Biogeosciences, 6, 2333–2353, 2009
www.biogeosciences.net/6/2333/2009/

Biogeosciences

Skill assessment of the PELAGOS global ocean biogeochemistry model over the period 1980–2000

M. Vichi and S. Masina

Centro Euro-Mediterraneo per i Cambiamenti Climatici, Bologna, Italy

Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy

Received: 24 March 2009 – Published in Biogeosciences Discuss.: 1 April 2009
Revised: 7 October 2009 – Accepted: 16 October 2009 – Published: 2 November 2009

Abstract. Global Ocean Biogeochemistry General Circulation Models are useful tools to study biogeochemical processes at global and large scales under current climate and future scenario conditions. The credibility of future estimates is however dependent on the model skill in capturing the observed multi-annual variability of firstly the mean bulk biogeochemical properties, and secondly the rates at which organic matter is processed within the food web. For this double purpose, the results of a multi-annual simulation of the global ocean biogeochemical model PELAGOS have been objectively compared with multi-variate observations from the last 20 years of the 20th century, both considering bulk variables and carbon production/consumption rates. Simulated net primary production (NPP) is comparable with satellite-derived estimates at the global scale and when compared with an independent data-set of in situ observations in the equatorial Pacific. The usage of objective skill indicators allowed us to demonstrate the importance of comparing like with like when considering carbon transformation processes. NPP scores improve substantially when in situ data are compared with modeled NPP which takes into account the excretion of freshly-produced dissolved organic carbon (DOC). It is thus recommended that DOC measurements be performed during in situ NPP measurements to quantify the actual production of organic carbon in the surface ocean. The chlorophyll bias in the Southern Ocean that affects this model as well as several others is linked to the inadequate representation of the mixed layer seasonal cycle in the region. A sensitivity experiment confirms that the artificial increase of mixed layer depths towards the observed values substantially reduces the bias. Our assessment results qualify the model for studies of carbon transformation in the surface ocean and metabolic balances. Within the limits of the model assumption and known biases, PELAGOS indicates a net heterotrophic balance especially in the more oligotrophic regions of the Atlantic during the boreal winter period. However, at the annual time scale and over the global ocean, the model suggests that the surface ocean is close to a weakly positive autotrophic balance in accordance with recent experimental findings and geochemical considerations.

1 Introduction

Ocean Biogeochemistry General Circulation Models (OBGCM) derive from the coupling of GCMs that solve the hydrodynamics of the ocean and biomass-based mathematical representations of the lower trophic levels of marine ecosystems. Given the limitation of the biomass-based mathematical definitions, OBGCMs are a rough approximation of the complexity observed in the global ocean ecosystem. The focus of these models is not the study of interactions at the community and ecosystem scales but the quantitative representation of the biogeochemical fluxes of major constituents among the lower trophic levels. The planktonic components are, despite their little standing stock, the basis of the biological pump in the ocean with a global estimated primary production ranging from less than 40 to more than 60 Pg C yr^{-1} (Carr et al., 2006). Even if the net role of the oceanic biological pump in the global carbon cycle is still to be clarified (e.g. Sarmiento et al., 1998), there are concerns that alterations in ocean features due to climate change may affect the ocean biogeochemistry with feedbacks into the climate system itself (see Denman et al., 2007, for a compendium on the current literature).

Correspondence to: M. Vichi
(vichi@bo.ingv.it)

OBGCMs are promising tools to study the functioning of global biogeochemical processes and to produce future projections driven by climate change scenarios. The credibility of projections is however dependent on the adequate simulation of the ocean biogeochemical features under current climate conditions. This requires an objective assessment of model skill against available data, a method that only recently has penetrated the community of biogeochemical modellers (Lynch et al., 2009). The comparison between observations and model results at the global scales is however both a technical and a scientific problem. On the one hand data availability and data gridding have practical limitations but on the other hand it is crucial that observations and model variables are compared appropriately by considering the underlying processes on both sides. .

This work presents an assessment study of a multi-component biogeochemical model that was originally developed for coastal regions and recently generalized to the global ocean (see Sect. 2.1). The paper is conceptually divided in two main parts. In Sects. 2.2, and 3 we introduce and use a set of multi-variate, multi-instrument observations collected over the last 20 years of the 20th century to objectively evaluate the results of a multi-annual simulation of the model. This part is meant to demonstrate the model validity at the global and local scales as a tool for process understanding and for use as a component of an Earth System Model for carbon cycle studies. We focus mostly on the assessment of primary production and microbial carbon transformation processes as these are the basic information used to derive carbon export and sequestration rates in the oceanic carbon cycle (e.g. Dunne et al., 2007). The assessment exercise implies the computation of adequate skill scores (Sect. 2.2) and the choice of acceptable levels of validity. To cover the possible range of currently available data, we focused on the evaluation of mean bulk variables at the global scale mostly using satellite-derived data (Sects. 3.1 and 3.2) and on multi-annual process variables at selected locations for testing the seasonal and inter-annual variability of organic carbon transformation rates at the microbial level (Sects. 3.3 and 3.4).

The discussion section introduces the second part of the paper (Sect. 4). Initially, the results are discussed in view of the model formulation and main assumptions, providing explanations of the major biases and recommendations for model-data comparisons. Several studies have pointed out the existence of biogeographical provinces and that physically distinct oceanic regions have different biogeochemical characteristics (Longhurst, 2007), and some specific regional parameterizations might be useful to capture, for instance, the satellite-derived chlorophyll variability (Tjiputra et al., 2007). However, the extrapolation at larger spatial and temporal scales of limited observations describing carbon cycle rates may lead to misrepresentation of the microbial processes over the annual scale (e.g. Maixandeau et al., 2005a). This is thus a valid argument for using models of adequate complexity to make this extrapolation, since a properly assessed model is expected to capture the major features of the ocean physical processes.

As an example of this methodology, we computed in Sect. 4.3 the metabolic state of the surface global ocean under simulated current climate conditions, using the ratio of net community production and bacterial carbon demand over net primary production in the euphotic zone as indicators of the biological pump efficiency in the model. The discussion on the metabolic state of the surface ocean (e.g. Del Giorgio and Duarte, 2002; Riser and Johnson, 2008) has important implications for the role of microbial biogeochemistry in the oceanic carbon balance. The estimates of net surface community production provided in Sect. 4.3 are baseline references to compare with other on-going experiments where PELAGOS is used as a component of a coupled carbon cycle climate model under future climate scenarios. Section 5 finally offers some methodological recommendations and a summary of the major conclusions.

2 Methods

2.1 Model description and setup

PELAGOS (PELAgic biogeochemistry for Global Ocean Simulations, Vichi et al., 2007a,b) is a coupling between the OPA (Océan PArallelise) general circulation model (Madec et al., 1999) and the global ocean version of the Biogeochemical Flux Model (BFM, http://bfm.cmcc.it) originally derived and modified from the ERSEM regional model (Baretta et al., 1995). The model grid is the irregular ORCA2 configuration (Madec and Imbard, 1996) with a nominal 2×2 degrees size and a refined latitudinal mesh of 0.5 degree in the equatorial regions.

The biogeochemical model implements a set of biomass-based differential equations that solves the fluxes of carbon, nitrogen, phosphorus, silica and iron among selected biological functional groups representing the major components of the lower trophic levels. The functional groups in the pelagic environment are represented by unicellular planktonic autotrophs (pico-, nano-phytoplankton and diatoms), zooplankton (nano-, micro- and meso-) and heterotrophic bacterioplankton. The model also simulates the dynamics of nitrate, ammonium, phosphate, biogenic silicate, iron, oxygen and has an explicit parameterization of the biochemical cycling of dissolved/particulate non-living organic matter.

The results analysed here are extracted from a multi-annual simulation over the period 1958–2001 forced with daily mean heat and momentum fluxes from the European Centre for Medium Range Weather Forecasting (ECMWF) 40-year re-analysis. The forcing functions and the results of a similar physical simulation are described in Bellucci et al. (2007). The ocean physics parameterizations are as in Vichi et al. (2007a) with sea surface temperature (SST) relaxed to the daily-interpolated value of the Reynolds data set

(Reynolds et al., 2002) with a coefficient of 40 W m^{-2}. The biogeochemical model is initialized as in Vichi et al. (2008) from the World Ocean Atlas nutrient data (Conkright et al., 2002) and with homogeneous low values for all the other biogeochemical variables. The model was not calibrated against the data used in this paper and the current set of parameter values was derived following the experience with the model forced by climatological atmospheric data, by means of a manual one-at-a-time modification starting from the values presented in Vichi et al. (2007a). The major changes involved an increase of the affinity constants for nutrients compared to the original values in Baretta-Bekker et al. (1995) and a general reduction of the iron limitation for all phytoplankton groups. A new list of parameter values is available as a supplemental table (http://www.biogeosciences.net/6/2333/2009/bg-6-2333-2009-supplement.pdf and in the PELAGOS page of the BFM web site (http://bfm.cmcc.it) with the differences with respect to Vichi et al. (2007a, also available on the same page).

2.2 Data sets and skill indicators

The biogeochemical data sets used in this assessment encompass the last 20 years of the 20th century with a focus on the data that offer multivariate information and especially biological rates for at least a decade (e.g. primary and bacterial production). Global coverage data are however related to derived parameters of phytoplankton biomass only, available mostly through satellites and ocean color products (e.g. the Sea Wide Field-of-view Sensor. SeaWiFS). Empirical data-models are required to translate sensor information into relevant properties such as chlorophyll-a concentration, and the quality of the reconstructed data is to be considered when comparing with deterministic models. For instance, Gregg and Casey (2004) report an average root mean square log error of 31% and a coefficient of determination of 0.76 for chl satellite estimates against in situ data over the whole global ocean. These scores meet the objectives of the SeaWiFS mission at the global scale. On the other hand, a thorough validation analysis against concurrent in situ chl data shows regional discrepancies with overestimation in the equatorial Atlantic and underestimation in the Southern Ocean. Satellite data and inherent optical properties can be further combined to estimate other important biogeochemical properties such as primary production, carbon content and plankton functional groups distributions (e.g. Behrenfeld and Falkowski, 1997; Alvain et al., 2005; Aiken et al., 2007). In this global assessment we used the estimates of primary production derived with the Vertically Generalized Production Model (VGPM) proposed by Behrenfeld and Falkowski (1997).

The other datasets used in this assessment are in situ observations that have been selected because of their temporal and spatial coverage of primary production and other relevant biogeochemical rates or biomass data. We focused on three publicly available datasets: the ClimPP dataset (Friedrichs et al., 2009) and the Joint Global Ocean Study (JGOFS) time series HOT (Hawaii Ocean Timeseries at Station ALOHA, Lukas and Karl, 1999) and BATS (Bermuda Atlantic Time-Series, Steinberg et al., 2001). These data are further described in their specific Sects. 3.3 and 3.4.

The choice of the performance indicators or scores is done according to recent works that focused on skill assessment (Allen et al., 2007; Lynch et al., 2009; Friedrichs et al., 2009; Stow et al., 2009). The suggested univariate indices comprise the measure of bias (B), average absolute error (AAE) and variability of the misfit measured as Root Mean Square Differences (RMSD, see Appendix A). Two additional performance indicators have been applied as suggested by Allen et al. (2007) and Stow et al. (2009): the Modelling Efficiency (MEF, Nash and Sutcliffe, 1970) and the Reliability Index (RI, Leggett and Williams, 1981), which are further described in the Appendix A. Regression analysis was also performed to evaluate the goodness-of-fit of prediction vs. observations taking into account the linear methods described in Smith and Rose (1995) and Pineiro et al. (2008).

To visualize the combination of the different indicators and compare the PELAGOS results with the other biogeochemical models presented in Friedrichs et al. (2009), we use the Multi-Dimensional Scaling (MDS) technique. The MDS is an iterative technique used to visualize proximities in a low-dimensional space first introduced in psychometrics (Borg and Groenen, 2005). This analysis carefully preserves the distance between items from multivariate datasets and allows the combined visualization of multiple information in one single plot (see also the Appendix A).

Score values and confidence intervals were also evaluated by means of empirical p-values estimates and bootstrap techniques. The probability of obtaining a score value better than the one achieved is generally termed p-value (Mason, 2008). The empirical distribution of score values was constructed with 10 000 random re-samplings of the observation (or simulation) time series and computing the verification index for each new set of model-data pairs. As pointed out by Mason (2008), p-values do not answer the question whether the score value is good, but rather they provide a degree of significance with respect to random combinations. The confidence interval is instead computed by means of the bootstrap technique, in which the choice of randomly permuted model-data pairs is done by replacing the extracted pairs in the original time series. This procedure ensures that the quality of the new randomly-generated time series is as high as the original model-data pairs to be evaluated. The 95% confidence limits are then empirically computed from the distribution of the score values.

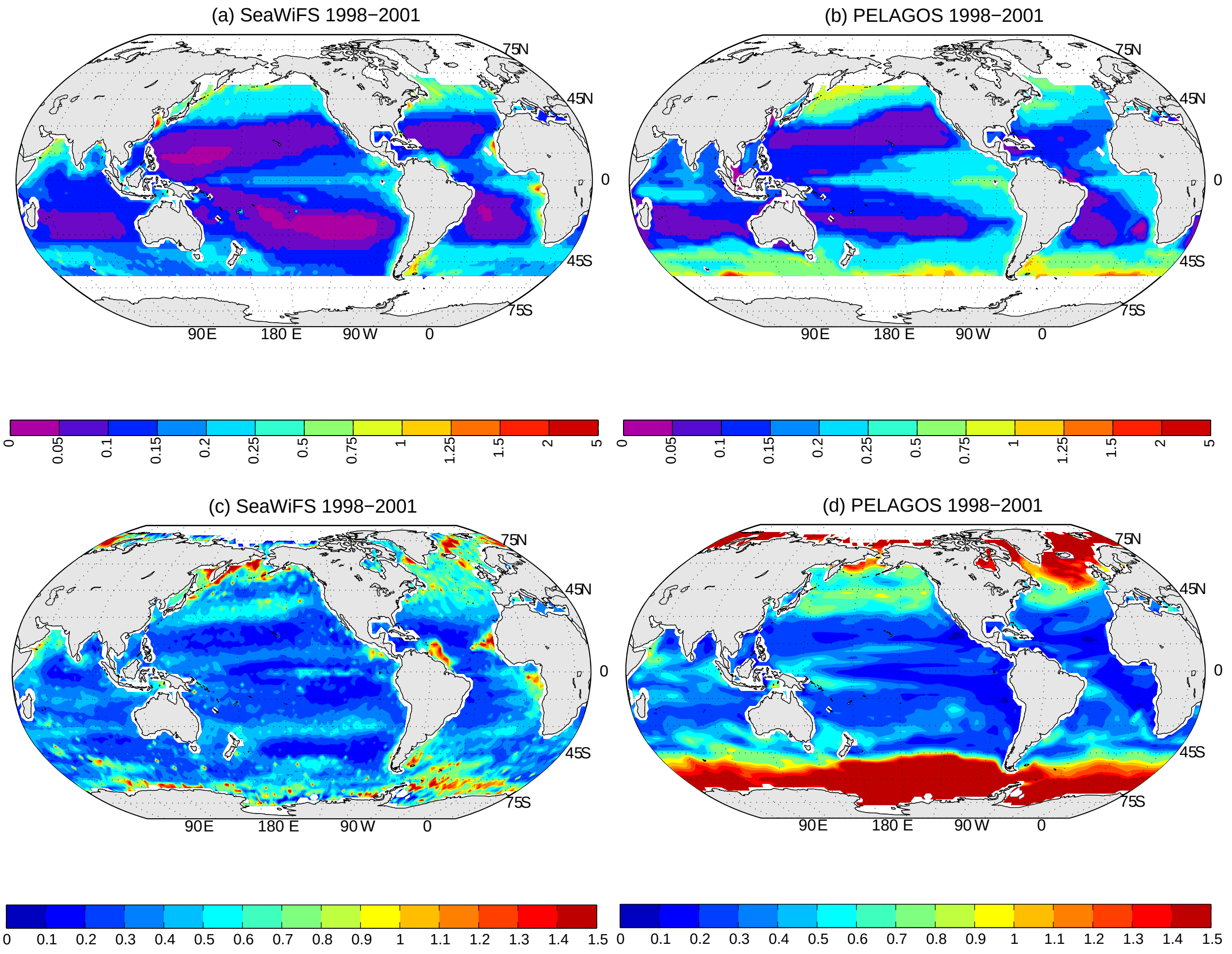

Fig. 1. Comparison of observed and simulated chlorophyll (mg chl m^{-3}) annual means and coefficient of variation (non-dimensional) for the period 1998–2001: **(a, c)** SeaWiFS; **(b, d)** PELAGOS (average over the euphotic zone depth).

3 Skill assessment

3.1 Global chlorophyll concentration

Figure 1 presents a visual comparison of simulated chlorophyll (chl) with satellite-derived estimates from SeaWiFS. Remotely-sensed chl represents the average chl concentration within one optical penetration depth, which in turn depends on the radiative extinction done by the chl concentration itself. When comparing model results with satellite data we need to consider both a data uncertainty, the 30% RMS log error described in Sect. 2.2, and the model representation of satellite chl, which may range between the simulated surface values and the vertically integrated chl concentration averaged over the euphotic depth (the latter shown in Fig. 1). For instance, in the center of the gyres and in coastal upwelling areas, the simulated surface values generally underestimate the observations more than the integrated values (not shown). The comparison is shown as annual means and variability. The variability was estimated with the anomaly coefficient of variation, which is the standard deviation of the anomaly fields computed with respect to and normalized by the monthly climatological averages.

The spatial distribution of the annual mean is visually captured in terms of maxima and minima. The modelled mean chl concentrations in the north Atlantic and Pacific have improved with respect to the climatological simulations in Vichi et al. (2007b), as well as the oligotrophic sub-tropical regions. Coastal maxima are instead unchanged with respect to the climatological runs, as the resolution of the ocean model did not change. The model is however capable of simulating the maxima of variation at the higher latitudes and the typical minima in the sub-tropical regions (Fig. 1d), with distinct intervals of high variability marking the borders of the Pacific cold tongue influence as found in the satellite data. The Northern Hemisphere higher-than-observed variability, particularly evident in the North Atlantic, is caused by too fast decrease of surface biomass after the spring bloom, a feature that has not improved from the climatological model results (Vichi et al., 2007b). The Southern Ocean is instead characterised by a marked positive bias, either in terms of variability and biomass.

A more objective way of looking at this latter feature is through the MEF index (Fig. 2). This index is extremely strict when applied to spatial fields, because it computes a point-to-point comparison on a reference grid. This analysis adds information on the seasonal dependence of the misfit that cannot be captured with maps alone. The misfits in the Northern Hemisphere and tropical ocean are stationary and linked to the seasonal cycle. The worst skill is found during the boreal winter while during summer the model is as good a predictor as the mean of the data (see Appendix A for the threshold limits of MEF). The lowest performances occur in the Southern Ocean during spring, due to a large simulated bloom in the frontal regions of the Southern Ocean.

The simulation of mixed layer depth (MLD) spatial and temporal evolution can partly explain the bias in the Southern Ocean. Figure 3 allows the visual comparison with objectively analyzed annual mean data (de Boyer Montégut et al., 2004). The MLD was evaluated with a temperature difference criterium of 0.2°C both in the data and in the model results. The mean annual mismatch in the Southern Ocean is extremely large, particularly south of 40° S as also evidenced in the zonal mean distribution (Fig. 3c). This discrepancy is mostly found during the onset of the stratification in October-November in the Subantarctic province (Fig. 3d; the Subantarctic province is defined according to Longhurst (1998) as the zone between the sub-tropical convergence around 35° S and the limit of the polar front at about 55° S).

3.2 Global primary production

The VGPM (Behrenfeld and Falkowski, 1997) is an empirical model that estimates net primary production (NPP) from satellite-derived chlorophyll using a temperature-dependent description of chlorophyll-specific photosynthetic efficiency. The NPP is thus computed using the observed chlorophyll concentration, SST and surface available light as inputs. The comparison with PELAGOS was done with a derived product of the VGPM (available at http://www.science.oregonstate.edu/ocean.productivity/) that implements the exponential dependence of production on temperature according to Eppley (1972). This exponential relationship is more similar to the one applied in PELAGOS (Vichi et al., 2007a).

This exercise is a model-to-model comparison (see also section below) therefore we focus here on maps of the annual means (Fig. 4) because they are expected to be better captured by the VGPM and satellite models in general (Campbell et al., 2002). The spatial variability is in fact very similar to the SeaWiFS chl variability (Fig. 1), since the major input of VGPM are the satellite-derived chlorophyll data. There is a good agreement in the spatial distribution of maxima, especially in the location of the frontal maximum in the Antarctic Circumpolar Current (AACC). This good result is obtained in spite of the large positive bias in the annual mean chl value (Fig. 1 and previous section). The coastal zone production is generally underestimated because of the low phytoplankton biomass. Nutrient availability plays a key role in these areas and the model is not able to fuel the surface coastal ocean with sufficient nutrient inputs. This is partly due to the resolution, as also evidenced by the mismatch in the subtropical and equatorial Atlantic and Indian Oceans, which are the smaller basins. The Mauritanian upwelling is in fact absent and the equatorial maximum in the Atlantic is closer to the South American continent than found in the VGPM estimates.

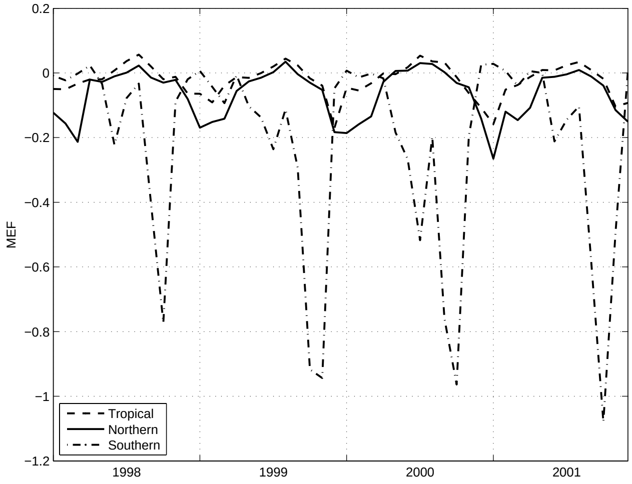

Fig. 2. MEF index for PELAGOS and SeaWiFS chlorophyll data over the period 1998–2001. The regions are defined as Tropical: 20° S–20° N; Northern Hemisphere: 20–60° N; Southern Hemisphere: 20–60° S.

3.3 Primary production in the equatorial Pacific

3.3.1 The ClimPP dataset

The comparison of model results with satellite-derived primary production is a valid assessment only if satellite-based NPP models are good in reproducing in situ observations. This kind of assessment was undertaken by the series of inter-comparison studies called Primary Production Assessment Round-Robin (PPARR, Campbell et al., 2002; Carr et al., 2006). In the latest published round, PPARR3, Friedrichs et al. (2009) collected a set of observations from the equatorial Pacific that were used as benchmarks for the reality check of satellite-based production models (SatPPM) and OBGCMs. One of their conclusions is that current state-of-the-art SatPPM are only slightly more skillful then prognostic OBGCMs and that the actual dominance depends on the choice of the assessment score. Some models are better than others in terms of bias, while others are better in terms

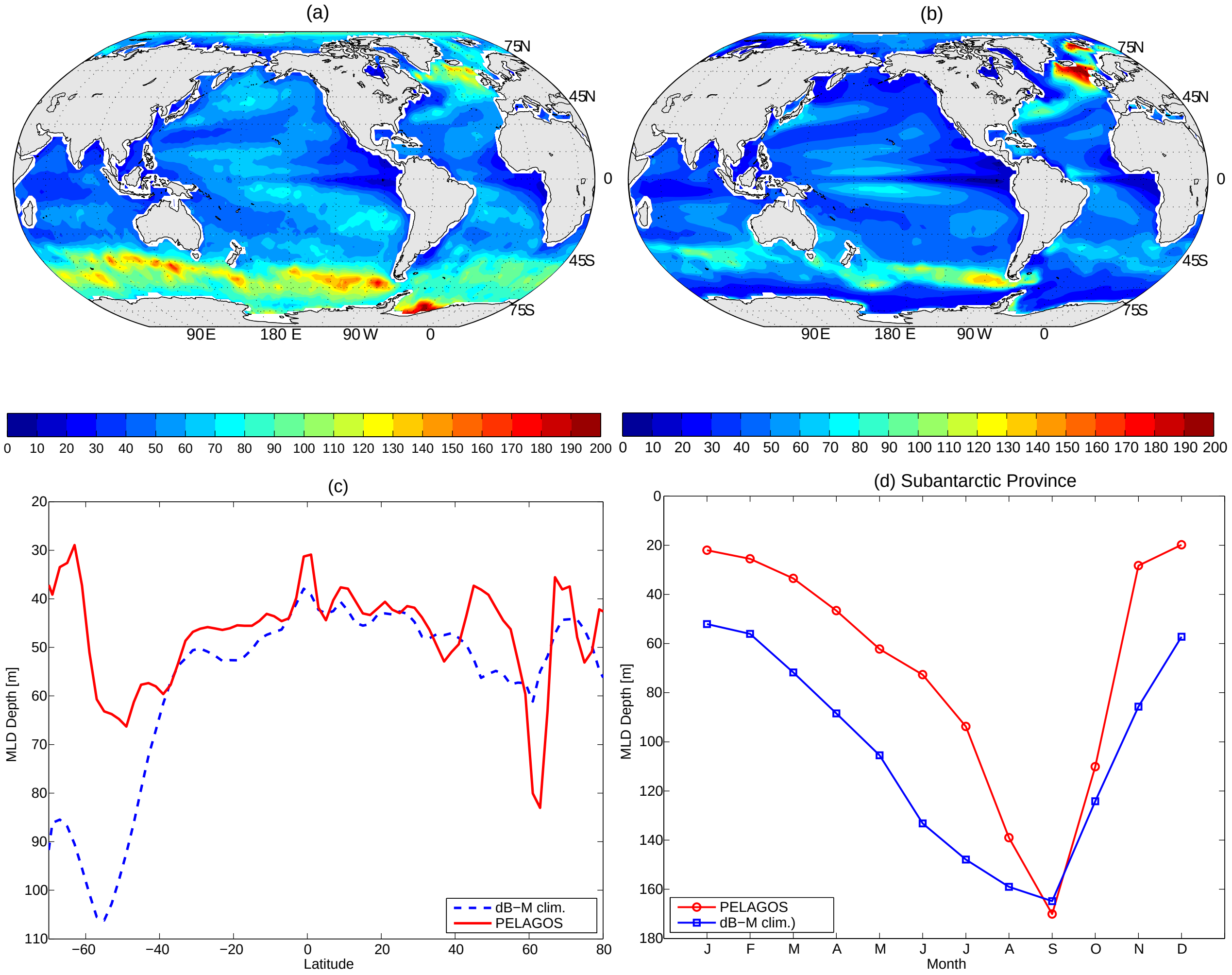

Fig. 3. Comparison of observed and simulated mixed layer depths (MLD in m, temperature criterion ΔT=0.2°C). **(a)** de Boyer Montégut et al. (2004) **(b)** PELAGOS means over the whole simulation period; **(c)** comparison of zonal averages **(d)** mean seasonal cycle in the sub-antarctic province.

of variability. Incidentally, one of the major outcomes of this project is the public availability of this quality-checked dataset, which can thus be used to test model performances. The dataset consists of vertically-integrated euphotic zone measurements of in situ primary production with a standardized ^{14}C methodology, combined with SST, chlorophyll and ancillary model data for MLD and surface irradiance, covering the whole tropical Pacific over the period 1983–1995. The NPP data are originally available in units of daily carbon production per unit area. A point-to-point comparison was done according to the same protocol described in Friedrichs et al. (2009) using monthly mean model data, and the assessment was done by considering the same suggested set of performance indicators.

3.3.2 PELAGOS results

Net primary production in PELAGOS is parameterized as a function of light, chlorophyll, iron cell-content and dissolved silicate concentration (Vichi et al., 2007a). The cell availability of N and P does not directly control photosynthesis, but the subsequent transformation of carbohydrates into proteins and cell material. A portion of photosynthesized carbon is therefore released as Dissolved Organic Carbon (DOC) exudates according to the internal nutrient quota (Baretta-Bekker et al., 1995; Vichi et al., 2007a).

During in situ incubation experiments performed with radiocarbon techniques, the contribution of this DOC part is only partly measured because of the filtering procedure that removes particles smaller than (usually) 2 μm. The colloidal High Molecular Weight (HMW) portion is likely to remain attached to cells, while the remainder Low Molecular Weight (LMW) part is released in the water. The percentage of colloidal HMW DOC during a bloom is estimated around 20% (Kepkay et al., 1993) while an overall bulk figure of LMW DOC varies from 65 to 80% (Ogawa and Tanoue, 2003), the highest loss likely to occur in oligotrophic waters. From a modelling point of view this distinction is not necessary, because the relevant input of organic carbon is the NPP computed as photosynthetic production minus the metabolic and/or activity respiration losses. Nevertheless, an unknown portion of the NPP that is lost to DOC should be removed when comparing the model output with data. To account for this effect, two different estimates of NPP have been used in the comparison: NPP1 is the total amount of organic carbon produced by autotrophs, while NPP2 considers an estimated conservative loss of 50% of the time-varying DOC production rate. The overall estimate of ClimPP parameters is presented in Table 1 and a graphical misfit analysis (Stow et al., 2009) of NPP is shown in Fig. 5.

The SST scores are very good, as it was expected because of the relaxation to observed data described in Sect. 2.1. Simulated MLD is also in accordance with the model data provided in the ClimPP datasets (MLD is on average deeper but with similar standard deviation). Friedrichs et al. (2009) report that the ClimPP MLD data are in the ±20 m range with respect to the JGOFS equatorial Pacific Process Study cruises, therefore our results fall within the same range. The low MEF however indicates that the two models give different time evolution and thus further independent data are needed to assess the quality of the vertical structure of the model with respect to the equatorial Pacific conditions. MLD in the equatorial Pacific is probably not the best indicator for production as the satPP models using MLD are not as skillful as the others. OBGCMs like PELAGOS use more information than the MLD and are capable to obtain better scores because they do not rely on this variable only.

Chlorophyll skill is good only regarding the indicators of average concentration. Bias and average errors are small, but the simulated standard deviation is much higher than observed and also the MEF indicates a poor if not bad predictive performance. NPP scores are instead much better than chlorophyll and in line with the results of the other PPARR3 models as further shown below. The NPP2 estimate of PP improves all the performance indices (Table 1): for instance, the bias is much reduced with respect to NPP1 and consequently the total RMSD. It is interesting to note that also some indices of variability improve, such as the standard deviation and the correlation coefficient.

A possible way to show the combined skill of PELAGOS in the framework of the other PPARR3 models is presented in Fig. 6 with the aid of the MDS ordination (Sect. 2.2). The two-dimensional distances between the multivariate set of indices is well represented (stress is close to 0) . The ClimPP data point is included in the ordination by assuming a set of indices with the highest score values (e.g. r=1, B=0, etc.). Two additional artificial data points that represent the worst cases have been added. They are obtained by combining the worst scores from all the models and by taking the highest and the lowest standard deviation values. This picture clearly shows that the NPP estimates from satPP models and OBGCMs are indeed comparable as initially suggested by Friedrichs et al. (2009). All the models lie at approximately the same distance from the data, the closest being a set of satPPMs. PELAGOS NPP2 is much better than NPP1 because it is located in the cluster of the best OBGCMs and satPPMs.

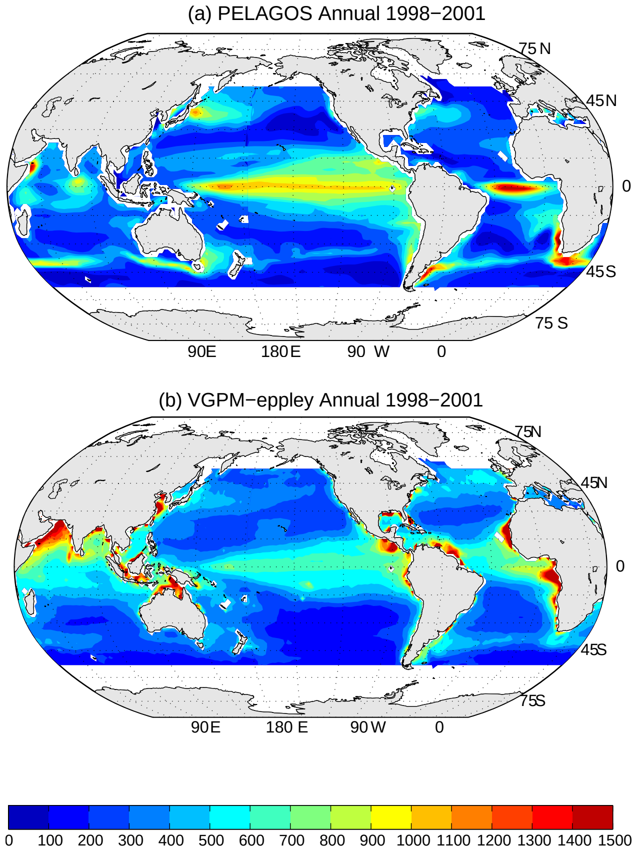

Fig. 4. Annual mean NPP over the euphotic zone in mg C m^{-2} d^{-1} for the period 1998–2001: **(a)** PELAGOS, **(b)** VGPM with Eppley temperature dependence.

The direct visual comparison of ClimPP and PELAGOS data and misfits (Fig. 5, NPP2 results) provides additional information on the goodness-of-fit of model results. The data show a small positive trend as reported by Friedrichs et al. (2009), though this is not statistically significant. The model-data misfit (Fig. 5b) apparently decreases with time though the model does not have a trend (Fig. 5a). This tendency suggests that observed primary production in the ’80s was relatively lower than the one in the ’90s. A similar evolution can be found in the model results if the cluster of high NPP after 1988 is not considered (not shown). The model in fact overestimates the production during and after the 1988 La Niña event.

Table 1. Skill assessment indices for the ClimPP dataset. The units apply to all indices except the correlation coefficient and MEF that are non-dimensional. Skill measures are defined in the Appendix A.

	SST [deg C]	MLD [m]	Chlorophyll	$\log_{10}$PP (NPP1)	$\log_{10}$PP (NPP2)
Pearson r	0.90	0.68	0.58	0.50	0.58
$\mathrm{RMSD_{tot}}$	0.89	19	0.11	0.38	0.27
B	−0.22	−13	−0.02	0.29	0.11
AAE	0.64	15	0.08	0.31	0.21
RMSD_{cp}	0.85	14	0.10	0.24	0.24
s.d. (PELAGOS)	1.96	15	0.12	0.17	0.25
s.d. (ClimPP)	2.03	19	0.01	0.28	0.28
MEF	0.81	0.02	−0.10	−0.82	0.10

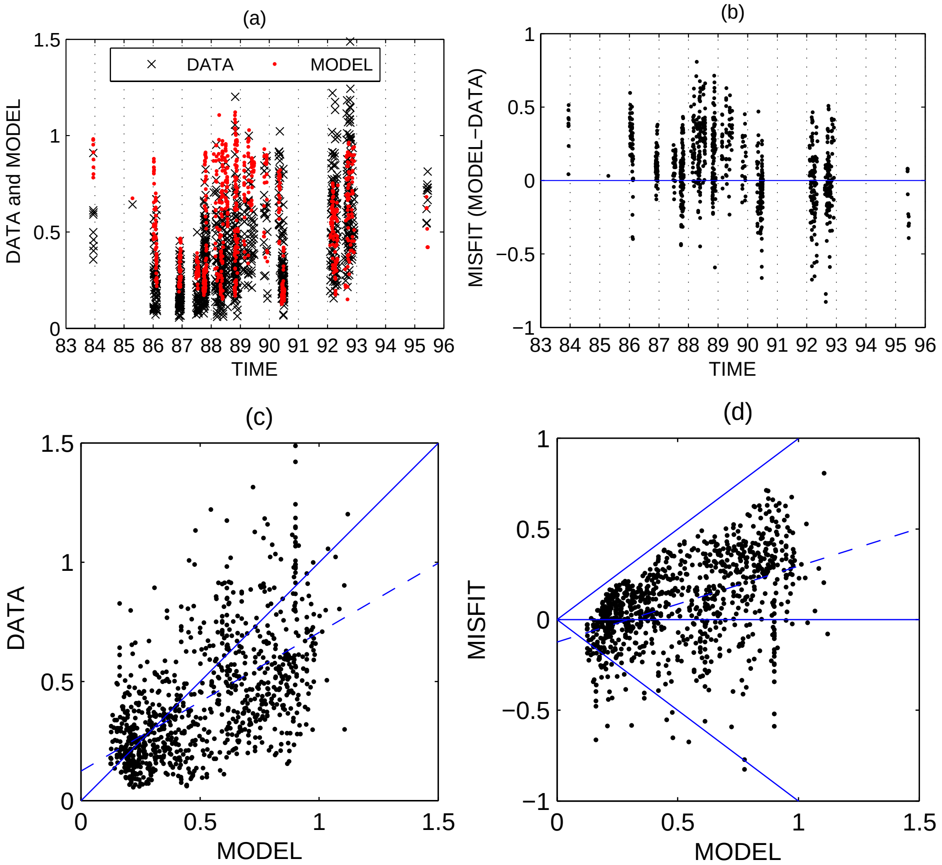

Fig. 5. Comparison of observed and simulated NPP values in g C m^{-2} d^{-1} for the data collected in the ClimPP dataset. **(a)** data and model time-series; **(b)** model misfit versus time; **(c)** data and model scatter plot with the 1:1 line and regression line $\hat{y}_D$=0.58y_M+0.12, r^2=0.33; **(d)** model misfit versus model (the continuous lines represent the 0 misfit value and the 1:1 lines where the misfit is equal to the model value; the regression line is dashed).

The goodness-of-fit between model and data values can be objectively assessed by means of linear regression on the scatter plot (Fig. 5c). The slope of the regression line is significantly different from 1 and the coefficient of determination is r^2=0.33. The lack of fit is however not caused by the bias, but mostly by unexplained variance and partly by the different slope. This confirms the results of Table 1 performed on the log-transformed data. The high dispersion might be due to natural variability in the observations, although it is also a possible indication of misspecification in the model. The misfit in fact increases with increasing model values (Fig. 5d) with a majority of overestimation. The hypothesis H_0 that the regression slope with coefficient 0.42 is equal 0 can be significantly rejected with p<0.01.

3.4 Multi-annual assessment at selected locations

The number of available long term stations is rather limited and the choice of the two major JGOFS stations in the Atlantic (Bermuda) and Pacific (station ALOHA) is almost mandatory as locations for model calibration and hypothesis testing (e.g. Hurtt and Armstrong, 1996; Spitz et al., 2001; Huisman et al., 2006; Brix et al., 2006). In this section we have focused on the set of data provided at the end of JGOFS by the modelling and synthesis group because they reflect a coherent dataset based on uniform unit and method conventions for both stations.

3.4.1 The JGOFS Station BATS

BATS lies at the western boundary of the north Atlantic subtropical gyre and, though being stratified for most of the time, it experiences winter mixing events as deep as 200 m. This seasonal signal is found in NPP data and well captured by the model (Fig. 7). The linear regression indicates goodness-of-fit because the value of the slope is statistically equal to 1 (Fig. 7b, the equation is given in the caption). The major discrepancy with the data is due to the bias, which is mostly found during summer when our model simulates higher than observed NPP, even if most of the previous modelling studies reported underestimation during this period (Brix et al., 2006). There is however a distinct difference between the surface and the vertically-integrated NPP values: the model overestimates NPP in the euphotic zone and underestimates the surface value (not shown). This implies that PELAGOS simulates a more productive deeper community during summertime with respect to the previous modelling results. These results are obtained with the NPP2 estimation (cf. Sect. 3.3.2), but since during summertime more colloidal DOC is exudated because of the oligotrophic conditions, it is likely that considering a constant HMW proportion in DOC when comparing with data is only partly sufficient. This issue is further discussed in Sect. 4.2.

Primary production peaks are linked to the maxima in MLD (Fig. 8) both in the data and in the model (linear correlation coefficient r=0.80 and r=0.65, respectively). The lag-correlation analysis performed on the model results shows a peak at 1 month and remain larger than 0.4 for an interval of about 3 months. This implies that production starts when the mixed layer is still deep and the peak of production is reached after the onset of stratification. MLD is visually well predicted by the model although the scatter plot (Fig. 8b) is not as significant as for NPP due to the wintertime bias. The underestimation of NPP during winter is likely due to the underestimation of MLD since the misfits are linearly correlated (r=0.56). The model simulates the NPP inter-annual variability quite well, particularly when linked to distinct physical features. This occurs for instance during the low-production event of winter 1994 when the observed MLD is shallower than other years and the model is able to simulate it correctly.

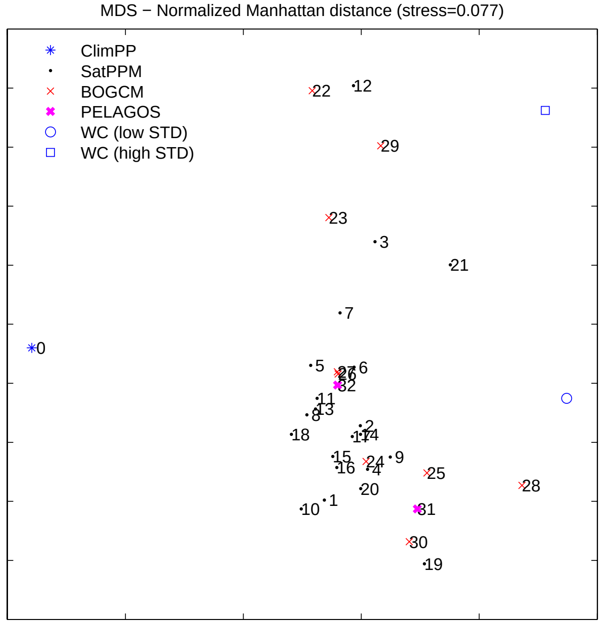

Fig. 6. MDS representation of the combined performance indices for the models participating to the PPARR3 and for PELAGOS. Numbering and classification (SatPPM and BOGCM) is done according to Friedrichs et al. (2009) with the addition of PELAGOS results (NPP1=31; NPP2=32) and the data (0). The points marked with WC represent artificial worst cases derived from the model results space and by assuming the lowest and highest standard deviations with respect to the data.

The values of the performance indicators at BATS indicate good skill of the model (Table 2). Means are correctly predicted and within the standard error for all the variables except nutrients, which are close to the detection limits in the observations and very close to 0 in the model. The linear correlation coefficients for both chl and NPP are high and significant (p<0.01) with confidence intervals of 0.50–0.73 and 0.53–0.78, respectively (both computed with the bootstrap method, Sect. 2.2). The MEF index is larger than 0 for NPP only (p<0.01, confidence interval 0.27–0.58), which is classified as a “good” score (see Appendix A). The RI is instead rather high due to the summer overestimation, implying that on average, the spread of the predicted production is large and can get more than twice as high than observed.

Bacterial biomass and production (BP) were compared with the observed surface values due to the larger data availability at this level. Biomass values were converted from cell counts using the cellular carbon content suggested by Gundersen et al. (2002, 10 fg C/cell). BP was converted from thymidine incorporation hourly rates into daily carbon production by means of the conversion factor suggested by

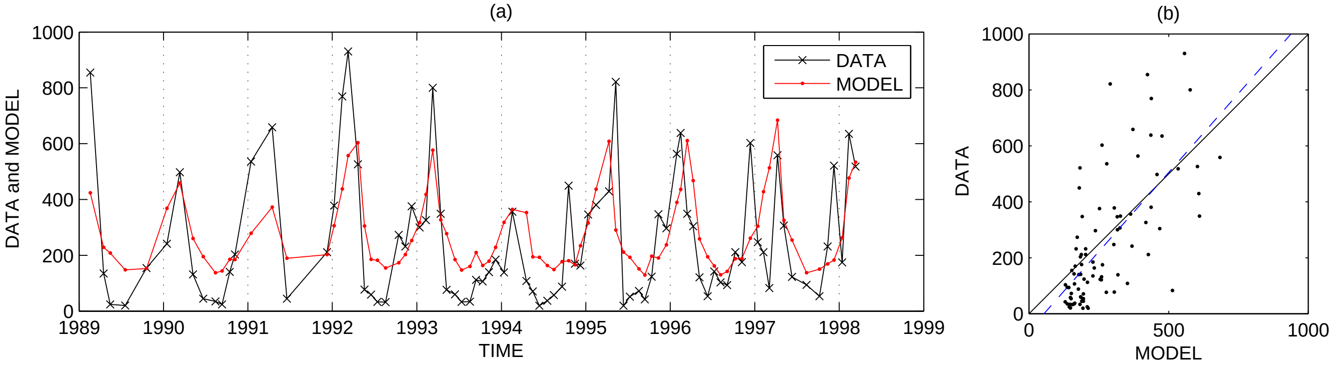

Fig. 7. Comparison of observed and simulated integrated primary production (in mg C m^{-2} d^{-1}) at BATS. **(a)** JGOFS BATS time series; **(b)** scatter plot with regression line $\hat{y}_D$=1.1y_M−60, r^2=0.46 (H_0: slope=1 cannot be significantly rejected, p=0.32).

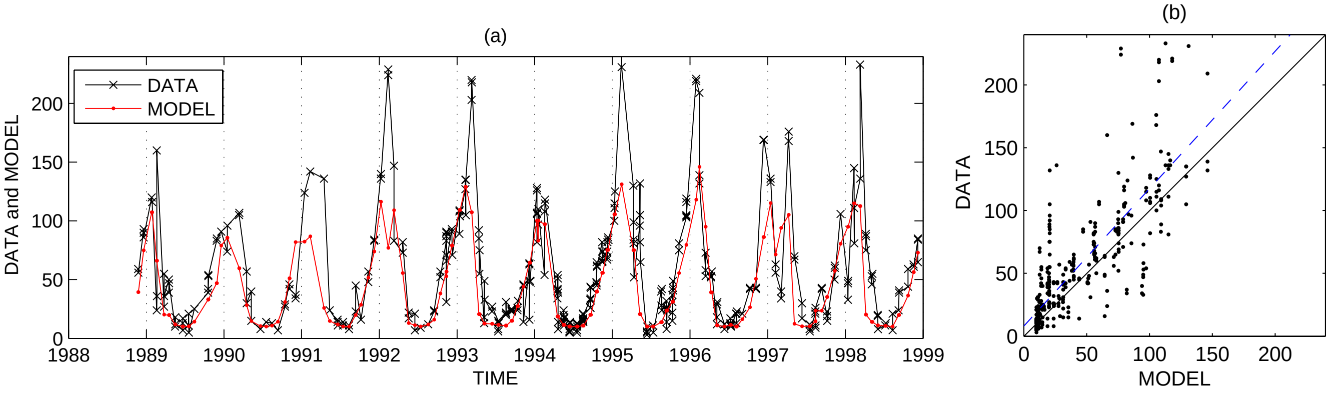

Fig. 8. Comparison of observed and simulated mixed layer depths (in m) at BATS. **(a)** JGOFS BATS time series; **(b)** scatter plot with regression line $\hat{y}_D$=1.1y_M+8, r^2=0.66 (H_0: slope=1 can be rejected according to the choice of the reference p-value since $p=0.02$).

Fuhrman and Azam (1982) and by considering a multiplicative factor of 12, which is in the range of the ratios between leucine and thymidine incorporations as measured by Ducklow et al. (2001) in the Arabian Sea. The mean bacterial biomass is well simulated by the model, as confirmed by the very low bias and RI value close to 1 (Table 2). The model is however not capable to simulate the variability (not shown), which is partly linked to the seasonal production. This is shown here by the high values of RMSD$_{cp}$ and the absolute average error. There is a small linear phase correlation (confidence interval 0.11–0.46, p<0.01) which is likely caused by the presence of a weak seasonal signal both in data and model, although the MEF index is still close to 0 confirming that the model can only capture the mean value. Similar considerations can be done for BP, which has no bias at all but is characterized by a higher RI due to the low variability predicted by the model. The choice of a constant scaling factor used to convert thymidine incorporation into daily production may also play a role, because in some seasons it can be 4 times as high as considered here (Ducklow et al., 2001). The linear correlation is slightly higher than for the biomass (confidence interval is 0.21–0.46, p<0.01) and the MEF is also positive (c.i. is 0.05–0.18, p<0.01) although still poor according to the indicative thresholds given in the Appendix A.

3.4.2 The JGOFS station ALOHA (HOT)

Station ALOHA is located in the sub-tropical Pacific Ocean north of Hawaii. It is characterized by a permanently stratified water column with a deep chlorophyll maximum (DCM) below 100 m, mostly composed of *Prochlorococcus* with temporary outbursts of diatoms and dinoflagellates (Karl et al., 2003a). The surface ocean is depleted in nutrients and the mixed layer depths occasionally reaches the location of the DCM during wintertime.

PELAGOS results at this station capture the typical low production conditions but miss the observed higher frequency variability (Fig. 9). The tendency of models to predict low NPP at this station has been reported by other authors (e.g. Ondrusek et al., 2001). The observations are bounded by the NPP1 and NPP2 estimates, which suggests that the choice of the fraction of colloidal DOC at Sta. ALOHA may not be the same as estimated with the ClimPP

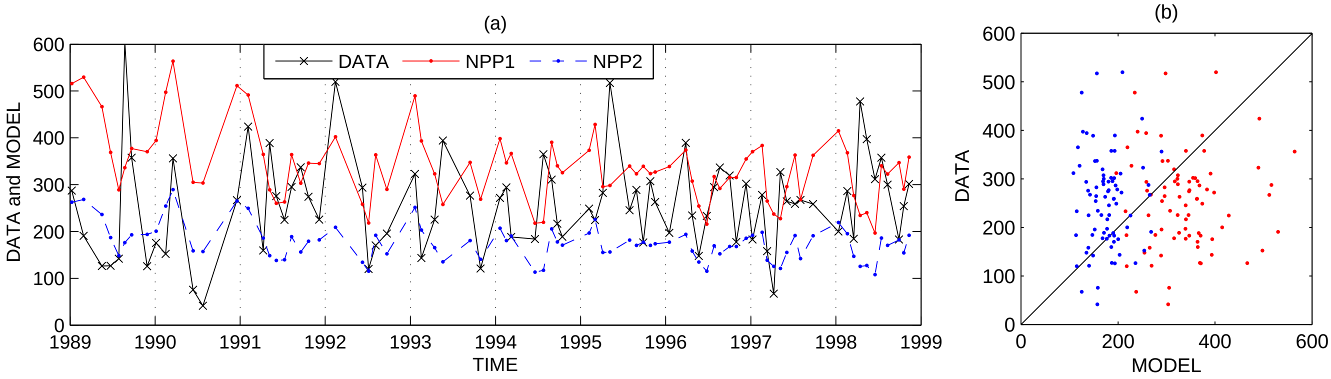

Fig. 9. Comparison of observed and simulated integrated primary production (mg C m^{-2} d^{-1}) at Sta. ALOHA. **(a)** JGOFS HOT time series for the NPP1 and NPP2 estimates (cfr. Sect. 3.3.2); **(b)** scatter plot of model vs. data for both estimates.

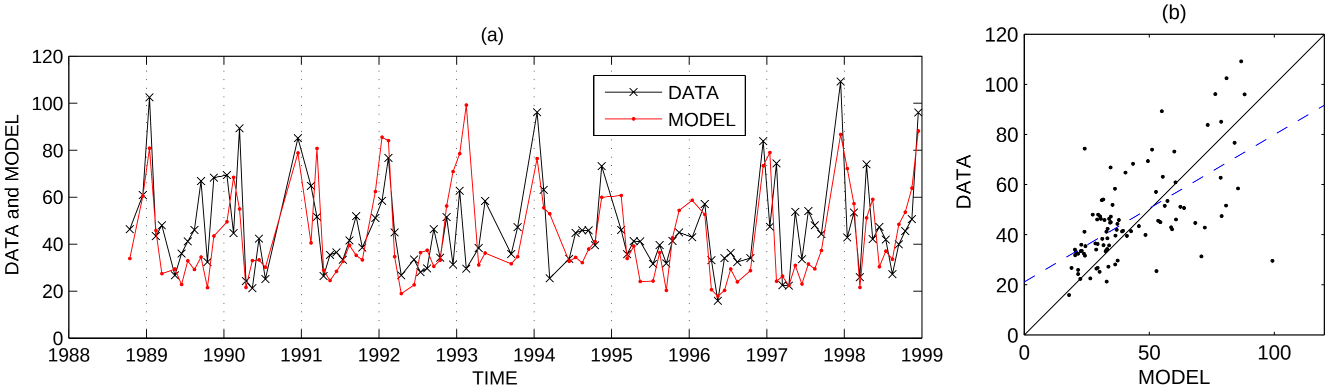

Fig. 10. Comparison of observed and simulated mixed layer depths (in m) at Sta. ALOHA. **(a)** JGOFS HOT time series; **(b)** scatter plot with regression line $\hat{y}_D$=0.59y_M+0.21, r^2=0.38 (H_0: slope=1 can be significantly rejected).

dataset and BATS (cf. Sect. 4.2). Both NPP1 and NPP2 however show a lack of fit as indicated by the scatter plot in Fig. 9b.

The MLD evolution is well-reproduced by the model (Fig. 10), but there is a less clear relationship with NPP as seen for instance at BATS (r=0.46 both in observations and model data; the lag-correlation in model data is higher with one month lag, r=0.57). It is likely that the NPP variability here is driven by small scale features, episodic but seasonally-recurrent nitrogen fixation events (Dore et al., 2008) or possibly by chaotic fluctuations (Huisman et al., 2006). None of these features can be reproduced by the current model design, either due to the coarse horizontal resolution or because of the absence of diazotrophs that are currently not considered in PELAGOS.

The values of the performance indicators of biological variables are less good than at BATS (Table 3) although SST and MLD are very well predicted both in terms of magnitude and variability. Chl is on average one third of the observations but standard deviations are comparable and the linear correlation is statistically significant (c.i. 0.17–0.59, p<0.05). The same considerations apply for bacterial biomass, whose bulk value is well predicted. The number of bacterial data at HOT is however smaller than at BATS, which makes the linear correlation and RI values less significant (p<0.1, c.i. −0.13∼0.57, RI=1.3∼1.4, respectively). The indicators of the standing stocks are thus generally acceptable but the NPP is underestimated. If we consider the NPP1 estimate of autotrophic production, PELAGOS overestimates the measured NPP with a mean value of 337 and s.d.=77. In this case the total amount of autotrophic carbon production is more similar to the values suggested by Ondrusek et al. (2001).

4 Discussion and applications

We presented here a selected set of results from a multiannual simulation of PELAGOS compared with publicly available global and local datasets. The technical aim of the previous part is to design a layout for model assessment, building on the general prescriptions suggested in Stow et al. (2009). The scientific aim is to demonstrate that this model

Table 2. Model skill assessment at BATS. The units apply to all indices except the correlation coefficient, MEF and RI. Variables refer to the surface except MLD and NPP, which is integrated over the MLD in the data and to the euphotic zone depth in the model (∼90 m). n.s.=non significant according to an empirical p-value estimate. Skill measures are defined in the Appendix A.

	SST [deg C]	MLD [m]	Chlorophyll [mg m^{-3}]	NPP [mg C m^{-2}d^{-1}]	Phosphate [mmol m^{-3}]	Nitrate [mmol m^{-3}]	Bacteria [mg C m^{-3}]	BP [mg C m^{-3} d^{-1}]
mean	23.0	43	0.14	275	$8\ 10^{-4}$	0.19	4.96	2.33
mean BATS	23.5	55	0.10	251	$8\ 10^{-3}$	0.04	4.58	2.33
s.d.	2.8	36	0.12	136	0.002	0.009	0.38	0.47
s.d. BATS	2.9	49	0.08	227	0.02	0.11	1.15	1.84
Pearson *r*	0.96	0.81	0.64	0.68	n.s.	n.s.	0.29	0.35
RMSD$_{\mathrm{tot}}$	0.94	31	0.10	169	0.02	0.18	1.16	1.73
B	−0.56	−12	0.04	24	−0.007	0.15	0.38	0.00
AAE	0.74	19	0.07	132	0.009	0.17	0.90	1.27
RMSD$_{cp}$	0.75	29	0.09	167	0.02	0.11	1.10	1.73
MEF	0.90	0.59	−0.80	0.44	−0.14	−1.6	−0.03	0.11
RI	1.0	1.7	2.2	2.5	n.s.	n.s.	1.4	2.2

Table 3. Model skill assessment at HOT. The units apply to all indices except the correlation coefficient, MEF and RI that are non-dimensional. n.s.=non significant according to an empirical p-value estimate. Skill measures are defined in the Appendix A.

	SST [deg C]	MLD [m]	Chlorophyll [mg m^{-3}]	NPP [mg C m^{-2}d^{-1}]	Phosphate [mmol m^{-3}]	Nitrate [mmol m^{-3}]	Bacteria [mg C m^{-3}]
mean	24.5	43	0.04	176	$2\ 10^{-4}$	$3\ 10^{-4}$	4.58
mean HOT	24.8	47	0.09	259	0.08	$5\ 10^{-3}$	4.17
s.d.	1.38	20	0.02	39	$5\ 10^{-4}$	$4\ 10^{-4}$	0.07
s.d. HOT	1.27	19	0.04	103	0.03	0.01	0.90
Pearson *r*	0.95	0.62	0.41	n.s.	n.s.	n.s.	0.24
RMSD$_{\mathrm{tot}}$	0.53	17	0.06	136	0.08	0.01	0.96
B	−0.33	−3	−0.05	−83	−0.07	−0.004	0.41
AAE	0.42	13	0.05	104	0.08	0.005	0.79
RMSD$_{cp}$	0.42	17	0.03	109	0.03	0.01	0.87
MEF	0.83	0.16	−1.9	n.s.	n.s.	n.s.	−0.18
RI	1.0	1.4	3.2	n.s.	n.s.	n.s.	1.2

can be used as a component of an Earth System Model to study climate change scenarios where modifications of the ocean large-scale features are expected to affect marine biogeochemical dynamics. It is therefore essential to assess the quality of the PELAGOS simulation in the context of the 20th century climate conditions. The debate about the meaning of validation in ecological models is still open (e.g. Rykiel, 1996) and it is not our aim to focus here on the more phylosophical questions of whether a model can ever be verified or only falsified. Our interest is in the demonstration of operational validity through the usage of objective indicators of performance. This implies the subjective judgment about the skill of the model (Lynch et al., 2009), which includes the definition of acceptable thresholds. The target is the description of marine carbon transformation rates under current climate conditions and therefore we are interested in describing the major distribution properties such as means, standard deviations and phase correlation. This implies that values of RI$<$2 and MEF$\geq$0 are the minimum requirements (see Appendix A for definitions), phase correlations should be significant and $\geq$0.4 and that goodness-of-fit tests with linear scatter plots should give $r^2\geq$0.6 and slope values of 1. Bias and RMSD values are very informative properties to qualify model behavior, but the definition of validity thresholds is even more based on subjective considerations and was not decided a priori. However, as suggested by Rose and Smith (1998), it is more scientific to debate measures and thresholds than graphical comparisons.

The problem of sample size is also relevant in case of biogeochemical data. We make here an additional substantial assumption that cannot be assessed yet: the data used in this work are considered sufficient to describe the mean state and variability of the current climate conditions from a biogeochemical point of view. This is a rather strong assumption given the little amount of data concerning carbon exchange rates and the lack of repeated measurements that are needed to define variability in a climatological sense. It is more likely to get good scores by accident with limited set of data

(Mason, 2008), such as the monthly observations in the HOT and BATS timeseries.

Scoring model performance is finally very useful in the context of model development or in the framework of multi-model comparison. These values can be used as benchmarks to check the efficacy of a new component addition, parameterization changes or newly available data. The heuristic nature of biogeochemical models implies that model formulations are parameterized on specific datasets or derived from general considerations on ecosystem functioning. Testing their genericity against a set of benchmarks as proposed here is therefore one possible way forward for building more robust formulations (see also discussions in Hood et al., 2006).

4.1 Analysis of major biases

The simulated mean chlorophyll field is visually acceptable in the northern mid-latitude regions and equatorial Pacific, but is markedly overestimated in the Southern Ocean (cfr. Fig. 1). This feature is found in other OBGCMs, either linked to the role of iron in this High-Nutrient Low-Chlorophyll (HNLC) region or to inadequate mixing (e.g. Schneider et al., 2008). The use of multi-annual atmospheric forcing functions instead of climatological have improved the model results especially in the Northern Hemisphere. However, the improvements in the North Atlantic and equatorial Pacific with respect to the simulations in Vichi et al. (2007b, especially in terms of diatoms, not shown) led to an increase of the chl bias in the Southern Ocean. We hypothesize that the early stratification in the Subantarctic province (Fig. 3d) favours the bloom of diatoms, which can maximise production through photo-acclimation in the illuminated shallow MLD utilizing the abundance of nutrients typical of this region. A sensitivity analysis has been performed to demonstrate the linkage between biomass production in the Southern Ocean and the seasonal cycle of the mixed layer. The results shown in Fig. 11 for year 2000 indicate that simulated chlorophyll is substantially reduced by artificially increasing of 2 orders of magnitude the minimum level of turbulent kinetic energy throughout the year in the Southern Ocean south of 50° S. Consequently, also the MEF index for the Southern Hemisphere shown in Fig. 2 increases to values close to 0 as for the other regions (not shown). This is however only a proof of concept and further work is needed both on the physical and biological components of the model. It is known that the Southern Ocean experiences local intense mixing events (Garabato et al., 2004) and that current MLD climatologies are inadequate to provide robust estimates of this region due to data limitations (de Boyer Montégut et al., 2004). Recently, a new climatology derived from ARGO float data indicated that the deepest mixed layers are located in the Subantarctic province, with maxima from June to October (Dong et al., 2008), similarly as obtained in Fig. 11. This experiment demonstrates that mixed layer depth controls the simulated annual evolution of phytoplankton in the Southern Ocean and that the energy transfer paramaterizations in the subsurface should be improved in the future in order to reduce the systematic bias in model results. It is also important to avoid overconstraints of parameters such as iron-limitation coefficients to force the model towards the observed concentrations.

Simulated primary production is comparable with satellite estimates at the global scale, though only a qualitative visual comparison has been performed in this case. In fact, the usage of satellite-derived products for model assessment is a model-to-model comparison and not an operational validation with observations. As demonstrated in Sect. 3.3.2 and also pointed out in Friedrichs et al. (2009), the predictability skill of satellite PP models and OBGCMs against in situ, quality-controlled observations is comparable, at least in the equatorial Pacific. This latter comparison has shown that PELAGOS has better predictability than the mean of the data (MEF$>$0), which implies that it can reproduce the long-term observed mean and part of the variability which is related to the climate variability in the region (not shown).

The independent test with the two long-term JGOFS stations increases the confidence in the model parameterizations of carbon cycling in the surface ocean. The model was not calibrated to the observations, therefore the skill in capturing the observed means is remarkable. All the computed scores are significant and indicative of some skill, mostly in predicting the mean state. Variability, measured as linear correlation and standard deviation, is well captured more at BATS than at HOT because of the presence of a more marked seasonal cycle. The observations at HOT indicates the presence of short-term and small-scale sources of variability that are unlikely to be simulated by the coarse resolution ocean model in a generally stable area such as the subtropical north Pacific gyre. The long-term means of observed nutrient concentrations is an order of magnitude higher than in the model simulations both at BATS and HOT, although all values are already very close to the detectable limits. In particular, phosphate is reported to be 0 for most of the sampled data in the JGOFS time series. However, the high observed s.d. indicates the presence of nutrient pulses, which at BATS occur during summer and not during the more mixed periods (not shown).

If we consider the MEF and RI indicators as overall measures, we may conclude that the model has skills for simulating primary production in the equatorial Pacific and in sub-tropical mid-latitude regions such as BATS. The latter occurs in spite of the poor performance in predicting nutrient concentrations, which implies that from a functional point of view there is a nutrient threshold below which carbon production can be quantitatively simulated unregarding the kind of the limiting nutrient. There is thus more work to be done for parameterizing multiple nutrient limitations in models , though the influence on primary production estimates are second order in PELAGOS.

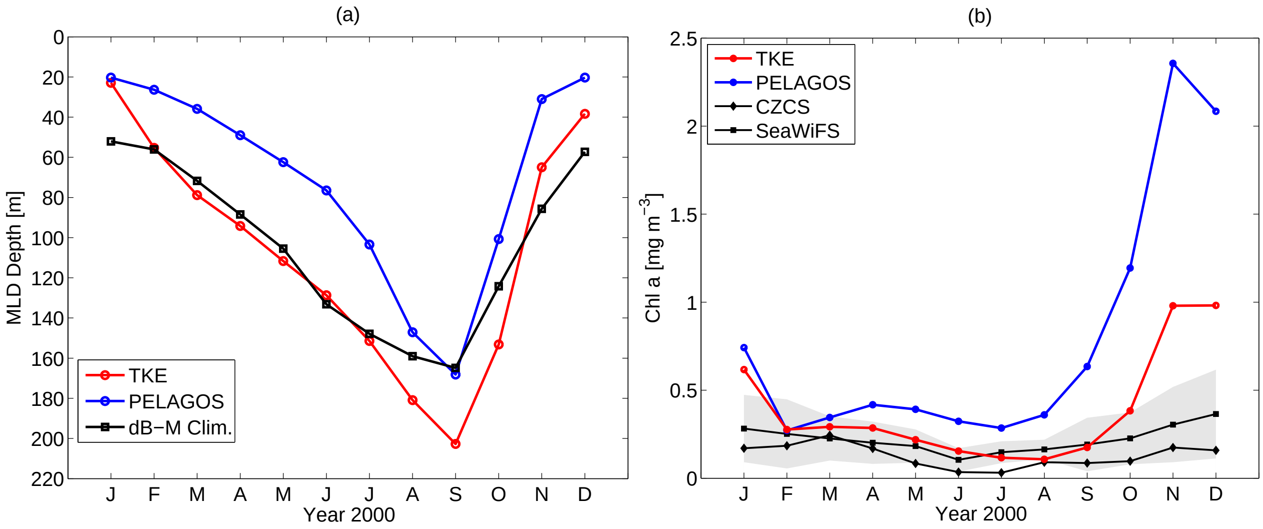

Fig. 11. Results of a sensitivity experiment on the artificial increase of TKE in the Southern Ocean (shown for year 2000). **(a)** Comparison of the seasonal cycle of mixed layer depths in the Subantarctic province with the observed climatology and reference simulation shown in Fig. 3d; **(b)** resulting mean chlorophyll concentrations and comparison with SeaWiFS and CZCS data. The gray-shaded area is the spatial standard deviation of SeaWiFS in the province. (b) can be directly compared with Fig. 9c in Vichi et al. (2007b).

4.2 DOC and primary production

The usage of objective skill indicators allowed us to demonstrate that NPP scores improve when the model variable is diagnosed by estimating net particle production (NPP2) and not the more typical difference between gross production and respiration losses (NPP1, Sect. 3.3.2). In fact, the removal of a constant portion of the produced carbon that is directly released as DOC improves the comparison with the ClimPP data set (Table 1 and Fig. 6) and also at BATS (not shown). It is known that a considerable fraction of primary production may be lost directly as dissolved organic carbon in nutrient-stressed conditions (Ogawa and Tanoue, 2003). Recently, a paper comparing 8 different methods of measuring primary production highlighted the role of dissolved organic matter, which may lead to experimental underestimates of ^{14}C NPP especially in the case of nutrient-stressed cells (Robinson et al., 2009). Our results indicate that considering this fraction when comparing with ^{14}C in situ primary production estimates considerably improves the results.

However, the choice of a constant fraction that fix the proportion between HMW and LMW DOC is still insufficient. In the more oligotrophic HOT data, the observations lie between the two NPP estimates (Fig. 9). If the amount of colloidal LMW DOC produced at Sta. ALOHA is higher than at BATS due to the more oligotrophic conditions, it is likely that a fraction higher than 50% (as estimated with the ClimPP dataset and used with BATS, cfr. Sect. 3.3.2) be retained by the filter and thus considered as particle production. Concurrent comparisons of ^{14}C NPP, DOC quality and oxygen production fluxes should help to clarify further this issue. Our experiments suggest that a dynamical parameterization of the quality of the exudate production may contribute to a more proper estimation of the observed production and export rates, since colloidal DOC may also increase the sinking velocity of organic matter through aggregation (e.g. Engel et al., 2004).

It is important to remember that the different estimates of NPP presented in this work do not change the other variable results; it only implies a different way of comparing data with a model like PELAGOS that implements a more sophisticated parameterization of primary production. This occurs because our model of primary production simulates the different carbon pathways and the same methods may not be applicable with specific models built to quantify net (particulate) primary production only. However, since most of the biogeochemical models aim at the estimation of net ecosystem production as a proxy to export production, by neglecting this fraction they may underestimate the flow of carbon through the food web.

4.3 Implications for the metabolic balance of the global ocean

Microbial production and consumption rates are central for evaluating the efficiency of the biological pump. The amount of organic carbon produced by autotrophs and consumed by the heterotrophic biota in the surface and deeper layers determine the actual export of organic matter. A key question here is whether the global ocean is net heterotrophic or how large the net heterotrophic regions are (e.g. Del Giorgio and Duarte, 2002; Riser and Johnson, 2008) and what will be the changes in future climate conditions.

Unfortunately, only sparse direct measurements of these rates exist at the scale of the global ocean. Brix et al. (2006) have analysed the BATS and HOT data to quantify export production in subtropical gyres. They diagnosed metabolic rates by means of the ratio of net community production (NCP) over NPP (e- or ef-ratio), the ratio of particle production over NCP (p-ratio) and the ratio of particle production over NPP (pe-ratio, see Table 1 in Brix et al., 2006, for further descriptions and values).

The long-term mean ratios simulated by the model at Sta. Aloha (Sect. 3.4.2) suggests that all NPP is utilised in the surface euphotic layer by heterotrophic respiration (e-ratio is 0.04, p-ratio is 0.43 and pe-ratio is 0.015). The e-ratio at Sta. ALOHA is much lower than estimated by Brix et al. (2006, e-ratio=0.22) but the ep-ratio, which is a proxy of the effective export from the euphotic zone is comparable. Notice that these ratios are independent of the choice of the NPP computation according to the DOC percentage discussed above, since modelled NCP also includes bacterial production that originate from the utilization of DOC released by phytoplankton. The estimates computed by Brix et al. (2006) might instead be affected by this factor. NPP measured through C^{14} bottle incubations may be higher than reported if there is extra-cellular DOC release, and therefore the e-ratio can be substantially lower.

BATS shows a higher export ratio (e-ratio=0.23), with p-ratio=0.23 and pe-ratio=0.05. PELAGOS is thus capable of predicting the higher p-ratio at BATS indicating that there is more particulate material production at BATS than at Sta. ALOHA. Further comparison with bacterial production and respiration data are still needed to assess whether the rates of bacterial carbon transformation predicted by the model are realistic. At the current stage these results imply that subtropical regions in the model are metabolically neutral from the point of view of the effects on the global carbon cycle because all primary produced carbon is utilized in the surface ocean.

The JGOFS stations give information on the temporal variability of carbon cycling in the subtropical gyre but lack any resolution of spatial variability. One of the largest available datasets covering basin-wide microbial carbon fluxes was presented by Hoppe et al. (2002) and it consists in data collected during the period November 1991–January 1992 along an Atlantic meridional transect. These data revealed that the tropical areas of the Atlantic with high SST values have Bacterial Carbon Demand (BCD) that exceeds NPP at the surface of the ocean. This is an indicator of net heterotrophy at the microbial level, that is likely to reduce the importance of the biological pump as a carbon sequestration process.

The direct comparison of bottle incubations with model data arises the same methodological issues as in Sect. 3.3.2. The model is parameterized to produce daily values of carbon production rates while incubation experiments have shorter time scales (usually 8–12 h). The strategy we adopted to reduce uncertainties in the extrapolation was to convert model results into data units by taking into account the experimental protocols applied during the campaign. The estimate of NPP by Hoppe et al. (2002) was done by neglecting the DOC exudation, which they estimated as 5–30% of the total.We present both the NPP1 and NPP2 model estimates when comparing with the observations (as done in Sec. 3.3.2), but used the net organic production NPP1 only for deriving integrated carbon production at the basin and global scales.

The model computes bacterial production (BP) from a constant bacterial growth efficiency (BGE) of 30% (Vichi et al., 2007a) and BCD is a function of the nutrient content of the available organic substrate. The comparison with observed BP values is thus the central assessment of model skill because this is the variable that is directly measured. The surface ratio of BP/NPP (Fig. 12a) computed along the meridional transect and during the same period of observations (Fig. 12c) shows a similar range of spatial variability as reported in Hoppe et al. (2002), with minima found at the higher latitudes and at the equator, and maxima in the tropics and at the southern boundary of the South-Pacific gyre. The model reproduces also the BCD/BP distribution derived from the data, with values above 100% in correspondence of the same maxima of BP/NPP from 20° N to 40° S. The values in the northern part, which are equal to the southern maximum and higher than observed, are likely due to spatial biases in the location of PP with respect to the estimates of satellite-based PP models (Fig. 4). Simulated NPP is in fact very low in the western part of the tropical Atlantic close to the Amazon river. Since the data from Hoppe et al. (2002) have been collected at the surface (although with some special treatments to account for low-light production), it is interesting to analyse the integrated model results over the euphotic zone depth (Fig. 12b). In this case the model predicts that the Atlantic north of 45° S is net heterotrophic during the boreal winter period, implying that BCD is sustained by additional non-local sources of organic matter. Since the model has no input of organic matter from land, these additional sources are provided by the more productive mid-latitude regions of the northern and southern Atlantic.

In the upper part of Table 4 we present a direct comparison of model results with averages from the northern and southern parts of the transect. The ranges of observations and model data overlap for all parameters and the spatial difference between north and south Atlantic is well captured. The southern part of the transects shows higher values of production mostly because of the sampling along the eastern shelf of South America and the crossing with the polar front. Surface chl is underestimated in the Northern Hemisphere but not in the southern part of the transect because the model simulates high biological activity in the frontal region that increases the mean value. The ratio BP/PP is also well simulated. The NPP2 estimates lead to higher ratios but also to larger standard deviations that still overlaps the observed values.

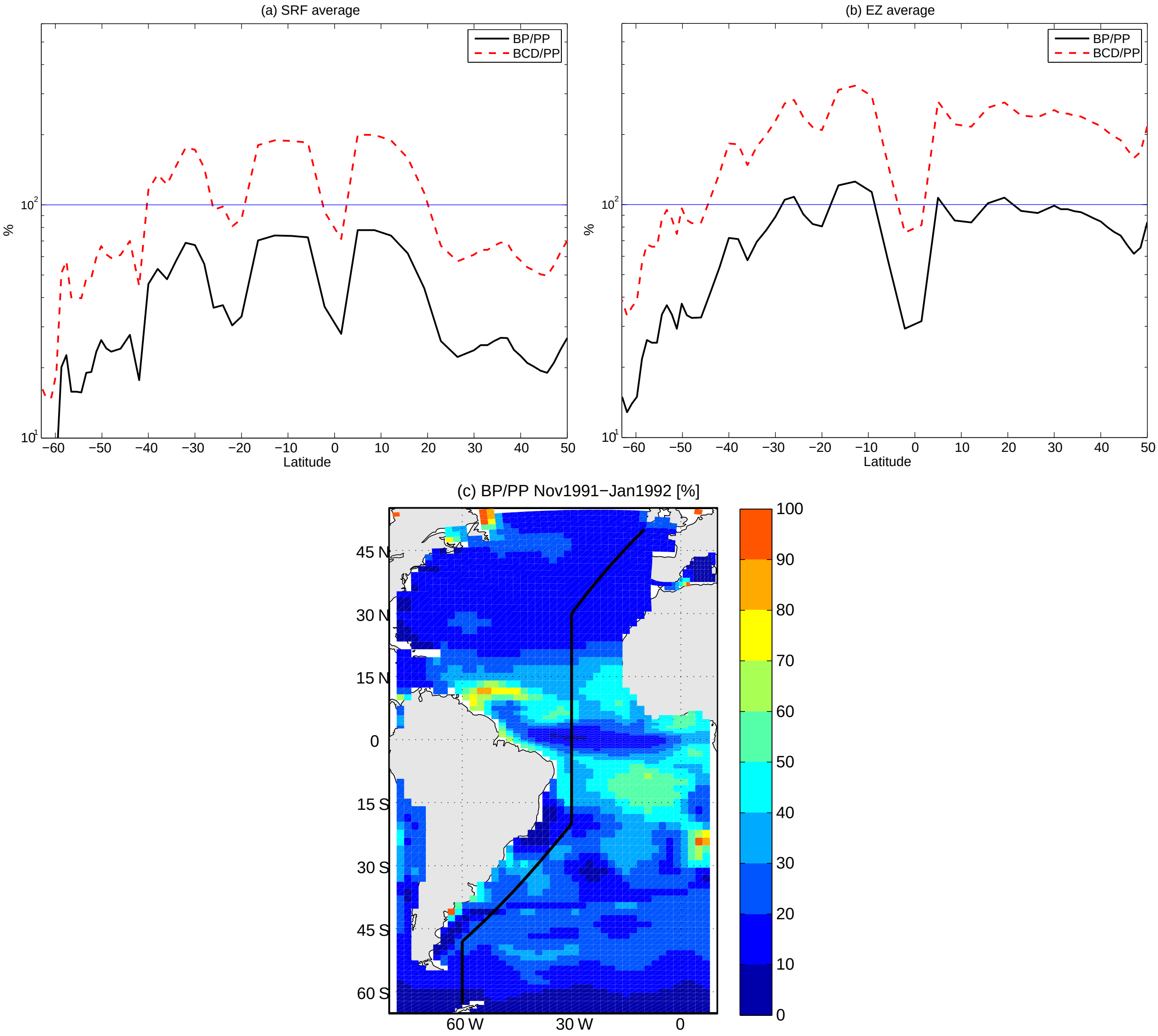

Fig. 12. Ratios of Bacterial Production (BP) and Bacterial Carbon Demand (BCD) to net primary production (NPP2) along the Atlantic track of Hoppe et al. (2002): **(a)** surface values; **(b)** values integrated over the euphotic zone depth; **(c)** map of the simulated mean value of BP/PP over the period of the measurements (November 1991 to January 1992) and location of the transect where model data have been extracted and compared with observations.

Given the good model skills on the transect data, it is interesting to derive basin scale estimates of the metabolic ratios first in the Atlantic and then in the global ocean. The results are shown in the lower part of Table 4 using the values integrated over the euphotic zone to provide a figure for the surface ocean. To check the relevance of seasonality, we have first extrapolated the boreal winter (NDJ) values computed over the whole simulation period 1980–2001 to the year length and compared it with the actual annual means. According to model results, if the boreal winter conditions are taken as representative of the mean microbial activity, the Atlantic ocean is net heterotrophic and the global ocean is slightly above 100%. The annual climatological means instead reveal that autotrophic carbon production is as important as heterotrophic processes, leading to values that are close to a neutral metabolic balance, if not slightly autotrophic. The overall figure of carbon production is in accordance with the satellite derived estimates (Falkowski et al., 2000; Carr et al., 2006). It is however likely that the model compensates for the low production in the gyres with the overestimation in the Southern Ocean (see Sect. 4.1), thus leading to a value that is comparable with the satellite estimates.

Table 4. Comparison of model results with data from Hoppe et al. (2002, Tab. 1) for the surface Atlantic in the period November 1991–January 1992. Mean values with standard deviation between brackets. The published bacterial production (BP) data estimated via leucine incorporation were converted to carbon using the factor of 1.5×10^{-3} μg C pmol^{-1}. n.a.=not available. The lower part of the table shows the spatially-integrated annual values of metabolic properties integrated over the euphotic zone (EZ) for the Atlantic and global ocean. Mean November-December-January (NDJ) values have been extrapolated to the annual value.

Data and model	SST [deg C]	Chl [μg l^{-1}]	PP [μg C l^{-1} h^{-1}]	BP [μg C l^{-1} h^{-1}]	BP/PP [%]	BCD/PP [%]
North Atlantic	20.8 (5.5)	0.35 (0.43)	0.72 (0.65)	0.05 (0.03)	18 (20)	n.a.
PELAGOS NPP1	19.9 (5.2)	0.14 (0.06)	0.77 (0.41)	0.07 (0.04)	21 (9)	53 (22)
NPP2			0.49 (0.22)		33 (21)	85 (49)
South Atlantic	14.6 (10.1)	0.86 (0.95)	1.31 (2.10)	0.09 (0.09)	23 (28)	n.a.
PELAGOS NPP1	15.0 (8.7)	0.72 (0.71)	1.91 (1.39)	0.15 (0.08)	21 (10)	54 (25)
NPP2			1.49 (1.24)		35 (22)	89 (56)
	Model (EZ)		PP [Pg C y^{-1}]	BP [Pg C y^{-1}]	BP/PP [%]	BCD/PP [%]
	Atlantic (NDJ, extrapolated)		11.31	4.93	44	112
	Global (NDJ, extrapolated)		56.40	22.70	40	103
	Atlantic (Annual)		11.48	4.33	38	97
	Global (Annual)		53.94	20.80	39	99

The net autotrophic balance suggested by the model is in accordance with geochemical evidence of oxygen production and recent direct observations (Najjar and Keeling, 2000; Riser and Johnson, 2008). The coarse resolution of the model cannot distinguish whether the autotrophic production is due to pulses of production as suggested by Karl et al. (2003b) or to a continuous contribution as recently suggested (Riser and Johnson, 2008). It is however clear from model results that the extrapolation of process rate variables from the local scale to the annual and basin scales may lead to misestimation of the metabolic state, with a tendency to show net heterotrophic conditions. This may even be more pronounced when the extrapolation is done using real data that are affected by mesoscale local processes more than our coarse model simulation (Maixandeau et al., 2005a) and it is thus important to consider an entire seasonal cycle to properly estimate the trophic state (Maixandeau et al., 2005b).

5 Conclusions

A set of objective assessment tools have been used to test the skill of the PELAGOS model over the last 20 years of the 20th century. The aims were twofold. Firstly, to evaluate the performance of the model under current climate conditions in view of its usage in climate change scenario simulations in the context of Earth System Models (ESM). The focus was thus on the production of organic carbon and its transformation along the microbial food web. Secondly, based on additional comparisons with measured basin-scale carbon exchange rates in the Atlantic, we computed the efficiency of the surface net community production taken as a proxy for the biological pump.

The skill of the model in terms of simulating NPP and carbon transformation in the surface ocean is adequate in some regions such as the equatorial Pacific and the north Atlantic subtropical gyre. The visual comparison with satellite-derived PP data is also qualitatively acceptable, although it has been demonstrated that satPP models and OBGCMs have similar skill scores with respect to in situ observations. Further independent data sets are thus needed to make sure that satellite-based products can be used to fill the observational gaps and robustly validate OBGCMs. It is also extremely important that long-term data series are maintained and new ones are implemented in other important regions of the ocean (for instance the northern Indian Ocean and the Southern Ocean, where the models shows the largest biases), in order to allow assessments as the one shown in Sect. 3.4

Our results underline the importance of suitable comparison between observations and model simulations of primary production. Experimental protocols and model variable (or process) definitions need to be properly considered to maximize the information that can be extracted from data and model results. In our specific case, the quality and quantity of DOC exudated from phytoplankton under oligotrophic conditions was found to be a key variable to improve the goodness-of-fit of the model against in situ primary production observations. It is therefore recommended that DOC quality measurements be taken during in situ incubations for NPP studies.

The comparison with completely independent data of the carbon fluxes through bacteria increase our degree of confidence in the model results, making it suitable for studying the degradation processes of organic matter under different oceanic conditions. Within the limits of the model assumption and known biases, we have thus used PELAGOS results to estimate the metabolic balance of the global ocean in the euphotic zone. The model predicts that in boreal winter conditions and in oligotrophic regions of the Atlantic there is a tendency towards net heterotrophy as observed in the field. However, in the annual mean and over the Atlantic basin up to the global scales, the surface ocean is close to a slightly positive autotrophic balance. It is therefore interesting to further investigate the behaviour of the model in case of climate change scenarios and assess whether the induced changes in the general circulation and water-mass properties might affect this state.

Appendix A

Univariate skill scores

The most simple measures of distance between a set of observations O_n and model predictions P_n, n=1,2,...,N are the bias

$$\mathrm{B}=\frac{1}{N}\sum_{n=1}^{N}P_n-\frac{1}{N}\sum_{n=1}^{N}O_n=\bar{P}-\bar{O}$$

and the absolute average error

$$\mathrm{AAE}=\frac{1}{N}\sum_{n=1}^{N}|O_n-P_n|.$$

The total Root Mean Square Difference (RMSD) is defined as

$$\mathrm{RMSD}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(O_n-P_n)^2}$$

which can be further separated into a component due to the bias and a centered-pattern (unbiased) difference

$$\mathrm{RMSD}_{CP}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[\left(O_n-\bar{O}\right)-\left(P_n-\bar{P}\right)\right]^2}$$

with the relationship

$$\mathrm{RMSD}^2=\mathrm{B}^2+\mathrm{RMSD}_{CP}^2.$$

The Nash-Sutcliffe Model Efficiency (Nash and Sutcliffe, 1970) is a measure of the ratio of the model error to the variability of the observations:

$$\mathrm{MEF}=1-\frac{\sum_{n=1}^{N}(O_n-P_n)^2}{\sum_{n=1}^{N}\left(O_n-\overline{O}\right)^2}$$

where O_n and P_n are the N pairs of observational data and predictions, respectively. Performance levels are usually categorised as follows: >0.65 excellent, 0.65–0.5 very good, 0.5–0.2 good, <0.2 poor. If the index is lower than 0, it means that the model is a worse predictor than the mean of the observations. If the index is close to 0, the model is as good a predictor as the data mean. This implies that the model correctly reproduces the mean but that the simulated variability is lower than observed.

The Reliability Index (RI, Leggett and Williams, 1981) measures the order of magnitude of model predictions with respect to data:

$$\mathrm{RI}=\exp\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\log\frac{O_n}{P_n}\right)^2}.$$

It was originally proposed as a statistical interpretation of log-normal distributed data, which is a typical distribution for most of the properties in ecology. The measure is interpreted as the value such that 68% of the model predictions fall within 1/RI and RI (Smith and Rose, 1995). This index thus does not distinguish whether the multiplicative factor is related to over- or underestimation therefore it requires the concurrent analysis of the bias.

Non-metric Multi-Dimensional Scaling (MDS)

Starting from a matrix of similarities/dissimilarities (resemblance matrix) between n items, this algorithm constructs a new set of data points in a low dimension space (usually 2-D) whose proximities are obtained through a minimization procedure that maintains the original distances in the resemblance matrix. The stress function that is minimized through iteration is the measure of the fit between proximities in the new low-dimensional space (d_{ij}) and the distances in the original data space (δ_{ij}) as, for instance:

$$\mathrm{stress}=\left[\frac{\sum_{i=1}^{n}\sum_{j>i}^{n}\left(d_{ij}-f\left(\delta_{ij}\right)\right)^2}{\sum_{i=1}^{n}\sum_{j>i}^{n}d_{ij}^2}\right]^{\frac{1}{2}}$$

where f is a non-metric, monotone transformation of the input data (distances). The minimization will lead to a final choice of f that reproduces the general rank-ordering of distances between the objects. A stress value lower than 0.2 is considered a sufficient description of the original proximities in the lower dimensional space. The distances between the coordinates of item i and j in the new space are Euclidean and depends on the chosen number of dimensions m, such as

$$d_{ij}=\left[\sum_{a=1}^{m}\left(x_{ia}-x_{ja}\right)^2\right]^{\frac{1}{2}}$$

where x_a, a=1,...,m, is the new coordinate system. MDS is a standard tool in many statistical software packages and in this particular case we used the MATLAB implementation. The multivariate distance metrics containing the combination of the skill score values can be computed in different ways. Given the non-linear nature of the skill scores, the resemblance matrix was built using the Manhattan (cityblock) distance:

$$\delta_{ij} = \sum_{k=1}^{K} \left| s_{ik} - s_{jk} \right|$$

where K is the number of normalized skill scores s defining each model object.

Acknowledgements. This work was supported by the Italian grant FISR given to the Centro Euro-Mediterraneo per i Cambiamenti Climatici (CMCC). SeaWiFS data used in this publication were produced by the SeaWiFS project at Goddard Space Flight Centre. SeaWiFS and VGPM data have been downloaded from http://www.science.oregonstate.edu/ocean.productivity. Use of this data is in accord with the SeaWiFS Research Data Use Terms and Agreements. JGOFS data have been downloaded from the US JGOFS Synthesis & Modeling Project web page. MV thanks M. Azzaro and R. LaFerla for the help with the bacterioplankton production rate data. We thank three anonymous reviewers who provided valuable comments on an early version of the manuscript.

Edited by: C. Heinze

References

Aiken, J., Fishwick, J. R., Lavender, S., Barlow, R., Moore, G. F., Sessions, H., Bernard, S., Ras, J., and Hardman-Mountford, N. J.: Validation of MERIS reflectance and chlorophyll during the BENCAL cruise October 2002: preliminary validation of new demonstration products for phytoplankton functional types and photosynthetic parameters, Int. J. Remote Sens., 28, 497–516, 2007.

Allen, J., Somerfield, P., and Gilbert, F.: Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models, J. Marine Syst., 64, 3–14, 2007.

Alvain, S., Moulin, C., Dandonneau, Y., and Breon, F. M.: Remote sensing of phytoplankton groups in case 1 waters from global SeaWiFS imagery, Deep-Sea Res. Pt. I, 52, 1989–2004, 2005.

Baretta, J., Ebenhöh, W., and Ruardij, P.: The European Regional Seas Ecosystem Model, a complex marine ecosystem model, J. Sea Res., 33, 233–246, 1995.

Baretta-Bekker, J., Baretta, J., and Rasmussen, E.: The microbial food web in the European Regional Seas Ecosystem Model, J. Sea Res., 33, 363–379, 1995.

Behrenfeld, M. and Falkowski, P.: Photosynthetic rates derived from satellite-based chlorophyll concentration, Limnol. Oceanogr., 42, 1–20, 1997.

Bellucci, A., Masina, S., Dipietro, P., and Navarra, A.: Using temperature-salinity relations in a global ocean implementation of a Multivariate data assimilation scheme, Mon. Weather Rev., 135, 3785–3807, 2007.

Borg, I. and Groenen, P.: Modern Multidimensional Scaling: theory and applications, Springer-Verlag, New York, 2nd edn., 2005.

Brix, H., Gruber, N., Karl, D. M., and Bates, N. R.: On the relationships between primary, net community, and export production in subtropical gyres, Deep-Sea Res. Pt. II, 53, 698–717, 2006.

Campbell, J., Antoine, D., Armstrong, R., Balch, W., Barber, R., Behrenfeld, M., Bidigare, R., Bishop, J., Carr, M.-E., Esaias, W., Falkowski, P., Hoepffner, N., Iverson, R., Kiefer, D., Lohrenz, S., Marr, J., Morel, A., Ryan, J., Vedernikov, V., Waters, K., Yentsch, C., and J., Y.: Comparison of algorithms for estimating ocean primary production from surface chlorophyll, temperature, and irradiance, Global Biogeochem. Cy., 16, 1035, doi:10.1029/2001GB001444, 2002.

Carr, M. E., Friedrichs, M. A. M., Schmeltz, M., Aita, M. N., Antoine, D., Arrigo, K. R., Asanuma, I., Aumont, O., Barber, R., Behrenfeld, M., Bidigare, R., Buitenhuis, E. T., Campbell, J., Ciotti, A., Dierssen, H., Dowell, M., Dunne, J., Esaias, W., Gentili, B., Gregg, W., Groom, S., Hoepffner, N., Ishizaka, J., Kameda, T., Le Quere, C., Lohrenz, S., Marra, J., Melin, F., Moore, K., Morel, A., Reddy, T. E., Ryan, J., Scardi, M., Smyth, T., Turpie, K., Tilstone, G., Waters, K., and Yamanaka, Y.: A comparison of global estimates of marine primary production from ocean color, Deep-Sea Res. Pt. II, 53, 741–770, 2006.

Conkright, M., Garcia, H., O'Brien, T., Locarnini, R., Boyer, T., Stephens, C., and Antonov, J.: World Ocean Atlas 2001, Volume 4: Nutrients, vol. NOAA Atlas NESDIS 52, U.S. Government Printing Office, Washington D.C., ftp://ftp.nodc.noaa.gov/pub/data.nodc/woa/PUBLICATIONS/woa01v4d.pdf, cD-ROMs, 2002.

de Boyer Montégut, C., Madec, G., Fischer, A., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: an examination of profile data and a profile-based climatology, J. Geophys. Res., 109, C12003, doi:10.1029/2004JC002378, 2004.

Del Giorgio, P. A. and Duarte, C. M.: Respiration in the open ocean, Nature, 420, 379–384, 2002.

Denman, K., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P., Dickinson, R., Hauglustaine, D., Heinze, C., Holland, E., Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P., Wofsy, S., and Zhang, X.: Couplings Between Changes in the Climate System and Biogeochemistry, in: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., M.Tignor, and Miller, H., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Dong, S., Sprintall, J., Gille, S. T., and Talley, L.: Southern Ocean mixed-layer depth from Argo float profiles, J. Geophys. Res., 113, C06013, doi:10.1029/2006JC004051, 2008.

Dore, J. E., Letelier, R. M., Church, M. J., Lukas, R., and Karl, D. M.: Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: Historical perspective and recent observations, Prog. Oceanogr., 76, 2–38, 2008.

Ducklow, H. W., Smith, D. C., Campbell, L., Landry, M. R., Quinby, H. L., Steward, G. F., and Azam, F.: Heterotrophic bacterioplankton in the Arabian Sea: Basinwide response to year-round high primary productivity, Deep-Sea Res. Pt. II, 48, 1303–1323, 2001.

Dunne, J. P., Sarmiento, J. L., and Gnanadesikan, A.: A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor, Glob. Biogeochem. Cy., 21, GB4006, doi:10.1029/2006GB002907, 2007.

Engel, A., Thoms, S., Riebesell, U., Rochelle-Newall, E., and Zondervan, I.: Polysaccharide aggregation as a potential sink of marine dissolved organic carbon, Nature, 428, 929–932, 2004.

Eppley, R.: Temperature and phytoplankton growth in the sea, Fishery Bulletin, 70, 1063–1085, 1972.

Falkowski, P., Scholes, R., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Högberg, P., Linder, S., Mackenzie, F., III, B. M., Pedersen, T., Rosenthal, Y., Seitzinger, S., Smetacek, V., and Steffen, W.: The global carbon cycle: a test of our knowledge of Earth as a system, Science, 290, 291–296, 2000.

Friedrichs, M. A. M., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. A., Asanuma, I., Behrenfeld, M. J., Buitenhuis, E. T., Chai, F., Christian, J. R., Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J., Gentili, B., Gregg, W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Melin, F., Moore, J. K., Morel, A., O'Malley, R. T., O'Reilly, J., Saba, V. S., Schmeltz, M., Smyth, T. J., Tjiputra, J., Waters, K., Westberry, T. K., and Winguth, A.: Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean, J. Mar. Sys., 76, 113–133, doi:10.1016/j.jmarsys.2008.05.010, http://www.sciencedirect.com/science/article/B6VF5-4SMDYY6-8/2/5d74d6d829a3d2b9874a5b1e3ec1fd91, 2009.

Fuhrman, J. A. and Azam, F.: Thymidine incorporation as a measure of heterotrophic bacterioplankton production in marine surface waters - evaluation and field results, Mar. Biol., 66, 109–120, 1982.

Garabato, A. C. N., Polzin, K. L., King, B. A., Heywood, K. J., and Visbeck, M.: Widespread intense turbulent mixing in the Southern Ocean, Science, 303, 210–213, 2004.

Gregg, W. W. and Casey, N. W.: Global and regional evaluation of the SeaWiFS chlorophyll data set, Remote Sensing of Environment, 93, 463–479, 2004.

Gundersen, K., Heldal, M., Norland, S., Purdie, D. A., and Knap, A. H.: Elemental C, N, and P Cell Content of Individual Bacteria Collected at the Bermuda Atlantic Time-Series Study (BATS) Site, Limnol. Oceanogr., 47, 1525–1530, 2002.

Hood, R., Laws, E., Armstrong, R., Bates, N., Brown, C., Carlson, C., Chai, F., Doney, S., Falkowski, P., Feely, R., et al.: Pelagic functional group modeling: Progress, challenges and prospects, Deep-Sea Res. Pt. II, 53, 459–512, doi:10.1016/j.dsr2.2006.01.025, 2006.

Hoppe, H.-G., Gocke, K., Koppe, R., and Begler, C.: Bacterial growth and primary production along a north-south transect of the Atlantic Ocean., Nature, 416, 168–171, doi:10.1038/416168a, 2002.

Huisman, J., Pham Thi, N. N., Karl, D. M., and Sommeijer, B.: Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum., Nature, 439, 322–325, doi:10.1038/nature04245, 2006.

Hurtt, G. and Armstrong, R.: A pelagic ecosystem model calibrated with BATS data., Deep-Sea Res. Pt. II, 43, 653–683, 1996.

Karl, D., Bates, N., Emerson, S., Harrison, P., Jeandel, C., O., L., Liu, K.-K., Marty, J.-C., Michaels, A., Miquel, J., Neuer, S., Nojiri, Y., and Wong, C.: Temporal studies of biogeochemical processes determined form ocean time-series observations during the JGOFS era, in: Ocean Biogeochemistry, edited by: Fasham, M., chap. 10, 239–267, Springer, Berlin, 2003a.

Karl, D. M., Laws, E. A., Morris, P., Williams, P. J. L., and Emerson, S.: Global carbon cycle: metabolic balance of the open sea., Nature, 426, 32 pp., 2003b.

Kepkay, P. E., Niven, S., and Milligan, T. G.: Low molecular weight and colloidal DOC production during a phytoplankton bloom, Mar. Ecol. Prog. Ser., 100, 33–244, 1993.

Leggett, R. and Williams, L.: A reliability index for models, Ecol. Model., 13, 303–312, 1981.

Longhurst, A. R.: Ecological geography of the sea, Academic Press, San Diego, London, 1st edn., 1998.

Longhurst, A. R.: Ecological geography of the sea, Academic Press, Burlington, San Diego, London, 2nd edn., 2007.

Lukas, R. and Karl, D.: Hawaii Ocean Time-series (HOT), 1988–1998: A decade of interdisciplinary oceanography., CD-ROM 99-05, School of Ocean and Earth Science and Technology, University of Hawaii, http://hahana.soest.hawaii.edu/hot/hot, 1999.

Lynch, D. R., McGillicuddy Jr., D. J., and Werner, F. E.: Skill assessment for coupled biological/physical models of marine systems, J. Mar. Sys., 76, 1–3, doi:10.1016/j.jmarsys.2008.05.002, http://www.sciencedirect.com/science/article/B6VF5-4SJP7C0-1/2/623892e349a763de24ad6f4bd3f0814c, 2009.

Madec, G. and Imbard, M.: A global ocean mesh to overcome the North Pole singularity, Clim. Dynam., 12, 381–388, 1996.

Madec, G., Delecluse, P., Imbard, M., and Levy, C.: OPA8.1 ocean general circulation model reference manual, Notes du pole de modelisation, IPSL, France, http://www.lodyc.jussieu.fr/opa, 1999.

Maixandeau, A., Lefevre, D., Fernandez, I. C., Sempere, R., Sohrin, R., Ras, J., Van Wambeke, F., Caniaux, G., and Queguiner, B.: Mesoscale and seasonal variability of community production and respiration in the surface waters of the NE Atlantic Ocean, Deep-Sea Res. Pt. I, 52, 1663–1676, 2005a.

Maixandeau, A., Lefevre, D., Karayanni, H., Christaki, U., Van Wambeke, F., Thyssen, M., Denis, M., Fernandez, C. I., Uitz, J., Leblanc, K., and Queguiner, B.: Microbial community production, respiration, and structure of the microbial food web of an ecosystem in the northeastern Atlantic Ocean, J. Geophys. Res., 110, C07S17, doi:10.1029/2004JC002694, 2005b.

Mason, S. J.: Understanding forecast verification statistics, Meteorol. Appl., 15, 31–40, doi:10.1002/met.51, 2008.

Najjar, R. and Keeling, R.: Mean annual cycle of the air-sea oxygen flux: A global view, Glob. Biogeochem. Cy., 14, 573–584, 2000.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models, part 1 – a discussion of principles, J. Hydrol., 10, 282–290, 1970.

Ogawa, H. and Tanoue, E.: Dissolved organic matter in oceanic waters, J. Oceanogr., 59, 129–147, 2003.

Ondrusek, M. E., Bidigare, R. R., Waters, K., and Karl, D. M.: A predictive model for estimating rates of primary production in the subtropical North Pacific Ocean, Deep-Sea Res. Pt. II, 48, 1837–1863, 2001.

Pineiro, G., Perelman, S., Guerschman, J. P., and Paruelo, J. M.: How to evaluate models: Observed vs. predicted or predicted vs. observed?, Ecol. Model., 216, 316–322, 2008.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., and Wang, W.: An Improved In Situ and Satellite SST Analysis for Climate, J. Climate, 15, 1609–1625, 2002.

Riser, S. C. and Johnson, K. S.: Net production of oxygen in the subtropical ocean, Nature, 451, 323–325, http://dx.doi.org/10.1038/nature06441, 2008.

Robinson, C., Tilstone, G. H., Rees, A. P., Smyth, T. J., Fishwick, J. R., Tarran, G. A., Luz, B., Barkan, E., and David, E.: Comparison of in vitro and in situ plankton production determinations, Aquat. Microb. Ecol., 54, 13–34, 2009.

Rose, K. A. and Smith, E. P.: Statistical assessment of model goodness-of-fit using permutation tests, Ecol. Model., 106, 129–139, 1998.

Rykiel, E. J.: Testing ecological models: the meaning of validation, Ecol. Model., 90, 229–244, doi:10.1016/0304-3800(95)00152-2, http://www.sciencedirect.com/science/article/B6VBS-3VWK7V1-7/2/6a5c8fc2b57707e1f028e00432baa7b2, 1996.

Sarmiento, J., Hughes, T., Stouffer, R., and Manabe, S.: Simulated response of the ocean carbon cycle to anthropogenic climate warming, Nature, 393, 245–249, 1998.

Schneider, B., Bopp, L., Gehlen, M., Segschneider, J., Frölicher, T. L., Cadule, P., Friedlingstein, P., Doney, S. C., Behrenfeld, M. J., and Joos, F.: Climate-induced interannual variability of marine primary and export production in three global coupled climate carbon cycle models, Biogeosciences, 5, 597–614, 2008, http://www.biogeosciences.net/5/597/2008/.

Smith, E. P. and Rose, K. A.: Model goodness-of-fit analysis using regression and related techniques, Ecol. Model., 77, 49–64, 1995.

Spitz, Y., Moisan, J., and Abbott, M.: Configuring an ecosystem model using data from the Bermuda Atlantic Time Series (BATS), Deep-Sea Res. Pt. II, 48, 1733–1768, 2001.

Steinberg, D. K., Carlson, C. A., Bates, N. R., Johnson, R. J., Michaels, A. F., and Knap, A. H.: Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry, Deep-Sea Res. Pt. II, 48, 1405–1447, 2001.

Stow, C. A., Jolliff, J., McGillicuddy Jr., D. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A., and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems, J. Marine Syst., 76, 4–15, doi:10.1016/j.jmarsys.2008.03.011, http://www.sciencedirect.com/science/article/B6VF5-4SKB3BN-4/2/31bd1e91528321dc08996d9687069afb, 2009.

Tjiputra, J. F., Polzin, D., and Winguth, A. M. E.: Assimilation of seasonal chlorophyll and nutrient data into an adjoint three-dimensional ocean carbon cycle model: Sensitivity analysis and ecosystem parameter optimization, Global Biogeochem. Cy., 21, doi:10.1029/2006GB002745, http://dx.doi.org/10.1029/2006GB002745, 2007.

Vichi, M., Masina, S., and Navarra, A.: A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part II: numerical simulations, J. Mar. Sys., 64, 110–134, 2007b.

Vichi, M., Pinardi, N., and Masina, S.: A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part I: theory, J. Mar. Syst., 64, 89–109, 2007a.

Vichi, M., Masina, S., and Nencioli, F.: A process-oriented model study of equatorial Pacific phytoplankton: The role of iron supply and tropical instability waves, Progr. Oceanogr., 78, 147–162, 2008.