Geostatistical approach for meteo-oceanographic variables evaluation at the Brazilian coast

Diogo de Jesus Amore[1]

[1]National Institute for Space Research
PO box 515 – ZIP 12227-010 - São José dos Campos - SP, Brasil
amore@dsr.inpe.br

Abstract. The spatial correlation among meteo-oceanografic variables within the Brazilian coast is here investigated. MODIS/*Aqua* Level 3 products for chlorophyll-a (chla), sea surface temperature (SST), and photosynthetic active radiation (PAR) were used for the geographically weighted regression (GWR) analysis performed within a 150-km buffer of the Brazilian coast for the time period ranging from 2002/07 till 2014/04. The variables correlation was between SST or PAR as the predictors and chla as the regressed variable. Both a GWR and a bayesian GWR (BGWR) were used for evaluating the variables. Colored matrices were plotted for displaying beta values, significance (mean squared errors), residuals, and *t-statistics*. R^2 were also computed for all months. Also, a ratio for the GWR beta estimates over the 95% confidence interval BGWR estimates was carried out. Results showed overall better R^2 for SST than for PAR regression but also showed better beta estimates for PAR than for SST in relation to BGWR beta significance range. Mostly, Northern regions of the Brazilian coast presented lower statistical significant values, and the months of July presented lowest GWR beta values and best significance, and January presented the highest beta values and worst significance, April, and October presented highly variable results.

Palavras-chaves: GWR, Geostatistics, SST; PAR, chlorophyll-a

1. Introduction

Environmental variables such as sea surface temperature (SST), photosynthetic active radiation (PAR) are crucial parameters for the comprehension of ocean primary production (OPP). OPP itself acts as proxy for fishery activity dynamics as well as biological activity quality indicator. Economic exclusive zones (EEZ) are a 200-mile buffer zone of highly economic importance for coastal states mostly due to fishery activity. In such scenario, OPP and ocean environmental variables dynamics comprehension are most important for any EEZ country aiming to properly maintain jurisdiction over its economically productive coast.

Geostatiscal techniques are frequently used for parameter estimation wherever data are not sufficiently available for the parameters examination of the study area. Moreover, spatial regression such as geographic weighted regression (GWR) can be useful for the correlation assessment of two or more variables. That way one can tackle issues of spatial auto-correlation which are encountered in a dataset, as well as be able to investigate the spatial distribution and significance of prediction parameters such as beta coefficients. Moreover, a Bayesian GWR approach can further evaluate the estimation significance of the beta coefficient by simulating a probability distribution function for each sample. Thence, a more statistically robust parameter estimation could be conducted.

Therefore, this study aimed to correlate SST and PAR with chlorophyll-*a* (chla) via GWR techniques and assess whether GWR provides a theoretical framework such that PAR presents a more robust statistical spatial correlation with chla than presented in the literature, which usually indicates SST as a better predictor. Also, it aimed to investigate how the GWR estimated parameters behave spatio-temporally over the study area. Lastly, this study aimed to evaluate the efficacy of the BGWR in relation to GWR and whether it can provide statistical improvement over the latter.

2. Materials and Methods

2.1. Study Area

The area considered for this study is the Brazilian coast ranging from the Equator latitude to the southernmost part of Rio Grande do Sul (30S – 50W) as shown in Figure 1. In coastal areas and in the Amazon basin fish consumption is much higher than in inland regions. Fish consumption has increased substantially in recent years as a result of massive campaigns to promote fish consumption. Annual per caput consumption was estimated at about 8.9 kg in 2010, with a rapid increase from the level of 6.0 kg in 2005 and earlier. In such scenario, it is crucial the comprehension of SST, PAR, and chla in the Brazilian coast as OPP-forcing parameters which ultimately control fishery production. One notes how the northernmost region of the Brazilian coast is not included in the analysis. These pixels were excluded because they were not common pixels among all months, which is a common feature in Amazonian regions due to cloud presence. Also, at the southernmost region of the study area, the Uruguayan coast was included in the analysis due to important spatial variations that might occur in those regions.

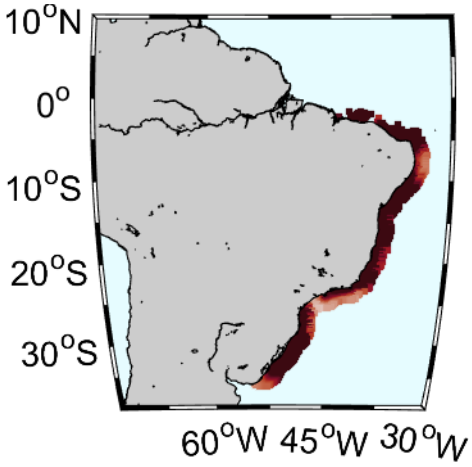

Figure 1. Spatial extent considered in this study represented by the red-shaded color.

2.2. MODIS/*Aqua* Level 3 products

R_{rs} data were acquired from the *MODerate resolution Imaging Spectroradiometer* (MODIS) sensor onboard *Aqua* satellite, available at the *OceanColor data* (http://oceandata.sci.gsfc.nasa.gov) website, referring to the time period spanning 2002/07 a 2014/04. Level 3 (L3) data products were acquired for each parameter with a 9-km spatial resolution. The products did not need any further processing because they already were in units of interest for this study.

Prior to the application of the GWR model, the data statistics were investigated for via the Shapiro-Wilk test for normality, Breusch-Pagan test for heterocedasticity. Also, Moran' I was applied on the entry dataset in order to evaluate the spatial auto-correlation, and whether the GWR would be a better choice of correlation analysis than an Ordinary Least Squares (OLS). A further test on the latter assumption was the application of an F-test on the residuals for variance equality between an OLS and the GWR as to investigate whether to carry on with GWR rather than OLS.

2.3. Geographically weighted regression (GWR)

The GWR model extends the traditional regression framework by allowing parameters to be estimated locally so that the model can be expressed as

$$Y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \quad i = 1, \ldots, n$$

where (u_i, v_i) denotes the coordinates of the point i in space, $\beta_0(u_i, v_i)$ represents the intercept value, and $\beta_k(u_i, v_i)$ is a set of values of parameters at point i. Unlike the 'fixed' coefficient estimates over space in the global model, this model allows the parameter estimates to vary across space and is therefore likely to capture local effects (Huang, 2010).

To calibrate the model, it is assumed that the observed data close to point i have a greater influence in the estimation of the $\beta_k(u_i, v_i)$ parameters than the data located farther from observation i. The estimation of parameters $\beta_k(u_i, v_i)$ is given by

$$\hat{\beta}(u_i, v_i) = [X^T W(u_i, v_i) X\,]^{-1} X^T W(u_i, v_i) Y$$

where $W(u_i, v_i)$ is an n x n matrix whose diagonal elements denote the geographical weighting of observation data for observation i, and the off-diagonal elements are zero. The weight matrix is computed for each point i at which parameters are estimated.

2.4. Weighting matrix specification

The weight matrix in GWR represents the different importance of each individual observation in the data set used to estimate the parameters at location i. In general, the closer an observation is to i, the greater the weight. Thus, each point estimate i has a unique weight matrix.

In essence, there are two weighting regimes that can be used: fixed kernel and adaptive kernel. For the fixed kernel, distance is constant but the number of nearest neighbors varies. For the adaptive kernel, distance varies but the number of neighbors remains constant. The most commonly used kernels are Gaussian distance decay-based functions (Fotheringham et al.2002) and which has been used in this study was

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{h^2}\right)$$

where h is a non-negative parameter known as bandwidth (in this study 0.36), which produces a decay of influence with distance and d_{ij} is the measure of distance between location i and j. Using point coordinates $(x_i\ ,\ y_i)$ and $(x_j,\ y_j)$, the distance is usually defined as a Euclidean distance

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

To avoid exaggerating the degree of non-stationarity present in the areas where data are sparse or mask subtle spatial non-stationarity where the data are dense (Paez et al. 2002), adaptive weighting functions are used to change the kernel size to suit localized observation patterns. Kernels have larger bandwidths where the data points are sparsely distributed and smaller ones where the data are plentiful. By adapting the bandwidth, the same number of nonzero weights is used for each regression point i in the analysis. For example, the adaptive bi-square weighting function is the following

$$W_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h_i}\right)^2\right]^2, & \text{if } d_{ij} < h_i \\ 0, & \text{otherwise} \end{cases}$$

where h_i stands for the bandwidth particular to location i.

2.5. Bayesian Geographically weighted regression (BGWR)

Bayesian GWR (BGWR) consists in applying the concepts of Bayes theorem into the GWR modeling. We used Gibbs sampling to estimate the BGWR model. This approach is particularly attractive in this application because the conditional densities all represent know distributions that are easy to obtain. To implement the Gibbs sampler we need to derive and draw samples from the conditional posterior distributions for each group of parameters, β_i,σ,δ, and V_i in the model. Let $P(\beta_i|\sigma,\delta,V_i,\gamma)$ denote the conditional density of β_i, where γ represent the values of other β_i $j \neq i$. Using similar notation for the other conditional densities, the Gibbs sampling process can be viewed as follows:

1. start with arbitrary values for the parameters $\beta_i^0, \sigma^0, \delta^0, V_i^0, \gamma^0$
2. for each observation $i = 1, \dots, n$,
   a) sample a value, β_i^1 from $P(\beta_i|\sigma^0, \delta^0, V_i^0, \gamma^0)$
   b) sample a value, V_i^1 from $P(V_i|\beta_i^0, \sigma^0, \delta^0, \gamma^0)$
3. use the sampled values β_i^1, $i = 1, \dots, n$ from each of the n draws above to update γ^0 to γ^1.
4. Sample a value, $\sigma^1$from $P(\sigma|\delta^0, V_i^1, \gamma^1)$
5. Sample a value, $\delta^1$from $P(\delta|\sigma^1, V_i^1, \gamma^1)$
6. Go to step 1 using $\beta_i^1, \sigma^1, \delta^1, V_i^1, \gamma^1$ in place of the arbitrary starting values.

The sequence of draws outlined above represents a single pass through the sampler, and we make a large number of passes to collect a large sample of parameter values from which we construct our posterior distributions. In this study a total of 500 draws were applied to the dataset. The BGWR modelling rely on the compact statement of the BGWR model expressed in the equation below to facilitate presentation of the conditional distributions that we rely on during the sampling:

$$\tilde{y}_i = \tilde{X}_i \beta_i + \varepsilon_i$$

Where the definitions of the matrices are:

$$\tilde{y}_i = W_i^{1/2} y$$

$$\tilde{X}_i = W_i^{1/2} X$$

Figure 2 summarizes the BGWR approach via Gibbs´s sampling Markov Chain Monte Carlo (MCMC). In short, the sampler does pseudo-random intelligent guesses for the different parameters. Then, the sampler tells a prior function a given parameter sampled value, then, the prior-calculation module throws back a prior probability for that parameter. Also, the sampler tells a preconceived data model (e.g., in this study a Gaussian model) a given parameter value, then, the model combined with the input data throws back the likelihood. The product of the prior probability and the likelihood gives us the posterior. Then, small jumps are taken, for every iteration, aiming towards higher-probability posterior distributions where convergence occurs, and thus, the distribution for that given location for that given parameter is achieved. In the case of this study, the distribution of betas was evaluated.

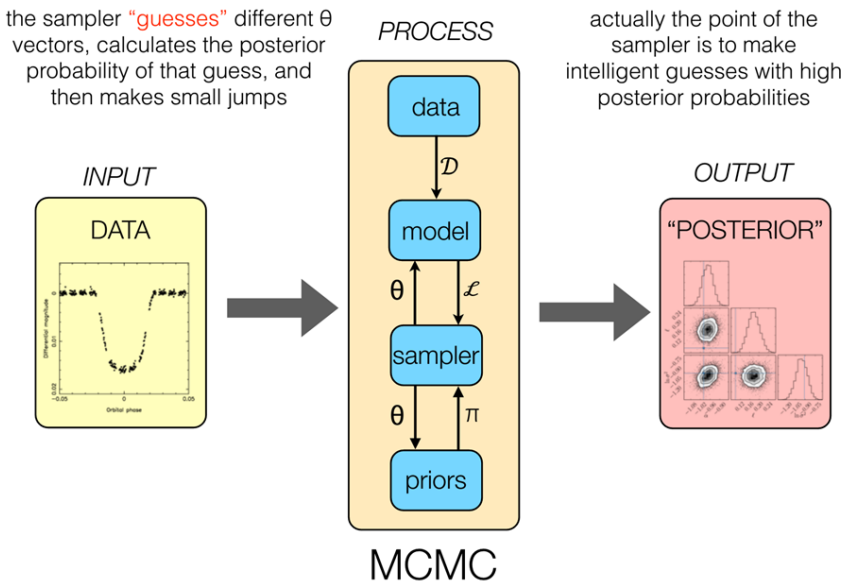

Figure 2. Spatial extent considered in this study represented by the red-shaded color.

1. Results and Discussion

Shapiro-Wilk test for PAR/SST/chla rejected the null hypothesis that the distribution is normal with unspecified mean and variance at .05. Breusch-pagan test showed positive test for heterocedasticity (i.e., very low p-values). Thence, a boxcox transform was applied to the dataset in order to normalize it. The Boxcox lambda value indicates the power to which all data should be raised. Figure 3 presents the results of the boxcox lambda for each variable in this study. It is noted that for most months evaluated in this study, chla presented a reciprocal square root transform (e.g., average lambda value of -0.3), SST and PAR presented, mostly, a log transform (e.g., lambda value of zero) with some extremer value representing a power transform (Figure 3).

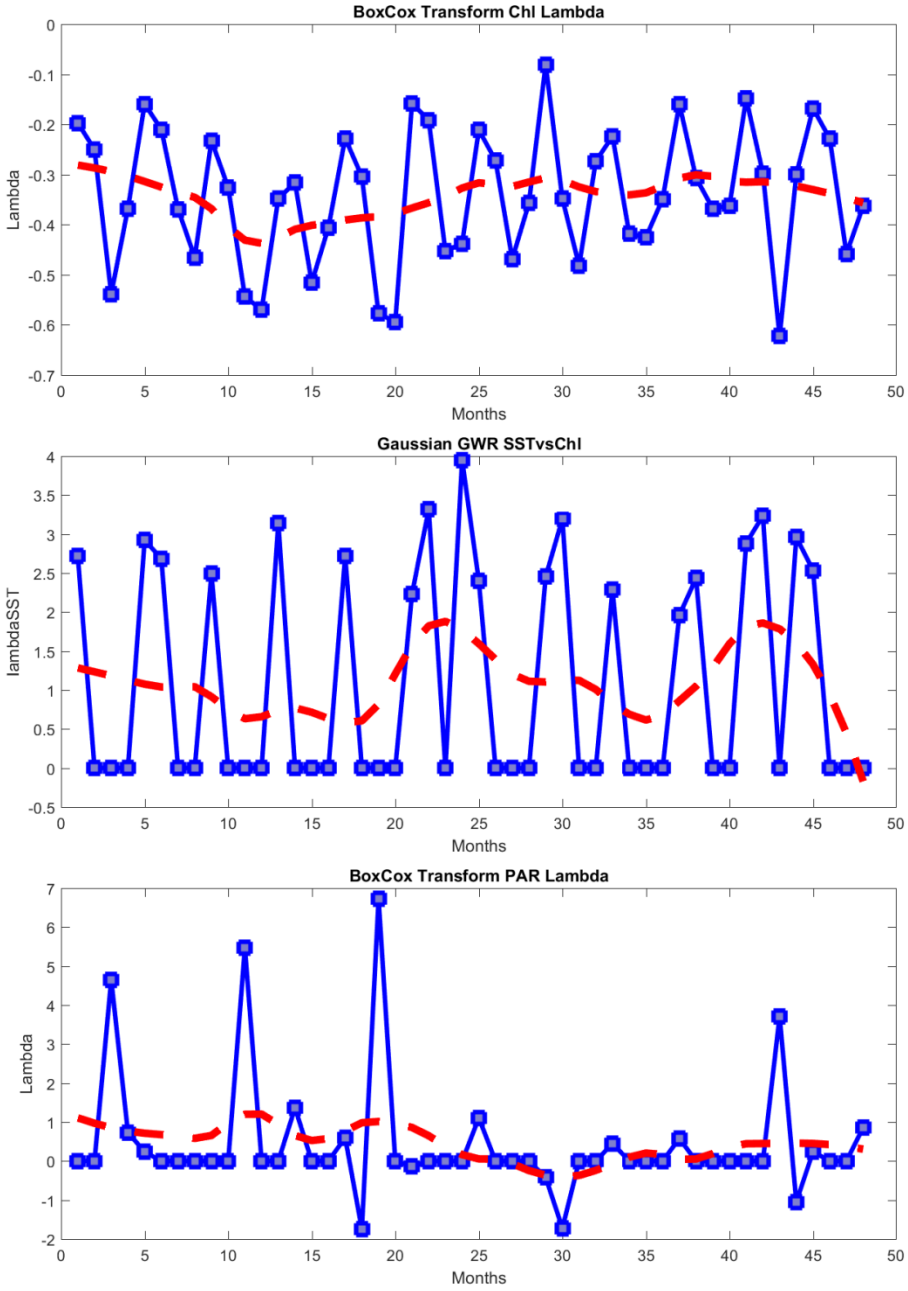

Figure 3. Boxcox lambda for each variable (chla, SST, PAR) for all 48 months considered in this study.

Moran's I test was applied to the dataset in order to verify spatial auto-correlation, which was confirmed from the test results (0.95-0.97). The I-Stat value ranged from 123 to 126 which is far larger than the threshold test of 1.96, meaning the null hypothesis of null spatial correlation could be rejected at a .05 significance. The F-test results for the GWR and OLS residuals were 1, implying there is sufficient evidence that they come from different distributions and the null hypothesis can be rejected. If there were to be no evidence to reject this hypothesis, it would mean

that an OLS regression model would be an adequate descriptor of the data. Therefore, in this study, GWR presented itself as a better fitting model than OLS.

The spatial distribution of beta coefficients, significance (Mean Squared Error, MSE), residuals, and *t-statistic* (measure of signal to noise ratio) for an example month are depicted in figures 4 and 5 for chla versus SST, and chla versus PAR, respectively. That way one can have a general idea of the spatial distribution of results as a function of latitude and longitude within the study area. However, in order to make the comparison of all months regressed via GWR most readable more concisely, matrix-like graphs were plotted, and are depicted from Figures 6 through to Figure 11.

Figure 6 depicts the beta coefficient results. It is important to clarify the structure of the matrix presented in Figures 6 to 11 as containing 5702 rows, and 48 columns. The rows represent all pixels accounted for in the study starting from the northernmost part of the study area and the columns represent the 48 eight seasonal months evaluated. The row dimension of the matrix considered the spatial distribution of the data in following manner: for each latitude every pixel along the longitude dimension was selected and place in the first n rows of the matrix; then, for the second latitude the same procedure was applied placing the next range of pixels below the first range, and this procedure was performed till all pixels were accounted for. Overall, the matrix pixel number range in relation to map geographical degree location are approximately as follows: pixel 0 to 2000 accounts for the 0° to 12-14° south latitude, pixel 2001 to 3500 accounts for the 12-14° to 22-24° south latitude, and pixels 3500-5702 accounts for 22-24° to 35-37° south latitude.

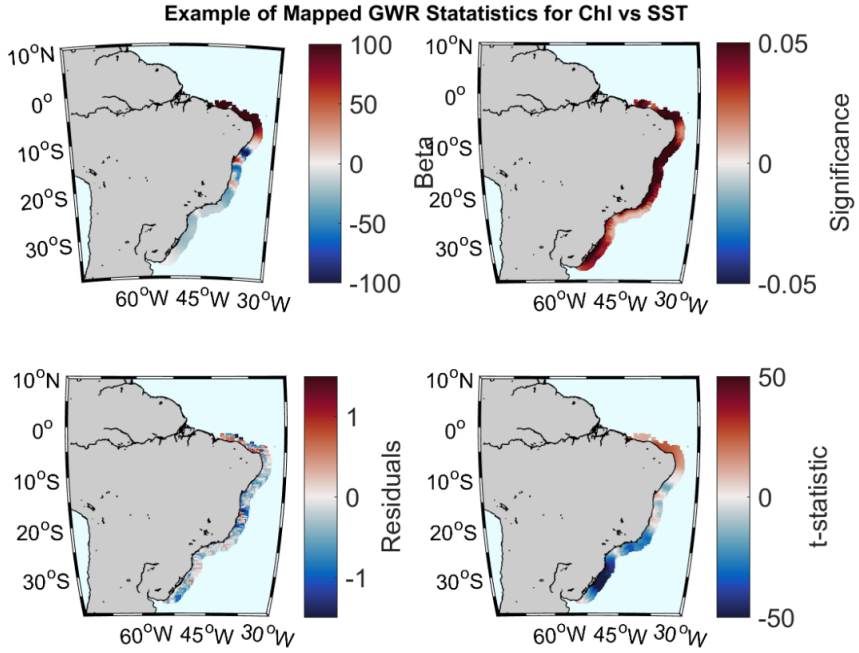

Figure 4. Spatial distribution od points used in GWR for chla versus SST.

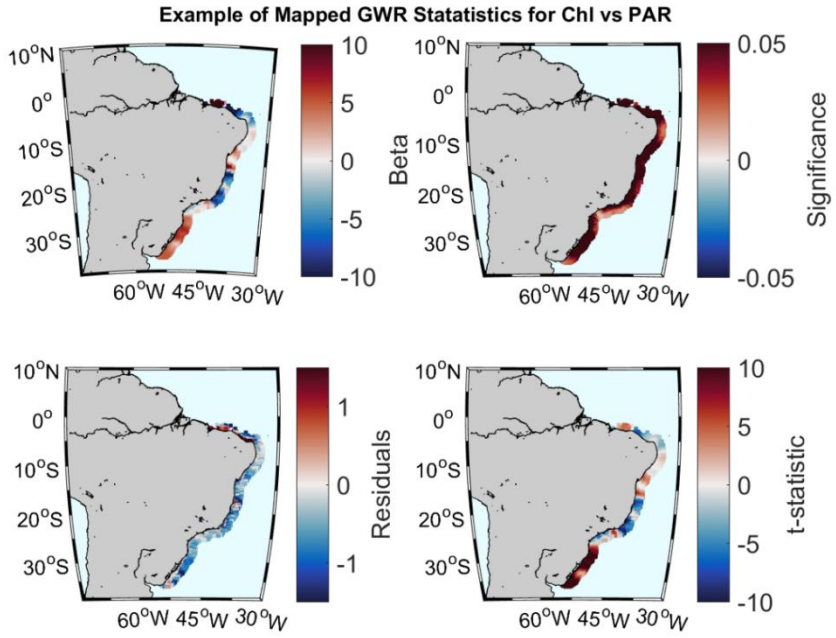

Figure 5. Spatial distribution of points used in GWR for chla versus PAR.

Still in Figure 6, it is notable for the chla-SST beta values, how the month of July, among most years, presented very low beta values near zero for all pixels, indicating very low correlation between the variables for all pixels for the austral winter. In the southern region of the coast, this process could be due to the downwelling of the water bodies which occur more frequently in austral winter due to conld fronts reversing wind direction (Paul et al., 2009) which ultimately lead to Ekman transport towards the coast. The downwelling process decrease the nutrient content in the water body which feeds the phytoplankton, and therefore, increasing the chla concentration. As the chla concentration increase does not occur substantially during winter the chla signal in the dataset might be affected given rise to noise data, decreasing chla predictability potential.

Another important feature clear from the GWR chla-SST beta values is that most values were within the negative range indicating an inverse correlation between the variables. This is an expected result since the lower the SST value the higher the chla magnitude due to upwelling high-nutrient lower-temperature water masses. However, at the northernmost part of the study area, a positive correlation occurs due to the lesser influence of upwelling waters and due to higher influence of local higher SST due to lower latitude. That way, wherever upwelling occurs more intensely, a stronger negative correlation for beta coefficient would be identified.

Considering the chla-PAR beta values, one notes that for the months of October/2006 and 2009 the beta values were very large, indicating that little variation in the predictor (PAR) implied large variation in the regressed variable (chla). Other than that, the overall chla-PAR beta value magnitude range presented an inverse correlation to that of chla-SST beta. Regarding the low correlation beta values (i.e., near zero), no specific pattern could be identified with respect to which month presented an overall tendency towards low beta values.

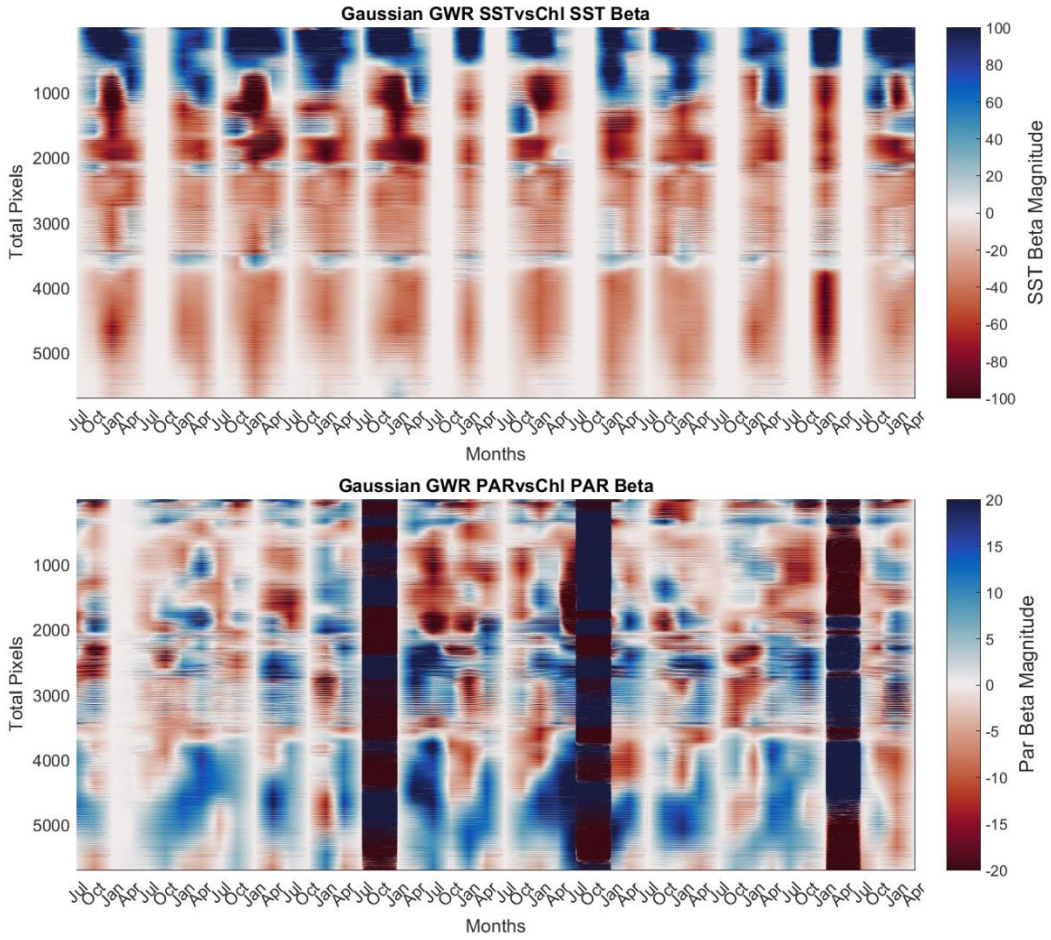

Figure 6. Beta values for chla-SST and chla-PAR GWR.

Considering the beta values for the constant term of the GWR (Figure 7), one notes how their results pattern equal in magnitude to that of the beta for the predictor variables. However, they were the inverse of each other with respect to the magnitude sign. This was a natural consequence of the regression line position within the regression attribute space in which the positive correlation presented negative constant term values whereas the negative correlation presented positive constant term values.

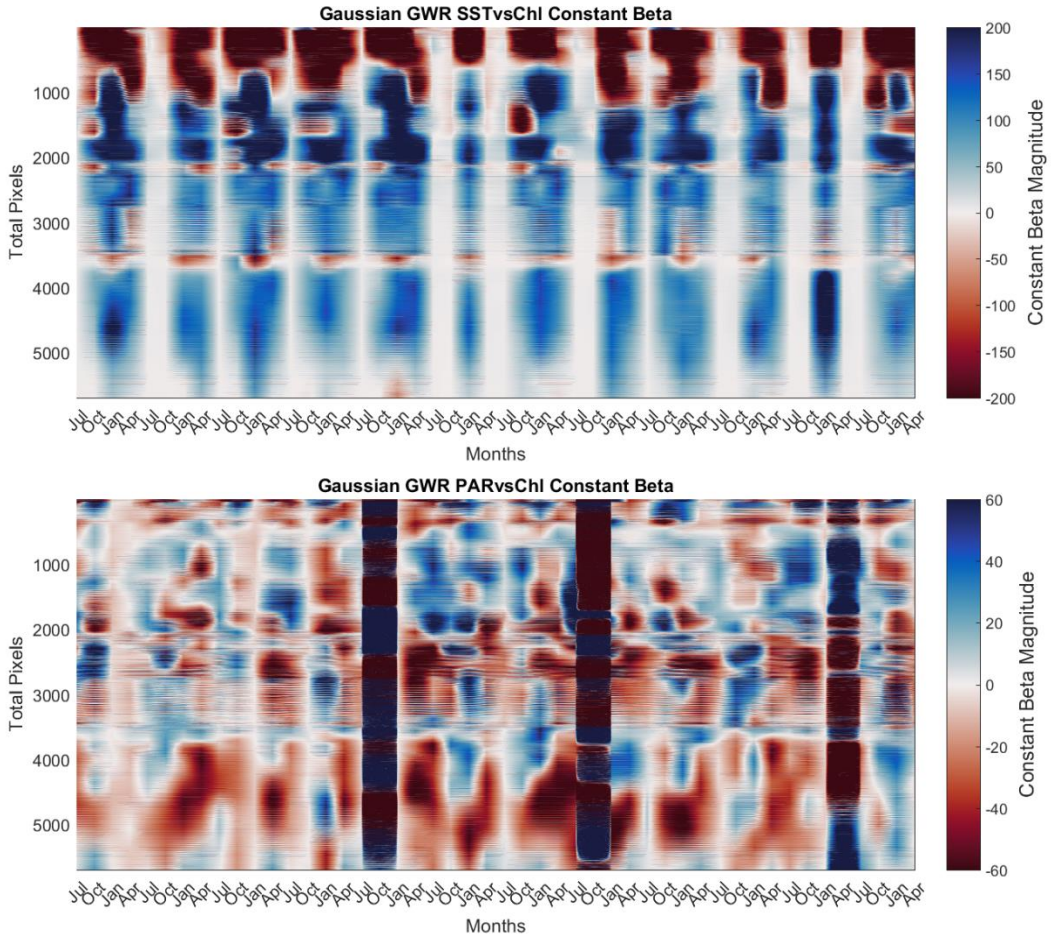

Figure 7. Beta values for constant component of the GWR for chla-SST and chla-PAR.

Figure 8 depicts the significance or mean squared error (MSE) values for the GWR results. The MSE is a measure of the quality of an estimator - it is always non-negative, and values closer to zero are better. Therefore, one notes how there is an annual cyclical pattern of the significance variance throughout the years despite not clearly showing a specific month with consistent lowest value. Nevertheless, the month of July presented the lowest value, and, thus, best estimation quality. The months of January presented the weakest value. Finally PAR presented a slightly weaker significance than SST results.

Despite Figures 6 and 7 present very low beta values for the month of July, and therefore, little correlation, it also displayed strongest significance (Figure 8) mostly due to less deviation from the GWR estimation. Some October months also presented that pattern. This feature is further corroborated by Figure 9 residual results in which those months presented low residual values. Opposed to that, the month of January presented, overall, the highest beta values, the weakest significance, and the highest residuals. And again, the months of April and October yielded results with not-so-defined patterns mostly due to the fact they represent Spring and Autumn which are seasons in which both summer and winter patterns can be found due to seasonal variability transition.

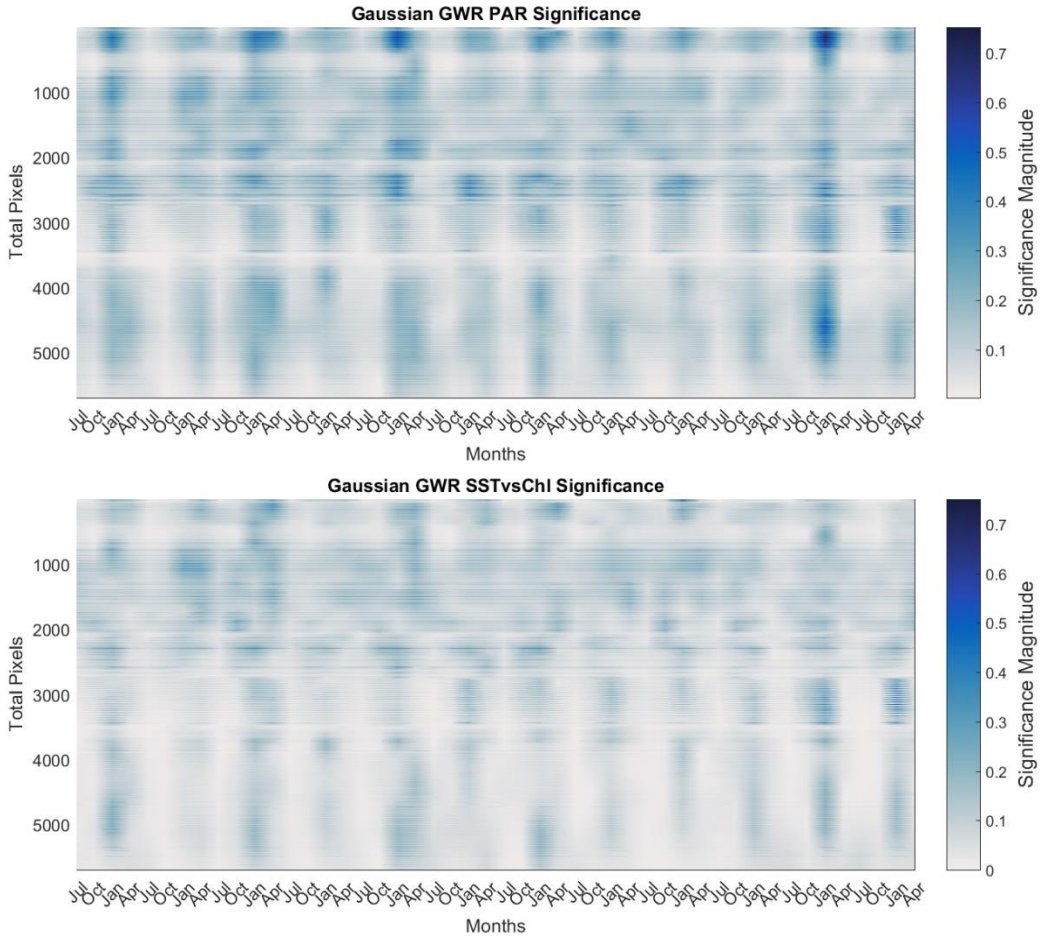

Figure 8. Significance or MSE values of the GWR for chla-SST and chla-PAR.

Another important point to emphasize is the fact that the pixels located at the northernmost latitude also presented highest residuals, weakest significance. Again, due to higher temperature in those areas, the OPP is not as influencing as in sourthern regions where upwelling occurs more frequently. As a consequence, lower chla value occur in water tend to be clearer (other than areas where sediments are more intensely discharged over the sea). And, low chla values tend to introduce noise in the data where prediction is being attempted.

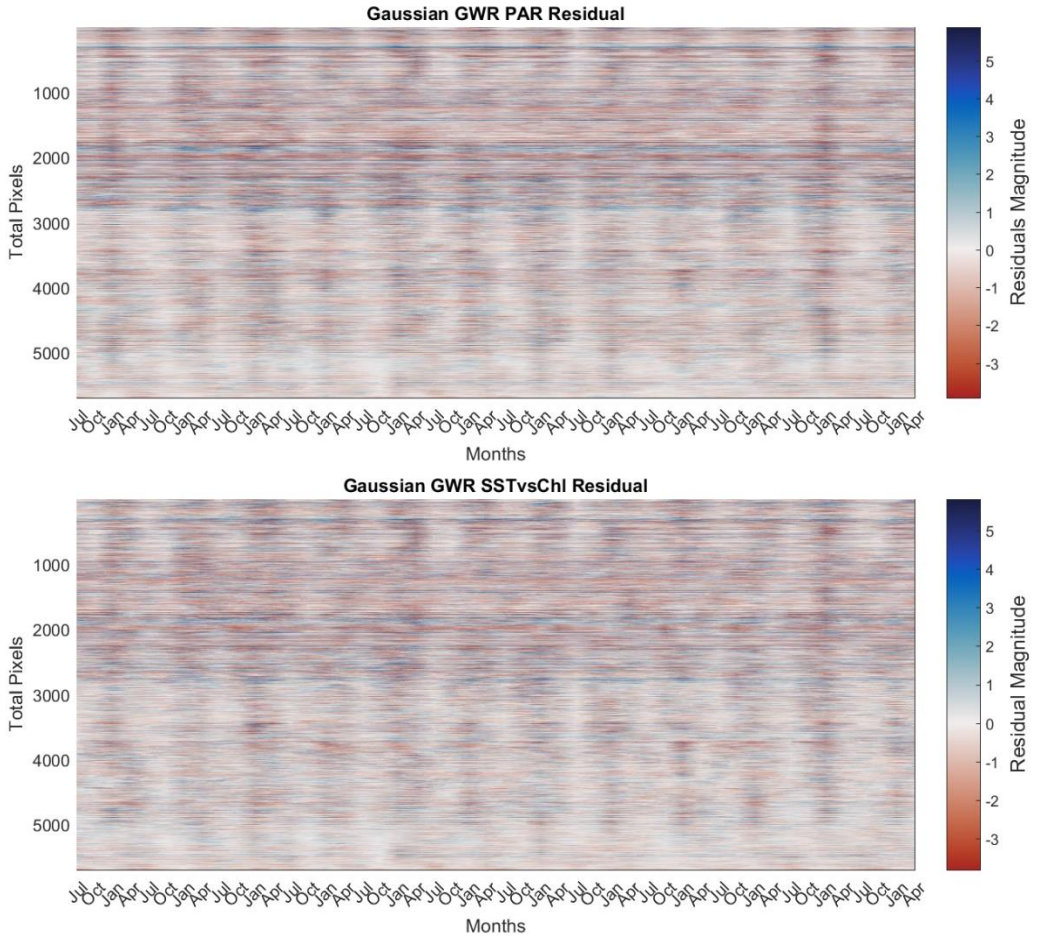

Figure 9. Residual values of the GWR for chla-SST and chla-PAR.

Figures 10 and 11 depict the same patterns for *t-statistics* as encountered for the beta values. Similar magnitude range, however, opposite signs. As *t-statistics* is a measure of signal-to-noise ratio, if the signal is weak relative to the noise, the (absolute) size of the *t-value* will be smaller as encountered for the northernmost regions just as it was found for significance and residuals. So the difference is not likely to be statistically significant. Positive values reflect a sample mean larger than your hypothesized mean, negative values reflect a sample mean smaller than your hypothesized mean. *t-statistic* is similar to a *Z-score*. In both cases, large values can be positive or negative, and the farther from 0 a value is the lower the probability beyond that value. In short, whether positive or negative, large *t-statistics* will always correspond to small p-values.

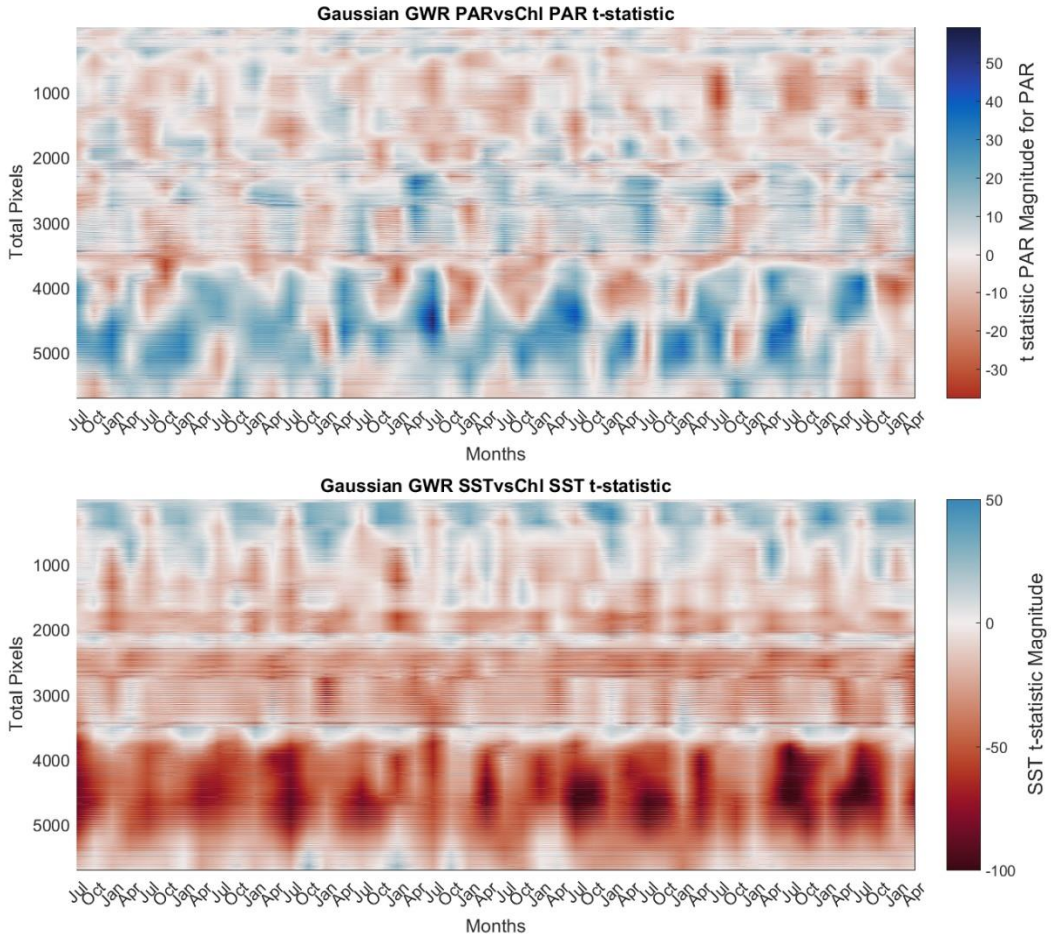

Figure 10. *t-statistic* values for predictor component of the GWR for chla-SST and chla-PAR.

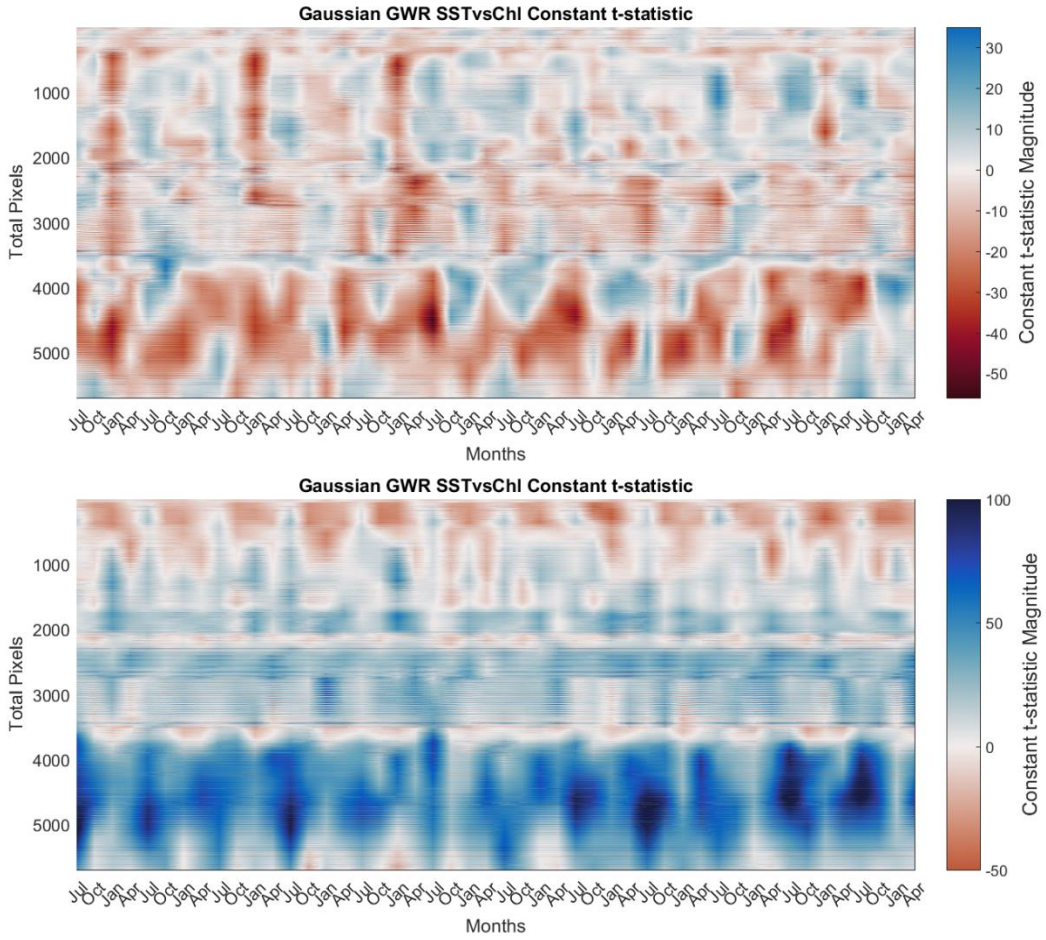

Figure 11. *t-statistic* values for constant component of the GWR for chla-SST and chla-PAR.

Figure 12 depicts the resulting R^2 value for the for both chla versus SST and chla versus PAR GWR analyses. The overall trend presented in the R^2 is that chla versus SST and chla versus PAR showed an approximate mean value of .86, and .8, respectively. This indicates that the correlation between chla and SST to be slightly stronger that that of chla versus PAR. Such result demonstrate similar correlation for both variable (SST and PAR). Overall, January presented the lowest R^2 values.

Overall, SST is expected to have a better correlation than PAR because upwelling events bring with them cold water which directly influence the dynamics of chla within OPP. However, since solar light is a crucial component of OPP PAR should also present strong correlation with chla concentration, and a robust dataset both spatially- and temporally-constructed could appreciate that fact. Since this study aimed to cover the entire Brazilian coast, the dataset appeared sufficient to account for the impact of each variable in the final regression result, aimed to account for the spatial correlation by using the GWR technique along with all data treatment prior to model analysis.

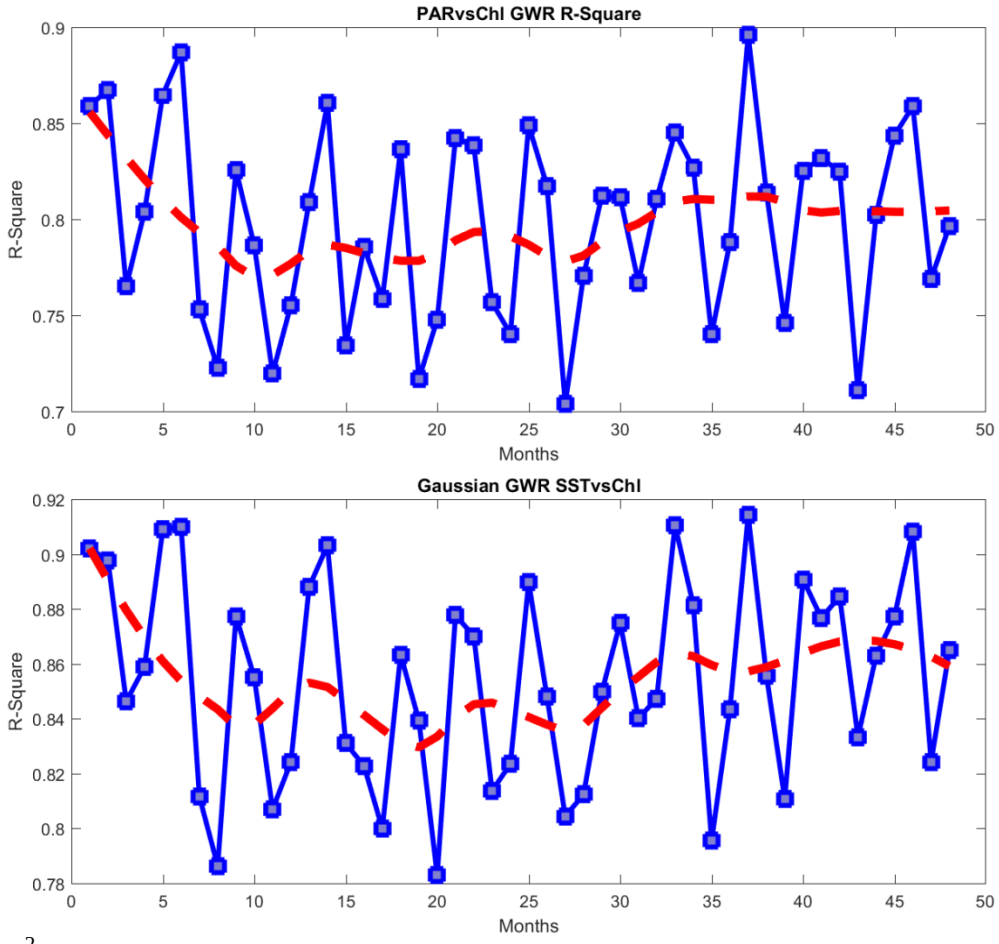

Figure 12. R^2 values chla-SST and chla-PAR.

Figure 13 depicts the beta value range results from BGWR compared to the beta estimates from GWR. Only results for one month are depicted in Figure 13 for simplicity. Nevertheless, one notes how a significant range from the GWR beta estimation falls outside the 95% confidence interval estimated by the BGWR. Thus, in order to properly account for such variability through all months, a histogram was created for the ratio between the amount of pixels from GWR estimates that fall within the BGWR 95% confidence interval range and the total of pixels available for analysis (Figure 14).

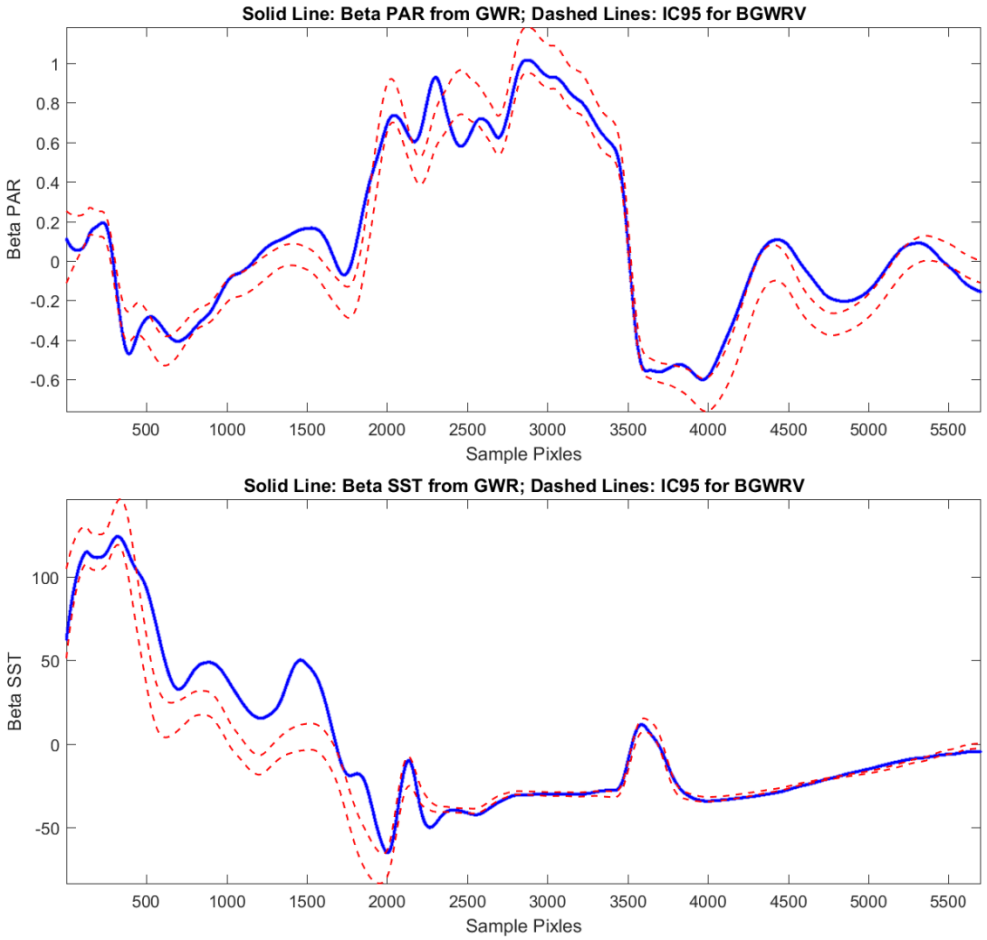

Figure 13. Beta values estimated via GWR in blue, and 95% confidence interval range from BGWR in dashed red.

Overall, it is evident from both histogram that the average percent ratio are very low (mean of .45 fort chla versus PAR GWR; and .37 for chla versus SST). These results demonstrate the importance of stochastic processes in quantify large scale spatio-temporal analysis for GWR as it might introduce significant amount of error even after accounting for the regression prerequisites. Also, another important feature observed from Figure 14 is that PAR results appeared slightly better than SST results. This can be an indication that although PAR R^2 values being slightly worse, its spatio-temporal beta distribution was more statistically robust, indicating that it might be an important variable such as SST. And depending on the statistical parameter of interest, PAR can be a statistically more significant variable.

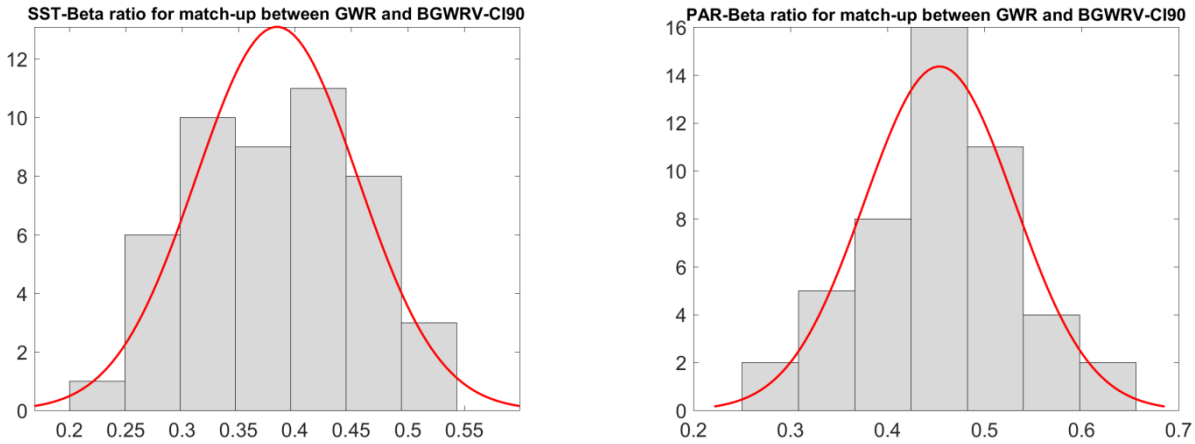

Figure 14. Ratio between beta values estimate via GWR and beta values estimated via the BGWR.

Lastly, Figure 15 depicts the matrix containing binary values of either 0 or 1, indicating the absence or presence of that pixel from a given month into the 95% confidence interval of the BGWR result. One notes how the northern regions have a tendency to fall outside the BGWR confidence interval range corroborating results from Figure 8 and 9 which indicate higher error and higher residual values, respectively. And again, the month of July mostly fell outside the BGWR confidence interval range for most months and pixels, further indicating the low predictability efficacy for that month.

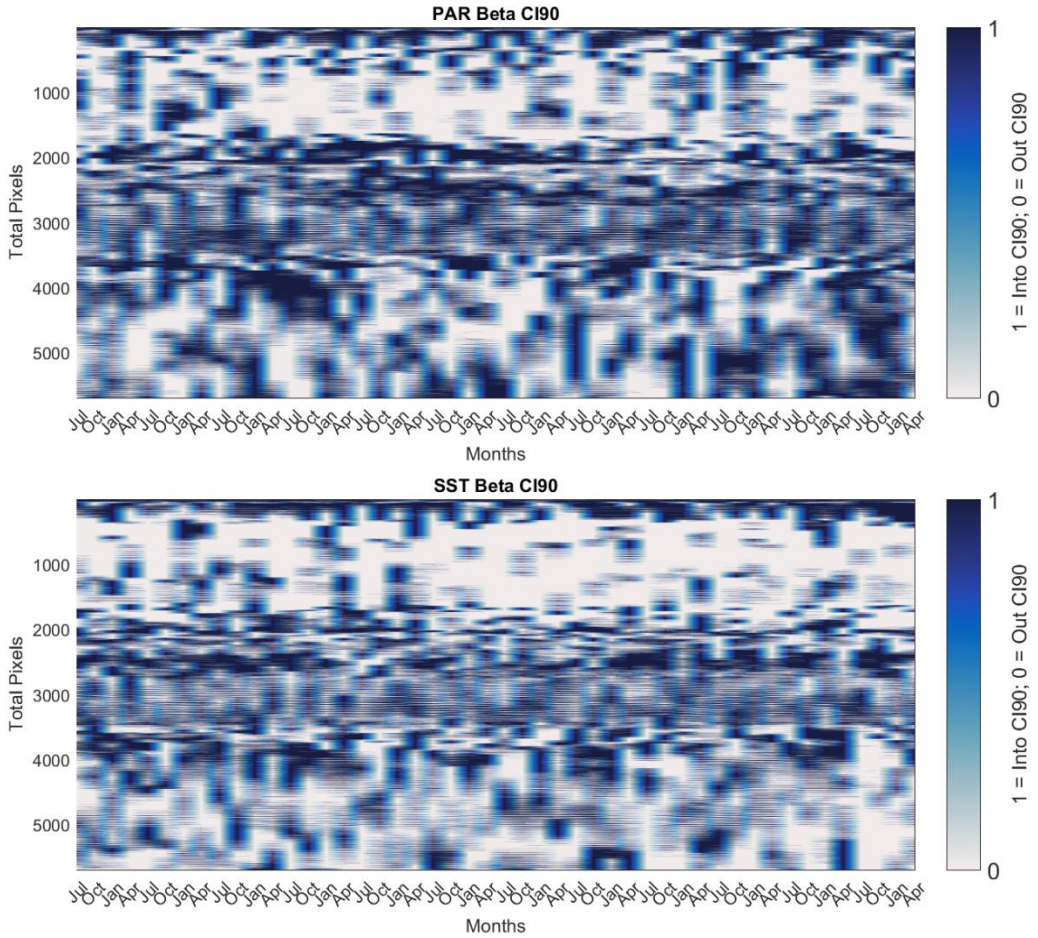

Figure 15. Matrix of binary inclusion (blue) or exclusion (white) of a pixels of each month into the BGWR 95% confidence interval.

3. Conclusion

This study aimed to investigate whether SST or PAR presented best correlation with chla concentration evaluating them over a time span of 48 months along the Brazilian coast, and also tested for a BGWR technique as to compare GWR results with a stochastic approach. Overall, the results did not seem significantly different, with slightly better R^2 values for chla versus SST and slightly more significant beta results for chla versus PAR. Therefore, despite the fact that SST might have a more direct effect on chla concentration, PAR also play a major part in dictating chla concentration over the Brazilian coast. Also, BGWR showed that GWR results was not highly robust indicating the importance of a stochastic approach.

References

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2002.Geographically weighted regression. Chichester, UK: John Wiley and Sons.

Huang, B., Wu, B., Barry, M., 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int. J. Geogr. Inf. Sci. 24 (3), 383–401.

Paez, A., Uchida, T., and Miyamoto, K., 2002. A general framework for estimation and inference of geographically weighted regression models: location-specific kernel bandwidths and a test for local heterogeneity.Environmental and Planning A, 34, 733–754.

Paul M. Markowski; Yvette P. Richardson, 2011. Mesoscale Meteorology in Midlatitudes. John Wiley and Sons. p. 120. ISBN 978-1-119-96667-8.