An analysis of the interaction between plankton time series and climate via stochastic differential equations

Alessandro De Gregorio
Department of Statistics, Probability and Applied Statistics
P.le Aldo Moro 5, 00185 Rome-Italy
alessandro.degregorio@uniroma1.it

Alessandra Conversi
CNR, ISMAR-La Spezia
Forte S.Teresa, 19032 Pozzuolo di Lerici (SP)-Italy
a.conversi@ismar.cnr.it

Alessandra Micheletti, Daniela Morale
Department of Mathematics "F.Enriques"
Via C.Saldini 50, 20133 Milan-Italy
alessandra.micheletti@unimi.it,daniela.morale@unimi.it

Abstract

The Gulf of Trieste, North Adriatic (eastern Mediterranean), hosts a multidecadal (January 1986 -December 2005, monthly) mesozooplankton time series, which is ideal, for time span and frequency, for investigating the impact of climate change on zooplankton populations. In this study we analyze selected copepod abundance time series to investigate whether a stochastic differential equation approach can provide tools for predicting the pattern of plankton abundance in function of climatic proxies variability. We consider a system of stochastic differential equations and estimate their parameters by using the Euler scheme. Furthermore, we carry out a forecasting analysis and compare the expected values obtained with the real data in order to evaluate the reliability of the proposed models. The results suggest that this approach is promising.

Keywords: Euler method, forecast, principal component analysis, stochastic differential equations, zooplankton, decadal variability, climate impact.

1 Introduction

Copepods are a group of small crustaceans, primarily marine, but also found in nearly every freshwater habitat. They are probably the most numerous multicellular organisms on earth (Mauchline, 1998), and constitute the biggest source of protein in the oceans. Planktonic copepods (drifting in sea waters) are important to global ecology and the carbon cycle. They are usually the dominant members of the zooplankton, and represent the basis for higher trophic levels in the marine chain. Indeed, they are major food organisms for small fish, and thus are responsible for transferring phytoplankton carbon to higher trophic levels, such as fish, birds, and marine mammals. Although they can undertake substantial vertical migrations, they are basically passively transported by currents (in fact, the word plankton derives from the greek word *plank= to drift*). This role is particularly important in the North Adriatic, as this is one of the few regions of high permanent production in the Mediterranean Sea (Buljan, 1964; Fonda-Umani et al., 1992; Fonda-Umani et al., 2004). The Gulf of Trieste is the northernmost section of the Adriatic Sea, and is characterized by an overall shallowness and by large and variable freshwater inputs (Fonda-Umani et al., 1992; Russo and Artegiani, 1996). For this work we have used the copepod monthly time series in the Gulf of Trieste, one of the longest series in the Mediterranean sea (1970-todate). The copepod community in the Gulf of Trieste is characterized by a few (approximately 30) species, which in turn can exhibit high dominance (Kamburska and Fonda-Umani 2006, Conversi et al., 2009).

The aim of the study is the investigation of the impact of climate changes on the abundance of copepod populations. We start our analysis from monthly time series (subset January 1986 - December 2005; chosen because the series contained no gaps) of the abundance of the nineteen copepod taxa. Furthermore we consider three climate predictors, sea surface temperature (SST), north hemisphere temperature (NHT) and east atlantic pattern (EA).

We face this problem by modelling the phenomenon via stochastic differential equations (SDE). Stochastic differential equation models are widely used in different fields. These models have a variety of applications in many disciplines and emerge naturally in the study of many phenomena. In literature, examples of applications to physics (see e.g. Papanicolaou, 1995 for a review), astronomy (Schuecker *et al.*, 2001), social sciences (Cobb, 1981), mathematical finance (Hull, 2000), geology (Ditlevsen *et al.*, 2002), ecology (Holmes, 2004), biology (Ricciardi, 1977, Morale, Capasso, 2000, 2005), can be found. Gutierrez *et al.* (2008) proposed a bivariate stochastic diffusion process to analyze the interdependence between Gross Domestic Product and CO_2 emissions in Spain. The reader interested to study in depth the probabilistic aspects of these processes may consult for example Karatzas and Shreve (2001) or Capasso and Bakstein (2004).

The widespread use of these models and their growing importance in many experimental fields has led to develop their theoretical and practical statistical aspects. In the last decades the research in this field moved along two different lines related to the observation schemes. From one side, models related to the experimental observations of the whole sample path of the process are investigated. The interested reader may refer to Kutoyants (1984, 2004). Recently many statisticians and probabilists have been dealing with models related to a partially observed framework, in which the sample paths of continuum processes are observed only at discrete

instants of time. This sampling scheme is often closer to the reality, since most of times during experiments, processes may be observed only a finite number of times; see Sørensen (2004) for a review.

The model developed as part of this work represents each time series as a discretely observed diffusion process. In order to reduce the dimension of the problem and thus the related computational costs in the estimation procedure, we perform the analysis only on a subset of the plankton species, which may be considered the most representative of the whole plankton population. Then we propose to model trend-cyclical behaviors of the abundances by different stochastic differential equations. The possible dependence of the abundance on the climate predictors are modelled in the drift of the diffusion processes. By using a pseudo-likelihood approach, we estimate the parameters of the model. Furthermore we carry out a forecasting analysis to evaluate its performance.

The paper is organized as follows. In Section 2 we present an explorative analysis, performed via a Principal component analysis, to reduce the data set to a lower dimension. This means we have selected the more representative species of copepods, i.e. the species which best explains the variance in the data. In Section 3 we model each chosen variable via a stochastic differential equation with a linear diffusion. Moreover we present both theoretical and numerical aspects of a pseudo-likelihood method based on the Euler scheme, used for estimating the parameters. The results obtained for the estimates are quoted as well. In Section 4 we carry out a forecasting analysis in order to evaluate the reliability of the adopted model. Finally in Section 5 we summarize and discuss the results. All the numerical and statistical analysis of this work have been performed using R statistical environment (R Development Core Team, 2008).

2 Data Description and Explorative Analysis

Twenty year (January 1986 - December 2005) of monthly abundances (number of individuals per unit volume) of nineteen copepod species, have been used for this work. This subset of the original series (1970-todate) was chosen because it did not incorporate missing data. Mesozooplankton has been collected by vertical hauls from bottom (18 m) to surface with a WP2 net, at a 200 m offshore station (C1) located at 45 42' 03" N, 13 42' 36" E in the Gulf of Trieste - North Adriatic Sea (Italy), eastern Mediterranean (Conversi et al., 2009). The observed species, and their abbreviations (which will be used in the following text) are listed in Table 1. As climatic predictors three specific indices were taken into account:

- Sea Surface Temperature (SST) for the Gulf of Trieste region, derived from the long-term meteorological station of CNR100 ISMAR-Trieste, located at at 453834N, 134514E (Caterini et al., 2007).
- East Atlantic pattern (EA), downloaded from the web site *http://www.cpc.noaa.gov*,
- Northern Hemisphere Temperature (NHT), downloaded from the web site *http://www.cru.uea.ac.uk/*.

While the first index is a local index, the sea surface temperature in the Gulf of Trieste, the latter two are large scale indices. In particular, the Northern Hemisphere Temperature index has

been associated to changes in copepod abundance and diversity in the North Sea (Beaugrand et al., 2008). The Eastern Atlantic pattern is the second prominent mode of low-frequency variability over the North Atlantic, after the North Atlantic Oscillation (NAO), and consists of a north-south dipole of anomaly centers spanning the North Atlantic from east to west, which are displaced southeastward relative to the NAO centers.

For all climate proxies we have used monthly time series from 1986 to 2005, in analogy with biological time series.

Diaixis pygmoea	(Dp)	*Calanus*	(CA)
Oithona	(OI)	*Pseudocalanus elongatus*	(Pe)
Temora	(TE)	*Temora longicornis*	(Tl)
Centropages typicus	(Ct)	*Centropages*	(CE)
Acartia clausi	(Ac)	*Calanus helgolandicus*	(Ch)
Temora stylifera	(Ts)	*Paracalanus parvus*	(Pp)
Euterpina acutifrons	(Ea)	*Oncaea*	(ON)
Centropages kroyeri	(Ck)	*Corycaeus*	(CO)
Harpacticoida	(HA)	*Ctenocalanus vanus*	(Cv)
Clausocalanus	(CL)		

Table 1: List of copepods species and taxa recorded the Gulf of Trieste time series.

In order to simplify our analysis and to reduce the dimension of the problem, we need to select a few species which can be highly representative (in variability terms) of the whole considered plankton population. For this aim, we used the classical Principal Component Analysis (PCA) (see e.g. Mardia et al. 1979) which provides us the eigenvectors listed in Tables 2-3. From the PCA emerges that the variance explained by the first three principal components is

$$\frac{Var(\text{Comp.1}) + Var(\text{Comp.2}) + Var(\text{Comp.3})}{Var(\sum_{i=1}^{19} \text{Comp.i})} = 0.86$$

where Comp.i indicates the i-th principal component. Then it is sufficient to choose, with a suitable method, three biological variables to explain the 86% of the total variability. To discard the variables we start dealing with the eigenvectors in Table 3 corresponding to the smallest eigenvalue (in this case Comp.19) and eliminating the variable with the largest score in absolute value (in this case Dp). The same operation is iterated considering the eigenvector related to the next smallest eigenvalue and rejecting the variables with highest score. Ultimately, the above procedure selects three plankton species: *Acartia clausi*, *Oithona* and *Oncea*. From an ecosystem point of view, these three taxa are rather relevant in this area: *Acartia clausi* dominates the Gulf of Trieste most of the year, comprising at some points >80% of the total biomass, but its dominance has decreased in the last two decades, while the small copepods *Oithona* and *Oncea* have become the next most abundant taxa for the selected period of time (Conversi et al., 2009)

Obviously both plankton and climate variables might be affected by seasonality. In order to consider only the trend-cyclical component, we have then filtered linearly the data by means of

a moving average operator of the type

$$T_t = \frac{1}{2a+1} \sum_{t=-a}^{a} X_t, \tag{2.1}$$

where X_t is the value of the abundance at time t. We applied Filter (2.1) to the plankton and to the climate time series with $a = 12$. In other words we eliminated the seasonal component by averaging over 25 contiguous terms for each value. Hence, we obtained a new dataset on which we will work in the rest of the paper. It's clear that in this way the data of the years 1986 and 2005 have been eliminated, in fact the first data available after the application of Operator (2.1) correspond to January 1987 and last one correspond to December 2004.

We denote by X_1 ="*Oithona* abundance", X_2 ="*Acartia clausi* abundance", X_3 ="*Oncaea* abundance", u_1 ="SST", u_2 ="NHT", u_3 ="EA", the trend-cycle component of the biological and climate variables. It will be useful in the next section to consider the correlation matrices for $\mathbf{X} = (X_1, X_2, X_3)$ and $\mathbf{u} = (u_1, u_2, u_3)$, respectively

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} 1.00 & 0.62 & \text{-}0.06 \\ 0.62 & 1.00 & 0.09 \\ \text{-}0.06 & 0.09 & 1.00 \end{pmatrix} \tag{2.2}$$

$$\Sigma_{\mathbf{u}} = \begin{pmatrix} 1.00 & \text{-}0.01 & 0.42 \\ \text{-}0.01 & 1.00 & 0.68 \\ 0.42 & 0.68 & 1.00 \end{pmatrix} \tag{2.3}$$

Looking at $\Sigma_{\mathbf{X}}$ we observe that the variables X_1 and X_2 have a positive correlation, while X_3 seems to be scarcely correlated with the other plankton species. From $\Sigma_{\mathbf{u}}$ it can be observed that the variables u_1, u_3 and u_2, u_3 are positively correlated.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Dp	0.00	-0.00	-0.00	0.00	0.00	0.00	-0.01	0.00	-0.00	0.01
CA	-0.00	0.01	0.00	0.00	-0.11	0.02	0.03	-0.00	-0.02	0.00
OI	-0.30	-0.74	-0.60	-0.07	-0.00	0.01	0.02	0.02	0.02	0.01
Pe	-0.01	-0.01	-0.01	0.01	0.01	-0.00	-0.04	-0.03	-0.04	-0.04
TE	-0.01	-0.00	-0.03	0.01	0.00	-0.05	-0.05	0.17	-0.55	-0.38
Tl	-0.01	-0.02	-0.00	0.01	0.00	-0.01	-0.02	-0.02	-0.16	-0.10
Ct	-0.09	0.07	0.03	-0.27	-0.02	0.80	0.23	0.44	-0.08	-0.02
CE	-0.04	0.00	0.02	-0.02	-0.00	-0.32	-0.17	0.70	0.01	0.16
Ac	-0.93	0.22	0.17	0.23	-0.01	-0.02	0.01	-0.07	-0.01	-0.01
Ch	-0.00	-0.00	-0.01	0.00	0.00	0.01	0.01	0.02	-0.01	-0.06
Ts	0.01	0.03	-0.04	0.01	-0.97	0.01	-0.01	0.01	0.16	0.02
Pp	-0.17	0.24	-0.11	-0.90	0.03	-0.24	0.03	-0.15	0.07	0.02
Ea	0.02	0.05	-0.08	0.00	-0.02	-0.02	0.02	-0.02	0.22	-0.89
ON	0.05	0.58	-0.76	0.23	0.06	0.03	0.07	0.05	0.01	0.09
Ck	-0.04	0.02	0.02	-0.01	0.00	-0.32	-0.05	0.49	0.15	-0.12
CO	0.01	0.03	-0.05	-0.07	-0.21	-0.15	0.03	-0.08	-0.75	0.03
HA	-0.00	0.00	-0.00	-0.01	0.00	-0.04	0.01	0.05	-0.01	0.03
Cv	-0.01	0.01	-0.02	0.02	0.01	-0.01	-0.05	0.02	-0.04	0.00
CL	-0.02	0.06	-0.07	-0.08	-0.01	0.26	-0.95	-0.06	-0.01	-0.01

Table 2: Eigenvectors associated to the principal components for the plankton data.

	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18	Comp.19
Dp	0.00	-0.01	-0.03	0.01	0.03	-0.01	0.01	-0.00	1.00
CA	0.07	0.03	-0.78	0.58	-0.02	-0.06	-0.19	0.00	-0.02
OI	-0.01	-0.02	0.00	0.01	-0.00	-0.01	0.00	-0.00	-0.00
Pe	0.04	0.08	-0.23	-0.49	0.35	0.33	-0.65	0.20	-0.00
TE	0.58	0.20	0.24	0.23	0.16	0.07	-0.02	0.00	0.00
Tl	0.27	0.22	-0.37	-0.54	-0.57	-0.24	0.16	-0.02	0.01
Ct	-0.07	-0.08	-0.01	-0.05	-0.01	-0.02	-0.02	0.01	0.00
CE	-0.31	0.49	-0.02	0.00	0.01	-0.01	-0.03	-0.14	0.00
Ac	-0.02	0.01	0.01	0.01	-0.00	0.00	0.01	0.00	0.00
Ch	-0.01	0.01	-0.29	-0.11	0.20	0.68	0.63	-0.00	-0.01
Ts	0.12	0.06	0.10	-0.06	0.03	0.01	0.03	-0.00	0.00
Pp	0.08	0.06	0.00	0.01	0.02	0.01	0.00	-0.02	0.00
Ea	-0.37	0.07	-0.03	0.00	-0.01	-0.08	-0.01	-0.02	0.01
ON	0.03	0.00	-0.00	-0.00	-0.02	0.01	-0.01	0.00	-0.00
Ck	0.22	-0.73	-0.06	-0.07	-0.11	0.02	-0.03	0.12	-0.01
CO	-0.52	-0.29	-0.05	-0.05	-0.07	-0.01	-0.01	0.00	0.00
HA	-0.07	0.14	-0.01	0.03	0.13	-0.24	0.22	0.92	-0.00
Cv	0.06	-0.11	-0.21	-0.24	0.67	-0.55	0.24	-0.27	-0.04
CL	-0.04	-0.09	-0.01	0.04	-0.05	0.01	0.01	0.04	-0.01

Table 3: (continued) Eigenvectors associated to the principal components for the plankton data.

3 SDE models and inference

In order to study the behaviour of the plankton species selected in the previous section, we assume that the variables X_1, X_2, X_3 (which evolve in continuous time) are solutions of a SDE. In other words the underlying random model which explains the abundance of the plankton is a diffusion process (see Karatzas and Shreve, 2001 for an extensive theoretical treatment of this topic). We assume that the trend, that is the drift in the SDE framework, is defined by means of a linear combination of the climatic variables and a possible interaction with one of the other plankton species. Furthermore we will use a diffusion term with linear unknown parameters. Therefore, by setting $\theta_k = (\alpha_{1k}, \alpha_{2k}, \beta)$ and $\mathbf{u}(t) = (u_1(t), u_2(t), u_3(t))$, we assume that $X_k(t), t \geqslant 0$, $k = 1, 2, 3$, satisfies the following SDE system

$$\mathrm{d}X_k(t) = b(\theta_k, \mathbf{u}(t), X_k(t), X_j(t))\mathrm{d}t + \sigma_k \sqrt{X_k(t)}\mathrm{d}W_k(t) \tag{3.1}$$

where the drift term is equal to

$$b(\theta_k, \mathbf{u}(t), X_k(t), X_j(t)) = X_k(t)\left[1 + \alpha_{1k}u_1(t)u_3(t) + \alpha_{2k}u_2(t)u_3(t) + \beta_k X_j(t)\right]$$

with $k, j = 1, 2, 3,\ k \neq j$ and W_k is a standard Brownian motion, independent from $W_j, k \neq j$.

The parametric estimation problem in our case concerns a partially observed (at equidistant discrete times) diffusion process. For an account of the statistical methods in this context the reader can consult Sørensen (2004) and Iacus (2008). To estimate the parameters θ_k and σ_k we use a pseudo-likelihood method based on the Euler scheme. The discretization reads

$$X_k(t_i) - X_k(t_{i-1}) = b(\theta_k, \mathbf{u}(t_{i-1}), X_k(t_{i-1}), X_j(t_{i-1}))\Delta_n + \sigma_k \sqrt{X_k(t_{i-1})}\Delta_i W_k$$

where $X_k(t_i), \mathbf{u}(t_{i-1})$ represent the values of X_k and $\mathbf{u}(t)$ at time t_i, $\Delta_n = t_i - t_{i-1}$ and $\Delta_i W_k = W_k(t_i) - W_k(t_{i-1})$. The increments $X_k(t_i) - X_k(t_{i-1})$ are independent gaussian random variables and the likelihood function of the sample becomes

$$\begin{aligned} L_k(\vartheta_k) &= \frac{1}{\prod_{i=1}^n (2\pi\sigma_k^2 X_k(t_i)\Delta_n)^{1/2}} \\ &\times \exp\left\{ -\sum_{i=1}^n \frac{[X_k(t_i) - X_k(t_{i-1}) - b(\theta_k, \mathbf{u}(t_{i-1}), X_k(t_{i-1}), X_j(t_{i-1}))\Delta_n]^2}{2\sigma_k^2 X_k(t_{i-1})\Delta_n} \right\} \end{aligned}$$

where $\vartheta_k = (\theta_k, \sigma_k)$. Now we define the approximated log-likelihood in the following way

$$\begin{aligned} l_k(\vartheta_k) &= -\frac{1}{2}\sum_{i=1}^n \frac{[X_k(t_i) - X_k(t_{i-1}) - b(\theta_k, \mathbf{u}(t_{i-1}), X_k(t_{i-1}), X_j(t_{i-1}))\Delta_n]^2}{\sigma_k^2 X_k(t_{i-1})\Delta_n} \\ &\quad -\frac{1}{2}\sum_{i=1}^n \log(2\pi\sigma_k^2 X_k(t_{i-1})\Delta_n) \end{aligned}$$

and the maximum likelihood estimator is

$$\hat{\vartheta}_k = (\hat{\theta}_k, \hat{\sigma}_k) = (\hat{\alpha}_{1k}, \hat{\alpha}_{2k}, \hat{\beta}_k, \hat{\sigma}_k) = \arg\max_{\vartheta_k} l_k(\vartheta_k) \tag{3.2}$$

From a numerical point of view, the choice of the mesh Δ_n is crucial. In our case we are considering monthly observations so that we fix $\Delta_n = 1/12$. We recall that the Euler method works with small values of Δ_n. For more information on numerical implementation of this method we refer to Kloeden and Platen (1994) and Iacus (2008).

It's important also to observe that to compute the values of Estimator (3.2), we will use a numerical optimization method ("L-BFGS-B") developed in Byrd et al. (1995). This method allows box constraints, that is each variable can be given a lower and/or upper bound. The initial value must satisfy the constraints. We treat separately the SDE's corresponding to the three species and use for our implementation the data from January 1987 to December 2003. This is because, in the last section, we will use the observed data of the last year 2004 for a comparison (and thus a validation) with the forecasting of the models.

3.1 Parametric estimation for X_1

The correlation matrix $\Sigma_{\mathbf{X}}$ suggests us to consider an interdependence between X_1 and X_2, then from (3.1) for X_1 the following model is assumed

$$\mathrm{d}X_1(t) = b(\theta_1, \mathbf{u}(t), X_1(t), X_2(t))\mathrm{d}t + \sigma_1\sqrt{X_1(t)}\mathrm{d}W_1(t). \tag{3.3}$$

We estimate the parameters of (3.3) by using Euler's method with the optimization procedure described above. The estimated value for β_1 is $\hat{\beta}_1 = 0$. In other words the contribute of X_2 to explain X_1 seems to be negligible. For this reason we consider the following alternative stochastic differential equations in which the interaction with X_2 is absent:

- SDE.A1:

$$\mathrm{d}X_1(t) = X_1(t)\left[1 + \alpha_{11}u_1(t)u_3(t) + \alpha_{21}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_1\sqrt{X_1(t)}\mathrm{d}W_1(t) \tag{3.4}$$

- SDE.A2:

$$\mathrm{d}X_1(t) = X_1(t)\left[1 + \alpha_{11}u_1(t)u_3(t)\right]\mathrm{d}t + \sigma_1\sqrt{X_1(t)}\mathrm{d}W_1(t) \tag{3.5}$$

- SDE.A3:

$$\mathrm{d}X_1(t) = X_1(t)\left[1 + \alpha_{21}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_1\sqrt{X_1(t)}\mathrm{d}W_1(t) \tag{3.6}$$

and estimate their parameters applying again the same method. We observe that for the models (3.5) and (3.6) it's easy to define in closed-form the maximum likelihood estimator $\hat{\vartheta}_k$, i.e.

$$\hat{\alpha}_{k1} = \frac{\sum_{i=1}^{n} u_k(t_{i-1})u_3(t_{i-1})\left[X_1(t_i) - 2X_1(t_{i-1})\right]}{\sum_{i=1}^{n}(u_k(t_{i-1})u_3(t_{i-1}))^2 X_1(t_i)}$$

$$\hat{\sigma}_1^2 = \frac{1}{2n\Delta_n}\sum_{i=1}^{n}\frac{\left[X_1(t_i) - X_1(t_{i-1}) - X_1(t_{i-1})(1 + \hat{\alpha}_{k1}u_k(t_{i-1})u_3(t_{i-1}))\Delta_n\right]^2}{X_1(t_{i-1})}$$

with $k = 1, 2$. It's clear that the models (3.5) and (3.6) are embedded in (3.4). The results of the estimates have been listed in Table 4.

3.2 Parametric estimation for X_2

The same remarks raised for X_1 hold also for X_2. That is we can write

$$\mathrm{d}X_2(t) = b(\theta_2, \mathbf{u}(t), X_2(t), X_1(t))\mathrm{d}t + \sigma_2\sqrt{X_2(t)}\mathrm{d}W_2(t), \tag{3.7}$$

where the interaction with X_1 emerges. The estimated value of β_2 is $\hat{\beta}_2 = 0$, so that the contribute of X_1 in explaining X_2 is negligible. Thus the models become

- SDE.B1:

$$\mathrm{d}X_2(t) = X_2(t)\left[1 + \alpha_{12}u_1(t)u_3(t) + \alpha_{22}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_2\sqrt{X_2(t)}\mathrm{d}W_2(t) \tag{3.8}$$

- SDE.B2:

$$\mathrm{d}X_2(t) = X_2(t)\left[1 + \alpha_{12}u_1(t)u_3(t)\right]\mathrm{d}t + \sigma_2\sqrt{X_2(t)}\mathrm{d}W_2(t) \tag{3.9}$$

- SDE.B3:

$$\mathrm{d}X_2(t) = X_2(t)\left[1 + \alpha_{22}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_2\sqrt{X_2(t)}\mathrm{d}W_2(t) \tag{3.10}$$

We observe that for the models (3.9) and (3.10) the estimator $\hat{\vartheta}_k$ is defined in closed form, i.e.

$$\hat{\alpha}_{k2} = \frac{\sum_{i=1}^{n} u_k(t_{i-1})u_3(t_{i-1})\left[X_2(t_i) - 2X_2(t_{i-1})\right]}{\sum_{i=1}^{n}(u_k(t_{i-1})u_3(t_{i-1}))^2 X_2(t_i)}$$

$$\hat{\sigma}_2^2 = \frac{1}{2n\Delta_n}\sum_{i=1}^{n}\frac{\left[X_2(t_i) - X_2(t_{i-1}) - X_2(t_{i-1})(1 + \hat{\alpha}_{k2}u_k(t_{i-1})u_3(t_{i-1}))\Delta_n\right]^2}{X_2(t_{i-1})}$$

with $k = 1, 2$. The results of the estimates are listed in Table 5.

3.3 Parametric estimation for X_3

From the correlation matrix $\Sigma_{\mathbf{X}}$ it emerges that the variable X_3 is uncorrelated with X_1 and X_2; therefore we don't consider a possible interaction with the previous species. Thus in this case, we consider directly X_3 as solution of the following stochastic differential equations

- SDE.C1:

$$\mathrm{d}X_3(t) = X_3(t)\left[1 + \alpha_{13}u_1(t)u_3(t) + \alpha_{23}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_3\sqrt{X_3(t)}\mathrm{d}W_3(t) \tag{3.11}$$

- SDE.C2:

$$\mathrm{d}X_3(t) = X_3(t)\left[1 + \alpha_{13}u_1(t)u_3(t)\right]\mathrm{d}t + \sigma_3\sqrt{X_3(t)}\mathrm{d}W_3(t) \tag{3.12}$$

- SDE.C3:

$$\mathrm{d}X_3(t) = X_3(t)\left[1 + \alpha_{23}u_2(t)u_3(t)\right]\mathrm{d}t + \sigma_3\sqrt{X_3(t)}\mathrm{d}W_3(t) \tag{3.13}$$

Analogously to X_1 and X_2, for (3.12) and (3.13) we have that

$$\hat{\alpha}_{k3} = \frac{\sum_{i=1}^{n} u_k(t_{i-1})u_3(t_{i-1})\left[X_3(t_i) - 2X_3(t_{i-1})\right]}{\sum_{i=1}^{n}(u_k(t_{i-1})u_3(t_{i-1}))^2 X_3(t_i)}$$

$$\hat{\sigma}_3^2 = \frac{1}{2n\Delta_n}\sum_{i=1}^{n}\frac{\left[X_3(t_i) - X_3(t_{i-1}) - X_3(t_{i-1})(1+\hat{\alpha}_{k3}u_k(t_{i-1})u_3(t_{i-1}))\Delta_n\right]^2}{X_3(t_{i-1})}$$

with $k = 1, 2$. Table 6 contains the estimates of the parameters of (3.11), (3.12) and (3.13).

4 Forecasting results

In order to compare and evaluate the performance of the SDE models introduced for X_1, X_2, X_3 in the previous section, we perform a forecasting analysis by Monte Carlo simulations. In this way we are also able to select the model which better predicts the real data. We want to forecast the values of the biological time series during 2004. To simulate the sample paths of the process we use the estimates $(\hat{\theta}_k, \hat{\sigma}_k)$ obtained in the previous Section and the Euler scheme as follows

$$X_k(t_i) = X_k(t_{i-1}) + b(\hat{\theta}_k, \mathbf{u}(t_{i-1}), X_k(t_{i-1}))\Delta_n + \hat{\sigma}_k\sqrt{X_k(t_{i-1})}\Delta_n Z \tag{4.1}$$

where $Z \sim N(0,1)$. So we fix the starting point at January 2004. Hence, we simulate 12 (one for each month) values of the random variable Z and calculate, via Expression (4.1), the values of the variable X_k. For each variable X_k, we repeat these steps 10000 times for the three proposed diffusion models. The results of our forecasting analysis are depicted in Figures 1-6.

In order to measure the discrepancy between the simulated values and the real data, we introduce the following percentage index

$$\mathcal{D}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_k(t_i) - \bar{X}_k(t_i)|}{X_k(t_i)} * 100, \quad k = 1, 2, 3, \tag{4.2}$$

with $\bar{X}_k(t_i) = \frac{1}{10000}\sum_{j=1}^{10000} X_k^j(t_i)$, where $X_k^j(t_i)$ represents the predicted value for $X_k(t_i)$ by the j-th simulation. The index $\mathcal{D}_k$ gives us the mean relative distance between the real data and the average of the simulated values at time t_i.

As one can see in Figure 1 for X_1 (*Oithona*) the three sequences of average values $\bar{X}_1(t_i)$ capture the global trend of the real time series during the year. However, the model that seems to provide us the better forecasting is (3.6); in fact for SDE.A3 the discrepancy index reads

$$\mathcal{D}_1 = 4.27\%$$

whilst for the models (3.4) and (3.5) we have respectively $\mathcal{D}_1 = 15.18\%$ and $\mathcal{D}_1 = 7.74\%$. Furthermore, if one observes the boxplots in Figure 2, the real values always fall into the confidence interval containing the 50% of the empirical distributions emerging from SDE.A3. These considerations show that, from a forecasting point of view, the best model is (3.6), which considers only the combination of the NHT and EA indeces to explain the *Oithona*'s abundance.

Similar considerations can be done with X_2 (*Acartia clausi*). In fact, with respect to the models SDE.B1 ($\mathcal{D}_2 = 7.56\%$) and SDE.B2 ($\mathcal{D}_2 = 13.2\%$), the prediction (Figure 3) emerging from Model SDE.B3 is better, being in this case

$$\mathcal{D}_2 = 6.93\%.$$

Furthermore, Figure 4 shows that the true observations (with the exception of May and June) are between the first and third quartile of the empirical distributions emerging from (3.10). Hence, X_1 and X_2 have the same relationships with the climatic variables.

For the last variable X_3 (*Oncea*) we can conclude that the best model to make our forecast is SDE.C1, being

$$\mathcal{D}_3 = 5.22\%$$

whilst for the models SDE.C2 and SDE.C3 we have respectively $\mathcal{D}_3 = 13.9\%$ and $\mathcal{D}_3 = 17.73\%$. Figure 6 shows that the true observations (with the exception of March and April) are between the first and third quartile of the empirical distributions emerging from (3.11). Therefore, in this case the trend component is defined by a linear combination between the couples SST, EA and EA, NHT.

5 Conclusions

To conclude it is opportune to summarize the results obtained in this paper. We have analyzed the copepod population in the Gulf of Trieste (Northern Adriatic Sea) and selected the following representative representative taxa: *Acartia clausi* (the dominant species), *Oithona*, and *Oncea* (taxa that have shown a large increment in numbers). We have focused our attention on the dependence of the abundance of these taxa on climatic proxies and a possible interaction with one of the other plankton species. We have evaluated the feasibility of using SDEs models to forecast plankton abundance. Stochastic models have been used for addressing prey-predator relationships (e.g., marine plankton predation by planktivorous fish, Mukhopadhyay and Bhattacharyya, 2008), but climate-plankton relationships are usually investigated using deterministic models, such as regression and correlations (e.g., Drinkwater et al., 2003, Beaugrand et al., 2008). However, given the high stochasticity of climate processes, SDE models should provide better forecasting capabilities. Our results suggest that SDEs are a suitable mathematical tool to study the evolution of the plankton abundance. We have estimated the parameters of the introduced models and used the estimates to forecast one year. Specifically, the following considerations hold:

- interactions among different species seem to be negligible, and the major predictors are associated to climate variables. Specifically:
- *Oithona*: a linear combination of the indexes EA, NHT seems to explain the trend of this species;
- *Acartia clausi*: the same conclusions for *Oithona* hold;

- *Oncea*: in this case the best model combines in the drift an interaction between EA, NHT and SST, EA.

The above results suggest that SDE's models can provide a useful tool to forecast plankton variations as functions of climate, which may be applied in developing climate-plankton scenarios.

Acknowledgments

We are indebted to Serena Fonda-Umani for providing the copepod time series and for sharing her knowledge on the planktonic community in the Gulf of Trieste, and to Fulvio Crisciani and Renato Colucci for providing us with the SST data. We are thankful to Tiziana Peluso for her help with data handling and for constructive criticism. This work has been supported by the Italian Project VECTOR (VulnErability of the Italian coastal area and marine Ecosystems to Climatic changes and Their rOle in the Mediterranean caRbon cycles; http://vector-conisma.geo.unimib.it/).

References

[1] Beaugrand G., Edwards M., Brander K., Luczak C., Ibanez F. (2008) Causes and projections of abrupt climate-driven ecosystem shifts in the North Atlantic, *Ecology Letters*, 11, 1157-1168

[2] Boi, S., V. Capasso, Morale, D. (2000) Modeling the aggregative behavior of ants of the species Polyergus rufescens, *Nonlinear Analysis: Real World Applications*, I, 163-176

[3] Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16, 11901208.

[4] Buljan M. (1964). An estimate of productivity of the Adriatic Sea made on the basis of its hydrographic properties, *Acta Adriat.*, 11, 4, 35-45.

[5] Capasso V., Bakstein D. *An Introduction to Continuous-Time Stochastic Processes - Theory, Models and Applications to Finance, Biology and Medicine.* Birkhäuser, Boston.

[6] Caterini, E., Crisciani F., Ferraro S., Immediato F., Maselli M., Raicich F. (2007) Dati meteorologici di Trieste - Riassunti mensili anno 1996, CNR - Istituto Sperimentale Talassografico di Trieste, LINT Trieste s.r.l.

[7] Cobb, L. (1981) Stochastic Differential Equations for the Social Sciences, in *Mathematical Frontiers of the Social and Policy Sciences*, Cobb & Thrall eds, Westview Press, 1-26.

[8] Conversi A., Peluso T., Fonda-Umani S. (2009), The Gulf of Trieste: a changing ecosystem. *J.Geophys. Res.* (in press).

[9] Ditlevsen, P.D., Ditlevsen, S., Andersen, K. K. (2002) The fast climate fluctuations during the stadial and interstadial climate states, *Ann. Glaciol.*, 35, 457-462.

[10] Drinkwater, K., Belgrano A., Borja, A., Conversi, A., Greene, C., Ottersen, G., Pershing, A. (2003). The response of marine ecosystems to climate variability associated with the North Atlantic Oscillation. In: The North Atlantic Oscillation: climatic significance and environmental impact, edited by J. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck, American Geophysical Union, Washington DC, Geophysical Monograph Series, Vol. 134, pp. 211-234.

[11] Fonda-Umani S., Franco, P., Ghirardelli, E., Malej, A. (1992), Outline of oceanography and the plankton of the Adriatic Sea, in Marine Eutrophication and Population Dynamics, edited by G. Colombo, I. Ferrari, V.U. Ceccherelli and R. Rossi, pp. 347365, Olsen and Olsen.

[12] Fonda Umani, S., A. Beran, S. Parlato, D. Virgilio, T. Zollet, A. De Olazabal, B. Lazzarini, and M. Cabrini (2004), Noctiluca scintillans MACARTNEY in the Northern Adriatic Sea: long-term dynamics, relationship with temperature and eutrophication, and role in the food web, *J. Plankton Res.*, 26, 1-17.

[13] Gutierrez, R., Gutierrez-Sanchez, R., Nafidi, A. (2008) A bivariate Gompertz diffusion model: statistical aspects and application to the joint modelling of the Gross Domestic Product and CO_2 emissions in Spain, *Environmetrics*, 19, 643-658.

[14] Holmes, E.E. (2004) Beyond theory to application and evaluation: di?usion approximations for population viability analysis, *Ecological Applications*, 14, 1272-1293.

[15] Hull, J. (2000) *Options, Futures and Other Derivatives, Prentice-Hall*, Englewood Cliffs, NJ.

[16] Iacus, S.M. (2008) *Simulation and Inference for Stochastic Differential Equations*, Springer, New York.

[17] Karatzas, I., Shrevre, S.E. (2001) *Brownian motion and stochastic calculus*, Springer, New York.

[18] Kamburska, L., and S. Fonda-Umani (2006), Long-term copepod dynamic in the Gulf of Trieste (Northern Adriatic Sea). Recent changes and trends, *Clim. Res.*, 31, 195-203.

[19] Kutoyants, Yu. (1984) *Parametric estimation for stochastic processes*. Heldermann-Verlag. Berlin.

[20] Kessler, M. (1997) Estimation of an ergodic diffusion from discrete observations, *Scand. J. Statist.*, 24, 211-229.

[21] Kloden, P., Platen, E. (1999) *Numerical Solution of Stochastic Differential Equations*, Springer, New York.

[22] Kutoyants, Yu. (2004) *Statistical inference for ergodic diffusion processes*. Springer, New York.

[23] Papanicolaou, G. (1995) Diffusions in random media, in *Surveys in Applied Mathematics*, J.B. Keller, D. McLaughin and G. Papanicolaou (Eds), Plenum Press, 205-255.

[24] Mardia, K.V., Kent, J.T., Bibby, J.M. (1979) *Multivariate Analysis*, Academic Press, New York.

[25] Mauchline J. (1998). Advances in marine biology - The biology of calanoid copepods - VOL 33, Academic Press, San Diego, 710 pp.

[26] Morale, D., Capasso, V. Oelschlaeger K. (2005) An interacting particle system modelling aggregation behavior: from individuals to populations. *J. Mathematical Biology*, 50, 1, 49 - 66.

[27] R Development Core Team (2008) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[28] Ricciardi, L.M. (1977) *Diffusion Processes and Related Topics in Biology*, Lecture Notes in Biomathematics, Springer, New York.

[29] Russo, A., and A. Artegiani (1996), Adriatic Sea hydrography, *Sci. Mar.*,60, 3-43.

[30] Schuecker, P., Bohringer, H., Arzner, K., Reiprich, T. H. (2001) Cosmic mass functions from Gaussian stochastic diffusion processes, *Astronomy & Astrophysics*, 370, 715-728.

[31] Sørensen, H. (2004) Parametric inference for diffusion processes observed at discrete points time: a survey. *International Statistical Review*, 72, 337-354.

	Est.	Std. Err.
α_{11}	-0.20	0.04
α_{21}	3.54	1.30
σ_1	8.56	0.43

	Est.	Std. Err.
α_{11}	-0.10	0.01
σ_1	8.71	0.43

	Est.	Std. Err.
α_{21}	-3.26	0.43
σ_1	9.17	0.46

Table 4: Estimates of the parameters from left to right respectively of the models (3.4), (3.5) and (3.6).

	Est.	Std. Err.
α_{12}	-0.19	0.04
α_{22}	3.34	1.28
σ_2	9.68	0.48

	Est.	Std. Err.
α_{12}	-0.10	0.01
σ_2	9.84	0.49

	Est.	Std. Err.
α_{22}	-3.26	0.36
σ_2	10.34	0.51

Table 5: Estimates of the parameters from left to right respectively of the models (3.8), (3.9) and (3.10).

	Est.	Std. Err.
α_{13}	-0.26	0.05
α_{23}	5.34	1.78
σ_3	7.75	0.38

	Est.	Std. Err.
α_{13}	-0.11	0.01
σ_3	7.92	0.39

	Est.	Std. Err.
α_{23}	-3.46	0.42
σ_3	8.23	0.41

Table 6: Estimates of the parameters from left to right respectively of the models (3.11), (3.12) and (3.13).

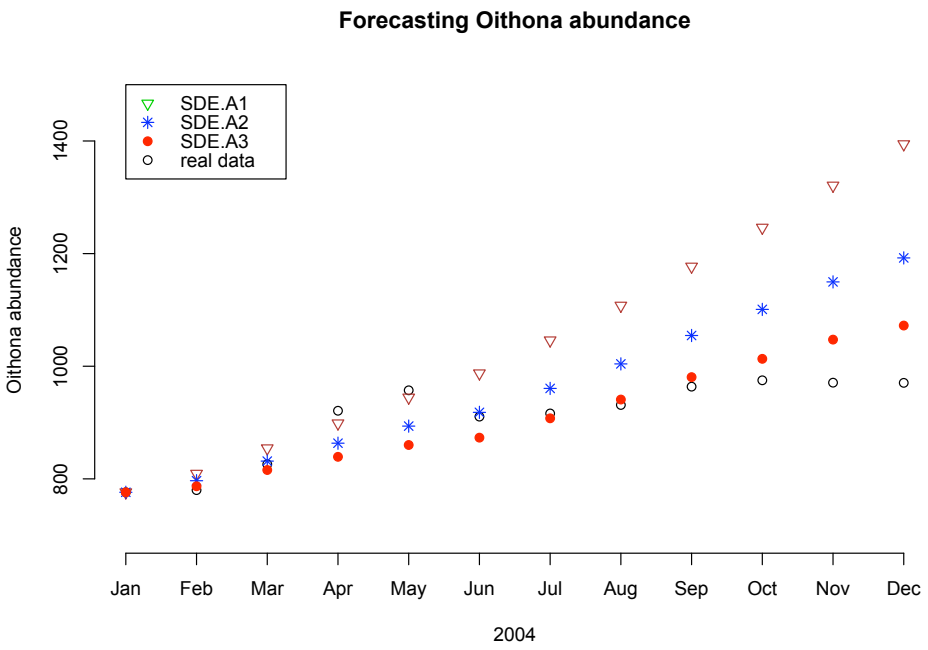

Figure 1: Average simulated values $\bar{X}_1$ for every month on 10000 replications obtained by using SDE.A1, SDE.A2, SDE.A3. The figure also shows the real data for *Oithona* abundance regarded in the year.

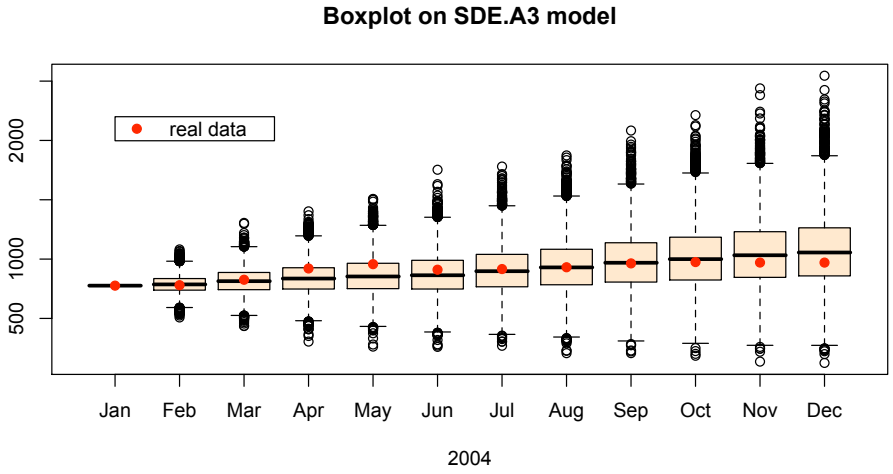

Figure 2: Boxplots for the simulated values derived by SDE.A3 model. The figure also shows the real data.

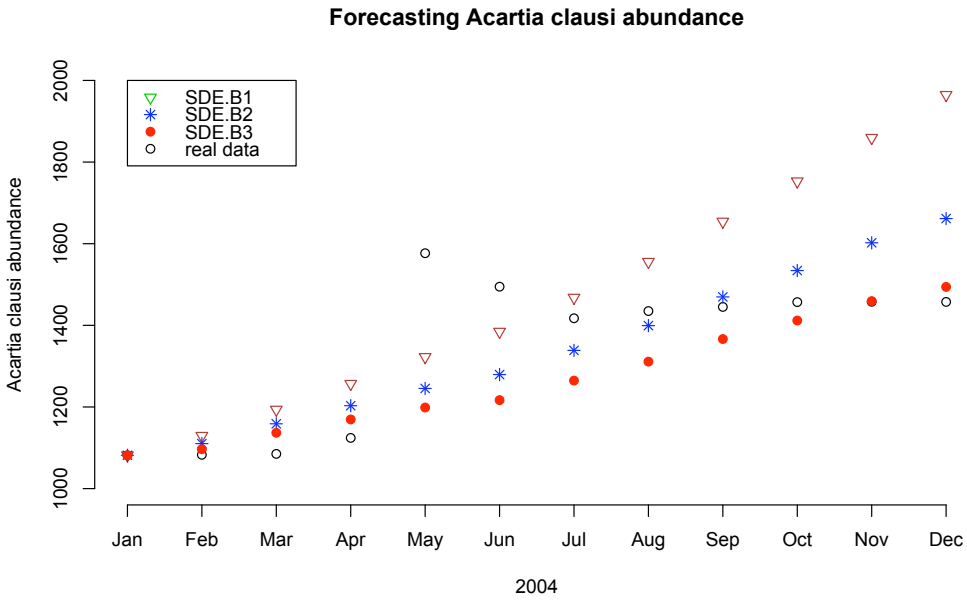

Figure 3: Average simulated values $\bar{X}_2$ for every month on 10000 replications obtained by using SDE.B1, SDE.B2, SDE.B3. The figure also shows the real data regarded in the year.

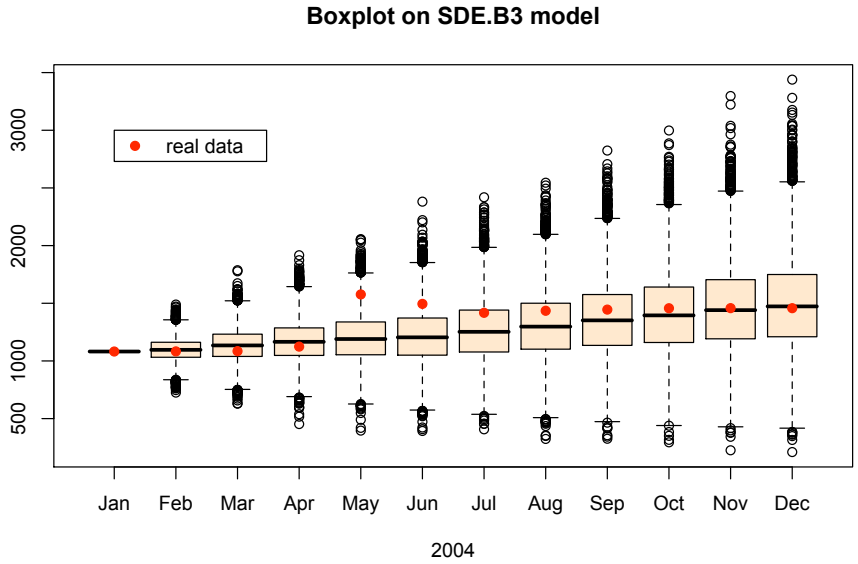

Figure 4: Boxplots for the simulated values derived by SDE.B3 model. The figure also shows the real data.

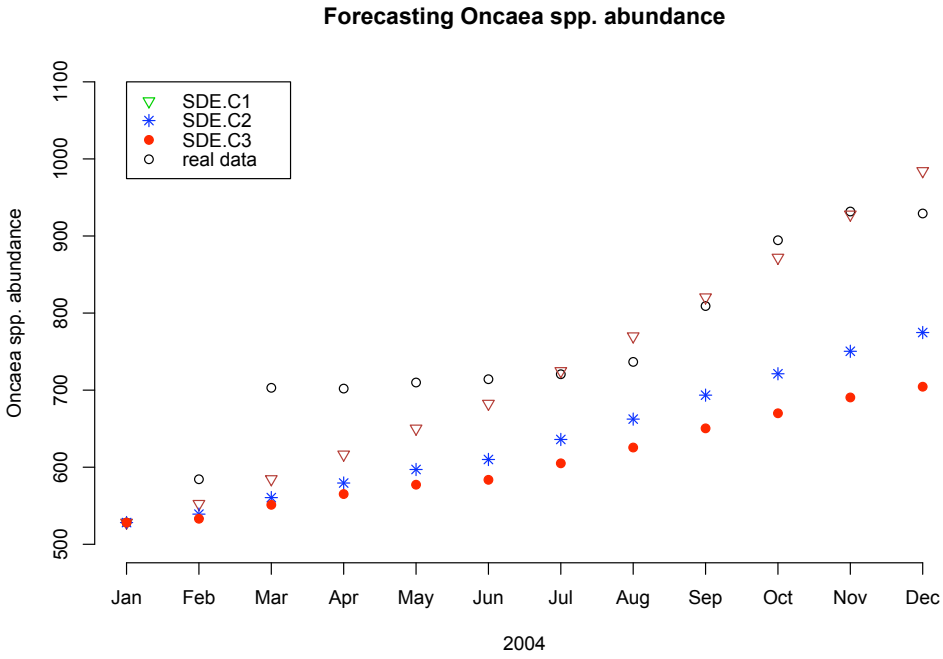

Figure 5: Average simulated values $\bar{X}_3$ for every month on 10000 replications obtained by using SDE.C1, SDE.C2, SDE.C3. The figure also shows the real data regarded in the year.

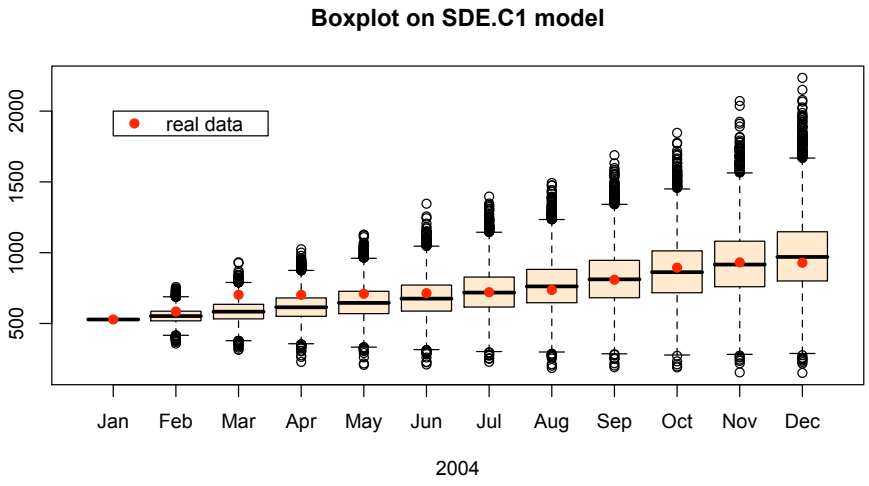

Figure 6: Boxplots for the simulated values derived by SDE.C1 model. The figure also shows the real data.