[Click here to view linked References](#)

1

2

3 1

4

5 2

6

7 3

8 4

9

10 5

11 6

12

13 7

14 8

15 9

16 10

17 11

18 12

19 13

20 14

21 15

22 16

23 17

24 18

25 19

26 20

27 21

28 22

29 23

30 24

31 25

32 26

33 27

34 28

35 29

36 30

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

Efficient Ensemble Forecasting of Marine Ecology with Clustered 1D Models and Statistical Lateral Exchange: Application to the Red Sea

Denis Dreano¹, Kostas Tsiaras², George Triantafyllou², Ibrahim Hoteit^{1,3*}

- ¹ Applied Mathematics and Computational Sciences, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
- ² Hellenic Centre for Marine Research, Institute of Oceanography, Anavyssos, Attica, Greece
- ³ Earth Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

* Corresponding author
Email: ibrahim.hoteit@kaust.edu.sa (IH)

Abstract

Forecasting the state of large marine ecosystems is important for many economic and public health applications. However, advanced three-dimensional (3D) ecosystem models, such the European regional seas ecosystem model (ERSEM), are computationally expensive, especially when implemented within an ensemble data assimilation system requiring several parallel integrations. As an alternative to 3D ecological forecasting systems, we propose to implement a set of regional one-dimensional (1D) water-column ecological models that run at a fraction of the computational cost. The 1D models domains are determined using a Gaussian mixture model (GMM)-based clustering method and satellite chlorophyll-a (Chl-a) data. Regionally averaged Chl-a data is assimilated into the 1D models using the Singular Evolutive Interpolated Kalman filter (SEIK). To laterally exchange information between sub-regions and improve the forecasting skills, we introduce a new correction step to the assimilation scheme, in which we assimilate a statistical forecast of future Chl-a observations based on information from neighbouring regions. We apply this approach to the Red Sea and show that the assimilative 1D ecological models can forecast surface Chl-a concentration with a high accuracy. The statistical assimilation step further improves the forecasting skill by as much as 50%. This general approach of clustering large marine areas and running several interacting 1D ecological models is very flexible. It allows many combinations of clustering, filtering and regression technics to be used and can be applied to build efficient forecasting systems in other large marine ecosystems.

1. Introduction

Marine ecosystem models have important applications to monitor algal blooms that can harm human and marine life (Pettersson and Pozdnyakov 2013) and affect the operation of desalination plants (Richlen et al. 2010). They are also used for fisheries management, by providing biomass forecasts under varying fishing stress and environmental conditions (Latour et al. 2003). Modelling marine ecosystems through deterministic differential equations has a long history, dating back at least to the 1950's (see Fennel and Neumann (2004) for an introduction). In these models, the flows of material and energy through the biota trophic web is represented by a set of differential equations (Baretta et al. 1995; Blackford et al. 2004). Ecosystem models can vary widely in complexity depending on how many physical, chemical and biological processes are represented. They can be as

1
2
3 1 simple as the nutrient-phytoplankton-zooplankton (NPZ) model (Anderson 2005), where the pelagic ecosystem
4
5 2 is represented by only two trophic levels, or as complex as the European regional seas ecosystem model
6
7 3 (ERSEM), which includes more than 80 state variables (Baretta et al. 1995; Triantafyllou et al. 2014). Due to the
8
9 4 large number of variables and processes represented, complex ecological models describe the ecosystem
10
11 dynamics better, but are also harder to interpret than simple ones and require important computational resources
12 5
13
14 6 to integrate them.

15
16 7 Due to their complexity and the scarcity of available data, complex ecological models are difficult to
17
18 8 parameterise and initialise (Anderson 2005), leading to important uncertainties in their forecasts (Edwards et al.
19
20 2015). Data assimilation is an important tool to overcome these limitations and enhance the forecasting skills of
21 9
22 ecological models with available observations (Edwards et al. 2015). Data assimilation techniques are deployed
23 10
24 within operational prediction systems; for example, to assess the impact of human activities on the ecosystem of
25 11
26 the Pagasitikos Gulf (Korres et al. 2012). A similar forecasting system in the Red Sea is currently under
27 12
28 development (Triantafyllou et al. 2014). Most of these systems use the Ensemble Kalman filter (EnKF)
29 13
30 assimilation scheme (Ciavatta et al. 2011; Ciavatta et al. 2014), or the deterministic variants that are more
31
32 14 efficient with small ensemble sizes relative to the number of observations, such as the Singular Evolutive
33
34 15 Interpolated Kalman filter (SEIK) (Pham 2001; Hoteit et al. 2002; Triantafyllou et al. 2003; Korres et al. 2012).
35
36 16 These ensemble-based filters are Monte Carlo implementations of the Kalman filter, suitable for sequential data
37
38 17 assimilation into large scale nonlinear dynamical models. They integrate an ensemble of system states forward in
39
40 18 time with the ecological model. This forecasting step is followed by a Gaussian Kalman correction step of the
41
42 19 ensemble as new observations become available (Hoteit et al. 2015).
43
44 20
45 20
46

47 21 The development of assimilative marine ecosystem system is however often hindered by various limitations.
48
49 22 One of the important issues reside in the dynamical consistency between the dynamical model and the updates
50
51 23 after assimilations. Such inconsistencies can arise at the onset of blooms, that cause large discrepancies between
52
53 24 the data and the model prediction and can result in inconsistent cross-covariance estimations (Gharamti et al.
54
55 25 2017). In addition, the assimilation update may trigger earlier blooming of some of the ecosystem state variables,
56
57 26 which also affect the estimation of state covariance statistics. The scarcity of biogeochemical observations
58
59 27 available for assimilation, especially for the subsurface may further result in ecologically inconsistent forecasts
60
61
62
63
64
65

1
2
3 1 in poorly observed areas and variables (Edwards et al. 2015). The poor knowledge of the model and
4
5 2 observational error statistics and the strongly nonlinear and intermittent nature of the ecosystem blooms are also
6
7 3 important limiting factors in the assimilation performance (Triantafyllou et al 2007; Hoteit et al. 2005)”.
8

9
10 4 Assimilative three-dimensional (3D) marine ecosystem models are also limited by the large computational
11
12 5 resources required for the implementation of state-of-the-art ensemble-based data assimilation techniques, which
13
14 6 are necessary to achieve sufficient forecasting skills. In this study, we propose a new approach to efficiently
15
16 7 forecast the ecology of large marine ecosystems by running several one-dimensional (1D) regional models
17
18 8 water-column in parallel, which run at a fraction of the computational cost of integrating a 3D model. Using
19
20
21 9 remotely sensed chlorophyll-a (Chl-a) data, we first apply a clustering method to divide a given ocean domain
22
23 10 into smaller eco-regions. The ecology of each of these regions is then simulated by a 1D ecological model,
24
25 11 constrained via assimilation of Chl-a data. To represent the lateral interactions between eco-regions, such as the
26
27 12 horizontal advection of nutrients and biomass, we also assimilate statistical forecasts of the future Chl-a
28
29
30 13 observations based on current observations in neighbouring regions. This novel approach aims at improving the
31
32 14 forecast of the regional 1D models by exploiting available information in adjacent regions. We demonstrate the
33
34 15 efficiency of the proposed framework to forecast the ecosystem of the Red Sea from 2003 to 2004.
35

36 16 The article is organized as follows. Section 2 presents the Red Sea ecosystem, the Chl-a data and the
37
38
39 17 clustering approach that is applied to divide the basin into smaller eco-regions. Section 3, describes the 1D
40
41 18 ERSEM models that are implemented to simulate the ecology of each of the clustered eco-regions, the SEIK
42
43 19 ensemble assimilation scheme, and the proposed approach to improve the forecast of the regional 1D models by
44
45 20 exchanging information between neighbouring clusters through the assimilation of statistical predictions. Section
46
47 21 4 outlines the experimental setup and presents and discusses the results of assimilation experiments. Section 5
48
49
50 22 concludes the work with a summary and a discussion of the results and perspectives for future works.
51

52 23 53 54 24 **2. Study Region and Data**

55 56 57 25 *2.1. Study Region*

58
59 26 We demonstrate the principle of clustering a large marine ecosystem into smaller eco-regions and exchanging
60
61 27 information between 1D ecological models in the Red Sea. The Red Sea is an ideal test region since its primary
62

productivity is can be strongly affected by horizontal advection of water masses, which are not accounted for by 1D models. This elongated basin (~2250 km length, 200-300 km width) is situated between Africa and the Arabic peninsula. In general, Red Sea waters are considered to be deficient in primary nutrients (nitrate, ammonium, phosphate, and silicate) and their concentrations show an increase from North to South (Weikert 1987). The only significant exchange of water takes place in the Gulf of Aden through the strait of Bab-el-Mandeb, at the basin southern boundary (Sofianos and Johns 2007). The northward intrusion of fresher, cooler and nutrient-rich water from the Gulf of Aden determines a marked north-south gradient of biophysical parameters along its axis (Fig 1) (Kürten et al. 2014; Nanninga et al. 2014). In winter, the monsoonal winds blow northward in the southern half of the Red Sea (Yao et al. 2014a), enhancing the intrusion of nutrient-rich Gulf of Aden surface water into the Red Sea and triggering intense phytoplankton blooms (Raitzos et al. 2013). Accordingly, the winter phytoplankton bloom timing follows a clear south to north gradient with distinct time lags: the bloom in the southern area starts in January, followed one month later by the bloom in the central area, and later, during March, in the northern area (Triantafyllou et al. 2014).

2.2. Data

We acquired 8-day level 3 remotely sensed surface Chl-a measurements at 4 km resolution from the OC-CCI dataset (www.esa-oceancolour-cci.org) over the Red Sea (period 1997-2012). This dataset is produced and validated by the European Space Agency and provides the most complete and consistent time series of multi-sensor global Chl-a data by combining datasets from the Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua, the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS), and the Medium Resolution Imaging Spectrometer (MERIS). The merging of data from these sensors produces high-resolution Chl-a data with a substantially improved coverage compared to single-sensor datasets, particularly in the southern half of the Red Sea region where the presence of clouds and haze during summer has resulted in very few observations (Racault et al. 2015; Brewin et al. 2015). For instance, the merged CCI product has average percentage data coverage of ~64% in the southern Red Sea for July (2002-2012) compared to ~5% for single-sensor MODIS (Dreano et al. 2016).

Since the clustering (described in Section 2.3) requires complete datasets, we use the DINEOF (Data Interpolating Empirical Orthogonal Function) method (Beckers and Rixen 2003; Alvera-Azcárate et al. 2009) to

fill in the missing values in the OC-CCI Chl-a dataset for the period 1997-2012. DINEOF is a parameter-free method that has been used to reconstruct incomplete chlorophyll datasets in many regions of the ocean such as the South Atlantic bight (Miles and He 2010) and the Red Sea (Dreano et al. 2015). It proceeds by initially filling in missing values in the data matrix with arbitrary values, and then recursively applying a singular value decomposition (SVD) followed by a truncated reconstruction until the data matrix converges. Further details on the method and its implementation in the Red Sea can be found in (Dreano et al. 2015).

2.3. Regional Clustering

After filling in the missing values of the OC-CCI dataset using DINEOF, we apply a clustering technique to divide the Red Sea into eco-regions. Since the distribution of Chl-a data tends to be log-normally distributed (Brewin et al. 2013), we use a Gaussian Mixture Model (GMM) clustering approach (Fraley and Raftery 2002). In this approach, the time series $\mathbf{y}_i$ of weekly observations of Chl-a log-concentrations at any given location (i) are assumed to be samples of a mixture of multivariate Gaussian distributions of the form:

$$p(\mathbf{y}_i) = \sum_{k=1}^K \tau_k f(\mathbf{y}_i | \boldsymbol{\mu}_k, \Sigma_k), \quad (1)$$

where τ_k is the *a priori* probability of a pixel location (i) to belong to the cluster, or component, of the mixture k , and $f(\cdot | \boldsymbol{\mu}_k, \Sigma_k)$ is a multivariate normal distribution of mean $\boldsymbol{\mu}_k$, and covariance matrix Σ_k . The parameters of the mixture model (τ_k , $\boldsymbol{\mu}_k$ and Σ_k) are then estimated using the expectation-maximisation (EM) algorithm (Fraley and Raftery 2002), and each pixel location is attributed a cluster k based on a maximum likelihood criterion. EM is an iterative algorithm that approximates the maximum likelihood estimator of distribution parameters when the data is incomplete, which is the case here where we do not know the cluster to which a sample belongs. It has been shown to converge under mild regularity conditions (Dempster et al. 1977; Fraley and Raftery 2002).

Since the gradient of Chl-a log-concentrations along the latitudinal axis of the Red Sea varies widely, being steep in the south and smoother in the north, a direct application of the GMM clustering method results in one large cluster for the northern and central Red Sea and many smaller clusters in the southern Red Sea. To divide the Red Sea into regions of similar sizes, we employ a dichotomic approach where we use GMM to cluster the Red Sea into 2 regions, then cluster the largest of the so far obtained regions into two other clusters, and so forth,

1
2
3 1 recursively, until we obtain 4 provinces. Using this approach, we cluster the Red Sea into four eco-regions (Fig
4
5 2 1b), which are very consistent with the ones suggested by (Raitsos et al. 2013) on the basis of physical and
6
7 3 biological considerations (Fig 1a).

9 4 The 1D ecological models are implemented based on the resulting regions. The northern Red Sea (NRS) is
10
11 significantly influenced by its predominant cyclonic circulation and is colder than the other regions (Raitsos et
12 5
13 al. 2013). The NRS Chl-a concentrations are slightly higher, as compared to the more oligotrophic northern-
14 6
15 central Red Sea, and are characterised by distinct winter blooms that are triggered by nutrients brought to the
16 7
17 euphotic zone through deep mixing (Raitsos et al. 2013; Triantafyllou et al. 2014). During winter, the NRS
18 8
19 appears to export nutrients to the NCRS following the overturning circulation, while in the summer the situation
20
21 9 is more or less balanced (Triantafyllou et al. 2014). The NCRS is the most oligotrophic region of the Red Sea,
22
23 10 while the southern-central Red Sea (SCRS) is the second most productive behind the southern Red Sea (SRS)
24
25 11 (Raitsos et al. 2013). Both the NCRS and the SCRS exhibit more pronounced inter-annual variability in the
26
27 12 strength of the winter blooms, which are driven by the intrusion of surface water from the Gulf of Aden (Raitsos
28
29 13 et al. 2013). The southern Red Sea is the most productive province due to its proximity with the strait of Bal-el-
30
31 Mandeb (Raitsos et al. 2013), from which it receives significant quantities of dissolved inorganic nutrients and
32 14
33 organic carbon (Triantafyllou et al. 2014). The SRS is characterised by a winter bloom driven by surface
34 15
35 intrusion and a summer bloom driven by a subsurface intrusion of intermediate water from the Gulf of Aden
36 16
37 (Dreano et al. 2016).

38 17
39
40 In this study, we demonstrate the efficiency of the proposed approach of assimilating statistical predictions of
41 18
42 neighbour regions to improve the forecast of 1D ecological models in the NRS, NCRS and SCRS. The SRS is
43 19
44 not investigated in the present study because of the relatively high rate of missing data over this region (Racault
45 20
46 et al. 2015). Chl-a data are averaged over the NRS, NCRS and SCRS to obtain 8-day time series, which are
47 21
48 directly assimilated into the 1D ecological models (Section 3.2), and also used to fit the linear models that are
49
50 22 used to exchange information between the three considered clusters (Section 3.3).
51
52 23
53
54 24
55
56 25
57
58

59 26 **3. Methods**

60 27 *3.1. 1D ERSEM Models*

Three coupled 1D water-column biogeochemical models were developed to describe each region ecosystem dynamics (NRS, NCRS, SCRS). The biogeochemical model is based on the biogeochemical ERSEM model (Baretta et al. 1995), a comprehensive and generic model that has been applied in a variety of different marine ecosystems, such as the Mediterranean (Zavatarelli et al. 2000; Petihakis et al. 2002; Petihakis et al. 2009; Tsias et al. 2014), the North Sea (Patsch and Radach 1997), and the Arabian Sea (Blackford and Burkill 2002). The state variables have been chosen, to keep the model relatively simple without omitting any component that exerts a significant influence on the energy balance of the system. The dynamics of biological functional groups are described in the model, taking into account the most important population (growth, migration, and mortality) and physiological (respiration, grazing, ingestion, lysis, excretion and egestion) processes.

Three broad functional types are used to classify the ecosystem's biota: producers, decomposers, and consumers, with a further subdivision based on their size class or feeding method. Producers include four phytoplankton groups (diatoms, dinoflagellates, nano- and pico-phytoplankton), decomposers consist of bacteria, while consumers are represented by three zooplankton groups (heterotrophic nanoflagellates, microzooplankton, mesozooplankton). Bacteria use dissolved organic matter and are responsible for the degradation of particulate organic matter. They also compete for dissolved inorganic nutrients with phytoplankton. Heterotrophic nanoflagellates prey on bacteria and picophytoplankton and are grazed by microzooplankton that also consumes diatoms and nanophytoplankton. Mesozooplankton preys also on heterotrophic nanoflagellates and large phytoplankton (diatoms, dinoflagellates).

The bio-carbon dynamics are loosely coupled with the chemical dynamics of nitrogen, phosphorus, silicate, and oxygen, through dynamically varying C:N:P:S elemental ratios of the different functional groups. This allows the model to be adjusted to spatial and temporal variations in carbon and nutrient availability and to reproduce the different types of ecosystem behaviours. Here we focus on the open sea water column of each region and therefore use a simple benthic returns model instead of the standard ERSEM dynamical benthic model (Baretta-Bekker et al. 1997).

The initial conditions for the biogeochemical variables in the three different water columns were obtained from a 3D Red Sea coupled model simulation (Triantafyllou et al. 2014). The hydrodynamic properties of the water column (temperature, vertical diffusivity) in the three areas were obtained on a daily basis from a 3D MIT-

GCM simulation output (Yao et al. 2014a; Yao et al. 2014b) at specified points (NRS: 35.0°E 26.6°N, NCRS: 36.9°E 23.5°N, SCRS: 38.8°E 20.0°N) over 2003–2004 and were used to drive the 1D biogeochemical models.

3.2 Ensemble Assimilation: SEIK

The Singular Evolutive Interpolated Kalman (SEIK) filter is an ensemble square-root implementation of the Kalman filter (KF). In the forecast step, it propagates the KF estimate and its assumed low-rank (l) error covariance matrix by integrating an ensemble of $N = l + 1$ state vectors, called interpolated states or ensemble members, with the dynamical (ecosystem) model forward in time. This ensemble is randomly sampled so that its sample mean and covariance exactly match those of the KF. The forecast state and its error covariance matrix, which are taken as the mean and covariance of the forecast ensemble, are then updated with a KF correction step every time a new observation is available. The successive sampling, forecast and analysis steps of SEIK are summarized below. The reader is referred to Pham (2001) and Hoteit et al. (2002) for a full description of the SEIK algorithm.

(i) Sampling step

Starting from an available analysis state $x^a(t_k)$ and a low rank (l) error covariance $P^a(t_k) = L_k U_k L_k^T$ at a given time t_k , an ensemble of N state vectors $x_1^a(t_k), \dots, x_N^a(t_k)$ is randomly sampled after every analysis step such that:

$$x^a(t_k) = \frac{1}{N} \sum_{i=1}^N x_i^a(t_k), \quad (1)$$

$$P^a(t_k) = \frac{1}{N} \sum_{i=1}^N [x_i^a(t_k) - x^a(t_k)][x_i^a(t_k) - x^a(t_k)]^T. \quad (2)$$

U_k and L_k are $l \times l$ and $n \times l$ matrices, where n is the system state dimension. To generate the ensemble members $x_i^a(t_k)$, we use the second-order exact sampling technique (Pham, 2001; Hoteit et al., 2002);

$$x_i^a(t_k) = x^a(t_k) + \sqrt{l+1} L_k (\Omega_{k,i} C_k^{-1})^T, \quad (3)$$

where C is the square root matrix of U_k and $\Omega_{k,i}$ denotes the i th row of a randomly generated matrix W_k , with columns orthonormal and orthogonal to the vector $[1 \cdots 1]^T$.

(ii) Forecast step

The ensemble members are propagated forward with the dynamical model M to the time of the next available observation to compute the forecast ensemble as $x_i^f(t_{k+1}) = M(t_{k+1}, t_k) x_i^a(t_k)$. The forecast state $x^f(t_{k+1})$ and its error covariance matrix $P^f(t_{k+1})$ are then estimated as the sample mean and covariance of the $x_i^f(t_{k+1})$, respectively. One can then decompose $P^f(t_{k+1})$ as:

$$P^f(t_{k+1}) = L_{k+1} \hat{C} N T T^T \hat{C}^{-1} L_{k+1}^T, \quad (4)$$

with:

$$L_k = [x_1^f(t_k), \dots, x_{l+1}^f(t_k)] \cdot T, \quad (5)$$

and T a $(l+1) \times l$ matrix with zero column sums (Hoteit et al., 2002).

(iii) Analysis step

Once a new observation y_{k+1} becomes available, the analysis state and its error covariance matrix are computed as:

$$x^a(t_{k+1}) = x^f(t_{k+1}) + K_{k+1} [y_{k+1} - H_{k+1} (x^f(t_{k+1}))], \quad (6)$$

$$P^a(t_{k+1}) = L_{k+1} U_{k+1} L_{k+1}^T, \quad (7)$$

where $K_{k+1} = L_{k+1} U_{k+1} (HL)_{k+1}^T R_{k+1}^{-1}$ is the so-called Kalman gain, H_{k+1} is the observational operator, which computes the observation prediction from the forecast state. R_{k+1} is the observational error covariance matrix and U_{k+1} is computed from:

$$U_{k+1}^{-1} = r \frac{1}{N} (T^T T)^{-1} + (HL)_{k+1}^T R_{k+1}^{-1} (HL)_{k+1}. \quad (8)$$

r is a forgetting factor, that takes values between 0 and 1. It is used to inflate the forecast error covariance by $1/r$ to account for various sources of uncertainties in the system and the filter (Hoteit, Pham and Blum, 2004).

3.3. Assimilation of Statistical Forecasts

We propose to improve the forecast of the SEIK assimilation scheme presented above by introducing a “statistical” assimilation step right after the usual dynamical forecast step (and before the analysis step with the data of that same region), as depicted in Figure 2. During this step, we assimilate a statistical prediction $\tilde{y}_{k+1}$ of the future Chl-a observation y_{k+1} calculated based on current data from neighbouring regions. In a given region, the forecasted observation is computed using a linear model of the form $\tilde{y}_{k+1} = \sum_j a_j z_k^{(j)} + b$, where b and a_j are the coefficients of the linear model (fitted by minimising the least-square error) and $z_k^{(j)}$ are the Chl-a satellite observations in the neighbouring region j at time step k . The coefficients of the linear model are fitted offline by minimisation of the least-square error on the Chl-a data for the period 1997 to 2002. The prediction $\tilde{y}_{k+1}$ thus incorporates the influence of neighbouring regions on the future Chl-a concentrations in the region under consideration. Specifically, the regression models can represent phenomena such as the horizontal advection of nutrients and plankton and the covarying seasonal patterns of chlorophyll concentrations.

We use the same SEIK filtering scheme to assimilate the prediction of the observation $\tilde{y}_{k+1}$ in order to produce an improved forecast $\tilde{x}^f(t_{k+1})$ of the state. $\tilde{x}^f(t_{k+1})$ thus incorporates information from neighbouring regions at time t_k . At the following time step, the real observation y_{k+1} becomes available and is assimilated in turn using the SEIK filter as described above. Here, the sequential assimilation of $\tilde{y}_{k+1}$ and then y_{k+1} is not problematic as the statistical forecasted observation is computed using independent observations from other provinces.

After fitting the statistical model, we run SEIK scheme with assimilation of statistical forecasts in each of the three regions (NRS, NCRS, SCRS) over the period 2003-2004. We calculate the root-mean-square error (RMSE) between the (dynamical and enhanced with statistical predictions) forecasts of Chl-a and the actual observations y_k over 2004. The covariance $\tilde{R}$ of the statistical observation error is chosen such that $\tilde{R} \geq R$ (as we expect the statistical model to have add larger errors than the observations it is trying to predict) and $\tilde{R}$ minimises the RMSE between improved forecast and satellite observations over 2004.

4. Numerical Experiments and Results

4.1. Experimental Setup

We will have three experimental setups per region (NRS, NCRS and SCRS): the free model run, the run with the standard SEIK assimilation of satellite Chl-a data, and the run with SEIK assimilation of satellite and statistically predicted Chl-a data. A free model run of the 1D ecological models in each region, initialised as described in Section in 3.1, is performed from 2003 to 2004, with 2003 considered as a spin up period. The root mean-square error (RMSE) between the forecasted Chl-a levels and the satellite observations are then computed for each region over 2004.

The standard SEIK assimilation scheme (SEIK_{STD}, Section 3.2) assimilates remotely sensed Chl-a data averaged over each of the three regions (Section 2.1) every 8 days. After verifying that the results of the filtering were not very sensitive to the size of the ensemble, we set the number of ensemble members to 25 in all our experiments. Similarly, we set the inflation factor to $\rho = 0.2$. The RMSE between satellite observations and the model forecast is computed over 2004.

The SEIK assimilation scheme with assimilation of statistical information (SEIK_{STAT}, Section 3.3) is configured similarly as SEIK_{STD} with 25 ensemble members and an inflation factor of $\rho = 0.2$ at each assimilation (of statistical and satellite observations). The statistical predictions are assimilated every 8 days right after the forecast of the 1D ecological model (Correction step, Fig 2). The RMSE is computed over 2004 both for the 1D ecological model forecast (Dynamical Forecast step, Fig 2) and for the forecast corrected with statistical observations (Corrected step, Fig 2).

4.2. Northern Red Sea (NRS)

Without assimilation, the 1D model reproduces relatively well the Chl-a seasonal succession in the NRS as observed from satellite (Fig 3b), with very low Chl-a concentrations in summer and a sharp increase during winter. However, the Chl-a concentrations obtained from the model free-run exhibit substantial errors against the observed satellites data, with an average RMSE of 0.18 mg / m³ over 2014, compared with a mean Chl-a concentration of 0.17 mg / m³. During the winter bloom, the 1D ecological model tends to predict Chl-a levels twice as high (~ 0.5 mg / m³) as compared to the values typically observed by remote sensing (~ 0.25 mg / m³). Assimilating the Chl-a satellite measurements with the SEIK_{STD} filter (without incorporating information from the neighbour subdomains) leads to a substantial improvement of the model behaviour, decreasing the model forecast error by as much as $\sim 60\%$. We further test different values of the observational error variance R to

assess the sensitivity of the assimilation results to R , and eventually, tune its value (minimise the RMSE). Table 1 outlines the resulting average RMSEs over 2014, suggesting that the assimilation results are not very sensitive to the value of R , with $R = 1/200$ (standard deviations of $0.071 \text{ mg} / \text{m}^3$) providing the best results with respect to RMSE.

The only region neighbouring the NRS is the NCRS (Fig 1). We therefore fit a linear model predicting the NRS Chl-a log-concentration (at time $t + 1$) as a function of the NCRS Chl-a log-concentration (a time t , *i.e.* one 8-day period beforehand) as described in Section 3.3. The model is fitted using Chl-a satellite data over the 1997-2002 period ($r^2 = 0.62$, $p < 0.001$, Fig 3a). The linear regression model is then used to predict the NRS Chl-a concentrations over the 2003-2004 period (Statistical Forecast step, Fig 2).

These statistical predictions of Chl-a are considered as observations and assimilated with SEIK_{STAT} (Correction step, Fig 2) as described in Section 3.3 to complement the satellite Chl-a with information from the neighbour subdomain for improving the forecast. We also test several values for the observational error variance $\tilde{R}$ of the statistical observations (Table 2) and choose the one that minimises the RMSE of the improved forecast (Forecast 2). The best value is again obtained with $\tilde{R} = 1/200$, leading to an RMSE of $0.038 \text{ mg} / \text{m}^3$. This represents an important 46% error reduction compared to the RMSE obtained with SEIK without the assimilation of statistical information ($0.070 \text{ mg} / \text{m}^3$). Additionally, the forecast error before the assimilation of statistical observations (Forecast 1) is rather stable (RMSE of $0.075 \text{ mg} / \text{m}^3$), indicating that the assimilation of the statistical information is consistent the dynamics of the 1D ecological model.

Figure 3b shows that the model forecasts with assimilation were substantially improved over the winter bloom period compared to the Chl-a satellite data. In contrast to the free model run, the 1D ecological model forecasts of Chl-a levels with assimilation are quite close to the satellite observations. However, strong variations in the forecast error can still be noticed during the winter bloom (Fig 3c). The main contribution of the statistical correction of the forecast by the neighbour information (Correction step, Fig 2) is the reduction of these variations, leading to a large (46%) improvement of the RMSE. During the oligotrophic summer period, the improved forecast (Forecast 2) with the assimilation of statistical predictions (Correction step, Fig 2) is comparable to the forecast from the 1D ecological model (Forecast 1, Dynamical Forecast step, Fig 2). The larger improvement of the statistical assimilation during winter can be linked to the enhanced horizontal

advection of nutrient-rich water masses during this period (see Section 2.3) and shows that the principle of assimilating statistical observations represents this interaction between the NRS and NCRS regions successfully.

4.3. Northern-Central Red Sea (NCRS)

The free-run of the 1D ecological model in the northern-central Red Sea matches well the cycle of remotely-sensed Chl-a data, with the succession of low-concentrations in summer and high-concentrations during the winter blooming period (Fig 4b). Overall, the RMSE with respect to satellite observations is $0.096 \text{ mg} / \text{m}^3$ over 2004, which is not excessive compared with the average Chl-a concentration over the same period ($0.156 \text{ mg} / \text{m}^3$). The 1D ecological model systematically underestimates the Chl-a concentration in summer, while it overestimates it during the winter bloom period. SEIK_{STD} (without assimilation of statistical observations) is first used to enhance the model forecasting skill by assimilating Chl-a satellite observations. We test the filter performances with different values of R (Table 3) and again find that the system is not very sensitive to the choice of this parameter. We choose the value $1/R = 100$ (standard deviation of $0.10 \text{ mg} / \text{m}^3$), which reduces the RMSE of the forecast over 2004 by half ($0.049 \text{ mg} / \text{m}^3$, Table 3).

The NCRS region is surrounded by the NRS and the SCRS (Fig 1). We therefore fit a bivariate regression model that predicts the Chl-a log-concentration (at time $t + 1$) in the NCRS, with the NRS and the SCRS Chl-a log-concentrations as predictors (at time t , *i.e.* one 8-day period beforehand). We fit the model's coefficients as explained in Section 3.3 using Chl-a satellite data over the 1997-2002 period ($r^2 = 0.61$ and $p < 0.001$ for both predictors, Fig 4a). We then use this multivariate regression model to predict the NCRS Chl-a concentration over the 2003-2004 period (Statistical Forecast step, Fig 2).

In $\text{SEIK}_{\text{STAT}}$ (Section 3.3), the statistical predictions are assimilated as observations in an analysis step (Correction step, Fig 2), before the assimilation of satellite Chl-a data. Several values of the error variance $\tilde{R}$ of this statistical observation are tested (Table 4), according to which we choose $1 / \tilde{R} = 100$ (standard deviation of $0.10 \text{ mg} / \text{m}^3$), which is the value that minimises the RMSE of the corrected forecast (Forecast 2) over 2004. With this value of $\tilde{R}$ the assimilation of statistical observations reduces the RMSE to $0.024 \text{ mg} / \text{m}^3$ (Correction step, Fig 2), which corresponds to a reduction of 51% compared to the forecast of Chl-a in the SEIK_{STD} assimilation scheme. As for the NRS, the forecast error from the 1D ecological model (Dynamical Forecast step,

Fig 2) is stable ($0.054 \text{ mg} / \text{m}^3$ compared to $0.049 \text{ mg} / \text{m}^3$ for the forecast in SEIK_{STD}) showing that the 1D ecological model is not perturbed by the Correction step (Fig 2) in $\text{SEIK}_{\text{STAT}}$.

Figure 4b shows the Chl-a satellite data and the forecasts from the model. The assimilation improves the fit of the 1D ecological model forecast to the Chl-a satellite data and corrects the biases in winter and summer. As in the NRS, the forecast errors are more important during the winter bloom (Fig 4c), when the variability of Chl-a concentration is the highest. The assimilation of statistical predictions (Forecast 2, Correction step, Fig 2) reduces the errors over the whole period compared to the dynamical forecast (Forecast 1) and is very effective in reducing large errors during the winter blooming period. The large improvement in the forecast with assimilation of statistical information (55%) can be linked to the efficient representation of lateral interactions between water masses in the NCRS and its neighbours (NRS, SCRS).

4.4. Southern-Central Red Sea (SCRC)

After approximately 4 months of simulations, the Chl-a levels predicted by a free-run of the 1D ecological model converge toward values close to the satellite observations and then depict relatively accurately their seasonal variability (Fig 5b). The free model run exhibits moderate errors on the predicted Chl-a concentration (RMSE of $0.21 \text{ mg} / \text{m}^3$ over 2004) with respect to the observed average Chl-a concentration ($0.32 \text{ mg} / \text{m}^3$). In general, we observe that the 1D ecological model tends to underestimate the level of Chl-a and to have a noticeable lag in the prediction of the start of the winter bloom. Assimilating Chl-a satellite observations with SEIK_{STD} reduces the forecast RMSE by approximately half. The reduced forecast RMSE is outlined in Table 5 for different values of the observation error covariance R . The best filtering performances were obtained with $1 / R = 200$ (standard deviation of $0.071 \text{ mg} / \text{m}^3$), which reduces the RMSE to $0.105 \text{ mg} / \text{m}^3$.

The SCRS is bordered by the NCRS in the north and the SRS in the south (Fig 1). Since we excluded the data from the SRS (see Section 2.2.), we fit a linear regression model to predict the SCRS Chl-a log-concentration (at time $t + 1$) from the NCRS Chl-a log-concentration (at time t , *i.e.* one 8-day period beforehand). We use remotely-sensed Chl-a data over the 1997-2002 period to fit the model. The model fit in this case is less good than with the other regions ($r^2 = 0.28$, $p < 0.001$). We then use the fitted linear model to predict the NRS Chl-a concentration over the 2003-2004 period (Statistical Forecast step, Fig 2).

As above, we assimilate these statistical observations to improve the forecast (Correction step, Fig 2). We choose the error variance $\tilde{R}$ of the statistical observation by testing several values and selecting the one that minimises the RMSE of the corrected forecast (Forecast 2) over 2004 (Table 6). The best values resulted from $\tilde{R} = 1/50$ (standard deviation of 0.14 mg / m³), which leads to an RMSE of 0.078, corresponding to a reduction of 22% of the error compared with the forecast obtained with SEIK without assimilation of statistical observations (SEIK_{STD}). As noted in the other regions, the RMSE of the forecast error before assimilation of statistical information (Forecast 1, 0.100 mg / m³) is stable compared with the forecast of SEIK_{STD} (0.105 mg / m³), showing that the statistical forecast is consistent with the dynamics of the 1D ecological model.

Overall, the assimilation improves the fit of the 1D ecological model to the Chl-a data and removes the bias towards higher Chl-a concentrations. In contrast to the NRS and the NCRS, the forecast errors are better spread over the whole period and do not seem to exhibit seasonal behaviours. The statistical correction of the forecast (Forecast 2, Correction step, Fig 2) very consistently improves the forecast results over the 1D ecological model forecast (Forecast 1, Dynamical Forecast step, Fig 2). In the SCRS, the improvement of the forecast using the statistical assimilation (22%) is less than that in the NRS (46%) and the NCRS (51%). This can be attributed to the lesser quality of the linear regression model used to predict the future observation ($r^2 = 0.28$, compared to $r^2 \sim 0.60$ in the other regions). We can expect that the most important lateral interaction for the SCRS takes place with the SRS, through which large quantities of nutrient from the Indian are advected (Section 2.3). However, this interaction is not represented in the current statistical model as we excluded data from the SRS.

4.5. Assessing Impact of Assimilation on Nutrients

A key challenge when assimilating satellite Chl-a into marine ecosystem models is to preserve the dynamical consistency of the non-observed state variables after the filter's update; particularly the dissolved inorganic nutrients which are the main drivers of primary production in the oligotrophic Red Sea environment. We examined the impact of assimilation on the dissolved inorganic nutrients variables nitrates (NO₃) and phosphates (PO₄) and compared the model simulated PO₄ and NO₃ to the mean annual profiles from the World Ocean Atlas (WOA) (Garcia et al. 2014), which were used to initialise the 1D ecological models. As shown in Fig 6, in most cases the assimilation of satellite surface Chl-a data has a relatively weak impact on the simulated PO₄ and NO₃. The slight deviation of deep water nutrients from WOA after assimilation of surface Chl-a in NRS and

NCRS is to a point expected, given the limitations of the 1-D model configuration not allowing for a feedback from horizontal processes that would for example counterbalance the increase in the NCRS with the decrease in NRS, following the thermohaline circulation. In the SCRS, the underestimation of nutrients by the free run in the subsurface layer (30-100m) is partially corrected in the assimilation run. In general, the simulated annual profiles remain close to the WOA profiles, both in the free and assimilation runs. Overall, and combined with the demonstrated improved forecasting skills, our analysis suggest that the data assimilation of the statistical forecasts does not result in any distortion of the dynamics of these (non-observed) variables.

The impact of assimilation on the different phytoplankton functional types relative ratio has been further checked and was found to be rather weak (e.g in the NRS diatoms+dinoflagellates/total phyto changed from 16% to 18% in the assimilation). This is expected given the correlation between Chl-a and phytoplankton groups.

5. Summary and Future Perspectives

Forecasting the state of marine ecosystems using ecological models has promising applications for fisheries management and harmful algal bloom mitigation. However, operational forecasting systems of large scale marine ecosystems require computationally demanding coupled 3D biogeochemical models. Here we propose a new approach to efficiently forecast the ecological state of a large marine ecosystem using a cluster of much cheaper 1D water-column regional ecological models. The regions are determined based on a data-driven clustering approach and remotely-sensed chlorophyll data; a 1D ecological model is then separately implemented for each identified eco-region. Chl-a data is assimilated into these 1D models every 8-day period using a deterministic ensemble Kalman filter (SEIK). To exchange information between adjacent regions, we introduced the concept of assimilating statistical information from neighbour regions. The idea consists in using Chl-a observations in neighbouring clusters to predict, using a linear regression model, the future observation in the region under consideration, and then assimilate the predicted “statistical observations” into the corresponding 1D model using the SEIK analysis step. The linear regression models represent the lateral interactions between clusters such as the advection of water masses that transports plankton and nutrients.

We implemented and tested the proposed framework in the Red Sea, where the clustering method divided the Red Sea into four regions, matching the biological clustering of (Raitso et al. 2013) into NRS, NCRS, SCRS,

1
2
3 1 and SRS. We then implemented 1D ecological models in the NRS, NCRS and SRS. These models successfully
4
5 2 reproduced the main features and seasonal variability of the surface Chl-a concentrations, as inferred from
6
7 3 satellite observations. Paired with the assimilation of Chl-a data, the models were also capable of providing
8
9
10 4 good-quality forecasts of future observations. Furthermore, we demonstrated that these forecasts could be
11
12 5 considerably improved (by nearly 50% in the NRS and NCRS) by exchanging lateral information between
13
14 6 clusters through the assimilation of statistical observations. This improvement did not come at the cost of
15
16 7 sacrificing the stability of the ecological models nor affecting the consistency of the nutrient profiles, suggesting
17
18 8 that the proposed method can be easily applied to improve the forecasting skill of other assimilative marine
19
20
21 9 ecosystem forecasting systems. We notice that the exchange of information yields the largest improvements in
22
23 10 the NRS and in the NCRS during winter, when large amounts of nutrients are advected along the axis of the Red
24
25 11 Sea, indicating that the method effectively represents interactions between neighbouring clusters.

27 12 The proposed approach is portable and should be readily applicable to model the ecology of any large marine
28
29
30 13 ecosystem using a set of easily parallelisable regional 1D ecosystem models. It is also very flexible in that one
31
32 14 may use other clustering algorithms and datasets to divide large marine ecosystems into sub-regions.
33
34 15 Additionally, the model used to predict the statistical observation can be improved by constructing more
35
36 16 sophisticated regression models, such as Gaussian additive models, support vector machines or neural networks
37
38
39 17 (James et al. 2013). Statistical models can also be fitted to forecasts future observations several time steps ahead
40
41 18 and be used to improve ecological model forecasts at more distant time-horizons. Additional covariates could
42
43 19 also be exploited in the statistical model to improve the prediction of future observations, either by taking into
44
45 20 account further away regions, or by considering other ocean variables such as temperature or sea surface height.
46
47 21 This would enable the model to better represent the effect of the physical environment on phytoplankton growth.
48
49
50 22 Finally, another way to exchange information between regions could be to assimilate statistical observations
51
52 23 predicted based on the state variables of neighbouring clusters. Such an approach is less straightforward to apply
53
54 24 since the data used for offline fitting of the statistical models will not be sampled from the same distribution as
55
56 25 the one used online to predict future observation, due to the interaction between models through the assimilation
57
58
59 26 of these quantities. Finally, the use of a cluster of 1D ecosystem models also opens the door for applying fully
60
61 27 non-Gaussian data assimilation techniques, that are known to require large ensembles, such as the Particle and

Gaussian Mixture filters, which are expected to be more performant than the Gaussian-based ensemble Kalman filters, especially with dynamics exhibiting rapid and nonlinear changes as those of a marine ecosystem (Hoteit et al. 2005; Triantafyllou et al. 2013).

Acknowledgements

This research was funded by King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. This research made use of the resources of the supercomputing laboratory at KAUST. We thank the ESA Ocean Colour CCI Team for providing OC-CCI chlorophyll data.

Bibliography

- Alvera-Azcárate A, Barth A, Sirjacobs D, Beckers JM (2009) Enhancing temporal correlations in EOF expansions for the reconstruction of missing data using DINEOF. *Ocean Sci* 5:475–485.
- Anderson TR (2005) Plankton functional type modelling: running before we can walk? *J Plankton Res* 27:1073–1081. doi: 10.1093/plankt/fbi076
- Baretta JW, Ebenhöh W, Ruardij P (1995) The European regional seas ecosystem model, a complex marine ecosystem model. *Netherlands Journal of Sea Research* 33:233–246. doi: 10.1016/0077-7579(95)90047-0
- Baretta-Bekker JG, Baretta JW, Ebenhöh W (1997) Microbial dynamics in the marine ecosystem model ERSEM II with decoupled carbon assimilation and nutrient uptake. *J Sea Res* 38:195–211. doi: 10.1016/S1385-1101(97)00052-X
- Beckers JM, Rixen M (2003) EOF calculations and data filling from incomplete oceanographic datasets. *J Atmos Ocean Tech* 20:1839–1856. doi: 10.1175/1520-0426(2003)020<1839:Ecadff>2.0.Co;2
- Blackford JC, Allen JI, Gilbert FJ (2004) Ecosystem dynamics at six contrasting sites: a generic modelling study. *J Marine Syst* 52:191–215. doi: 10.1016/j.jmarsys.2004.02.004
- Blackford JC, Burkill PH (2002) Planktonic community structure and carbon cycling in the Arabian Sea as a result of monsoonal forcing: the application of a generic model. *J Marine Syst* 36:239–267. doi: 10.1016/S0924-7963(02)00182-3
- Brewin RJW, Raitsos DE, Dall'Olmo G, et al (2015) Regional ocean-colour chlorophyll algorithms for the Red Sea. *Remote Sensing of Environment* 165:64–85. doi: 10.1016/j.rse.2015.04.024
- Brewin RJW, Raitsos DE, Pradhan Y, Hoteit I (2013) Comparison of chlorophyll in the Red Sea derived from MODIS-Aqua and in vivo fluorescence. *Remote Sensing of Environment* 136:218–224. doi: 10.1016/j.rse.2013.04.018
- Ciavatta S, Torres R, Martinez-Vicente V, et al (2014) Assimilation of remotely-sensed optical properties to improve marine biogeochemistry modelling. *Prog Oceanogr* 127:74–95. doi: 10.1016/J.Pocean.2014.06.002
- Ciavatta S, Torres R, Saux-Picart S, Allen JI (2011) Can ocean color assimilation improve biogeochemical hindcasts in shelf seas? *J Geophys Res-Oceans* 116:C12043. doi: 10.1029/2011JC007219
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met* 39:1–38.
- Dreano D, Mallick B, Hoteit I (2015) Filtering remotely sensed chlorophyll concentrations in the Red Sea using a space-time covariance model and a Kalman filter. *Spat Stat* 13:1–20. doi: 10.1016/j.spasta.2015.04.002
- Dreano D, Raitsos DE, Gitting J, et al (2016) The Gulf of Aden Intermediate Water Intrusion regulates the Southern Red Sea Summer Phytoplankton Blooms.
- Edwards CA, Moore AM, Hoteit I, Cornuelle BD (2015) Regional ocean data assimilation. *Ann Rev Mar Sci* 7:21–42. doi: 10.1146/annurev-marine-010814-015821
- Fennel WW, Neumann TT (2004) Introduction to the modelling of marine ecosystems. Elsevier Amsterdam Boston
- Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. *J Am Stat*

Assoc 97:611–631. doi: 10.1198/016214502760047131

Garcia HE, Locarnini RA, Boyer TP, et al (2014) *World Ocean Atlas 2013, Volume 4: Dissolved Inorganic Nutrients (Phosphate, Nitrate, Silicate)*. NOAA Atlas NESDIS

Gharamti ME, Samuelsen A, Bertino L, et al (2017) Online tuning of ocean biogeochemical model parameters using ensemble estimation techniques: Application to a one-dimensional model in the North Atlantic. *J Marine Syst* 168:1–16. doi: 10.1016/j.jmarsys.2016.12.003

Hoteit I, Pham DT, Blum J (2002) A simplified reduced order Kalman filtering and application to altimetric data assimilation in Tropical Pacific. *J Marine Syst* 36:101–127. doi: 10.1016/S0924-7963(02)00129-X

Hoteit I, Pham DT, Gharamti ME, Luo X (2015) Mitigating Observation Perturbation Sampling Errors in the Stochastic EnKF. *Mon Weather Rev* 143:2918–2936. doi: 10.1175/MWR-D-14-00088.1

Hoteit I, Triantafyllou G, Petihakis G (2005) Efficient data assimilation into a complex, 3-D physical-biogeochemical model using partially-local Kalman filters. *Ann Geophys-Germany* 23:3171–3185.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning: with Applications in R*. Springer, New York

Korres G, Triantafyllou G, Petihakis G, et al (2012) A data assimilation tool for the Pagasitikos Gulf ecosystem dynamics: Methods and benefits. *J Marine Syst* 94:S102–S117. doi: 10.1016/J.Jmarsys.2011.11.004

Kürten B, Al-Aidaros AM, Struck U, et al (2014) Influence of environmental gradients on C and N stable isotope ratios in coral reef biota of the Red Sea, Saudi Arabia. *J Sea Res* 85:379–394. doi: 10.1016/j.seares.2013.07.008

Latour RJ, Brush MJ, Bonzek CF (2003) Toward Ecosystem-Based Fisheries Management. *Fisheries* 28:10–22. doi: 10.1577/1548-8446(2003)28[10:TEFM]2.0.CO;2

Miles TN, He R (2010) Temporal and spatial variability of Chl-a and SST on the South Atlantic Bight: Revisiting with cloud-free reconstructions of MODIS satellite imagery. *Continental Shelf Research* 30:1951–1962. doi: 10.1016/j.csr.2010.08.016

Nanninga GB, Saenz-Agudelo P, Manica A, Berumen ML (2014) Environmental gradients predict the genetic population structure of a coral reef fish in the Red Sea. *Mol Ecol* 23:591–602. doi: 10.1111/mec.12623

Patsch J, Radach G (1997) Long-term simulation of the eutrophication of the North Sea: temporal development of nutrients, chlorophyll and primary production in comparison to observations. *J Sea Res* 38:275–310. doi: 10.1016/S1385-1101(97)00051-8

Petihakis G, Triantafyllou G, Allen IJ, et al (2002) Modelling the spatial and temporal variability of the Cretan Sea ecosystem. *J Marine Syst* 36:173–196. doi: 10.1016/S0924-7963(02)00186-0

Petihakis G, Triantafyllou G, Tsiaras K, et al (2009) Eastern Mediterranean biogeochemical flux model – Simulations of the pelagic ecosystem. *Ocean Sci* 5:29–46.

Pettersson LH, Pozdnyakov DVDV (2013) *Monitoring of harmful algal blooms*. Chichester, UK : Springer, published in association with Praxis Publishing

Pham DT (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon Weather Rev* 129:1194–1207. doi: 10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2

- 1
- 2
- 3 1 Racault M-F, Raitsos DE, Berumen ML, et al (2015) Phytoplankton phenology indices in coral reef ecosystems:
4 2 Application to ocean-color observations in the Red Sea. *Remote Sens Environ* 160:222–234. doi:
5 3 10.1016/j.rse.2015.01.019
6
- 7 4 Raitsos DE, Pradhan Y, Brewin RJW, et al (2013) Remote sensing the phytoplankton seasonal succession of the
8 5 Red Sea. *PloS ONE* 8:e64909. doi: 10.1371/journal.pone.0064909
9
- 10 6 Richlen ML, Morton SL, Jamali EA, et al (2010) The catastrophic 2008-2009 red tide in the Arabian gulf region,
11 7 with observations on the identification and phylogeny of the fish-killing dinoflagellate *Cochlodinium*
12 8 *polykrikoides*. *Harmful Algae* 9:163–172. doi: 10.1016/J.Hal.2009.08.013
14
- 15 9 Sofianos SS, Johns WE (2007) Observations of the summer Red Sea circulation. *J Geophys Res-Oceans*. doi:
16 10 10.1029/2006jc003886
17
- 18 11 Triantafyllou G, Hoteit I, Petihakis G, Dounas K (2003): An interpolated Kalman filter to assimilate in-situ data
19 12 into a complex 3-D model of the Cretan sea ecosystem. *Journal of Marine Systems*, 40-41, 213-231,.
20
- 21 13 Triantafyllou G, Korres G, Hoteit I, Petihakis G, Banks AC (2007) Assimilation of ocean colour data into a
22 14 biochemical flux model of the Eastern Mediterranean Sea. *Ocean Science*, 3, 397-410.
23 15
- 24 16 Triantafyllou G, Hoteit I, Luo X, et al (2013) Assessing a robust ensemble-based Kalman filter for efficient
25 17 ecosystem data assimilation of the Cretan Sea. *J Marine Syst* 125:90–100. doi:
26 18 10.1016/J.Jmarsys.2012.12.006
28
- 29 19 Triantafyllou G, Yao F, Petihakis G, et al (2014) Exploring the Red Sea seasonal ecosystem functioning using a
30 20 three-dimensional biophysical model. *J Geophys Res-Oceans* 119:1791–1811. doi: 10.1002/2013jc009641
31
- 32 21 Tsiras KP, Petihakis G, Kourafalou VH, Triantafyllou G (2014) Impact of the river nutrient load variability on
33 22 the North Aegean ecosystem functioning over the last decades. *J Sea Res* 86:97–109. doi:
34 23 10.1016/j.seares.2013.11.007
36
- 37 24 Weikert H (1987) Plankton and the pelagic environment. In: Edwards AJ, Head SM (eds) *Key Environments:*
38 25 *Red Sea*. Pergamon Books, pp 90–111
39
- 40 26 Yao F, Hoteit I, Pratt LJ, et al (2014a) Seasonal overturning circulation in the Red Sea: 2. Winter circulation. *J*
41 27 *Geophys Res-Oceans* 119:2263–2289. doi: 10.1002/2013jc009331
42
- 43 28 Yao F, Hoteit I, Pratt LJ, et al (2014b) Seasonal overturning circulation in the Red Sea: 1. Model validation and
44 29 summer circulation. *J Geophys Res-Oceans* 119:2238–2262. doi: 10.1002/2013jc009004
45
- 46 30 Zavatarelli M, Baretta JW, Baretta-Bekker JG, Pinardi N (2000) The dynamics of the Adriatic Sea ecosystem.
47 31 An idealized model study. *Deep-Sea Res Pt I* 47:937–970. doi: 10.1016/S0967-0637(99)00086-2
49

LIST OF FIGURES

Figure 1. Clustering of the Red Sea into four eco-regions. a) The four provinces defined in (Raitsos *et al.* 2013) along with the average Chl-a concentration in the Red Sea averaged between 1998 and 2011 computed from 8-day Chl-a OC-CCI aggregates. The four provinces follow the Chl-a average concentration patterns. The SRS is the most productive region, the SCRS is a transition region between the SRS and the NCRS, which is the most oligotrophic region. The NRS is the most isolated region and is more productive than the NCRS due to the winter deep mixing. **b)** The four eco-regions obtained using the GMM clustering method. The results resemble closely in size and location the four provinces defined in (Raitsos *et al.* 2013).

Figure 2. Workflow representing the assimilation scheme integrating the exchange of information between clusters through the assimilation of statistical observations.

Figure 3. Results of the statistical assimilation in the northern Red Sea. a) The linear regression model (red line) used to forecast the NRS Chl-a level from the NCRS Chl-a level, one 8-day period beforehand (Statistical Forecast step, Fig 2). The circles represent the training data used to fit the model (period 1997-2002). **b)** Chl-a level in the NRS predicted by the 1D ecological model without assimilation (Free) over the 2003-2004 period; forecast from the 1D model with assimilation (Forecast 1, Dynamical Forecast step, Fig 2); corrected forecast obtained by assimilating the statistical prediction (Forecast 2, Correction step, Fig 2); Chl-a satellite observation. **c)** Absolute value of the errors (with respect to satellite Chl-a observations) given by the model forecast in the assimilation cycle (Forecast 1, Dynamical Forecast step, Fig 2) and the corrected forecast with the statistical observations (Forecast 2, Correction step, Fig 2).

Figure 4. Results of the statistical assimilation in the northern-central Red Sea. **a)** The linear regression model (solid plane) used to forecast the NCRS Chl-a level from the NRS and SCRS Chl-a levels, one 8-day period beforehand (Statistical Forecast step, Fig 2). The solid dots represent the training data used to fit the model (period 1997-2002). **b)** Chl-a level in the NCRS predicted by the 1D ecological model without assimilation (Free) over the 2003-2004 period; forecast from the 1D model with assimilation (Forecast 1, Dynamical Forecast step, Fig 2); corrected forecast obtained by assimilating the statistical prediction (Forecast 2, Correction step, Fig 2); Chl-a satellite observation. **c)** Absolute value of the errors (with respect to satellite Chl-a observations) given by the model forecast in the assimilation cycle (Forecast 1, Dynamical Forecast step, Fig 2) and the corrected forecast with the statistical observations (Forecast 2, Correction step, Fig 2).

Figure 5. Results of the statistical assimilation in the southern-central Red Sea. a) The linear regression model (red line) used to forecast the SCRS Chl-a level from the NCRS Chl-a level, one 8-day period beforehand (Statistical Forecast step, Fig 2). The circles represent the training data used to fit the model (period 1997-2002). b) Chl-a level in the SCRS predicted by the 1D ecological model without assimilation (Free) over the 2003-2004 period; forecast from the 1D model with assimilation (Forecast 1, Dynamical Forecast step, Fig 2); corrected forecast obtained by assimilating the statistical prediction (Forecast 2, Correction step, Fig 2); Chl-a satellite observation. c) Absolute value of the errors (with respect to satellite Chl-a observations) given by the model forecast in the assimilation cycle (Forecast 1, Dynamical Forecast step, Fig 2) and the corrected forecast with the statistical observations (Forecast 2, Correction step, Fig 2).

Figure 6. Nutrient profiles for the three eco-regions given by WOA, the models free run (averaged over 2003-2004), and the improved forecast (Statistical Correction Step, Fig 2, averaged of 2003-2004). a) Phosphate (PO₄). b) Nitrate (NO₃).

LIST OF TABLES

Table 1. RMSE (with respect to satellite observations) of NRS Chl-a concentrations forecasts using $SEIK_{STD}$ assimilation (without assimilation of statistical predictions) for different values of the observation error covariance.

$1/R$	50	100	200	300	500
RMSE	0.081	0.072	0.070	0.074	0.077

Table 2. RMSE of NRS Chl-a (with respect to satellite observations) with $SEIK_{STAT}$ assimilation, as forecasted by the 1D ecological model (Forecast 1) and after the assimilation of the statistical predictions (Forecast 2), as well as the percentage improvement between the improved forecast (Forecast 2) and the forecast without assimilation of statistical information ($SEIK_{STD}$, Table 1).

$1/\tilde{R}$	50	100	150	200
RMSE Forecast 1	0.072	0.073	0.074	0.075
RMSE Forecast 2	0.055	0.046	0.041	0.038
% Improvement	21	34	41	46

Table 3. RMSE (with respect to satellite observations) of NCRS Chl-a concentrations forecasts using $SEIK_{STD}$ assimilation (without assimilation of statistical predictions) for different values of the observation error covariance.

$1/R$	50	75	100	150	200	400
RMSE	0.049	0.049	0.049	0.050	0.051	0.054

Table 4. RMSE of NCRS Chl-a (with respect to satellite observations) with SEIK_{STAT} assimilation, as forecasted by the 1D ecological model (Forecast 1) and after the assimilation of the statistical prediction (Forecast 2), as well as the percentage improvement between the improved forecasts (Forecast 2) and the forecast without assimilation of statistical information (SEIK_{STD}, Table 3).

$1/\tilde{R}$	25	50	75	100
RMSE Forecast 1	0.051	0.052	0.053	0.054
RMSE Forecast 2	0.034	0.028	0.026	0.024
% Improvement	31	43	47	51

Table 5. RMSE (with respect to satellite observations) of SCRS Chl-a concentrations forecasts using SEIK_{STD} assimilation (without assimilation of statistical predictions) for different values of the observation error covariance.

$1/R$	50	100	200	400
RMSE	0.116	0.112	0.105	0.103

Table 6. RMSE of SCRS Chl-a (with respect to satellite observations) with SEIK_{STAT} assimilation, as forecasted by the 1D ecological model (Forecast 1) and after the assimilation of the statistical prediction (Forecast 2), as well as the percentage improvement between the improved forecasts (Forecast 2) and the forecast without assimilation of statistical information (SEIK_{STD}, Table 3).

$1/\tilde{R}$	25	50	100	150	200
RMSE Forecast 1	0.102	0.100	0.100	0.099	0.099
RMSE Forecast 2	0.079	0.078	0.082	0.084	0.087
% Improvement	21	22	18	16	13