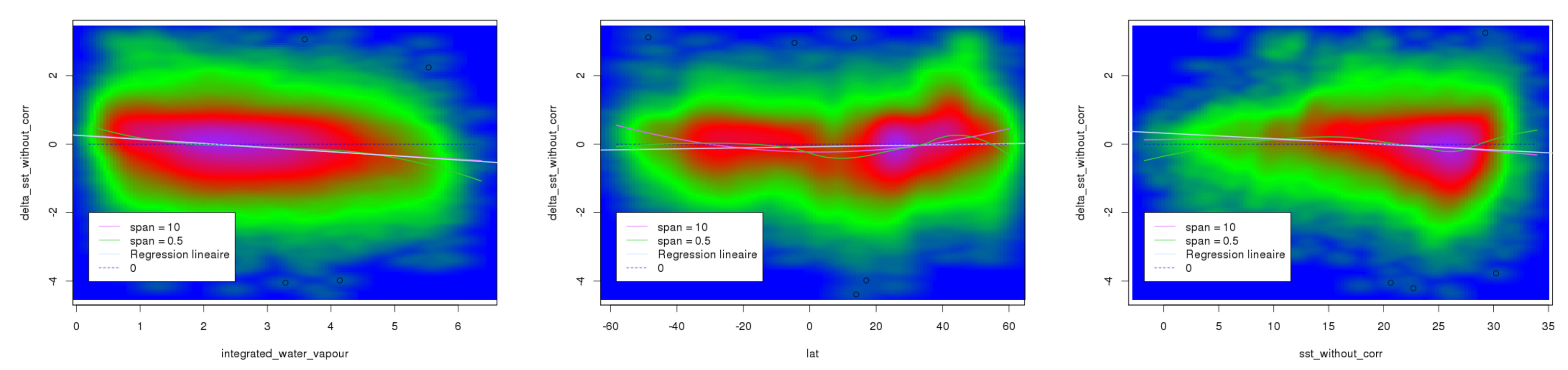
A machine learning approach for MSG/SEVIRI SST bias estimation

B. Gausset [1] S. Saux Picart [1] P. Tandeo [2] E. Autret [3]
[1]Météo-France [2]IMT-Atlantique [3]Ifremer

It is increasingly important for applications such as data assimilation or climate studies to have some knowledge about the uncertainties associated with the data being used. The GHRSST has for a long time recommended Sea Surface Temperature (SST) data producers to include Single Sensor Error Statistics (SSES) within their SST products. However there is no consensus as to which method may be used to provide SSES. They are usually understood as the mean and standard deviation of the difference between satellite retrieval and a reference.
This work is an attempt at using advanced statistical methods of machine learning to predict the bias between Ocean and Sea Ice (OSI SAF) Meteosat Second Generation (MSG) SST products and ground truth considered to be drifting buoy measurements. OSI SAF MSG current product is elaborated using a multilinear algorithm using 10.8 and 12m channels to which a correction is applied in the case of high concentration of atmospheric Saharan dusts. An algorithm correction method based on radiative transfer simulation is also used to account for seasonal and regional biases. However, for this study, the two corrections mentioned above have been removed. This was done to simplify interpretation of the results of statistical models for predicting bias in retrieved SST.

Objective

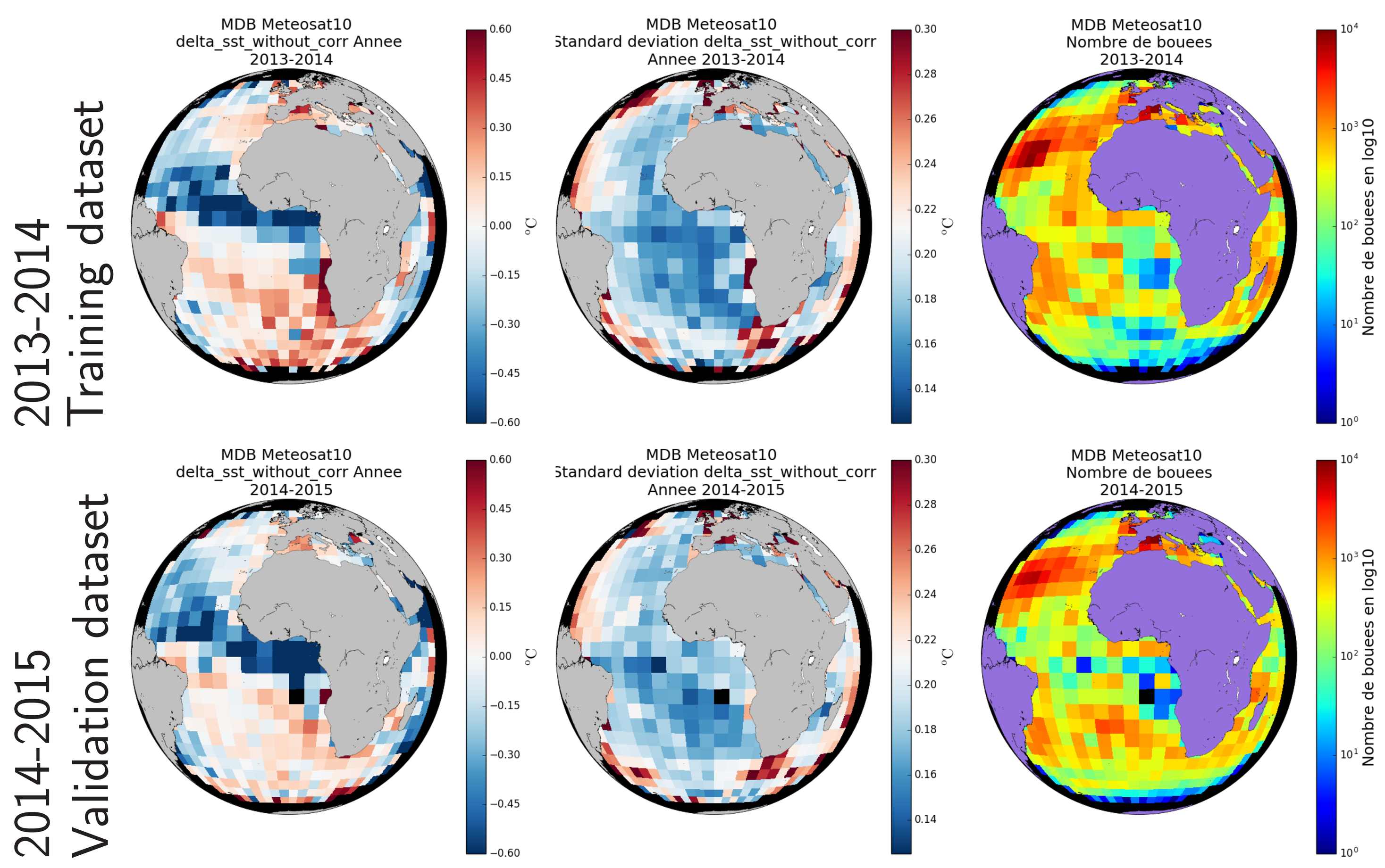
Design statistical models to represent $\Delta SST = SST_{sat} - SST_{buoys}$ with a set of explaining variables. We consider in-situ measurements from drifting buoys to be the ground truth. Such a model, once defined, could be used to operationally adjust estimates of SST.



ΔSST plotted against integrated water vapour (left), latitude (centre) and SST (right).

Data

This study uses a matchup dataset for Meteosat 10 satellite from August 2014 to July 2015. It regroups collocations between drifting buoy observations and SST estimated from SEVIRI instrument together with ancillary information such as model outputs. This matchup dataset include nearly 250000 entries.



ΔSST mean (left), STD (centre) and number of points (right).

Selected variables are:

Name	Description
Latitude	Measurement latitude
Wind speed	Near surface wind speed (ECMWF)
Solar zenith angle	Angle between zenith and sun position
Satellite zenith angle	Angle between zenith and satellite position
Integrated water vapour	Integrated water vapour in the atmosphere
IR_039 IR_087	Difference between channel 3.9μm and 8.7μm averaged in 5×5 pixels box
IR_108 IR_120	Difference between channel 10.8μm and 12.0μm smoothed by 5×5 pixels box
Number of valid pixels	Number of valid retrieval (quality level 3, 4 or 5) in 5×5 pixels box
SST STD	Standard deviation of SST in 5×5 box
SST	SST retrieved from SEVIRI

Methodology

Several models were tested:

- Simple linear regression: $\Delta SST = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_i + \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_{i,j} X_i X_j$
- LASSO (Least Absolute Shrinkage and Selection Operator): same as above but some α_i are null.
- GAM (Generalized Additive Model):
  $\Delta SST = \alpha_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{i=1}^{p}\sum_{j=1}^{p} f_{i,j}(X_i X_j)$, where f_i and $f_{i,j}$ are non-linear functions adjusted by local linear regression.
- Random Forest: $\Delta SST = \frac{1}{N}\sum_{i=1}^{N} t_i(X_1, \ldots, X_p)$, where t_i are regression trees.

Evaluation of the models: adjusted R^2.

$$R^2_{adj} = 1 - \frac{\mathrm{Var}_{\mathrm{error}}}{\mathrm{Var}_{\mathrm{total}}}$$

$$\mathrm{Var}_{\mathrm{error}} = \frac{\sum(\Delta SST - \widehat{\Delta SST})^2}{n-p-1}$$

$$\mathrm{Var}_{\mathrm{total}} = \frac{\sum(\Delta SST - \overline{\Delta SST})^2}{n-1}$$

where $\widehat{\Delta SST}$ is the model estimate of ΔSST, $\overline{\Delta SST}$ is the average of ΔSST, n is the number of observation and p is the number of explaining variables.

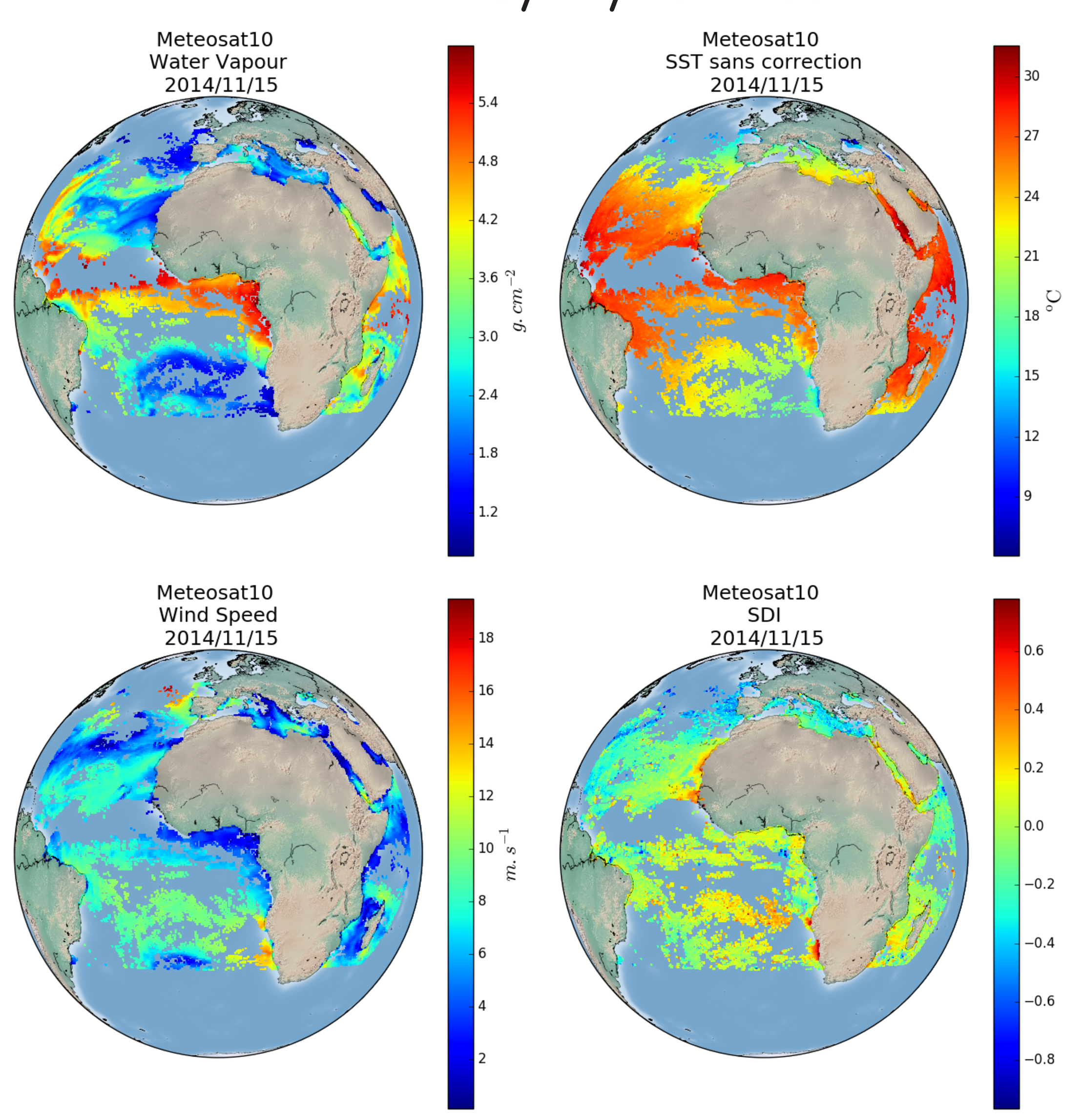
Results

Testing the models:

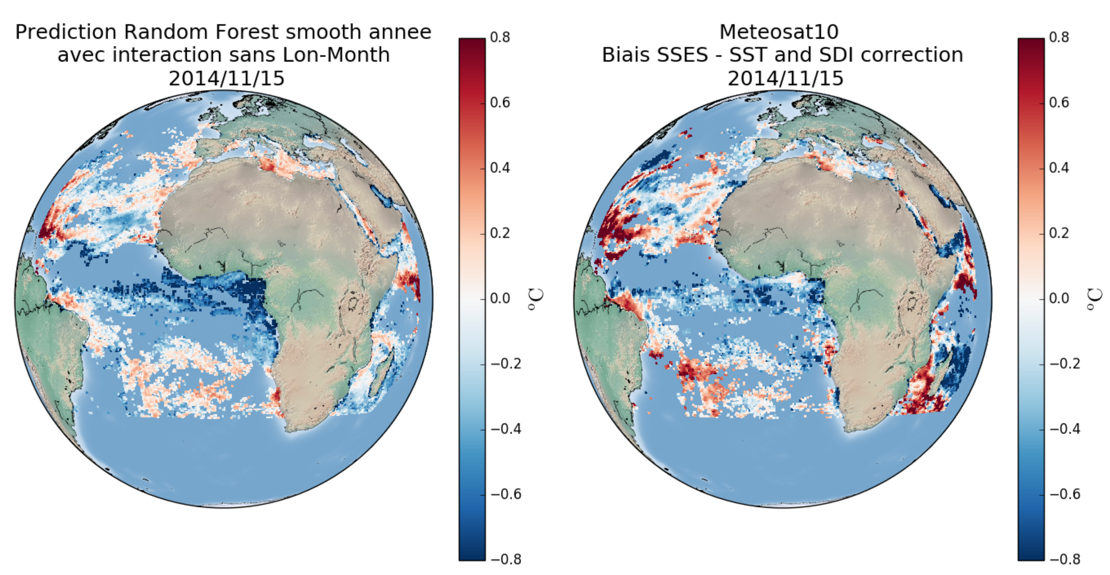
	Rg. Lin.	LASSO	Random Forest	GAM	SVM
R^2_{adj}	20,69 %	24,43 %	30,96 %	28,44 %	29,29 %

An example using random forest model: 15/11/2014 00h.

Water vapour (top-left), SST (top-right), wind speed (bottom-left), Saharan Dust Index (bottom-right).



Predicted bias from random forest model (left), SSES bias (in operational products, right).



Conclusion

- Best results were obtained with random forest model.
- Random forest is fast to run, which is potentially very interesting in an operational context.
- Amplitude of the modelled bias is lower than operational SSES.
- Some more in situ data would need to be included in areas were drifting buoys don't go (for example Namibia coast).
- Only night-time analysis has been conducted: we don't know how the model will cope with diurnal warming conditions.

stephane.sauxpicart@meteo.fr