Reconstructed Solar-Induced Fluorescence Based on AVHRR and MODIS RED NIR Reflectance Data

Yu Huang (yh3019)
Weiwei Zhan (wz2483)

1. Background

Solar-induced chlorophyll fluorescence (SIF) is an optical signal emitted during the light reactions of photosynthesis (Porcar-Castell et al., 2014). When plants absorb sunlight through the chlorophyll, there are three pathways for the absorbed photons: photochemical quenching for photosynthesis, non-photochemical quenching for heat dissipation, and fluorescence (Genty et al., 1989). The leaf-level photosynthesis and fluorescence share the same energy source from photosynthetically active radiation (PAR) and there is a close linkage between fluorescence and carbon assimilation rate (Krause and Weis, 1991). Therefore, successful retrievals of SIF from satellites could provide a new promising way to quantify gross primary production (GPP) at regional to global scales (Frankenberg et al., 2011).

However, SIF is a very weak signal, the intensity of which is less than 2% of the background reflected sunlight (Frankenberg and Berry, 2018). Since the sensors used to retrieve SIF were not designed to estimate SIF at first, the satellite retrieval of SIF usually has a large footprint and big uncertainties (Frankenberg et al., 2014; Joiner et al., 2013). To reduce the large measurement noises, existing satellite SIF products are aggregated to coarse temporal and spatial resolutions. For instance, the footprint of Global Ozone Monitoring Experiment-2 (GOME-2) is 40km $\times$40km, and if averaged to global uniform grids, the finest spatial resolution of GOME-2 is 0.5°, usually at biweekly or monthly scale. SIF products generated from the Orbiting Carbon Observatory-2 (OCO-2) partially alleviated this problem via a much smaller footprint (1.3km 2.25km), higher signal-to-noise ratio and more numbers of observations per day (Frankenberg et al., 2014; Sun et al., 2018). However, because of the sparse sampling strategy and long revisit cycle, there are wide gaps between nearby swaths in OCO-2 SIF data. And OCO-2 SIF products are usually averaged to 1° 1° spatial resolution at a monthly scale. The coarse resolutions inhibit the direct comparison between SIF from satellite retrievals and GPP from flux towers due to the spatial inconsistency, which also hinders our understanding of the SIF-GPP relationship.

A SIF dataset with high spatiotemporal resolutions is needed, and efforts have been made to reconstruct global SIF. Gentine and Alemohammad (2018) applied machine learning to generate global SIF normalized by clear sky PAR based on Moderate Resolution Imaging Spectroradiometer (MODIS) reflectance and GOME-2 SIF. Zhang et al (2018) constructed contiguous SIF datasets from OCO-2 SIF retrievals also based on four bands of MODIS reflectance accompanied with solar zenith angle (SZA) data. The above studies show the possibility of using broadband reflectance and PAR information to estimate SIF. However, the time span of MODIS is limited (2000-present) and we couldn't obtain the SIF values beyond the temporal coverage of MODIS data. A contiguous SIF dataset with a longer time span and high spatial-temporal resolutions is required to greatly expand the application of SIF. The Advanced Very High Resolution Radiometer (AVHRR) remains an important data source for the long-term dynamics of the vegetated land surface as it provides the longest time-series of global satellite measurements starting from 1981 with daily temporal resolution (Zhang et al., 2015). However, only two bands of AVHRR, red band and near-infrared (NIR) band respectively, are located in the visible and near-infrared spectral range which provide the main information of vegetation and applicable for SIF predictions. It remains unknown if only red and NIR bands accompanied by PAR proxy could still successfully estimate SIF.

Therefore, our project tries to explore:

(1) Are red and NIR reflectances of MODIS sufficient to accurately estimate SIF?
(2) If the first step is feasible, could red and NIR reflectances of AVHRR successfully estimate SIF to generate longer SIF time series?

By exploring the above two key questions, we try to find a possible approach to reconstruct a global contiguous SIF dataset with longer temporal coverage.

2. Data and Methodology

2.1 Data Introduction

We tried to use reflectance data combined with SZA to predict SIF. SIF data were obtained from OCO-2 SIF products at 757nm from 2015 to 2016. The data preprocessing process of SIF is the same as that in Zhang et al (2018). The spatial resolution of SIF data is 0.05°. For reflectance data, we use red and NIR channels from MODIS and AVHRR respectively. The MODIS reflectance data is obtained from nadir bidirectional reflectance distribution adjusted reflectance (NBAR) product (MCD43C4

V006), which has the 0.05° daily resolution. The NBAR product collected the reflectance at a nadir viewing angle at local solar noon. The AVHRR reflectance data were from NOAA Climate Data Record (CDR) and could be downloaded at https://www.ncei.noaa.gov/data/avhrr-land-surface-reflectance/access/. The AVHRR reflectance also has a 0.05° daily resolution.

One thing should be noted is that we only use the SIF and reflectance under clear-sky conditions so that the relationship is not affected by cloud-related artifacts. And cos(SZA) was used as a proxy of the incoming PAR at top of the canopy. Since we did not consider the cloud and aerosol attenuation of the PAR, the predicted SIF was actually the “clear-sky instantaneous SIF”.

Two main issues were noticed when we compared the AVHRR reflectance against MODIS. The first challenge is the greatly decreased size of available samples. The original sample size of paired OCO-2 SIF and MODIS reflectance was about 1,600,000. However, after we extracted the overlap between OCO-2 SIF, MODIS, and AVHRR, the sample size was decreased to about 56,000 and the spatial coverage of overlap datapoint was limited to -40° ~ 45°, which may hinder the following training of the neural network. The second evident problem is the observed inconsistency between MODIS and AVHRR reflectance (Figure 2.1.1). A large number of AVHRR reflectance shows high values even though the MODIS reflectance is much smaller. Considering MODIS NBAR product should be more stable and consistent due to the removal of angle effects (Schaaf et al., 2002), the main cause for the inconsistency may be the relative worse data quality of AVHRR caused by incomplete removal of atmospheric effect, cloud and cloud shadow effect. Although we tried to use quality flags specified in original AVHRR dataset to delete the bad data points, the filtering does not well: it could not effectively remove the data points with abnormally high reflectance values, but deleted many data points within the normal range and greatly decreased the number of available samples (8282 samples left). Therefore, we tried to explore other methods to delete outliers in the AVHRR reflectance.

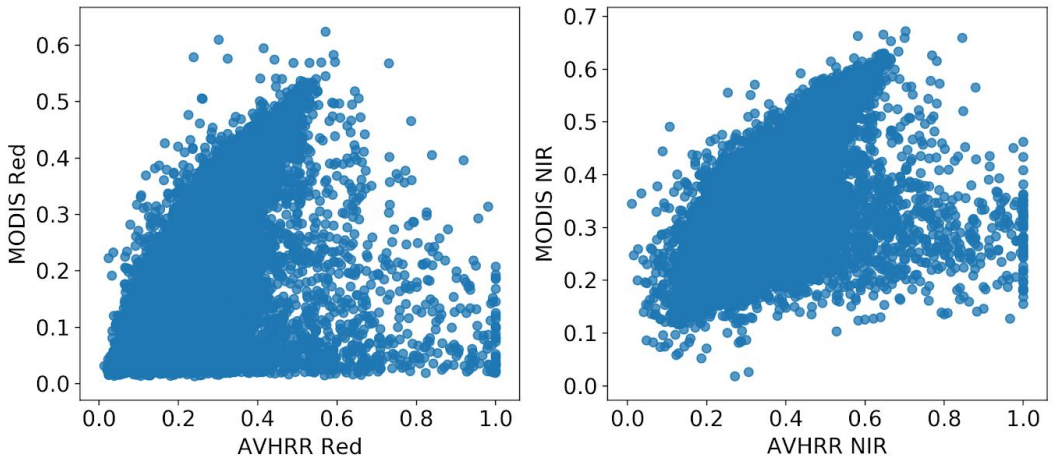

Figure 2.1.1 The comparison between MODIS and AVHRR reflectance (Number of data points: 56052).

2.2 Clustering and Outlier Removal

Because the AVHRR data does not have good quality and it is not helpful to apply the quality flag variable to filter the raw data, we proposed a few unsupervised clustering methods or outlier detection methods to split all the data into two groups. In one group, AVHRR reflectance data has high consistency with MODIS reflectance data (labeled as cluster '1', or regarded as inliers), and in the other group, AVHRR reflectance has a big bias when compared with MODIS data (labeled as cluster '0', or regarded as outliers). The bias could have been caused because the data was contaminated by some weather factors, such as clouds and fog, or some systematic shift.

We tried to cluster the data based on AVHRR & MODIS red and NIR bands. We also included AVHRR brightness temperature at 3.75 microns and brightness temperature at 12.0 microns (both are within IR bands) in clustering. In some cases, using AVHRR itself is sufficient to classify the inliers and outliers, while in some other cases, using both MODIS and AVHRR data can do a better job. Which variables give better cluster outcomes depends on the cluster methods to use. Our clustering results are elaborated as below.

2.2.1 K-means Clustering

K-means clustering is one of the most popular methods to use when grouping data. It aims to put the data with similar means into the same group and thus it is important to find the cluster centroids correctly. In high-dimensional problems, the algorithm is often presented as assigning objects to the nearest cluster by distance

(https://en.wikipedia.org/wiki/K-means_clustering, last access: 2019, December 19). We first applied k-means clustering on AVHRR & MODIS red and NIR bands (see Figure 2.2.1). When AVHRR and MODIS data is more consistent with each other, the data points should be more concentrated around the 1:1 line on Figure 2.2.1. It is obvious to tell that the two clusters are mistakenly grouped by k-means clustering.

This unsatisfying results given by k-means clustering is due to the fact that k-means clustering is a method that is very easy to be influenced by outliers. This algorithm updates the cluster centers by taking the average of all the data points that are closer to each cluster center. When there are outliers, this can affect the average calculation of the whole cluster. As a result, this could push the cluster center closer to the outliers. This is especially a problem in high-dimensional problems in that some outlier points (should be labeled as cluster '0') could have the same distance to the center of cluster '1' as some other good points in cluster '1' even though the outliers are not concentrated around the inlier cluster center, vice versa.

One way to improve the results of k-means clustering is to only use the AVHRR dataset itself, but include the two brightness temperatures retrieved at IR bands. Our paired data only ranges in low and mid latitudes (no snow or ice covered region). Thus these two variables contain the information of cloud, fog and so on in that when there are clouds over a region, the brightness temperature in infrared radiation bands would be much lower than that in the clear-sky condition, while the reflectance in red and NIR bands is super high.

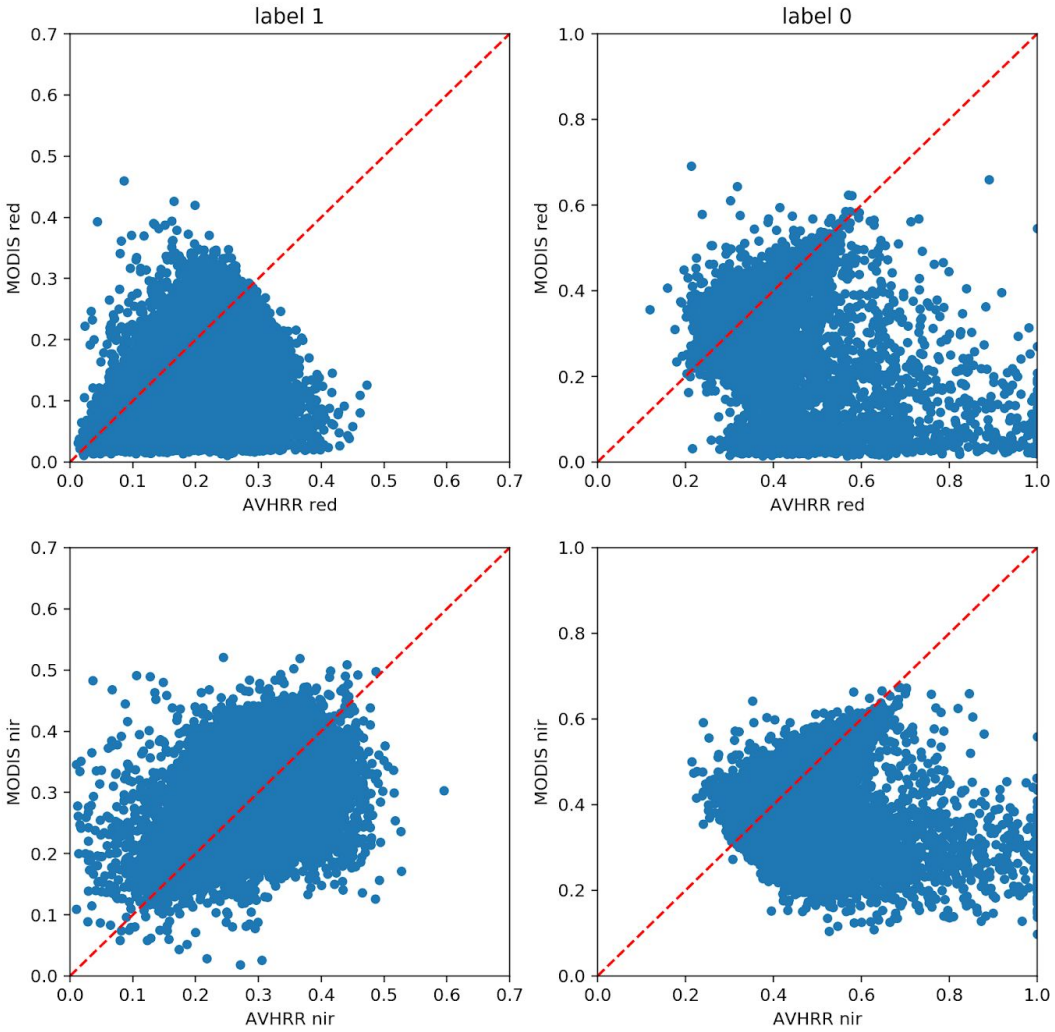

Figure 2.2.1 Two clusters grouped by k-means clustering using AVHRR & MODIS red and NIR bands.

Even though there is some improvement when using the new variables, k-means clustering has one more problem when applied in our case. There are many good-quality data samples mistakenly put into cluster ‘0’ (outlier cluster). That is because k-means clustering tends to find clusters of comparable spatial extent and shapes. While here we prefer to find a big dataset with as many good quality data points as possible along with a high sample density, and a dataset mostly with outliers or data with low consistency with MODIS, which should have a much smaller size and sparse sample distribution.

In order to continue to improve the clustering effects, we also performed One Class Support Vector Machine (OneClassSVM) (Schölkopf, B. et al., 2000), by which the

categories are divided by estimating a function (namely a separating hyperplane in data space) which is positive for data in one cluster and negative for data in the other cluster; the Density-based spatial clustering of applications with noise (DBSCAN) (Martin Ester et al., 1996), a density-based clustering algorithm whose core idea is that the outliers lie alone in low-density regions when compared with inners; robust covariance estimation by using Minimum Covariance Determinant estimator (MCD) and outlier detection (Devlin, S. J. et al., 1975; Rousseeuw, P. J., 1984; Hubert, M., & Debruyne, M. 2010), which selects a fixed fraction of observations (out of the whole dataset) whose covariance matrix has the lowest determinant and regards the left observations as outliers. In the end, we found the robust covariance estimation and outlier detection returns the most ideal results of clustering.

2.2.2 Robust Covariance Estimation and Outlier detection

As mentioned before, we carried out several different unsupervised clustering methods. The one that achieved the best performance is to assume the data has a Guassian distribution and estimate a robust matrix covariance by finding the minimum covariance determinant when the outlier fraction (10% is used in this study) is settled. The clustering results are shown in Figure 2.2.2 and Figure 2.2.3. Most sparse samples are marked out.

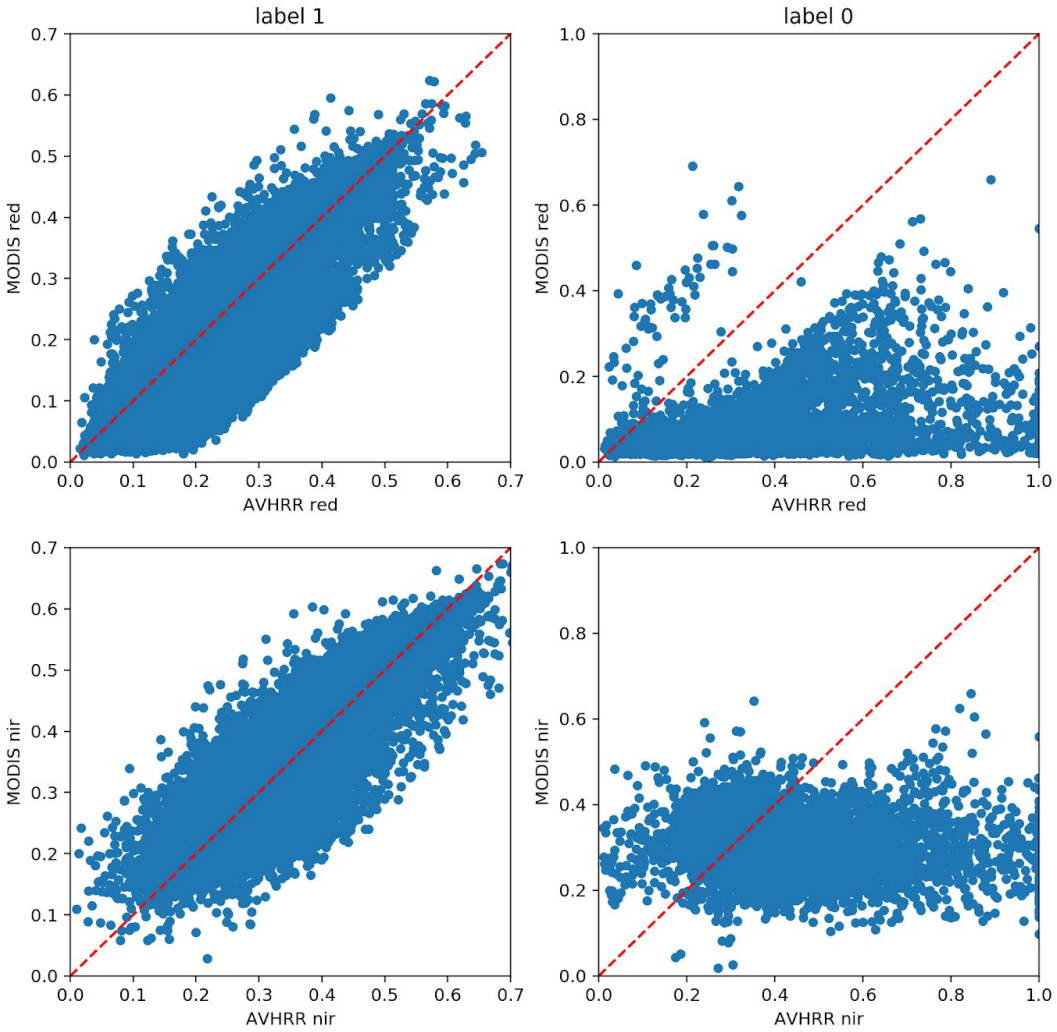

Figure 2.2.2 Two clusters (label 1: inliers, label 0: outliers) defined by Minimum Covariance Determinant estimator using AVHRR & MODIS red and NIR bands.

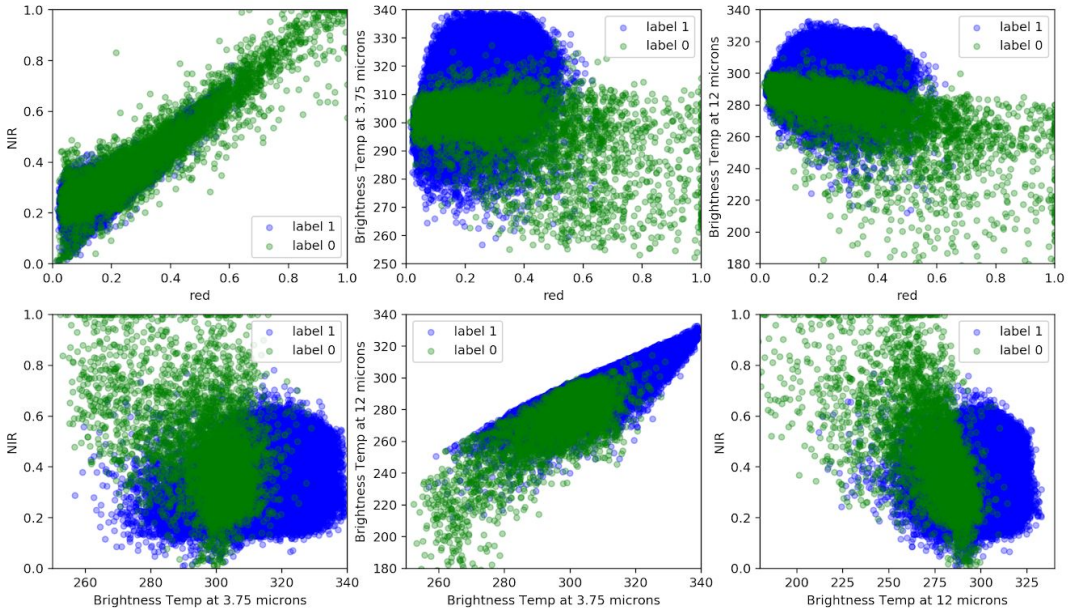

Figure 2.2.3 Scatterplots of different bands' information for two clusters (label 1: inliers and label 0: outliers) defined by Robust Covariance Estimation and Outlier Detection.

One advantage of this method is that it allows us to decide how much data should be outliers and leaves us with more data samples (90% of original paired data) to be further used in artificial neural networks. The other advantage is that the two clusters are reasonably separated in that almost all the data that promises consistency in MODIS & AVHRR correspondent bands are divided into cluster 1. The clustering outcome also generally makes sense in physics: almost all data with high reflectance but low brightness temperature are pointed out as outliers.

2.3 Artificial Neural Networks

We used feed-forward neural networks to predict SIF. We test networks with one (when data size is small) to two hidden layers and 128 neurons for each layer and found the number of neurons could affect the prediction results obviously when the data is abundant while have little impact on the prediction when the data size is limited. Thus, we chose the simplest structure (one hidden layer, 128 neurons) for SIF estimation. Since the neural network is not deep and there is no sign of overfitting, we didn't use any dropout or regularization methods during the training. The rectified linear unit (ReLU) was used as the activation function. We used the mean square error (MSE) as the loss function and Adam as the optimization algorithm. The learning rate was set to 0.0005. We trained the neural network for 40 epochs with a batch size of 1024.

To train the neural network, we used SZA and reflectance of red and NIR bands as predictors to estimate SIF. The above data were divided into two subsets: training and validation constituting 70, 30 % of all data. Before training, each predictor was normalized by its mean and standardised deviation. We trained the neural network multiple times for reflectance data preprocessed by different methods, and more detailed information for each training process could be found at the table 3.1.

3. Results and Conclusion

After the preprocessing of data (including pairing and clustering), we finally applied a feedforward neural network to predict SIF with satellite red and NIR bands reflectance. We tested the neural network performance for a few datasets (listed in Table 3.1) whose major difference lies in the data size and sample quality, and named the models from 1 to 4. There is no sign of overfitting in all models but the Model 1, 2, 4 is superior to Model 3 from the view of loss and validation coefficient of determination.

Table 3.1 Detailed information of the training processes for different input data

Model Name	Data Used	Sample Size	Loss	R^2 of validation
Model 1	MODIS	Without pairing: 1,675,514	0.034	0.860
Model 2	AVHRR	After pairing 56,052	0.071	0.635
Model 3	MODIS	After pairing: 56,052	0.032	0.841
Model 4	AVHRR	Only label = 1 48,885	0.041	0.699

In the following parts, we first confirmed that the red and NIR reflectance data of MODIS is sufficient to accurately estimate SIF and then showed the red and NIR reflectance data of AVHRR is able to successfully estimate SIF and makes it possible to generate longer SIF time series.

3.1 SIF Predictions Based on MODIS Reflectance

Figure 3.1.1 shows the comparison between OCO-2 SIF and predicted SIF based on MODIS red and NIR reflectances (Model 1). The loss of Model 1 is the lowest among all the models, with a value of 0.034. It also has a *validation* coefficient of determination (

R^2) (see Table 3.1) up to 0.86, which is comparable with the *training* coefficient of determination in the original paper of Zhang et al (2018) (around 0.8) using four MODIS bands to predict SIF. SIF has a skewed distribution that has a very high density in the low values. In the satellite observations, there are negative values of SIF, which is not reasonable. We are happy to see the neural network is able to calibrate the predicted SIF to have more positive values.

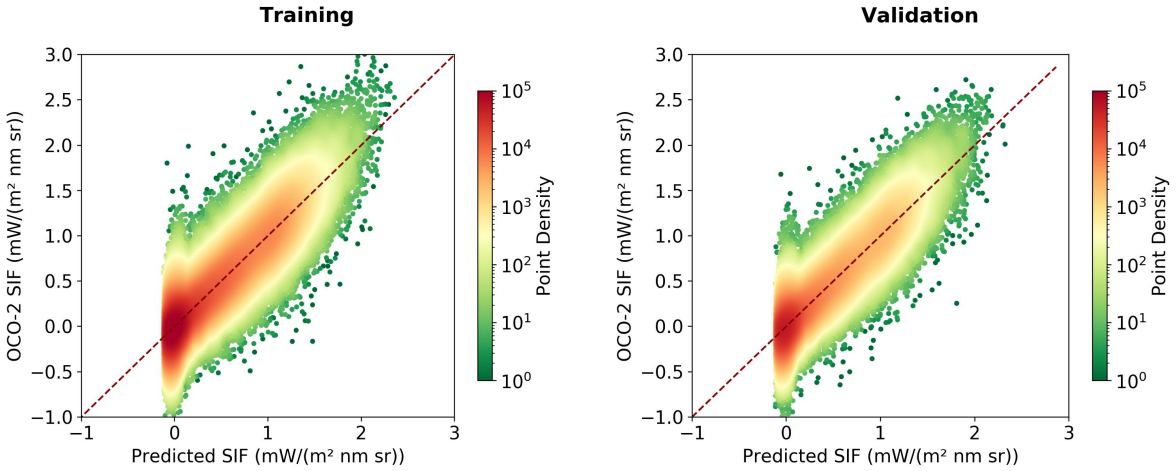

Figure 3.1.1 The comparison between OCO-2 SIF and predicted SIF based on MODIS red and NIR reflectance (Model 1).

In order to check the spatial distribution pattern is correctly predicted by the neural network, we compared the temporal mean of OCO-2 SIF and predicted SIF from Model 1 (Figure 3.1.2 shows the results of training dataset, the results of validation dataset is very similar). The predicted SIF from Model 1 generally captures the spatial characteristics as the observations. Model 1 underestimates SIF in Europe, coastal areas of East and South Asia, as well as the east coast of the US; overestimates SIF in the inner continent of Africa and East Asia.

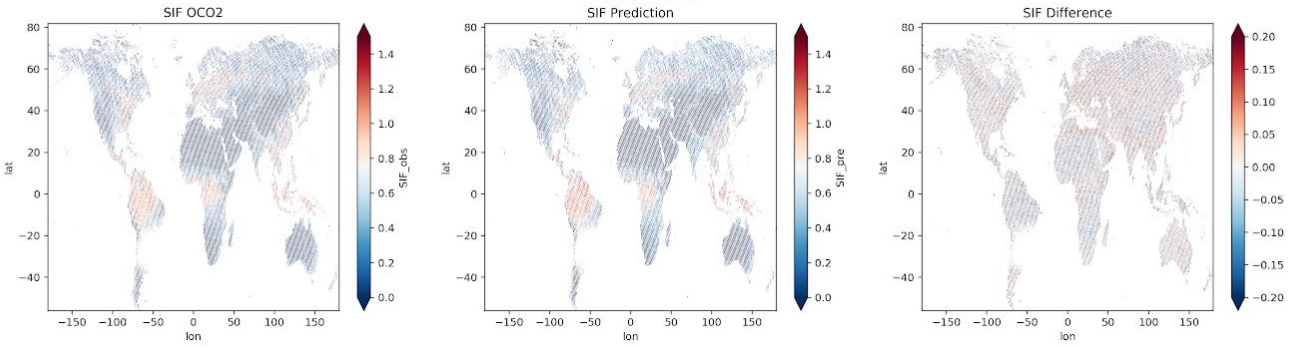

Figure 3.1.2 The comparison of temporal mean between OCO-2 SIF and predicted SIF based on MODIS red and NIR reflectance (Model 1)

Figure 3.1.3 shows the time series of OCO-2 SIF and Model 1 predicted SIF in two regions with area of 2x2 degrees, around Boston and in the center of Amazon respectively. The seasonal variation of SIF is clearly presented by the predicted SIF from Model 1 and the predicted values and observation values are very close.

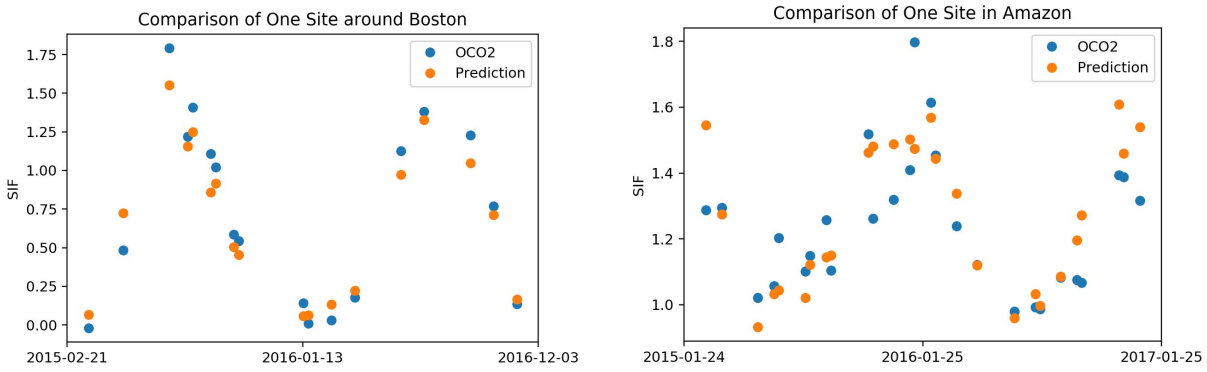

Figure 3.1.3 The comparison of temporal variations between OCO-2 SIF and predicted SIF based on MODIS red and NIR reflectance (Model 1). The extracted data points are from two 2-by-2-degree areas located around Boston, US and the center of Amazon respectively.

In conclusion, the red and NIR band reflectances of MODIS is able to predict SIF with a very high accuracy. The spatial and temporal variations are also captured. The next step is to test if the AVHRR two bands reflectance data is also sufficient to predict SIF well.

3.2 SIF Predictions Based on AVHRR Reflectance without Outlier Removal

The data sample size decreases drastically from about 1,675,514 to about 56,052 after purely matching the date, latitude and longitude of AVHRR, MODIS and OCO-2 data. We first directly applied the artificial neural network to the AVHRR red and NIR reflectances paired (Model 2). Table 3.1 and Figure 3.2.1 both illustrate the fact that the Model 2 prediction (validation $R^2 = 0.635,\ Loss\ = 0.071$) is less robust than that of previous Model 1. The predicted SIF sparsely (and almost randomly) spreads around the true values, especially for the high SIF values. For example, when the OCO-2 SIF is around 1.5, the predicted values range from 0 to 1.75, which means the neural network does not work properly.

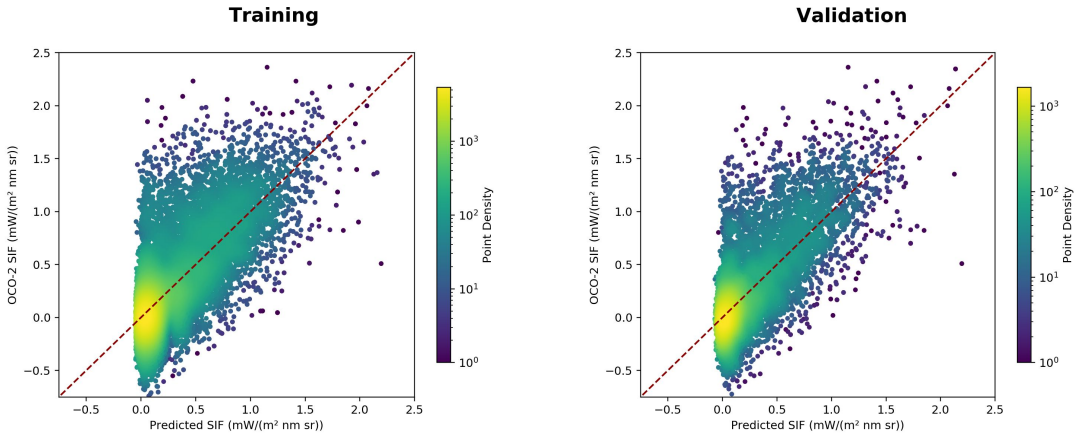

Figure 3.2.1 The comparison between OCO-2 SIF and predicted SIF based on AVHRR red and NIR reflectance data without outlier removal (Model 2).

In order to know whether the decreased dataset size or the inconsistency between MODIS and AVHRR data (namely the low quality samples in AVHRR) causes the relevant poor accuracy of the neural network, we carried out the same model structure on MODIS red and NIR reflectances again, but this time the MODIS dataset has exactly the same sample dates and locations as AVHRR dataset (Model 3). Table 3.1, Figure 3.2.1 and Figure 3.2.2 show that Model 3 has much better performance than Model 2. So, it is the poor quality of some samples in AVHRR dataset itself, instead of the small data size that caused the unsatisfactory results of Model 2. Thus, it is necessary for us to remove the outliers before the application of artificial neural network.

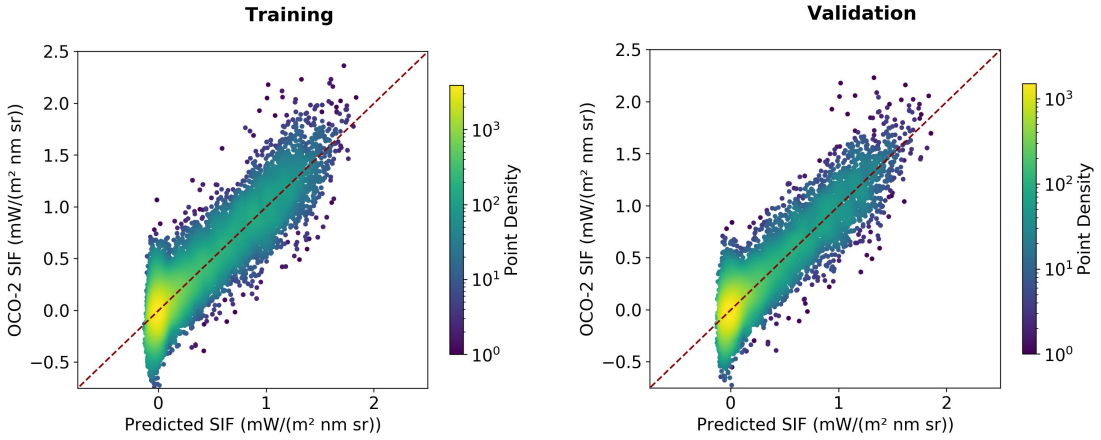

Figure 3.2.2 The comparison between OCO-2 SIF and predicted SIF based on MODIS red and NIR reflectance data (Model 3). Model 3 has exactly the same sample dates and locations as Model 2.

3.3 SIF Predictions Based on AVHRR Reflectance with Outlier Removal

In the report, we only show the results from the neural network using AVHRR red and NIR bands reflectances with outlier removal by Minimum Covariance Determinant estimator (MCD) and outlier detection as the input (Model 4). Although Model 4 does not have as high accuracy as the models using MODIS red and NIR reflectances as the input (see Table 3.1), it outperforms than Model 2. Figure 3.3.1 reveals the higher agreement of predicted SIF from Model 4 with the observations from OCO-2. This outcome indicates that removing outliers from AVHRR is helpful in enhancing the performance of the neural network.

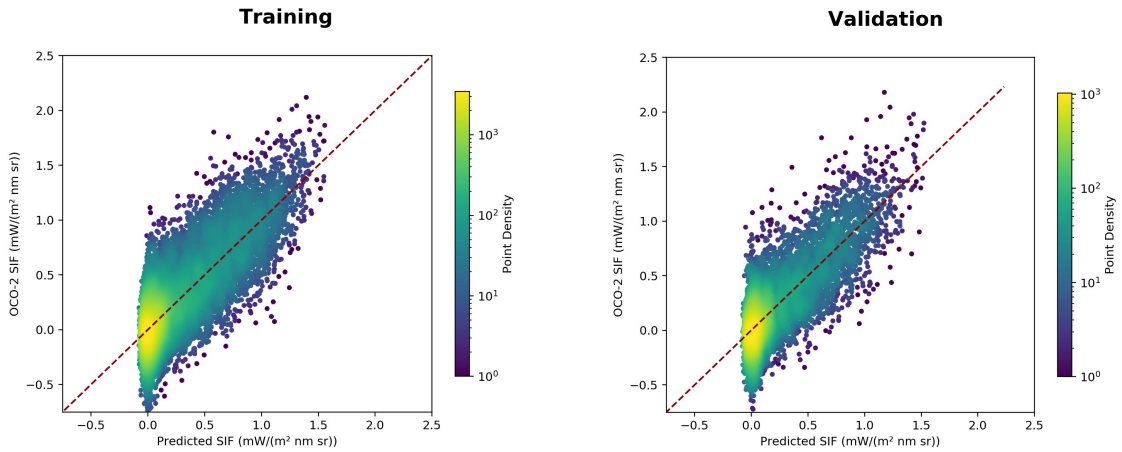

Figure 3.3.1 The comparison between OCO-2 SIF and predicted SIF based on AVHRR red and NIR reflectance data with outlier removal (Model 4).

In conclusion, two bands (red and NIR) reflectances from MODIS are sufficient to model and predict SIF using a simple artificial neural network and the predicted SIF is as good as that predicted from four bands reflectances (centered at 645, 858, 469 and 555 nm, respectively) and a more complicated neural network. However, when we substitute the MODIS data with AVHRR data, the validation coefficient of determination decreases by 25% and the loss is doubled. The consistency of predicted SIF and observed SIF also declines. Outlier removal from AVHRR could improve the model results to some extent and makes it possible for us to reconstruct the SIF series with a longer time period when MODIS data is not available. The model is hopefully to be further improved if the AVHRR outliers can be selected out with a better statistical method or if the AVHRR data can be calibrated closer to MODIS data. Researchers need to be very cautious with the data quality when using AVHRR dataset.

Appendix of Data and Codes

1. The AVHRR reflectance data were from NOAA Climate Data Record (CDR) and could be downloaded at: https://www.ncei.noaa.gov/data/avhrr-land-surface-reflectance/access/.

2. The example codes are stored in the below GitHub repository: https://github.com/YuHuang3019/EAEEE9305_Final_Project

3. The paired OCO-2 and MODIS data are stored at: https://drive.google.com/file/d/1JnV-22pm0MZatj_oKTUqiz85muq3nC_v/view?usp=sharing

Reference

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62(3), 531-545.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Frankenberg, C., & Berry, J. (2018). Solar induced chlorophyll fluorescence: origins, relation to photosynthesis and retrieval.

Frankenberg, C., Fisher, J. B., Worden, J., Badgley, G., Saatchi, S. S., Lee, J. E., . . . Kuze, A. (2011). New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity. Geophysical Research Letters, 38(17).

Frankenberg, C., O'Dell, C., Berry, J., Guanter, L., Joiner, J., Köhler, P., . . . Taylor, T. E. (2014). Prospects for chlorophyll fluorescence remote sensing from the Orbiting Carbon Observatory-2. Remote Sensing of Environment, 147, 1-12.

Genty, B., Briantais, J.-M., & Baker, N. R. (1989). The relationship between the quantum yield of photosynthetic electron transport and quenching of chlorophyll fluorescence. Biochimica et Biophysica Acta (BBA)-General Subjects, 990(1), 87-92.

Gentine, P., & Alemohammad, S. H. (2018). Reconstructed solar-induced fluorescence: A machine learning vegetation product based on MODIS surface reflectance to

reproduce GOME-2 solar-induced fluorescence. Geophysical research letters, 45(7), 3136-3146.

Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. Wiley interdisciplinary reviews: Computational statistics, 2(1), 36-43.

Joiner, J., Guanter, L., Lindstrot, R., Voigt, M., Vasilkov, A., Middleton, E., . . . Frankenberg, C. (2013). Global monitoring of terrestrial chlorophyll fluorescence from moderate-spectral-resolution near-infrared satellite measurements: methodology, simulations, and application to GOME-2. Atmospheric Measurement Techniques, 6(10), 2803-2823.

Krause, G., & Weis, E. (1991). Chlorophyll fluorescence and photosynthesis: the basics. Annual review of plant biology, 42(1), 313-349.

Porcar-Castell, A., Tyystjärvi, E., Atherton, J., Van der Tol, C., Flexas, J., Pfündel, E. E., . . . Berry, J. A. (2014). Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: mechanisms and challenges. Journal of experimental botany, 65(15), 4065-4095.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American statistical association, 79(388), 871-880.

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., ... & Lewis, P. (2002). First operational BRDF, albedo nadir reflectance products from MODIS. Remote sensing of Environment, 83(1-2), 135-148.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (2000). Support vector method for novelty detection. In Advances in neural information processing systems (pp. 582-588).

Sun, Y., Frankenberg, C., Jung, M., Joiner, J., Guanter, L., Köhler, P., & Magney, T. (2018). Overview of Solar-Induced chlorophyll Fluorescence (SIF) from the Orbiting Carbon Observatory-2: Retrieval, cross-mission comparison, and global monitoring for GPP. Remote Sensing of Environment, 209, 808-823.

Wikimedia Foundation. (2019, December 19). K-means clustering. Retrieved from https://en.wikipedia.org/wiki/K-means_clustering.

Zhang, X. (2015). Reconstruction of a complete global time series of daily vegetation index trajectory from long-term AVHRR data. Remote Sensing of Environment, 156, 457-472.

Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S., & Gentine, P. (2018). A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. Biogeosciences, 15(19), 5779-5800.