Towards a data assimilation system for the Cretan Sea ecosystem using a simplified Kalman filter

Hoteit Ibrahim\textsuperscript{a,}\textsuperscript{*}, Triantafyllou George\textsuperscript{b}, Petihakis George\textsuperscript{b}

\textsuperscript{a}Laboratoire de Modélisation et calcul, Tour IRMA BP 53, Grenoble, cedex 9, F-38041, France
\textsuperscript{b}Institute of Marine Biology of Crete, P.O. Box 2214, Iraklion, 71003 Crete, Greece

Received 30 November 2001

Abstract

With the aim of using data assimilation techniques for state estimation in marine ecosystem models, a singular evolutive extended Kalman (SEEK) filter was used to assimilate real in situ data in a water column marine coupled physical-biogeochemical model describing the Cretan sea ecosystem. The biogeochemistry of the ecosystem is described by the European Regional Sea Ecosystem Model (ERSEM), while the Princeton Ocean Model (POM) describes the physical forcing. In the SEEK filter, the error statistics are parameterised by means of a suitable set of empirical orthogonal functions (EOFs). Numerical experiments were conducted to evaluate the performance of this assimilation system. In this context, sensitivity studies to the observations are also presented and discussed.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Ecosystem modelling; Cretan Sea; Data assimilation; SEEK filter

1. Introduction

Coastal areas are highly dependent on the health of the environment since most economic activities are tightly coupled to it. Thus the management and protection of coastal areas is an important issue requiring comprehensive knowledge of the functioning of the system as well as efficient mechanisms for the forecasting and the assessment of possible environmental impacts. Although since the early 20th century mathematical models have been recognized for their power in aiding understanding, explaining and exploring hypotheses, their use was limited to the scientific community. The rapid growth of computers in conjunction with increasing effort in the measurement and monitoring of ecosystem parameters in the last few years has led to the development of the so-called pre-operational models which can be used satisfactorily in both scientific research studies and management policies. However, for the forecast of harmful events and the monitoring and management of sensitive areas fully operational models are required. Such models once adapted, tuned and validated to a particular system, will be able through the assimilation of data to reproduce the system dynamics and forecast the future evolution. In the last few years, the acquisition of field measurements in real time has been made possible through the development of autographic in situ instruments and satellite observa-
tions each complementing the other. Several papers have already resorted to data assimilation techniques for estimating poorly known parameters in the ecosystem models (Prunet et al., 1996; Hurtt and Armstrong, 1996). Recently, focus also turned on the use of data assimilation for state estimation of ecosystem models, using either variational methods (Natvick and Evensen, 2001) or Kalman filtering methods (Eknes and Evensen, 2002; Camillet et al., 2001).

During the Mediterranean Forecasting System Pilot Project (1998–2001) a buoy deployed in an offshore station in the Cretan Sea intensively sampled a number of parameters delivering a valuable data set. A singular evolutive extended Kalman (SEEK) filter, which is a simplified Kalman filter recently introduced by Pham et al. (1997), was used in this study to assimilate available data from the buoy in an existing complex 1D ecosystem model adapted to the Cretan Sea, describing satisfactorily the ecosystem dynamics. In the following sections, the model and the available data are briefly described in Sections 2 and 3. The SEEK filter is presented in Section 4. The validation of the assimilation scheme with pseudo and real observations is presented in Section 5. We conclude by discussing some of the implications of our work including extensions to more realistic situations.

2. Model description and configuration

The model used in this study is a generic 1D ecosystem model originally designed for the North Sea (Baretta et al., 1995) and later adapted to the Cretan Sea ecosystem (Triantafyllou et al., 2001, 2003). It consists of modules describing the biological and chemical processes in the water column, which may be stratified or mixed. It describes the significant biogeochemical processes affecting the flow of carbon, nitrogen, phosphorus, and silica. Each module consists of a coupled set of ordinary differential equations, which may be solved by a straightforward explicit method (Euler integration) or by an implicit higher order Runge–Kutta method. The ecosystem is subdivided into three functional types, producers (phytoplankton), decomposers (bacteria) and consumers (zooplankton). State variables have been chosen in order to keep the model relatively simple without omitting any component that exerts a significant influence upon the energy balance of the system. The organisms constituting the food web (Triantafyllou et al., 2002) are organized into functional groups while feeding/grazing relationships in the pelagic subsystem are generally restricted to the next-smallest functional group and to the same functional group. Due to the oligotrophic nature of the system under study and the significance of the microbial loop, a modified bacteria module (Allen et al., 2002) was used.

The vertical diffusion sub-model of the Princeton Ocean Model (POM) (Blumberg and Mellor, 1987) is used to provide the physical forcing to the ecological model while a turbulence closure model (Mellor and Yamada, 1982) determines the vertical temperature, turbulent kinetic energy, and diffusion coefficient profiles generated by a surface heat flux, salinity and wind stress.

The combination of food web with coupled nutrient dynamics allows the model to adjust to spatial and temporal variations and carbon and nutrient availability and to reproduce the different types of ecosystem behaviour. Versions of the model have been implemented in a wide variety of regimes from the coastal eutrophic (Allen, 1997; Allen et al., 1998; Allen et al., 1998; Vichi et al., 1998; Zavatarelli et al., 2000; Triantafyllou et al., 2001) to offshore oligotrophic (Triantafyllou et al., 2003; Petihakis et al., 2002) and closed systems (Petihakis et al., 1998; Triantafyllou et al., 2000).

The application area is located north of Heraklion at the deployment site of the Multi-Sensor Buoy (M3A) north of Heraklion at 35°40N and 25°E at an approximate depth of 1050 m described in detail in Nittis et al. (2003). The discretization of the model is 40 boxes in the vertical with a finer resolution at the euphotic zone in order to simulate fine scale phenomena.

The physical model is forced with real time wind speed and air temperature data for the computation of wind stresses and relaxed to mean monthly sea surface temperature and salinity obtained by the POSEIDON system during the period of the M3A deployment (January 2000–April 2001). The incident short-wave radiation at the sea is calculated from the latitude and modified by the cloud cover data using the methods of Patsch (1995).
3. The data

In this study, high frequency in situ data collected during the MFSPP project at the north of Crete during 30th January 2000 to 20th April 2001 are used (Fig. 1). The parameters measured on a daily basis were temperature, salinity, oxygen, nitrate and chlorophyll concentrations at 40, 65, 90 and 115 m as described in Nittis et al. (2003).

Dissolved oxygen sensors gave reliable data for the first 6 months but an attempt for in situ repair was not successful and thus data are missing after August 2000 (Nittis et al., 2002). The nitrate data were also fragmented since there were problems with the instrument after the 7th July 2000. Chlorophyll-a, dissolved oxygen and nutrients data were post-calibrated using in situ bottle measurements obtained during maintenance trips. As described in Nittis et al. (2002) for the calibration procedure, water samples were collected at the respective depths during each maintenance cruise, once during the recovery of the instruments and once during their deployment 2–3 days later. CTD measurements down to 1000 m were also carried out during each cruise. The calibration coefficients for the M3A sensors were calculated separately for each period between maintenance cruises, using the reference data collected during each re-deployment when all sensors had been cleaned. Possible sources of error in the measurements are the vertical movements of the mooring caused by the increased water currents, which in some cases were significant (115 m on 15/12/2000).

Fig. 1. Map of M3A mooring.
4. The assimilation scheme: the SEEK filter

The method used to assimilate real in situ data in our 1D ecological model is the singular evolutive extended Kalman filter, called SEEK filter, which is a sequential data assimilation scheme derived from the extended Kalman filter. This filter has been already implemented successfully in several ocean general circulation models (Pham et al., 1997; Verron et al., 1998; Brasseur et al., 1999; Hoteit et al., 2002).

The sequential data assimilation consists of the estimation of the state of the system at each observation time, using only observations up to this time. In the linear case, this problem has been solved by the well-known Kalman filter. In the nonlinear case, one often linearizes the model around the current estimated state vector, which yields to the so-called extended Kalman (EK) filter (see for example Ghil and Malanotte-Rizzoli, 1991 and De Mey, 1997 for a review). However, brute-force implementation of the EK filter is not possible in practice: computational requirements are excessive because of the huge number of the state variables while knowledge of the requisite error statistics is lacking.

The SEEK filter is aimed to reduce the prohibitive cost of the EK filter (for meteorological and oceanographical applications). The main idea is to view the system error covariance matrix as singular with a low rank. This leads to a filter in which the errors correction is made only along certain directions parallel to a linear subspace of dimension r. In its general form, the “correction directions” are made to evolve in time according to the dynamics of the model (see Pham et al., 1997 for more details). However, following Brasseur et al. (1999) and Hoteit et al. (2002), these directions were not updated in these numerical experiments in order to avoid expensive calculations. The covariance matrix is then represented by a limited number of three-dimensional multivariate empirical orthogonal functions (EOFs) of an approximation of the system error covariance matrix, describing the dominant modes of the system’s variability and defining in this way the structure of the correction directions.

For the presentation of the algorithm of the SEEK filter, the notation proposed by Ide et al. (1997) was adopted. Thus consider a physical system described by:

\[ X'(t_k) = M(t_k, t_{k-1})X'(t_{k-1}) + \eta_k \]  \hspace{1cm} (4.1)

where \( X'(t_k) \) denotes the vector representing the true state at time \( t \), \( M(s, t) \) is an operator describing the system transition from time \( s \) to time \( t \) and \( \eta \) is the system noise vector. At each time \( t_k \), one observes

\[ Y_k = H_kX'(t_k) + v_k \]  \hspace{1cm} (4.2)

where \( H_k \) is the observational operator and \( v_k \) is the observational noise. The noises \( \eta_k \) and \( v_k \) are assumed to be independent random vectors with mean zero and covariance matrices \( Q_k \) and \( R_k \), respectively.

The SEEK filter proceeds in two stages in the same way as the EK filter, excluding the initialisation stage. To initialise it we resort to an objective analysis, based on the first observation \( Y_0 \); we take as the initial analysis state vector

\[ X^a(t_0) = X(t_0) + LU_0L^TH_0R_0^{-1}(Y_0 - H_0X(t_0)) \]  \hspace{1cm} (4.3)

where \( L \) is an \( n \times r \) matrix containing the \( r \) retained EOFs on its columns,

\[ U_0 = (L^TH_0^TR_0^{-1}H_0L)^{-1} \]  \hspace{1cm} (4.4)

\( X(t_0) \) is an estimation of the model state at the time of the first available observation and \( H_0 \) is the gradient of \( H_0 \) evaluated at \( X(t_0) \). The initial analysis error covariance matrix may then be taken as:

\[ P^a(t_0) = LU_0L^T \]  \hspace{1cm} (4.5)

Note that we have used the first observation for the initialisation. The algorithm actually starts with the next observation.

4.1. Forecast stage

At time \( t_{k-1} \), it is assumed that one has an estimate \( X^a(t_{k-1}) \) of the system state and its corresponding error covariance matrix \( P^a(t_{k-1}) \), in the factorised form:

\[ P^a(t_{k-1}) = LU_{k-1}L^T \]  \hspace{1cm} (4.6)
where the matrix $U_{k-1}$ is of dimension $r \times r$. The model (4.1) is used to forecast the state as:

$$X^f(t_k) = M(t_k, t_{k-1})X^a(t_{k-1}). \quad (4.7)$$

The corresponding forecast error covariance matrix can then be approximated by:

$$P^f(t_k) = LU_{k-1}L^T + Q_k. \quad (4.8)$$

### 4.2. Correction stage

The new observation $Y^a_0$ at time $t_k$ is used to correct the forecast according to:

$$X^a(t_k) = X^f(t_k) + G_k(Y^a_0 - H_kX^f(t_k)) \quad (4.9)$$

where $G_k$ is the gain matrix given by:

$$G_k = LU_kL^TH_k^{-1}R_k^{-1}H_kL \quad (4.10)$$

with $H_k$ the gradient of $H_k$ evaluated at $X^f(t_k)$ and $U_k$ computed from:

$$U_k^{-1} = U_{k-1} + (LT)^{-1}LTQ_kL(LT)^{-1}$$

$$+ LT^TH_k^{-1}R_k^{-1}H_kL \quad (4.11)$$

The corresponding filter error covariance matrix is then equal to:

$$P^a(t_k) = LU_kL^T. \quad (4.12)$$

Since Eqs. (4.8) and (4.12) are not needed in the algorithm, the SEEK filter drastically reduces the computational cost with respect to the EK filter. Basically, it requires only one integration of the numerical model to compute the forecast state.

In order to treat the model and observation errors, the approach described by Pham et al. (1997) was followed. A so-called compensation technique is therefore used to parameterise the model error. The term $(LT)^{-1}LTQ_kL(LT)^{-1}$ expressing this error in the Eq. (4.11) is then taken into account by means of a forgetting factor $\rho$. This equation is rewritten as:

$$U_k^{-1} = \rho U_{k-1}^{-1} + LT^TH_k^{-1}R_k^{-1}H_kL \quad (4.13)$$

Such a factor replaces the contribution of the imperfect model by amplifying the already existing modes of the background error. Concerning the observation errors, it is reasonable to assume it to be a multiple of the identity: $R_k = \sigma^2I_d$. It can then be easily seen that only $U_k/\sigma^2$ enters the computation, as usually $U_o$ is very large with respect to $\sigma^2$. The reader is referred to Pham et al. (1997) for more details.

### 5. Assimilation experiments

Before assimilating real data, a series of twin experiments was conducted in which pseudo-observations were extracted from a previous model run, considered as the “truth”. These experiments are used to examine the impact of the assimilation on non-observed variables, which further enables us to study more rigorously the sensitivity of the filter to the number of retained EOFs and to the observations.

#### 5.1. Construction of the correction directions

Following Pham et al. (1997), the computation of the EOFs is made through a simulation of the model itself. Firstly, the model has been spun up for 30 years to reach a statistically quasi-steady state. Following, another integration of 5 years is carried out to generate a historical sequence $H_S$ of model realizations. A sequence of 600 vectors was retained by storing one state vector, defined by the 115 biogeochemical variables of the ecosystem model in the 40 layers, every 3
days to reduce the calculations since successive states are quite similar. Since the variables of the state vector of the model are not of the same nature, a multivariate EOF analysis was applied. Each state variable has then been normalized by the inverse of the spatial average (over all boxes) of its components variance. The correction directions were then obtained via a multivariate EOF analysis on the sample $H_S$. Fig. 2 plots the number of EOFs and the percentage of variability (or inertia) contained in the sample $H_S$ they explain.

5.2. Twin experiments

In these experiments, the “truth” provided by the model itself is assumed to be known. Therefore, a reference experiment was performed and the reference state $X_t$ was retained to be compared later with

![Graphs showing RMS evolution over days for various parameters](image.png)

Fig. 3. Evolution in time of the RMS of the filter with 20 EOFs and assimilation of chlorophyll data from twin experiments for phosphate, nitrate, silicate, diatoms, nanoplankton and dinoflagellates. The dashed line represents the RMS misfit between reference and the model free run solutions.
the fields produced by the filter. More precisely, a set of 90 state vectors was retained every day (which corresponds to a period of 3 months) during 1st January 2000 to 1st March 2000. The assimilation experiments are performed using pseudo-measurements, which are extracted from the reference states, on a daily basis for nitrate at 40 m, for oxygen at 65 and 115 m and for chlorophyll at 40, 65 and 115 m (according to the available in situ measurements) with a nominal accuracy of 5%. For the initial state of the filter, the model state of a previous run corresponding to 1st April 1999 is chosen as a first guess. Thus the misfit between the first guess and the reference state is quite large initially, expected to decrease during simulation. The relative efficiency of data assimilation for the various twin experiments is measured by computing the relative mean square error (RMS) misfit between the reference and the assimilated solutions (which will be called RMS of the filter). Note that after a series of empirical trials, the forgetting factor was set to 0.9.

5.2.1. Validation of the assimilation system

A first experiment was conducted to show the usefulness of our assimilation system by only assimilating pseudo-observations of chlorophyll at 40, 65 and 115 m. For this experiment, 20 EOFs were retained. Fig. 3 shows the evolution of the RMS of the filter and compares it to the RMS misfit between the reference and the model free run solutions for phosphate, nitrate, silicate, diatoms, nanoplankton and dinoflagellates. It can be seen that the filter performs very well, leading to a substantial diminution of the RMS misfit with regard to the free run simulation. It was further noted that the RMS of the filter was completely stabilized for almost all the ecosystem variables after only a few assimilation steps. Also the reference and the filter solutions for nitrate, phosphate, chl-a concentrations and bacteria biomass at 70 m were plotted (Fig. 4). In all cases, the

1 Without assimilation, starting from the state used to initialize the filter.
Performance of the assimilation scheme was very satisfactory showing chl-a increased values during spring and autumn with subsequent decrease in phosphate and nitrate. Bacteria follow the chl-a distribution with a short time lag indicating a rather tight coupling as expected in such an oligotrophic system. Finally, the RMS of the filter and the free run for the above variables was plotted (Fig. 5) as a function of the layers number after 60 assimilation steps in an attempt to demonstrate the capacity of the filter in dynamically transferring data into the lower layers of the model.

5.2.2. Sensitivity to the number of retained EOFs

Having explored the efficiency of the assimilation system, a second experiment using the same setup as
previously was conducted examining the sensitivity of the filter to the number of retained EOFs, and fixing the number of EOFs to be used. Fig. 6 shows the evolution in time of the RMS misfit with four different values of the number of retained EOFs: 5, 10, 15 and 20 which explain 87.9%, 97.1%, 98.9% and 98.9% of the model variance, respectively. The chosen rank of the initial error covariance matrix has a direct effect on the performance of the filter, and using more than 15 EOFs does not really improve the assimilation performance. This observation was also made by Cane et al. (1996) and Verron et al. (1998). The same general behaviour was observed when other (combination) observations were assimilated. From these experiments, the number of retained EOFs was set to 15 in all subsequent experiments.

Fig. 6. Evolution in time of the RMS of the filter with 5, 10, 15 and 20 EOFs and assimilation of chlorophyll data from twin experiments for phosphate, nitrate, silicate, diatoms, nanoplankton and dinoflagellates.
5.2.3. Sensitivity to the observations

The last experiment using pseudo-observations was conducted to study the sensitivity of the assimilation system to the observations. Four strategies to assimilate available observations (M3A data) were examined: (i) chlorophyll, (ii) nitrate and oxygen, (iii) chlorophyll and nitrate and finally (iv) chlorophyll and oxygen (Fig. 7). It can be seen that the assimilation of only chlorophyll data gives rather satisfactory results. Adding oxygen data to chlorophyll seems to improve the assimilation results, which means that oxygen data provides information not contained in the chlorophyll data. However, the assimilation of nitrate data did not produce any significant improvement with regard to the results obtained from the assimilation of chlorophyll data.

![Graphs showing RMS evolution over time for different data assimilation strategies](image)

**Fig. 7.** Evolution in time of the RMS of the filter with 15 EOFs and assimilation of (i) chlorophyll data, (ii) oxygen and nitrate data, (iii) chlorophyll and oxygen data and (iv) chlorophyll and nitrate data from twin experiments for phosphate, nitrate, silicate, diatoms, nanoplanckton and dinoflagellates.
only. This is rather expected since the oligotrophic ecosystem of the Cretan Sea is mainly phosphate limited and therefore it is the phosphate cycle which is strongly coupled to the dynamics of the biology. In contrast, nitrogen variability is a combined result of both biological and physical processes rendering its assimilation inferior. Unfortunately high frequency phosphate field measurements were not performed during the particular experiment due to inexistence of in situ analyzers able to measure such low concentrations at that time. Finally one can note that the assimilation of oxygen and nitrate data is not satisfactorily. The above suggest that among all available observations, chlorophyll data is the most important, tightly coupled to the ecosystem functioning and thus should be assimilated.

5.3. Assimilation of real observations

Finally real observations of nitrate at the 45 m layer, of oxygen at 65 and 115 m layers and of chlorophyll at 40, 65 and 115 m layers over a period of 65 days (from\(^2\) 5/3/2000 to 8/5/2000) were used. Based on the results obtained from the twin experiments, chlorophyll and oxygen data were assimilated and the filter was validated by comparing the field of nitrate at 40 m estimated by the filter to the observations, as well as to the field obtained from the model free run (without assimilation). For these experiments 15 EOFs were retained and several values of forgetting factor were tested. For the first guess, the filter was initialised by the state vector provided by the model free run on the time of the first available observation (5/3/2000).

Fig. 8 shows the RMS of the filter obtained with three different forgetting factors: 0.25, 0.5 and 0.75 and the RMS between the observations and the model (without assimilation). In all cases, the RMS of the filter is quite better compared to the RMS of the model. The assimilation system is shown to perform remarkably well within the period of 1 to 30 assimilation steps. The error increases afterwards reaching a pick at the 50th assimilation day. This might be attributed to model error, which seems to be particularly significant for this period (Fig. 9). It can also be seen (Fig. 8) that the use of a large forgetting factor improves the assimilation at the first steps where the ecological model behaves relatively well. However, a small forgetting factor significantly improves the assimilation results towards the end of the assimilation period where the model fails to simulate correctly the nitrate at 40 m. These results can be explained by the fact that the error covariance matrix needs to be amplified in order to make the correction of the filter more consequent when the model error is large. Finally, the filter and model free run solutions of nitrate concentrations at 40 m are compared to the observations (Fig. 9). The assimilation improves the model analysis especially towards the end of the assimilation period.

---

\(^2\) Corresponding to the longest period during which observations of these variables are available.
assimilation period where the solution of the filter seems to converge towards the observations. One can also notice how the assimilation scheme drew the model solution towards the observations.

6. Discussion

The imperfection of numerical marine ecosystem models and the relative sparseness of biogeochemical data necessitate the use of data assimilation techniques for realistic estimates of the ecosystem. In this study, a singular evolutive extended Kalman (SEEK) filter was used to assimilate real in situ data in a complex vertically resolved hydrodynamic ecosystem model of the Cretan Sea. In this filter, the concept of order reduction via empirical orthogonal functions (EOFs) analysis was adopted to simplify the time integration of the forecast error.

Numerical experiments were performed following either a twin experiments approach with simulated data to examine the efficiency of the assimilation system on nonobserved variables and to fix the filter parameters, or assimilating real in situ data from the M3A buoy. In the "perfect" context of twin experiments, the effectiveness of the filter is quite clear with regard to all variables of the model. By using such an approach, it was found that about only 15 EOFs are necessary to represent the model variability. Moreover, it was shown that the assimilation of chlorophyll and oxygen data significantly improves the performance of the model with regard to all the other combinations of available observations. Applications with real data over the period of 5/3/2000 to 8/5/2000 revealed a clear improvement in the model behaviour with respect to a reference simulation run of the derived nitrate variable. Several experiments conducted with different values of the forgetting factor showed the usefulness of such factors, which revealed the utmost importance of the model error on the filter performance. However, it is expected that the use of a "more realistic" model error would certainly improve the filter performance.

In order to enhance the performance of the assimilation system, continuation of this study will focus on the improvement of the model behaviour, and the estimation of the model error covariance matrix. The use of evolutive correction directions and therefore of the statistics of the estimation error according to the model dynamics is also expected to improve the performance of the assimilation system. However, the results so far obtained in this study are quite encouraging suggesting the implementation of more realistic applications. These preliminary experiments were thus necessary to establish a basic assimilation system before more realistic applications could take place. The case of a three-dimensional complex marine ecosystem model and the use of ocean colour data will be considered in the near future.

Acknowledgements

This work has been supported by Mediterranean Forecasting System Pilot Project MAS3-PL97-1608. The authors would like to thank Mrs. A. Pollani for her substantial help, Professor A. Eleftheriou for his constructive criticism during the preparation of this work, Mrs. M. Eleftheriou for her help in editing this text, and K. Georgiou for software assistance.

References

Cane, M.A., Kaplan, A., Miller, R.N., Tang, B., Hackert, E.C.,


