**Stochastic Environmental Research and Risk Assessment**
August 2011, Volume 25, Issue 6, Pages 793-804
http://dx.doi.org/10.1007/s00477-010-0442-8






# Linear Gaussian state-space model with irregular sampling: application to sea surface temperature

Pierre Tandeo[a, *], Pierre Ailliot[b, *] and Emmanuelle Autret[a, *]

[a] Laboratoire d'Océanographie Spatiale, IFREMER, Plouzané, France
[b] Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France

*: Corresponding authors : Pierre Tandeo, email address : pierre.tandeo@ifremer.fr , Pierre Ailliot, email address : pierre.ailliot@univ-brest.fr , Emmanuelle Autret, email address : emmanuelle.autret@ifremer.fr

**Abstract :**

Satellites provide important information on many meteorological and oceanographic variables. Statespace models are commonly used to analyse such data sets with measurement errors. In this work, we propose to extend the usual linear and Gaussian state-space to analyse time series with irregular time sampling, such as the one obtained when keeping all the satellite observations available at some specific location. We discuss the parameter estimation using a method of moment and the method of maximum likelihood. Simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the Expectation–Maximisation algorithm, which is combined with a standard numerical optimization procedure to maximize the likelihood function. The model is validated on sea surface temperature (SST) data from a particular satellite. The results indicate that the proposed methodology can be used to reconstruct realistic SST time series at a specific location and also give useful information on the quality of satellite measurement and the dynamics of the SST.

**Keywords :** State-space model - Irregular sampling - Ornstein–Uhlenbeck process - EM algorithm - Sea surface temperature

## 1. Introduction

Sea surface temperature (SST) is an important oceanographic variable for many applications (see e.g. [7] and references therein). Several satellites and buoy networks provide continuous observations of this variable, leading to a huge amount of data. Statistical methods are then needed to combine all this information and provide realistic SST analysis at any date and any location in the ocean.

State-space models provide a exible methodology for analysing such complex environmental data sets, and they have already been used in a wide range of problems (see

e.g. [13]) including meteorological and oceanographic applications (see e.g. [1], [11], [25] and [16]). The basic idea of these models consists in introducing the "true" value of the physical variable of interest as a hidden variable (the "state"). Then, stochastic models are used both to describe the dynamics of the state and to relate the observations to the state. When linear Gaussian models are used, we get the so-called linear Gaussian state-space model which has been extensively studied in the literature (see e.g. [8] and references therein). Note that [14] proposed unified notations for state-space models and data assimilation in oceanography and meteorology which are partially adopted here.

In this work we analyse satellite SST data at a single location, where buoy data is available for comparison, and we consider the time series obtained by keeping all the satellite data available nearby this location. It leads to a time series with irregular time-step, with generally several data each day but also sometimes gaps of several days with no data. We adopt a continuous-time state-space model to analyse this time series in which the state is supposed to be an Ornstein-Uhlenbeck process. It leads to a simple generalization of the usual linear Gaussian state-space model with regular time-step.

The most usual method for estimating the parameter in models with latent variable consists in computing the maximum likelihood estimates using the Expectation-Maximisation (EM) algorithm. In this work, we propose to improve the numerical efficiency of the EM algorithm by combining it with a method of moment and a standard numerical optimization procedure. The method of moment is used to provide realistic starting values to the EM algorithm with the extra benefit of providing graphical tools which permit to asses the realism of the model. The standard numerical optimization procedure is used to accelerate the convergence of the EM algorithm near the maxima and provide estimates of the observed information matrix and thus important information on the variance of the estimates.

The paper is organised as follows. The SST data and the model are introduced in Section 2. Then, the parameter estimation is discussed in Section 3: after describing the practical implementation of the various methods, we asses the efficiency of the whole procedure through simulations. In Section 4, we discuss the results obtained on the data with the proposed methodology. Conclusions are drawn in Section 5.

# 2 Data and model

Several instruments on-board satellites provide measurements of SST over the entire surface of the ocean with different spatial and temporal resolutions. In this work, we focus on the data provided by the infrared Advanced Very High Resolution Radiometer instrument on-board the METOP satellite (see [17] for more details). This satellite covers the global ocean with a spatial resolution of 0.05 degree and provides two SST observations per day at the most in optimal conditions. In this paper, we first consider the data available at a given location, with geographical coordinates ($0^0$N, $23^0$W), in the tropical region of the Atlantic Ocean. More precisely, we consider two years of data, from 11-Jul-2007 to 18-Jun-2009, which are representative of the variability of the SST conditions at this location. Hereafter, $(t_1, ..., t_n)$ denotes the times at which the METOP satellite data are

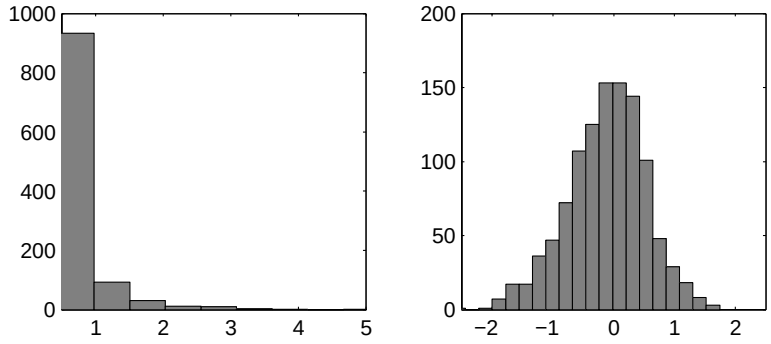


**Figure 1:** Histogram of the time lags $\Delta_i$ in days (left) and of the SST anomalies $\{y_{t_i}\}$ in degree Celsius (right).

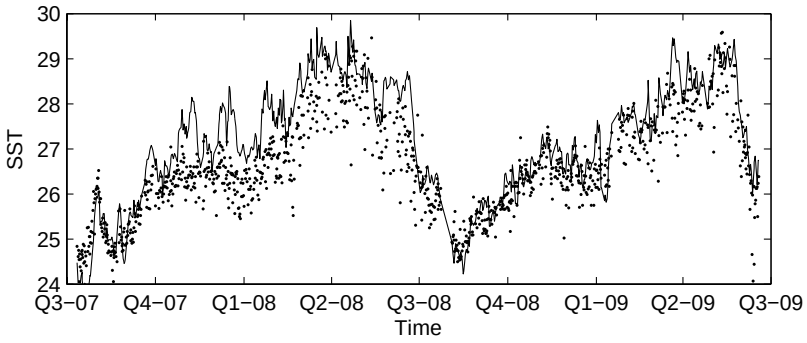


**Figure 2:** Raw METOP SST (in degree Celsius) time series (dotted line) and OIV2 SST analysis (full line).

available, with $n = 1087$ the total number of observations. Since satellite observations may be contaminated by atmospheric conditions (e.g. cloud coverage), some data are missing and the time difference $\Delta_i = t_i - t_{i-1}$ between two consecutive observations may vary from a half day to a few days (see Figure 1).

The resulting time-series is clearly non-stationary (see Figure 2) with in particular important seasonal components. The non-stationary components have complex features and we could not find any appropriate parametric model to describe them. We have thus decided to use the SST analysis produced by the National Climatic Data Center (NCDC) (daily "OIV2 analysis" with 0.25 degree spatial resolution) to remove these components. These analysis are derived from different satellite sources independent of METOP data (see [21]) and we assume that they provide a good estimate of the low-variations of the SST conditions. Both data sources METOP and OIV2 are available at the URL *http://www.hrdds.net*.

Then we consider the time series $y_{t_1}^{t_n} = (y_{t_1}, ..., y_{t_n})$ obtained by removing the OIV2 analysis from the METOP data (see Figure 3). We assume that this new time-series, referred as the SST anomaly hereafter, is a discrete-time realization of a continuous-time stationary process $\{Y_t\}$. Modelling the time series $\{y_{t_1}^{t_n}\}$ may provide important information on the small scale variability of SST and also on the quality of METOP

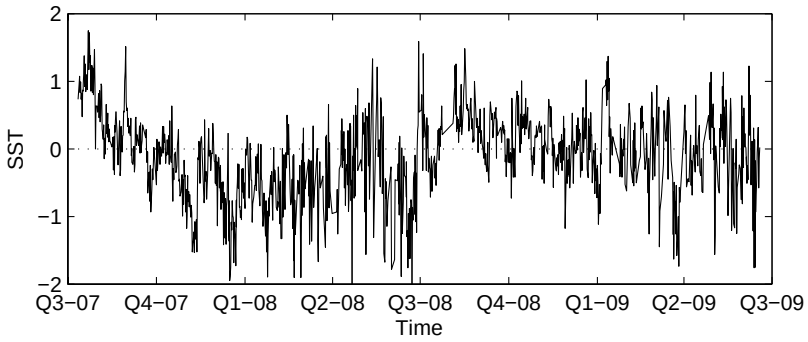


**Figure 3:** SST anomalies (in degree Celsius) obtained by removing the OIV2 analysis from METOP data.

measurements and OIV2 analysis as it will be shown in Section 4 and finally lead to a better assimilation of these data into numerical models.

The model that we consider for $\{Y_t\}$ is introduced below. First, we assume that the observed SST anomaly at time $t$, $Y_t$, is related to the "true" SST anomaly at time $t$, denoted $X_t$, by the measurement equation below:

$$Y_t = HX_t + \sqrt{R}\epsilon_t \tag{1}$$

where $\{\epsilon_t\}$ is a Gaussian white noise sequence with zero mean and unit variance. In practice $R$ represents the variance of the observation error and $H$ allows a transformation between the state and the observations. For the particular METOP measurements considered in this paper (we keep only the best quality data), the standard deviation of the observation error has been estimated globally to 0.5 degree Celsius, but it is known that it may vary according to the retrieval algorithm (day-time and night-time), the region and the season (see [17] for more details). The observation equation (1) could be modified to take into account these fluctuations in the accuracy of the data. In the same way, we could include the various covariates which alter the quality of the satellite measurements (see [24]) or assume that the parameters $H$ and $R$ depend on the satellite if the observed time series was obtained by mixing data from different satellites.

Then we assume that the latent process $\{X_t\}$ is a simple Ornstein-Uhlenbeck process, that is a stationary solution of the following stochastic differential equation:

$$dX_t = -\lambda X_t dt + \tau dW_t \tag{2}$$

where $\{W_t\}$ denotes a standard Brownian motion. A physical justification of using this model to describe the local dynamics of the SST, when neglecting horizontal transport and heat exchange, is given in [10]: $\lambda > 0$ is the time correlation (in day) or feedback parameter which represents the slowly evolving transfer of heat and $\tau > 0$ the variability coming from weather fluctuations (see also [20], [18]).

Hereafter, we denote $\sigma^2 = Var\,(X_t) = \frac{\tau^2}{2\lambda}$ the variance of the stationary distribution. $\{X_t\}$ is a Markov process which satisfies, for $i \in \{2, ..., n\}$,

$$X_{t_i} = M_{\Delta_i} X_{t_{i-1}} + \sqrt{Q_{\Delta_i}}\eta_{t_i} \tag{3}$$

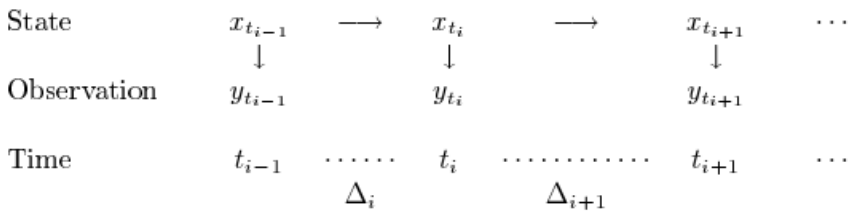


**Figure 4:** Directed acyclic graph for the linear Gaussian state-space model with irregular time step.

with $M_{\Delta_i} = \exp(-\lambda\Delta_i)$, $Q_{\Delta_i} = \sigma^2\left(1 - {M_{\Delta_i}}^2\right)$ and $\{\eta_{t_i}\}_{i\in\{2\ldots n\}}$ a Gaussian white noise sequence with zero mean and unit variance independent of $\{\epsilon_{t_i}\}_{i\in\{1\ldots n\}}$. In the particular case when the temporal sampling is regular, i.e. when $\Delta_1 = ... = \Delta_n$, we retrieve a standard $AR(1)$ process and the usual linear Gaussian state-space model. Here again, more complicated models could be considered, with for example non-linear dynamics, but this would complicate the statistical inference methods discussed in the next Section.

Finally, the various conditional independence assumptions which imply the particular Markovian structure of the state-space model, when observed at discrete time $t_1, ..., t_n$, are summarized on the directed acyclic graph shown on Figure 4.

# 3 Parameter estimation

The estimation of the unknown parameters in Gaussian linear state-space models observed at regular time step has been addressed by many authors and the most usual method consists probably in computing the maximum likelihood (ML) estimates using the EM algorithm (see e.g. [8]).

However, before computing the ML estimates, it is important to check the identifiability of the parameters. For the particular model under consideration, it is possible to show that the observations follow a multivariate Gaussian distribution with an explicit covariance function. Using this result, we can give conditions on the parameters which ensure identifiability and also propose a first method based on the moments to estimate the parameters. The corresponding estimates will be denoted MOM estimates hereafter. This is discussed in Section 3.1. Then, in Section 3.2, we detail the practical implementation of the EM algorithm for the Gaussian linear state-space model with irregular time-step. We discuss how it can be combined with the method of moment and a more standard numerical optimization procedure proposed in [15] to get a computationally efficient and numerically robust estimation procedure. Finally, this is illustrated in Section 3.3 through simulations.

## 3.1 Covariance function

With the various assumption made in the previous section, $\{Y_t\}$ is a stationary Gaussian process with zeros mean and covariance function

$$Cov\left(Y_t, Y_{t'}\right) = H^2\sigma^2 \exp\left(-\lambda\left|t-t'\right|\right) + R\mathbb{1}_{\{0\}}(t-t') \tag{4}$$

We deduce that the distribution of the observed sequence $(y_{t_1}, ..., y_{t_n})$ is a multivariate Gaussian distribution with zeros mean and covariance matrix which can be expressed from the unknown parameter $H$, $R$, $\sigma^2$ and $\lambda$. According to (4), this covariance matrix depends on the parameters $H$ and $\sigma^2$ only through the product $H^2\sigma^2$ and thus we need to add a constraint in order to ensure identifiability of the parameters. Hereafter, we fix $H = 1$ and denote $\theta = (\lambda, \sigma^2, R) \in (0, +\infty)^3$ the unknown parameters.

The covariance function (4) corresponds to a classical model in spatial statistics since we retrieve an exponential model with nugget $R$, sill $\sigma^2 + R$ and range $1/\lambda$. Usual methods in geostatistics permit to compute an empirical estimate of the variogram from the data (see e.g. [3],pp.69). The variogram is directly related to the covariance function for second order stationary processes and the empirical variogram can be used to check the realism of the parametric model 4 and also fit it using the weighted least square method. Here the weights depend on the number of pairs of time points which are available to estimate the empirical variogram as discussed in [3],pp.96. The corresponding estimates will be denoted MOM estimates hereafter.

## 3.2 Maximum likelihood estimation

Alternatively, the parameters can be estimated by computing the ML estimates. According to the conditional independence assumptions shown on Figure 4, the complete log-likelihood, based on both the latent and observed sequences, is given by

$$\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) = \log(p(x_{t_1})) + \sum_{i=2}^{n} \log\left(p(x_{t_i}|x_{t_{i-1}}; \theta)\right) + \sum_{i=1}^{n} \log\left(p(y_{t_i}|x_{t_i}; \theta)\right)$$

where the conditional distributions $p(x_{t_i}|x_{t_{i-1}}; \theta)$ and $p(y_{t_i}|x_{t_i}; \theta)$ are Gaussian distributions which characteristics are given respectively by (3) and (1). Hereafter we will assume that the initial distribution $p(x_{t_1})$ is a Gaussian distribution with known mean $x^{(b)}$ and variance $B$ and in practice these values will be estimated using historical data. Thus, apart from a constant, we obtain

$$\begin{aligned}\log\left(p(x_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta)\right) = &-(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n} \log\left(1 - \exp(-2\lambda\Delta_i)\right) \\ &- \frac{1}{2\sigma^2}\sum_{i=2}^{n} \frac{(x_{t_i} - \exp(-\lambda\Delta_i)x_{t_{i-1}})^2}{(1 - \exp(-2\lambda\Delta_i))} \\ &- \frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}(y_{t_i} - x_{t_i})^2\end{aligned} \tag{5}$$

The ML estimates $\hat{\theta}$ is the value of $\theta$ that maximises the (incomplete) likelihood of the observations $y_{t_1}^{t_n}$ formed by integrating the complete likelihood (5) over the missing variables.

In this paper, the EM algorithm due to [4] is used to compute $\hat{\theta}$. This recursive algorithm computes successive approximations $\hat{\theta}_k = (\lambda_k, \sigma_k^2, R_k)$ of $\hat{\theta}$ by cycling through the following steps.

**E-step:** Compute $U(\theta|\hat{\theta}_k) = E(\log(p(X_{t_1}^{t_n}, y_{t_1}^{t_n}; \theta))|y_{t_1}^{t_n}; \hat{\theta}_k)$ as a function of $\theta$.
**M-step:** Determine the updated parameter estimate $\hat{\theta}_{k+1} = \arg\max_{\theta} U(\theta|\hat{\theta}_k)$.

Under certain general conditions it can be shown that the sequence of estimates $\hat{\theta}_n$ yields monotonically increasing values of the incomplete likelihood, and converges to a maximum of this function (see [26]). Thus the EM algorithm provides an alternative method of maximising the incomplete log-likelihood which is commonly used in models with hidden or latent variables such as the model proposed here. The EM algorithm directly utilises the hidden structure and, as a consequence, is often more robust in practice to the choice of starting values than usual numerical optimization methods. Its computational efficiency is enhanced if the E and M steps are readily evaluated. Various authors have discussed the practical implementation of these steps for linear Gaussian state-space models with regular time sampling ([6], [5], [23] and [2],pp.384-388). Hereafter, we discuss the extension to the case with irregular sampling.

**E step** To determine $U(\theta|\hat{\theta}_k)$ as a function of $\theta$ we need to compute the following smoothing probabilities, for $i = 1, \ldots, n$:

$$x_{t_i}^{(s)} = E(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k), \quad x_{t_i,t_i}^{(s)} = E(X_{t_i}^2|y_{t_1}^{t_n}; \hat{\theta}_k), \qquad x_{t_{i-1},t_i}^{(s)} = E(X_{t_{i-1}}X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) \tag{6}$$

These quantities can be computed using the Kalman recursions described hereafter. This is a particular case of the general Kalman recursions given for example in [23] and [2],pp.127-147.

- **Kalman filter.** Let us denote

$$x_{t_i}^{(f)} = E(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k), \quad P_{t_i}^{(f)} = Var(X_{t_i}|y_{t_1}^{t_{i-1}}; \hat{\theta}_k)$$

  the mean and the variance of the forecast probabilities and

$$x_{t_i}^{(a)} = E(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k), \quad P_{t_i}^{(a)} = Var(X_{t_i}|y_{t_1}^{t_i}; \hat{\theta}_k)$$

  the mean and the variance of the filtering probabilities. These quantities can be computed using the recursion below.

  **Initialization:** compute the Kalman filter gain $K_{t_1} = \frac{B}{B+R}$ and

$$x_{t_1}^{(a)} = x^{(b)} + K_{t_1}(y_{t_1} - x^{(b)}), \quad P_{t_1}^{(a)} = (1 - K_{t_1})B$$

  where the parameters $x^{(b)} = E[X_{t_1}]$ and $B = Var(X_{t_1})$ of the initial distribution are supposed to be known.

  **Recursion:** for $i = 2, ..., n$

– **Time update:**

$$x_{t_i}^{(f)} = M_{\Delta_i} x_{t_{i-1}}^{(a)}, \quad P_{t_i}^{(f)} = M_{\Delta_i}^2 P_{t_{i-1}}^{(a)} + Q_{\Delta_i}$$

– **Observation update:** compute the Kalman filter gain $K_{t_i} = \frac{P_{t_i}^{(f)}}{P_{t_i}^{(f)}+R}$ and

$$x_{t_i}^{(a)} = x_{t_i}^{(f)} + K_{t_i}\left(y_{t_i} - x_{t_i}^{(f)}\right), \quad P_{t_i}^{(a)} = (1 - K_{t_i})\, P_{t_i}^{(f)}$$

- **Kalman smoother.** Let us denote

$$P_{t_i}^{(s)} = Var(X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k)$$

the variance of the smoothing probabilities at time $t_i$. These quantities and the conditional expectation $x_{t_i}^{(s)}$ define in (6) can be computed using the backward recursions below.

**Initialization:**

$$x_{t_n}^{(s)} = x_{t_n}^{(a)}, \quad P_{t_n}^{(s)} = P_{t_n}^{(a)}$$

**Recursion:** for $i = n-1, ..., 1$ compute the Kalman smoother gain $K_{t_i}^{(s)} = \frac{P_{t_i}^{(a)} M}{P_{t_{i+1}}^{(f)}}$ and

$$x_{t_i}^{(s)} = x_{t_i}^{(a)} + K_{t_i}^{(s)}\left(x_{t_{i+1}}^{(s)} - x_{t_{i+1}}^{(f)}\right), \quad P_{t_i}^{(s)} = P_{t_i}^{(a)} + \left(K_{t_i}^{(s)}\right)^2\left(P_{t_{i+1}}^{(s)} - P_{t_{i+1}}^{(f)}\right)$$

Finally $U(\theta|\hat{\theta}_k)$ can be computed from the quantities computed with the Kalman smoother above and the relations

$$x_{t_i,t_i}^{(s)} = P_{t_i}^{(s)} + \left(x_{t_i}^{(s)}\right)^2, \quad x_{t_{i-1},t_i}^{(s)} = Cov(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) + x_{t_{i-1}}^{(s)} x_{t_i}^{(s)}$$

where

$$Cov(X_{t_{i-1}}, X_{t_i}|y_{t_1}^{t_n}; \hat{\theta}_k) = (1 - K_{t_i})\, M P_{t_{i-1}}^{(a)} + \frac{P_{t_i}^{(s)} - P_{t_i}^{(a)}}{P_{t_{i-1}}^{(a)}}(1 - K_{t_i})\, M P_{t_i}^{(a)}$$

**M step**

The function $U(\theta|\hat{\theta}_k)$ can be decomposed as

$$U(\theta|\hat{\theta}_k) = U_X(\lambda, \sigma^2|\hat{\theta}_k) + U_{Y|X}(R|\hat{\theta}_k)$$

where

$$U_X(\lambda,\sigma^2|\hat{\theta}_k) = -(n-1)\log(\sigma) - \frac{1}{2}\sum_{i=2}^{n}\log\left(1-\exp(-2\Delta_i\lambda)\right) - \frac{1}{2\sigma^2}\sum_{i=2}^{n}\frac{x^{(s)}_{t_i,t_i} - 2\exp(-\Delta_i\lambda)\,x^{(s)}_{t_{i-1},t_i} + \exp(-2\Delta_i\lambda)\,x^{(s)}_{t_{i-1},t_{i-1}}}{1-\exp(-2\Delta_i\lambda)}$$

and

$$U_{Y|X}(R|\hat{\theta}_k) = -\frac{n}{2}\log(R) - \frac{1}{2R}\sum_{i=1}^{n}\left\{y^2_{t_i} - 2y_{t_i}x^{(s)}_{t_i} + x^{(s)}_{t_i,t_i}\right\}$$

The second term $U_{Y|X}$ is similar to the case with regular sampling and the maximum is obtained for $R = R_{k+1}$ with

$$R_{k+1} = \frac{1}{n}\sum_{i=1}^{n}\left\{y^2_{t_i} - 2y_{t_i}x^{(s)}_{t_i} + x^{(s)}_{t_i,t_i}\right\}$$

The first term $U_X$ is specific to the case with irregular sampling and numerical optimisation procedures have been used to compute $(\lambda_{k+1},\sigma^2_{k+1})$ since we could not derive analytic expressions these quantities. Here the relation

$$\sigma^2_{k+1} = \frac{1}{n-1}\sum_{i=2}^{n}\frac{x^{(s)}_{t_i,t_i} - 2\exp(-\Delta_i\lambda_{k+1})\,x^{(s)}_{t_{i-1},t_i} + \exp(-2\Delta_i\lambda_{k+1})\,x^{(s)}_{t_{i-1},t_{i-1}}}{1-\exp(-2\Delta_i\lambda_{k+1})}$$

has been used to transform the initial two-dimensional optimization problem into a simple one-dimensional optimisation problem and reduce computational time.

The EM algorithm has several well known limitations. First it may converge to a non-interesting local maximum of the likelihood function depending on the starting value $\hat{\theta}_0$, and thus it is important to provide realistic initial parameter values. Here we have used the estimates obtained using the method of moment described in Section 3.1. Indeed the various tests that we have done indicate that this method leads to robust estimates and generally provide a good starting value to the EM algorithm with low numerical cost (see Section 3.3). This is particularly useful to avoid numerical problem when fitting the model to a large number of data sets for regional studies such as the one performed in Section 4.4.

Another limitation of the EM algorithm is its slow convergence near the maxima where using a standard optimization algorithm is generally far more efficient, at least when it is possible to compute the incomplete likelihood function quickly. For the model under consideration, the incomplete likelihood function is a sub-product of the Kalman filter since we have

$$p(y^{t_n}_{t_1};\theta) = \prod_{i=2}^{n} p(y_{t_i}|y^{t_{i-1}}_{t_1})$$

where the conditional distribution $p(y_{t_i}|y_{t_1}^{t_{i-1}})$ is a Gaussian distribution with mean $E(X_{t_i}|y_{t_1}^{t_{i-1}})$ and variance $Var(X_{t_i}|y_{t_1}^{t_{i-1}}) + R$ and these quantities are computed recursively in the Kalman filter (see Section 3.2). Eventually, the gradient of the log-likelihood function could also be computed to accelerate the convergence of the numerical optimization procedure. In this work, we did not provide the gradient to the Matlab function used for the numerical optimization but we did not encounter any numerical problem and the computational efficiency was good enough.

Another advantage of switching from the EM algorithm to a quasi-Newton algorithm close to the maxima is that quasi-Newton algorithms provide an approximation of the Hessian of the log-likelihood function, and thus useful information on the variance of the ML estimates (see Section 3.3).

## 3.3 Simulations

In this section, the relative performances of ML and MOM estimates are assessed through simulations. More precisely, for various values of $n \in \{200, 300, ..., 2000\}$, we have simulated $N = 1000$ sequences of length $n$ using the scheme described below:

1. Simulate the time lags $(\Delta_i)_{i\in\{2...n\}}$ as an i.i.d. sample from the empirical distribution of the time lags for satellite data (see Figure 1).

2. Simulate the initial state $x_{t_1}$ as a Gaussian variable with mean $x^{(b)}$ and variance $B$ and then recursively $(x_{t_i})_{i\in\{2...n\}}$ according to (3).

3. Simulate the observed process $(y_{t_i})_{i\in\{1...n\}}$ using (1).

The following parameters values have been chosen for the numerical experiment: $\lambda = 0.5$, $B = \sigma^2 = 0.05$, $R = 0.5$ and $x^{(b)} = 0$. It corresponds to realistic values for the application discussed in the next Section.

Then, for each simulated sequence the ML and MOM estimates have been computed. In practice, ML and MOM estimates have been computed using a quasi-Newton algorithm with the true values of the parameters as initial value. Although such initialization is not possible for practical applications, it permits to avoid convergence to non interesting local maxima of the likelihood function and a fair comparison of the two estimates. Figure 5 shows the empirical estimate of the bias and variance of the estimates computed from these simulations. As expected, the ML estimates generally outperform the MOM estimates in terms of both bias and variance. However, the MOM estimates give satisfactory results for the different values of $n$ and have the advantage of being computed with low computational costs and less sensitive to the choice of realistic starting values than the EM algorithm. For comparison purpose, the variances computed from the inverse of the observed information matrix are also shown on Figure 5. The agreement with the empirical variances of the ML estimates is generally good, especially for large sample size as expected from the general asymptotic theory for the ML estimates.

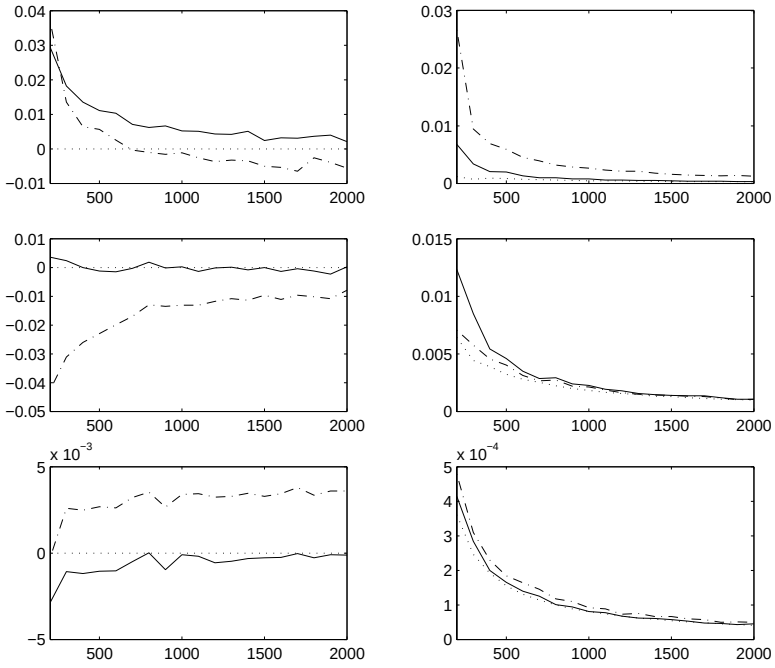


**Figure 5:** Plot of the simulated bias (left) and variances (right) for the MOM (dashed-dotted line) and ML estimates (full line) for different length sequences $n$ (x-axis). Estimate of $\lambda$ (top panel), of $\sigma^2$ (middle) and $R$ (bottom). The dotted lines on the right panel is the variance computed from the information matrix (empirical mean over the different simulations). The simulated results are based on $N = 1000$ replications.

# 4 Application to SST data

In this section the model is first fitted and validated on the SST data introduced in Section 2. The original time series has been divided into two consecutive parts: the first one $(y_{t_1}, ..., y_{t_{n_1}})$ for estimating the parameters and second one $(y_{t_{n_1+1}}, ..., y_{t_n})$ for validating the model. In practice, we used $n_1 = 725$ observations to fit the model, a reasonable amount of data according to the simulation results given in Section 3.3. It corresponds to a proportion of about two-thirds of the data (more than one year).

In Section 4.1, we first discuss the results obtained when fitting the model on the training data set. Then the model is validated using cross-validation on the validation data set in Section 4.2 and by comparison to buoy data in Section 4.3. Finally, in Section 4.4, the methodology is applied to data at many locations on a regular grid covering the Atlantic ocean and the spatial behaviour of the parameter estimates is discussed.

## 4.1 Parameter estimation

The parametric covariance model (4) has been fitted to the empirical estimate of the autocovariance function of the SST anomaly using weighted least square method leading to the MOM estimates (see Section 3.1). The corresponding variograms are shown in Figure 6. The overall agreement is good, except maybe a five days component which is visible on the empirical variogram function (see [12] for a discussion on the existence of peak frequencies in SST time series). This indicates that the assumptions made on the shape of the covariance function is realistic, at least when focussing to time lags up to 40 days. Let us remark that according to Figure 1, it seems also reasonable to assume that the marginal distribution is approximately Gaussian except maybe the lower tail of the distribution.

Starting from the MOM estimates obtained by fitting the covariance function, we have run the EM algorithm. The first iterations are efficient and the likelihood function increases rapidly (see Figure 7) but after some iterations the convergence becomes rather slow, and switching to a standard numerical optimisation procedure permits to save computational time. According to Table 1, the ML estimate of $\lambda$ is significantly lower than the MOM estimate and the ML estimates of $\sigma^2$ and $R^2$ are higher than the corresponding MOM estimates, although the differences for $\sigma^2$ and $R^2$ do not seem to be statistically significant if we compare the differences in the parameter values to the standard deviations given in Table 1. ML estimates identify a second-order structure with a higher sill, which better coincides with the empirical variance of the time series (about 0.47), and also a higher range. Despite these differences in the parameters values, the agreement between the covariance functions is good for time lags less than 10 days (see Figure 6) and thus we may expect that we would get similar results if using the model with the MOM instead of the ML estimates for estimating the true SST in Sections 4.2 and 4.3.

The final parameter values are in good agreement with our knowledge of the physical process under consideration. In particular, according to [17], the standard deviation of the measurement error of the METOP data considered in this paper may vary between

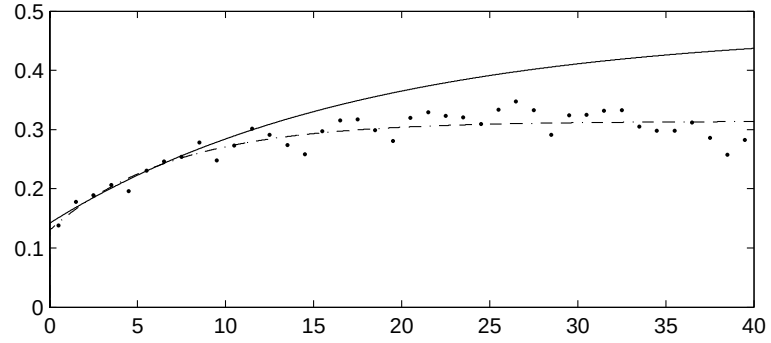


**Figure 6:** Empirical (dotted line) and fitted theoretical variogram for the MOM (dashed-dotted line) and ML (full line) estimates. Results obtained on the training data set. The x-axis is the time lag (in days).

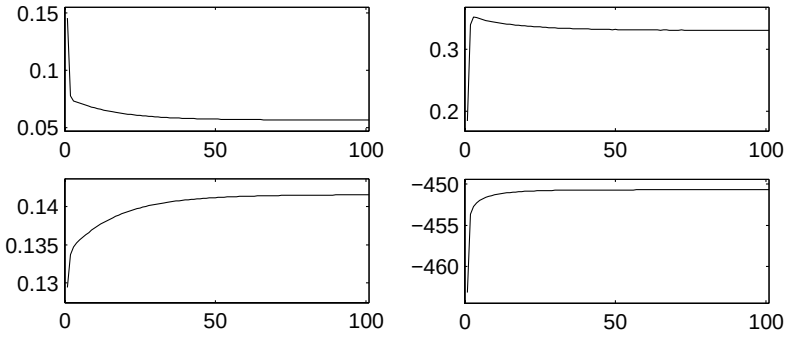


**Figure 7:** Evolution of the parameters values during the 100 iterations (x-axis) of the EM algorithm: $\hat{\lambda}$ (top-left), $\hat{\sigma^2}$ (top-right), $\hat{R}$ (bottom-left). The bottom-right panel shows the increase of the log-likelihood function.

| | Method of moments | Maximum likelihood | | Standard deviation |
|---|---|---|---|---|
| | | EM algorithm | Quasi-Newton | |
| $\hat{\lambda}$ $(day^{-1})$ | 0.145 | 0.057 | 0.056 | 0.019 |
| $\hat{\sigma^2}$ | 0.184 | 0.329 | 0.330 | 0.094 |
| $\hat{R}$ | 0.129 | 0.141 | 0.141 | 0.010 |
| Log-likelihood | −463.15 | −450.69 | −450.68 | |

**Table 1:** Parameter value after the different steps of the fitting procedure: method of moment (first column), 100 iterations of the EM algorithm (second column) and numerical optimization of the likelihood function with a quasi-Newton algorithm (third column). The last column gives an estimate of the standard deviation of the ML estimates computed from the information matrix. Results obtained on the training data set.

0.33 and 0.51 degree Celsius depending on the conditions. This range matches with the 95% confidence interval for the standard deviation of the observation error (we get approximately the interval between 0.35 and 0.40 degree Celsius). Then, the low value of $\lambda$ imply an important temporal persistence of the SST conditions and is coherent with the climatology of the place of interest were SST anomaly is known to have a strong temporal correlation. Finally, comparing the variance of the innovation of the dynamics for a time lag of one day ($\hat{Q}_1 = 0.04$) with the one of observation error $\hat{R}$ indicates that more weights will generally be given to the previous analysis than to the current observation in the Kalman recursions.

## 4.2 Cross-validation

In this section, we validate the model using cross-validation on the validation data set. For each $i \in \{n_1+1, ...n\}$, the observation at time $t_i$ is removed and the Kalman recursions are used to compute

$$x^{(s)}_{t_i|i}(\hat{\theta}) = E(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta}), \quad P^{(s)}_{t_i|i}(\hat{\theta}) = Var(X_{t_i}|y^{t_{i-1}}_{t_{n_1+1}}, y^{t_n}_{t_{i+1}}; \hat{\theta})$$

If the various assumptions made in Section 2 are valid, then the conditional distribution of $Y_{t_i}$ given the past observations $y^{t_{i-1}}_{t_{n_1+1}}$ and the future observation $y^{t_n}_{t_{i+1}}$ should be approximately Gaussian with mean $x^{(s)}_{t_i|i}(\hat{\theta})$ and variance $P^{(s)}_{t_i|i}(\hat{\theta}) + \hat{R}$. An histogram of the standardized residuals

$$\frac{y_{t_i} - x^{(s)}_{t_i|i}(\hat{\theta})}{\sqrt{P^{(s)}_{t_i|i}(\hat{\theta}) + \hat{R}}}$$

is shown on Figure 8 together with the probability density function of the standard normal distribution and a normal probability plot (formal goodness of fit test are hard to implement since the residuals are not independent). The fit is generally good except again for the lower part of the distribution and this indicates that there are too many low residuals. According to Figure 9, it corresponds to breaks in the observed time series at date when the SST anomaly suddenly drops. It is known that various factors (aerosol optical depth, wind speed or proximity to clouds for example) may perturb the quality of the data and a careful examination of these factors at the dates when the SST drops has been done. We could not identify anything special at these dates and thus we believe that the drops are due to non-linearities in the dynamics of the true SST anomaly. It indicates that using a non-linear model instead of (2) may be more appropriate. Let us remark that the standardized residuals may also provide useful information on outliers.

## 4.3 Comparison with buoy data

Using the model proposed in this work and the Kalman smoother on SST anomaly derived from satellite data, we can estimate the "true" SST anomaly at any time and thus emulate a virtual buoy. In order to check the realism of such virtual buoy, we have been compared the result with SST buoy measurements available at high temporal resolution

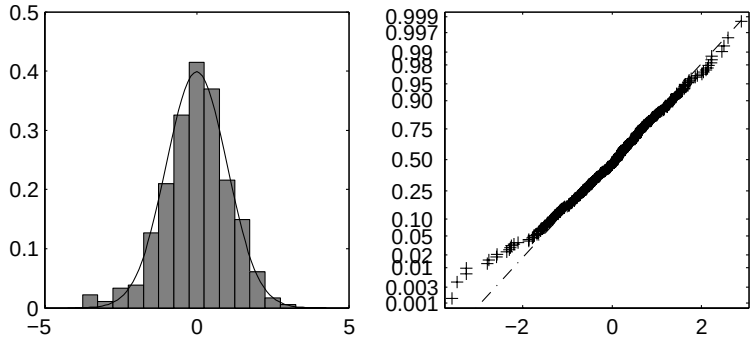


**Figure 8:** Left panel: histogram of the standardised residuals obtained by cross-validation on the validation data set and probability density function of the standard Gaussian distribution (full line). Right panel: normal quantile-quantile plot of the standardised residuals.

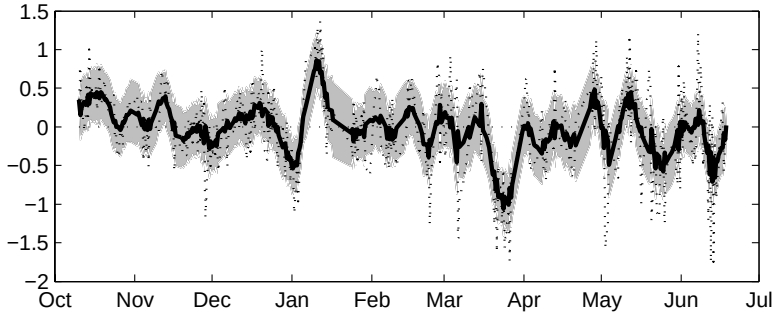


**Figure 9:** Raw (dotted line) and interpolated (full line) satellite SST anomalies (in degree Celsius) together with a 95% fluctuation interval for the smoothing probabilities (grey). Results obtained by cross-validation on the validation data set.

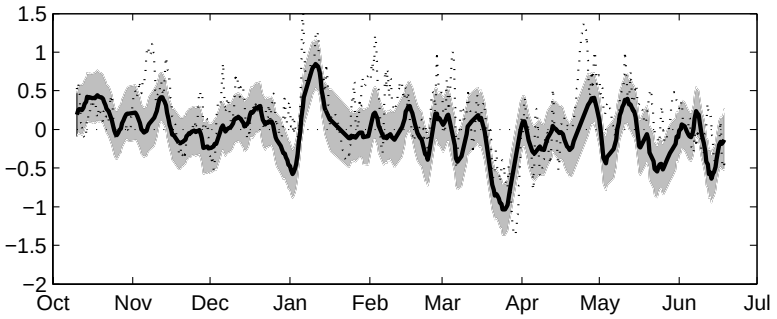


**Figure 10:** Buoy SST anomalies (dotted line) and smoothed satellite SST anomalies (full line) in degree Celsius together with a 95% fluctuation interval for the smoothing probabilities (grey). Results obtained on the validation data set.

| | Bias | Standard deviation | RMSE |
|---|---|---|---|
| Raw satellite data | −0.22 | 0.47 | 0.52 |
| Smoothed satellite data | −0.22 | 0.31 | 0.38 |

**Table 2:** Difference between satellite (raw and smoothed) and buoy SST (bias, standard deviation and root mean square error) computed on the validation data set.

(10 minutes) from the Pilot Research Moored Array in the Tropical Atlantic (PIRATA, see [22]) at the same location ($0^0$N, $23^0$W). According to Figure 10, the virtual buoy obtained by smoothing satellite data has some similarities with buoy data, but there are also important differences (only 63% of buoy measurements are contained in the 95% fluctuation intervals for the smoothing probabilities). However, Table 2 indicates that using the model proposed in this paper permits to improve the quality of the original satellite data and decrease the standard deviation of the error but can not correct the negative bias present in the original satellite data (underestimation of the SST measured at the buoy).

Since the results given in the previous sections indicate that the state-space model proposed in this paper is realistic for satellite data, we may conclude that the significant differences between the buoy and the virtual buoy are due to differences in the satellite and buoy data. A first reason may be the well know depth-to-skin bias discussed in [17]: METOP satellite measures the skin SST (the temperature of the sea in the first $\mu$m) whereas the buoy measures the temperature at a depth of about 1 meter and the temperature gradient evolves strongly in this surface layer. A second possible reason is the difference in the scale of the measurements: buoy data are local measurements and are able to identify small scale variation whereas METOP data describes larger scale variations since they retrieve the mean SST over a $5 \times 5$ km$^2$ surface.

Finally, these results highlight the difficulties of building a realistic SST time series from satellite data. Possible improvements are discussed in the conclusion.

### 4.4 Generalization to the Atlantic ocean

The methodology introduced above for the point with geographical coordinates ($0^0$N, $23^0$W) has been applied to locations on a regular grid with 1° resolution in both latitude and longitude covering the Atlantic Ocean. The state-space model is fitted at each point on the time series of SST anomalies obtained by removing OIV2 analysis from METOP data. The length $n$ of the time series depends on the location of interest and varies from 100 to 900 (see Figure 11). According to the simulation results given in Section 3.3, this may lead to estimates with high variance at locations with poor satellite coverage.
The spatial behaviour of the parameter estimates shown on Figure 11 gives important information on the small-scale variability of SST and also on the quality of METOP data and OIV2 analysis. First, the feedback parameter $\lambda$ (expressed in $day^{-1}$) informs us about the heat transfer at the surface of the ocean. In order to facilitate the interpretation, we have chosen to represent the spatial evolution of $M_1 = \exp(-\lambda)$ which corresponds to the autoregression coefficient for a time lag of one day between two observations. The estimate of $M_1$ mainly depends on the latitude with longer range temporal dependence in the inter-tropical convergence zone (ITCZ) than in the mid-latitudes. Then, the variance of the stationnary distribution of the state $\sigma^2$ informs us about the variability of the SST anomaly. According to Figure 11, the areas with high variability correspond to places, like the Falkland area off the Brazilian coast and the Gulf Stream off the Canadian coast, with strong sea-surface currents and wind conditions. Moreover, the more important upwelling systems of the Atlantic ocean can also be identified, e.g. the Canary and Benguela regions which are areas with strong winds yielding to a mixing of the ocean layer. In the rest of the Atlantic ocean, the variance is about 0.1. Finally, the value of the parameter $R$ is the variance of the measurement errors of the METOP sensor. Estimate of this variance were provided in a previous study ([17]) by comparing METOP observations to data from drifting buoys. Unfortunately, the number of buoys is limited and covers a small part of ocean. The approach presented in this paper, based only on remotely sensed data, presents a global view of the spatial distribution of $R$. According to Figure 11, the principal sources of contamination of METOP infra-red sensor seem to be the aerosol of the Saharan dust (see [9]) and the wildfire off the Angola coast.

## 5 Conclusion and perspectives

In this paper, we propose an extension of the usual linear and Gaussian state space model to analyse satellite data at irregular time step. We propose to combine various methods and algorithms to estimate the parameters efficiently. Indeed, simulation results indicate that the method of moment leads to a computationally efficient and numerically robust estimation procedure suitable for initializing the EM algorithm. A standard numerical optimization procedure is then used in the vicinity of the maximum of the likelihood function identified by the EM algorithm. It permits to accelerate the convergence of the EM algorithm with the extra benefit of giving as output an estimate of the information matrix which provide an estimate of the variance of the estimates.

This paper focus on SST data from the METOP satellite and the various results given in

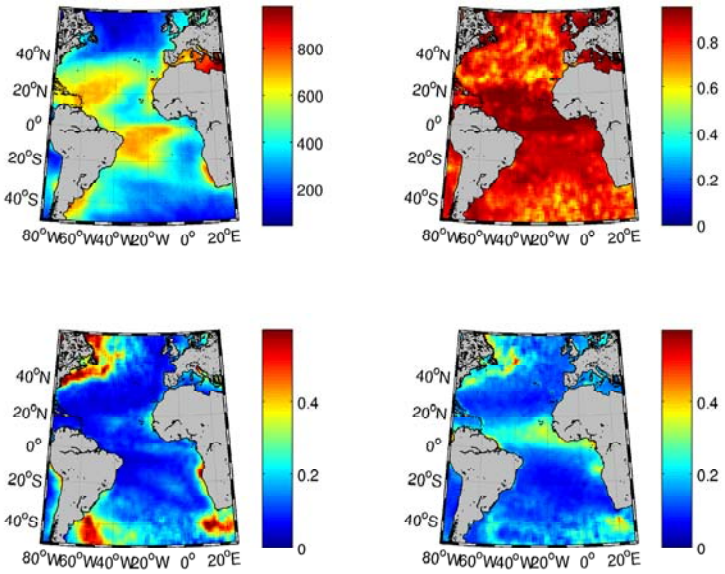


**Figure 11:** Top-left : number of METOP data at each grid point. Top-right : spatial evolution of the estimate of the one day autocorrelation coefficient $\hat{M}_1 = \exp(-\hat{\lambda})$. Bottom-left : spatial evolution of the estimate of the variance of the stationary distribution $\hat{\sigma}^2$. Bottom-right : spatial evolution of the estimate of the variance of the measurement error $R$.

this paper indicate that the model is appropriate for describing some important properties of this data set such as the temporal structure and the measurement errors. Comparison with buoy data indicates that there is work to be done in order to estimate realistic SST conditions from METOP data. Nevertheless, we think that the state-space formulation adopted in this work is an appropriate method. In order to reconstruct realistic SST maps, we plan to extend the formulation in space and time to handle SST data from various satellites with their own accuracies and space-time resolutions. Indeed, using such formulation has several benefits. First, it allows modelling flexibility. For example, non-linear dynamics, which incorporate the effects of advection and diffusion (see [19] and references therein) or non-linear evolution in the atmospheric variability can be considered. We also plan to investigate more elaborated measurement equations and include covariates to model the changing biases and variances of the different satellites (see e.g. [24]). Then, the Markovian structure of the model leads to efficient methods for the statistical inference. In particular, it allows to compute the maximum likelihood estimates and it is shown that these estimates are more efficient than the ones obtained using the method of moment commonly used in geostatitics with kriging. The Kalman recursions used to compute the smoothing probabilities take also benefit of the Markovian properties of the model and permit to save computational time compared to space-time kriging where high dimensional linear systems need to be solved.

# 6 Acknowledgements

We would like to thank the TAO Project Office of the National Oceanic and Atmospheric Administration/Pacific Marine Environmental laboratory (NOAA/PMEL), the National Climatic Data Center (NCDC) and the Godae High Resolution Sea Surface Temperature Pilot Project (GHRSST-PP) for respectively providing in situ PIRATA, real-time OIV2 SST analysis and satellite METOP SST measurements. We are grateful to the Meteo-France Lannion Team, A. Bentamy and O. Talagrand for their expertise and valuable comments on this work.

# References

[1] L. Bertino, G. Evensen, and H. Wackernagel. Sequential Data Assimilation Techniques in Oceanography. *International Statistical Review*, 71:223–241, 2003.

[2] O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Verlag, 2005.

[3] N.A.C. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, 1993.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] L. Deng and X. Shen. Maximum likelihood in statistical estimation of dynamic systems: Decomposition algorithm and simulation results. *Signal Processing*, 57(1):65 – 79, 1997.

[6] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition. 1(4):431–442, Oct. 1993.

[7] C. J. Donlon, P. J. Minnett, C. Gentemann, T. J. Nightingale, I. J. Barton, B. Ward, and M. J. Murray. Toward Improved Validation of Satellite Sea Surface Skin Temperature Measurements for Climate Research. *Journal of Climate*, 15:353–369, February 2002.

[8] J. Durbin and S.J. Koopman. *Time series analysis by state space methods*. Oxford University Press, 2001.

[9] G.R. Foltz and M.J. McPhaden. Impact of Saharan dust on tropical North Atlantic SST. *J. Climate*, 21:5048–5060, 2008.

[10] C. Frankignoul and K. Hasselmann. Stochastic climate models. II- Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, 29:289–305, 1977.

[11] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in geophysics*, 33:141–266, 1991.

[12] S.A. Grodsky, J.A. Carton, C. Provost, J. Servain, J. Lorenzzetti, and M.J. McPhaden. Tropical instability waves at 0N, 23W in the Atlantic: A case study using Pilot Research Moored Array in the Tropical Atlantic (PIRATA) mooring data. *Journal of Geophysical Research*, 110, 2005.

[13] AW Heemink and AJ Segers. Modeling and prediction of environmental data in space and time using Kalman filtering. *Stochastic Environmental Research and Risk Assessment*, 16(3):225–240, 2002.

[14] K. Ide, P. Courtier, M. Ghil, and A.C. Lorenc. Unified notation for data assimilation: Operational, sequential and variational. *Practice*, 75(1B):181–189, 1997.

[15] R.H. Jones and F. Boadi-Boateng. Unequally spaced longitudinal data with AR (1) serial correlation. *Biometrics*, 47(1):161–175, 1991.

[16] A. Kaplan, M.A. Cane, Y. Kushnir, A.C. Clement, M.B. Blumenthal, and B. Rajagopalan. Analyses of global sea surface temperature 1856-1991. *Journal of Geophysical Research*, 103(18):567–18, 1998.

[17] P. Le Borgne, G. Legendre, and A. Marsouin. Operational SST Retrieval from MetOp/AVHRR. In *Proc. 2007 EUMETSAT Conf., Amsterdam, the Netherlands*, 2007.

[18] C. Penland. A stochastic model of IndoPacific sea surface temperature anomalies. *Physica D: Nonlinear Phenomena*, 98(2-4):534–558, 1996.

[19] L.I. Piterbarg and A.G. Ostrovskii. *Advection and diffusion in random media: implications for sea surface temperature anomalies*. Kluwer Academic Publishers, 1997.

[20] RW Reynolds. Sea surface temperature anomalies in the North Pacific Ocean. *Tellus*, 30:97–103, 1978.

[21] R.W. Reynolds, T.M. Smith, C. Liu, D.B. Chelton, K.S. Casey, and M.G. Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007.

[22] J. Servain, A. J. Busalacchi, M. J. McPhaden, A. D. Moura, G. Reverdin, M. Vianna, and S. E. Zebiak. A Pilot Research Moored Array in the Tropical Atlantic (PIRATA). *Bulletin of the American Meteorological Society*, 79:2019–2032, October 1998.

[23] R.H. Shumway and D.S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

[24] P. Tandeo, E. Autret, J. F. Piolle, J. Tournadre, and P. Ailliot. A multivariate regression approach to adjust aatsr sea surface temperature to in situ measurements. *Geoscience and Remote Sensing Letters, IEEE*, 6(1):8–12, Jan. 2009.

[25] C. K. Wikle and N. Cressie. A Dimension-Reduced Approach to Space-Time Kalman Filtering. *Biometrika*, 86:815–829, 1999.

[26] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.