CHALMERS | GÖTEBORG UNIVERSITY

PREPRINT 2006:11

Spatio-temporal Statistical Modelling of Significant Wave Height

A. BAXEVANI
S. CAIRES
I. RYCHLIK

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2006

Preprint 2006:11

Spatio-temporal Statistical Modelling of Significant Wave Height

A. Baxevani, S. Caires, I. Rychlik

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, June 2006

Preprint 2006:11
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2006

Spatio-temporal statistical modelling of significant wave height

A. Baxevani[1]*, S. Caires[2] and I. Rychlik[3]†

[1] Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden

[2] Meteorological Service of Canada and Royal Netherlands Meteorological Institute

[3] Centre for Mathematical Sciences, Lund University, Lund, Sweden

Abstract

In this paper, we construct a homogeneous spatio-temporal model to describe the variability of significant wave height over small regions of the sea and over short periods of time. Then, the model is extended to a non-homogeneous one that is valid over larger areas of the sea and for time periods of up to ten hours. To validate the proposed model, we reconstruct the significant wave height surface under different scenarios and then compare it to satellite measurements and the C-ERA-40 field.

KEY WORDS: Significant wave height, random surface, satellite data, Gaussian random fields, stationary.

1 Introduction

Significant wave height, H_s, is traditionally defined as the average of the one third highest wave heights observed at sea. An alternative definition, valid under the assumption that the sea is well modelled using Gaussian fields, is that H_s equals four times the standard deviation of the vertical displacement of the sea surface, and hence H_s^2 is proportional to the average wave energy.

*corresponding author

†Research partially supported by the Gothenburg Stochastic Centre

Mathematically speaking, H_s is defined assuming stationary sea conditions. However in reality, the sea conditions change both in time and space and hence H_s may be regarded as the parameter that describes the evolution in time and space of the local wave energy. Although H_s is a parameter, it changes in a random way and hence its variability can be modelled by means of random three dimensional fields. The dimension of the fields may be reduced to two, if we consider the sea surface at fixed time or even to one if we instead consider the sea surface at a fixed point.

The properties of the random fields vary with the geographical location and time of the year. Hence in general, the fields should be non-homogeneous both in time and space. However, as it was shown in Baxevani et al. (2006), the sea surface may be modelled with sufficient accuracy by means of homogeneous (isotropic and stationary) Gaussian fields over restricted regions usually rectangles with sides of about four degrees and during limited time period.

In this paper, we propose a non-homogeneous Gaussian field that describes the variability of H_s over large areas and which, when the field is restricted to smaller regions becomes homogeneous. Within this model, we allow time evolution of up to ten hours. Longer time periods are not considered here since this would require taking into account the swell that travels over long distances, and hence a more complicate dependence structure.

The proposed model is parametric, and the majority of the parameters is estimated applying the method developed in Baxevani et al. (2006), on satellite altimeter data. The temporal dynamics though are modelled using buoy measurements and C-ERA-40 reanalysis data.

Models of significant wave height in space and time have the potential to be applied in various areas. These models may be used in modelling wave loads acting on marine structures and computing probabilities of risks associated with marine operations; for example, ship stability, coastal erosion or oil spill motion. Another area of application is fatigue analysis. The long term fatigue of a ship's hull depends on the wave climate along the ship's routes and on the wave induced response of the ship, see Baxevani and Rychlik (2006a). Another application is in computing the probability that the maximum value of a random field exceeds a certain threshold. This is an important problem in safety and reliability analysis of structures exposed to environmental loads, see Baxevani and Rychlik (2006b).

In section 2 we construct a homogeneous spatio-temporal model with a specific covariance structure, which is then extended to a non-homogeneous one. In section 3 we demonstrate how the proposed structure can be used to model the logarithmic values of the satellite estimates of significant wave height, H_s. Finally in section 4, we apply the model to data from the TOPEX-Poseidon satellite over an area of the North Atlantic. The derived model is then validated by reconstructing the significant wave height surface and comparing it to the satellite measurements and the C-ERA-40 data.

2 Model formulation

In this section we construct a homogeneous spatio-temporal model for a zero-mean Gaussian three dimensional field, with a specific spatial covariance structure. The model is constructed through a recursive formula. Then, the homogeneous model is extended to a non-homogeneous one that is valid over larger areas and for longer periods of time. In the latter case, the recursive formula is defined with the help of diffeomorphisms that are solutions to the general transport equation.

2.1 Homogeneous model

2.1.1 Spatial model

We commence our construction of the model by considering the spatial case first. Fix a time point $t_0 \in \mathbb{R}^+$. Then for any point $\mathbf{p}_0 \in \mathbb{R}^2$ denote the region of stationarity around this point by $\eta^{t_0}(\mathbf{p}_0) \equiv \eta(\mathbf{p}_0) \subset \mathbb{R}^2$. For simplicity of presentation we assume the origin lies inside the region $\eta(\mathbf{p}_0)$. Let $X(\mathbf{p}),\ \mathbf{p} \in \eta(\mathbf{p}_0)$ be a homogeneous, real valued, zero mean, Gaussian random field with covariance function

$$r(\mathbf{p}) = \mathrm{Cov}(X(\mathbf{0}), X(\mathbf{p})) = \sigma^2 e^{-\frac{|\mathbf{p}|^2}{2L^2}}, \tag{1}$$

where σ^2 is the field variance, $|\cdot| : \mathbb{R}^2 \to \mathbb{R}$ denotes the euclidean distance and L is the so-called correlation length. Here we should notice that although not explicit in the notation the parameters in (1) may depend on time t_0 and location $\mathbf{p}_0$. The corresponding spectral density (spectrum), is given by

$$S(\boldsymbol{\omega}) = S(\boldsymbol{\omega}; \sigma^2, L) = \frac{\sigma^2 L^2}{2\pi} e^{-L^2 \frac{|\boldsymbol{\omega}|^2}{2}}, \quad \boldsymbol{\omega} = (\omega_1, \omega_2) \in \mathbb{R}^2. \tag{2}$$

It follows from the theory of Hilbert spaces that the field $X(\mathbf{p})$ has the following spectral representation

$$X(\mathbf{p}) = \lim_{\Delta\omega \to 0} (\Delta\,\omega)^2 \sum_{i,j=-\infty}^{+\infty} R_{ij} \sqrt{S(\boldsymbol{\omega})} \cos(\boldsymbol{\omega} \cdot \mathbf{p} + e_{ij}), \tag{3}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2), \mathbf{p} = (p_1, p_2)$ and " $\cdot$ " denotes the inner product between two vectors, i.e., $\boldsymbol{\omega} \cdot \mathbf{p} = \omega_1 p_1 + \omega_2 p_2$. Moreover $\{R_{ij}\}$ and $\{e_{ij}\}$ are sequences of mutually independent random variables distributed as Rayleigh and uniform in $[0, 2\pi)$, respectively. Also, $\{\omega_i\}_{-\infty}^{\infty}$ is a partition of $\mathbb{R}$ and $\Delta\omega = \omega_{i+1} - \omega_i$ is the constant grid step.

2.1.2 Temporal model.

Let now $X(t)$ denote the field at time $t \in \mathbb{R}$ at a fixed position $\mathbf{p} \in \eta^{t_0}(\mathbf{p}_0)$. We assume that the temporal covariance of the process (one-dimensional field) is given by

$$r(t; T) = \mathrm{Cov}(X(0), X(t)) = \sigma^2 e^{-\frac{t^2}{2T^2} - \frac{|t|}{2C}}, \tag{4}$$

for some parameters T and C that depend on both time t_0 and location $\mathbf{p}_0$. Moreover the parameter T is related to the correlation length L through the relation $T = \frac{L}{|\mathbf{v}|}$, where $\mathbf{v}$ denotes the velocity the field $X(\mathbf{p})$ is drifting with, see Baxevani et al. (2003) for a discussion on velocities defined on random surfaces that evolve with time. If there are no dynamics, i.e. if $\mathbf{v} = \mathbf{0}$ or equivalently $T = \infty$, the covariance (4) simplifies to

$$r(t; \infty) = \sigma^2 e^{-\frac{|t|}{2C}},$$

which is a covariance function of the Ornstein-Uhlenbeck type, i.e. $X(t)$ is a Markov process. Similarly for the case $C = \infty$

$$r(t; T) = r(-\mathbf{v}t) = \mathrm{Cov}(X(\mathbf{p}), X(\mathbf{p} - \mathbf{v}t)),$$

which suggests the time variability is a consequence of the field drifting with constant velocity $\mathbf{v}$.

2.1.3 Spatio-temporal model.

Let us now denote by $X(\mathbf{p}, t)$ the field at position $\mathbf{p}$ and time t with $\mathbf{p}, t \in \eta(\mathbf{p}_0)$. A spatio-temporal covariance function can be constructed by combining the functions in (1) and (4);

$$r(\mathbf{p}, t) = \mathrm{Cov}(X(\mathbf{0}, 0), X(\mathbf{p}, t)) = \sigma^2 e^{-\frac{|t|}{2C} - \frac{|\mathbf{p}-\mathbf{v}t|^2}{2L^2}} = \rho(t) r(\mathbf{p} - \mathbf{v}t), \tag{5}$$

where as before $\mathbf{v}$ is the velocity the field $X(\mathbf{p})$ is moving with inside the region $\eta(\mathbf{p}_0)$, and $\rho(t) = e^{-\frac{|t|}{2C}}$ denotes the temporal correlation. Notice that $r(\mathbf{p}, 0) = r(\mathbf{p})$ as in (1) and $r(\mathbf{0}, t) = r(t; T)$ as in (4).

For simulation purposes it is convenient to consider the time evolution of the field $X(\mathbf{p}, t)$ at discrete times $t = i\,dt$ where $dt > 0$ is a suitably chosen time lag and provide with a recursive scheme that allows for sampling. For any $t = i\,dt,\ i > 0$ let

$$X(\mathbf{p}, 0) = X_0^*(\mathbf{p}) \text{ and } X(\mathbf{p}, t) = \rho X(\mathbf{p} - \mathbf{v}\,dt, t - dt) + \sqrt{1-\rho^2} X_i^*(\mathbf{p}) \tag{6}$$

where $X_i^*(\mathbf{p})$ are independent, homogeneous, identically distributed random fields with the covariance structure given in (1) that can be simulated using formula (3) and ρ denotes the constant temporal correlation step $\rho(dt) = e^{-\frac{dt}{2C}}$.

Remark 1 *It is easy to see that the recursive scheme in (6) can be also given in the following non-recursive form, with* $t = i\,dt,\ \ dt > 0$,

$$X(\mathbf{p}, t) = \rho^i X_0^*(\mathbf{p} - \mathbf{v}t) + \sum_{k=1}^{i} \rho^{i-k} \sqrt{1-\rho^2} X_k^*(\mathbf{p} - \mathbf{v}(i-k)\,dt), \tag{7}$$

where $X(\mathbf{0}, 0) = X_0^*(\mathbf{0})$.

Consequently

$$Cov(X(\mathbf{0}, 0), X(\mathbf{p}, t)) = Cov(X_0^*(\mathbf{0}), \rho^i X_0^*(\mathbf{p} - \mathbf{v}t)) = \rho^i r(\mathbf{p} - \mathbf{v}t),$$

and hence we have demonstrated that the field defined in (6) or (7) has the covariance function given in (5), since $\rho(t) = \rho^i \equiv \rho^i(dt)$.

2.2 Non-homogeneous model

The homogeneous field $X(\mathbf{p}, t)$ constructed in section 2.1.3, is valid in a relatively small neighborhood of $\mathbf{p}_0$ and t_0 denoted by $\eta(\mathbf{p}_0, t_0)$. (The parameters L and σ^2 in (1), depend usually on both $\mathbf{p}_0$ and t_0.) In this section we extend $X(\mathbf{p}, t)$ to a non-homogeneous Gaussian field that is valid in areas that are larger than $\eta(\mathbf{p}_0, t_0)$.

2.2.1 Spatial model

Fix $t_0 \in \mathbb{R}^+, \mathbf{p} \in \mathbb{R}^2$ and denote by $\tilde{\eta}$ an area that contains $\eta(\mathbf{p}_0, t_0)$, and for $\mathbf{p} \in \tilde{\eta}$, introduce the slowly varying functions $L(\mathbf{p})$ and $\sigma^2(\mathbf{p})$. (By slowly varying we mean functions that are almost constant inside $\eta(\mathbf{p}_0, t_0)$.)

Now let $X(\mathbf{p})$ be a zero mean Gaussian field with the following spatial covariance function

$$r(\mathbf{p}, \mathbf{q}) = \mathrm{Cov}(X(\mathbf{p}), X(\mathbf{q})) = \frac{2\sigma(\mathbf{p})\sigma(\mathbf{q})L(\mathbf{p})L(\mathbf{q})}{L^2(\mathbf{p}) + L^2(\mathbf{q})} e^{-\frac{|\mathbf{p}-\mathbf{q}|^2}{L^2(\mathbf{p})+L^2(\mathbf{q})}}, \tag{8}$$

for $\mathbf{p}, \mathbf{q} \in \tilde{\eta}$. Obviously the parameters of the covariance depend also on time t_0, although it is not explicit in the notation. When both points $\mathbf{p}$ and $\mathbf{q}$ lie in the same region of stationarity $\eta(\mathbf{p}_0, t_0)$, the covariance in (8) is close to the covariance in (1), since $L(\mathbf{p})$ and $\sigma^2(\mathbf{p})$ are assumed to be almost constant on any set $\eta(\mathbf{p}_0, t_0)$.

Next, consider the non-negative bounded real-valued function

$$S(\boldsymbol{\omega}; \sigma^2(\mathbf{p}), L(\mathbf{p})) = \frac{\sigma^2(\mathbf{p})L^2(\mathbf{p})}{2\pi} e^{-L^2(\mathbf{p})\frac{|\boldsymbol{\omega}|^2}{2}}, \quad \boldsymbol{\omega} = (\omega_1, \omega_2) \in \mathbb{R}^2, \tag{9}$$

which obviously depends on the position $\mathbf{p}$. Notice that the function in (9) is only locally a spectrum, i.e., if $\tilde{\eta} = \eta(\mathbf{p}_0, t_0)$, then $S(\boldsymbol{\omega}; \sigma^2(\mathbf{p}), L(\mathbf{p}))$ coincides with the spectrum in (2).

Finally, the field $X(\mathbf{p}, t)$ also assumes a spectral representation analogous to that in (3),

$$X(\mathbf{p}, t) = \lim_{\Delta\omega \to 0} (\Delta\,\omega)^2 \sum_{i,j=-\infty}^{+\infty} R_{ij} \sqrt{S(\boldsymbol{\omega}; \sigma^2(\mathbf{p}), L(\mathbf{p}))} \cos(\boldsymbol{\omega} \cdot \mathbf{p} + e_{ij}),$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2), \mathbf{p} = (p_1, p_2)$ and $\{R_{ij}\}$ and $\{e_{ij}\}$ are again sequences of independent random variables that are distributed as Rayleigh and uniform in $[0, 2\pi)$ respectively.

2.2.2 Spatio-temporal model

In this section we model the time evolution of the non-homogeneous field $X(\mathbf{p})$ during short periods of time. Let $t, s \in \tilde{\eta} \supset \eta(\mathbf{p}_0, t_0)$ and assume that $L(\mathbf{p})$ and $\sigma^2(\mathbf{p})$ inside $\tilde{\eta}$ depend only on the location. Moreover, the velocity $\mathbf{v}$ the field is moving with inside the region $\tilde{\eta}$ has to be modelled by means of a deterministic field $\mathbf{v}(\mathbf{p},t)$. For simplicity of notation let $t_0=0$ and $0\leq t\leq s\leq 1$ which can be achieved by a suitable choice of units. Then the motion of the field is modelled by means of a flow of diffeomorphisms

$$\phi:\mathbb{R}^2\times[0,1]^2\to\mathbb{R}^2$$

that satisfy $\phi(\mathbf{p},0,1)=\phi(\mathbf{p}),\quad \phi(\mathbf{p},s,s)=\mathbf{p},\quad \phi(\cdot,t,s)=\phi(\cdot,u,s)\circ\phi(\cdot,t,u)$ and are the solution to the transport equation

$$\phi(\mathbf{p},t,s)=\mathbf{p}+\int_t^s \mathbf{v}(\phi(\mathbf{p},t,u),u)du,\quad t<s. \tag{10}$$

Clearly $\mathbf{p}=\phi(\mathbf{q},t,s)$ is the position at time s of the point that at time t was at $\mathbf{q}$. The point $\mathbf{q}$ will be denoted by $\mathbf{p}_{ts}$, i.e.

$$\mathbf{p}=\phi(\mathbf{p}_{ts},t,s),\qquad \mathbf{p}_{ts}=\phi^{-1}(\mathbf{p},t,s),$$

for any $t<s$.

The field dynamics are modelled by generalising the recursion formula in (6). Let $t_0=0$ and $t=i\,dt,\ dt>0$. Let also $X_k^*(\mathbf{p}),\ k=0,1,\dots i$, denote independent, zero-mean, Gaussian fields having the covariance function given in (6) with the parameters $\sigma^2(\mathbf{p}),L(\mathbf{p})$ taking values at $t_0=0$. Then, the field $X(\mathbf{p},t)$ is defined as follows

$$X(\mathbf{p},t)=\rho^i X_0^*(\mathbf{p}_{0t})+\sum_{k=1}^{i}\rho^{i-k}\sqrt{1-\rho^2}X_k^*(\mathbf{p}_{ut}),\qquad u=k\ dt. \tag{11}$$

Remark 2 *At time $t_0=0$, the spatial covariance of the field $X(\mathbf{p},t)$ is defined by means of (8) with certain parameters $\sigma^2(\mathbf{p})$ and $L(\mathbf{p})$. However, at time t these parameters are different and so is the spatial covariance of $X(\mathbf{p},t)$. Considering this model inside regions with sides around 4 degrees minimises this difference.*

The field $X(\mathbf{p},t)$ defined in (11) has the covariance funcion

$$\begin{aligned}\mathrm{Cov}(X(\mathbf{p},t),X(\mathbf{q},s)) &= \rho^{i+j}r(\mathbf{p}_{0t},\mathbf{q}_{0s})+\sum_{k=1}^{i}\rho^{i+j-2k}(1-\rho^2)r(\mathbf{p}_{ut},\mathbf{q}_{us})\\ &= e^{-\frac{t+s}{2C}}r(\mathbf{p}_{0t},\mathbf{q}_{0s})+\sum_{k=1}^{i}e^{-\frac{t+s-2u}{2C}}\left(1-e^{-\frac{dt}{C}}\right)r(\mathbf{p}_{ut},\mathbf{q}_{us}),\end{aligned} \tag{12}$$

for $t=i\,dt, s=j\,dt, u=k\,dt$ and $i\leq k\leq j$.

The formula in (12) will be used in the coming sections to reconstruct and predict values of the logarithms of the significant wave height field. However, when the velocity field $\mathbf{v}$ is not constant the covariance function depends on the discretisation step dt. Letting $dt \to 0$, the resulting covariance function is a generalisation of the homogeneous case in (5). Moreover, for small values of dt, $1 - \rho^2 \approx dt/C$, and therefore

$$\lim_{dt \to 0} \operatorname{Cov}(X(\mathbf{p}, t), X(\mathbf{q}, s)) = e^{-\frac{t+s}{2C}} r(\mathbf{p}_{0t}, \mathbf{q}_{0s}) + \frac{1}{C} \int_t^s e^{-\frac{t+s-2u}{2C}} r(\mathbf{p}_{ut}, \mathbf{q}_{us})\, du, \tag{13}$$

where r denotes the non-homogeneous covariance function in (8).

3 Modelling significant wave height in the North Atlantic

In this section, we demonstrate how we can model the logarithmic values of the satellite estimates of significant wave height using the structures introduced in section 2.

3.1 Homogeneous field

Let us denote by $Y(\mathbf{p}, t)$ the field consisting of the logarithmic values of significant wave height, H_s, at position $\mathbf{p}$ and time t, i.e., $Y(\mathbf{p}, t) = \log(H_s(\mathbf{p}, t))$. For a fixed point $\mathbf{p}_0$, we take $\eta(\mathbf{p}_0)$ to be a small square with side approximately 4 degrees. Then, using the methodology developed in Baxevani et al. (2006), the mean value of the field $Y(\mathbf{p}, t)$,

$$m(\mathbf{p}, t) = m_0(\mathbf{p}) + m_1(\mathbf{p}, t)\cos(\omega t) + m_2(\mathbf{p}, t)\sin(\omega t), \quad \omega = \frac{2\pi}{365.2}$$

is removed using non-linear regression. Note the time t is measured in days starting from the 1^{st} of January.

For fixed time t_0 the residual field $\epsilon(\mathbf{p}, t_0) = Y(\mathbf{p}, t_0) - m(\mathbf{p}, t_0)$ has been shown to be homogeneous over the area $\eta(\mathbf{p}_0, t_0)$. Moreover, it has been shown that $\epsilon(\mathbf{p}, t_0)$ can be considered as a real-valued Gaussian random field that is the sum of three independent homogeneous zero mean real-valued Gaussian fields:

$$\epsilon(\mathbf{p}, t_0) = \sqrt{p} X_l(\mathbf{p}, t_0) + \sqrt{1-p} X_s(\mathbf{p}, t_0) + X_e(\mathbf{p}, t_0), \tag{14}$$

where $p \in [0, 1]$ is a mixing parameter that describes the energy contribution by each one of the fields X_l and X_s. The physical interpretation of the model (14) is that the logarithmic values of H_s exhibit spatial variability at three different scales: small, medium and long. The small-scale variability, denoted by X_e, is interpreted as measuring error and should be removed using some smoothing filter. The long-scale X_l (which describes the size of a storm and the propagation of waves), and the medium-scale X_s (which is some type of correlated noise), are modelled using a special type of covariance function. For a detailed treatment of the problem see Baxevani et al. (2006). Additionally, we also assume that each one of the fields X_l, X_s and X_e can be modelled according to sections 2.1 and 2.2.

Combining equations (7) and (14) for constant velocity field $\mathbf{v}$ inside $\eta(\mathbf{p}_0, t_0)$, the field $\epsilon(\mathbf{p}, t)$ can be writen for $t = i\,dt$, in the following way:

$$\begin{aligned}\epsilon(\mathbf{p},t) &= \rho^i \left[\sqrt{p}X^*_{l,0}(\mathbf{p}-\mathbf{v}t) + \sqrt{1-p}X^*_{s,0}(\mathbf{p}) + X^*_{e,0}(\mathbf{p})\right] \\ &+ \sum_{k=1}^{i} \rho^{i-k}\sqrt{1-\rho^2}\left[\sqrt{p}X^*_{l,k}(\mathbf{p}-\mathbf{v}(t-u)) + \sqrt{1-p}X^*_{s,k}(\mathbf{p}) + X^*_{e,k}(\mathbf{p})\right],\end{aligned} \tag{15}$$

where $u = k\,dt$ and $t_0 = 0$. The fields $X^*_{l,k}(\mathbf{p}), X^*_{s,k}(\mathbf{p})$ and $X^*_{e,k}(\mathbf{p})$, $k = 0, 1, \ldots i$, are independent Gaussian fields with the covariance function in (1) for suitable parameters L and σ^2 taken at time t_0. Notice that only the long-scale field X_l is moving with the velocity $\mathbf{v}$.

3.2 Non-homogeneous field

A natural extension of the field $\epsilon(\mathbf{p}, t)$, $t_0 = 0$, $t = i\,dt$, $i > 0$ defined in (15), can be derived combining equations (11) and (14). Exactly as in the homogeneous case, only the long-scale component is moving with velocity that is the solution to the transport equation in (10).

$$\begin{aligned}\epsilon(\mathbf{p},t) &= \rho^i \left[\sqrt{p}X^*_{l,0}(\mathbf{p}_{0t}) + \sqrt{1-p}X^*_{s,0}(\mathbf{p}) + X^*_{e,0}(\mathbf{p})\right] \\ &+ \sum_{k=1}^{i} \rho^{i-k}\sqrt{1-\rho^2}\left[\sqrt{p}X^*_{l,k}(\mathbf{p}_{ut}) + \sqrt{1-p}X^*_{s,k}(\mathbf{p}) + X^*_{e,k}(\mathbf{p})\right],\end{aligned}$$

with $u = k\,dt$. For $t = i\,dt, s = j\,dt,\ i < j$, the covariance function is given by

$$\begin{aligned}\mathrm{Cov}(\epsilon(\mathbf{p},t),\epsilon(\mathbf{p},s)) &= \rho^{i+j}\left[pr_l(\mathbf{p}_{0t},\mathbf{q}_{0t}) + (1-p)r_s(\mathbf{p},\mathbf{q}) + r_e(\mathbf{p},\mathbf{q})\right] + \\ + &\sum_{k=1}^{i} \rho^{i+j-2k}(1-\rho^2)\left[pr_l(\mathbf{p}_{ut},\mathbf{q}_{us}) + (1-p)r_s(\mathbf{p},\mathbf{q}) + r_e(\mathbf{p},\mathbf{q})\right] = \\ = &\ p\mathrm{Cov}(X_l(\mathbf{p},t), X(\mathbf{q},s)) + \rho^{j-i}\left[(1-p)r_s(\mathbf{p},\mathbf{q}) + r_e(\mathbf{p},\mathbf{q})\right],\end{aligned}$$

where the covariance $\mathrm{Cov}(X_l(\mathbf{p},t), X(\mathbf{q},s))$ is given by formula (12) while the covariances $r_s(\mathbf{p},\mathbf{q})$ and $r_e(\mathbf{p},\mathbf{q})$ by (8) and the subindices denote the source of variability. Finally, using formula (13) and letting $dt \to 0$ we obtain the limiting covariance function

$$\lim_{dt\to 0} \mathrm{Cov}(\epsilon(\mathbf{p},t), \epsilon(\mathbf{p},s)) = pe^{-\frac{t+s}{2C}} r_l(\mathbf{p}_{0t}, \mathbf{q}_{0s}) + \\ + \frac{p}{C}\int_0^t e^{-\frac{t+s-2u}{2C}} r_l(\mathbf{p}_{ut}, \mathbf{q}_{us})\, du + e^{-\frac{s-t}{2C}} \left[(1-p) r_s(\mathbf{p},\mathbf{q}) + r_e(\mathbf{p},\mathbf{q})\right]. \tag{16}$$

4 Model presentation for the North Atlantic

In this section we apply the model presented in section 3, to altimeter data from the TOPEX-Poseidon satellite over an area of the North Atlantic that extends between 42 and 62 degrees in latitude and -48 to -12 degrees in longitude. The results are presented for the month of December. The choice of month is arbitrary. For a detailed description of the data see Baxevani et al. (2006).

4.1 Spatial model

We commence the analysis by defining as areas of stationarity squares with sides of approximately 4 degrees. Then the mean value, $m(\mathbf{p},t)$, is estimated using non-linear regression over the data set consisting of the logarithm of the first observation from every satellite passage. The spatial covariance parameters $\sigma^2(\mathbf{p},t), p(\mathbf{p},t), \sigma_e^2(\mathbf{p},t), L_l(\mathbf{p},t)$ and $L_s(\mathbf{p},t)$ are estimated by applying the method developed in Baxevani et al. (2006). The values for the different regions are presented in Fig. 1. The parameter L_e is arbitrarily taken to be $5km$, i.e., the field X_e is treated as white noise.

4.2 Temporal model

To model the time dynamics we need to estimate the parameters C and T. We expect them to depend on both time and position but also vary slowly enough so they can be considered as constants over the region of stationarity and for a period of time that does not exceed ten hours.

The estimation of the parameter C_l (where the subindex l indicates the long-scale component field), is essential. It reflects the dynamics of the atmospheric conditions leading to creation of storms and it cannot be estimated using satellite measurements.

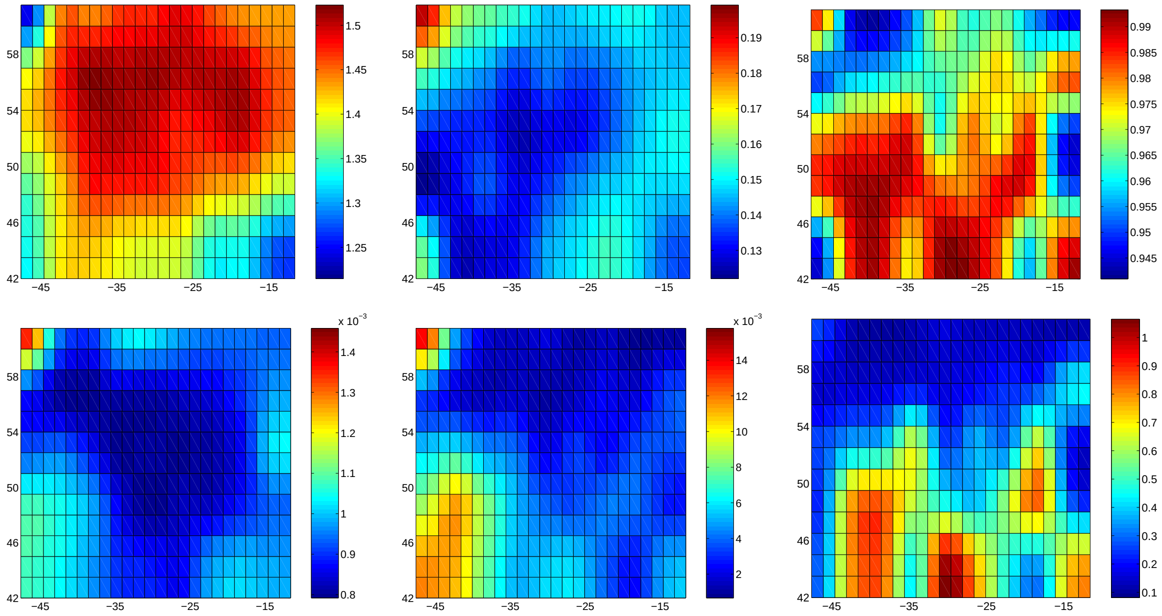

Figure 1: The model for $Y(\mathbf{p}, t)$ for the month of December. The x-axis indicate longitute and the y-axis latitude. (*Top-left*) - $m(\mathbf{p}, t)$; (*Top-middle*) - $\sigma^2(\mathbf{p}, t)$; (*Top-right*) - $p(\mathbf{p}, t)$. (*Bottom-left*) - $\sigma_e^2(\mathbf{p}, t)$; (*Bottom-middle*) - $L_l(\mathbf{p}, t)$ in degrees; and (*Bottom-right*) - $L_s(\mathbf{p}, t)$ in degrees. (All the plots are smoothed using a Gaussian kernel with window 20° in order to ensure the smoothness of the parameter values. The values close to the boundary are used in the model unsmoothed.)

Data from other sources have to be used since the temporal resolution of the satellite data is not of the right order. Buoys have the right temporal resolution but unfortunatelly are usually located along the coast. Hence, we additionally assume the temporal correlation along the coast is the same as in the middle of the ocean and use buoy data to estimate the parameters C_l and T_l. In table 1 we present the results from 20 NOOA deep water buoys and in Fig. 3 (*Left*), we fit the temporal covariance to data from the buoy 46003.

The study of the temporal correlation of the buoy data reveals that although the parameter C_l does not depend on time it varies significantly with location. At present we arbitrarily assign to C_l the value 35 hours, however there are indications that it can be smaller especially when we depart from the coastal areas. In practice we are interested in the time correlation for $t > 0.33$ hours ($t_0 = 0$) since at a fixed location we usually assume the sea conditions to be homogeneous during a period of about 20 min. Although the satellite data indicate that for $t > 0.33$ hours, the fields $X_s(\cdot, t_0)$ and $X_s(\cdot, t_0 + t)$ are practically independent, in the case we are interested in time correlation at shorter time lags the parameters C_s and C_e should also be estimated. The magnitude of C_s and C_e is of the order of a few minutes and seconds and here is arbitrarily assigned to be 2 minutes and 7 seconds respectively.

What remains to be estimated is the velocity field $\mathbf{v}$. Estimation of the parameter T_l ($T_l = \frac{C_l}{|\mathbf{v}|}$), although provides with information on the speed of the wave propagation does not give any insight on its direction. Therefore, since it is well known that the direction of wave propagation depends on the wind direction history, the variability of $\mathbf{v}$ should be modelled as a random field. However, we may assume that $\mathbf{v}$ varies so slowly inside $\eta(\mathbf{p}_0, t_0)$ for a time period less than 10 hours that can be additionally assumed to be constant. We then let $\mathbf{v}$ equal its mean value which is a function of both space and time, $(\mathbf{p}_0, t_0)$ and the diffeomorphism in (10), simplifies to

$$\phi(\mathbf{p}, t, s) \approx \mathbf{p} + \mathbf{v}(\mathbf{p})(s - t).$$

Furthermore, since in 10 hours the distance between two points cannot exceed 3 degrees, the points $\mathbf{p}_{0t}, \mathbf{p}_{ut},\ 0 \leq u \leq t$ lie inside the same stationarity region $\eta(\mathbf{p}_0, t_0)$. Consequently, $\mathbf{v}(\mathbf{p}_{ut}) = \mathbf{v}(\mathbf{p}), \forall\ 0 \leq u \leq t$, and hence $\mathbf{p}_{ut} = \mathbf{p} - \mathbf{v}(\mathbf{p})(t - u)$. Therefore, combining formulas (8) and (16) we obtain

$$\begin{aligned}\lim_{dt\to 0} \mathrm{Cov}(\epsilon(\mathbf{p}, t), \epsilon(\mathbf{q}, s)) = p e^{-\frac{t+s}{2C}} \frac{2\sigma(\mathbf{p}_{0t})\sigma(\mathbf{q}_{0s})L(\mathbf{p}_{0t})L(\mathbf{q}_{0s})}{L^2(\mathbf{p}_{0t}) + L^2(\mathbf{q}_{0s})} e^{-\frac{|\mathbf{p}-\mathbf{q}|^2}{2C}} + \\ \frac{p}{C} \frac{2\sigma(\mathbf{p})\sigma(\mathbf{q})L(\mathbf{p})L(\mathbf{q})}{L^2(\mathbf{p}) + L^2(\mathbf{q})} \int_0^t e^{-\frac{t+s-2u}{2C}} e^{-\frac{|\mathbf{p}-\mathbf{v}(\mathbf{p})(t-u)-\mathbf{q}+\mathbf{v}(\mathbf{q})(s-u)|^2}{L^2(\mathbf{p})+L^2(\mathbf{q})}} du \\ +\ \ e^{-\frac{s-t}{2C}} \left[(1-p) r_s(\mathbf{p}, \mathbf{q}) + r_e(\mathbf{p}, \mathbf{q})\right]. \qquad (17)\end{aligned}$$

The covariance function in (17) for $\mathbf{p} = \mathbf{p}_0$ and $\mathbf{q} \in \tilde{\eta}$ and $t = s = t_0 = 0$ can be seen in Fig. 2 (*Left*), and for $t = 0$ and $s = 10$ hours in Fig. 2 (*Right*).

In the following examples the mean value of $\mathbf{v}$ is estimated using the ERA-40 data, a description of which can be found in section 6 and its variability with location can be seen in Fig. 3 (*Right*). Obviously the dependence between $\mathbf{v}(\mathbf{p}, t)$ and $\epsilon(\mathbf{p}, t)$ has to be further investigated, but this lies outside the scope of this paper.

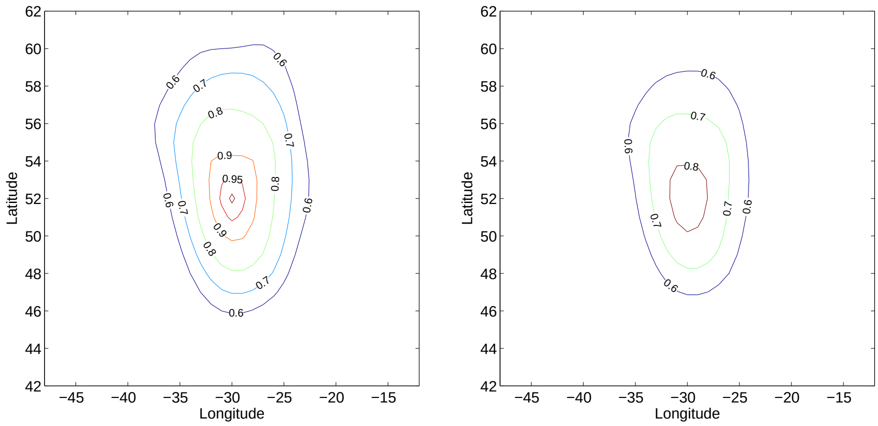

Figure 2: (*Left*) - The covariance in (17) between the region and the central point at time $t = 0$. (*Right*) - The covariance between the central point and the region 10 hours later.

4.3 Model Validation

The proposed spatial model has been locally validated for the North Atlantic, see Baxevani et al. (2006). What remains is to validate the temporal component and the global spatial model. To do so, we reconstruct the H_s surface in the region $[-48, -12] \times [42, 62]$ on the 17th of December 1996 at 12:00, using measurements from satellite passages that took place between 10:00 and 20:00 that day, under different scenarios and then compare it to satellite measurements and the C-ERA-40 field. The positions and times of the 893 recorded satellite measurements can be found in Fig. 4 (*Top-left*). The longest passage took place at 11:24 with orientation from SW to NE and consisted of 43 not equally spaced measurements. We start by reviewing some facts about Gaussian fields.

4.3.1 Surface reconstruction

In this section we discuss the reconstruction of the unknown H_s surface. We may think of it as a random field with a certain distribution conditionally on the H_s measurements along the satellite tracks.

Let us denote by $\epsilon(\mathbf{p}, t)$ the unknown zero-mean surface consisting of the logarithmic values of H_s with covariance function $\mathrm{Cov}(\epsilon(\mathbf{p}, t), \epsilon(\mathbf{q}, s))$ given in (17). Assume we want to reconstruct the surface at position $\mathbf{p}_0$ and time t_0. Let us also denote by $(\mathbf{p}_i, t_i)$, $i = 1, \ldots, K$, the coordinates of the satellite measurements that fall inside $\tilde{\eta}$. (Obviously K

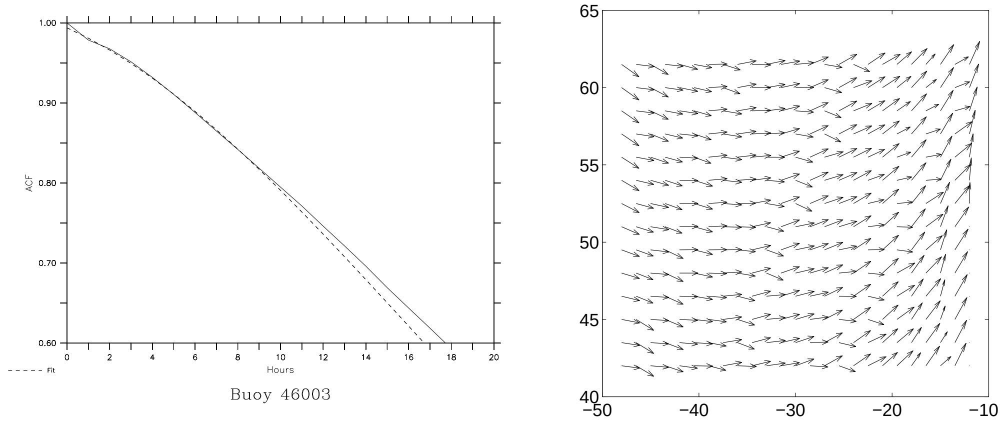

Figure 3: (*Left*) - Fit of the covariance function in (4) to data from buoy 46003. (*Right*) - The average velocity field $\mathbf{v}$.

depends on both $\mathbf{p}_0$ and t_0.)

The satellite measurements form a column vector $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^K$, with $\epsilon_i = \log(h_s(\mathbf{p}_i, t_i)) - m(\mathbf{p}_i, t_i)$, zero mean and covariance matrix $\Sigma = [\sigma_{ij}^2]_{i,j=1}^K$, with entries $\sigma_{ij}^2 = \mathrm{Cov}(\epsilon(\mathbf{p}_i, t_i), \epsilon(\mathbf{p}_j, t_j))$, $i, j = 1, \ldots, K$. Furthermore, let us denote by $\mathbf{C}$ the row vector of cross-covariances with entries $c_i = \mathrm{Cov}(\epsilon(\mathbf{p}_0, t_0), \epsilon(\mathbf{p}_i, t_i)), i = 1, \ldots, K$. The field $\epsilon(\mathbf{p}_0, t_0)$ conditionally on the vector $\boldsymbol{\epsilon}$ is a Gaussian variable with mean $\hat{\epsilon}$ and variance $\hat{\sigma}^2$ given by

$$\hat{\epsilon}(\mathbf{p}_0, t_0) = \mathbf{C}\Sigma^{-1}\boldsymbol{\epsilon}, \qquad \hat{\sigma}^2(\mathbf{p}_0, t_0) = \sigma^2(\mathbf{p}_0, t_0) + \sigma_e^2 - \mathbf{C}\Sigma^{-1}\mathbf{C}^T, \tag{18}$$

where $\mathbf{C}^T$ denotes the transpose of the matrix $\mathbf{C}$.

Hence the optimal prediction of the significant wave height value $\hat{H}_s(\mathbf{p}_0, t_0)$, and an approximately 99% prediction interval $\hat{I}$ are given by

$$\hat{H}_s(\mathbf{p}_0, t_0) = e^{m(\mathbf{p}_0, t_0) + \hat{\epsilon}}, \qquad \hat{I} = \left[\hat{H}_s e^{-1.96\hat{\sigma}}, \hat{H}_s e^{1.96\hat{\sigma}}\right], \tag{19}$$

where $m(\mathbf{p}_0, t_0)$ is the value of the deterministic mean value field at $(\mathbf{p}_0, t_0)$. The surface at $(\mathbf{p}_0, t_0)$ should be approximated using only the mean value $m(\mathbf{p}_0, t_0)$ if there are no satellite observations in $\tilde{\eta}$. Notice that the probability coverage 0.99 is exact under the assumption the model is correct, and the width of the prediction interval is considerably wider if the uncertainty of the parameters is also taken into account.

4.3.2 Validation of the model using satellite measurements of H_s.

As our initial analysis, we decided to use 19 of the satellite measurements at 11:24, in order to predict the significant wave height values at 24 future locations and inside a time span of 8 hours.

The predictor $\hat{\epsilon}(\mathbf{p}_0, t_0)$ and its variance $\hat{\sigma}^2(\mathbf{p}_0, t_0)$ are given by (18). An estimate of the field ϵ at each point $(\mathbf{p}_0, t_0)$ and a 99% confidence interval of the significant wave height are given by (19).

In Fig. 4 (*Bottom-left*), "*" indicates the 19 satellite measurements that were used in the analysis. The predicted 24 values are marked by "." and should be compared to the satellite measurements at the same locations indicated by the irregular line. The 99% prediction band is indicated by the thicker lines.

Notice that the width of the prediction band, about 5 meters, can be contributed to a few factors. Namely to the uncertainty of the model parameters, the time elapsed between the two groups of measurements and the distance between the locations of the measurements and the predictions.

4.4 Model validation using the C-ERA-40 H_s field

In this section we compare the H_s surface given by C-ERA-40 data field at 12:00 on the 17th of December 1996 with the surface predicted using first 19 and then all 43 satellite measurements of section 4.3.2. We consider three types of analysis: Validation of the C-ERA-40 field by means of satellite measurements that are close in time and space; consistency checks of the proposed model by means of predictions when the measurements and the predicted values are further apart and finally prediction in time and space in large areas.

4.4.1 Close in time prediction - comparison with C-ERA-40 data

In this section we compare the H_s values predicted in section 4.3.2, to the C-ERA-40 H_s values.

The 19 satellite measurements used in the analysis of section 4.3.2, were chosen so that the locations and the times of the measurements are close enough to the locations and times of the available C-ERA-40 data. Indeed, the satellite passage took place at 11:24 while the C-ERA-40 data were given at 12:00. Additionally, the distance between the locations of the satellite measurements and the correspnding C-ERA-40 grid points are less that 80 km. Since both the C-ERA-40 and the satellite provide with measurements of the same H_s field and the positions and times of the measurements are quite close, one should expect that the two sources give comparable results.

In Fig. 5, the satellite measurements are marked by '*', the C-ERA-40 values by '+', the predicted values by "." and the irregular lines indicate the 99% confidence intervals. The C-ERA-40 H_s values with the exception of one location appear to be systematically higher than the satellite measurements. The question that arises is if the differences between these two data sets are due to the random character of the H_s field or the two data sets are significantly different. We turn to this problem next.

We start by predicting the H_s values at the locations of the C-ERA-40 data using only the first 19 measurements and the model presented in section (3). The comparison of the two data sets can be found in Fig. 5 (*Left*). With the exception of two locations the C-ERA-40 data fall inside the 99% confidence interval.

In Fig. 5 (*Right*), the H_s values were predicted using all 43 satellite measurements. The last 24 measurements were recorded about 9 hours after the first group of 19 measurements. So, we are also interested in checking if we can satisfactory predict the H_s surface using satellite measurements that are close in space but far apart in time. Remember that buoy data indicate the correlation between measurements at the same location and 10 hours apart is about 0.8. The two sets of predictions are very close to each other and we claim that at least these particular C-ERA-40 values agree well with the measured H_s values and the differences are due to parameter uncertainty and possible model error.

Note that the prediction intervals are quite wide (about one meter), even for predictions that are made in nearby locations and only 40 minutes ahead. Moreover, the assumption the field is log-normal, results into wider prediction intervals for high predicted values.

4.4.2 Prediction of H_s in large regions

We turn now to the prediction of the H_s surface at all 350 grid points of the ERA-40 field at 12:00 on the 17th of December 1996. The predictors $\hat{H}_s(\mathbf{p}, t)$ are based on the 43 satellite measurements analyzed in section 4.3.2. The positions $\mathbf{p}_i$, $i = 1, \ldots, 43$, of the measurements are marked as dots on Fig. 6. The times t_i can be found in Fig. 4 (*Top-left*). Among the 350 grid points only 19 of them have satellite measurements that are at distance less than 80 kilometers and less than 30 minutes before 12:00. Here we would like to remind that even at that case the prediction interval was between one and two meters wide. The values of H_s at the remaining 331 locations are much harder to predict.

The prediction is based on the spatio-temporal model with space parameters presented in Fig. 1 and time correlation length $C_l = 35$ hours. To predict the value $\hat{H}_s(\mathbf{p}, t)$ at position $\mathbf{p}$ and time t, we use satellite measurements that fall inside a discus of radius about 9 degrees, or equivalently 1000 km and times $|t_i - t| < 10$ hours.

As it has been already observed, the C-ERA-40 H_s values at the selected 19 grid points are generally higher than the H_s satellite measurements and in three locations the difference is slightly above the upper 99% prediction interval. In Fig. 6, the contour lines of the difference between the C-ERA-40 data and the $\hat{H}_s(\mathbf{p}, t)$ are given. One can see the biggest difference, of the order of 4 meters, appears in a region close to Greenland where modelling is more difficult. (There is a region where no ERA-40 data are given and hence prediction is not possible.) In general, the agreement between the C-ERA-40 data and the predictions is very good.

Finally, in Fig. 7 (*Left*), the predicted surface can be seen together with the satelletite measuremets that were used in the computation. In Fig. 7 (*Right*) we present the upper bound of the prediction interval, i.e. the surface in Fig. 7 (*Left*) is multiplied by a factor $e^{1.96\cdot\sigma}$. Also notice, that the prediction intervals are computed for individual locations and hence the probability the true surface (assuming the model is correct) exceeds the prediction intervals at some point in the region is much higher than 0.01, which is the probability that this happens at one fixed location.

Finally, we conclude that the agreement between the C-ERA-40 field and the predicted field based on the model and the 43 satellite measurements is acceptable. The uncertainty of the predictions is illustrated in Fig. 7 (*Right*), where the factor needed to multiply with the predicted value in order to get the upper prediction interval is given. In the regions were no observations are available the factor is as high as 2.2.

Acknowledgement

We are indebted to our data sources, namely the ECMWF for the ERA-40 data, Helen Snaith of SOC for the altimeter data, and the American National Oceanographic Data Center for the buoy data.

5 References

[1] Baxevani, A., Podgórski, K., Rychlik, I. (2003). Velocities for moving random surfaces, *Probabilistic Engineering Mechanics* **18**, 251-271.

[2] Baxevani, A. and Rychlik, I., (2006a). Fatigue life prediction for a vessel sailing the North Atlantic route, *Department of Mathematical Sciences,Chalmers University of Technology* PREPRINT 2006:10.

[3] Baxevani, A. and Rychlik, I., (2006b). Maxima for Gaussian seas, *Ocean Engineering*, 33, 895-911.

[4] Baxevani, A., Rychlik, I., Wilson, R. J., (2006). A new method for modeling the space variability of significant wave height. To appear in *Extremes*.

[5] Caires, S. and Sterl, A., (2003). On the estimation of return values of significant wave height data from the reanalysis of the European Centre for Medium-Range Weather Forecasts. Safety and Reliability, *Bedford and van Gelder (Eds.)*, ISBN 9058095517, 353-361.

[6] Caires, S., Sterl., A., Bidlot, J.-R., Graham, N. and Swail, V. (2004). Intercomparison of different wind wave reanalyses *J. Clim.,* 17(10), pp.1893-1913.

[7] Caires, S. and Sterl, A., (2005). A new non-parametric method to correct model data: Application to significant wave height from the ERA-40 reanalysis. *J. Atmospheric and Oceanic Tech.,* 22(4), 443-459.

[8] Challenor, P. and Cotton, D., (1999). Trends in TOPEX significant wave height mea-

surements. Available as PDF document *http://www.soc.soton.ac.uk/JRD/SAT/TOPtren/TOPtren.pdf*(6 pp).

[9] Janssen, P.A.E.M., Doyle, J.D., Bidlot, J., Hansen, B., Isaksen, L., and Viterbo, P., (2002). Impact and feedback of ocean waves on the atmosphere. In W.A. Perrie (Ed.) *Atmosphere-Ocean Interactions*(**I**, 155-197.) Advances in Fluid Mechanics

[10] Komen, G. J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S. and Janssen, P. A. E. M. (1994). *Dynamics and Modelling of Ocean Waves.* Cambridge Univ. Press

6 Appendix

In this study we use data of three different kinds: altimeter, buoy and reanalysis data.

6.1 Altimeter

The TOPEX/Poseidon along track altimeter measurements of significant wave height, H_s, were taken at discrete locations along one-dimensional tracks over the oceans, at different times from October 1992 until January 1999. The data were obtained from the Southampton Oceanography Centre (SOC) (GAPS interface[1]), The TOPEX wave height observations for 1997 to 1999, (cycles 170-253) have drifted; the drift was corrected according to Challenor and Cotton (1999) and Caires and Sterl (2003) using a functional relationship model, $H_s^{buoy} = 1.05H_s^{topex} - 0.07$.

6.2 Buoy

For years, buoy observations were considered as the most reliable existing wave observations. Unfortunatelly they are limited to locations usually along the coast and mainly in the Northern hemisphere and are available only at a small number of locations before 1978. From 1978 and onwards, buoy observations from the American National Data Buoy Center (NDBC-NOAA), from locations off the coast of North America became available. In this work, we have used the NDBC-NOAA deep water buoy data from 20 locations, see Fig. 6.2.

[1] http://www.soc.soton.ac.uk/ALTIMETER/

Buoy	n	$\bar{H}_s$	m_0	m_1	m_2	σ^2	p	C_l	T_l
32302	61084	2.2	0.737	-0.056	-0.123	0.0597	0.9574	116.3	26.7
41001	125580	2.0	0.575	0.144	0.309	0.2042	0.9990	42.0	19.6
41002	92769	1.8	0.489	0.104	0.277	0.1834	0.9988	41.3	21.3
41006	84734	1.7	0.396	0.068	0.323	0.1700	0.9963	51.0	25.0
41010	88333	1.6	0.321	0.051	0.306	0.1811	0.9955	60.2	26.7
42001	135927	1.1	-0.104	0.141	0.413	0.3413	0.9989	48.5	25.0
42002	160398	1.2	0.021	0.152	0.332	0.2995	0.9996	43.1	22.4
42003	145379	1.1	-0.118	0.139	0.410	0.3159	0.9998	36.2	25.0
44004	139634	2.0	0.528	0.137	0.370	0.2485	1.0023	33.8	16.2
46001	166388	2.7	0.879	0.018	0.418	0.1792	0.9994	32.6	17.7
46002	144449	2.7	0.889	0.107	0.372	0.1436	0.9932	43.9	25.0
46003	136120	3.0	0.999	0.056	0.418	0.1614	0.9974	30.9	18.3
46004	63617	2.9	0.960	0.134	0.454	0.1693	0.9982	31.8	23.6
46005	151355	2.7	0.891	0.107	0.441	0.1541	0.9952	32.7	22.4
46006	138596	2.8	0.889	0.162	0.475	0.1514	0.9931	44.6	25.0
46059	42620	2.8	0.916	0.101	0.329	0.1312	0.9896	57.5	23.6
51001	120231	2.5	0.847	0.097	0.247	0.0752	0.9818	45.0	21.3
51002	107611	2.4	0.842	0.070	0.137	0.0581	0.9742	54.3	31.6
51003	109531	2.2	0.778	0.091	0.186	0.0557	0.9724	35.0	23.6
51004	108087	2.4	0.854	0.060	0.139	0.0487	0.9673	56.2	28.7

Table 1: Time correlation for 20 NOAA deep water buoys. $m(\mathbf{p}_0, t) = m_0 + m_1 \cos(\omega t) + m_2 \sin(\omega t)$, is the seasonally varying mean, with phase $\omega = \frac{2\pi}{365.2}$ and time measure in days starting from the 1^{st} of January.

6.3 Reanalysis

Recently, a wave reanalysis data set on a global 1.5 × 1.5 latitude/longitude grid covering the period of 1957 to 2001 has been made available - the ERA-40 dataset. This reanalysis was carried out by the European Centre for Medium-Range Weather Forecasts (ECMWF), using its Integrated Forecasting System, a coupled atmosphere-wave model with variational data assimilation. A distinguishing feature of ECMWF's model is its coupling, through the wave height dependent Charnock parameter (see Janssen et al., 2002), to a third generation wave model, the well-known WAM (Komen et al., 1994), which makes wave data a natural output of ERA-40. A large subset of the complete ERA-40 data set, including H_s, can be freely downloaded and used for scientific purposes from the website http://data.ecmwf.int/data/.

The results of ERA-40 have been extensively validated against observations (Caires and Sterl, 2005) and other reanalysis data sets (Caires et al., 2004). These studies concluded that the ERA-40 data set, although severely underestimating high sea states, compares better with the observations in terms of root mean square error and scatter index than other available datasets. Besides the underestimation of high percentiles, the ERA-40 data set has another limitation that seriously discourages its use in direct studies of climate variability and trends: the existence of inhomogeneities in time due to the assimilation of different altimeter H_s data sets in the ERA-40 computations. These two limitations in the ERA-40 H_s data set motivated their correction by Caires and Sterl (2005). These authors corrected the data using a nonparametric regression method, the main idea of which was to estimate the expected error between ERA-40 H_s and "true" H_s conditional on past (up to 12 hours) and present values of the former, using data from locations at which both ERA-40 and Topex measurements were simultaneously available, and then to use this conditional expected value to correct the whole ERA-40 data. The result was a new 45-year global 6-hourly dataset - the C-ERA-40 dataset. Comparisons of the C-ERA-40 data with measurements from in-situ buoy and global altimeter data show clear improvements in both bias, scatter and percentiles in the whole range of values and the removal of the inhomogeneities present in the ERA-40 dataset. This data set can also be freely obtained for scientific purposes from the authors.

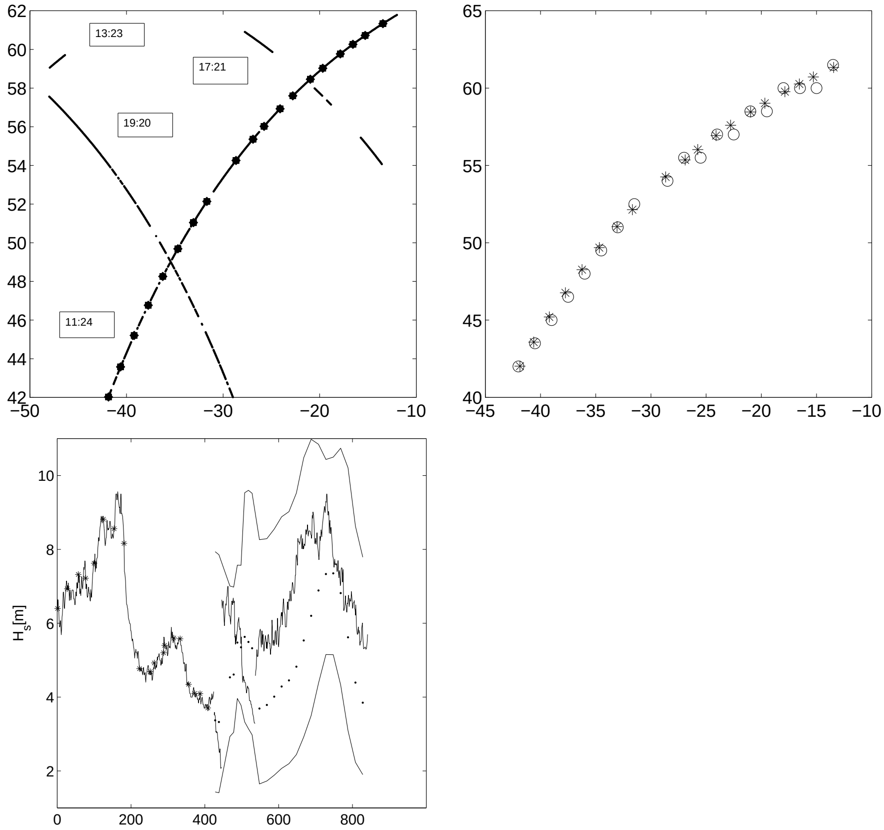

Figure 4: (*Top-left*) - Locations of the satellite measurements between 10:00 and 20:00 on 17th December 1996, the thick black dots mark the 19 locations used to predict the future values of H_s. Times of the passages are given in the boxes. (*Top-right*) - The locations of the ERA-40 H_s values (rings) and satellite measurements (dots). (*Bottom-left*) - Satellite measurements are the irregular line, stars for the 19 values used to predict the future observations of H_s, the 24 predicted values are black dots and the wide lines are the 99% prediction bands. The velocity $\mathbf{v}$ is given in Fig. 3 (*Right*).

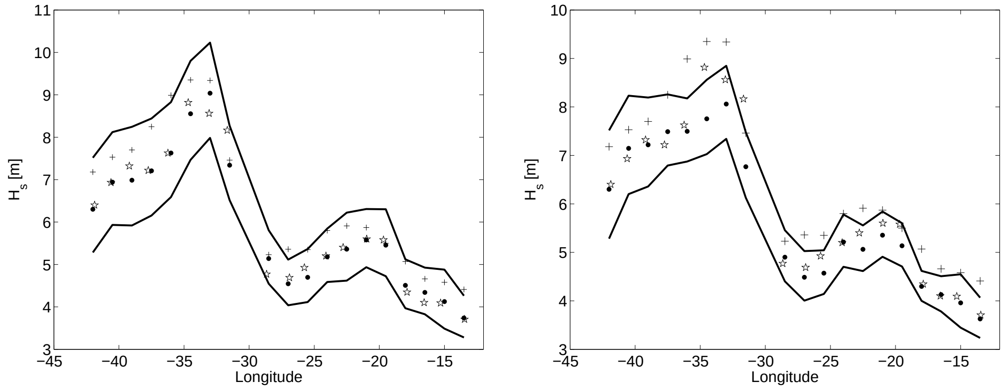

Figure 5: (*Left*) - Values of ERA-40 H_s at 12:00 ('+'); satellite measurements of H_s about 40 minutes earlier (stars); predictions $\hat{H}_s$ at 12:00 at the locations of C-ERA-40 using the 19 satellite observations (dots), black solid lines the 99% prediction bands. (*Right*) - Same as in the (*Left*) plot with $\hat{H}_s$ computed using 43 satellite measurements.

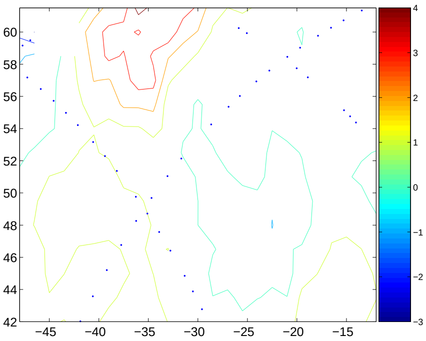

Figure 6: Difference between the C-ERA-40 data and predictions of H_s in the region at 12:00 based on 43 satellite measurements at locations marked by dots. The model has the time correlation length $C_l = 35$ hours and the velocity field in Fig. 2 (*Right*).

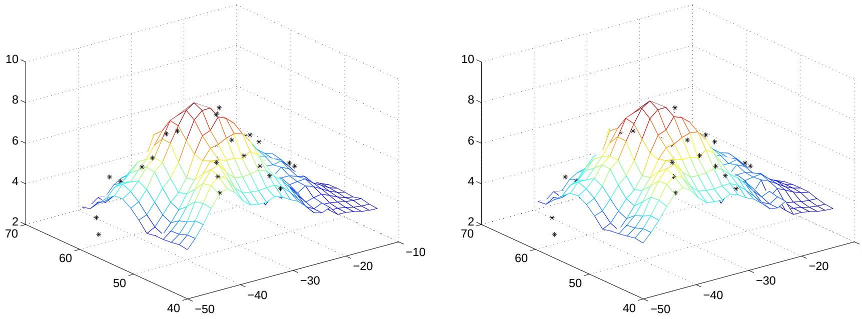

Figure 7: (*Left*) - The predicted surface with satellite measurements marked as stars. (*Right*) - The upper prediction interval $\hat{H}_s \exp(1.96 \cdot \sigma)$.

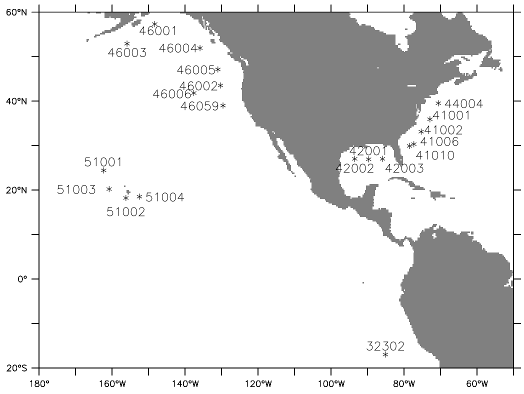

Figure 8: Buoy codes and locations.