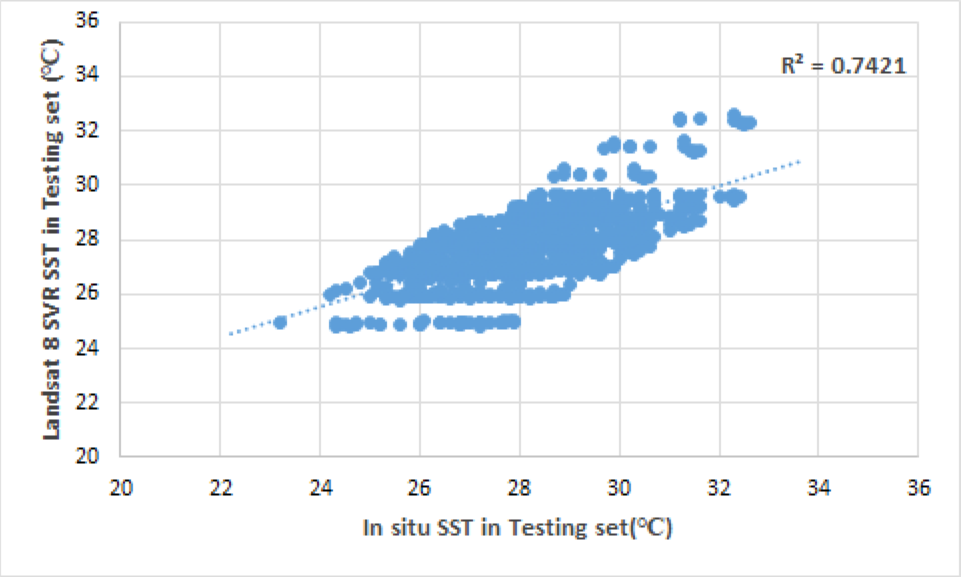


Figure.3: Representation of In situ and SVR SST in Testing set in Bay of Bengal

4.1 Inter-comparison of MODIS SVR SST and Landsat 8 SVR SST

The results of inter-comparison between **MODIS SVR SST** and **Landsat 8 SVR SST** generated wrt in-situ data in the study region is given in Table 2 below. In the BoB region, SVR SST generated in this research has lower R^2, higher RMSE and higher MAE compared the MODIS SVR SST generated by Sunder et al.,2020. The error values indicate that the MODIS SVR SST has lesser error. Moreover, it was observed that the RMSE value of the Landsat 8 SVR SST is almost thrice the RMSE of MODIS SVR SST in BoB region. It should be noted that the SVR SST derived in this study is of 30 m resolution, which is much finer than the MODIS SST SST which has 1km resolution. In this research, the data set was taken only for a span of 2 years for training and 1 year data for testing. But for MODIS SVR SST estimation, the dataset taken for training was for 8 years. And the testing set was data for 2 years. This may be the main reason for the difference seen in the result comparison between Landsat 8 SVR SST and MODIS SVR SST. Also the statistics of the in-situ data shows that there's a considerable difference in their mean minimum and maximum values. So the training couldn't reflect all ranges of in -situ data. If the dataset was taken for a longer period of time say, 10 years, the result could have been different and more efficient.

Therefore, even though MODIS SVR SST is slightly performing better. SVR SST generated in this research will be useful for studies which require high resolution SST.

Table 2. Performance of MODIS SVR SST and Landsat 8 SVR SST with respect to in situ data for BoB for the period 2015.

Statistics	Landsat 8 SVR SST	MODIS SVR SST
R^2	0.7421	0.9676
RMSE(℃)	1.0907	0.4575
MAE(℃)	0.6989	0.258

5 CONCLUSION AND RECOMMENDATIONS

This study is unique as it tries to estimate cloud -free high resolution SST from Landsat 8 data using Machine learning technique for the first time. The support vector regression technique(SVR) was selected for the study. The SVR model was trained and tested for the study region, Bay of Bengal (BOB).The obtained results when analysed with respect to in-situ data as well the MODIS SVR SST developed by Sunder et al.,2020. It was observed that the MAE value of the SVR algorithm is higher compared to the MODIS SVR SST in BoB so as the RMSE values. This is majorly due to statistical difference in in situ data used for training and testing (major difference between mean minimum and maximum) and the size of training set and testing set taken for study. It is relatively smaller compared to the dataset taken for MODIS SVR SST estimation. It should be noted that the Landsat 8 SVR SST derived in this study is of 30m resolution, which is much closer to the MODIS SVR SST which has 1km resolution. Also, despite the high cloud coverage in both training and testing set, ML technique was able to produce high resolution cloud-free SST with high accuracy using Landsat 8 dataset. The SVR SST generated in this study can be used for studies that require very high resolution. Further advanced techniques such as deep learning could be used to improve the accuracy of the algorithm in the future.

REFERENCES

Alavi, Amir H., Amir H. Gandomi, and David J. Lary. 2016. “Progress of Machine Learning in Geosciences: Preface.” *Geoscience Frontiers* 7 (1): 1–2. doi:10.1016/j.gsf.2015.10.006.

Autret, E. and Piolle,J.F., 2011. Product User Manual for ODYSSEA Level 3 and 4 global and regional products. MYO-PUM-SST-TAC-ODYSSEA, Ifremer/CERSAT.[Available online at:http://projets.ifremer.fr/cersat/Data/Discovery/By-parameter/Seasurfacetemperature/ODYSSEA-Global-SST-Analysis].

Baith, K.,Lindsay, R., Fu, G. and McClain, C.R., 2001. Data analysis system developed for ocean color satellite sensors. Eos, Transactions American Geophysical Union, 82(18), pp.202-202.
.

Barton, Ian J.2001. “Interpretation of Satellite-Derived Sea Surface Temperatures.” Advances in Space Research 28 (1): 165–70. doi:10.1016/S0273 1177(01)00337-4.
.

Brasnett, B.,2008." The impact of satellite retrievals in a global sea‐surface temperature analysis". Quarterly Journal of the Royal Meteorological Society, 134(636), pp.1745-1760.
.

Buongiorno Nardelli, B., C. Tronconi, A. Pisano, and R. Santoleri. 2013. “High and Ultra-High Resolution Processing of Satellite Sea Surface Temperature Data over Southern European Seas in the Framework of MyOcean P 604 project.” Remote Sensing of Environment(February): 1–16. doi:10.1016/j.rse.2012.10.012.

CERSAT.2018, Sea Surface Temperature In Situ Data [online].available at http://cersat.ifremer.fr/data/tools-and-services/match-up-databases/item/298-seasurfacetemperature-in-situ-data·

Chao, Yi, Zhijin Li,John D. Farrara, and Peter Hung. 2009. “Blending Sea Surface Temperatures from Multiple Satellites and in Situ Observations for Coastal Oceans.” Journal of Atmospheric and Oceanic Technology 26 (7): 1415–26. doi:10.1175/2009JTECHO592.1.
.

Deng, C. and Wu, C.,2013. The use of single-date MODIS imagery for estimating large-scale urban impervious surface fraction with spectral mixture analysis and machine learning techniques. ISPRS Journal of Photogrammetry and Remote Sensing, 86, pp.100-110.
Fang, B., Li, Y., Zhang, H. and Chan, J.C.W., 2020. Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples. ISPRS Journal of Photogrammetry and Remote Sensing, 161, pp.164-178.
·

Hydrography and biogeochemistry of the north western Bay of Bengal and the north eastern Arabian Sea during winter monsoon. Journal of Marine Systems, 73(1-2), pp.76-86.Imagery." Continental Shelf Research 30 (18). 1951–62. doi:10.1016/j.csr.2010.08.016.

Gabriele,Morie and Sebastiano B Serpico. 2009. "Automatic Parameter Optimization for Support Vector Regression for Land and Sea Surface Temperature Estimation From Remote Sensing Data", IEEE Transactions on Geoscience and Remote Sensing , 47 (3): 909–21.

Kamir, E., Waldner, F. and Hochman, Z., 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS Journal of Photogrammetry and Remote Sensing, 160, pp.124-135.

LaCasse, Katherine M.,Michael E. Splitt, Steven M. Lazarus, and William M. Lapenta.656 2008."The Impact of High-Resolution Sea Surface Temperatures on the Simulated Nocturnal Florida Marine Boundary Layer." Monthly Weather Review 136 (4): 1349–72.doi:10.1175/2007mwr2167.1.

Lary, David J., Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. 2016. "Machine Learning in Geosciences and Remote Sensing." Geoscience Frontiers 7 (1). 3–10. doi:10.1016/j.gsf.2015.07.003.·

Liu, Meiling, Xiangnan Liu, Da Liu, Chao Ding, and Jiale Jiang. 2015. "Multivariable Integration Method for Estimating Sea Surface Salinity in Coastal Waters from in SituData and Remotely Sensed Data Using Random Forest Algorithm." Computers & Geosciences 75.44–56. doi:10.1016/j.cageo.2014.10.016.

Maturi, Eileen, Andy Harris, Chris Merchant, Jon Mittaz, Bob Potash, Wen Meng, and John Sapper. 2008. "NOAA's Sea Surface Temperature Products from Operational Geostationary Satellites." Bulletin of the American Meteorological Society 89 (12):1877–88.doi:10.1175/2008BAMS2528.1.
·

NASA Goddard Space Flight Center, Ocean Biology Processing Group. 2014. Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua Level 0 Data; NASA OB.DAAC,Greenbelt,MD, USA. Available at https://oceandata.sci.gsfc.nasa.gov/MODIS-Aqua/L0. Maintained by NASA Ocean Biology Distributed Active Archive Center OB.DAAC), Goddard Space Flight Center, Greenbelt MD.Newman, S. M., J. a. Smith, M. D. Glew, S. M. Rogers, and J. P. Taylor. 2005. "Temperature and Salinity Dependence of Sea Surface Emissivity in the Thermal Infrared." Quarterly Journal of the Royal Meteorological Society 131: 2539–57. doi:10.1256/qj.04.150.

O'Carroll, A.G., Armstrong, E.M., Beggs, H., Bouali, M., Casey, K.S., Corlett, G.K., Dash, P., Donlon, C., Gentemann, C.L., Høyer, J.L., Ignatov, A., 2019. Observational needs of sea surface temperature. Frontiers in Marine Science, 6, p.420.Picsart, Stéphane Saux, Pierre Tandeo, Emmanuelle Autret, and Blandine Gausset. 2018.“Exploring Machine Learning to Correct Satellite-Derived Sea Surface Temperatures.” Remote Sensing 10 (2): 1–11. doi:10.3390/rs10020224.

·

Sunder et al.,2020. Machine Learning Techniques for Regional Scale Estimation of High-Resolution Cloud-Free Daily Sea Surface Temperatures from MODIS Data.

·

Witten, Ian H, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

ACKNOWLEDGEMENTS

The authors also thank CERSAT, Ifermer for providing the in situ reference data. Additionally, authors would like to acknowledge Google earth engine for providing Landsat 8 L0 dataset. Authors would like to thank the GEE developers group for the criticisms.

APPENDICES

Source code:

Google earth engine

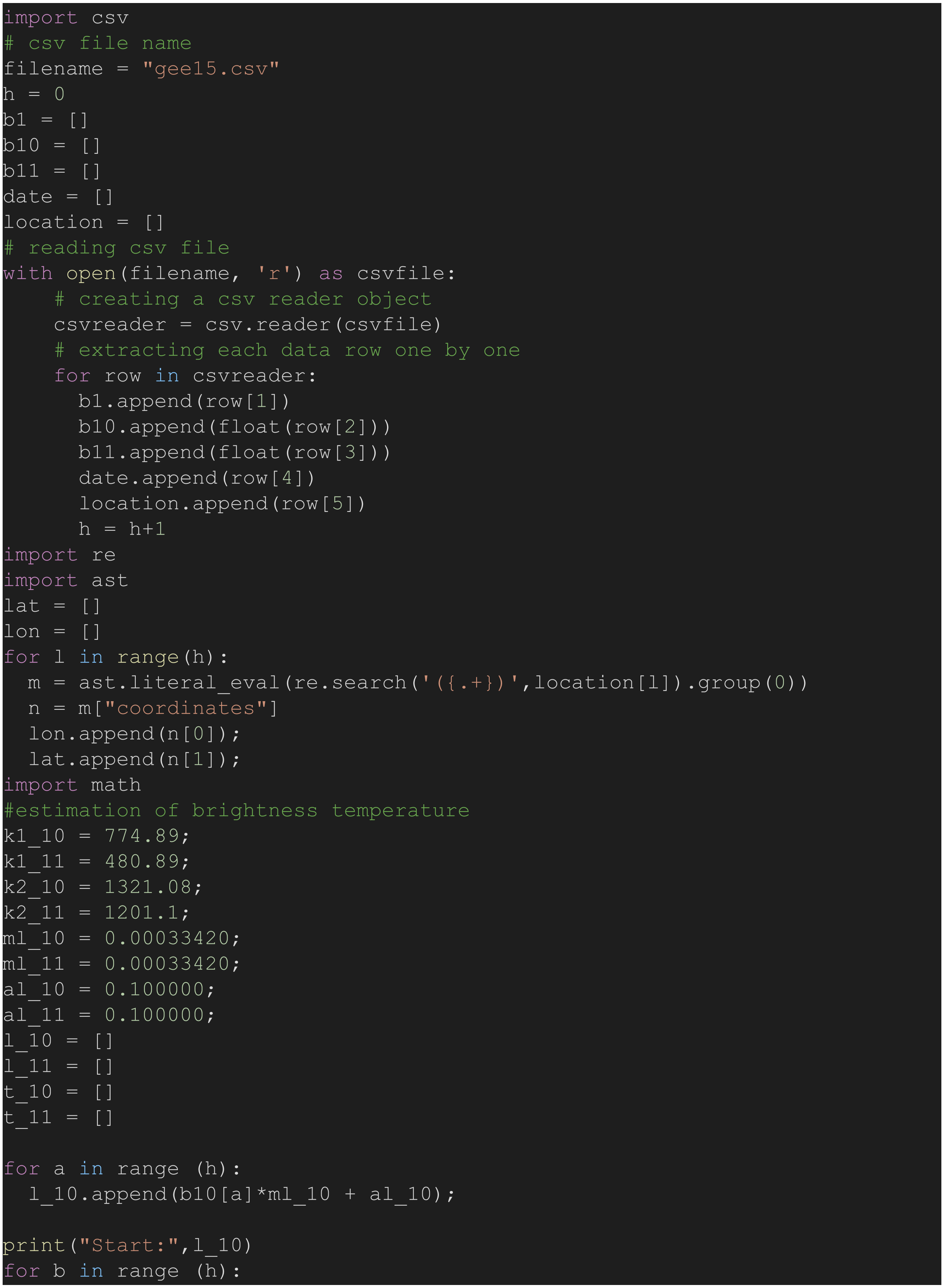
Imports

```
var table :Table users/kavya2000raja/data_2014
```

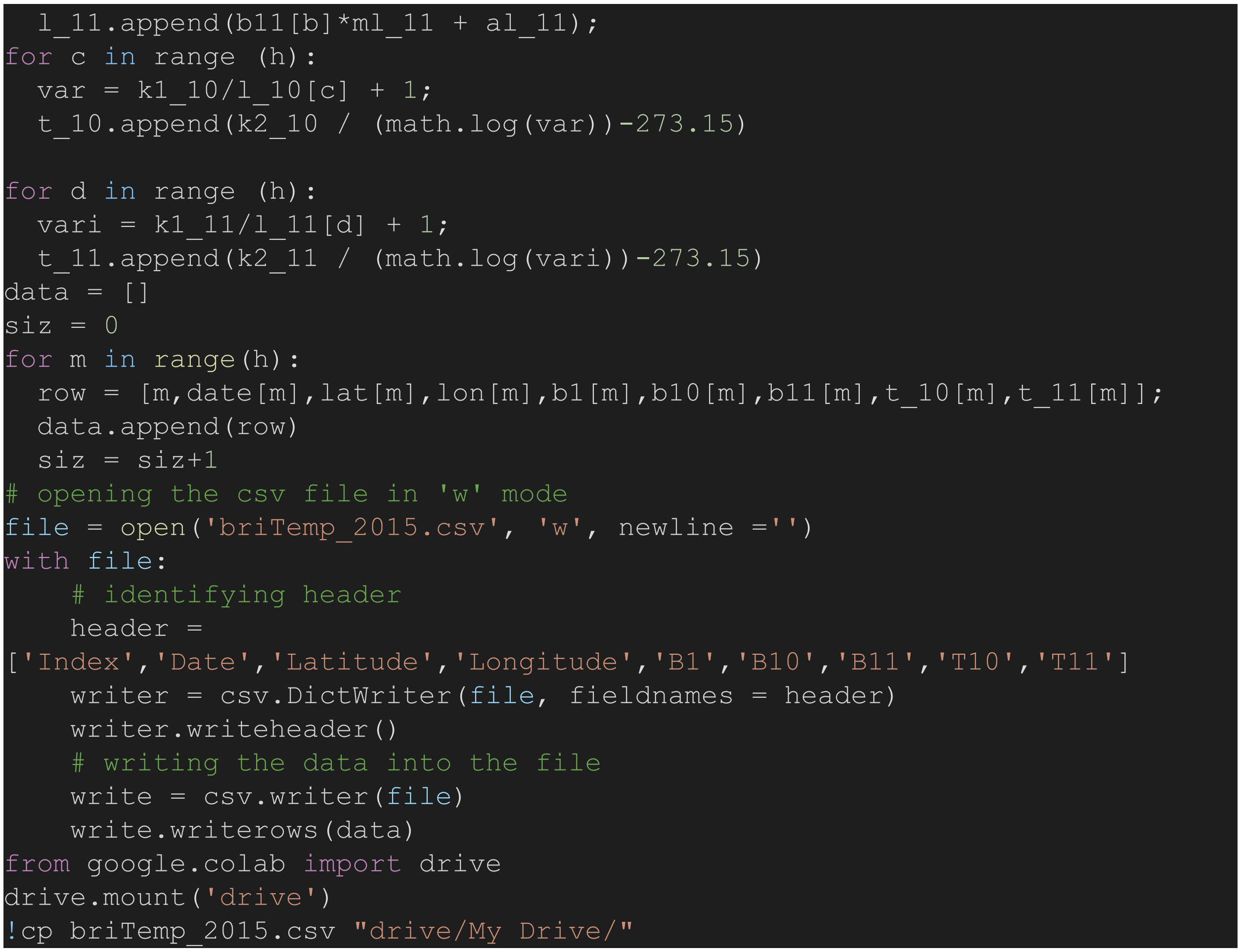
```
print(table.aggregate_array('end_date').size())
var i =0;
var sdate = table.aggregate_array('start_date').slice(i,i+1000);
var edate = table.aggregate_array('end_date').slice(i,i+1000);
print(edate)
var lon = table.aggregate_array('i_lon').slice(i,i+1000);
var lat = table.aggregate_array('i_at').slice(i,i+1000);
var dates = [];
for(var i = 0 ; i<1000;i++){
  dates[i] = [sdate.get(i),edate.get(i)];
}

var c1 = lon.zip(lat);
var map = function (coord){
  return ee.Geometry.Point(coord)
}
var c2 = ee.List(c1).map(map);
var newft
var newft1=ee.FeatureCollection([])
var main = ee.List([]);
for (var i = 0;i<1000;i++){
  var ls = dates[i];
   var start =ls[0];
  var end = ls[1];
  var polygon =
ee.Geometry.Polygon([[[95.944,22.516],[83.825,22.043],[85.825,12.225],[95.944,1
2.225]]]);

  var landsat =
ee.ImageCollection("LANDSAT/LC08/C01/T1").select(['B1','B10','B11']);

  var l1 = landsat.filterBounds(polygon);

  var cc =c2.get(i);
  var ccc = c1.get(i);
  var l2  =l1.filterDate(start,end);
  var l3 = l2.filterBounds(cc);
  var l4 = l3.select(['B1','B10','B11']);
  var ft = ee.FeatureCollection(ee.List([]));
  var newft = ee.FeatureCollection(ee.List([]));
    var fill = function(img, ini) {
    var inift = ee.FeatureCollection(ini)
    var c = ee.FeatureCollection(ee.Geometry.Point(c1.get(i)));
    var ft2 = img.reduceRegions(c, ee.Reducer.first(),30)
    var date = img.date().format()
    var ft3 = ft2.map(function(f){return f.set("date", date)})
        return inift.merge(ft3)
  }
  newft=ee.FeatureCollection(l4.iterate(fill, ft));
  newft1 = newft1.merge(newft);

}
print(newft1)

Export.table.toDrive({
 collection:newft1,
 description: 'exportTable',
 fileFormat: 'CSV'
});
```

Google colab

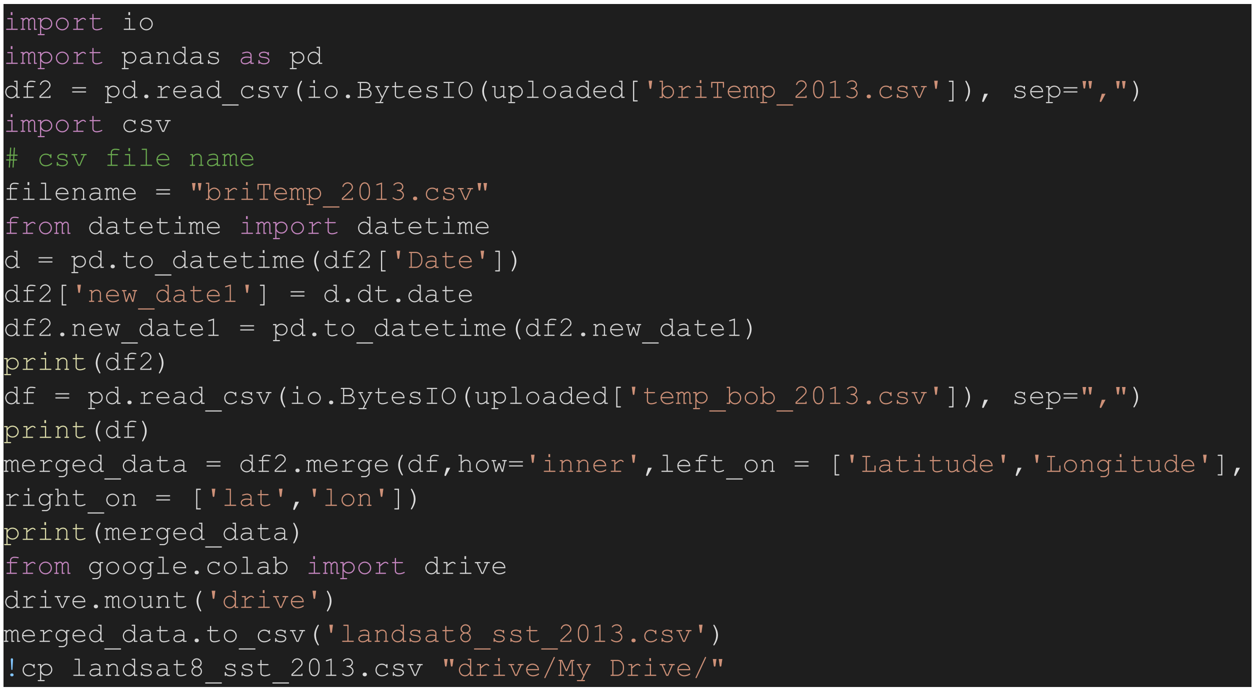
```
from google.colab import files
uploaded = files.upload()
import io
import pandas as pd
df2 = pd.read_csv(io.BytesIO(uploaded['gee15.csv']), sep=",")
```



```
import csv
# csv file name
filename = "gee15.csv"
h = 0
b1 = []
b10 = []
b11 = []
date = []
location = []
# reading csv file
with open(filename, 'r') as csvfile:
    # creating a csv reader object
    csvreader = csv.reader(csvfile)
    # extracting each data row one by one
    for row in csvreader:
      b1.append(row[1])
      b10.append(float(row[2]))
      b11.append(float(row[3]))
      date.append(row[4])
      location.append(row[5])
      h = h+1
import re
import ast
lat = []
lon = []
for l in range(h):
  m = ast.literal_eval(re.search('({.+})',location[l]).group(0))
  n = m["coordinates"]
  lon.append(n[0]);
  lat.append(n[1]);
import math
#estimation of brightness temperature
k1_10 = 774.89;
k1_11 = 480.89;
k2_10 = 1321.08;
k2_11 = 1201.1;
ml_10 = 0.00033420;
ml_11 = 0.00033420;
al_10 = 0.100000;
al_11 = 0.100000;
l_10 = []
l_11 = []
t_10 = []
t_11 = []

for a in range (h):
  l_10.append(b10[a]*ml_10 + al_10);

print("Start:",l_10)
for b in range (h):
```



```
  l_11.append(b11[b]*ml_11 + al_11);
for c in range (h):
  var = k1_10/l_10[c] + 1;
  t_10.append(k2_10 / (math.log(var))-273.15)

for d in range (h):
  vari = k1_11/l_11[d] + 1;
  t_11.append(k2_11 / (math.log(vari))-273.15)
data = []
siz = 0
for m in range(h):
  row = [m,date[m],lat[m],lon[m],b1[m],b10[m],b11[m],t_10[m],t_11[m]];
  data.append(row)
  siz = siz+1
# opening the csv file in 'w' mode
file = open('briTemp_2015.csv', 'w', newline ='')
with file:
    # identifying header
    header = 
['Index','Date','Latitude','Longitude','B1','B10','B11','T10','T11']
    writer = csv.DictWriter(file, fieldnames = header)
    writer.writeheader()
    # writing the data into the file
    write = csv.writer(file)
    write.writerows(data)
from google.colab import drive
drive.mount('drive')
!cp briTemp_2015.csv "drive/My Drive/"
```



```
import io
import pandas as pd
df2 = pd.read_csv(io.BytesIO(uploaded['briTemp_2013.csv']), sep=",")
import csv
# csv file name
filename = "briTemp_2013.csv"
from datetime import datetime
d = pd.to_datetime(df2['Date'])
df2['new_date1'] = d.dt.date
df2.new_date1 = pd.to_datetime(df2.new_date1)
print(df2)
df = pd.read_csv(io.BytesIO(uploaded['temp_bob_2013.csv']), sep=",")
print(df)
merged_data = df2.merge(df,how='inner',left_on = ['Latitude','Longitude'],
right_on = ['lat','lon'])
print(merged_data)
from google.colab import drive
drive.mount('drive')
merged_data.to_csv('landsat8_sst_2013.csv')
!cp landsat8_sst_2013.csv "drive/My Drive/"
```