engineering proceedings

Proceeding Paper

Chloroophyll Estimation from Multivariate Regression Analysis and Deep Learning Using Remote Sensing Data [†]

Sriniketan Sridhar [1], Carlos del Castillo [2] and Vidya Manian [3]

[1] Southwestern Education Society, Mayaguez, PR, USA; 26017@sesolion.com
[2] The Ocean Ecology Laboratory, NASA, USA; caros.e.delcastillo@nasa.gov
[3] Department of Electrical and Computer Engineering, University of Puerto Rico, Mayaguez, PR 00681, USA; vidya.manian@upr.edu
* Correspondence:

[†] Presented at the 9th International Electronic Conference on Sensors and Applications, 1–15 November 2022; Available online: https://ecsa-9.sciforum.net/.

Abstract: The Orinico river is in Venezuela and flows into the Carribbean sea. The chlorophyll concentration in the Ocean delta changes due to the dust deposition from the Orinoco river which affects the primary productivity. The wet and dry deposition measurements are obtained from MERRA a NASA climate reanalysis of meteorology, atmospheric chemistry, land, ocean, and aerosols data on a broad range of weather and climate time scales and places. Researchers are not sure how wet and dry deposition from the Orinoco river affects the chlorophyll concentration in the ocean. Aerosol optical depth (AOD), dry and wet deposition data are obtained from MERRA. Altimetry data of the Orinoco river and Chlorophyll concentration data are also obtained from the Giovanni database from 2016 to March, 2022. Linear regression analysis of altimetry and chlorophyll concentration show that the later does not depend on the water levels. Univariate models for each of the parameters of AOD, wet, and dry deposition are done. Bivariate models are done adding one additional variable at a time, and finally a multivariate model is built for prediction of chlorophyll concentration. From the analysis, it is seen that the multivariate models have higher correlation between chlorophyll and the independent variables. Of all the variables wet deposition is a better predictor of chlorophyll concentration. A deep learning neural network architecture is developed for performing forecasting of chlorophyll concentration from past values.

Keywords:

Citation: Sridhar, S.; del Castillo, C.; Manian, V. Chlorophyll Estimation from Multivariate Regression Analysis and Deep Learning Using Remote Sensing Data. *Eng. Proc.* **2022**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Francisco Falcone

Published: 1 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Primary productivity refers to how energy is converted to organic substances. Primary productivity usually occurs due to the absorption of sunlight which is an important role to produce certain nutrients needed for the development of a plant. Primary productivity is usually measured by the increase of carbon dioxide or the output of oxygen. In the ocean a type of plant known as phytoplankton is one of the two ways primary productivity occurs in the ocean. Phytoplankton uses chlorophyll to absorb sunlight in this case use photosynthesis. When the phytoplankton's chlorophyll absorbs sunlight carbon dioxide is combined with water which produces oxygen. Primary productivity is sometimes at risk due to dust deposition by river flow. Dust is usually important for plant productivity due to it having important nutrients such as iron. Due to river flow scientists and researchers are skeptical due to the increase of dust deposition in the ocean. Researchers and scientists are asking the question if dust deposition affects chlorophyll levels in the Orinoco river. Chlorophyll prediction using deep learning has been done from satellite ocean color images [1]. These predictions are done only for current values and do not forecast chlorophyll concentration into the future.

In this paper, we present multivariate regression analysis for predicting chlorophyll based on water level, and dust. We then propose a deep learning architecture for chlorophyll forecasting using past levels of chlorophyll. Section 2 presents the methods, Section 3 the results and discussion, and Section 4 the conclusions.

2. Materials and Methods

Materials: The water altimetry, chlorophyll, aerosol optical depth, MERR II wet and dry deposition, data are obtained from the website Giovanni [2]. AOD MODIS 0.55 um which refers to the optical scattering of airborne atmospheric particles. MERR II Dry dep refers to dust deposition in the Orinoco River. MERRA II Wet dep refers to water deposition in the Orinoco river. River flow can affect the amount of dust and water that is deposited in the Orinoco river. The time series data are downloaded for dates from 07/04/02 to 2/1/22. The total number of data points in each time series is 153. The univariate and multivariate regression analysis are done in Microsoft Excel, and the deep learning LSTM architecture for chlorophyll forecasting is implemented in Matlab.

Methods: Linear regression, multivariate analysis, and deep learning neural network are used for prediction of chlorophyll level. Univariate analysis which is the simplest form of analyzing data since it only involves one variable. Prediction of chlorophyll is done from employing water flow, AOD, wet or dry deposition as one independent variable. Multivariate analysis which involved multiple forms of data sets and information was also used to create the linear regression. Univariate analysis uses the equation $y = mx + c$ where x refers to the independent variable, and y the dependent variable. The equation used for multivariate regression analysis is $y = b_1x_1+b_2x_2+b_3x_3+b_4x_4+ c$. We have used up to four independent variables, x_1 to x_4. For the univariate regression we used water flow, AOD, MERRA II dry dep or wet. LSTM or long and short-term memory is a deep learning neural network architecture commonly used for time series prediction or forecasting. LSTM is a type of Recurrent Neural Network (RNN) that uses a hidden state vector to represent context based on prior inputs and outputs, to be considered along with the current state when generating an output. The output vector is produced after a series of transformations of the input vector. Because this is advantageous in terms of network accuracy, RNNs are useful for analyzing time-series data [3]. Long Short-Term Memory (LSTM) Neural Networks are a type of RNN that attempts to solve the "vanishing gradient" (very small gradients don't allow distant input nodes to be considered) problem. The basic unit of an LSTM network is a memory cell, which has an input gate, an output gate, and a forget gate, which control information flow into the system. It contains a pointwise multiplication operation and a sigmoid neural net layer that assist the mechanism. The cell determines the fate of the information it holds. The memory cell is also called 'cell state' which maintains its state over time. This is determined by an independent set of weights pertaining to the memory cell, which are adjusted by gradient descent and backpropagation. Figure 1 shows the structure of the LSTM cell. LSTM has feedback connections, and it can process the entire sequence of data, apart from single data points such as images. The LSTM equations are given in [4]. In this research, LSTM is used to predict chlorophyll concentration based on past values. LSTM architecture is trained with 90% of data and 10% is used for prediction. The total amount of data used by the LSTM is 137 samples for training and the remaining 16 for testing. It is more accurate than regular models and can be used for analyzing and predicting multiple complex data sets. For the univariate regression, 108 samples are selected randomly for estimation, and the remaining samples for prediction.

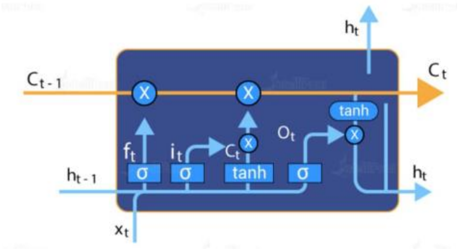


Figure 1. LSTM cell.

3. Experimental Results and Discussion

Figure 2 shows the output of linear regression with water flow into the Ornico river as the independent variable and chlorophyll as the dependent variable. The equation of the obtained line is: y = 0.0057x + 0.0509.

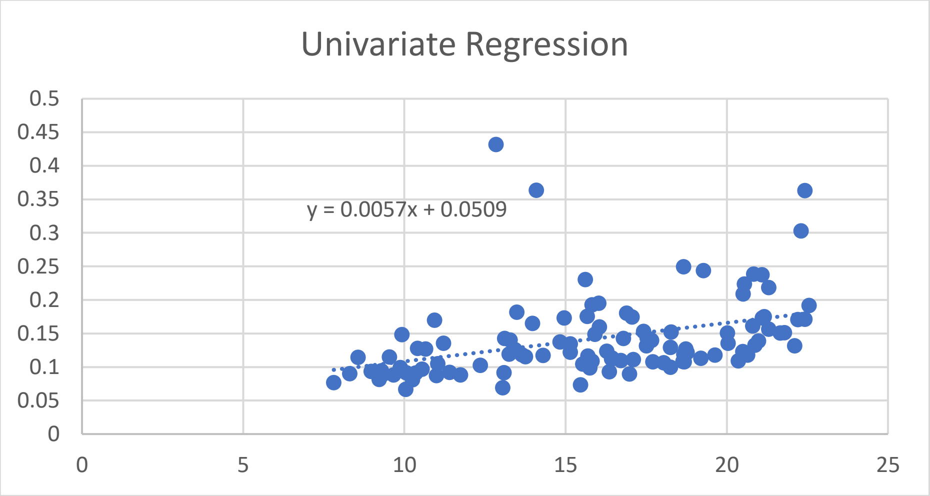
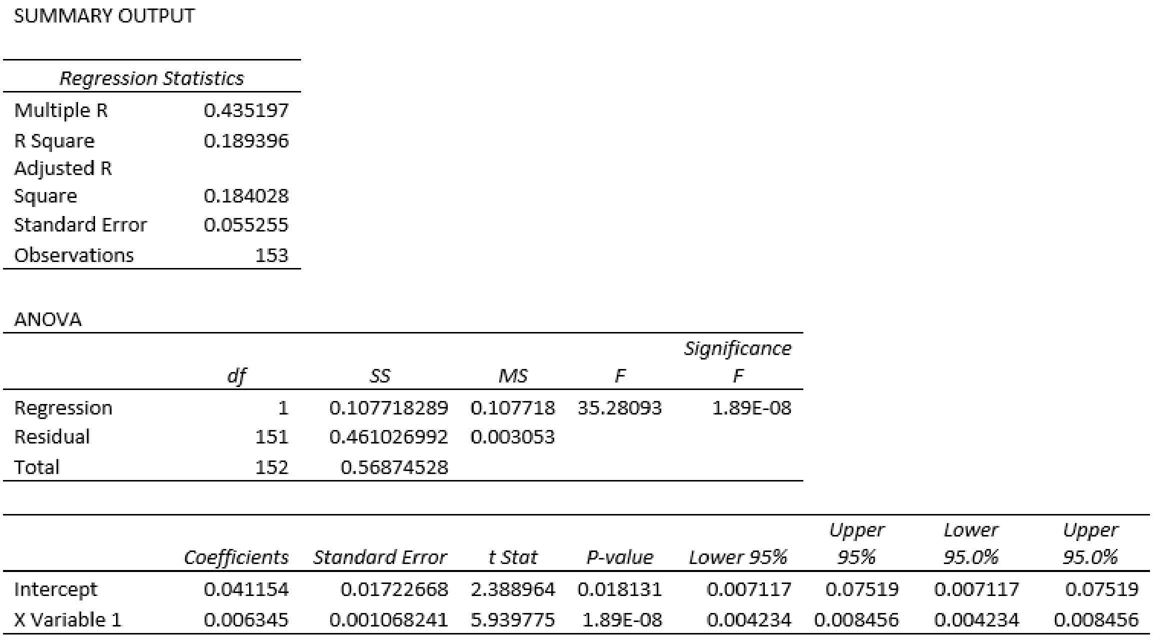


Figure 2. Univariate regression.

Figure 3 summarizes the regression and ANOVA analysis for the univariate model. The univariate analysis is done with each of the AOD, MERR II wet and dry deposition as independent variables. The summary outputs for each of them are given below in Figures 4–6. We used nearest neighbor interpolation for filling the missing river flow values for the regression analyses. Figure 7 summarizes the output from regression and ANOVA analysis for the multivariate model with four independent variables.



SUMMARY OUTPUT

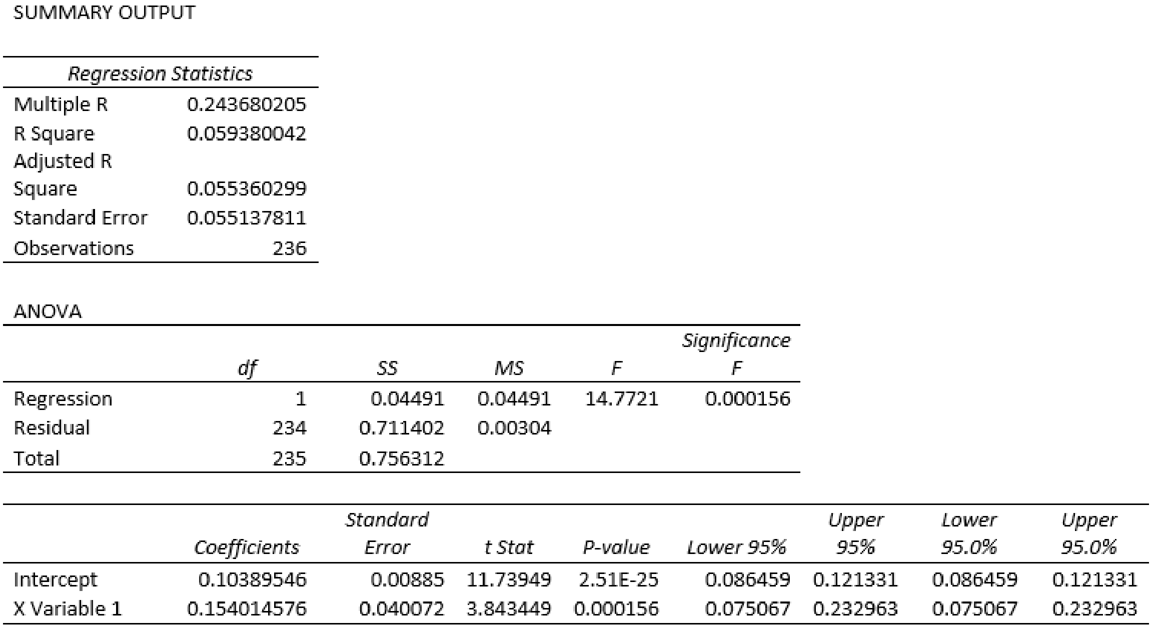
Regression Statistics	
Multiple R	0.435197
R Square	0.189396
Adjusted R Square	0.184028
Standard Error	0.055255
Observations	153

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.107718289	0.107718	35.28093	1.89E-08
Residual	151	0.461026992	0.003053		
Total	152	0.56874528			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.041154	0.01722668	2.388964	0.018131	0.007117	0.07519	0.007117	0.07519
X Variable 1	0.006345	0.001068241	5.939775	1.89E-08	0.004234	0.008456	0.004234	0.008456

Figure 3. Analysis summary for chlorophyll prediction from river flow.



SUMMARY OUTPUT

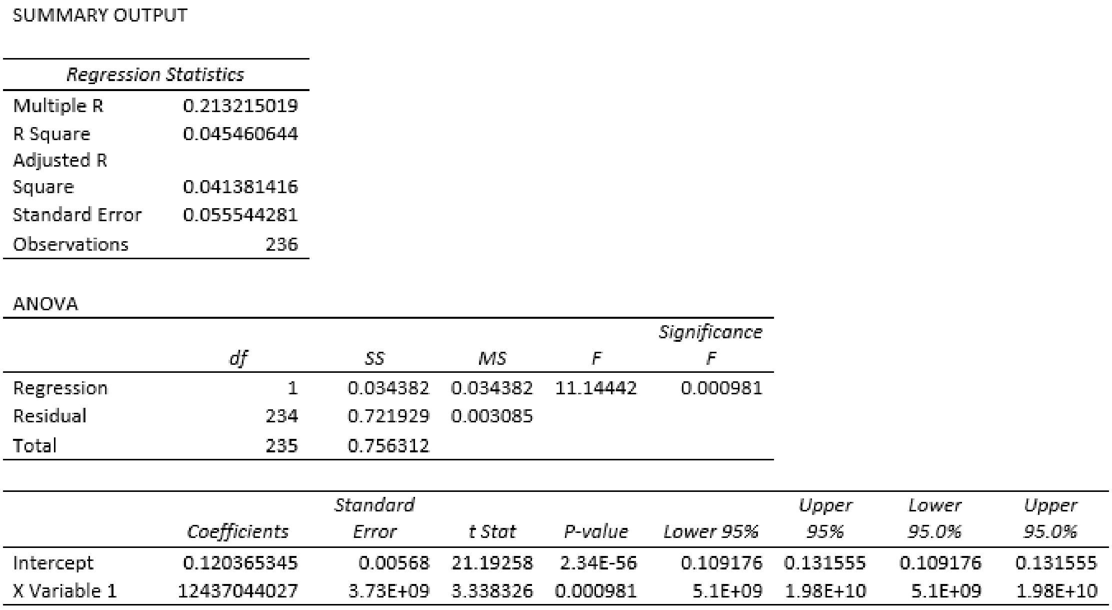
Regression Statistics	
Multiple R	0.243680205
R Square	0.059380042
Adjusted R Square	0.055360299
Standard Error	0.055137811
Observations	236

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.04491	0.04491	14.7721	0.000156
Residual	234	0.711402	0.00304		
Total	235	0.756312			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.10389546	0.00885	11.73949	2.51E-25	0.086459	0.121331	0.086459	0.121331
X Variable 1	0.154014576	0.040072	3.843449	0.000156	0.075067	0.232963	0.075067	0.232963

Figure 4. Analysis summary for chlorophyll prediction from AOD.



SUMMARY OUTPUT

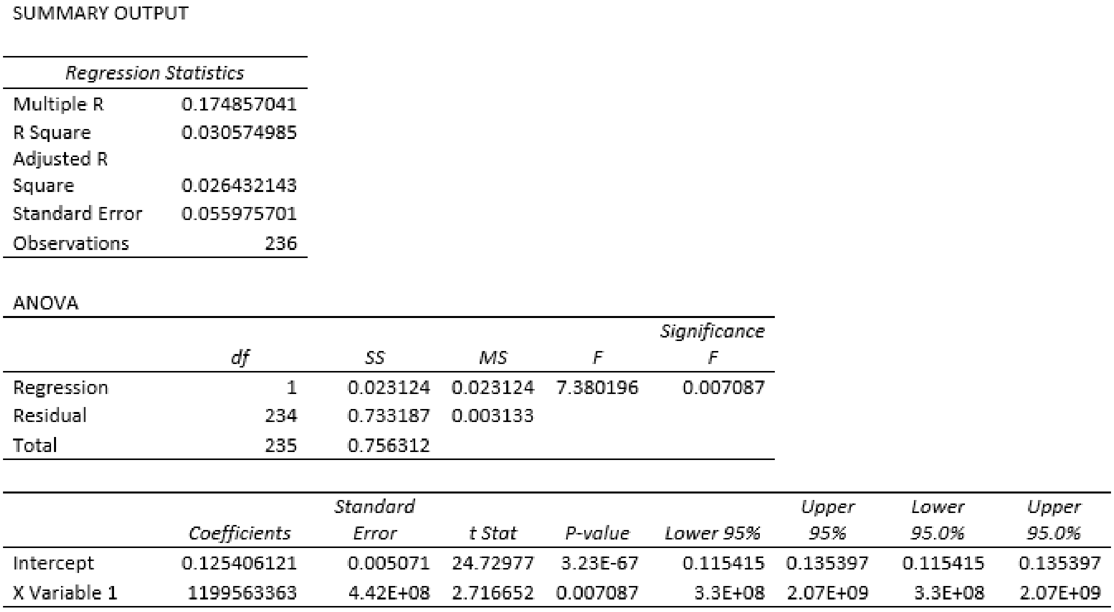
Regression Statistics	
Multiple R	0.213215019
R Square	0.045460644
Adjusted R Square	0.041381416
Standard Error	0.055544281
Observations	236

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.034382	0.034382	11.14442	0.000981
Residual	234	0.721929	0.003085		
Total	235	0.756312			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.120365345	0.00568	21.19258	2.34E-56	0.109176	0.131555	0.109176	0.131555
X Variable 1	12437044027	3.73E+09	3.338326	0.000981	5.1E+09	1.98E+10	5.1E+09	1.98E+10

Figure 5. Analysis summary for chlorophyll prediction from MERRA II dry deposition.



SUMMARY OUTPUT

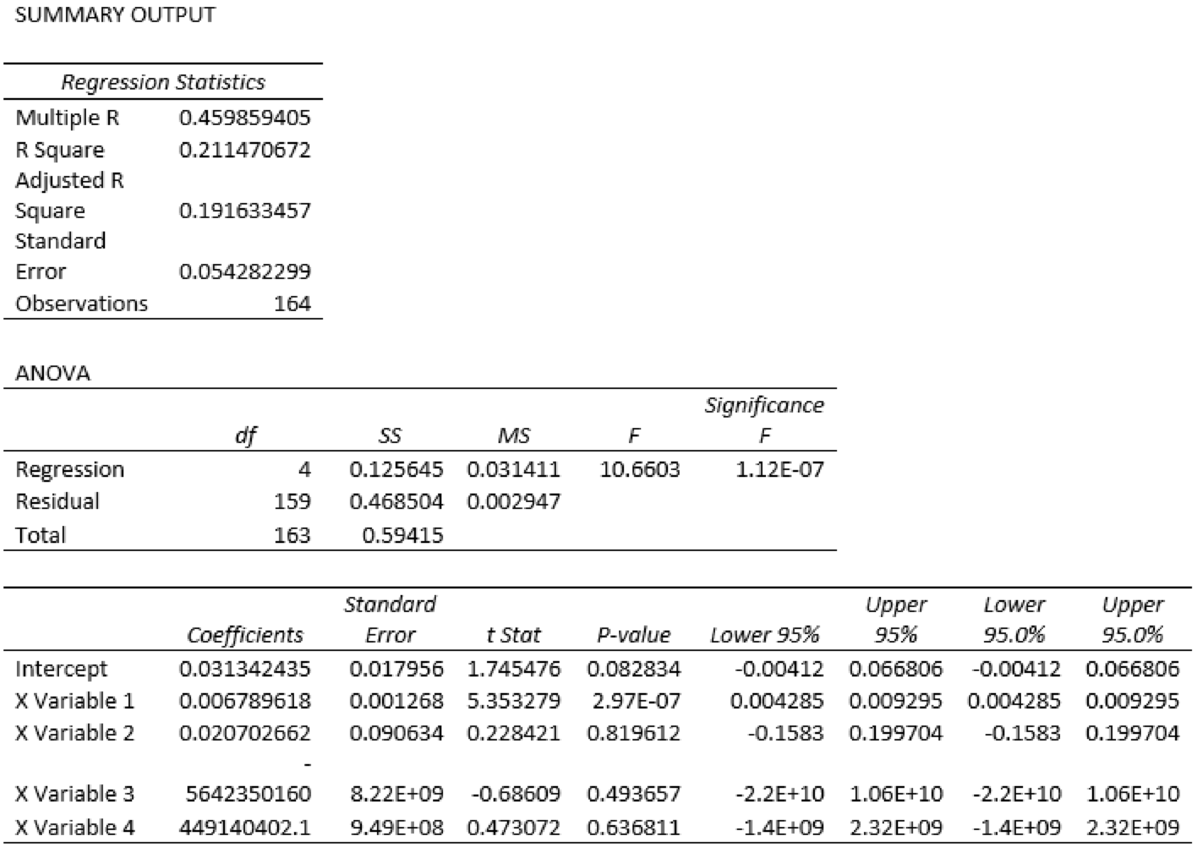
Regression Statistics	
Multiple R	0.174857041
R Square	0.030574985
Adjusted R Square	0.026432143
Standard Error	0.055975701
Observations	236

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.023124	0.023124	7.380196	0.007087
Residual	234	0.733187	0.003133		
Total	235	0.756312			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.125406121	0.005071	24.72977	3.23E-67	0.115415	0.135397	0.115415	0.135397
X Variable 1	1199563363	4.42E+08	2.716652	0.007087	3.3E+08	2.07E+09	3.3E+08	2.07E+09

Figure 6. Analysis summary for chlorophyll prediction from MERRA II wet deposition.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.459859405
R Square	0.211470672
Adjusted R Square	0.191633457
Standard Error	0.054282299
Observations	164

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	0.125645	0.031411	10.6603	1.12E-07
Residual	159	0.468504	0.002947		
Total	163	0.59415			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.031342435	0.017956	1.745476	0.082834	-0.00412	0.066806	-0.00412	0.066806
X Variable 1	0.006789618	0.001268	5.353279	2.97E-07	0.004285	0.009295	0.004285	0.009295
X Variable 2	0.020702662	0.090634	0.228421	0.819612	-0.1583	0.199704	-0.1583	0.199704
X Variable 3	-5642350160	8.22E+09	-0.68609	0.493657	-2.2E+10	1.06E+10	-2.2E+10	1.06E+10
X Variable 4	449140402.1	9.49E+08	0.473072	0.636811	-1.4E+09	2.32E+09	-1.4E+09	2.32E+09

Figure 7. Chlorophyll prediction from river flow, AOD, wet and dry deposition.

We can see that the adjusted R^2 value is 0.0264 for chlorophyll prediction using MERRA II wet deposition. We also combined two to four maximum independent variables that resulted in a standard error of 0.0538. Figure 8 shows each of the time series data. Figure 9 shows the time series values for chlorophyll used for training and prediction using the LSTM. Figure 9a shows the chlorophyll time series, and Figure 9b the predicted or forecast chlorophyll values. LSTMs are useful for making accurate predictions of a time series into the future. Figure 10 shows the training progress for the LSTM network. The network consist of 100 neuron units in the hidden layer, uses gradient descent for training with a learning rate of 0.005, and a piecewise learning rate. The maximum number of epochs is 250. The network consists of four layers: sequence input layer, LSTM layer, fully connected layer, and regression layer.

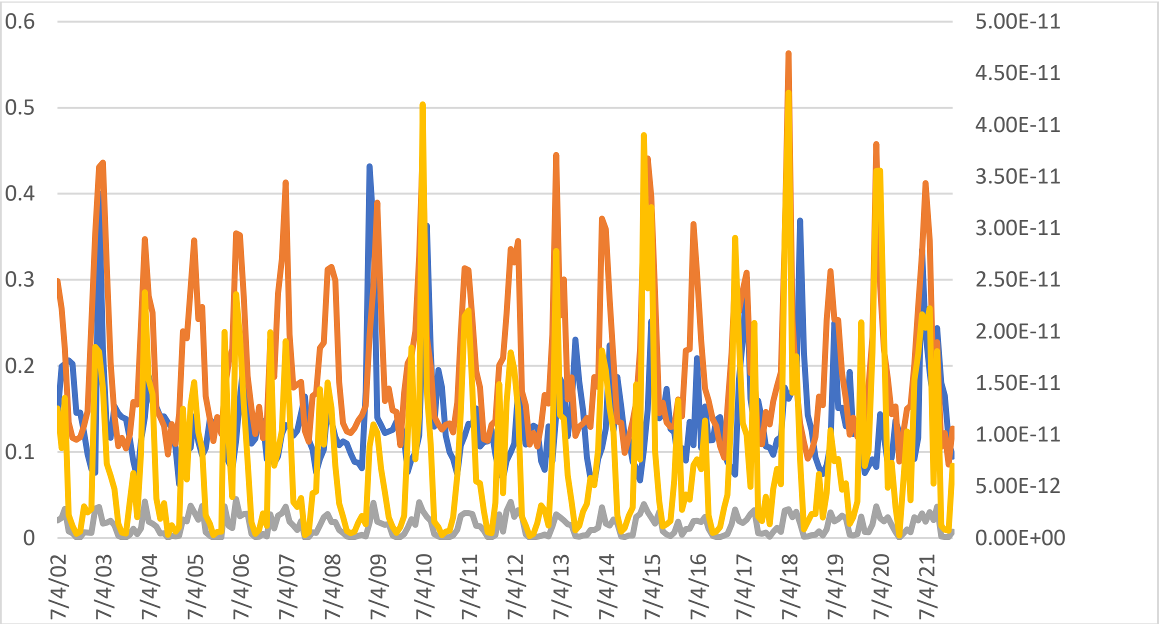
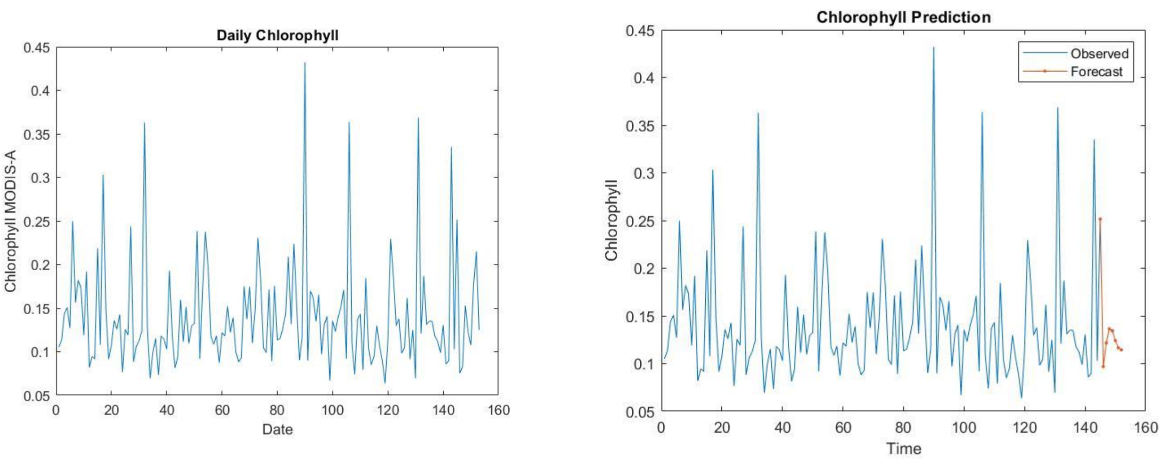
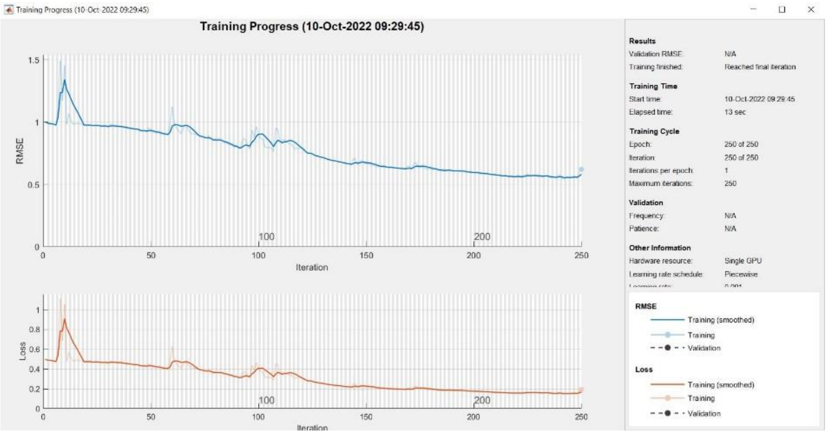


Figure 8. Time series data obtained from Giovanni (Blue—Chlorophyll MODIS -A, Orange—AOD MODIS, Grey—MERRA II Dry deposition, yellow—MERRA II Wet deposition).



(**a**) (**b**)

Figure 9. Chlorophyll time series forecasting (**a**) original times series, (**b**) time series with forecasted values.


Training Progress (10-Oct-2022 09:29:45)
Training Progress (10-Oct-2022 09:29:45)
RMSE
Loss
Iteration
1.5
1
0.5
0
0.8
0.6
0.4
0.2
50
100
150
200
250

Results
Validation RMSE: N/A
Training finished: Reached final iteration

Training Time
Start time: 10-Oct-2022 09:29:45
Elapsed time: 13 sec

Training Cycle
Epoch: 250 of 250
Iteration: 250 of 250
Iterations per epoch: 1
Maximum iterations: 250

Validation
Frequency: N/A
Patience: N/A

Other Information
Hardware resource: Single GPU
Learning rate schedule: Piecewise
Learning rate: 0.001

RMSE
Training (smoothed)
Training
Validation

Loss
Training (smoothed)
Training
Validation

Figure 10. Training curve for the LSTM.

Figure 11 gives the Root Means Square Error (RMSE) between the predicted and original values of chlorophyll concentration. The error is 0.045862 which is less than the standard error obtained by linear regression.

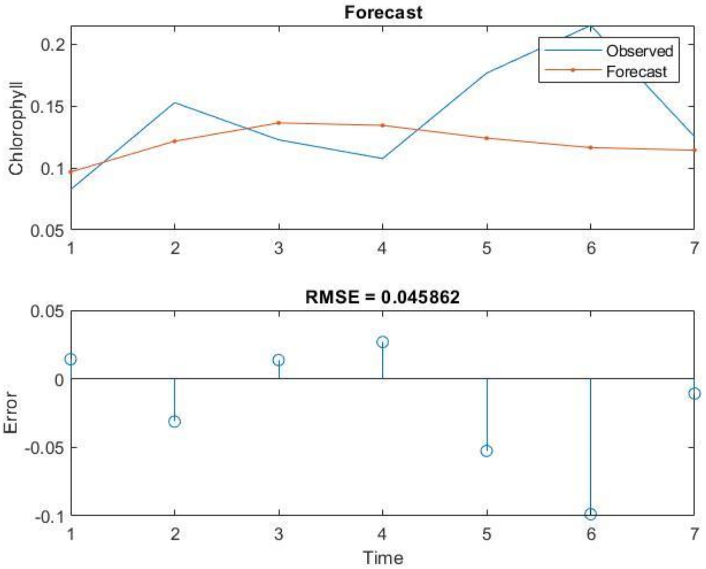


Figure 11. RMSE for chlorophyll prediction using LSTM.

4. Conclusions and Future Work

We have performed univariate and multivariate regression analysis for chlorophyll prediction from river flow, AOD, wet and dry depositions. A new LSTM algorithm is presented for chlorophyll forecasting from observed values of chlorophyll alone. The LSTM model is not affected by the correlation between the variables, and its predictions are based on past values of chlorophyll concentration. However, the model can be modified to include more variables for chlorophyll forecasting and further reduce the RMSE. Further, the architecture can be improved with optimal network design parameters.

Author Contributions:

Funding:

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest:

References

1. Jin, D.; Lee, E.; Kwon, K.; Kim, T. A deep learning model using satellite ocean color and hydrodynamic model to estimate chlorophyll-a concentration. *Remote Sens.* **2021**, *13*, 2003. https://doi.org/10.3390/rs13102003.
2. NASA Data, E. Giovanni The Bridge Between Data and Science v 4.37. Available online: https://giovanni.gsfc.nasa.gov/giovanni/(accessed on).
3. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* **2018**, *8*, 1–12. https://doi.org/10.1038/s41598-018-24271-9.
4. Sridhar, S.; Manian, V. Eeg and deep learning based brain cognitive function classification. *Computers* **2020**, *9*, 1–18. https://doi.org/10.3390/computers9040104.