

Global Patterns and Predictions of Seafloor Biomass Using Random Forests

Chih-Lin Wei^{1*}, Gilbert T. Rowe², Elva Escobar-Briones³, Antje Boetius⁴, Thomas Soltwedel⁴, M. Julian Caley⁵, Yousria Soliman⁶, Falk Huettmann⁷, Fangyuan Qu^{1,8}, Zishan Yu⁸, C. Roland Pitcher⁹, Richard L. Haedrich¹⁰, Mary K. Wicksten¹¹, Michael A. Rex¹², Jeffrey G. Baguley¹³, Jyotsna Sharma¹⁴, Roberto Danovaro¹⁵, Ian R. MacDonald¹⁶, Clifton C. Nunnally¹, Jody W. Deming¹⁷, Paul Montagna¹⁸, Mélanie Lévesque¹⁹, Jan Marcin Weslawski²⁰, Maria Wlodarska-Kowalczyk²⁰, Baban S. Ingole²¹, Brian J. Bett²², David S. M. Billett²², Andrew Yool²², Bodil A. Bluhm²³, Katrin Iken²³, Bhavani E. Narayanaswamy²⁴

1 Department of Oceanography, Texas A&M University, College Station, Texas, United States of America, **2** Department of Marine Biology, Texas A&M University at Galveston, Galveston, Texas, United States of America, **3** Instituto de Ciencias del Mar y Limnología, Universidad Nacional Autónoma de México, México D.F., México, **4** Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany, **5** Australian Institute of Marine Science, Townsville, Queensland, Australia, **6** Biological and Environmental Sciences, Qatar University, Doha, Qatar, **7** Biology and Wildlife Department, Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, Alaska, United States of America, **8** College of Marine Life Science, Ocean University of Qingdao, Qingdao, China, **9** CSIRO Marine and Atmospheric Research, Cleveland, Queensland, Australia, **10** Department of Biology, Memorial University, St. John's, Newfoundland and Labrador, Canada, **11** Department of Biology, Texas A&M University, College Station, Texas, United States of America, **12** Department of Biology, University of Massachusetts, Boston, Massachusetts, United States of America, **13** Department of Biology, University of Nevada Reno, Reno, Nevada, United States of America, **14** Department of Biology, University of Texas at San Antonio, San Antonio, Texas, United States of America, **15** Department of Marine Sciences, Polytechnic University of Marche, Ancona, Italy, **16** Department of Oceanography, Florida State University, Tallahassee, Florida, United States of America, **17** Department of Oceanography, University of Washington, Seattle, Washington, United States of America, **18** Harte Research Institute, Texas A&M University-Corpus Christi, Corpus Christi, Texas, United States of America, **19** Institut des sciences de la mer de Rimouski, Université du Québec à Rimouski, Rimouski, Québec, Canada, **20** Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland, **21** National Institute of Oceanography, Dona Paula, Goa, India, **22** National Oceanography Centre, Southampton, Southampton, United Kingdom, **23** School of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, Alaska, United States of America, **24** Scottish Association for Marine Science, Scottish Marine Institute, Oban, United Kingdom

Abstract

A comprehensive seafloor biomass and abundance database has been constructed from 24 oceanographic institutions worldwide within the Census of Marine Life (CoML) field projects. The machine-learning algorithm, Random Forests, was employed to model and predict seafloor standing stocks from surface primary production, water-column integrated and export particulate organic matter (POM), seafloor relief, and bottom water properties. The predictive models explain 63% to 88% of stock variance among the major size groups. Individual and composite maps of predicted global seafloor biomass and abundance are generated for bacteria, meiofauna, macrofauna, and megafauna (invertebrates and fishes). Patterns of benthic standing stocks were positive functions of surface primary production and delivery of the particulate organic carbon (POC) flux to the seafloor. At a regional scale, the census maps illustrate that integrated biomass is highest at the poles, on continental margins associated with coastal upwelling and with broad zones associated with equatorial divergence. Lowest values are consistently encountered on the central abyssal plains of major ocean basins. The shift of biomass dominance groups with depth is shown to be affected by the decrease in average body size rather than abundance, presumably due to decrease in quantity and quality of food supply. This biomass census and associated maps are vital components of mechanistic deep-sea food web models and global carbon cycling, and as such provide fundamental information that can be incorporated into evidence-based management.

Citation: Wei C-L, Rowe GT, Escobar-Briones E, Boetius A, Soltwedel T, et al. (2010) Global Patterns and Predictions of Seafloor Biomass Using Random Forests. *PLoS ONE* 5(12): e15323. doi:10.1371/journal.pone.0015323

Editor: Tamara Natasha Romanuk, Dalhousie University, Canada

Received: August 10, 2010; **Accepted:** November 8, 2010; **Published:** December 30, 2010

Copyright: © 2010 Wei et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by a grant from the Census of Marine Life (CoML) and the Sloan Foundation to G. T. Rowe and E. Escobar-Briones. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: weic@tamug.edu

Introduction

Rationale

A ‘census’, according to our dictionaries, was originally a counting of individuals for the purpose of taxation. The Census of Marine Life (CoML) is on the other hand an attempt to make a comprehensive assessment of what lives in the world’s oceans. CoML is attempting to document, describe, list, archive and map

as many species of organisms as possible in all marine ecosystems, independent of an individual species’ population size. A natural by-product of CoML however has been new tabulations of animal abundances and biomass by CoML field projects. The purpose of this CoML biomass synthesis has been to capture all the new information on biomass that has been uncovered during CoML into a single data base, independent of species composition. This project has thus archived and mapped a broad spectrum of

biomass data from CoML projects from around the world, added data from a number of previous comprehensive reviews, and, as a result, produced maps of biomass of a limited number of size groups living on the sea floor on a world wide basis.

While the causes of biodiversity remain obscure to a large degree, there is general agreement that biomass is a function of food supply to or within any particular habitat. As a result, standing stock biomass has been used as a surrogate for biomass production and carbon flow to and through an ecosystem, without necessarily defining the taxa contributing to the biomass. On the other hand, by analyzing the statistical relationships of diversity to biomass, it might be possible to make some practical inferences about the effects that productivity might have on diversity [1], as this is an open question that has generated considerable conjecture [2]. While the biomass census is not related to ‘taxation’ in the classic sense, it directly links marine populations to carbon as an ecosystem model currency. Inorganic carbon is fixed into organic-rich compounds by photosynthesis and then transferred through food webs where it has a variety of fates, usually a return to CO₂. However, it is also harvested by fishers and it thus ends up in markets around the world. A biomass census therefore has relevance to societies because human populations are putting a ‘tax’ on the ocean biota in the form of valuable protein in fisheries products.

Historical background

The earliest quantitative sampling of the sea floor began at the beginning of the 20th century as an attempt to determine food resources available to bottom-dwelling fish in European waters [3,4]. A good review of the mechanical instruments developed for the early shallow-water surveys [5] pictures a wide variety of grab-like samplers, many still in use today. By the middle of the 20th century, the macrofauna of many continental shelves and estuaries had been sampled quantitatively by a relatively standard set of instruments. For demersal fishes and vagile megafaunal invertebrates, the most common sampling methods are trawling and photography. Both methods have weaknesses: for example, trawling tends to capture only surface-dwelling and slow species. It may be impossible to positively identify animals to species from photographs. However, to this day neither is fool proof. With smaller forms (meiofauna, microfauna, bacteria and viruses), sampling problems are solved seemingly easily by utilizing small-diameter cores, but care has to be taken not to lose organisms by either washing or bow-wake of sampling devices. For these groups, the problem is that they have not yet been sampled comprehensively on global or ocean-basin scales.

Generalizations about the controls of sea floor biomass began to emerge by the middle of the 20th century. Expeditions sponsored by Union of Soviet Socialist Republics (USSR: dissolved in 1991) reached every corner of the globe. This large body of work concluded that biomass declines sharply with depth and with distance from land. They observed that high latitudes tended to have higher biomass than low latitudes. The major food supply to both pelagic and sea floor communities was the rain of particulate detritus, enhanced by a ladder of vertical migration [6]. Sea floor biomass likewise declines precipitously with depth, but is also tightly coupled to primary production in surface layers. Regression equations of the variation in benthic biomass as a function of depth and primary production established in the 1970’s initially (reviewed in [7]) are still reasonable estimates of deep benthic biomass today [8]. The slopes of the biomass regressions have been equated to the rate at which the delivery of POC to the sea floor declines, but the height or zero intercept of the regression line is a function of the mean primary production in the photic zone.

Previous reviews of seafloor standing stocks focused on bathymetric standing stock patterns in which the distribution of biomass and abundance was fitted to a linear function of water depth or direct measurement of sinking particle flux [7,8,9,10]. Applying such equations is conceptually intuitive but the relationships tend sometimes to fall apart in large scale predictive mapping. In this paper, we explore a novel machine-learning algorithm, Random Forests [11], to model the complex and potentially non-linear relationships between oceanic properties and seafloor standing stocks. Random Forests (RF) is a data mining method widely used in the fields of bioinformatics [12], speech recognition [13], and drug design and development [14]. Recently RF is gaining popularity in terrestrial ecology [15,16,17]; however, so far, only a handful of studies have applied RF in marine ecosystems [18,19]. In short, RF, as the name suggested, is an ensemble of many decision trees with binary divisions. Each tree is grown from a bootstrap sample of response variable and each node is guided by a predictor value to maximize differences in offspring branches. The fit of the tree is examined using the data not in the bootstrap selection; hence, cross-validation with external data is not necessary. Predictive accuracy requires low bias and low correlation between decision trees [11]. RF achieves these by growing a large number of trees and then averaging the predictions. At the same time, the node decision is chosen from a random subset of predictors to make the trees look as different as possible. RF does not assume any data distribution and does not require formal selection of predictors. RF is robust to outlier and unbalanced data, making it a better choice than traditional statistical methods [12].

Materials and Methods

Response Variables

Biomass and abundance of bacteria, meiofauna, macrofauna, megafauna (invertebrates+fishes), invertebrates, and fishes were assembled from literature and the Census of Marine Life (CoML) field projects (Figure 1 and Appendix S1). The “CoML Fresh Biomass Database” includes 4872 biomass records, 5511 abundance records, and 4196 records with both biomass and abundance from 175 studies. Additional datasets include nematodes (230 records from 10 studies) and pelagic decapods (17 records from 1 study); however, they were not included in this analysis. The complete list of references and detailed data information are available in Appendix S1 and File S1.

Categories of benthic fauna are usually defined by size classes. In this paper, we refer to the term “bacteria” to include both bacterial and archaeal domains. We have not included viruses. The metazoan meiofauna and macrofauna are small infauna invertebrates sampled by core or grab devices and retained on 20 to 74- μ m and 250 to 520- μ m sieves, respectively. Megafauna refers to large epibenthic invertebrates and demersal fishes (usually larger than 1 cm) caught or recorded by bottom trawling and photographic survey. Many studies deal with trawl invertebrates and fishes separately; hence, 3 categories were created for the megafauna, including the invertebrates plus fishes, invertebrates, and fishes. Here the “megafauna” dataset includes both invertebrates and fishes. Estimates of meiofaunal and macrofaunal standing stocks are affected by the gear design, sampling area, and sieve sizes [7,20,21,22]. These factors however have been suggested to be minor compared to water depth at a global scale and do not significantly affect the overall level and pattern of stock-depth relationships [2,8]. Only studies reporting standing stocks for the whole assemblage of a size category were used in these analyses. Benthic foraminiferans were not included due to

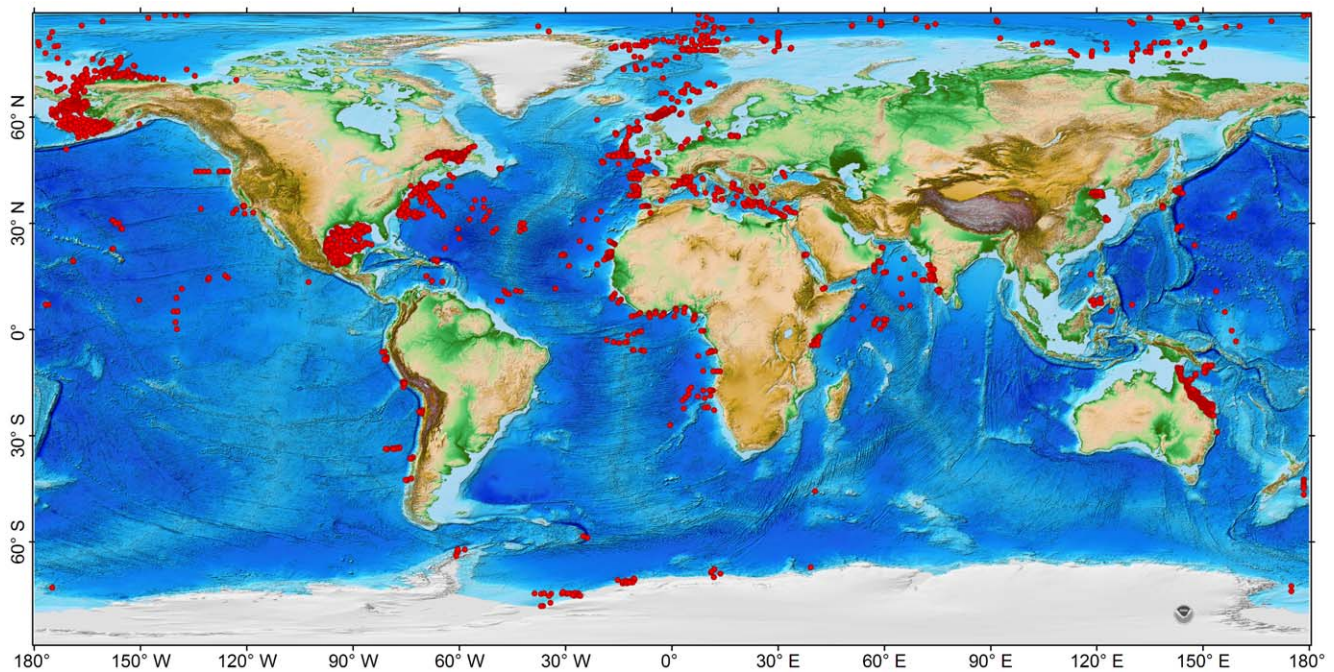


Figure 1. Distribution of abundance and biomass records in the “CoML Fresh Biomass Database”. References and locations for each size class are given in Appendix S1 and File S1. Bathymetric layer uses NOAA ETOPO 1 Global Relief Model [26]. doi:10.1371/journal.pone.0015323.g001

difficulty differentiating between living biomass from empty tests or shells [8,9]. Throughout this analysis, the abundance was standardized to cells (for bacteria) or individuals (for meiofauna, macrofauna, and megafauna) per square meter. The biomass was standardized to milligrams carbon per square meter using appropriate conversion factors from wet or dry weight to organic carbon weight [7,9].

Environmental Predictors

Environmental variables with global coverage were utilized to characterize 1) the surface ocean climate relating to phytoplankton production, 2) water column processes associated with export POC flux, 3) bottom water properties characterizing the seafloor habitats, and 4) seafloor relief (water depth) as a proxy of declining export POC flux arriving at the ocean floor (Table 1 and Figure S1). Contemporaneous environmental and standing stocks data were not always available; therefore, mean and standard deviation (S.D.) of the predictors were calculated for the longest time periods possible. The variables are listed as:

- 1) Primary productivity variables: Decadal mean and standard deviation (S.D.) of monthly net primary production (NPP) models (cbpm, vgpm), and the data inputs for the NPP models [23,24] including chlorophyll concentration (chl), sea surface temperature (sst), photosynthetic available irradiance (par), mixed layer depth (mld), particle backscatter (bbp), phytoplankton growth rate (growth), and carbon concentration (carbon), all calculated between years of 1998 and 2007. The monthly data were obtained from the Ocean Productivity Group, Oregon State University, as products of the Sea Viewing Wide Field-of-view Sensor (SeaWiFS r2009.1) and Advanced Very High Resolution Radiometer (AVHRR).
- 2) Water column processes: Decadal mean of water-column integrated total carbon (int.c) and nitrogen (int.n), detrital

carbon (det.c) and nitrogen (det.n), phytoplankton (phyt) and zooplankton (zoop), as well as export flux of detrital carbon (det.flx.c) and nitrogen (det.flx.n), obtained from a 10-year simulation of monthly model outputs from 1995 to 2004 using Ocean Circulation and Climate Advanced Model (OCCAM) driven by a nitrogen based Nutrient Phytoplankton Zooplankton Detritus (NPZD) Model [25].

- 3) Bottom water properties: Annual mean and seasonal standard deviation (S.D.) of bottom water temperature, salinity, oxygen, nitrate, phosphate, and silicate concentration were obtained from World Ocean Atlas 2009, NOAA National Oceanographic Data Center.
- 4) Global ocean depths were obtained from the ETOPO1 Global Relief Model, NOAA National Geophysical Data Center [26].

Data Analyses and Modeling

We used partial regression analysis to examine the relationships between standing stocks and depth when the latitude and longitude are held constant. The multiple regression residuals of stocks against latitude and longitude were used as dependent variables to regress against depth. To bring the dependent variable back to an appropriate scale, the y-intercept from the multiple regression was added to the residuals. The partial regression was also used in the pre-treatment of the depth-integrated bacteria data to standardize sediment penetration depths (from 0.5 to 29.5 cm; >83% are between 5 and 15 cm). Similar approaches has been developed and tested in Rex et al. [8].

A stochastic model between standing stocks and 39 environmental predictors (Table 1 and Figure S1) was constructed using Random Forests (RF) [11]. RF is a member of Regression Tree Analyses (RTA) [27]. In RTA, the response variable (standing stocks) is recursively partitioned into small successive binary splits. Each split is based on a single value of predictor from an

Table 1. Global datasets of environmental predictors.

Data Type	Data Source	Res.	Cell	Abbrev.	Variable	Unit
Primary Production	<u>Ocean Productivity, OSU</u>	5 minutes	3×3	chl	Chlorophyll a concentration (SeaWiFS r2009.1)	mg m ⁻³
	Decadal mean & standard deviation of monthly data from January 1998 to December 2007	5.3 minutes	3×3	sst	Sea Surface Temperature (AVHRR)	°C
		5 minutes	3×3	par	Photosynthetically available radiation (SeaWiFS r2009.1)	Einstein m ⁻² day ⁻¹
		5 minutes	3×3	bbp	Particulate backscatter (SeaWiFS r2009.1)	m ⁻¹
		10 minutes	1×1	mld	Mixed layed depth	m
		5 minutes	3×3	growth	Phytoplankton growth rate	divisions day ⁻¹
		5 minutes	3×3	carbon	Carbon concentration	mg m ⁻³
		5 minutes	3×3	vgpm	Chlorophyll based net primary production	mg C m ⁻² day ⁻¹
Water column	Yool et al. [25]	1 degree	1×1	int.c	Integrate C to 500 m above seafloor	mg C m ⁻²
	Decadal mean of monthly model simulation from January 1995 to December 2004	1 degree	1×1	int.n	Integrate N to 500 m above seafloor	mg N m ⁻²
		1 degree	1×1	det.c	Integrate detrital C to 500 m above seafloor	mg C m ⁻²
		1 degree	1×1	det.n	Integrate detrital N to 500 m above seafloor	mg N m ⁻²
		1 degree	1×1	phyt	Integrate phytoplankton to 500 m above seafloor	mg N m ⁻²
		1 degree	1×1	zoop	Integrate zooplankton to 500 m above seafloor	mg N m ⁻²
		1 degree	1×1	det.c.flx	Detrital C flux at 500 m above seafloor	mg C m ⁻² day ⁻¹
		1 degree	1×1	det.n.flx	Detrital N flux at 500 m above seafloor	mg N m ⁻² day ⁻¹
Bottom Water	<u>World Ocean Atlas 2009</u>	1 degree	1×1	temp	Temperature	°C
	Annual mean & seasonal standard deviation	1 degree	1×1	salin	Salinity	ppm
		1 degree	1×1	oxyg	Oxygen concentration	milliliters liter ⁻¹
		1 degree	1×1	nitra	Nitrate concentration	micromoles liter ⁻¹
		1 degree	1×1	phos	Phosphate concentration	micromoles liter ⁻¹
		1 degree	1×1	si	Silicate concentration	micromoles liter ⁻¹
Water Depth	<u>ETOPO1 Global Relief</u>	1 minute	N.A.	depth	Water depth	m

The mean value was extracted for abundance and biomass records with catchment area of 3×3 or 1×1 cells. The datasets are divided into 4 categories, including 1) primary productivity variables, 2) water column variables, 3) bottom water properties, and 4) water depth. The table abbreviations follow: Res. = data resolution, Cell = cell size for extraction, Abbrev. = variable abbreviation.

doi:10.1371/journal.pone.0015323.t001

exhaustive search of all available predictors to maximize the differences between the offspring branches. In RF, the response variable was bootstrap resampled before conducting RTA to generate large numbers of un-pruned decision trees (1000 trees in this study). Unlike traditional RTA, the RF algorithm searches the best split from a random subset of predictors (1/3 of all variables) and the prediction can be made from new data (environmental) by averaging the model outputs of all trees. At each bootstrap resampling step, 2/3 of the data (in-bag) were selected to build the decision tree. The other 1/3 of the data (out-of-bag, or OOB) were used to carry out an internal examination of model (decision tree) prediction error and estimate variable importance. The OOB data can generate predictions using the tree grown from the in-bag data. These OOB predictions were aggregated (by averaging the outputs of all trees) to compare with the observations and estimated the prediction error. The performance of the RF model was examined as percent variance explained: $R^2 = 1 - \text{MSE}_{\text{OOB}} / \text{observed variance}$, where MSE_{OOB} is the mean square error between observations and OOB predictions. Predictor Importance was determined by how much worse the OOB predictions can be if the data for that predictor are randomly permuted. This

essentially mimicked what would happen with or without the help of that predictor. The increase of prediction error (MSE_{OOB}) after the permutation was used to measure its contribution to the prediction accuracy. This accuracy importance measure (increase of MSE_{OOB}) was computed for each tree and averaged over the forest (1000 trees).

Construction of Random Forest Models

Standing stocks (biomass and abundance) were logarithm (base 10) transformed before conducting RF analysis. Environmental data were extracted based on the latitude and longitude of the stock records by averaging a box of size 3×3 or 1×1 cells (Table 1). Mean value of the box was matched to the corresponding stocks record. RF algorithm was then run independently on each of the 12 datasets. Most primary productivity predictors have declining temporal coverage at the high latitudes between years of 1998 and 2007 due to prolonged winter darkness or cloud cover preventing SeaWiFS ocean color measurements (Figure S2). This can be a source of error during the RF modeling, because decadal mean and standard deviation of the predictors was only calculated from the available monthly data. In order to evaluate the model

stability, we conducted 4 RF simulations for each dataset. The simulations were based on different data selection scenarios, including: 1) all standing stocks and environmental data were included; 2) only data calculated from >30 months of SeaWiFS measurements were included; 3) only data calculated from >60 months of SeaWiFS measurements were included; 4) only data calculated from >90 months of SeaWiFS measurements were included. In other words, Scenario 1 retained all the data and Scenario 4 excluded much of the high latitude data (>50°N or S, see Figure S2). The mean and standard deviation (S.D.) of the model performance (R^2) and variable importance were calculated to evaluate the model sensitivity. In the following text, the “simulations” refer to the RF runs under the 4 data selection scenarios.

Global Prediction of Seafloor Standing Stocks

Environmental data were averaged to the same grid resolution (1 arc degree grids) before using them as model inputs for global standing stocks predictions (Figure S1). For each dataset, 4 global predictions were generated from RF simulations. The mean and coefficient of variation (S.D./mean * 100%) were calculated for each grid to optimize the predictions and examine the output stability. In order to produce a smooth predicted surface, the predictions were interpolated to 0.1 degree cell resolution using Inverse Distance Weighting (IDW). The predicted map of standing stocks is displayed in color classes using Jenks Natural Breaks Optimization method to maximize the differences between the classes. The global integral of benthic biomass was integrated from each cell value multiplying the cell area on predicted map based on equidistant cylindrical projection. The calculations were based on the formula: Global integral = \sum map cell value (in per unit area) * cell area at equator ($\sim 12343 \text{ km}^2$) * cosine (latitude). Statistical analyses and RF modeling used R 2.11.0 [28] and R package randomForest [29]. Geostatistical analyses and mapping used ESRI® ArcMap™ 9.2 and R package sp [30].

Results

Partial linear regressions

Our results confirmed the conclusions of Rex et al. [8] and suggested significantly negative log-linear relationships of biomass, abundance, and body size for 3 large size classes with depth; however, none of these parameters showed statistically significant depth dependency for bacteria (Table 2). We adapted figure legends from Rex et al. [8] and raised the y-intercepts of their regression equations 3 orders of magnitude (converting the unit from g C m^{-2} to mg C m^{-2}) for comparison with our current results. Our regression y-intercepts were slightly lower than the previous synthesis (2.4 vs. 2.5 for bacteria; 2.2 vs. 2.3 for meiofauna; 3.1 vs. 3.2 for macrofauna; 1.8 vs. 2.3 for megafauna.), while the rate of decline biomass with depth was steeper for meiofauna (-2.4×10^{-4} vs. -1.7×10^{-4}) and macrofauna (-5.2×10^{-4} vs. -4.5×10^{-4}), but more gradual for megafauna (-3.1×10^{-4} vs. -3.9×10^{-4} , Table 2). The biomass hierarchy among size groups was similar between the 2 studies: macrofauna dominated the shelves and bacteria and meiofauna dominated the abyssal plains (Figure 2). The only apparent difference was a cross of the regression lines between macrofauna and megafauna at ~ 6000 m depth, or a reversal of their biomass hierarchies. The lower y-intercepts and steeper slopes for the meiofauna and macrofauna suggested that the biomass levels were lower in this study than in the previous synthesis. The rate of declining biomass with depth was highest for macrofauna, followed by megafauna and meiofauna. Except for meiofauna, the y-intercept of the

Table 2. Regression analyses of biomass, abundance, and body size against depth for bacteria, meiofauna, macrofauna, and megafauna.

Regression	Equations	N	F
Log10 Biomass (mg C m^{-2})			
Bacteria	$Y = 2.4 - (1.22e-06) X$	525	<0.01 n.s.
Meiofauna	$Y = 2.18 - (2.39e-04) X$	689	244.1***
Macrofauna	$Y = 3.05 - (5.15e-04) X$	2552	1885***
Megafauna	$Y = 1.81 - (3.07e-04) X$	282	136.2***
Log 10 Abundance (individual m^{-1})			
Bacteria	$Y = 13.27 - (3.58e-05) X$	515	2.82 n.s.
Meiofauna	$Y = 5.73 - (1.25e-04) X$	1148	184.7***
Macrofauna	$Y = 3.5 - (1.95e-04) X$	2734	618.2***
Megafauna	$Y = -0.68 - (2.82e-04) X$	253	32.92***
Log 10 Body Size ($\mu\text{g C individual}^{-1}$)			
Bacteria	$Y = -7.79 + (1.35e-05) X$	451	2.28 n.s.
Meiofauna	$Y = -0.61 - (6.81e-05) X$	616	27.6***
Macrofauna	$Y = 2.62 - (3.63e-04) X$	2393	637.3***
Megafauna	$Y = 6.17 - (1.57e-04) X$	136	43.58***

Response variables are \log_{10} transformed biomass (mg C m^{-2}), abundance (individual m^{-1}), and body size ($\mu\text{g C individual}^{-1}$). Predictor is depth (m). Scatter plots of the response variables against predictor and regression lines are given in Figures 2, 3, 4. Abbreviations: N = number of samples; ***denotes $P < 0.001$; n.s. = not significant.
doi:10.1371/journal.pone.0015323.t002

abundance-depth regressions were slightly lower in this study (13.3 vs. 14.1 for bacteria; 3.5 vs. 3.6 for macrofauna; -0.7 vs. -0.3 for megafauna.) while the slopes were more gradual (-2×10^{-4} vs. -2.8×10^{-4} for macrofauna; -2.8×10^{-4} vs. -3.7×10^{-4} for megafauna, Table 2). The rate of declining abundance with depth was sharpest for megafauna, followed by macrofauna and meiofauna (Figure 3, Table 2). Average body size for each size class was calculated as biomass divided by abundance. The average sizes of all 3 large groups showed significant depth dependency with the rates of declining mean size with depth being the most rapid for macrofauna, followed by megafauna and meiofauna (Table 2 and Figure 4). The rapid decline in average macrofaunal size was likely overestimated at abyssal depths, because the regression line was apparently higher at shelf depths due to extremely large values ($> 10 \text{ mg C individual}^{-1}$) at high latitude areas.

Random Forests

On average, RF models explained 78% to 81% of total variance (R^2) for bacteria, meiofauna, and macrofauna biomass (Figure 5a). Compared to the small size classes, the RF performance was subordinate for megafauna, invertebrates, and fishes, in which the models only explained 63% to 68% of the observed biomass variance. The RF algorithm appears to perform better for abundance with the models explaining 77% to 88% of total variance for each size class. The RF performance among different simulation scenarios was generally stable (S.D. $\leq 1\%$). The variability was only slightly higher for macrofauna and invertebrates with S.D. between 2% to 3%. A scatter plot between observed and predicted biomass (Figure 5b) suggests that the OOB predictions were in proper scale (regression slopes = ~ 1) with modest deviations from the observations.

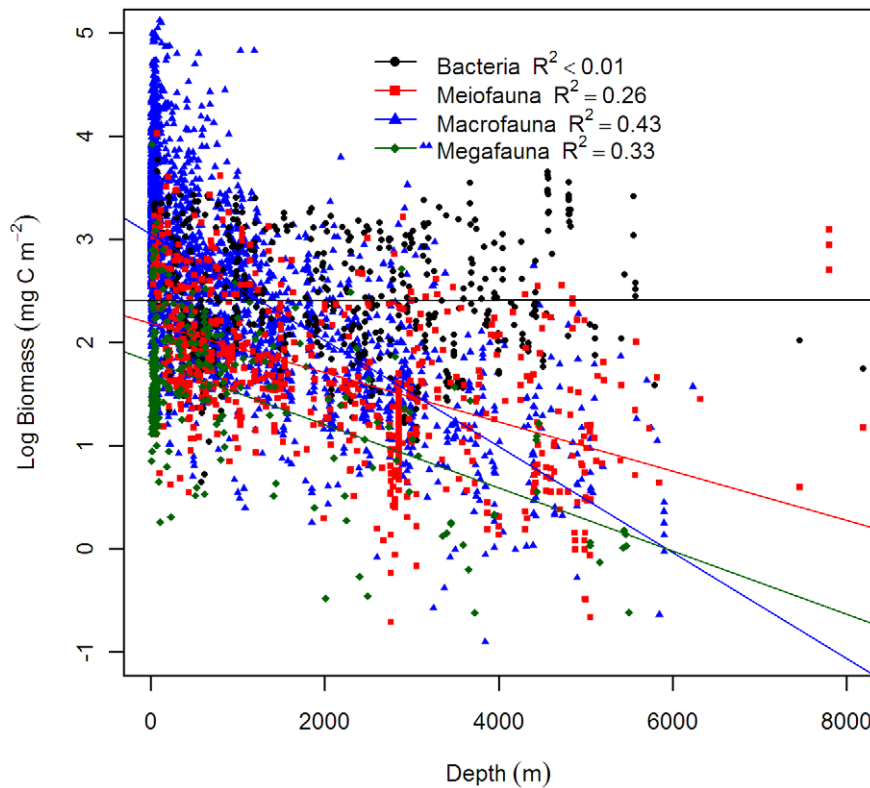


Figure 2. Biomass as a function of depth for bacteria, meiofauna, macrofauna, and megafauna. Biomass was \log_{10} transformed and the effects of latitude and longitude were removed by partial regression. Figure legend follows Rex et al. [8] for comparison. References of data source are available in Appendix S1 and File S1. Regression equations and test statistics for each size categories are available in Table 2. doi:10.1371/journal.pone.0015323.g002

We combined predictor importance from bacteria, meiofauna, macrofauna, and megafauna (Figure S3) to examine the predictor importance on total benthic biomass. This was only done for the biomass datasets because they were converted to a unified currency in mg C per square meter. With the exception of bacteria, depth was ranked highly important for the 3 larger size classes (Figure 6). To our surprise, neither net primary production (vgpm, cbpm) nor flux of detrital organic matter to seafloor (det.c.flx, det.n.flx) was considered the most important for the total benthic biomass. Instead, water depth and the data inputs for the NPP models (carbon, bbp, sst, par, mld, chl) were among the top 10 most important variables. Nonetheless, when the predictor importance was examined for the size classes, NPP models (vgpm, cbpm) had considerable importance for bacterial, meiofaunal, and macrofaunal biomass but appeared less important for megafaunal biomass. Decadal mean and S.D. of the predictors generally ranked in similar orders suggesting high correlation between them; however, it may also suggest that overall levels and seasonal fluctuations of the predictors were both important in predicting the biomass. The predictors associated with water column processes (Table 1) appeared not significant to the total biomass; however, the decadal mean of water column-integrated zooplankton (zoop.mean), total organic matter (int.c.mean, int.n.mean), and detrital organic matter (det.c.mean, det.n.mean), were among the most important predictors for megafaunal standing stocks, especially for abundance (see Figure S3d and Figure S4d). Annual mean salinity (salin.mean) was the only bottom water property ranked within the top 10 most important predictors for the total biomass (Figure 6).

Patterns of Predicted Biomass

No biomass predictions were given near the northern tip of the Arctic Ocean and part of the Antarctic shores due to a lack of SeaWiFS satellite data as a result of permanent sea ice cover (Figure S2). The predictions of major size classes (Figure S5a, b, c, d) were combined to estimate the total benthic biomass. The maximum biomass of 2.6 to 10 g C per square meter occurred on the shelves of the north frigid zones (e.g. Kara Sea, Siberian Sea, and Chukchi/Bering Sea) and temperate areas (e.g. Yellow sea and North Sea, see Figure 7, red color). These predictions however were lower than the empirical maximum found in the Chukchi/Bering Sea, where the infauna biomass as high as 40 to 100 g C m^{-2} were reported [31]. The discrepancy is probably associated with high prediction uncertainty in the areas (C.V. = 15% to 22%, Figure 8) or unexplained variability in the models (Figure 5a). The weaker maximum (orange color) between 1.3 to 2.5 g C per square meter occurred on the polar to temperate shelves and subtropical coastal areas (e.g. East/South China Sea, Arabian Sea, and Persian Gulf). The lowest biomass prediction between 30 and 80 mg C per square meter occurred on the abyssal plains of the Pacific, Atlantic, and Indian Ocean; however, relatively higher biomass was predicted on the seafloor of the east side of Pacific and Atlantic basins under the productive equatorial divergence and coastal upwelling areas [32]. For these largest ocean areas, the model outputs were stable among 4 RF simulations with S.D. less than 10% of the mean predictions (Figure 8, light blue to dark blue colors). Any high uncertainties were usually associated with high predicted biomass. The Southern Ocean for example had the highest uncertainty with S.D. between 15% and 26% of the mean

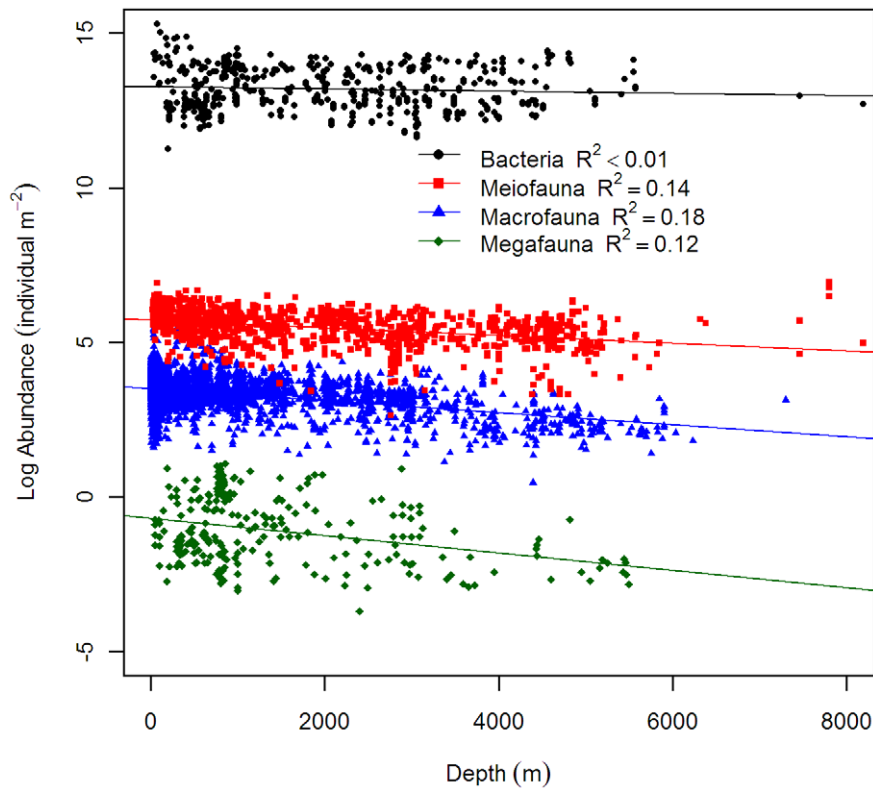


Figure 3. Abundance as a function of depth for bacteria, meiofauna, macrofauna, and megafauna. Abundance was \log_{10} transformed and the effects of latitude and longitude were removed by partial regression. Figure legend follows Rex et al. [8] for comparison. References of data source are available in Appendix S1 and File S1. Regression equations and test statistics for each size category are available in Table 2. doi:10.1371/journal.pone.0015323.g003

(yellow to red class), where most of the uncertainty was derived from the unstable predictions for macrofauna biomass (Figure S6). The S.D. of some Arctic shelves were slightly lower than the Southern Ocean, mostly between 11% and 18% of the mean (green to yellow class, Figure 8). The \log_{10} predicted biomass (mg C m^{-2}) and abundance (individual m^{-2}) for each size class are available in File S2 and File S3, respectively. Global maps showing the mean of abundance prediction and coefficient of variation for each size class are given in Figure S7 and Figure S8, respectively.

A total of 110.3 ± 48.2 (Mean \pm S.D. from 4 RF simulations) megatons of living carbon biomass were estimated based on the global integral of the predicted map cells (Figure 7), in which bacteria, meiofauna, macrofauna, and megafauna contributed 31.4%, 12.9%, 50.7%, and 5% of the global integral, respectively. Previous workers estimated that global POC flux to the seafloor was 3.76 to 3.91 megaton C day^{-1} [25,33] and carbon burial was about 0.82 megaton C day^{-1} [33]. By dividing the total mass by the flux [34,35], we estimated that the mean residence time for the seafloor living carbon was 36.6 ± 16 days (mean \pm S.D.). Generally, the predictions were highest on the continental shelves, which account for 21.1% of the global integral biomass but cover merely 5.9% of the total seafloor area (≤ 200 m water depth, Figure 9a). Water depths deeper than 3000 m harbor more than 50% of the global benthic biomass due to their vast area (covering $>75\%$ of seafloor). The predictions were also high at high latitudes ($> 60^\circ\text{N}$ or S) and the tropical ocean ($< 23.5^\circ\text{N}$ or S) of the northern and southern hemisphere, in which the biomass contributed 25.4% and 28.8% of the global integral on 13.4% and 40.7% of the ocean area, respectively (Figure 9b). As a rule of thumb, the total biomass of all size classes (except for bacteria) dissipates along the

continental margins to the abyssal plains (Figure 2) but this is accompanied by a major shift in size classes in the predictions, with the biomass composition changing from metazoan dominated (meiofauna + macrofauna) for the first couple hundred-meter zonal integrals to bacteria dominated on the abyssal plain (Figure 9a). Along the latitudinal zonal integrals, the biomass composition also shifted from the majority of large-size macrofauna at high latitudes to the small-size meiofauna and bacteria dominated at the tropics (Figure 9b).

Regional variability among the major ocean basin is apparent when predicted biomass was plotted against depth (Figure 10). Generally, the declining trends of biomass with depth were similar but the overall levels differed by basin, with the predictions bounded between the higher end of the Southern Ocean and the lower end of the Mediterranean Sea (Figure 10h). In the Atlantic and Arctic Ocean, high density at bathyal depths near the upper end of the biomass–depth distribution (Figure 10a, e) appeared responsible for elevated biomass levels above the Pacific, Indian Ocean, and Gulf of Mexico (Figure 10). These high values corresponded to the high biomass predictions in the North Atlantic to Arctic Ocean (Figure 7) under the productive subpolar gyre north of the Gulf Stream [32]. The high density at the bottom of the biomass–depth distribution for the Atlantic and Pacific Oceans (Figure 10a, b) illustrates the low predicted biomass on the vast abyssal plains. In the Indian Ocean, the extraordinary high predicted values between ~ 1200 to 3000-m water depths (Figure 10c) single out the Oman and Pakistan Margin, where the benthic biomass between 1.3 and 2.5 g C per square meter is as high as continental shelf values (Figure 7, orange color). We believe that the high predictions derive mainly from the monsoon

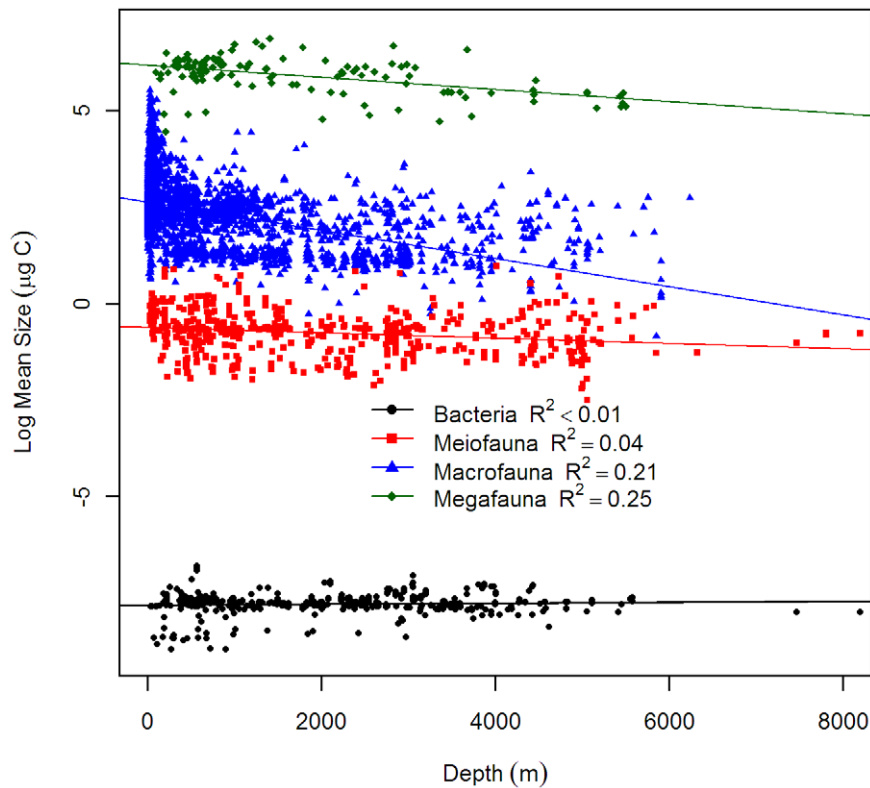


Figure 4. Average body size as a function of depth for bacteria, meiofauna, macrofauna, and megafauna. The average size was calculated by dividing biomass with abundance. The body size was \log_{10} transformed and the effects of latitude and longitude were removed by partial regression. Figure legend follows Rex et al. [8] for comparison. References of data source are available in Appendix S1 and File S1. Regression equations and test statistics for each size categories are available in Table 2. doi:10.1371/journal.pone.0015323.g004

dynamics and seasonal fluctuation of export POC flux [36] rather than the mid-water Oxygen Minimum Zones (OMZ), because resolution of our bottom oxygen data (Table 1) is probably not sufficient to detect OMZ influences. At hadal depths (>6000 m), the biomass predictions were meager in general ($<0.2 \text{ g C m}^{-2}$, Figure 10a, b); however, relatively high values ($0.5\sim 0.7 \text{ g C m}^{-2}$) were predicted near the Kurile-Kamchatka Trench of the Northwest Pacific Basin (Figure 10b) and the South Sandwich Trench near the southern tip of the South America and Antarctic Peninsula (Figure 10d).

Discussion

Observed and Predicted Patterns

In this study, classic log-linear declines of seafloor biomass and abundance with depth were demonstrated for meiofauna, macrofauna, and megafauna [7,8,9]. These widely recognized patterns have been attributed to the decreasing quantity and quality of sinking phytodetritus with increasing depth and distance from the productive coastal waters and river runoff [7,37]. Although the selection pressure (food limitation) may be the same, responses differed among the size groups along the depth gradients, showing disparate rates of declining biomass and shifts of biomass hierarchy from macrofauna domination on the shelves and upper slope to meiofauna and bacteria domination on the abyssal plains [8,34,35,38]. Figure 4 suggests that these observed biomass patterns among size groups are governed by the rate of declining average body size rather than by the rate of declining abundance with depth. The decrease of animal size in the deep-sea

has been explained by energy constraints and the need to maintain viable density for successful reproduction [8,39]. Recent evidence from terrestrial environments also suggests a potential link between the animal body size and food quality [40]. It has been suggested that the macrofauna may compete for fresh settled phytodetritus with bacteria [41,42,43,44], while the meiofauna may prefer bacterial carbon over phytodetritus [45]. Hence, the rapid decline of macrofaunal average size with depth could be related to the exponential decrease of sinking detrital carbon or the refractory organic matter in the deep-sea sediments. The meiofauna, on the contrary, may be less affected by the deterioration of the food influx and experienced a relatively gradual decline of average size with depth; however, the actual causes of this discrepancy in size-structure remain unclear.

Interestingly, our predicted biomass not only has captured the shifts of dominant size groups with depth but also with latitude (Figure 9), supporting the dominance of macrofaunal biomass [31,46] and meager importance of bacteria at the high latitudes [47] due, potentially, to strong benthic-pelagic coupling, short food chain, and weaker microbial loop in the overlying water [48,49]. Other intriguing features from our predictions include the apparent increase of bacterial, meiofaunal, and decrease of macrofaunal biomass integrals from high latitudes toward the tropical oceans (Figure 9b). In fact, the increasing bacterial and meiofaunal integrals were a function of the increasing cell areas toward the equator due to the map projection, which in turn makes the decrease of macrofaunal integrals seemingly even more convincing. This cross-latitude comparison however could be biased by a potential interaction with water depth, because the

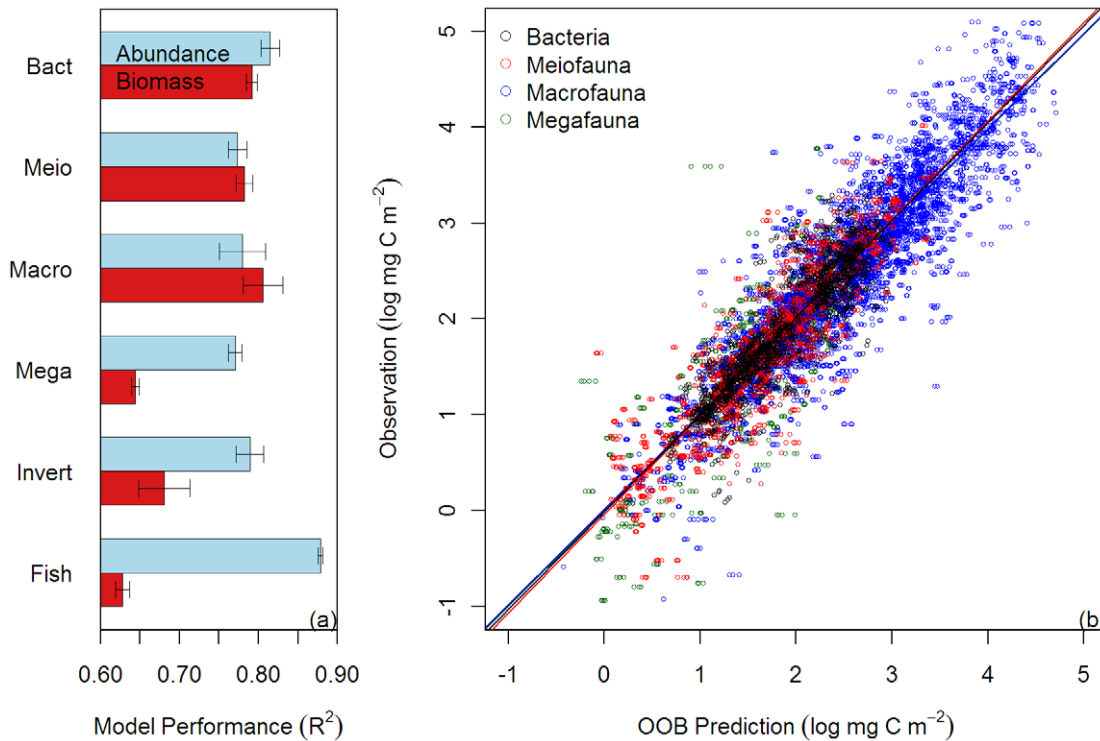


Figure 5. Random Forests (RF) performance on biomass and abundance of each size class. (a) Mean percent variance explained by the RF model \pm S.D. (error bar) from 4 RF simulations. Abbreviations: Bact = bacteria, Meio = meiofauna, Macro = macrofauna, Mega = megafauna, and invert = invertebrates. (b) Observed against OOB predicted biomass from the 4 RF simulations. Color legends indicate 4 major size classes. doi:10.1371/journal.pone.0015323.g005

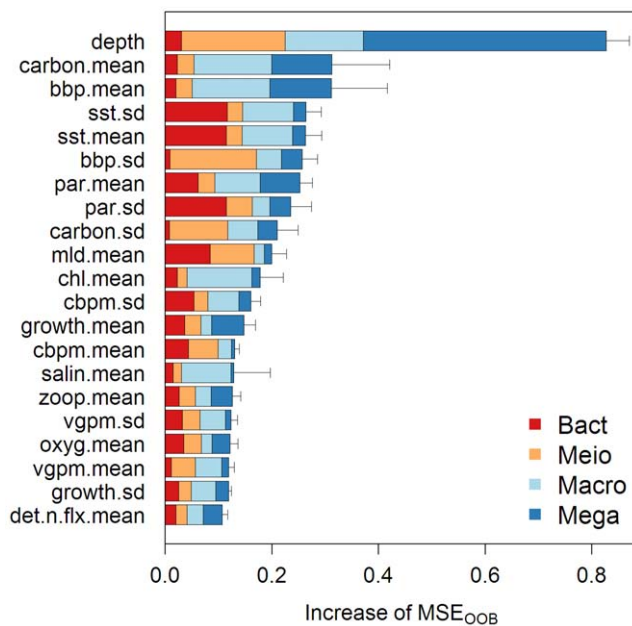


Figure 6. Mean predictor Importance on total seafloor biomass. The predictor importance of major size classes were combined (Figure S3) and mean \pm S.D. (error bar) was calculated from 4 RF simulations. The top 20 most important variables are shown in descending order. Increase of mean square error (MSE_{OOB}) indicates the contribution to RF prediction accuracy for that variable. doi:10.1371/journal.pone.0015323.g006

tropical oceans comprise many deep basins and the high latitudes, such as Chukchi/Bering Sea, have extended shelf areas. We tested this by using partial regression to statistically remove the effect of water depth and longitude. When depth was held constant, macrofaunal biomass could be fitted to a positive parabolic function of latitude ($R^2 = 0.17$, $P < 0.001$), supporting the elevated macrofaunal biomass at high latitudes [7].

From a global perspective, the results of regressions (Figures 2, 3, 4) reinforced the weak to no depth-dependency of bacterial standing stocks [8,50,51]. Despite immense variation in declining POC flux at depth, the surface sediments supported a remarkably constant bacterial stock spanning only ~ 2 orders of magnitude difference worldwide (30 to 2220 $mg\ C\ m^{-2}$ and 1.3×10^{12} to 1.9×10^{14} cells m^{-2} , 5th to 95th percentile, $n = 525$); nonetheless, regional and local studies in our database do indicate dependency of bacterial standing stocks with depth or POC flux [10,52,53]. The high bacterial stocks at the supposedly depauperate abyssal depths have been attributed to their barophilic adaption [54,55]. As bacteria are transported with phytoditrital aggregates to the deep sea [56], a large number of the bacteria could be dormant or inactive because of the extreme pressure and frigid temperature [57,58], while the active microbes are supported by carbon deposition flux [43], viral lysis of the infected prokaryotes [59], extracellular enzymatic activities [60,61], and benthic metazoan sloppy feeding [44]. It is worth noting that many studies have applied a uniformed conversion factor to estimate the biomass from bacterial numbers, which may be the main reason that no statistical relationship was detected between the bacterial cell size and water depth (Figure 4). Based on direct measurements of the cell volume over a wide range of water depths in the northern Gulf of Mexico, Deming and Carpenter [52] concluded that the greater ocean depths generally harbored smaller bacterial cells despite the

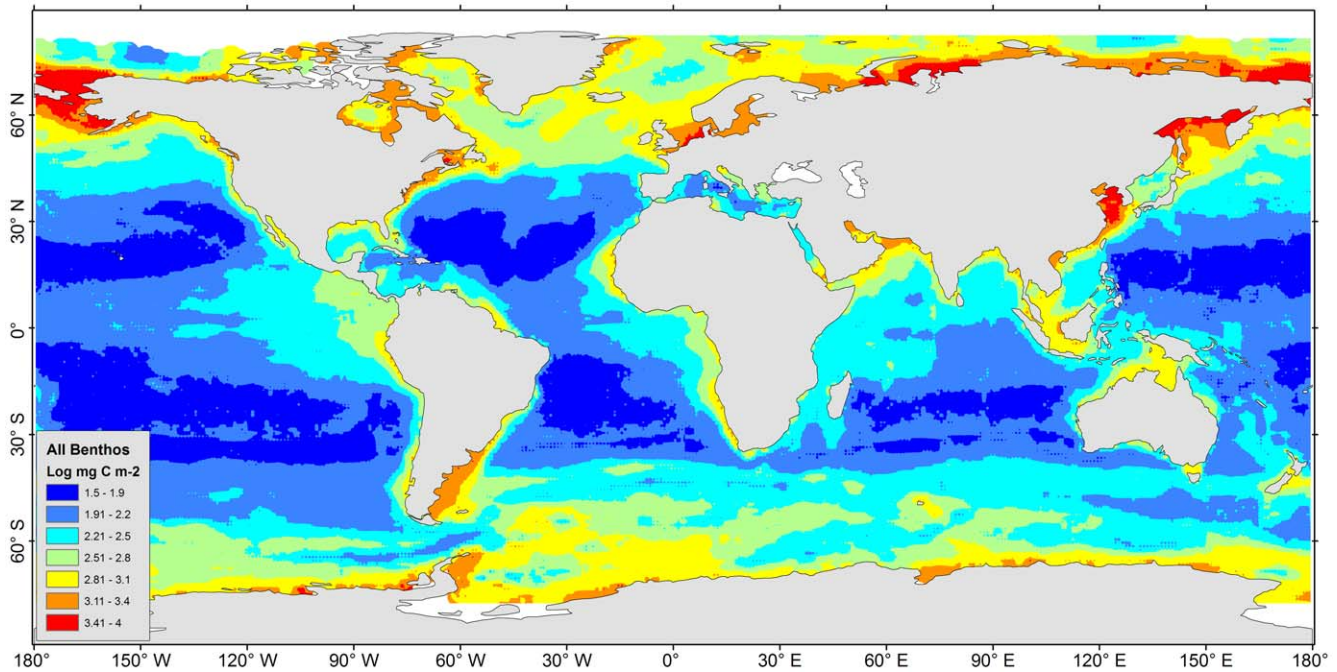


Figure 7. Distribution of seafloor biomass predictions. The total biomass was combined from predictions of bacteria, meiofauna, macrofauna, and megafauna biomass (Figure S5a, b, c, d). Map was smoothed using Inverse Distance Weighting interpolation to 0.1 degree resolution and displayed in logarithm scale (base of 10). doi:10.1371/journal.pone.0015323.g007

abundance remaining constant. That is, the constancy of bacterial biomass with depth that we observed here could be an artifact because the cell volumes were not measured directly at all depths. To our surprise, even though no depth-dependency was evident for the bacterial standing stocks, the RF algorithm performed well

in predicting the bacterial biomass ($R^2 = 79 \pm 0.6\%$) and abundance ($R^2 = 81 \pm 1.2\%$, mean \pm S.D, $n = 4$). High predictor importance of sea surface temperature (sst), irradiance (par), mixed layer depth (mld), and carbon-based primary production model (cbpm) support the idea that the sedimentary bacterial biomass may be

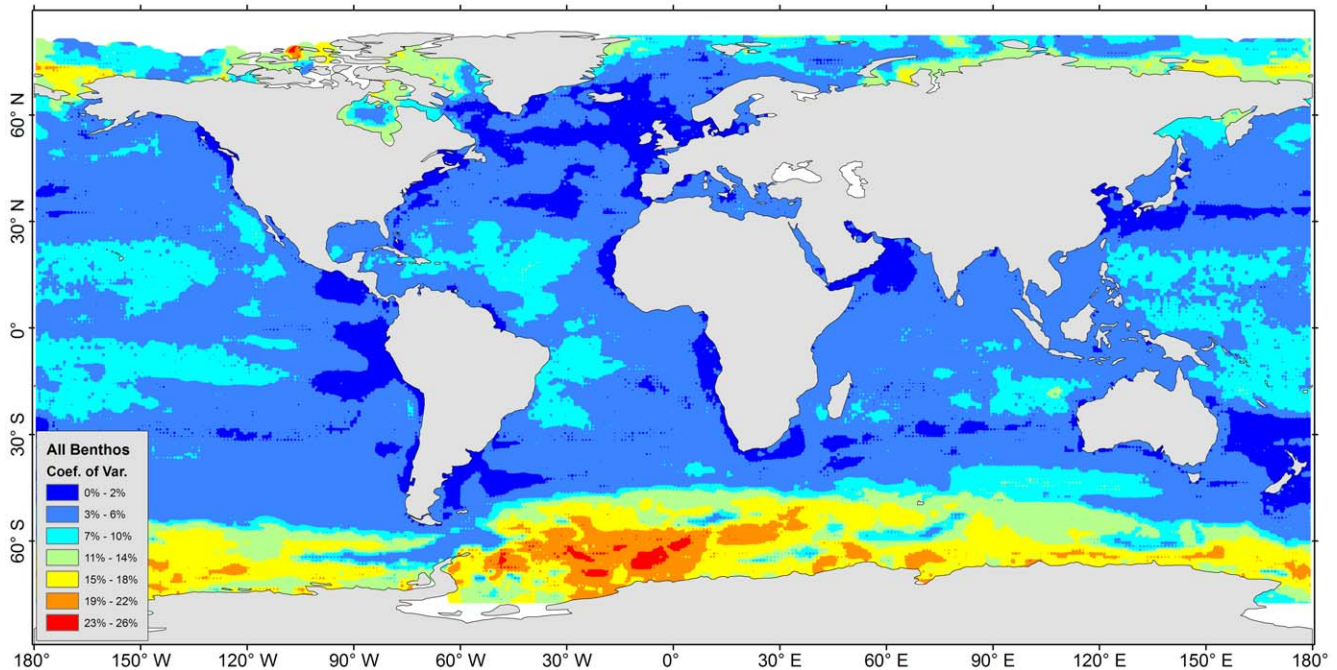


Figure 8. Coefficient of variation (C.V.) for mean seafloor biomass prediction. The C.V. was computed as S.D./mean * 100% from 4 RF simulations. Map was smoothed using Inverse Distance Weighting interpolation to 0.1 degree resolution. doi:10.1371/journal.pone.0015323.g008

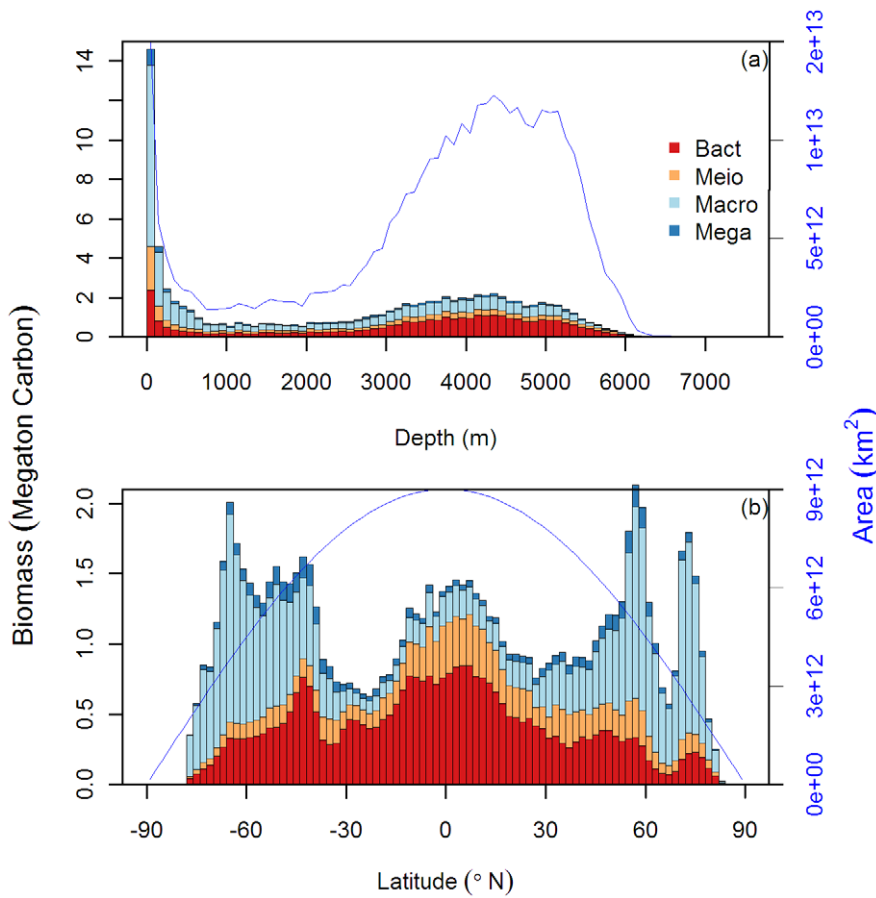


Figure 9. Global zonal integrals of benthic biomass (bars) in unit of megaton carbon based on 100-m bins (a) and 2-latitude-degree bins (b). The blue line shows integrals of seafloor area in unit of square kilometer. Color legends indicate 4 major size classes. doi:10.1371/journal.pone.0015323.g009

imported in the form of sinking particles [54,56]. The high bacterial biomass predictions on the abyssal plains of semi-enclosed basins, such as the Gulf of Mexico, Arabian Sea, and East Mediterranean (Figure S5a), supported potential lateral advection of detritus from the margins due to relatively large area of shelves and margins compared to basin volume [52].

Anomalies not explained by Random Forests

Although multiple predictors were obtained to cover as many aspects and processes that could affect the distribution of benthic standing stocks, around 19% to 36% of observed variances are still unexplainable in the current RF models. Some important predictors, such as sediment grain size [62], organic composition [63], bioturbation [64,65], and community oxygen demand [66,67], were not included due to sparse data availability; others such as oxygen minimums [68,69] or abrupt changes in thermal dynamic regimes [70], could also be left undetected due to the coarse resolution in available hydrographic data. Nevertheless, the largest unexplained variability was probably derived from our non-contemporaneous predictors that do not account for the seasonal or inter-annual changes of benthic standing stocks as a result of climate-induced variations on productivity and export POC flux [71,72]. The seafloor organisms depend on diverse sources of energy [73], including large food falls [74], hydrocarbons from cold seeps and hydrothermal vents [75,76], lateral resource advection from continental margins [77], accumulation of organic matter in submarine canyons [78] and trenches [79],

and rapid energy transfers on seamounts [80]. In addition, benthic foraminifera, sometimes accounting for more than 50% of eukaryote biomass [81], are not included in our datasets. These anomalies are not in the scope of this analysis and should be estimated separately elsewhere in a global context. For example, at the head of the New Zealand’s Kaikoura Canyon (data not in the database), the extremely high macrofauna and megafauna biomass (89 g C m⁻²) was about 100-fold more than our total biomass prediction (0.94 g C m⁻²) [82]. Within the datasets, extraordinary high “total biomass” was also reported at the head of the Mississippi Submarine Canyon [35] due to dominance of a “carpet of worms” [83]. The observed biomass was still more than 4-fold higher than our prediction. This is partially because the Gulf of Mexico basin had very high background bacterial biomass [52]. When the bacteria component is removed, the prediction still under estimates the observed biomass by about 50%. Hence, the total living carbon prediction in this study (Figure 7) should be considered as a conservative estimate for the soft bottom communities solely relying on sinking phytodetritus, with the anomalies causing the observed biomass to deviate from this baseline (Figure 5b).

Predictor Importance

We tested the RF algorithm using only the primary productivity predictors (decadal mean and S.D. of chl, sst, par, bbp, mld, growth, carbon, vgpm, and cbpm) and depth (Table 1). We found that the reduced models only experienced modest deterioration in perfor-

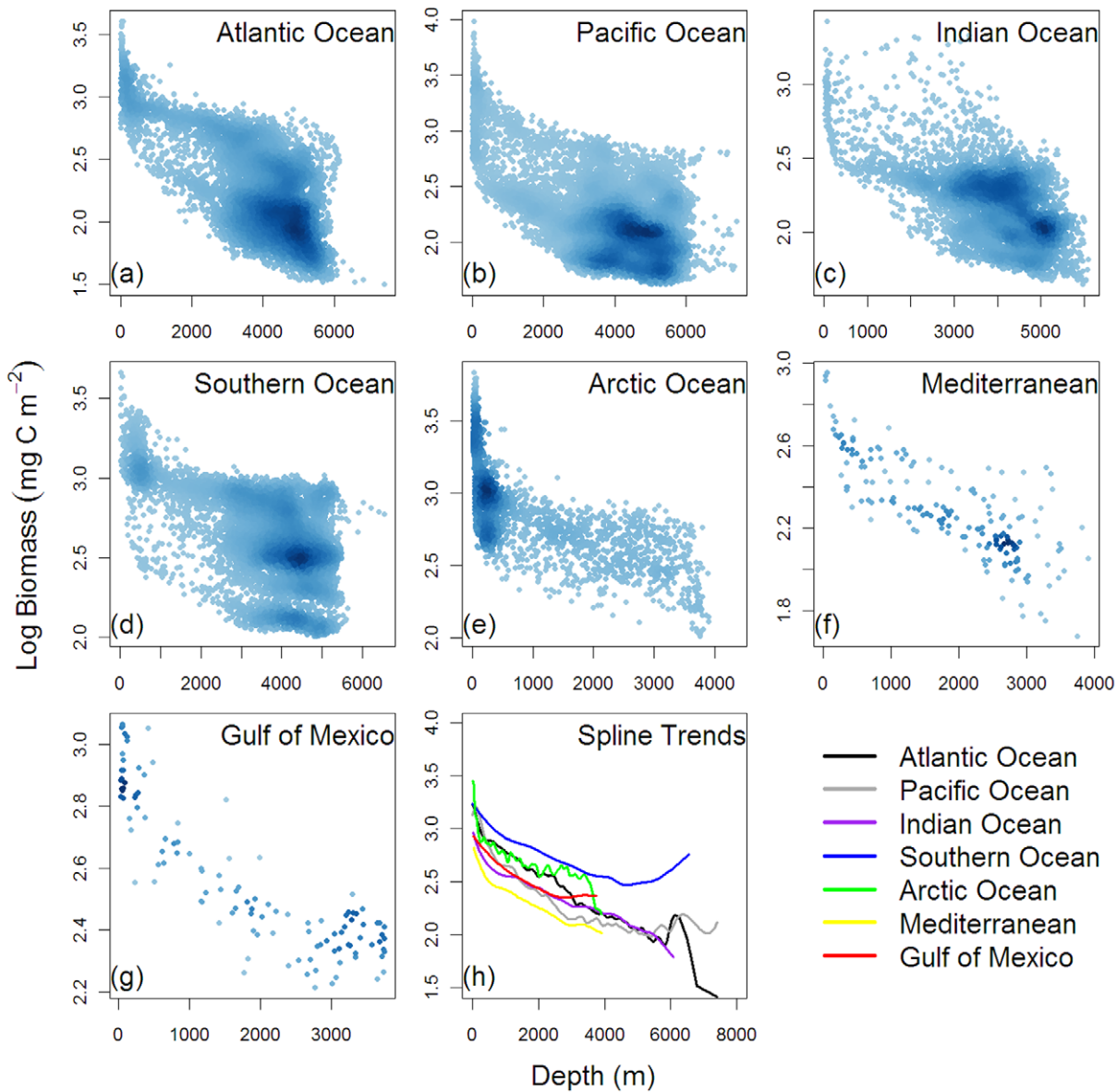


Figure 10. Seafloor biomass predictions against depths for the (a) Atlantic Ocean, (b) Pacific Ocean, (c) Indian Ocean, (d) Southern Ocean, (e) Arctic Ocean, (f) Mediterranean Sea, and (g) Gulf of Mexico. Blue color gradient indicates kernel density estimates. Panel (h) shows the regional predicted trends based on smoothing spline function. Color legend indicates the spline trends for each basin. doi:10.1371/journal.pone.0015323.g010

mance ($R^2_{\text{reduced}} = 75\%–80.3\%$ for biomass of 3 small size classes and $R^2_{\text{reduced}} = 63\%$ for megafauna biomass; $R^2_{\text{reduced}} = 76.3\%–80.6\%$ for abundance of 4 major size classes), suggesting that these productivity/depth predictors alone can explain much of the observed stock variances. It is also evident that these satellite-based ocean color parameters and depth are among the most important predictors when the full RF models were constructed (Figure 6). Their importance was even greater than the model estimates of export phytodetritus flux (det.c.flx & det.n.flx, Table 1) that have been considered important for benthic communities [43,56,84,85]. One possibility is that not all export flux is utilized by the benthos [35] and the combination of productivity/depth predictors simply explain the stock variances better; however, the spurious correlations between these predictors could also make them all rank highly important. Strobl et al. [86] recommended “Conditional Permu-

tation” while calculating the variable importance to reduce the effect of spurious correlations. We did not attempt this analysis because our focus was on prediction rather than pinpointing the exact contribution of each predictor.

Conclusions

The fate of sinking phytodetritus flux to the ocean floor and the energy transfer to the benthos is a complex biogeochemical process. The combination of mechanistic primary productivity models [23,24] and empirical relationship of export POC flux at depth [87] may not properly reflect the actual benthic food influx and consumption. In this study, we demonstrated that the combination of multivariate predictors and machine-learning algorithm was superior to conventional regression models using only water depth or export POC flux to predict benthic standing

stocks [8,88]. Conceptually, the RF predicted biomass presented here (Figure 7) can be seen as non-linear transformation of the surface primary production through a sophisticated decision network and is thus potentially a more realistic reflection of benthic food supply or utilization. Benthic biomass is essential to understand the dynamic processes of global carbon cycling [89] and productivity-diversity relationship in the deep sea [1,2]. Predictive mapping of this kind can fill the gaps where critical biomass information is lacking, since a true ‘census’ of global living carbon is expensive and practically impossible. Accurate prediction of benthic biomass can facilitate Ecosystem Based Management (EBM) on socioeconomically important species [90]. It is also extremely useful for generating and testing large scale hypotheses (e.g. latitudinal and cross-basin comparison) and planning shipboard surveys. Moreover, the reduced RF models mentioned above can be used to perform fine-scale predictions with high resolution ocean color images (5 arc minute grids) and the global relief model (1 arc minute grids, Table 1), and potentially reveal more heterogeneous biomass patterns at local scale than the current coarse analysis framework. The ocean color/depth predictors also make it possible to do contemporaneous modeling with recent sampling (SeaWiFS data are only available since 1997) or data collected in the future. This study presents an initial framework for archiving the seafloor standing stock data. More training datasets from diverse environments matched in space and time are urgently needed to improve the model performance and prediction accuracy, and perhaps, in due course, the seafloor standing stocks can be now-casted using the current ocean climate or even forecasted under the future climate scenarios [91].

Supporting Information

File S1 Google Earth file for the “CoML fresh biomass database”.
(KML)

File S2 Global seafloor biomass predictions. Predicted biomass (mg C m^{-2}) is in global 1×1 degree grids. Data fields include latitude, longitude, depth, and biomass of each size class. The biomass data are in logarithm scale (base 10).
(CSV)

File S3 Global seafloor abundance predictions. Predicted abundance (individual m^{-2}) is in global 1×1 degree grids. Data fields include latitude, longitude, depth, and abundance of each size class. The abundance data are in logarithm scale (base 10).
(CSV)

Appendix S1 The complete list of references for the “CoML Fresh Biomass Database”.
(DOC)

Figure S1 Environmental predictors for Random Forest models. Data were logarithm transformed (base 10) and scaled to between 0 (minimum value) and 1 (maximum value). Detail description of the variable is given in Table 1. Abbreviations: mean = decadal or annual mean; sd = decadal or seasonal standard deviation.
(TIFF)

Figure S2 Temporal coverage of primary productivity predictors between years of 1998 and 2007. Color ramp shows the sample size from 0 to 120 months of measurements. Detail description of the variable is given in Table 1. Abbreviations: n = sample size.
(TIFF)

Figure S3 Mean predictor Importance for biomass of (a) bacteria, (b) meiofauna, (c) macrofauna, and (d) megafauna. The mean \pm S.D. (error bar) were calculated from 4 RF simulations. The top 20 most important variables are shown in descending order. Increase of mean square error (IncMSE) indicates the contribution to RF prediction accuracy for that variable. Detail description of the variable is given in Table 1. Abbreviations: mean = decadal or annual mean; sd = decadal or seasonal standard deviation.
(TIFF)

Figure S4 Mean predictor Importance for abundance of (a) bacteria, (b) meiofauna, (c) macrofauna, and (d) megafauna. The mean \pm S.D. (error bar) were calculated from 4 RF simulations. The top 20 most important variables are shown in descending order. Increase of mean square error (IncMSE) indicates the contribution to RF prediction accuracy for that variable. Detail description of the variable is given in Table 1. Abbreviations: mean = decadal or annual mean; sd = decadal or seasonal standard deviation.
(TIFF)

Figure S5 Distribution of mean biomass predictions for (a) bacteria, (b) meiofauna, (c) macrofauna, (d) megafauna, (e) invertebrates, and (f) fishes. The mean biomass was computed from 4 RF simulations. Predictions were smoothed by Inverse Distance Weighting interpolation to 0.1 degree resolution and displayed in logarithm scale (base of 10).
(TIFF)

Figure S6 Coefficient of variation (C.V.) for mean biomass predictions of each size class. The C.V. was computed as $\text{S.D./mean} * 100\%$ from 4 RF simulations. The abbreviations are: bact = bacteria, meio = meiofauna, macro = macrofauna, mega = megafauna, inv = invertebrates, fis = fishes.
(TIFF)

Figure S7 Distribution of mean abundance predictions for (a) bacteria, (b) meiofauna, (c) macrofauna, (d) megafauna, (e) invertebrates, and (f) fishes. The mean abundance was computed from 4 RF simulations. Predictions were smoothed by Inverse Distance Weighting interpolation to 0.1 degree resolution and displayed in logarithm scale (base of 10).
(TIFF)

Figure S8 Coefficient of variation (C.V.) for mean abundance predictions of each size class. The C.V. was computed as $\text{S.D./mean} * 100\%$ from 4 RF simulations. The abbreviations are: bact = bacteria, meio = meiofauna, macro = macrofauna, mega = megafauna, inv = invertebrates, fis = fishes.
(TIFF)

Acknowledgments

We thank Paul Snelgrove and Michele Durant for overseeing the Census of Marine Life (CoML) synthesis groups and putting their trust and confidence in a graduate student (C.-L. Wei). We also thank an anonymous reviewer for comments to improve the manuscript. We thank Nikolaos Lampadariou, Agnes Muthumbi, Karline Soetaert, Jan Vanaverbeke, and Maria Franco for providing nematode data and Tone Falkenhaug for providing pelagic decapod data. We are grateful to John Dunne for generously providing export POC model from his publication, and to Peter Lawton, Nick Ellis, Stephen Smith, Lewis Incze, Michelle Greenlaw, Nicholas Wolff, and Tom Shirley for discussing modeling methods with the first author (C.-L. Wei).

Author Contributions

Conceived and designed the experiments: CLW GTR EEB. Analyzed the data: CW. Contributed reagents/materials/analysis tools: CLW GTR EEB

AB TS MJC YS FH FQ ZY CRP RLH MKW MAR JGB JS RD IRM CCN JWD PM ML JMW MWK BSI BJB DSMB AY BAB KI BEN. Wrote the paper: CLW GTR.

References

- Rowe GT, Wei C-L. Biodiversity of Deep-Sea Macrofauna as a Function of Food Supply. In preparation.
- Rex M, Etter R (2010) Deep-sea biodiversity: pattern and scale. Cambridge MA: Harvard University Press. 354 p.
- Petersen CGJ (1918) The sea-bottom and its production of fish-food. Report of the Danish Biological Station 25: 1–62.
- Petersen CGJ (1913) Valuation of the sea, Part 2. The animal communities of the sea-bottom and their importance for marine zoogeography. Report of the Danish Biological Station 21: 1–43.
- Holme N, McIntyre A (1971) Methods for the study of marine benthos. IBP Handbook No. 16 Oxford, UK: Blackwell.
- Vinogradov M, Tseitlin V (1983) Deep-sea pelagic domain (aspects of bioenergetics). In: Rowe GT, ed. Deep-sea biology. New York: Wiley-Interscience. pp 123–165.
- Rowe GT (1983) Biomass and production of deep-sea macrobenthos. In: Rowe GT, ed. Deep-Sea Biology. New York: Wiley-Interscience. pp 97–121.
- Rex MA, Etter RJ, Morris JS, Crouse J, McClain CR, et al. (2006) Global bathymetric patterns of standing stock and body size in the deep-sea benthos. *Marine Ecology Progress Series* 317: 1–8.
- Soltwedel T (2000) Metazoan meiobenthos along continental margins: a review. *Progress In Oceanography* 46: 59–84.
- Deming J, Yager P (1992) Natural bacterial assemblages in deep-sea sediments: towards a global view. In: Rowe GT, Pariente V, eds. Deep-sea food chains and the global carbon cycle. Dordrecht, The Netherlands: Kluwer. pp 11–27.
- Breiman L (2001) Random forests. *Machine Learning* 45: 5–32.
- Cutler A, Stevens JR (2006) Random Forests for Microarrays. In: Alan K, Brian O, eds. *Methods in Enzymology*: Academic Press. pp 422–432.
- Xu P, Jelinek F (2004) Random forests in language modeling. *Proceedings of EMNLP 2004*. Barcelona, Spain: Association for Computational Linguistics. pp 325–332.
- Svetnik V, Liaw A, Tong C, Culbertson JC, Sheridan RP, et al. (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* 43: 1947–1958.
- Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, et al. (2007) Random forests for classification in ecology. *Ecology* 88: 2783–2792.
- Prasad A, Iverson L, Liaw A (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* 9: 181–199.
- De'ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88: 243–251.
- Oppel S, Huettmann F (2010) Using a Random Forest Model and Public Data to Predict the Distribution of Prey for Marine Wildlife Management. In: Cushman SA, Huettmann F, eds. *Spatial Complexity, Informatics, and Wildlife Conservation*. pp 151–163.
- Pitcher CR, Lawton P, Ellis N, Smith SJ, Incze LS, et al. Exploring the role of environmental variables in shaping patterns of biodiversity composition in seabed assemblages. In preparation.
- Bett B, Vanreusel A, Vincx M, Soltwedel T, Pfannkuche O, et al. (1994) Sampler bias in the quantitative study of deep-sea meiobenthos. *Marine Ecology Progress Series* 104: 197–197.
- Gage JD, Hughes DJ, Vecino JLG (2002) Sieve size influence in estimating biomass, abundance and diversity in samples of deep-sea macrobenthos. *Marine Ecology Progress Series* 225: 97–107.
- Gage J, Bett B (2005) Deep-sea benthic sampling. In: Eleftheriou A, McIntyre A, eds. *Methods for the Study of the Marine Benthos*. Oxford, UK: Blackwell. pp 273–325.
- Behrenfeld MJ, Falkowski PG (1997) Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography* 42: 1–20.
- Westberry T, Behrenfeld M, Siegel D, Boss E (2008) Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles* 22: GB2024.
- Yool A, Shepherd JG, Bryden HL, Oeschle A (2009) Low efficiency of nutrient translocation for enhancing oceanic uptake of carbon dioxide. *J Geophys Res* 114: C08009.
- Amante C, Eakins BW (2009) ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24. 19 p.
- Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Belmont CA: Wadsworth.
- R Development Core Team (2010) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2: 18–22.
- Bivand RS, Pebesma EJ, Gomez-Rubio V (2008) Applied spatial data analysis with R. New York: Springer.
- Grebmeier JM, Cooper LW, Feder HM, Sirenko BI (2006) Ecosystem dynamics of the Pacific-influenced Northern Bering and Chukchi Seas in the Amerasian Arctic. *Progress In Oceanography* 71: 331–361.
- Williams RG, Follows MJ (2003) Physical transport of nutrients and the maintenance of biological production. In: Fasham M, ed. *Ocean Biogeochemistry: The role of the ocean carbon cycle in global change*: Springer. pp 19–51.
- Dunne J, Sarmiento J, Gnanadesikan A (2007) A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor. *Global Biogeochem Cycles* 21: GB4006.
- Rowe GT, Sibuet M, Deming JW, Khrifounoff A, Tietjen J, et al. (1991) "Total" sediment biomass and preliminary estimates of organic carbon residence time in deep-sea benthos. *Marine Ecology Progress Series* 79: 99–114.
- Rowe GT, Wei C-L, Nunnally C, Haedrich R, Montagna P, et al. (2008) Comparative biomass structure and estimated carbon flow in food webs in the deep Gulf of Mexico. *Deep Sea Research Part II: Topical Studies in Oceanography* 55: 2699–2711.
- Pfannkuche O, Lochte K (2000) The biogeochemistry of the deep Arabian Sea: overview. *Deep Sea Research Part II: Topical Studies in Oceanography* 47: 2615–2628.
- Gage JD, Tyler PA (1991) Deep-sea biology: A natural history of organisms at the deep-sea floor. Cambridge: Cambridge University Press. 504 p.
- Heip CHR, Duineveld G, Flach E, Graf G, Helder W, et al. (2001) The role of the benthic biota in sedimentary metabolism and sediment-water exchange processes in the Goban Spur area (NE Atlantic). *Deep Sea Research Part II: Topical Studies in Oceanography* 48: 3223–3243.
- Thiel H (1975) The size structure of the deep-sea benthos. *Internationale Revue der gesamten Hydrobiologie und Hydrographie* 60: 575–606.
- Ho CK, Pennings SC, Carefoot TH (2010) Is Diet Quality an Overlooked Mechanism for Bergmann's Rule? *The American Naturalist* 175: 269–276.
- Sweetman A, Witte U (2008) Response of an abyssal macrofaunal community to a phytodetrital pulse. *Marine Ecology Progress Series* 355: 73–84.
- van Nugteren P, Herman P, Moodley L, Middelburg J, Vos M, et al. (2009) Spatial distribution of detrital resources determines the outcome of competition between bacteria and a facultative detritivorous worm. *Limnol Oceanogr* 54: 1413–1419.
- Witte U, Wenzhofer F, Sommer S, Boetius A, Heinz P, et al. (2003) In situ experimental evidence of the fate of a phytodetritus pulse at the abyssal sea floor. *Nature* 424: 763–766.
- Rowe GT, Deming JW. An Alternative View of the Role of Bacteria in the Cycling of Organic Matter in Deep-sea Sediments. *Marine Biology Research*: Submitted.
- Ingels J, Van den Driessche P, De Mesel I, Vanhove S, Moens T, et al. (2010) Preferred use of bacteria over phytoplankton by deep-sea nematodes in polar regions. *Mar Ecol Prog Ser* 406: 121–133.
- Clough L, Renaud P, Ambrose Jr. W (2005) Impacts of water depth, sediment pigment concentration, and benthic macrofaunal biomass on sediment oxygen demand in the western Arctic Ocean. *Canadian Journal of Fisheries and Aquatic Sciences* 62: 1756–1765.
- Rowe GT, Boland GS, Escobar Briones EG, Cruz-Kaegi ME, Newton A, et al. (1997) Sediment community biomass and respiration in the Northeast water polynya, Greenland: a numerical simulation of benthic lander and spade core data. *Journal of Marine Systems* 10: 497–515.
- Kirchman DL, Moran XAG, Ducklow H (2009) Microbial growth in the polar oceans - role of temperature and potential impact of climate change. *Nature Reviews Microbiology* 7: 451–459.
- Grebmeier JM, Barry JP (2007) Chapter 11 Benthic Processes in Polynyas. In: Smith WO, Barber DG, eds. *Elsevier Oceanography Series*: Elsevier. pp 363–390.
- Deming J, Baross J (1993) The early diagenesis of organic matter: bacterial activity. In: Engel MH, Macko SA, eds. *Organic geochemistry*. New York: Plenum Press. pp 119–144.
- Dixon JL, Turley CM (2000) The effect of water depth on bacterial numbers, thymidine incorporation rates and C:N ratios in northeast Atlantic surficial sediments. *Hydrobiologia* 440: 217–225.
- Deming JW, Carpenter SD (2008) Factors influencing benthic bacterial abundance, biomass, and activity on the northern continental margin and deep basin of the Gulf of Mexico. *Deep Sea Research Part II: Topical Studies in Oceanography* 55: 2597–2606.
- Lochte K (1992) Bacterial standing stock and consumption of organic carbon in the benthic boundary layer of the abyssal North Atlantic. In: Rowe GT, Pariente V, eds. *Deep-sea food chains and the global carbon cycle*. Dordrecht, The Netherlands: Kluwer. pp 1–10.

54. Deming JW, Colwell RR (1985) Observations of Barophilic Microbial Activity in Samples of Sediment and Intercepted Particulates from the Demerara Abyssal Plain. *Appl Environ Microbiol* 50: 1002–1006.
55. Patching JW, Eardly D (1997) Bacterial biomass and activity in the deep waters of the eastern Atlantic—evidence of a barophilic community. *Deep Sea Research Part I: Oceanographic Research Papers* 44: 1655–1670.
56. Lochte K, Turley CM (1988) Bacteria and cyanobacteria associated with phytodetritus in the deep sea. *Nature* 333: 67–69.
57. Deming J, Baross J (2000) Survival, dormancy, and nonculturable cells in extreme deep-sea environments. In: Colwell RR, Grimes DJ, eds. *Nonculturable Microorganisms in the Environments*. Washington, DC: American Society for Microbiology Press. pp 147–197.
58. Quéric N-V, Soltwedel T, Arntz WE (2004) Application of a rapid direct viable count method to deep-sea sediment bacteria. *Journal of Microbiological Methods* 57: 351–367.
59. Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, et al. (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* 454: 1084–1087.
60. Boetius A, Lochte K (1994) Regulation of microbial enzymatic degradation of organic matter in deep-sea sediments. *Marine Ecology-Progress Series* 104: 299–299.
61. Vetter Y, Deming J, Jumars P, Krieger-Brockett B (1998) A predictive model of bacterial foraging by means of freely released extracellular enzymes. *Microbial ecology* 36: 75–92.
62. Flach E, Muthumbi A, Heip C (2002) Meiofauna and macrofauna community structure in relation to sediment composition at the Iberian margin compared to the Goban Spur (NE Atlantic). *Progress In Oceanography* 52: 433–457.
63. Danovaro R, Croce ND, Eleftheriou A, Fabiano M, Papadopoulou N, et al. (1995) Meiofauna of the deep Eastern Mediterranean Sea: distribution and abundance in relation to bacterial biomass, organic matter composition and other environmental factors. *Progress In Oceanography* 36: 329–341.
64. Clough L, Ambrose W (1997) Infaunal density, biomass and bioturbation in the sediments of the Arctic Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* 44: 1683–1704.
65. Smith C (1992) Factors controlling bioturbation in deep-sea sediments and their relation to models of carbon diagenesis. In: Rowe GT, Pariente V, eds. *Deep-sea food chains and the global carbon cycle*. Dordrecht, The Netherlands: Kluwer. pp 375–393.
66. Rowe GT, Morse J, Nunnally C, Boland GS (2008) Sediment community oxygen consumption in the deep Gulf of Mexico. *Deep Sea Research Part II: Topical Studies in Oceanography* 55: 2686–2691.
67. Smith K, Hinga K (1983) Sediment community respiration in the deep sea. In: GTR, ed. *Deep Sea Biology*. New York: Wiley-Interscience. pp 331–370.
68. Levin L, Gage J, Martin C, Lamont P (2000) Macrobenthic community structure within and beneath the oxygen minimum zone, NW Arabian Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* 47: 189–226.
69. Quiroga E, Quiñones R, Palma M, Sellanes J, Gallardo VA, et al. (2005) Biomass size-spectra of macrobenthic communities in the oxygen minimum zone off Chile. *Estuarine, Coastal and Shelf Science* 62: 217–231.
70. Narayanaswamy BE, Bett BJ, Hughes DJ (2010) Deep-water macrofaunal diversity in the Faroe-Shetland region (NE Atlantic): a margin subject to an unusual thermal regime. *Marine Ecology* 31: 237–246.
71. Smith KL, Ruhl HA, Bett BJ, Billett DSM, Lampitt RS, et al. (2009) Climate, carbon cycling, and deep-ocean ecosystems. *Proceedings of the National Academy of Sciences* 106: 19211–19218.
72. Billett DSM, Bett BJ, Reid WDK, Boorman B, Priede IG (2010) Long-term change in the abyssal NE Atlantic: The 'Amperima Event' revisited. *Deep Sea Research Part II: Topical Studies in Oceanography* 57: 1406–1417.
73. Rowe GT, Staresinic N (1979) Sources of organic matter to the deep-sea benthos. *Ambio Special Report*. pp 19–23.
74. Smith CR, Maybaum HL, Baco AR, Pope RH, Carpenter SD, et al. (1998) Sediment community structure around a whale skeleton in the deep Northeast Pacific: Macrofaunal, microbial and bioturbation effects. *Deep Sea Research Part II: Topical Studies in Oceanography* 45: 335–364.
75. Brooks J, Kennicutt M (1987) Deep-sea hydrocarbon seep communities: evidence for energy and nutritional carbon sources. *Science* 238: 1138.
76. Van Dover C (2000) *The ecology of deep-sea hydrothermal vents*. Princeton NJ: Princeton University Press.
77. Rowe GT, Boland GS, Phoel WC, Anderson RF, Biscaye PE (1994) Deep-sea floor respiration as an indication of lateral input of biogenic detritus from continental margins. *Deep Sea Research Part II: Topical Studies in Oceanography* 41: 657–668.
78. Rowe GT, Polloni PT, Haedrich RL (1982) The deep-sea macrobenthos on the continental margin of the northwest Atlantic Ocean. *Deep Sea Research Part A Oceanographic Research Papers* 29: 257–278.
79. Danovaro R, Gambi C, Della Croce N (2002) Meiofauna hotspot in the Atacama Trench, eastern South Pacific Ocean. *Deep Sea Research Part I: Oceanographic Research Papers* 49: 843–857.
80. Boehlert G, Genin A (1987) A review of the effects of seamounts on biological processes. In: Keating BH, ed. *Seamounts, islands and atolls: American Geophysical Union*. pp 319–334.
81. Gooday A, Levin L, Linke P, Heeger T (1992) The role of benthic foraminifera in deep-sea food webs and carbon cycling. In: Rowe G, Pariente V, eds. *Deep-sea food chains and the global carbon cycle*. Dordrecht, The Netherlands: Kluwer. pp 63–91.
82. De Leo FC, Smith CR, Rowden AA, Bowden DA, Clark MR (2010) Submarine canyons: hotspots of benthic biomass and productivity in the deep sea. *Proceedings of the Royal Society B: Biological Sciences*. In press.
83. Soliman Y, Wicksten M (2007) *Ampelisca mississippiana*: a new species (Crustacea: Amphipoda: Gammaridea) from the Mississippi Canyon (Northern Gulf of Mexico). *Zootaxa* 1389: 45–54.
84. Smith CR, Berelson W, Demaster DJ, Dobbs FC, Hammond D, et al. (1997) Latitudinal variations in benthic processes in the abyssal equatorial Pacific: control by biogenic particle flux. *Deep Sea Research Part II: Topical Studies in Oceanography* 44: 2295–2317.
85. Wei C-L, Rowe G, Hubbard G, Scheltema A, Wilson G, et al. (2010) Bathymetric zonation of deep-sea macrofauna in relation to export of surface phytoplankton production. *Marine Ecology Progress Series* 399: 1–14.
86. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307.
87. Pace ML, Knauer GA, Karl DM, Martin JH (1987) Primary production, new production and vertical flux in the eastern Pacific Ocean. *Nature* 325: 803–804.
88. Johnson NA, Campbell JW, Moore TS, Rex MA, Etter RJ, et al. (2007) The relationship between the standing stock of deep-sea macrobenthos and surface production in the western North Atlantic. *Deep Sea Research Part I: Oceanographic Research Papers* 54: 1350–1360.
89. Rowe GT, Pariente V, eds (1992) *Deep-sea food chains and the global carbon cycle*. Dordrecht, The Netherlands: Kluwer. 412 p.
90. Pikitch EK, Santora C, Babcock EA, Bakun A, Bonfil R, et al. (2004) *ECOLOGICAL: Ecosystem-Based Fishery Management*. *Science* 305: 346–347.
91. IPCC 2007 *Climate Change, editor* (2007) *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* Cambridge, UK and New York, USA: Cambridge University Press.