Deriving minimal sea surface temperature monitoring networks from remote sensing data using coherency analysis

Francesco Finazzi*
University of Bergamo, Bergamo, Italy - francesco.finazzi@unibg.it

Marian Scott
University of Glasgow, Glasgow, UK - marian.scott@glasgow.ac.uk

Abstract

This paper proposes a methodology to synthesize sea surface temperature data collected by remote sensing, in order to identify homogeneous areas of the ocean surface and to derive a minimal network for point monitoring. A recently introduced clustering algorithm is adopted, which is based on state-space modelling and which enables clustering of millions of time series with respect to their temporal pattern.

When the clustering algorithm is applied to sea surface temperature time series and the clusters mapped in space, we observe that the ocean surface divides into a relatively small number of clusters, with time series in each cluster sharing the same temporal pattern.

In this work, sea surface temperature clustering for the North Atlantic basin and time period 2003-2009 is provided. The clustering result is used to define the minimal network, in terms of number of buoys and their spatial locations. The minimal network derived following this approach consists of 25 nodes, each located within a cluster. Spatial representativeness of data collected using such network has been validated using remote sensing data from 2010 to 2012, computing bias and root mean squared prediction errors over space.

Keywords: climate; essential climate variables; clustering; remote sensing

1. Introduction and aims

Sea surface temperature (SST) is recognised as an essential climate variable by the Global Climate Observing System programme. Therefore, SST is extensively measured at the global level using both monitoring networks (buoys) and satellites (Emery, 2015). The buoy systems measure SST at a few thousand spatial locations scattered over the ocean surface. The remote sensing data, on the other hand, are available at a daily or weekly level and with high spatial resolution (e.g. 9 km grid cell). Remote sensing data sets are thus very large, with several million time series.

Temporal variability of SST is induced by many causes, some of them well known and others not fully understood or yet discovered. In this regard, remote sensing data represent a precious source of information as they cover the entire globe and they should help deepen the mechanisms at the basis of the SST variability. When remote sensing data sets are considered, however, it may not be easy to extract useful information as data are dense in both space and time. Spatially, SST data may also be redundant as points of the ocean only a few kilometers away are likely to behave in a very similar way. The aim of this work is twofold: first, to understand whether the ocean surface can be partitioned into homogeneous areas with respect to the SST temporal pattern, and second, to define a minimal network of buoys for point monitoring. The main idea is that, if the partition exists, than observations at the network nodes can well describe the SST variability over the entire ocean surface, with possibly a small error.

2. Data set

SST data are taken from the “MODIS Terra Level 3 SST Thermal IR Nighttime” product (persistent ID: PODAAC-MODST-T8D9N), covering the entire ocean surface with around 9 km spatial resolution and 8 day temporal resolution. In this work we focus on the North Atlantic basin and the temporal period $2003-2012$, for a total of 460 time steps. For each time step, data are provided as a gridded map with possible missing observations.

As data will be clustered with respect to their temporal pattern, it is useful to consider them as spatially registered time series, for a total of around $488'000$ time series covering North Atlantic. Time series with

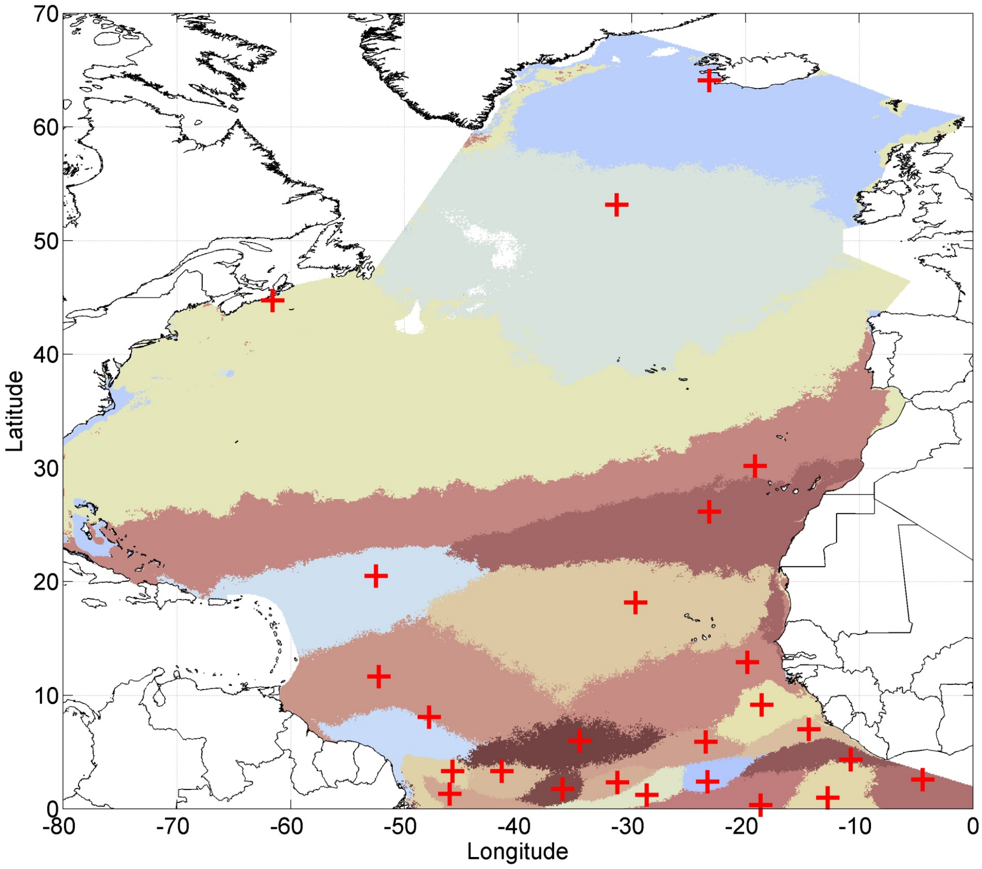

Figure 1: North Atlantic basin partition with respect to SST temporal coherence and minimal monitoring network (red crosses).

more than 60% of missing observations are removed from the data set as they may produce artifacts in the analysis results.

3. Methodology

This section describes the statistical method adopted to partition SST time series, the strategy to derive the minimal monitoring network and the validation approach. In order to validate the network using independent data, the data set described above is partitioned into the estimation data set $\mathbf{Y}_E$, from 2003 to 2009, and into the validation data set $\mathbf{Y}_V$, from 2010 to 2012.

3.1 Ocean surface partition

Let $y_{B_i}(t)$, $t = 1, ..., T$, be the time series of remote sensing SST measurements at the generic pixel $B_i \in \{B_1, ..., B_N\}$. The model-based clustering method introduced in Finazzi et al. (2015) is used here to cluster the time series of $\mathbf{Y}_E$ with respect to their temporal coherency. In particular, a group of time series are said to be jointly coherent when, apart from random noise, they share the same temporal pattern along the entire temporal frame of observation. As only the temporal pattern is of interest, each time series is standardized with respect to its own mean and variance.

The result provided by the clustering method consists of the number of clusters $\hat{p}$, the average time series $\hat{z}_j(t)$, $j = 1, ..., \hat{p}$ and the cluster membership of each time series $y_{B_i}(t)$. This gives rise to a partition of the

Data set	$\mathbf{Y}_E$		$\mathbf{Y}_V$	
Temporal period	2003-2009	2010	2011	2012
Bias - 1th perc.	0.00	-1.27	-0.95	-1.29
Bias - 50th perc.	**0.00**	**-0.09**	**-0.07**	**-0.26**
Bias - 99th perc.	0.00	0.93	1.02	0.93
RMSE - 1th perc.	0.37	0.36	0.33	0.40
RMSE - 50th perc.	**0.70**	**0.81**	**0.73**	**0.82**
RMSE - 99th perc.	2.00	2.31	2.30	2.40

Table 1: Minimal monitoring network validation. Percentiles of the bias and RMSE distributions over the 488'000 pixels and different temporal periods. All results in $°C$.

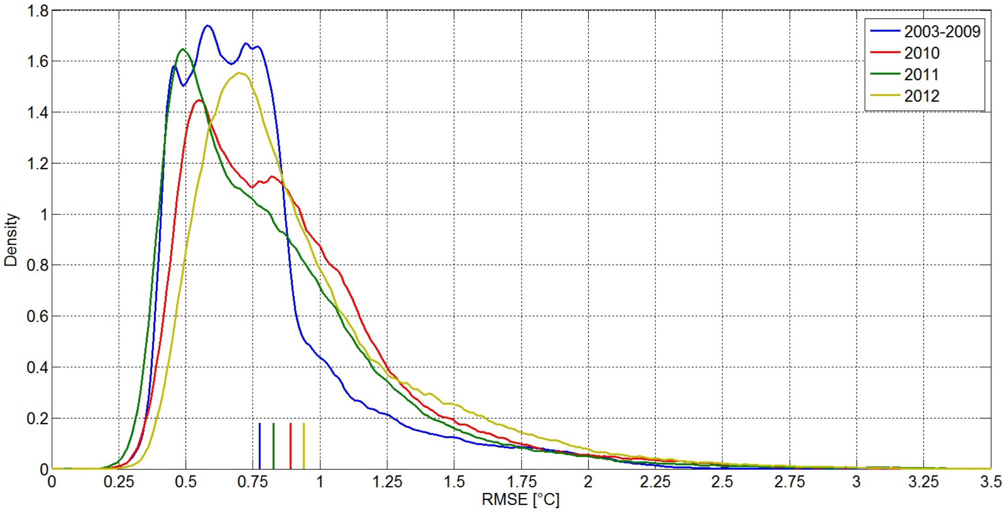

Figure 2: Density functions of the SST reconstruction RMSE for the North Atlantic basin and different temporal periods. Expected value of the distributions depicted by the vertical segments at the base of the x axis.

ocean surface as each $y_{B_i}(t)$ is assigned to one and only one cluster. Also note that each $\hat{z}_j(t)$ is related to a cluster but not to a precise point/pixel in space.

3.2 Minimal network

As the time series in each cluster share the same temporal pattern, the minimal network is simply derived selecting a reference pixel in each cluster. In particular, for each cluster j, the temporal correlation between $\hat{z}_j(t)$ and each $y_{B_i}(t)$ in the cluster is computed and weighted with respect to the fraction f_{B_i} of non-missing observations. The reference pixel obtained as

$$\hat{B}_j = \underset{B_i \in \mathcal{B}_j}{\arg\max}\, corr\left(y_{B_i}(t), \hat{z}_j(t)\right) \cdot f_{B_i}$$

where $\mathcal{B}_j$ is the set of pixels in the j-th cluster. It follows that the minimal network of $\hat{p}$ nodes is given by $\hat{\mathcal{B}} = \left\{\hat{B}_1, ..., \hat{B}_{\hat{p}}\right\}$.

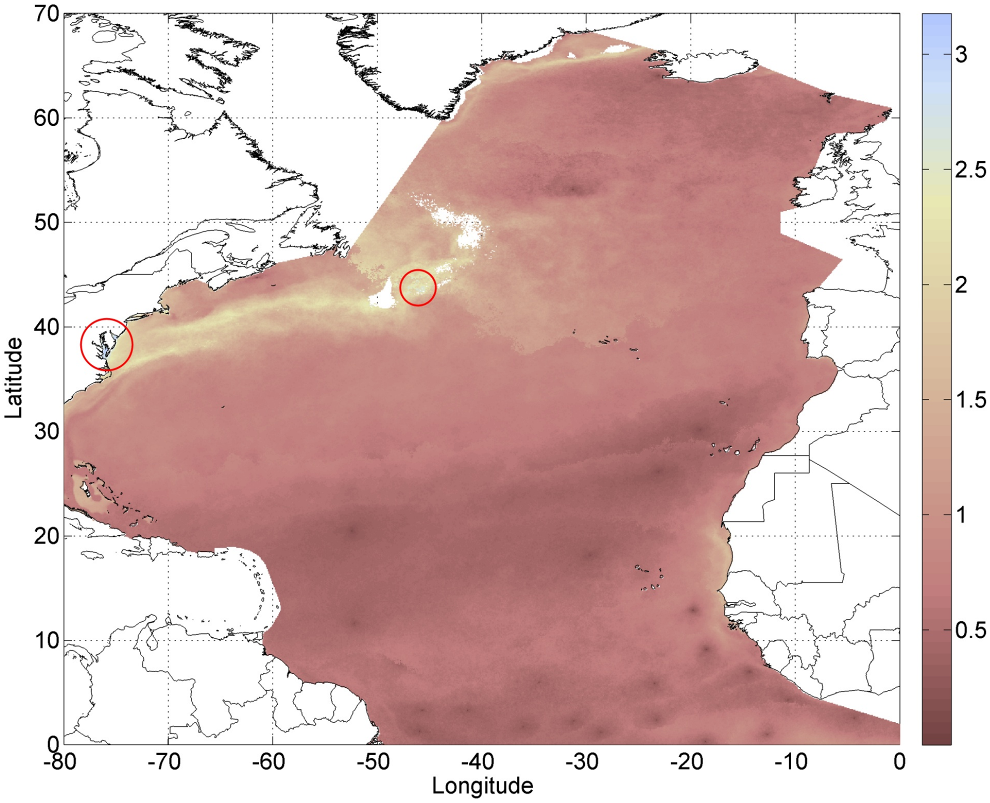

Figure 3: Spatial distribution of the SST reconstruction RMSE for the North Atlantic basin and the temporal period 2003 – 2009. Red circles depicts areas with RMSE higher than 3 $°C$

3.3 Network validation

The estimated minimal network $\hat{\mathcal{B}}$ is validated using both $\mathbf{Y}_E$ and $\mathbf{Y}_V$. First of all, for each $\mathcal{B}_j$ and each (non-standardised) time series $y_{B_i}(t)$, the following regression model is estimated

$$y_{B_i}(t) = b_{0,B_i} + b_{1,B_i} \cdot y_{\hat{B}_j}(t) + \varepsilon_{B_i}(t)\,,\ B_i \in \mathcal{B}_j$$

with b_{0,B_i} and b_{1,B_i} the regression coefficients and $\varepsilon_{B_i}(t) \sim N\left(0, \sigma^2_{B_i}\right)$. In practice, the time series $y_{\hat{B}_j}(t)$ at the reference pixel $\hat{B}_j$ is used to describe all the other time series in cluster j through a simple regression model with spatially varying coefficients. Bias and root mean squared error (RMSE) are eventually assessed at each B_i as

$$\begin{aligned} bias_{B_i} &= E\left[y_{B_i}(t) - \hat{y}_{B_i}(t)\right], && (1)\\ RMSE_{B_i} &= E\left[\left(y_{B_i}(t) - \hat{y}_{B_i}(t)\right)^2\right] && (2) \end{aligned}$$

where

$$\hat{y}_{B_i}(t) = \hat{b}_{0,B_i} + \hat{b}_{1,B_i} \cdot y_{\hat{B}_j}(t)\,,\ B_i \in \mathcal{B}_j.$$

Note that $\hat{b}_{0,B_i}$ and $\hat{b}_{1,B_i}$ are estimated using the outputs of the clustering approach when applied to $\mathbf{Y}_E$. When $\mathbf{Y}_V$ is considered, prediction bias and prediction RMSE are computed in the same way of (1) and

(2), considering the same minimal network $\hat{\mathcal{B}}$ and the same regression coefficients estimated on $\mathbf{Y}_E$. To all the effects, regression coefficients are associated with the network and they allow to reconstruct SST at each pixel using measurements at the network nodes.

4. Results

Clustering of the SST time series for the North Atlantic basin and the temporal period $2003-2009$ has been carried out using the D-STEM software (Finazzi and Fassò, 2014) and required 16 hours of computing time on a server machine equipped with 128 GB of RAM and 16 CPU cores.

Ocean surface partition and minimal monitoring network are depicted in Figure 1. The partition consists of 25 clusters and they are characterized by a quite compact structure in space. Following the strategy described above, network node spatial locations are identified within each cluster.

Bias and RMSE are computed for each of the $488'000$ pixels covering the basin, for the time period $2003-2009$ and for each year from 2010 to 2012. Results are reported in Table 1 in terms of three percentiles of the bias and the RMSE distributions. Considering the estimation data set $\mathbf{Y}_E$, SST time series of the North Atlantic basin are reconstructed with a median RMSE equal to 0.7 $°C$. As expected, both bias and RMSE slightly increase when the validation data set $\mathbf{Y}_V$ is considered but the minimal monitoring network seems to well describe SST even during the period $2010-2012$.

Figure 2 depicts the probability density functions of the RMSE obtained by kernel-smoothing the RMSE observed over the pixels. Distributions related to the validation period have a heavier tail but the average RMSE never exceeds 1 $°C$, with 2011 performing better than 2010 (see also Table 1).

Figure 3 shows the spatial distribution of the RMSE for the period $2003-2009$. It is possible to note that the RMSE is low for most of the basin with the exception of the ocean area affected by the Gulf Stream, where RMSE is around 2.2 $°C$. This value has to be compared with the SST standard deviation in the same area which ranges between 7 and 11 $°C$.

5. Conclusions

Preliminary results obtained for the North Atlantic basin are promising and they will help to derive a minimal network for SST monitoring at the global level. Prediction bias and RMSE assessed using validation data sets are particularly useful as they can help to identify areas not well represented, and thus to improve the network by eventually adding nodes where needed. Globally, the SST data set includes around 5 million time series and the clustering algorithm will be optimized to deal with this more challenging problem.

References

Emery, W.J. (2015). Sea Surface Temperature. Encyclopedia of Atmospheric Sciences, Air Sea Interactions, 136-143.

Finazzi, F. & Fassò, A. (2014). D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. Journal of Statistical Software, 62(6).

Finazzi, F., Haggarty, R., Miller, C., Scott, M. & Fassò, A. (2015). A comparison of clustering approaches for the study of the temporal coherence of multiple time series. Stochastic Environmental Research and Risk Assessment, 29(2), 463-475.