Uncertainty Characterisation in Ocean Colour Estimation

Kayvan Nejabati Zenouz[1]

University of Greenwich

The Leslie Comrie Seminar Series

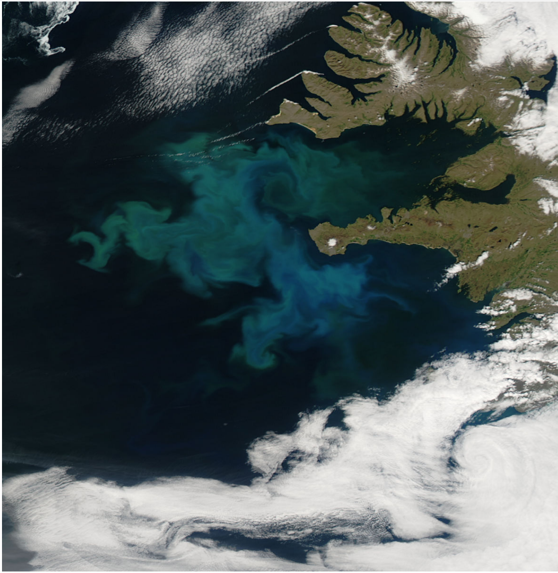
November 20, 2019

Joint work with Land, Peter E.; Bailey, Trevor C.; Taberner, Malcolm; Pardo, Silvia; Sathyendranath, Shubha; Brammall, Vicki; Shutler, Jamie D.; and Quartly, Graham D.

[1] Email: K.NejabatiZenouz@gre.ac.uk website: www.nejabatiz.com

Background: Thursday, 4 July 2002 through Sunday, 10 November 2019, https://oceancolor.gsfc.nasa.gov/cgi/browse.pl?sen=am

Content I



Earth Observatory[2]

[2] Phytoplankton Bloom off Iceland, Moderate Resolution Imaging Spectroradiometer on NASA's Aqua satellite, June 24, 2010.

Overview

- Introduce **statistical modelling** method in order to characterise **uncertainty** in **ocean colour** estimation.
- Modelling with Generalised Additive Models for Location, Scale, and Shape (GAMLSS).
- Data on ocean **chlorophyll concentrations** from the **MODIS** instrument aboard NASA's Terra and Aqua satellites.
- Match satellite and **in situ** measurements of oceanic chlorophyll concentrations.
- Take **explanatory variables** provided by **satellite** and model via GAMLSS.
- Find best-fitting model to explain the error and most **contributing** explanatory **variables**.
- This can be used to **improve** satellite instruments.

Introduction

- Ocean colour is determined with the **interaction** of **Sun** with substances in the **ocean**, one of which is chlorophyll produced by marine phytoplankton.
- Surface chlorophyll concentration is an **important** indicator of the **biology** and **physics** within the surface ocean and crucial for understanding of the **Earth System**.
- Ocean colour is estimated either **in situ** using **boats** or permanent observation stations or by using suitable **sensors** on board **satellites**.
- The **methods** for ocean colour estimation used by **NASA** have uncertainties which depend on
  - Sun-sensor geometry
  - Atmospheric aerosol load
  - Cloud contamination
- However, satellites covers the Earth in short time, so large data production!

Ocean Colour Processing Flags

- Several levels of **flags** based on **continuous** thresholds are used to exclude pixels from colour processing.
- In this way many outliers are removed from daily or monthly composites.
- At level 2 and 3 NASA satellite masks pixels with
  - **CLDICE:** suspected cloud or ice contamination,
  - **HILT:** high light, saturating one or more visible channels,
  - **HIGLINT:** strong sun glint,
  - **HISATZEN:** high satellite view zenith angle,
  - **HISOLZEN:** high solar zenith angle,
  - **STLIGHT:** stray light from nearby bright pixels.
- **The problem:** if a pixel is just **below** the **threshold** for each of the above, it will be **included**, but the final estimation may be **unreliable**!

Aims and Data

Aim

In this work we created a **statistical model** of the **difference** between **satellite** chlorophyll-a, chl_{SAT} reference or validation data in situ chlorophyll-a, chl_{IS}.

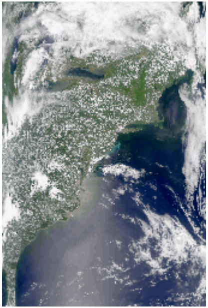
Data

Out **response variable** is from a skewed χ-squared distribution, so we need flexible regression techniques

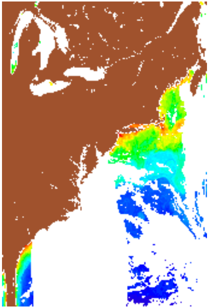
Generalised Gamma Distribution

offered through GAMLSS.

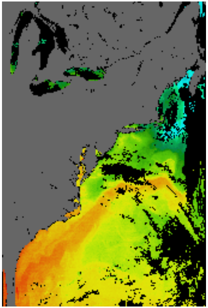
- First **dataset** was extracted from **NASA's** Ocean Color WEB level 1 and 2 browser.
- This data is a subset of that collected by the **MODIS** instrument aboard the Aqua satellite and was recorded between July 2002 and November 2011.
- Typical files: west Coast of US, Wednesday, 6 July 2016,



Quasi True Color



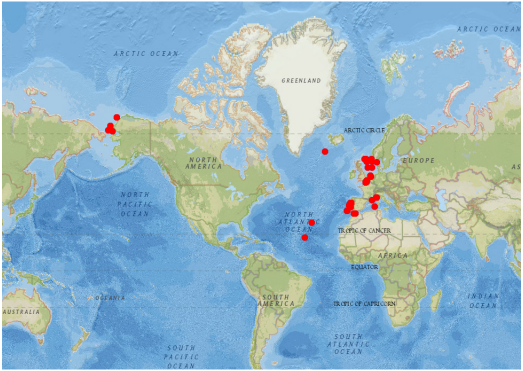
Chlorophyll



Sea Surface Temperature

In Situ Chlorophyll-a chl_{IS}

- Another **dataset** of 359 in situ High Performance Liquid Chromatography (HPLC) surface ocean chlorophyll-a (chl) measurements from 2002 to 2011.



http://rpubs.com/KayvanNejabati/551817

- Mostly from **European** shelf seas but including some data from the open **North Atlantic**, the Mediterranean, and the North Pacific.

Matching Data

- For each chl_{IS} measurement, we searched for all **overlapping** MODIS-Aqua overpasses within $\pm$12h.
- We use a subset, information about some of the variables
  - **In situ:** `timeI`, `lonI`, `latI`, `chlorI`.
  - **Satellite:** `satid`, `lonS`, `latS`, `chlorS`.
  - **Matching:** `distkm`, `timediffmin`.
  - **Pixels Quality:**
    - `sdlnchlor` standard deviation of the error,
    - `nchl` number of measurements, available each in a pack of 9.
  - **Spacial Variables:**
    - `senzr` the sensor view angle relative to the zenith,
    - `solzr` the angle of the Sun relative to the zenith,
    - `windspeed`, the speed of wind,
    - `tlg869` the specular reflection of the sea surface transmitted to the top of atmosphere,
    - `taua869` and many others.

Definition of Error

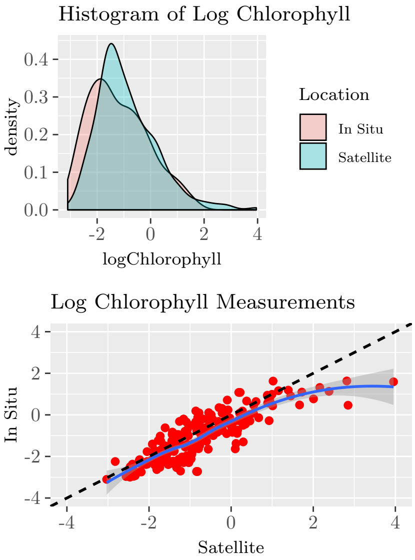
We defined it to be the difference squared of the log of the values of measurements

$$\text{Error} = \left(log\left(\text{chlorI}\right) - log\left(\text{chlorS}\right)\right)^2 = \left(log\left(\frac{\text{chlorI}}{\text{chlorS}}\right)\right)^2.$$

Distribution of Error

- It is thought that the distribution of chlorophyll-a is **log-normal**.
- Expect the error to be from a χ^2 distribution on one degree of freedom.
- May use a **Gamma distribution** to model the data.
- Though Error seems to be follow a **skewed** distribution.

Histogram of Chlorophyll-a



Requirements

- We need a suitable method of **modelling** which allows for **flexibility**
  - choice of the **distribution**,
  - **parameters** that need to be modelled,
  - **skewness** and **kurtosis** of data,
  - **smoothing** methods to be applied.
- Find the most suitable model.

A brief review of technology available to come...

Let the response variable be Y with r covariates $x_1, \ldots, x_r$ and sample size n.

Linear Regression

- In the linear regression model we assume $Y_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$, i.e.,

$$Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} + \epsilon_i$$

  for $i = 1, ..., n$, where

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- In particular, ϵ_i are i.i.d. from a **normal** distribution.
- We seek to estimate β_j for $j = 1, ..., r$ together with σ.

Estimation of Parameters

- Write $\boldsymbol{Y} = (Y_1, ..., Y_n)$. Design matrix $\boldsymbol{X}$ an $n \times (r+1)$ where $X_{i1} = 1$ and $X_{i(j+1)} = x_{ij}$ for $j = 1, ..., r$.
- We can write $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma \boldsymbol{I})$, i.e.,

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  with parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_r)$.
- **Estimate** for $\boldsymbol{\beta}$ is given through

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\beta)^{\boldsymbol{T}} (\boldsymbol{Y} - \boldsymbol{X}\beta) \Longrightarrow \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}$$

- An **unbiased** estimated for σ^2 by using $\hat{\boldsymbol{\beta}}$ is given by

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{n - r},$$

  where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}}$.

Developed 1972-1989 and Allows for

- Normal distribution to be replaced by **exponential family** of distributions,
$$Y_i \sim \mathcal{E}\left(\mu_i, \phi\right).$$
- A **link function** $g()$ is used to model the relationship of $E(Y)$ and covariates,
$$\eta_i = g(\mu_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir}.$$
- Parameter vector $\boldsymbol{\beta}$ are estimate through iteratively **weighted** least square method.
- **Exponential** family distribution $\mathcal{E}\left(\mu_i, \phi\right)$ is defined by probability **distribution** function
$$f(y \mid \mu, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right) \text{ where } \mu = b'(\theta).$$

Developed 1990-2006, a Smoothing Technique

Allows the data to determine the relationship between η and explanatory variables.

- As in GLM we have

$$\boldsymbol{Y} \sim \mathcal{E}\left(\boldsymbol{\mu}, \boldsymbol{\phi}\right).$$

- Link function $g()$ is used to model, however we assume

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + s_1(\boldsymbol{x_1}) + s_2(\boldsymbol{x_2}) + \cdots + s_J(\boldsymbol{x_J}),$$

- The terms s_j is **nonparametric** smoothing function applied to covariate $\boldsymbol{x_j}$ for $j = 1, ..., j$.
- **However**, all these methods are fixed with two parameter: location $\boldsymbol{\mu}$ and scale $\boldsymbol{\phi}$ and only regression on former.

GAMLSS 2005, Models with Skewness and Kurtosis

- The generalised additive model for location, scale and shape.
- Here we have,

$$\boldsymbol{Y} \sim \mathcal{D}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}\right),$$

  $\boldsymbol{Y}$ is from a four-parameter family of distributions.
- The parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ are related to location and shape, and $\boldsymbol{\nu}, \boldsymbol{\tau}$ are shape parameters.
- Models is extended by

$$\begin{aligned}
\boldsymbol{\eta_1} &= g_1(\boldsymbol{\mu}) = \boldsymbol{X_1\beta_1} + s_{11}(\boldsymbol{x_{11}}) + \cdots + s_{1J_1}(\boldsymbol{x_{1J_1}}),\\
\boldsymbol{\eta_2} &= g_2(\boldsymbol{\sigma}) = \boldsymbol{X_2\beta_2} + s_{21}(\boldsymbol{x_{21}}) + \cdots + s_{2J_2}(\boldsymbol{x_{2J_2}}),\\
\boldsymbol{\eta_3} &= g_3(\boldsymbol{\nu}) = \boldsymbol{X_3\beta_3} + s_{31}(\boldsymbol{x_{31}}) + \cdots + s_{3J_3}(\boldsymbol{x_{3J_3}}),\\
\boldsymbol{\eta_4} &= g_4(\boldsymbol{\tau}) = \boldsymbol{X_4\beta_4} + s_{41}(\boldsymbol{x_{41}}) + \cdots + s_{4J_1}(\boldsymbol{x_{4J_4}}).
\end{aligned}$$

GAMLSS Features

- Algorithm **maximises** a **penalised likelihood** function

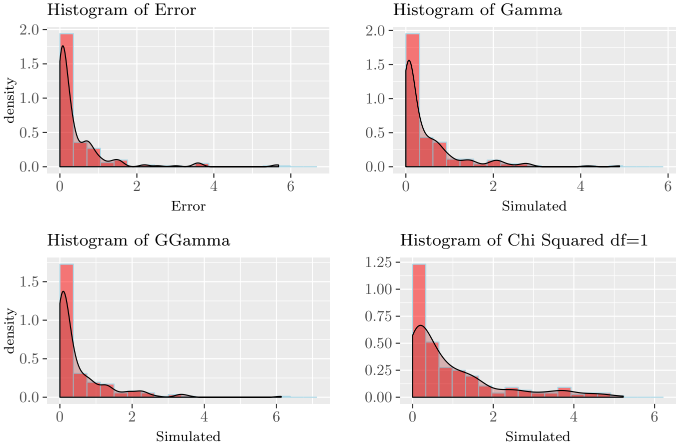
$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^T \boldsymbol{G}_{kj}(\lambda)\,\boldsymbol{\gamma}_{kj} \text{ where}$$

$$\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) = \sum_{i=1}^{n} \log f(y_i \mid \mu_i, \sigma_i, \nu_i, \tau_i).$$

- Implementation of in R `gamlss` **supports** 100 discrete, continuous, and mixed distributions.
- **Creating** new and **modifying** distributions is easy.
- Allows **linear** or **nonlinear** parametric functions, or **nonparametric** smoothing functions.
- The **additive** terms can be chosen from: P-splines, cubic splines, loess curve fitting, random effects.
- Further addition allow for **neural networks**, **decision tree**, **random effects**, **multidimensional smoother**.

We use a **Generalised Gamma** distribution as our modelling distribution, which has pdf

$$f(y|\mu,\sigma,\nu) = \frac{|\nu|}{\Gamma(\theta)} \left(\frac{\theta}{\mu^{\nu}}\right)^{\theta} y^{\theta\nu-1} \exp\left(-\frac{\theta y}{\mu^{\nu}}\right) \text{ with } \theta = \frac{1}{\sigma^2\nu^2}.$$



Methodology I

Strategy

- Start with **simple models** through `glm`, `gam`, etc...
- Find **significant** explanatory variables.
- **Compare** models through R^2 values, Akaike Information Criterion, Global Deviance, etc...
- **Check** residuals and model **diagnostic plots** for model validity.
- **Change** distribution to find a suitable one `gamlss`.
- **Regress** on all distribution parameters: location, scale, shape.

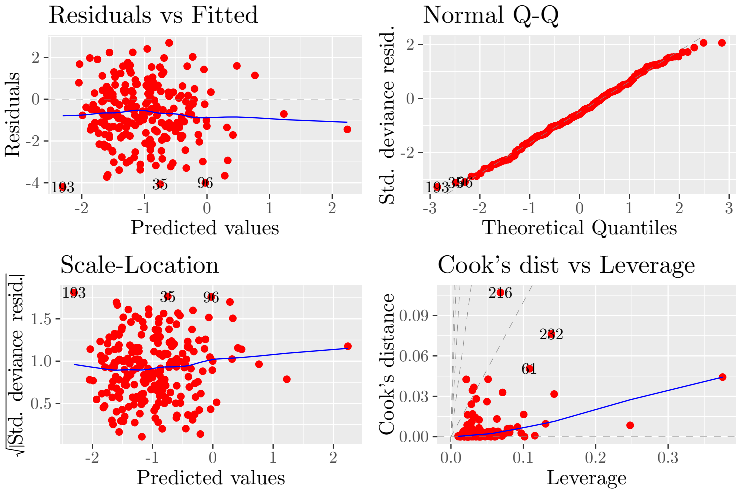
Modelling glm

```
m1<-glm(Error ~ distkm +  atimdifmin + sdlnchlor + nchlor +
     senzr + solzr + windspeed + taua869,
  family=Gamma(link="log"),
  data = matchup)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.5331	0.6742	-2.2739	0.0239
distkm	-1.1194	0.4138	-2.7051	0.0074
atimdifmin	0.0022	0.0005	4.2502	0.0000
sdlnchlor	2.9764	0.8041	3.7014	0.0003
nchlor	-0.0473	0.0498	-0.9515	0.3424
senzr	0.6653	0.3263	2.0392	0.0426
solzr	0.4934	0.4448	1.1091	0.2686
windspeed	0.0150	0.0426	0.3510	0.7259
taua869	-0.6294	1.6231	-0.3878	0.6985

Table: Coefficient Estimations

Model Checking GLM



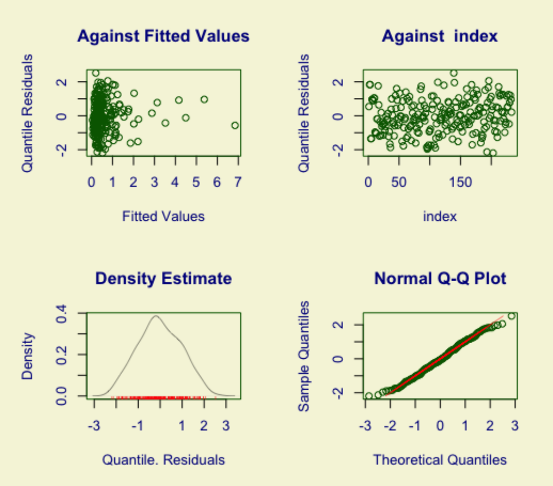
Shapiro-Wilk normality test on residuals:

$$W = 0.75227, \text{ p-value} < 10^{-16}$$

Modelling gamlss

```
m3gs<-gamlss(Error ~ cs(distkm) +  cs(atimdifmin) +
  cs(sdlnchlor) + nchlor + cs(senzr) + cs(solzr) + cs(
    windspeed) + cs(taua869),
  sigma.fo=~cs(distkm) +  cs(atimdifmin),
  nu.fo=~cs(distkm) +  cs(atimdifmin) +
  cs(sdlnchlor) + nchlor + cs(senzr) + cs(solzr) + cs(
    windspeed) + cs(taua869),
  family=GG(mu.link ="log"),
  control=gamlss.control(c.crit = 0.001, n.cyc = 40),
  data = matchup)
```

Model Checking



Shapiro-Wilk normality test on residuals:

$$W = 0.99116, \text{ p-value} = 0.1708$$

Results and Discussion

- The method was applied to a larger dataset (359 observations) with more explanatory variables
- Established a suitable model which explained around 67% variation as potentially correctable bias.
- However, the dataset still covers a limited geographical area.
- Potential models allowing for random effect can be though about.

Thank you for your attention![3]

[3] The orbiting Aqua/MODIS instrument found the above phytoplankton-brightened cyclonic eddy swirling in the Tasman Sea on the first day of November 2019.

Suggested Reading and References

See http://rpubs.com/KayvanNejabati/551817 for a summary of statistical models.

References: E. Land et al. (2018); Stasinopoulos et al. (2017).

E. Land, P., T. C. Bailey, M. Taberner, S. Pardo, S. Sathyendranath, **Nejabati Zenouz, Kayvan**, V. Brammall, J. D. Shutler, and G. D. Quartly May 2018. A Statistical Modeling Framework for Characterising Uncertainty in Large Datasets: Application to Ocean Colour. *Remote Sensing*, 10. https://doi.org/10.3390/rs10050695.

Stasinopoulos, M., R. Rigby, G. Heller, V. Voudouris, and F. De Bastiani 2017. *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman & Hall/CRC The R Series. CRC Press.