Ecological Modelling 382 (2018) 51–62

Contents lists available at ScienceDirect

Ecological Modelling

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

A depth-resolved artificial neural network model of marine phytoplankton primary production

F. Mattei[a,b,*], S. Franceschini[a,b], M. Scardi[a,b]

[a] Department of Biology, University of Rome "Tor Vergata", via della Ricerca Scientifica, 00133, Rome, Italy
[b] CoNISMa, Piazzale Flaminio, 9, 00196, Rome, Italy

ARTICLE INFO

Keywords:
Marine phytoplankton
Primary production
Artificial neural networks
Depth-resolved model

ABSTRACT

Marine phytoplankton primary production is an extremely important process and its estimates play a major role not only in biological oceanography, but also in a broader context, due to its relationship with oceanic food webs, energy fluxes, carbon cycle and Earth's climate.

The measurement of this process is both expensive and time consuming. Therefore, indirect methods, which can estimate phytoplankton primary production using only remotely sensed predictive information, have many advantages. We describe the development of a depth-resolved model based on an Artificial Neural Network for estimating global phytoplankton primary production. Furthermore, we applied two different approaches, based on input perturbation analysis and on connection weights, to assess the relative importance of the predictive variables. Finally, we compared the results of our depth-resolved model with a previous depth-integrated solution, showing that through the depth-resolution we gained not only useful information on the vertical distribution of the estimated primary production, but also an enhanced accuracy in its depth-integrated estimates.

1. Introduction

Phytoplankton primary production is a fundamental process due to its intimate relationship with oceanic biotic and abiotic processes, including biogeochemical cycles, especially the carbon one, and Earth's climate (Hattam et al., 2015). Furthermore, it represents the bulk of the whole oceanic autotrophic production (Duarte and Cebrián, 1996), which contributes roughly one-third of the global primary production.

For these reasons, the assessment of phytoplankton primary production and the study of its variability, both from a spatial and a temporal perspective, plays a fundamental role in marine ecological studies, from food webs (Richardson and Schoeman, 2004) to fisheries (Nixon, 1988; Holmlund and Hammer, 1999) and from large marine ecosystems to a better understanding of the relationships between fundamental and demand-derived ecosystem services (Costanza et al., 1997; Cloern et al., 2014; Hattam et al., 2015).

As the direct measurement of this biological process is not only difficult, but also expensive and time-consuming, the need for methods aimed at its indirect evaluation is evident. This is especially true if the space and/or time scale of a study is too large for direct measurements and obviously when the main objective is a global assessment. In fact, many models have been developed for the estimation of phytoplankton primary production and several among them are based on an empirical approach, i.e. they assume that primary production can be estimated as a function of other variables (Platt and Sathyendranath, 1988; Scardi, 1996; Morin et al., 1999). Some of these models are based on predictive variables that can be obtained from remote sensing of the ocean colour, which is a relatively cheap source of information on a global scale, while others combine information obtained from ocean circulation and biogeochemistry (Aumont et al., 2003; Buitenhuis et al., 2006). These approaches provide a way to avoid long and expensive sampling procedures, that would otherwise impose strict limits to the practical applications of primary production estimates (Clark et al., 2001; Low-Décarie et al., 2014).

The structure of these indirect methods widely varies in complexity, also depending on the required input variables and on the nature of the relationships between those variables and the desired output. For example, the empirical models proposed by Ryther and Yentsch (1957) and Smith and Baker (1978) took into account only chlorophyll concentrations and irradiance, while the model developed by Behrenfeld and Falkowski (1997) used a larger set of predictive variables, although derived from chlorophyll concentration, temperature and photosynthetically active radiation (PAR).

The broad variety of these techniques and their utility in different fields have led to various comparisons between models over the years, especially thanks to the Primary Productivity Algorithm Round Robin

* Corresponding author at: Department of Biology, University of Rome "Tor Vergata", via della Ricerca Scientifica, 00133, Rome, Italy.
E-mail address: francesco.mattei90@yahoo.it (F. Mattei).

https://doi.org/10.1016/j.ecolmodel.2018.05.003
Received 19 January 2018; Received in revised form 30 April 2018; Accepted 1 May 2018

(PPARR) (Campbell et al., 2002; Carr et al., 2006; Saba et al., 2011; Lee et al., 2015). The PPARR has been useful in order to assess and compare the accuracy of different models. Furthermore, comparisons have laid the basis for substantial improvements in this field, providing a common context in which a wide range of methods has been tested. In this way, the main differences among models have been outlined.

Among the various models tested in the PPARR framework, only one was based on an Artificial Neural Network (ANN), which was used to develop a depth-integrated model for the evaluation of the phytoplankton primary production (Scardi, 2001). As the ANN approach to the estimation of the primary production has shown good results in comparisons with other models (Friedrich et al., 2009; Saba et al., 2011; Lee et al., 2015). Moreover, if compared to other techniques ANNs can be easily re-calibrated as soon as new data become available.

In fact, empirical models built through this Machine Learning approach, if correctly trained, are able to reproduce the complex non-linear relations that underpin most natural processes, and phytoplankton primary production is no exception. Moreover, this method does not trade generality for simplicity, as it can involve complex calculations during its development (*training*, in ANN jargon) (Scardi, 1996; Lek et al., 1996b; Scardi and Harding, 1999; Scardi, 2001; Olden et al., 2008). While the computational burden needed to develop a model is not light, it only requires computational resources that are no longer a limiting factor because of advances in computing power. Once developed, an ANN model is as fast and easy to run as most other models.

The structure of an ANN model is not defined a priori, but it is determined during the training procedure. In fact, other types of methods try to describe analytically the major processes the primary production depends on, although this can be a very difficult task, especially if the objective is to build a model that is both general and accurate. On the contrary, the ANN approach tries to learn the relationships between the input variables and the primary production directly from the available data. This is a benefit not to be underestimated while dealing with natural processes that involve complex relationships, which in most cases are unknown or difficult to parameterize, especially on a global scale (Maier and Dandy, 2000). Moreover, a model based on ANNs or other Machine Learning approaches can be easily updated without rebuilding it from scratch. In fact, if a new data set becomes available at a later stage relative to the development of an ANN model, it can be easily exploited simply by performing a new training procedure for the ANN, without having to reinterpret the relationship between the variables given the empirical nature of the approach.

The advantage in using an ANN as a tool for ecological modelling becomes more evident if both the heterogeneity of remote sensing information and the incomplete understanding of the causal relationships between the variables of interest are taken into account. In fact, this methodology does not need any a priori knowledge in order to exploit the information contained in the input variables and, at the same time, it is robust enough with respect to a few redundant inputs, if any. These properties allow the use of a wide range of predictive variables with no need for a priori knowledge about the nature of the relationship with the output of the model, i.e. with primary production, thus enhancing the potential value of any data source, including remotely sensed data.

Indeed, the possibility to explore a wider range of information looking for potential input variables is a major advantage in ecological modelling, especially in the light of the lack of large data set and of the difficulties in obtaining some of the desired measures from which an ANN model can learn. In fact, primary production data sets are often incomplete especially in terms of spatial coverage. Some regions of the world are over-represented while others are under-represented or, even worse, not represented at all. These issues may affect the model capability of generalisation and thus interfere with the modelling accuracy on a broad spatial scale (Lek et al., 1996b; Maier and Dandy, 2000; Recknagel, 2001). In this context it is important to note that the performance of any kind of model is highly dependent on both the quality and the amount of available data. However, this is especially true for Machine Learning approaches, in which the data drive the training procedure without any explicit mathematical formulation between the predictive variables and the targets.

As the phytoplankton primary production estimates are useful in several research fields, such as the management of fishery resources (Nixon, 1988, 1992; Conti and Scardi, 2010), the assessment of the ocean oxygen production (Reuer et al., 2007), the study of climate change influences on the oceanic primary production (Behrenfeld et al., 2006) and both the understanding and the evaluation of various ecosystem services (Cloern et al., 2014; Hattam et al., 2015; Holmlund and Hammer, 1999; Richardson and Schoeman, 2004; Melaku Canu et al., 2015), any improvement in the accuracy of these appraisals could be a very important achievement. ANNs have proved to be valuable tools, providing good results with respect to the modelling of biological processes and they are widely open to experimentation and optimization in data preprocessing and in model fine-tuning (Scardi, 1996; Lek et al., 1996b; Maier and Dandy, 2000; Olden et al., 2008).

In this framework, our main goal was the development of a model which could use the available information more efficiently to improve both the accuracy and the granularity of the primary production estimates. We decided to opt for a depth-resolved solution using an empirical approach based on an ANN with the aim to describe not only the vertically integrated magnitude of the phytoplankton production but also its distribution along the water column. We also decided to rely upon surface data only as predictive information, thus assuring the widest applicability of the resulting model.

In fact, although a depth-resolved approach to the primary production evaluation using an ANN was already presented by Scardi (2003), that model was strictly local and mainly aimed at demonstrating the potential of the method. Therefore, it was developed and tested on a rather limited data set, thus embedding a small amount of heterogeneity and a limited set of structures of the primary production profiles. On the contrary, a much wider variability in vertical production profiles is one of the most challenging elements for the global phytoplankton primary production models.

The results of the depth-resolved model we present here were compared to those of the depth-integrated ANN model developed by Scardi (2001). The two models share the same computational method and the same spatial scale, while the data set upon which they were trained coincide almost completely. Therefore, their comparison can show if an enhancement in the accuracy of primary production estimates can be achieved through a depth-resolved approach and appropriate data management and preprocessing.

2. Materials and methods

2.1. Data preprocessing and partitioning

The data set used in this study includes 3304 vertical profiles of phytoplankton primary production that were acquired during oceanographic cruises carried out from 1954 to 1994, which have been obtained from http://www.science.oregonstate.edu/ocean.productivity/field.data.c14.online.php. The bulk of the sampling stations was located in three regions. The first one, and the most represented, corresponds to the North-Western Atlantic, off the coast of the United States, while the second one is situated in the Eastern Equatorial Pacific, off the Western coast of South America and the last one is off the West coast of the United States. The remaining part of the world oceans hosts only a few sparse stations, but their scarcity makes the information obtained from them even more relevant.