ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

BICEP
Biological Pump and Carbon Exchange Processes

DOC from space

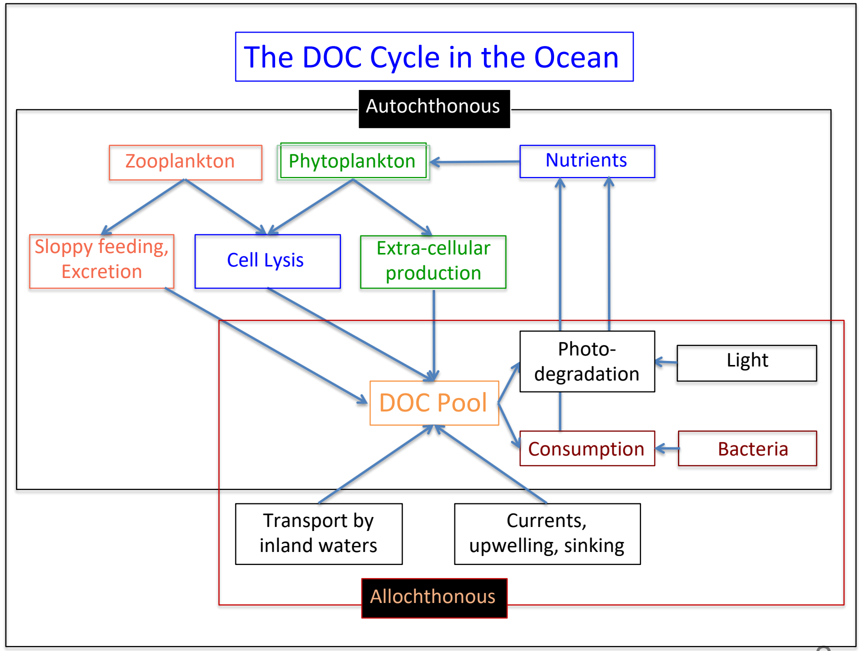
Marko Laine
Finnish Meteorological Institute

14.2.2022 **Ocean Carbon from Space Workshop**

Background

BICEP
Biological Pump and Carbon Exchange Processes

- ESA-BICEP project WP 3.4 Dissolved Organic Carbon.
- Use advanced non-linear statistical tools to elucidate the relationship between DOC and its satellite-based proxies.
- Here we concentrate on open sea DOC, using empirical approach with multiple satellite products.



ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Data sets used in the analysis

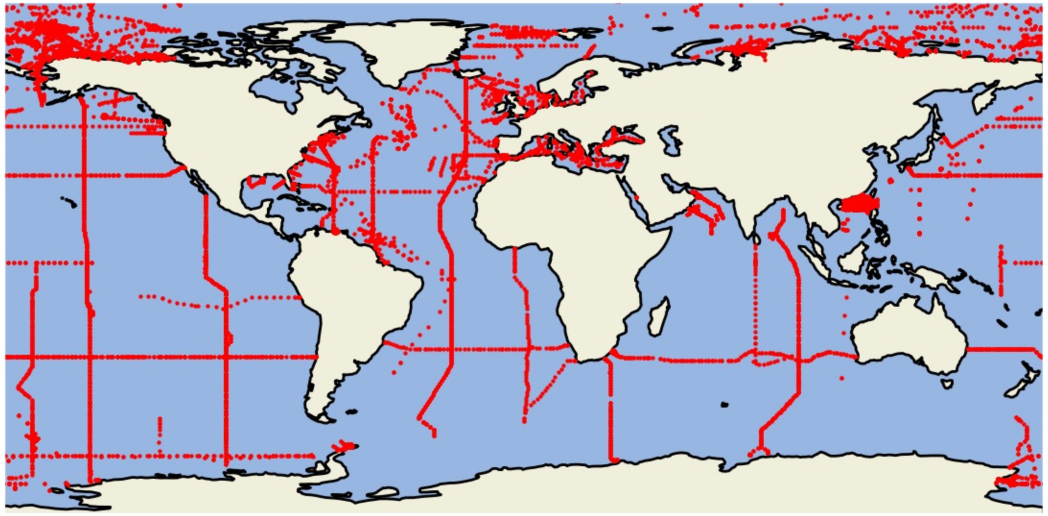
Data set	spatial resolution	temporal coverage	citation
Ocean Colour CCI	1/24°	1997–2019	Sathyendranath et al. [2019]
NPP	1/12°	1998–2018	Kulk et al. [2020]
Salinity CCI	1/12°	2010–2019	Boutin et al. [2020]
SST	1/25°	2007–2020	UK Met Office [2005]
in-situ		1994–2012	Aurin et al. [2018]
in-situ		1994–2020	Hansell et al. [2021]

Table 1: The data sets used. Except for the in-situ data, we are using monthly averaged data.

In-situ DOC data set by Hansel et al. 2021

- D. A. Hansell, et al.: Compilation of dissolved organic matter (DOM) data obtained from the global ocean surveys from 1994 to 2020, 2021. doi:10.25921/s4f4-ye35
- Merged with interpolated **monthly** Ocean Colour data and primary production.
- Total of 11200 DOC data points of which 8796 has OC and other regressor values.

In-situ DOC data locations



Different approaches to modelling DOC

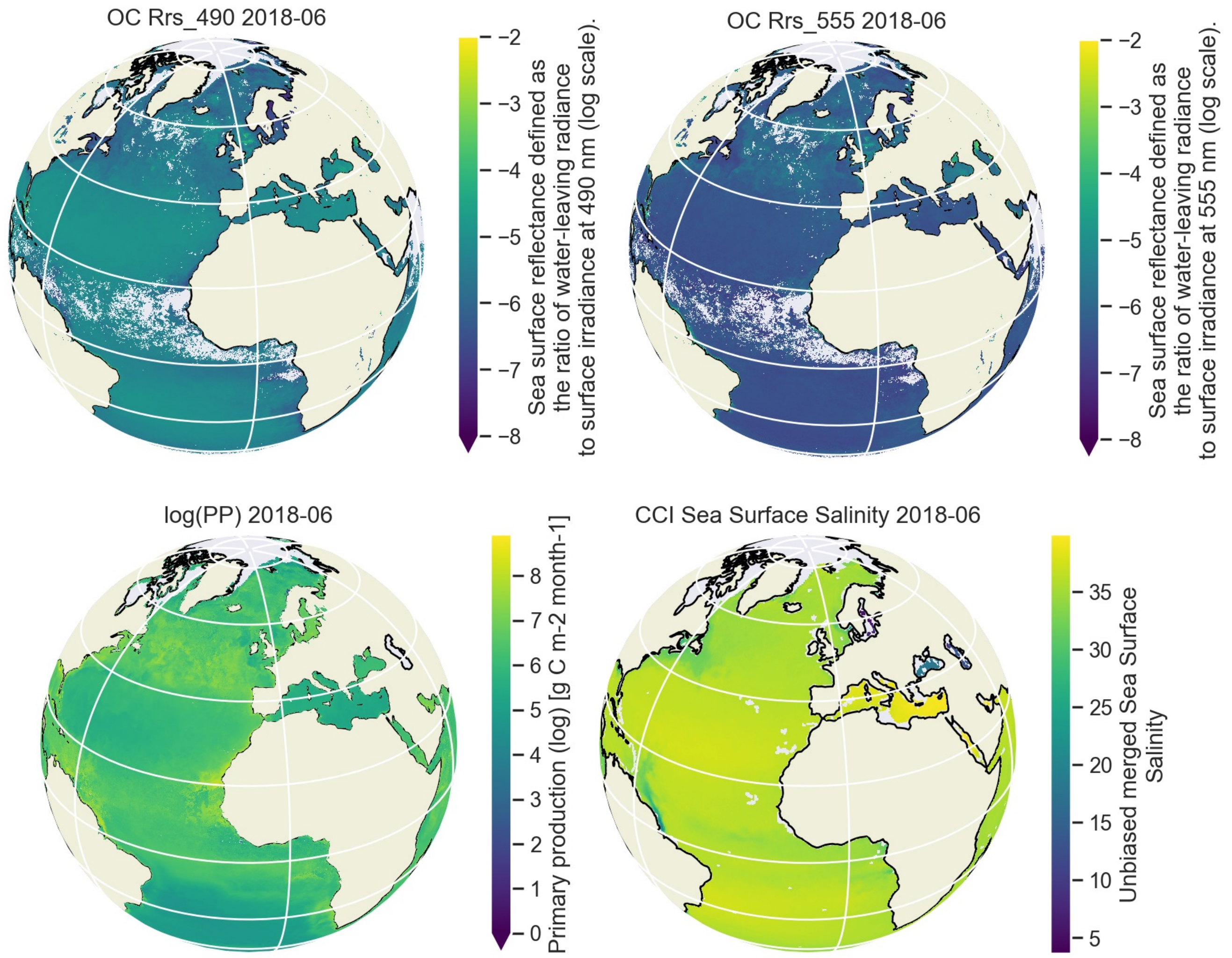
- Mechanistic/physical models (Hansel 2009, DeVries and Weber 2017).
- Statistical/empirical models (Mannino 2008, Hansel 2013, Aurin et al. 2018).
- Machine learning / AI (Aurin et al. 2018, Roshan and DeVries 2017).

- All are using more or less hybrid modelling, combining different approaches.

Empirical models for DOC using satellite based regressors

- Linear Regression.
- Neural Network.
- Gradient Boosting.
- Random Forest.

- DOC ~ Rrs_*nnn* + temp + salt + sqrt(pp) + lat + dts + depth + wclass



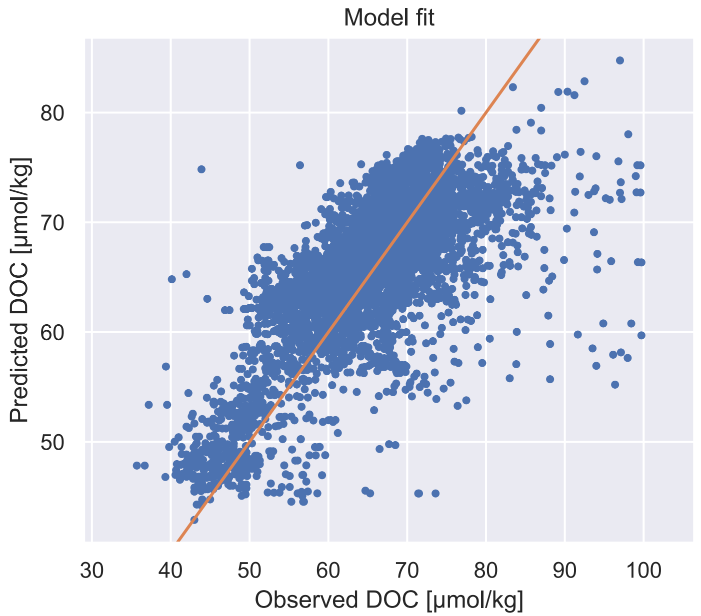
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

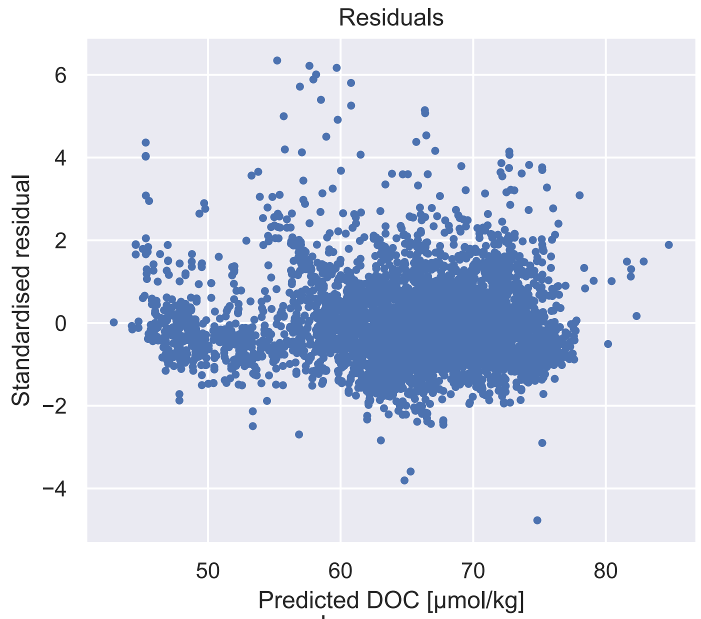
Kevin P. Murphy. Machine Learning A Probabilistic Perspective. The MIT Press, 2012.

Linear Regression

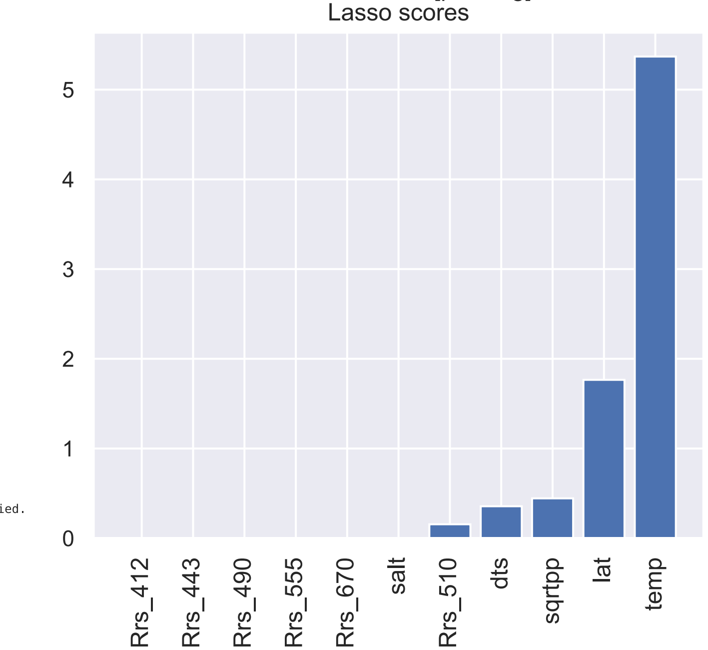
- Uses all Rrs variables, pp, salinity, temperature, latitude and distance to shore.
- R^2 = 53%.

Regression model output, > 300 km from shore, Hansel (2021) data set





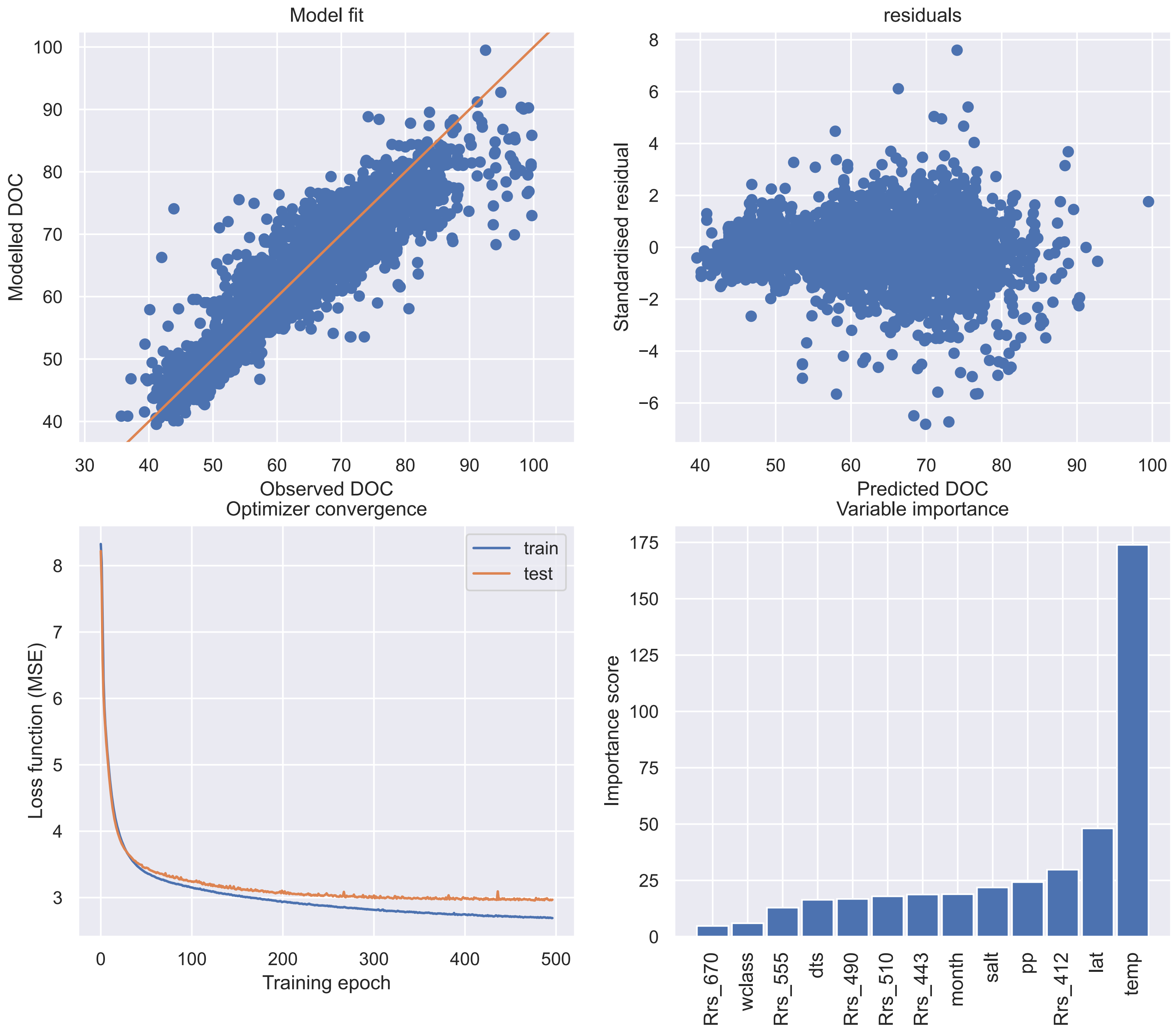

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    DOC   R-squared:                       0.529
Model:                            OLS   Adj. R-squared:                  0.528
Method:                 Least Squares   F-statistic:                     498.8
Date:                Fri, 11 Feb 2022   Prob (F-statistic):               0.00
Time:                        09:55:53   Log-Likelihood:                -16097.
No. Observations:                4897   AIC:                         3.222e+04
Df Residuals:                    4885   BIC:                         3.230e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     63.3736      0.124    512.225      0.000      63.131      63.616
sqrtpp         0.8534      0.177      4.809      0.000       0.505       1.201
salt          -1.7989      0.187     -9.597      0.000      -2.166      -1.431
temp           8.1014      0.229     35.404      0.000       7.653       8.550
dts           -1.0577      0.115     -9.178      0.000      -1.284      -0.832
lat            1.5079      0.163      9.225      0.000       1.187       1.828
Rrs_412        1.4957      0.863      1.733      0.083      -0.196       3.188
Rrs_443        2.9970      1.173      2.555      0.011       0.697       5.297
Rrs_490       -3.4711      0.888     -3.911      0.000      -5.211      -1.731
Rrs_510       -1.1438      0.644     -1.775      0.076      -2.407       0.120
Rrs_555        1.6828      0.416      4.044      0.000       0.867       2.498
Rrs_670        0.4927      0.186      2.652      0.008       0.129       0.857
==============================================================================
Omnibus:                     1069.144   Durbin-Watson:                   0.485
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3612.865
Skew:                           1.083   Prob(JB):                         0.00
Kurtosis:                       6.608   Cond. No.                         34.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



Neural Network

- Shallow net with one layer and 64 units.
- Uses water class and month as additional regressors.
- R^2: 80%

NN model output, > 300 km from shore, Hansel (2021) data set

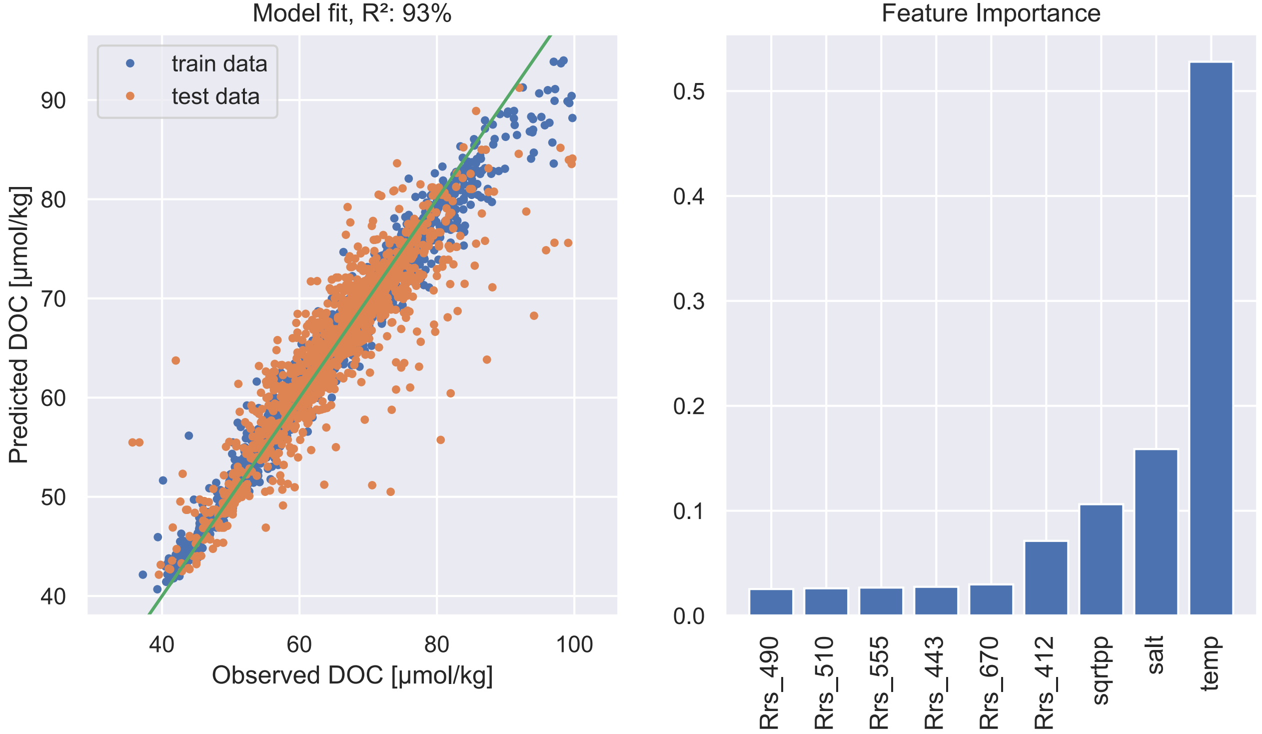


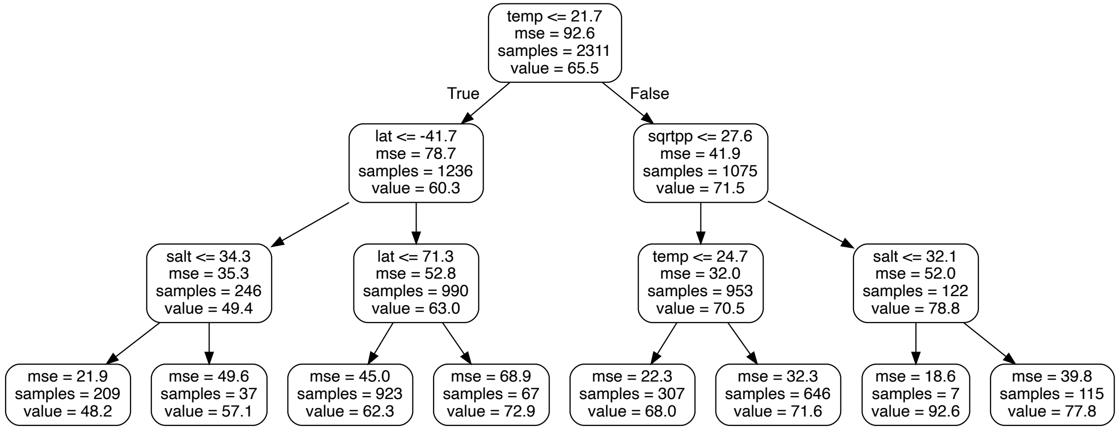
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Random Forest Regression

- Best results for both train and test data sets.
- Used to produce global DOC data set.
- R^2: 94%

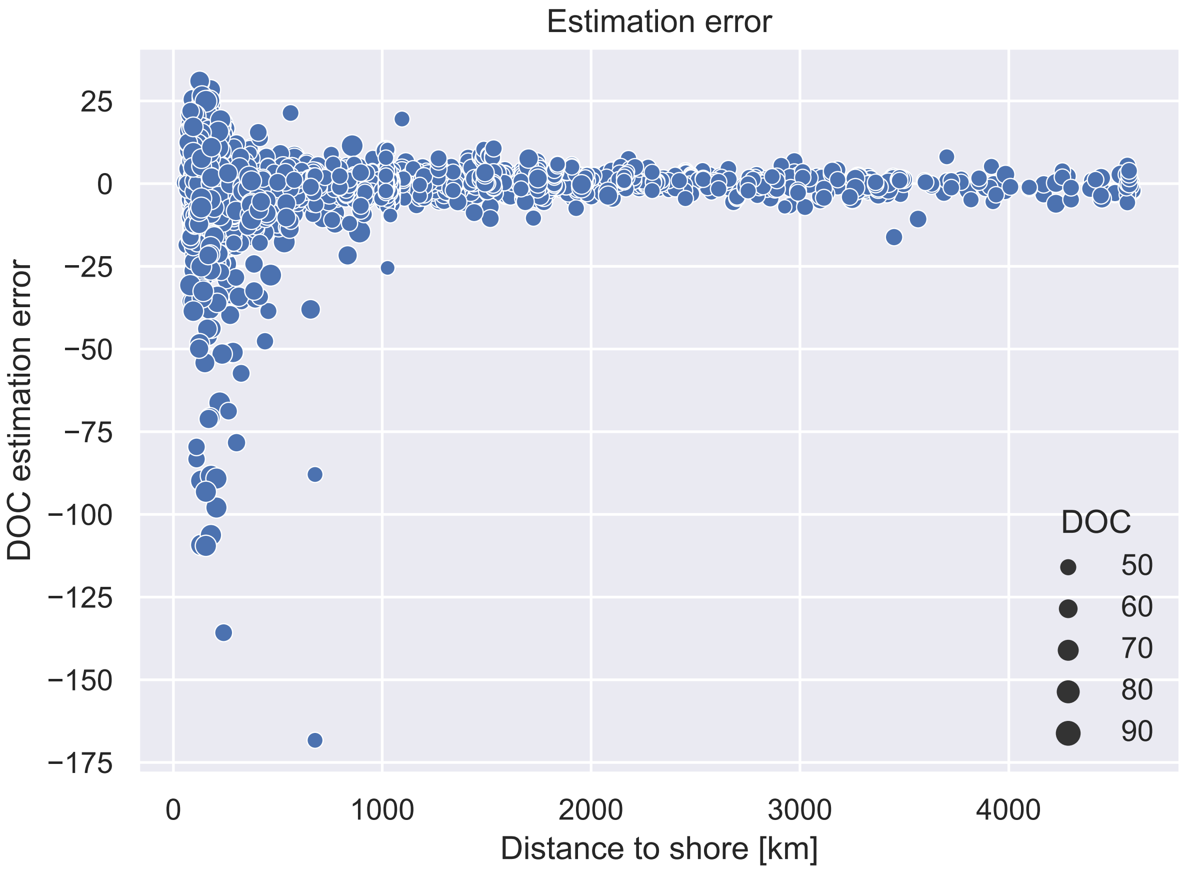
Random forest model output, > 300 km from the shore, Hansel (2021) data set

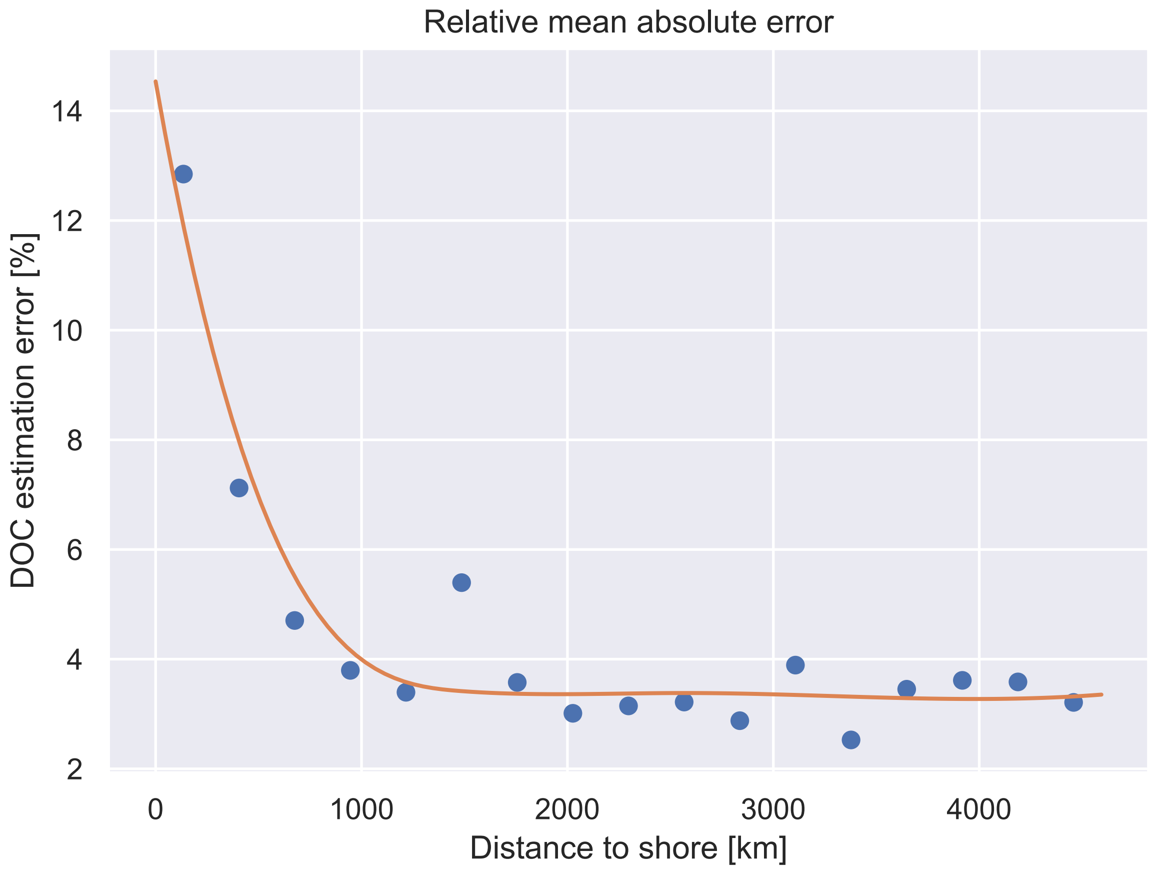




ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

How to estimate uncertainty?



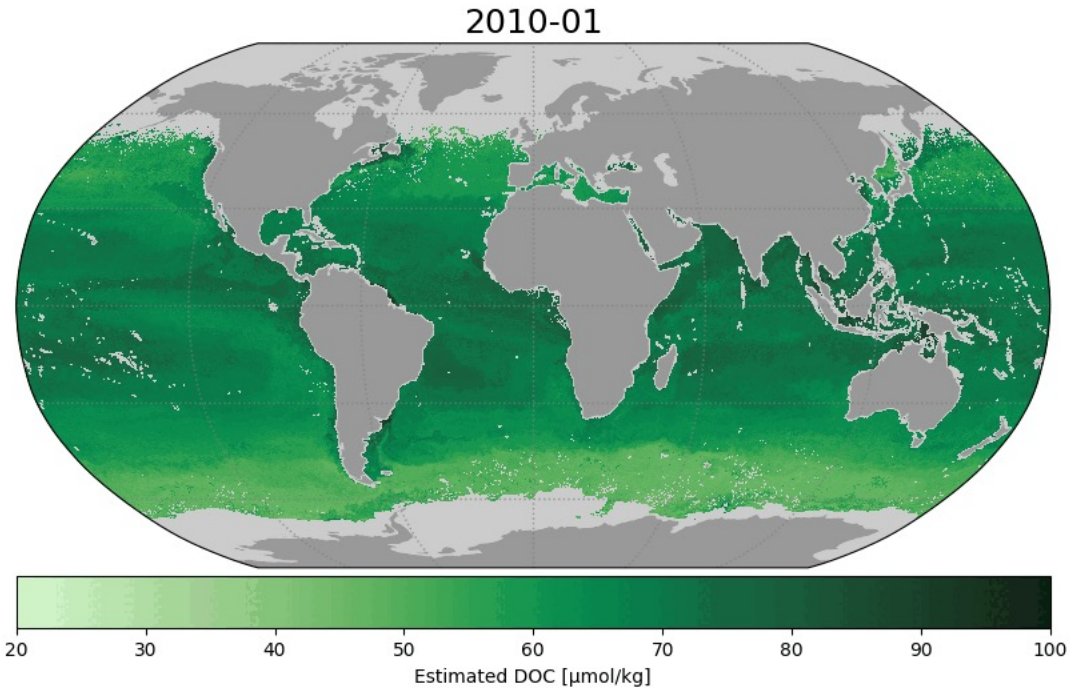


ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

Estimated monthly DOC 2010-2018

- Generated using the Random Forest model.
- 1/12° spatial resolution.
- Monthly values for 2010-2018.

- Missing values are optionally interpolated linearly.



In-situ locations for each month shown as dots

Conclusions

- The new in-situ compilation by Hansel was used with Ocean Colour and other satellite based data.
- Random forest regression provided the best results.
- Proper uncertainty quantification and validation are still the challenges.
- A hybrid empirical–physical model would perhaps be the best option here.