Sargassum: Detect & Predict (SDP)

Arturo Esquerra, Rani Fields, Kingsuk Maitra

***Abstract*—A simple, generalizable, and cost-effective machine learning based model is developed to enable quick and accurate detection of floating sargassum in earth's ocean systems. It is shown why existing Physics based, partially empirical models fall short in accurately discriminating between sargassum and other floating vegetative matter. MODIS remote sensing satellite data is used for training, and testing subsequent predictive accuracy. Exhaustive exploratory data analysis (EDA) is carried out and several insights are drawn along the way to highlight gaps and limitations in existing data collection methods and techniques. Economic and social value of this exercise is explicitly justified to emphasize its applicability across different industry and government verticals. A compact API endpoint for accessing and testing the model is also developed. Conclusions are drawn as to the suitability of generating pre-trained algorithms which may be deployed in edge devices in a high-volume inferencing context.**

***Index Terms*—Artificial Intelligence, Floating vegetation, Machine Learning, Modeling, MODIS, Remote Sensing, Sargassum**

I. Introduction

Sargassum proliferation is coastal areas are often cited as causes of economic, environmental, and social catastrophe in coastal areas which is usually under-appreciated or outright ignored. For example, recently, massive accumulations of sargassum seaweed in the coastal regions of the Caribbeans is identified as the primary cause of economic activity loss (by way of damage to the thriving tourism industry in the region) to the tune of $1.85 billion (USD). Coupled with massive ecological consequence alongwith a changing climate, Sargassum proliferation has the potential of displacing large segments of the world population from coastal habitats to other inland communities triggering a potential migrant crisis the world may not have seen. Apart from threating the $37 billion USD a year tourism industry in the Caribbean, the sargassum encroachment threatens pristine coral reefs, triggers beach fouling, causes irreversible damage to the all important fishing industry, and destroys the natural habitats of endangered species like sea turtles. The potential benefit of an early warning and detection system of potential sargassum proliferation cannot be overstated. Early detection would enable strategic deployment of clean-up crews and preventive actions in beaches and high-risk areas. Thus, any automated and reasonably accurate methodology of early detection is of paramount importance and value. Before, delving into further details of the methodology and algorithm development, a closer look is taken in this section of contextualizing the background with establishing clarity around definitions and terms that would be used extensively throughout this report.

The interested reader is first exposed to the Sargassum problem in this paragraph. *Pelagic Sargassum* is a type of brown seaweed distributed throughout tropical oceans notable for forming large floating mats called "golden tides" that provide a habitat for marine life in cold ocean waters. As mentioned in the preceding paragraph, starting in 2011, atypical massive shoals of Sargassum started to wash ashore at Caribbean coasts. The Caribbean is an oligotrophic sea (low nutrients), making the large quantities of biomass brought by golden tides a dangerous disruption to the ecosystem. The decomposition of large quantities of Sargassum that reach the beaches causes bioerosion, changes in the ocean's water chemistry, anoxia, foul odors, and the production of hydrogen sulfide. Interestingly, the causation of Sargassum attacks are not well understood-why it happens or when and where? Little is known about the cause or source of Sargassum outbreaks but many researchers hypothesize that it is related to changes in ocean characteristics, especially global warming. Though the connection with climate change is not clearly established, there is an emerging consensus among the scientific community that climate change may very well be a trigger for sargassum proliferation. Below, an anecdotal history of the key consequences of sargassum outbreak are sketched. It is by no means meant to be an exhaustive survey of the same, but is expected to be a representative window of the potential harm caused by sargassum damage from a historical perspective. For instance, between July and September 2018 between 24 and 48 tons of Sargassum washed ashore daily. Local Mexican government have removed more than 150,000 tons of Sargassum were removed in four months from Quintana Roo. Quintana Roo expects 20 million tourists in 2019 who would account for $8 billion (USD) in revenue. The purported opportunity cost implications from disruption of this key industry to the local economy is staggering to say the least. In the next section titled "State of the Art", the techniques used today to solve the Sargassum problem is summarized.

The authors are with the Department of Information Sciences at the University of California at Berkeley, Berkeley, CA as final semester graduate students in the Master of Information and Data Science (MIDS) program. AE is with Facebook, Inc. in Mexico City, Mexico; RF is with Paypal, Inc. in San Francisco Bay area, CA, USA; KM is Microsoft Corp. in SF bay area, CA, USA

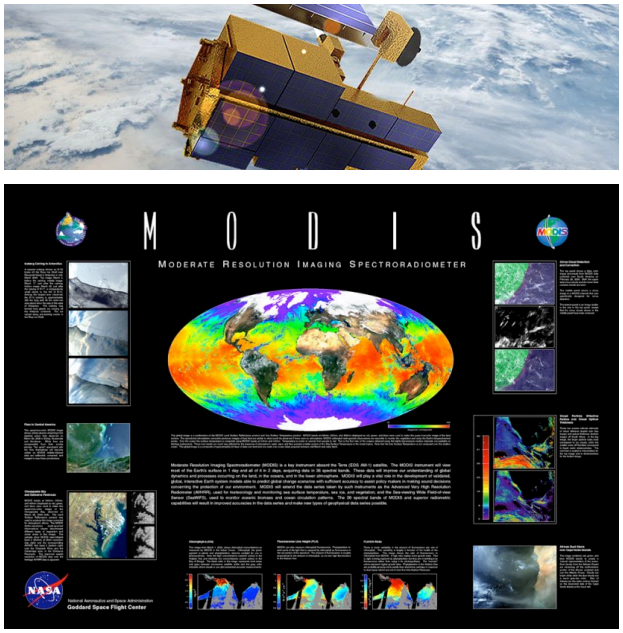

Figure 1: Specifications of the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors onboard of MODIS-Aqua and MODIS-Terra NASA satellites.

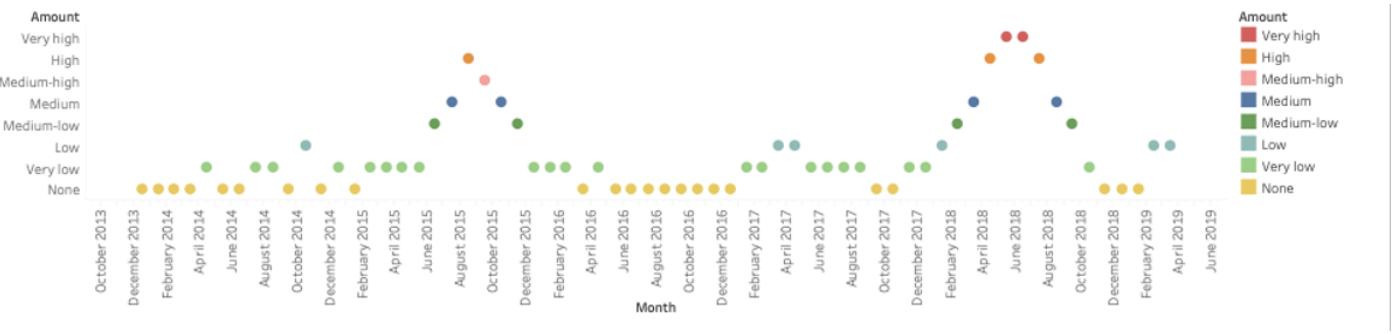

Figure 2: Sargassum levels in the Caribbean from 2014 to 2019.

II. STATE OF THE ART

The standard approach used by scientists to detect Sargassum outbreaks consists on relying on an optics analysis of satellite imagery from NAA satellites. (Figure 1 illustrates the details of collecting remote sensing satellite data). The Moderate Resolution Imaging Spectroradiometer (MODIS) is a sensor on board of the 2002 NASA Aqua satellite. It captures spectral bands from a wide range of wavelengths to provide measurements of large-scale dynamics and processes occurring in the oceans. The MODIS measurements on sea color, levels of chlorophyll, and sea surface temperature are thought to be useful to detect Sargassum.

The key metric used for ascertaining the "Sargassum" versus "No Sargassum" infested ocean waters is called "Floating Algae Index" or FAI which may be expressed as a function of wavelength (λ) of the light band under consideration, and the corresponding reflectance as follows [8, 9]:

$$\mathrm{FAI} = R_{rc}^{NIR} - \left[\frac{R_{rc}^{RED}(\lambda_{NIR} - \lambda_{SWIR}) + R_{rc}^{SWIR}(\lambda_{RED} - \lambda_{NIR})}{(\lambda_{RED} - \lambda_{SWIR})}\right]$$

Clearly, when pre-labelled data (marked as "Sargassum" vs. "No Sargassum") is used to calculate FAI, and plotted as shown below, the data fails to discriminate between the two categories. This is an indicator of the fact that a better technic based on machine learning may be suitable for discerning between ocean water infested with Sargassum versus No Sargassum. This is where the State of the ART Physics based models fall short.

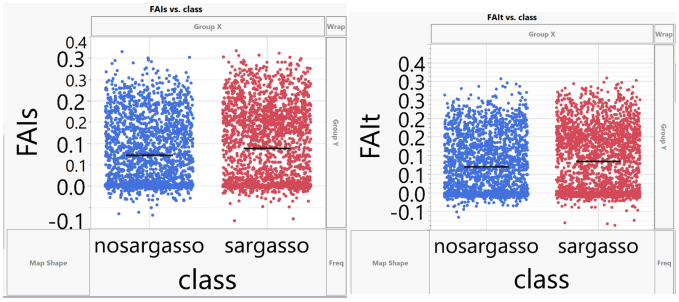

Figure 1(a): FAI's computed using pre-labelled Sargassum versus No Sargassum data. Clearly, they fall short of discriminating between the two.

III. DATA

Our project focuses on studying the Sargassum phenomenon through the two most recent outbreaks which happened in 2015 and 2017. This focus stems from the fact that while a Sargassum outbreak was reported in 2011 (the first one in modern days), the data quality for this period is significantly inferior as it provides sparse information about the beginning of the outbreak and does not have the granularity we needed to create robust models.

In order to obtain a workable model we needed to know the date and geographic location in which Sargassum washed ashore. Therefore, we combined official government information from the state government of Quintana Roo, Mexico (one of the regions that have been affected the most by the outbreaks) with news articles obtained by scraping Google for reports about Sargassum in the area, data obtained from the creators of ERISNet Dr. Javier Arellano-Verdejo and Dr. Hugo-Enrique Lazcano-Hernández, information about dates and locations from Professor Briggita I. van Tussenbroek, data from local businesses (hotels), from NGOs (United Nation Environment Program's Sargassum report), and peer-classified observations from iNaturalist.org. The combination of these sources of information provided the largest known Sargassum dataset which not only included dates and locations but also the

intensity of the outbreak in an 8-point scale from "none" to "very high" generated by the volume of Sargassum removed each month from the shores in the region.

Given the set of dates, locations, and outbreak's intensity we obtained satellite imagery from Moderate Resolution Imaging Spectroradiometer sensors from NASA's MODIS-Aqua and MODIS-Terra missions focusing on specific wavelengths of level 2 and level 3 data which had been shown to provide the most insightful data about sargassum presence based on an entropy analysis (ERISNet). In addition, and in contrast with previous research, we added images to our dataset of each location's Aerosol Optical Thickness (AOT), Sea Surface Temperature (SST), and Nighttime Sea Surface Temperature (NSST) which are algorithmically generated images based on reflectance levels of other spectrums but which provide insights of complex phenomenon that happen on the surface.

To efficiently obtain the data, we developed an API to locate, download, extract, and format the data in a scalable way using a five-step process:

1. Get sinusoidal tile ID from latitude/longiture coordinates
2. Find appropriate dataset file from data download directories (date + tile + product)
3. Shell: "wget" to handle redirects
4. GDAL: extract raw data as geotiff files to target directory
5. Matplotlib: Convert geotiff data to image files

IV. Exploratory Data Analysis

Based on a dataset provided by the ERISNet researchers consisting on reflectance values for seven MODIS bands for forty dates in which Sargassum presence was officially confirmed an Exploratory Data Analysis was conducted to determine important features in the data.

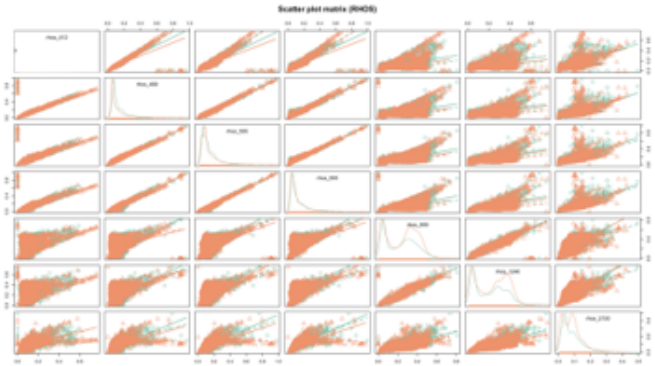

Figure 3: Scatterplot matrix of the reflectance levels of the ERISNet dataset.

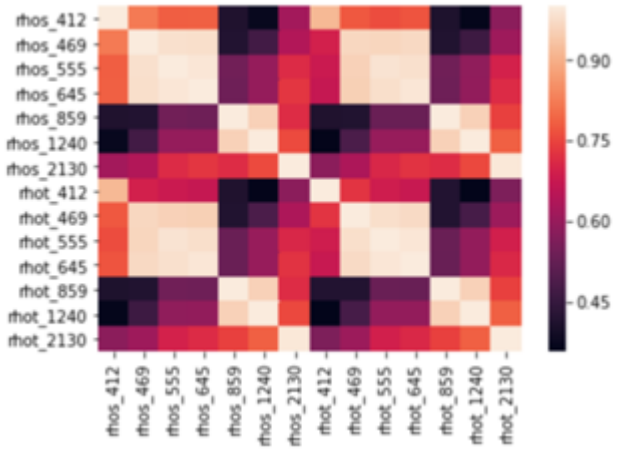

Figure 4: Correlation heatmap of the reflectance levels of the ERISNet dataset.

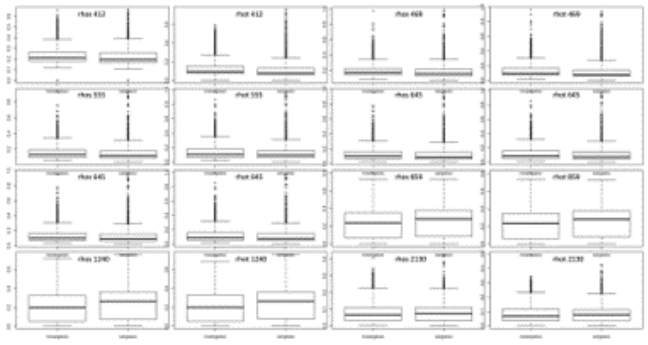

Figure 5: Boxplots of the reflectance levels of the ERISNet dataset.

The exploratory data analysis reflects large linear relationships between the reflectance values but not with the presence of sargassum. This behavior strongly suggests the presence of a highly non-linear relationship between the features and the outcome. In addition, Figure 5 shows the considerable imbalance of classes as well as the challenges of prediction with a linear model of the features. Exploratory Data Analysis also suggests that Principal Component Analysis (PCA) could be useful to reduce dimensionality and increase model performance.

V. SDP Model

What makes the SDP model unique is the amount of data that it uses to detect Sargassum outbreaks and also the focus on deriving predictions from the ocean characteristics at each point in time instead of using individual reflectance values obtained from pixels of low-resolution satellite imagery. This makes the model more robust to errors in data quality and also more insightful as it links the unique characteristics of the ocean to the presence or absence of Sargassum.

SDP uses information from the reflectance values of MODIS imagery for the wavelengths: rrs 412, rrs 443, rrs 469, rrs 488, rrs 531, rrs 547, rrs 555, rrs 645, rrs 667, rrs 678, sst, nsst, and aot 869, where sst (sea-surface temperature), nsst (night sea-surface temperature), and aot (aerosol optical thickness) are composite measures that summarize a diverse set of specific

aspects of the ocean. Furthermore, descriptive statistics (image pixel mean, standard deviation, maximum, and minimum values), are obtained for each map of a region of interest and missing pixel values due to tearing or presence of clouds are imputed through different approaches.

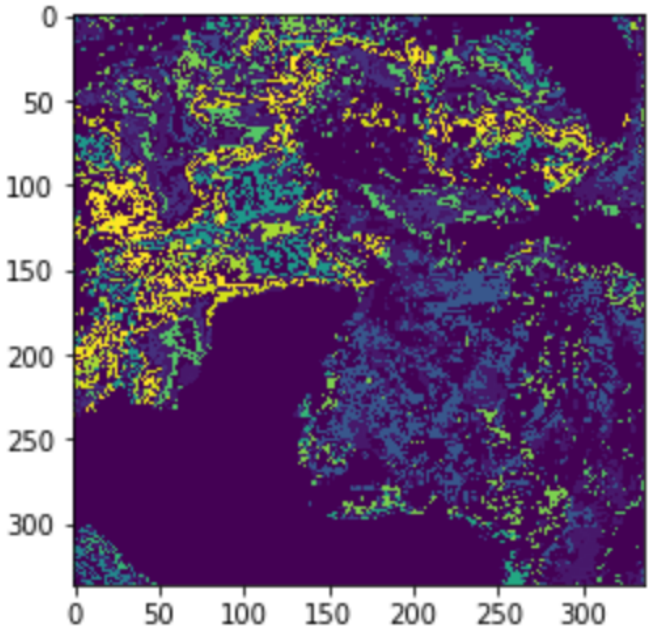

Figure 6: Sea surface image temperature of the Caribbean during a Sargassum outbreak.

A. Feature Engineering

Based on the wavelengths and composite measures images obtained from MODIS need to be pre-processed in order to accurately predict the presence of Sargassum. First, the pixel values are applied a power transformation (λ=4) to increase the variance in pixel values. Then, missing pixel, a considerable issue with the MODIS imagery, is addressed through either the IPPA or MRPF methods developed for SDP. Next, standardization within each image's context is applied and descriptive statistics for each image are calculated to be used for prediction.

1) Informed Per-Pixel Averaging (IPPA)

Informed Per-Pixel Averaging (IPPA) looks to substitute missing values through an average of past recent values at the same position. In particular, for each pixel (x, y) coordinate, IPPA tabulates a list of values from the past 30 days for that pixel (rolling mean of each pixel). Therefore, for each pixel, the values are averaged and missing data signals are dropped.

Given that ocean characteristics are relatively constant through time unless a serios disturbance is found, IPPA allows us to recover missing data based on the most recent images.

2) Most-Recent Pixel Fill (MRPF)

Most-Recent Pixel Fill (MRPF) is an alternative, but more computationally intensive approach we developed takes the most recent pixel for a missing spot from within the past 30 days, ignoring anything older for pixel interpolation.

B. Sargassum Models

Instead of simply modelling the presence or absence of Sargassum, SDP leverages the large amount of data paired with the intensity score developed for the model to predict the direction or trend of an outbreak. This allows additional actionability for the model as it expands its use from binary prediction to a forecast of whether a crisis will continue to increase in intensity or if it will dwindle down. Each model is estimated and validated on the raw image data, on the raw image data using Principal Component Analysis, with pixel-missing imputation (i.e. IPPA), and with imputation plus PCA. For prediction, logistic regression and Random Forests are of particular interest given that the former provide simplicity and interpretability that can help elucidate causes and strategies to combat Sargassum while the latter can provide good accuracy for production. We are dubious of the performance of CNNs due to the heavy presence of gradients. Features which are useful in common CNN use cases (e.g. edge detection) would not produce meaningful gains with our satellite data.

1) The Zero-Case

Given the considerable limitations of current spatial imagery from the MODIS missions in terms of spatial and temporal resolution, paired with imperfect labelling mechanisms that fail to detect the start of an outbreak until a critical mass of Sargassum has washed ashore, there is a large amount of heterogeneity among cases labeled as "no Sargassum", in particular during the beginning and end of golden tide episodes. Therefore, when these entries are included in the model, the error rate increases substantially, generating problems to train accurate models.

The Zero-Case is nontrivial to deal with since they are fundamentally different than missing data and that there is real "no sargassum" data in the dataset, however, it is mixed with missing data. Likewise, we cannot just treat all Zero-Cases as "no sargassum reported" without majorly increasing our error rate.

2) Binary Classifiers

Logistic regressions and Random Forest classifiers are applied to predict the binary outcome of whether an outbreak is increasing or decreasing in intensity. Leaving aside the Zero-Case.

3) Multinomial Classifiers

Taking advantage of the increases in performance offered by IPPA and MRPF, the multinomial approach includes the Zero-Case, extending the model to be able to predict not only whether the crisis is increasing or decreasing, but also when it starts and ends.

Table 1: Model results.

TABLE 1
Model Results

Model	Raw Image Data Only	Raw Image w/ PCA (47)	w/ IPPA*	IPPA* only	IPPA* only, PCA (21)
Logistic Regression (binary, macro avg. f1)	0.71	0.69	0.91	0.87	0.93
Random Forest Classifier (binary, macro avg. f1)	0.67	0.64	0.97	0.99	0.99
Logistic Regression (multinomial, macro avg. f1)	0.37	0.4	0.55	0.65	0.66
Random Forest Classifier (multinomial, macro avg. f1)	0.36	0.36	0.93	0.95	0.98

* IPPA = Informed Per-Pixel Averaging, rolling mean of each pixel for data above a certain value

VI. Conclusions

The sargassum phenomenon is highly complex and still not fully understood. Limitations in the spatial and temporal resolutions for current satellite sensors make it almost impossible to simply track sargassum. Therefore, models such as SDP provide a valuable resource to expand our understanding of golden tides as well as give the tools to predict them on time.

SDP has shown that it is possible to not only predict when an outbreak will start, but also to determine whether it will continue to grow or decline in intensity. This is an important breakthrough that hasn't been explored in the past and that can help stakeholders respond in a timely manner and manage their resources optimally. Furthermore, while imagery of ocean data is relatively stable through large periods of time, we found that as sargassum outbreaks develop, oceans start to show unique characteristics which allow us to predict golden tides accurately. This is important not only from a modeling perspective, but also because it helps us understand more about the phenomenon. For example, some of the leading experts on Sargassum believe that outbreaks are so recent and devastating because the seaweed has changed its behavior due to an important environment change such as due to climate change. SPD gives credence to these theories. Which in all fairness could be caused by the outbreak, but we have reason to believe the contrary given the temporal component in which that we can predict the start of an outbreak through oceans characteristics.

One of the biggest breakthroughs of the project came from how we collected, processed, and analyzed the data. SDP has the largest Sargassum dataset we are aware of both which focuses on the larger picture, the dynamics between the ocean's characteristics and Sargassum outbreaks, making it more robust than traditional models which have relied on pixel values imputed from low resolution satellite imagery. Moreover, the way in which we processed the data, handling incomplete and dirty information, allowed us to achieve great performance. All-in-all, this gives us a good level of confidence to say that the Sargassum problem can be solved with better data. In addition, we conclude that Machine-Learning approaches provide a much more accurate and saleable ways to study and predict the phenomenon compared to traditional methods. We believe that this is because advances in data processing, combined with large amounts of data, and improvements in modelling can help us cover some of the shortcomings of current satellite sensors (which, as we have mentioned, have low spatial and temporal resolutions) and imperfect data labelling systems.

Finally, this is still a very recent phenomenon of which we have only encountered three major crises, therefore, it is fundamental to validate it on future outbreaks, determining whether the model suffers from over-fitting.

VII. References

[1] Arellano-Verdejo J., Lazcano-Hernandez H., Cabanillas-Terán N. 2018. ERISNet: Deep learning network for Sargassum detection along the coastline of the Mexican Caribbean. PeerJ Preprints 6:e27445v1 https://doi.org/10.7287/peerj.preprints.27445v1

[2] Caribbean Alliance for Sustainable Tourism. Sargassum: A Resource Guide for the Caribbean.

[3] Hu C. 2009. A novel ocean color index to detect floating algae in the global oceans, Remote Sensing of Environment, Volume 113, Issue 10.

[4] Putman N., Goni G., Gramer L., Hu C., Johns E., Trinanes J., Wang M. 2018. Simulating transport pathways of pelagic Sargassum from the Equatorial Atlantic into the Caribbean Sea. Progress in Oceanography, Volume 165.

[5] Rodríguez-Martínez R., Tussenbroek B., Jordán-Dahlgren E. (2016). Afluencia masiva de sargazo pelágico a la costa del Caribe mexicano (2014-2015).

[6] United Nations Environment Programme (UNEP). 2018. Sargassum White Paper - Sargassum Outbreak in the Caribbean: Challenges, Opportunities and Regional Situation.

[7] Webster R.K., Linton T. (2013). Development and implementation of Sargassum Early Advisory System (SEAS). Shore & Beach.

[8] Dogliotti et. al, 1. Remote Sens. 2018, 10, 1140; doi:10.3390/rs10071140

[9] Hu et. al, 2009, Remote Sens. Environ., 2009, 113, 2118-2129