Progress in Oceanography 138 (2015) 58–76

Contents lists available at ScienceDirect

Progress in Oceanography

journal homepage: www.elsevier.com/locate/pocean

Combining shipboard in situ data with satellite data to estimate daily primary production in a coastal upwelling system: A data mining approach

Robert I. Williamson [a,b,*], John G. Field [a,b], Frank A. Shillington [a,c], Astrid Jarre [b,d], Anet Potgieter [e]

[a] Department of Oceanography, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
[b] Marine Research Institute, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
[c] Nansen Tutu Centre, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
[d] Department of Biological Sciences, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
[e] Agents Research Lab, 10 Bloemendal Road, Mowbray, 7700, South Africa

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form 8 September 2015
Accepted 8 September 2015
Available online 12 September 2015

ABSTRACT

This study classifies coastal time-series data according to subsurface phytoplankton vertical distributions to be able to capture the variability of primary production at fine spatial and temporal scales. Our method uses algorithms developed to extract patterns in large datasets of time-sequential data. We use short time-series of QuikSCAT surface winds, MODIS sea surface temperature and surface chlorophyll *a* associated with each in situ chlorophyll *a* profile, as well as the season and bottom depth of the in situ station to discover patterns that can be used to classify new data into 12 profile classes. We first fill in missing MODIS data using a conditional random field model so that cloudy days are not excluded. The most likely profile is then predicted using all the available data. We apply our method to the St Helena Bay area, a region within the productive Benguela Current upwelling system. A profile is predicted for each day and each pixel of 4 km resolution satellite image for 16 consecutive months. Each profile is used in a broad-band photosynthesis model to produce a daily three-dimensional estimate of gross primary production. An average production rate of 3.2 g C m^{-2} day^{-1} was obtained for the area, which shows very good agreement with other estimates from the region. The results show persistent high productivity near the surface throughout the year with the exception of the winter months. Deeper in the water column productivity is more seasonal. The 16 month time-series highlights the interannual, seasonal and daily variability of the system. By linking physical processes to the distribution of phytoplankton at appropriate spatio-temporal scales, we can now more rigorously investigate bottom-up driven impacts on ecosystems characterised by short-term variability.

1. Introduction

Photosynthesis by marine phytoplankton is the vital link in cycling carbon between its inorganic form and living marine organisms (Behrenfeld et al., 2006). Estimation of carbon mass entering the marine environment requires accurate measurement of carbon fixation by phytoplankton. To estimate total carbon production over the water column it is necessary to integrate with respect to depth through the euphotic zone. This requires information on the phytoplankton photosynthetic pigments (approximated by the biomass of chlorophyll *a*) and the photosynthetically available radiation (PAR) at depth. Resolving the vertical light and biomass distribution can greatly improve the accuracy of primary production models (Jacox et al., 2013).

Numerous publications on depth-integrated primary production estimated from ocean colour data incorporate the vertical distribution of chlorophyll *a* (chl *a*). The simplest approach is to assume a vertically homogenous biomass profile (André, 1992; Behrenfeld and Falkowski, 1997a; Platt, 1986). More complex methods estimate the vertical distribution using satellite-derived surface biomass (Antoine et al., 1996; Hoepffner et al., 1999; Morel and Berthon, 1989; Uitz et al., 2010) while others estimate the parameters of a Gaussian curve that describes the vertical distribution (Platt et al., 2008, 1991, 1988). A different approach is to adopt representative profiles for a given region and season

Abbreviations: SSC, sea surface chlorophyll *a*; CRF, conditional random field; CAR, class association rules.

* Corresponding author. Current address: Department for Biodiversity & Conservation Biology, University of the Western Cape, Private Bag X17, Bellville, 7535, South Africa.

E-mail address: ri.williamson@outlook.com (R.I. Williamson).

http://dx.doi.org/10.1016/j.pocean.2015.09.004
0079-6611/

following the biogeography concept of Platt and Sathyendranath (1988) (Longhurst et al., 1995; Sathyendranath et al., 1995). These methods originated with the advent of satellite ocean colour data and had the objective of obtaining large-scale estimates of primary production. At the basin to global scale, monthly, seasonal or regionally static representation of the biological structure of the water column may be an appropriate simplification. However, recent evaluations of ocean colour models showed a decrease in model skill towards the coast and in shallower regions, and an underestimation of the observed variability of primary production (Carr et al., 2006; Friedrichs et al., 2009; Saba et al., 2011). Kahru et al. (2009) applied commonly used models in the California Current System and found a systematic high productivity bias in all the tested models. Sources of error or uncertainty that have been suggested include the input data, an over estimation of the euphotic depth and the assumption of a steady state system (Carr et al., 2006; Kahru et al., 2009; Siegel et al., 2001).

Many coastal systems are complex due to interaction with the coastal topography and local wind forcing, and models that incorporate data that average over the appropriate spatial and temporal scales of the events of interest will not be able to reproduce the variability. This is particularly important for coastal upwelling systems that are characterised by higher rates of primary production, carbon export and fisheries yields. Demarcq et al. (2008) demonstrated the monthly variability of primary production in the Benguela upwelling system by predicting the chl *a* profile for each pixel of monthly SeaWiFS images at 4.5 km spatial resolution. Although their method highlighted the monthly variability of primary production in the system, working on monthly time-scales in upwelling areas may still misrepresent the physical–biological interaction (García-Reyes et al., 2014).

Previous work on primary production in the Benguela system has attempted to incorporate the variability of phytoplankton distributions. Silulwane et al. (2001) described characteristic profiles in the Benguela system according to environmental data associated with each profile class. Demarcq et al. (2008) progressed on this work by using environmental data to predict the monthly profile shape, and used the profiles in a broad-band light transmission model to estimate monthly primary production. In this paper we extend this work by linking the dynamic physical forcing of surface wind stress and time-series of daily surface data to each characteristic profile. Models of the effects of synoptic wind data on coastal upwelling production have previously shown the importance of wind variability (Botsford et al., 2006; García-Reyes et al., 2014).

To overcome the paucity of data in time-series of daily sea surface temperature (SST) and surface chl *a* (SSC) data, we apply a conditional random field (CRF) model. The CRF, as opposed to the commonly used methods based on empirical orthogonal functions (e.g. DINEOF (Alvera-Azcárate et al., 2007) and variations thereof), is a highly-flexible non-linear method based on probability theory. The CRF model uses available information to predict the most likely missing value in a sequence. CRFs provide a framework for labelling sequential data based on the conditional approach. By applying the statistical properties of the class labels conditioned on the observations, the label with the highest likelihood can be calculated. They have been shown to outperform the more commonly used hidden Markov model on a number of real-world sequential labelling tasks (Lafferty et al., 2001; Pinto et al., 2003; Sha and Pereira, 2003).

Data from the St Helena Bay area, a region within the southern Benguela upwelling system, are used to test the methodology. The St Helena Bay area includes an intense upwelling cell off Cape Columbine, an inshore region known for its high SSC due to retentive oceanographic processes, and a wide shelf region which has a fairly persistent shelf edge front with characteristic open ocean water beyond the front. Within the Benguela system, this area is one of the most productive in terms of phytoplankton productivity and fisheries yields and is an important nursery ground for local fish species (Hutchings et al., 2012).

2. Methods

Our method links the forcing (represented by a history of daily remotely-sensed surface winds) with an in situ snapshot of the subsurface vertical distribution of the phytoplankton biomass (chl *a* profile). Information on recent changes in remotely-sensed sea surface temperature (SST) or state of upwelling, and the recent cumulative biomass of remotely sensed phytoplankton or sea surface chlorophyll *a* (SSC) is included. Bottom depth and season are also taken into account.

To simplify the complex interactions of the variables, the data were pre-processed into discrete classes. A conditional random field (CRF) model was used to fill in any missing class labels in the SST and SSC sequences. The CRF model, which evaluates the conditional probability of all possible values for the missing data in a given sequence, is applied to the entire time-series at each pixel location. As this can be computationally expensive process on a standard desktop computer, the CRF model was applied to a limited 16 month daily time-series from a sub-region of the southern Benguela. With complete sequences of data, each profile class could be associated with a suite of relevant information on the concurrent and prior state of the environment that may influence the profile's shape. The statistical relationships between the time-series of the variables and the profiles are used to predict the profile for each pixel of a daily 4 km resolution satellite image. The profiles are used with a satellite-derived estimate of surface Photosynthetically Active Radiation (PAR) to calculate the subsurface light field and subsequently the gross vertical primary production distribution (Fig. 1).

2.1. Data pre-processing

2.1.1. Shipboard regional profiles (k-means clustering)

The South African Department of Environmental Affairs (DEA) has provided a record of over 6500 vertical profiles of phytoplankton fluorescence and temperature obtained from a thermistor and profiling fluorometer on a Seabird SBE CTD deployed from ships. These records date from 1988–2011 and cover the southern Benguela region. Data were collected during annual surveys of demersal and pelagic fish stocks. The season and the depth range of the surveys vary according to the target stock, which explains the inshore location of surveys in autumn and the few surveys in winter (Fig. 2). A total of 1590 profiles fall within the St Helena Bay region.

The profile data consist of three batches based on their processing. The first batch, which spans 1988–2001 and the second batch from 2001–2008 were recorded with a Chelsea Instruments Aqua-Tracka MK III. For the first batch, a calibration equation was obtained by regressing the recorded voltages from the fluorescence sensor against discrete depth samples beginning a few metres below the surface and at intervals of 10 m down to between 30 and 50 m depending on the profiling signature. Samples were filtered onto Whatman GF/F filters, which were placed in 90% acetone for 24 h to extract the pigments. Chl *a* was measured fluorometrically using a Turner Designs Model 10-000R fluorometer. The second batch was measured by reverse-phase high performance liquid chromatography (HPLC) method described by Barlow et al. (1997). The third batch spans 2009–2011 and was obtained from a WETLabs fluorometer. A standard fluorometer calibration from WETLabs software was used to obtain chl *a* concentrations. Profiles were sampled at 1 m depth intervals after being smoothed