# The Ocean Colour Climate Change Initiative: III. A round-robin comparison on in-water bio-optical algorithms

Robert J.W. Brewin[*,a,b], Shubha Sathyendranath[a,b], Dagmar Müller[c], Carsten Brockmann[d], Pierre-Yves Deschamps[e], Emmanuel Devred[f], Roland Doerffer[c], Norman Fomferra[d], Bryan Franz[g], Mike Grant[a], Steve Groom[a], Andrew Horseman[a], Chuanmin Hu[h], Hajo Krasemann[c], ZhongPing Lee[i], Stéphane Maritorena[j], Frédéric Mélin[k], Marco Peters[d], Trevor Platt[a], Peter Regner[l], Tim Smyth[a], Francois Steinmetz[e], John Swinton[m], Jeremy Werdell[g], George N. White III[n]

[a]*Plymouth Marine Laboratory (PML), Prospect Place, The Hoe, Plymouth PL1 3DH, UK*
[b]*National Centre for Earth Observation, PML, Plymouth PL1 3DH, UK*
[c]*Helmholtz-Zentrum Geesthacht, Max-Planck-Straße 1, 21502 Geesthacht, Germany*
[d]*Brockmann Consult, Max-Planck-Straße 2, D-21502 Geesthacht, Germany*
[e]*HYGEOS, 165 Avenue de Bretagne, 59000 Lille, France*
[f]*Université Laval, 2325, rue de l'Université, Québec G1V 0A6, Canada*
[g]*NASA Goddard Space Flight Center, Greenbelt, Maryland, USA*
[h]*College of Marine Science, University of South Florida, St. Petersburg, FL 33701, USA*
[i]*College of Science and Mathematics, University of Massachusetts Boston, Boston, MA 02125-3393, USA*
[j]*University of California Santa Barbara, Santa Barbara, CA 93106, USA*
[k]*European Commission - Joint Research Centre, Institute for Environment and Sustainability, Ispra, 21027, Italy*
[l]*European Space Agency, ESRIN, Via Galileo Galilei, Casella Postale 64 00044 Frascati, Italy*
[m]*Telespazio VEGA UK Ltd, 350 Capability Green, Luton, Bedfordshire, LU1 3LU, UK*
[n]*Ocean Science Division, Bedford Institute of Oceanography, Box 1006, Dartmouth, Nova Scotia, B2Y 4A2, Canada*

---

## Abstract

Satellite-derived remote-sensing reflectance ($R_{rs}$) just above the sea surface can be used for mapping biogeochemically relevant variables, such as the chlorophyll concentration and the Inherent Optical Properties (IOPs) of the water, at global scales for use in climate-change studies. Prior to generating such prod-

ucts, suitable algorithms have to be selected that are appropriate for the purpose. Algorithm selection needs to account for both qualitative and quantitative requirements. In this paper, we develop an objective methodology designed to rank the quantitative performance of a suite of bio-optical models. The objective classification is applied using the NASA bio-Optical Marine Algorithm Data set (NOMAD). Using *in situ* $R_{rs}$ as input to the models, the performance of eleven semi-analytical models, as well as five empirical chlorophyll algorithms and an empirical diffuse attenuation coefficient algorithm, are ranked for spectrally-resolved IOPs, chlorophyll concentration and the diffuse attenuation coefficient at 489 nm. The sensitivity of the objective classification and the uncertainty in the ranking is tested using a Monte-Carlo approach (bootstrapping). Results indicate that the performance of the semi-analytical models varies depending on the product and wavelength of interest. For chlorophyll retrieval, empirical algorithms perform better than semi-analytical models, in general. The performance of these empirical models reflect either their immunity to scale errors or instrument noise in $R_{rs}$ data, or simply that data used for model parameterisation were not independent of NOMAD. Nonetheless, uncertainty in the classification suggest the performance of some semi-analytical algorithms at retrieving chlorophyll were comparable with the empirical algorithms. For phytoplankton absorption at 443 nm, some semi-analytical models also performed with similar accuracy to an empirical model. We discuss the potential biases, limitations and uncertainty in the approach, as well as additional qualitative considerations for algorithm selection for climate change studies. Our classification has the

potential to be routinely implemented, such that the performance of emerging algorithms can be compared with existing algorithms as they become available. In the long-term, such an approach will further aid algorithm development for ocean-colour studies.

*Key words:* Phytoplankton, Ocean colour, Inherent Optical Properties, Remote sensing, chlorophyll-a

## 1. Introduction

Visible radiance received by satellite ocean-colour sensors over oceanic regions is essentially influenced by two components: the atmosphere and the ocean. Typically, the atmospheric component constitutes more than 80% of the signal received by the sensor, and it needs to be removed to isolate the signal from the ocean. The ocean-colour signal may then be used to quantify optically-significant water-constituents such as Coloured Dissolved Organic Matter (CDOM) and the abundance of particulate matter, inclusive of phytoplankton, indexed through their chlorophyll pigment concentration, and non-phytoplanktonic material (e.g. detrital and inorganic matter).

Phytoplankton are a key component of the Earth System and are recognised as an Essential Climate Variable in the Implementation Plan of the Global Climate Observing System (GCOS, 2011). Phytoplankton absorb light energy that is either dissipated as heat, directly influencing the physical properties of the

*Corresponding author. Plymouth Marine Laboratory (PML), Prospect Place, The Hoe, Plymouth PL1 3DH, UK

*Email address:* robr@pml.ac.uk (Robert J.W. Brewin)

oceans, or used for photosynthesis (primary production), by which light is converted into chemical energy and carbon converted from inorganic to organic form. It is estimated that phytoplankton fix approximately 50 gigatons of carbon per year, equivalent to net terrestrial primary production. Phytoplankton, together with physical processes, regulate the $CO_2$ concentration of the surface ocean and the rate of $CO_2$ exchanges between the atmosphere and ocean. They are at the base of the food web, providing sustenance for all pelagic marine life, and contribute to the biogeochemical cycling of a variety of climatically-relevant elements, such as silica, nitrate and phosphate, in addition to carbon. Monitoring the variability in phytoplankton distribution is vital to understanding how the ocean ecosystem is likely to respond to future changes in climate.

The concentration of CDOM, its photodegradation status and the concentration of detrital matter present in the water have a significant effect on phytoplankton photosynthesis, through their absorption of light at blue wavelengths of the visible spectrum, which corresponds to the main phytoplankton absorption peak. CDOM can also affect the transport and bioavailability of trace metals (Santschi et al., 1997; Guo et al., 2001), with possible implications for biological activity, and plays an important role in photochemistry and photobiology, with implications for ocean-climate connections (Nelson and Siegel, 2013). The presence of highly-scattering non-phytoplanktonic particulate material (e.g. detrital and inorganic matter) alters the spectral quality of the underwater light field and thus influences phytoplankton photosynthesis. The concentration of particulate material in the water is also important in coastal regions and has implications for

coastal protection, shipping and recreational activities. These are some of the reasons why the systematic monitoring of ocean colour is considered a requirement for climate research by GCOS (GCOS, 2011) and why it is a component of the Climate Change Initiative (CCI) of the European Space Agency (ESA).

The CCI programme was launched to realise the full potential of long-term, global, Earth Observation archives that ESA as well as its member states have established over the past 30-years, and to contribute to the Essential Climate Variable databases required by United Nations Framework Convention on Climate Change (UNFCCC). The Ocean Colour CCI (OC-CCI) project is one of 14 ESA funded CCI projects. The aims of OC-CCI are to create a long-term, consistent, error-characterised time series of ocean-colour products, for use in climate change studies. A key component of the programme is the selection of suitable algorithms that meet user requirements and project aims. The selection of algorithms for the OC-CCI project can be partitioned into two parts: (i) selection of algorithms that correct for atmospheric affects; and (ii) algorithms that convert the retrieved ocean-colour signal into biogeochemically relevant variables, hereafter referred to as atmospheric-correction and in-water algorithms respectively. This paper focuses on the development of an objective methodology designed to aid the selection of appropriate in-water algorithms for climate studies. For information regarding the selection of atmospheric-correction algorithms the reader is referred to Müller et al. (Submitted) in this issue.

Since the establishment of ocean-colour remote sensing from space, with the launch of the Coastal Zone Color Scanner (CZCS) of NASA on board the

Nimbus-7 satellite in 1978, blue-to-green ratios of water-reflectance have been used in empirical relationships to derive the total concentration of chlorophyll-a ($C$), an ubiquitous pigment present in phytoplankton. With the launch of the Sea-viewing Wide Field-of-view Sensor (SeaWiFS), the NASA successor to CZCS, NASA organised the SeaWiFS Bio-optical Algorithm Mini-workshop (SeaBAM; O'Reilly et al., 1998), designed to identify chlorophyll algorithms suitable for operational use for processing SeaWiFS data. A database was developed with simultaneous measurements of *in situ* chlorophyll and *in situ* measurements of remote-sensing reflectance just above the surface ($R_{rs}(\lambda)$). Based on the results from the workshop, an empirical blue-green band ratio algorithm, labelled the Ocean-Chlorophyll-2 (OC2) algorithm, was chosen as the operational algorithm for SeaWiFS. This was later updated to the Ocean-Chlorophyll-4 (OC4) algorithm (O'Reilly et al., 2000).

In Case-1 waters (Morel and Prieur, 1977) typically encountered in the open ocean, where variations in ocean-colour are driven primarily by the abundance of phytoplankton, with a co-varying influence from particulate matter and CDOM, empirical blue-green band-ratio algorithms were generally found to perform with reasonable accuracy. However, in more optically-complex waters (Case-2 waters according to Morel and Prieur, 1977), often encountered in coastal regions, where the concentrations of particulate matter and CDOM do not covary in a predictable manner with the abundance of phytoplankton, empirical blue-green band-ratio algorithms can give spurious results (e.g. Lavender et al., 2004).

Theoretical approaches have demonstrated that $R_{rs}(\lambda)$ is related to the In-

herent Optical Properties (IOPs) of seawater, the absorption and backscattering coefficients. The absorption coefficient can in turn be partitioned into the contributions from water itself, and the type and abundance of material present in the water, including phytoplankton, detrital matter and CDOM. The backscattering coefficient can be partitioned into contributions from pure seawater and particulate matter suspended in the water (which includes phytoplankton). IOPs can be used to infer biogeochemical processes and to estimate the concentrations of various optically-significant water constituents, such as chlorophyll. Theoretical approaches that derive IOPs from $R_{rs}(\lambda)$ may improve performance of algorithms in more optically-complex waters (see IOCCG, 2000), and a variety of semi-analytical approaches have been developed in this direction (see IOCCG, 2006).

Recently, NASA organised an international IOP algorithm workshop (Werdell, 2009), designed to provide data sets (Werdell and Bailey, 2005) and processing framework in an international forum within which a new generation of global IOP products can be developed and evaluated. The workshop aimed to: define the state of the art with regard to the application of semi-analytical models to satellite radiometry; identify similarities and differences between approaches; identify strategies to provide uncertainties in IOPs; and achieve community consensus toward the generation of global IOP products (Werdell, 2009). An output of the workshop was the development of a Generalised Inherent Optical Property model (GIOP), a test platform for algorithm development that offers freedom to specify various optimisation approaches and parameterisations (Franz

and Werdell, 2010; Werdell et al., 2013).

In contrast to the aims of the NASA GIOP workshop, but making use of progress made as a result of the workshop, and building on the report of the IOCCG working group on the topic (IOCCG, 2006), this paper aims to establish an objective methodology for algorithm selection for climate-change studies, and then to use the method to compare and rank a variety of algorithms. Both qualitative and quantitative considerations are examined. Qualitative considerations relate to the suitability of the algorithms for use in climate change studies and the quantitative considerations relate to algorithm performance. Qualitative algorithm considerations include the ability of the algorithm to:

- Create a long-term, consistent, error-characterised time series of ocean-colour products for use in climate-change studies;
- Generate products that best suit the requirements of the user community;
- Facilitate seamless merging of Case-1 (open-ocean) and Case-2 (coastal optically-complex) waters;
- Quantify a variety of properties of the marine ecosystem that are relevant to climate studies and accessible from satellite ocean-colour data and;
- Be robust against potential modifications in the marine ecosystem in a changing climate.

Ideally, the most suitable algorithm would meet all these requirements and compare well in statistical tests of performance. Using a suite of statistical tests,

and an *in situ* database of chlorophyll ($C$), the diffuse attenuation coefficient at 489 nm ($K_d(489)$), IOPs and $R_{rs}(\lambda)$, we evaluate the quantitative performance of a number of empirical and semi-analytical in-water bio-optical models. The limitations of the approach are discussed and additional challenges regarding the selection of in-water algorithms for climate studies are highlighted.

## 2. Data

To test in-water bio-opticals models, we made use of the publicly-available NASA bio-Optical Marine Algorithm Data set (NOMAD, Werdell and Bailey, 2005). NOMAD Version 2.0 ALPHA was compiled on 18 July 2008 by the NASA Ocean Biology Processing Group and source data is available online (http://seabass.gsfc.nasa.gov/seabasscgi/nomad.cgi), as is documentation related to IOPs (Werdell, 2005). The NOMAD database provides global *in situ* measurements of above-water spectral water-leaving radiance ($L_w(\lambda)$) and spectral surface irradiance ($E_s(\lambda)$), from which remote-sensing reflectance can be computed ($R_{rs}(\lambda) = L_w(\lambda)/E_s(\lambda)$), and coincident measurements of water constituents such as the chlorophyll-a concentration, IOPs and $K_d(489)$ (diffuse attenuation coefficient at 489 nm). The solar sun-zenith angle ($\theta$) was computed for each data point using information on time and location. Table 1 denotes the variables used in the comparison.

The OC-CCI project currently focuses on the use of three ocean-colour satellite platforms: the Medium Resolution Imaging Spectrometer (MERIS) of ESA; the Moderate Resolution Imaging Spectro-radiometer (MODIS) of NASA; and the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) of NASA, to create a

time-series of satellite data. Therefore, to be representative of the majority of wavelengths in all three satellite sensors, a common band set of 411, 443, 489, 510, 555, and 665 nm was chosen to maximise the amount of validation data points in NOMAD. Though there are some mis-matches (MERIS native 560 > 555 nm; MODIS native 547 < 555 nm and 531 nm excluded; and SeaWiFS 670 > 665 nm), this compromise was adopted to maximise the number of samples. The common band set used included six bands compatible with MERIS and SeaWiFS and five bands compatible with MODIS. Co-located *in situ* measurements of $R_{rs}(\lambda)$ were used as input to the models, as opposed to satellite-derived $R_{rs}(\lambda)$, to minimise mis-matches in spatial scales between input and output variables.

To maximise the number of $b_b(\lambda)$ samples, 670 nm was used where reflectance data at 665 nm were unavailable. Note that $b_b(\lambda)$, and the slope of $b_b(\lambda)$, denoted as $\gamma$ (Table 1), were used in this comparison as opposed to partitioning $b_b(\lambda)$ into the contribution from pure water ($b_{bw}$) and particles ($b_{bp}$), to avoid issues caused by different $b_{bw}$ spectra in different semi-analytical models. Remote sensing reflectance data, at various wavelengths, and solar-zenith angles were used as input to in-water algorithms to estimate IOPs, $C$ and $K_d(489)$ (Table 1, 2 and 3). Estimated variables using the models were then compared with *in situ* values in NOMAD, to determine the performance of the algorithms. Figure 1 shows the spatial coverage and number of samples for each variable used in the *in situ* database and the NOMAD record identifier for each measurement used in the comparison is provided as Supplementary Data.

## 3. Models

The following sections describe the semi-analytical models, designed to retrieve IOPs, and the chlorophyll models and the diffuse attenuation coefficient ($K_d$) models incorporated into the comparison. Tables 2 and 3 also provide a description of the output variables of each model and a summary listing key attributes of the various algorithms.

### *3.1. Semi-analytical models*

Semi-analytical models used in the comparison are described in this section. The term 'semi-analytical models' will be conventionally employed hereafter to describe Models A-K for the sake of brevity. However, we acknowledge that some of the models vary in their use of analytical and empirical solutions to solve for the IOPs. These semi-analytical models (A-K) are used to compute the total absorption coefficient ($a$), combined absorption by detritus and coloured dissolved organic matter or gelbsfoff ($a_{dg}$), absorption by phytoplankton ($a_{ph}$), total back-scattering coefficient ($b_b$), the spectral slope of the total backscattering coefficient ($\gamma$), the spectral slope of $a_{dg}$, denoted $S_{dg}$, and the ratio of phytoplankton absorption at 555 nm to that at 443 nm ($a_{ph}(555)/a_{ph}(443)$) (see Table 1 for all notations used). The ratio $a_{ph}(555)/a_{ph}(443)$ was used in this comparison as an index of the spectral shape of the phytoplankton absorption coefficient, an index of the community structure of the phytoplankton (Sathyendranath et al., 2001, 2004; Ciotti et al., 2002). The ratio of 555 nm to 443 nm was chosen as these wavelengths typically represent the minimum and maximum of the phytoplankton absorption spectra. However, we acknowledge that ratios of other

wavelengths could have also been used.

### *3.1.1. Model A*

Model A refers to the model of Smyth et al. (2006). It uses an algebraic approach for determining IOPs. The model uses spectral slopes for $a - a_w$ (where subscript $w$ stands for water) and $b_{bp}$ (total particulate backscattering) derived from field measurements, at the central wavelengths of 490 and 510 nm (or 531 for MODIS). Once the absorption and backscattering coefficients are known at these wavelengths, based on Morel (1980), and assuming a fixed spectral slope for $b_{bp}$, the absorption and backscattering coefficients across the spectrum can be determined. Once absorption and backscattering are determined spectrally, $a_{dg}$ and $a_{ph}$ can be determined using standard relationships and slopes between the wavelengths of 412 and 443 nm.

### *3.1.2. Model B*

Model B refers to the model of Smyth et al. (2006), as in Model A, but applying a new optical water classification, whereby the model parameters (spectral slopes in $a - a_w$ and $b_{bp}$) were computed for eight optical classes (see Moore et al., 2009). Based on the fuzzy-class-membership for each sample, determined from $R_{rs}$, the spectral slopes in $a - a_w$ and $b_{bp}$ are re-computed and implemented in the model of Smyth et al. (2006): $a_{dg}$ and $a_{ph}$ are then determined as in Model A.

### 3.1.3. Model C

Model C refers to the ocean-colour model of Devred et al. (2011) with some simplifications. This model is designed to derive in-water optical properties and water constituents from spectral water-leaving radiances, using non-linear optimisation procedures. The method makes use of a three-component model of phytoplankton absorption coupled to the reflectance model of Sathyendranath and Platt (1997). The model retrieves $b_{bp}(555)$ (assuming the slope of $b_{bp} = 1.03$ following Maritorena et al. (2002)), $a_{dg}(443)$ and $S_{dg}$ from $R_{rs}$, initially assuming that $a_{ph}$ can be expressed as the sum of the absorption coefficient of three phytoplankton size classes (pico-, nano- and micro-phytoplankton), each with its particular specific absorption spectrum ($a^*_{ph}$, phytoplankton absorption normalised by chlorophyll concentration) derived from the NOMAD dataset. Wavelengths from 443 to 555 nm were used in the inversion of Model C. Output variables were constrained to lie within the following range: $0.0 < a_{ph} < 100\,\mathrm{m}^{-1}$; $0.0 < a_{dg} < 100\,\mathrm{m}^{-1}$; and $0.0 < b_{bp} < 5.0\,\mathrm{m}^{-1}$.

### 3.1.4. Model D

Model D refers to the algebraic Quasi-Analytical Algorithm (QAA) of Lee et al. (2002). The model was designed to retrieve IOPs in optically-deep waters. The model inversion is based on two steps: the first involves partitioning water reflectance into $b_b$ and $a$ and the second decomposing $a$ into $a_{dg}$ and $a_{ph}$. The model is referred to as "Quasi-Analytical" as parts of the inversion are based on analytical, semi-analytical and empirical approximations. Model D uses the original parameterisation as described in Lee et al. (2002).

### 3.1.5. Model E

Model E refers to the Quasi-Analytical Algorithm (QAA) of Lee et al. (2002), as in Model D, but following an updated parameterisation (see Lee et al., 2009). This includes the use of measured $R_{rs}(670)$ in the calculation of $a(555)$, in contrast to Model D which instead uses $R_{rs}(640)$ in the calculation of $a(555)$, estimated empirically from other wavelengths when using data from SeaWiFS, MODIS, or MERIS.

### 3.1.6. Model F

Model F refers to the physics-based Hyperspectral Optimization Process Exemplar (HOPE) model of Lee et al. (1998, 1999). In this model, $R_{rs}$ is modelled as a function of IOPs, and when influencing the $R_{rs}$ signal, bottom depth and bottom albedo. Unknowns are derived from non-linear optimisation. The spectral shape of bottom albedo is pre-determined before the optimisation starts, with the choice of two shapes (one for sand, another for grass) automatically selected using the $R_{rs}$ spectrum. The phytoplankton absorption coefficients were constrained to lie within an upper and lower boundary (e.g. $0.002 < a_{ph}(443) < 1.0\,\mathrm{m}^{-1}$).

### 3.1.7. Model G

Model G refers to the semi-analytical Garver-Siegel-Maritorena (GSM) model, that was initially developed by Garver and Siegel (1997) and later updated by Maritorena et al. (2002). The GSM model retrieves simultaneous estimates of chlorophyll ($C$), $a_{dg}(443)$ and $b_{bp}(443)$ from $R_{rs}(\lambda)$, assuming an underlying bio-optical model and using non-linear optimisation. Global parameters

of the bio-optical model were initially assigned based on simulated annealing on a global quasi-real dataset, which are then used in the non-linear optimisation routine. These include a fixed chlorophyll-specific phytoplankton absorption coefficient ($a^*_{ph}$), $S_{dg}$ and the slope of $b_{bp}$. The chlorophyll ($C$), $a_{dg}(443)$ and $b_{bp}(443)$ are first retrieved by fitting the bio-optical model to the observed $R_{rs}(\lambda)$. IOPs at any wavelengths are then obtained using $C$, $a_{dg}(443)$ and $b_{bp}(443)$ and their specific shape function from the bio-optical model. For Model G, the output variables are constrained to lie within the range that was used to parameterise the model ($0.01 < C < 64\,\mathrm{mg\,m^{-3}}$; $0.0001 < a_{dg}(443) < 2.0\,\mathrm{m^{-1}}$; and $0.0001 < b_{bp}(443) < 0.1\,\mathrm{m^{-1}}$).

### *3.1.8. Model H*

Model H refers to the semi-analytical Garver-Siegel-Maritorena (GSM) model (Maritorena et al., 2002), as in Model G, but allowing the retrievals to have any value, thus removing the constraint imposed on Model G.

### *3.1.9. Model I*

Model I refers to a preliminary configuration of the Generalized Inherent Optical Property algorithm (GIOP; Franz and Werdell, 2010; Werdell et al., 2013). The GIOP model is designed as a test platform for algorithm development and was the result of a NASA IOP Algorithm Workshop (see Werdell, 2009; Werdell et al., 2013). Whereas the GIOP model offers the user freedom to specify different parameterisations and optimisation approaches, a preliminary configuration for GIOP is available which includes: an assigned $a^*_{ph}$ following Bricaud et al. (1995) but normalised by $0.055\,\mathrm{m^2\,(mg}C)^{-1}$; a spectral backscattering depen-

dency following the QAA; a fixed spectral slope for $a_{dg}(\lambda)$ of 0.018 $nm^{-1}$; Morel et al. (2002) $f/Q$ ratio for zero Sun angle and zero view angle, where $Q(\lambda)$ is the ratio of upwelling irradiance to upwelling radiance and $f(\lambda)$ captures the net effects of variation in sea state, illumination conditions, and water column content; and Levenberg-Marquardt optimisation. It is designed to retrieve spectral IOPs and chlorophyll, and it is worth noting that this preliminary configuration could be changed with time. All IOPs ($a_{dg}$, $a_{ph}$, $b_{bp}$, and $a_{dg} + a_{ph}$) were constrained to lie within -0.005 and 5 $m^{-1}$. Retrievals were excluded if the reconstructed $R_{rs}$ spectrum, between 411-555 nm, differed from the observed $R_{rs}$ spectrum by more than 33%.

### *3.1.10. Model J*

Model J refers to a Case-1 model, in which all IOPs are modelled as a function of the chlorophyll concentration ($C$) derived using the NASA OC4v6 empirical model (Model L). Once $C$ is estimated from $R_{rs}$, $C$ is used as input to estimate: $a_{ph}(\lambda)$ using a three-component model of phytoplankton absorption (Brewin et al., 2011); $a_g(\lambda)$ using a power-function of $C$ (Morel, 2009) with an exponential spectral slope ($S_g$) of 0.018 $nm^{-1}$; $a_d(\lambda)$ using a power-function of $C$ (Bricaud et al., 2010) with an exponential spectral slope ($S_d$) of 0.0094 $nm^{-1}$; $b_{bp}(\lambda)$ as a function of $C$ using the model of Huot et al. (2008); pure water absorption ($a_w$) according to Pope and Fry (1997); and pure-water backscattering ($b_{bw}$) according to Buiteveld et al. (1994). Components of absorption and backscattering are added to obtain the totals $a$ and $b_b$ respectively, from which $R_{rs}$ is computed using the model of Gordon et al. (1988).

### 3.1.11. Model K

Model K refers to a preliminary configuration of an in-water artificial Neural-Network (NN) (e.g. Doerffer and Schiller, 2000, 2006; Doerffer et al., 2002) which is used as the forward model within an optimisation procedure (Levenberg-Marquardt). The model computes IOPs from water-leaving radiance for all available multi-spectral ocean colour sensors as well as *in situ* measurements. The method was optimised to invert water-leaving radiance directly into spectral IOPs, with chlorophyll ($C$) parameterised as a function of phytoplankton absorption and $K_d(489)$ as a function of scattering and total absorption.

## 3.2. Chlorophyll ($C$) models

Chlorophyll ($C$) algorithms incorporated into the comparison are described in the following section. For semi-analytical Models C, G, H, I, and K, chlorophyll is an output from the models. For semi-analytical Models A, B, D, E, and F, chlorophyll is not an output. For the purposes of the comparison, we estimated chlorophyll as a function of $a_{ph}(443)$ using a power-law relationship (e.g. Bricaud et al., 1995), such that

$$C = \left[\frac{a_{ph}(443)}{A}\right]^{\frac{1}{B}}, \tag{1}$$

where, $A$ and $B$ are positive empirical parameters. The empirical parameters $A$ and $B$ were computed using the *in situ* NOMAD database (1042 samples), and set to $A = 0.0497$ and $B = 0.7575$. For semi-analytical Models A, B, D, E, and F, $a_{ph}(443)$ was first computed, then chlorophyll was computed using Eq. (1).

It is worth noting that the empirical conversion from $a_{ph}(443)$ to chlorophyll is merely introduced to facilitate the comparison, it is not a feature of the original algorithms. Note that Model J is not incorporated in the chlorophyll comparison as this model uses chlorophyll estimated from an empirical model (Model L) as input to compute IOPs. In addition to the semi-analytical models (A-I and K), a variety of empirical chlorophyll algorithms were also incorporated into the comparison and are described below.

### *3.2.1. Model L*

Model L refers to the NASA OC4 chlorophyll algorithm (O'Reilly et al., 2000). This is a polynomial algorithm that relates the log-transformed ratio of remote-sensing reflectances ($X$) to the chlorophyll concentration ($C$). The OC4v6 uses a four-band blue-green reflectance ratio such that:

$$X = \log_{10}\{[R_{rs}(443) > R_{rs}(489) > R_{rs}(510)]/R_{rs}(555)\}. \quad (2)$$

Chlorophyll ($C$) is estimated according to:

$$C = 10^{(a_0+a_1X+a_2X^2+a_3X^3+a_4X^4)}, \quad (3)$$

where, $a_0 = 0.3272$, $a_1 = -2.9940$, $a_2 = 2.7218$, $a_3 = -1.2259$ and $a_4 = -0.5683$ (NASA, 2010).

### 3.2.2. Model M

Model M refers to the NASA OC3S chlorophyll algorithm (O'Reilly et al., 2000). Like the OC4, this is a polynomial algorithm that relates the log-transformed ratio of remote-sensing reflectances ($X$) to the chlorophyll concentration ($C$). The OC3S uses a three-band blue-green reflectance ratio where

$$X = \log_{10}\{[R_{rs}(443) > R_{rs}(489)]/R_{rs}(555)\}, \tag{4}$$

and chlorophyll ($C$) is estimated according to Eq. (3) where, $a_0 = 0.2515$, $a_1 = -2.3798$, $a_2 = 1.5823$, $a_3 = -0.6372$ and $a_4 = -0.5692$ (NASA, 2010).

### 3.2.3. Model N

Model N refers to the NASA OC2S chlorophyll algorithm (O'Reilly et al., 2000). Like the OC4 and OC3S, this is a polynomial algorithm that relates the log-transformed ratio of remote-sensing reflectances ($X$) to the chlorophyll concentration ($C$). The OC2S uses a two-band blue-green reflectance ratio where

$$X = \log_{10}[R_{rs}(489)/R_{rs}(555)], \tag{5}$$

and chlorophyll ($C$) is estimated according to Eq. (3) where, $a_0 = 0.2511$, $a_1 = -2.0853$, $a_2 = 1.5035$, $a_3 = -3.1747$ and $a_4 = 0.3383$ (NASA, 2010).

### 3.2.4. Model O

Model O refers to the MERIS chlorophyll band-ratio algorithm (Morel and Antoine, 2011). Like the OC4, it is a four-band polynomial algorithm that relates

the log-transformed ratio of remote-sensing reflectance ($X$) to the chlorophyll concentration ($C$). Considering that a common-band set was chosen, not inclusive of 560 nm, the algorithm was implemented following Morel et al. (2007), such that the wavelength of 560 nm was replaced by 555 nm, and $X$ can be estimated following Eq. (2) and (3), where $a_0 = 0.4461529$, $a_1 = -3.291807$, $a_2 = 3.777216$, $a_3 = -4.172339$ and $a_4 = 1.415588$ (see Table 2 OC4Me555 of Morel et al., 2007).

### *3.2.5. Model P*

Model P refers to the chlorophyll algorithm of Hu et al. (2012). This empirical algorithm was designed to improve the estimate of chlorophyll ($C$) in the global ocean at concentrations $\leq 0.25\,\mathrm{mg\,m^{-3}}$. For low chlorophyll concentrations ($\leq 0.25\,\mathrm{mg\,m^{-3}}$), the algorithm uses a colour index (CI), which is defined as the difference between $R_{rs}$ in the green region of the visible spectrum and a reference formed linearly between $R_{rs}$ in the blue and red region of the visible spectrum. For high chlorophyll concentrations ($>0.3\,\mathrm{mg\,m^{-3}}$), Model P conforms to the OC4 algorithm (Model L), and for concentrations between $>0.25$ and $\leq 0.3\,\mathrm{mg\,m^{-3}}$ a mixture of the colour index (CI) and the OC4 algorithm (Model L) is used, allowing a smooth transition from the CI to the OC4 with increasing chlorophyll.

## *3.3. Diffuse attenuation models ($K_d$)*

Algorithms for computing the diffuse attenuation coefficient at 489 nm ($K_d(489)$) are described in the following section. For semi-analytical Models A to J, $K_d(489)$ was computed following Lee et al. (2005), with $a(489)$ and $b_b(489)$

computed according to the particular model (A-J) and the solar sun-zenith angle ($\theta$) as input, such that:

$$K_d(489) = [1 + (0.005\theta)]a(489) + 4.18\{1 - 0.52\exp[-10.8a(489)]\}b_b(489). \quad (6)$$

For semi-analytical Model K, $K_d(489)$ is an output, tied to scattering and total absorption. In addition to $K_d(489)$ estimates from semi-analytical models, an empirical algorithm was also incorporated into the comparison (Model Q).

### *3.3.1. Model Q*

Model Q refers to the NASA empirical algorithm for deriving $K_d(489)$ from SeaWiFS (KD2S). This is a polynomial algorithm that relates the log-transformed ratio of remote-sensing reflectances ($X$) to $K_d(489)$. The algorithm uses a two-band blue-green reflectance ratio to compute $X$ (see Eq. 5), and $K_d(489)$ is computed following:

$$K_d(489) = 10^{(a_0+a_1X+a_2X^2+a_3X^3+a_4X^4)} + 0.0166, \quad (7)$$

where, $a_0 = -0.8515$, $a_1 = -1.8263$, $a_2 = 1.8714$, $a_3 = -2.4414$ and $a_4 = -1.0690$ (NASA, 2009).

## 4. Methods

### *4.1. Statistical Tests*

To test the performance of the in-water algorithms the following univariate statistical tests were adopted that are commonly used in comparisons between

modelled and *in situ* data (e.g. Doney et al., 2009; Friedrichs et al., 2009).

### *4.1.1. Pearson correlation coefficient (r)*

The correlation coefficient $r$ (also called Pearson's product moment correlation) is calculated according to

$$r = \frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{X_i^M - \left(\frac{1}{N}\sum_{j=1}^{N} X_j^M\right)}{\{\frac{1}{N-1}\sum_{k=1}^{N}\left[X_k^M - \left(\frac{1}{N}\sum_{l=1}^{N} X_l^M\right)\right]^2\}^{1/2}}\right]\left[\frac{X_i^E - \left(\frac{1}{N}\sum_{m=1}^{N} X_m^E\right)}{\{\frac{1}{N-1}\sum_{n=1}^{N}\left[X_n^E - \left(\frac{1}{N}\sum_{o=1}^{N} X_o^E\right)\right]^2\}^{1/2}}\right] \tag{8}$$

where, $X$ is the variable and $N$ is the number of samples. The superscript $E$ denotes the estimated variable (from the model) and the superscript $M$ denotes the measured variable (from NOMAD). Note that the Pearson correlation coefficient assumes a linear relationship between variables and normal distributions. The correlation coefficient may take any value between -1.0 and 1.0.

### *4.1.2. Root Mean Square Error (Ψ)*

The absolute Root Mean Square Error (Ψ) is calculated according to

$$\Psi = \left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i^E - X_i^M\right)^2\right]^{1/2}. \tag{9}$$

### *4.1.3. The bias (δ)*

The bias between model and measurement can be expressed according to

$$\delta = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^E - X_i^M\right). \tag{10}$$

#### 4.1.4. The centre-pattern Root Mean Square Error ($\Delta$)

The absolute centre-pattern (or unbiased) Root Mean Square Error ($\Delta$) is calculated according to

$$\Delta = \left( \frac{1}{N} \sum_{i=1}^{N} \{ \left[ X_i^E - \left( \frac{1}{N} \sum_{j=1}^{N} X_j^E \right) \right] - \left[ X_i^M - \left( \frac{1}{N} \sum_{k=1}^{N} X_k^M \right) \right] \}^2 \right)^{1/2} . \quad (11)$$

It describes the error of the estimated values with respect to the measured ones, regardless of the average bias between the two distributions. It is related to $\Psi$ and $\delta$ according to $\Delta^2 = \Psi^2 - \delta^2$.

#### 4.1.5. Slope ($S$) and Intercept ($I$) of a Type-2 regression

The performance of a model with respect to *in situ* data can be tested using linear regression between the estimated variable (from the model) and the measured variable (*in situ* data), such that

$$X^E = X^M S + I. \quad (12)$$

A slope ($S$) close to one and an intercept ($I$) close to zero is an indication that the model compares well with the *in situ* data. Type-1 regression typically assumes the dependent variable (*in situ* data) is known infinitely well, when in reality the *in situ* data are also affected by uncertainties (e.g. problems with *in situ* data sampling techniques) that are difficult to quantify. Therefore, we adopted Type-2 regression (Glover et al., 2011, MATLAB function lsqfitma.m), which instead of minimising the vertical distance between independent data and linear fit (as in

Type-1 regression), minimises the perpendicular distance between independent data and linear fit.

#### *4.1.6. Percentage of possible retrievals ($\eta$)*

Considering that algorithms chosen for climate studies should perform routinely, and globally, and should not be a source of more gaps in the data than would be the case if other algorithms were used, the percentage of possible retrievals ($\eta$) is an important criterion that should be considered in the comparison, calculated according to

$$\eta = \frac{N^E}{N^M} 100, \tag{13}$$

where $N^E$ represents the number of retrievals using the model and $N^M$ represents the number of *in situ* data points.

All statistical tests described above were performed in $\log_{10}$ space, considering the majority of variables are approximately log-normally distributed, with the exception of $S_{dg}$, $\gamma$ and $a_{ph}(555)/a_{ph}(443)$ for which the analysis was performed in linear space.

### *4.2. Quantitative statistical methodology*

As with the OC-CCI comparison of atmospheric correction algorithms (Müller et al., Submitted), a points scoring classification was used in the in-water comparison to rank objectively the performance of the algorithms. Each variable was tested independently in the points scoring classification. For each variable, $R_{rs}(\lambda)$ values in the database were used as input to the algorithm to estimate the

variable, the estimated variable was then compared with the corresponding *in situ* value using each statistical test and a score was assigned for each test ranging from zero to two. These tests are described in the following sections. If the algorithm was not capable of estimating the variable, it was given zero points for that test.

In addition, a chi-square test was also performed separately on a selection of the semi-analytical models. This information was used to evaluate the goodness of fit of the computed spectral $R_{rs}$ values compared with the observed values. The samples were only compared when the measured and estimated variables conformed to the following requirements, which represent extreme upper and lower boundaries fixed to avoid the influence of spurious results on the statistical tests (note that algorithms were penalised (Eq. 13) for a higher number of spurious results):

- $C > 0.001$ and $< 200$ mg m$^{-3}$;
- $K_d > a_w$ (Pope and Fry, 1997) and $< 10.0$ m$^{-1}$;
- $a > a_w$ (Pope and Fry, 1997) and $< 10.0$ m$^{-1}$;
- $a_{dg} > 0.0001$ and $< 10.0$ m$^{-1}$;
- $a_{ph} > 0.0001$ and $< 10.0$ m$^{-1}$;
- $b_b > b_{bw}$ (Zhang et al., 2009) and $< 10.0$ m$^{-1}$;
- $\gamma > 0$ and $< 4.32$ (slope of pure water from Morel, 1974);
- $S_{dg} > 0$ and $< 0.05$ nm$^{-1}$;

- $a_{ph}(555)/a_{ph}(443) > 0$ and $< 5.0$

The lower boundaries for $a_{dg}$ and $a_{ph}$ were chosen based on the raw uncertainty of a WET-Labs ac9 in waters with low attenuation (WET-Labs, 2012), and lower boundaries for $C$ were based on the absolute accuracy for HPLC detection if all protocols are strictly followed (Aiken et al., 2009). The exclusion of spurious results was conducted on a variable-by-variable basis. For instance, for a given $R_{rs}$ spectra, if a semi-analytical model has one variable (e.g. $a_{ph}(443)$) that falls outside selected boundaries but another (e.g. $a(443)$) that falls within selected boundaries, the former would be excluded and the latter included.

#### *4.2.1. Pearson correlation coefficient (r) test*

The $r$ test involved determining whether the $r$-value for each model was statistically higher or lower than the mean $r$-value for all models. This was determined using the $z_{score}$. The $z_{score}$ may be used to determine if two correlation coefficients are statistically different from one another (Cohen and Cohen, 1983). Knowing the $r$-value for two respective models (say $r_1$ and $r_2$, for model 1 and 2 respectively) and knowing the number of samples used to determine the $r$-values (say $n_1$ and $n_2$) one can determine the $z_{score}$ using the Fisher's $r$-to-$z$ transformation. Making use of the sample size employed to obtain each coefficient, these $z$-scores of each $r$-value ($z_1$ and $z_2$) can be used to compute the overall $z_{score}$ (Cohen and Cohen, 1983), such that:

$$z_1 = 0.51\log(\frac{1 + r_1}{1 - r_1}), \tag{14}$$

$$z_2 = 0.51\log(\frac{1+r_2}{1-r_2}), \tag{15}$$

$$z_{score} = \frac{z_1 - z_2}{\{[1/(n_1-3)] + [1/(n_2-3)]\}^{1/2}}. \tag{16}$$

Having determined the $z_{score}$, this can be converted into a $p$-value assuming normal distribution. For the in-water comparison, a two-tailed test was used and if the $p$-value was <0.05, the $r$-values were deemed to be statistically different.

The mean $r$-value for all models was first determined by averaging the $r$-value of all the models being tested. The mean number of samples used to compute the $r$-value, was also determined by averaging all models being tested. The $r$-value and number of samples of a particular model were then compared with the mean value for all models, so as to determine if the model's $r$-value was statistically lower, similar or higher than the average value for all models. The following points for each model were awarded accordingly:

- 0 points = $r$-value for the model tested was statistically lower than the mean $r$-value for all models.
- 1 point = $r$-value for the model tested was statistically similar to the mean $r$-value for all models.
- 2 points = $r$-value for the model tested was statistically higher than the mean $r$-value for all models.

539 *4.2.2. Root Mean Square Error ($\Psi$) and centre-pattern Root Mean Square Error*
540 *($\Delta$) tests*

541 In addition to computing $\Psi$ and $\Delta$ for each model, it is possible to determine
542 the 95% confidence levels in the $\Psi$ and $\Delta$, which provide an indication of how
543 confident one is in $\Psi$ and $\Delta$ estimates. The 95% confidence levels can be com-
544 puted from the standard error of the mean percentage and the *t*-distribution of the
545 sample size. Confidence levels provide a very powerful way of showing differ-
546 ences and similarities between models. If the 95% confidence intervals of two or
547 more models overlap, then it can be assumed that the models have a statistically
548 similar $\Psi$ or $\Delta$.

549 For each model, the $\Psi$ and $\Delta$ were computed in addition to their 95% con-
550 fidence intervals. Furthermore, the average $\Psi$ and $\Delta$ value for all models tested
551 and the average 95% confidence interval on these values were also calculated.
552 The following points for each model were awarded accordingly:

553 • 0 points = $\Psi$ or $\Delta$ for the model tested was statistically higher than the
554 mean $\Psi$ or $\Delta$ for all models (95% confidence levels did not overlap).

555 • 1 point = $\Psi$ or $\Delta$ for the model tested was statistically similar to the mean
556 $\Psi$ or $\Delta$ for all models (95% confidence levels overlap with mean values).

557 • 2 points = $\Psi$ or $\Delta$ for the model tested was statistically lower than the mean
558 $\Psi$ or $\Delta$ for all models (95% confidence levels did not overlap).

559 Figure 2 shows an example of the points classification for models in the chloro-
560 phyll (*C*) comparison using $\Psi$.

#### *4.2.3. Bias ($\delta$) test*

The closer the model bias ($\delta$) is to the reference value of zero implies that the model corresponds well with the *in situ* data. However, a model could have a $\delta$ close to the reference value of zero, when compared with another model, but have a much larger 95% confidence interval, implying lower confidence in the retrieved $\delta$. Therefore, the following points classification was introduced for the bias:

- 0 points = the 95% confidence interval of $\delta$ for a particular model is higher than the mean 95% confidence interval for all models. In addition to this, the bias ± its 95% confidence interval did not overlap with zero ± the mean 95% confidence interval for all models.

- 1 point = either, the 95% confidence interval of $\delta$ for a particular model is lower than the mean 95% confidence interval for all models, or, the bias ± its 95% confidence interval overlaps with zero ± the mean 95% confidence interval, but not both cases.

- 2 points = the 95% confidence interval of $\delta$ for a particular model is lower than the mean 95% confidence interval for all models, and, the bias ± its 95% confidence interval overlaps with zero ± the mean 95% confidence interval.

#### *4.2.4. Slope ($S$) and Intercept ($I$) test*

In addition to computing the intercept ($I$) and the slope ($S$) from Type-2 regression, it is possible to compute the standard deviation on $I$ and $S$ (Glover

et al., 2011, MATLAB function lsqfitma.m). The closer the intercept ($I$) is to the reference value of zero and the closer the slope ($S$) is to the reference value of one, the better the fit between variables. However, a model could have an intercept closer to the reference value of zero and a slope closer to the reference value of one, when compared with another model, but have a much larger standard deviation on its retrieved parameters, implying lower confidence in the fit. Therefore, to account for both these possibilities the following points classification was introduced for the slope ($S$) parameter:

- 0 points = the standard deviation of the $S$ parameter for a particular model is higher than the mean standard deviation for all models. In addition to this, the $S$ parameter ± its standard deviation does not overlap with one ± twice the mean standard deviation for all models.

- 1 point = either, the standard deviation of the $S$ parameter for a particular model is lower than the mean standard deviation for all models, or, the $S$ parameter ± its standard deviation overlaps with one ± twice the mean standard deviation for all models, but not both cases.

- 2 points = the standard deviation of the $S$ parameter for a particular model is lower than the mean standard deviation for all models, and, the $S$ parameter ± its standard deviation overlaps with one ± twice the mean standard deviation for all models.

The following points classification was introduced for intercept ($I$) parameter:

- 0 points = the standard deviation of the $I$ parameter for a particular model is higher than the mean standard deviation for all models. In addition to this, the $I$ parameter ± its standard deviation does not overlap with zero ± twice the mean standard deviation for all models.

- 1 point = either, the standard deviation of the $I$ parameter for a particular model is lower than the mean standard deviation for all models, or, the $I$ parameter ± its standard deviation overlaps with zero ± twice the mean standard deviation for all models, but not both cases.

- 2 points = the standard deviation of the $I$ parameter for a particular model is lower than the mean standard deviation for all models, and, the $I$ parameter ± its standard deviation overlaps with zero ± twice the mean standard deviation for all models.

#### *4.2.5. Percentage of possible retrievals ($\eta$) test*

To compare the percentage of possible retrievals ($\eta$) between models, the average percentage of retrievals for all models was computed in addition to its standard deviation. The following points criteria were set-up:

- 0 points = $\eta$ of a model is less than the mean $\eta$ of all models − its standard deviation.

- 1 point = $\eta$ of a model overlaps with the mean $\eta$ for all models ± its standard deviation.

- 2 points = $\eta$ of a model is greater than the mean $\eta$ of all models + its standard deviation.

627 *4.2.6. Total points*

628 To rank the performance of each model with reference to a particular variable,
629 all points were summed over each statistical test. The total score for each model
630 was then normalised by the average score of all models being tested. A score of
631 one indicates the performance of a model is average with respect to all models,
632 a score greater than one indicates a model is performing better than the average
633 and a score less than one indicates the model is performing worse than average.
634 Figure 3 shows a flow-chart illustrating the methodology of the scoring system
635 used to intercompare models. Note that a doubling of points (say from 1 to 2)
636 does not imply an algorithm is twice as good; instead it implies that the difference
637 between the two models is statistically significant.

638 The stability of the scoring system, and the sensitivity of the scores, was
639 tested using the method of bootstrapping (Efron, 1979; Efron and Tibshirani,
640 1993). This involved using sampling with replacement to randomly re-sample
641 the *in situ* data (1000 times) creating 1000 new datasets the same size as the
642 original dataset but not identical. The quantitative statistical methodology was
643 then re-run for each new dataset (Monte-Carlo approach) and from the resulting
644 distribution of scores, a mean score for each model was computed. Additionally,
645 a 2.5% and a 97.5% interval on the bootstrap distribution was taken and assumed
646 to be the error-bars or confidence limits on the mean score for each model, rather
647 than standard deviations on the bootstrap distribution, to avoid misinterpretation
648 of results should the bootstrap distribution not follow a normal distribution or be
649 skewed, for instance from the presence of outliers in the data.

650 *4.2.7. Chi-square test*

651 In addition to the tests described above, a chi-square ($\chi^2$) test was also used to
652 compare performance of a selection of semi-analytical models. For each semi-
653 analytical model tested, a reconstructed reflectance spectrum was produced in
654 forward mode and compared with the *in situ* reflectance data. This was con-
655 ducted on 1713 samples ($K_d$(489) database) representative of a broad range of
656 oceanic environments inclusive of the major ocean basins (see Fig. 1). The test
657 is designed to examine how well each semi-analytical model performed at re-
658 producing the observations. The results from this test are not incorporated into
659 the points classification, as some semi-analytical models in the comparison are
660 algebraic (e.g. Models A, B, D and E) thus their $\chi^2$ values equal zero. However,
661 the information is useful to evaluate the performance of those semi-analytical
662 algorithms that are not algebraic (Models C, F, G, H, I, J and K). The chi-square
663 was computed for each of the 1713 spectra using the following formula:

$$\chi^2 = \sum_{i=1}^{N_\lambda} \left[ R_{rs}^M(i) - R_{rs}^E(i) \right]^2, \tag{17}$$

664 where, the super-script $M$ is the measured reflectance data, and the super-script
665 $E$ is the estimated reflectance data from the model. The lower the $\chi^2$ is, the better
666 the model reproduces the observed reflectance data.

667 ## 5. Results

668 ### *5.1. Chlorophyll comparison*

669 Figure 4 shows results of the quantitative comparison on chlorophyll concen-
670 tration. What is clear from the scatter plots in Fig. 4 is that all the algorithms
671 perform reasonably at estimating chlorophyll when compared with the *in situ*
672 data ($r$ >0.75). Secondly, a visual qualitative comparison of the scatter plots and
673 the results from the points classification score (bar chart in Fig. 4) reveals that
674 the objective points classification appears to be working consistently, such that
675 the models showing larger discrepancies between modelled and *in situ* data in
676 the scatter plots (e.g. Models C and K) have a low score, and models showing
677 a tighter relationship between modelled and *in situ* data in the scatter plots (e.g.
678 Models L to P) have a higher score.

679 Results from the classification in Fig. 4 (bar chart) highlight that the empiri-
680 cal chlorophyll models have the highest score (e.g. Model L, M, N and P). This is
681 not surprising considering that many of the *in situ* data used to parameterise these
682 empirical models are not independent of the *in situ* data used here to test these
683 models (see Table 3 and Section 6.1.1 for a discussion of this aspect). However, it
684 is worth noting that Model O, which is the same mathematical equation as Model
685 L, was parameterised using a theoretical model of ocean colour (Morel and Mar-
686 itorena, 2001) tuned using data gathered by the Laboratoire d'Océanographie de
687 Villefranche on $K_d$ and chlorophyll (see Morel and Antoine, 2011, for details),
688 data that are independent of the chlorophyll and $R_{rs}$ data used in this compari-
689 son. The high score by Model O support the results from Models L, M, and N, in

690 that the empirical (blue-green band-ratio) chlorophyll algorithms perform with
691 a high score in the quantitative comparison. The performance of the empirical
692 algorithms may reflect their immunity to scale errors in $R_{rs}$ data (e.g. band-ratio,
693 see Fig. 14) or errors induced by instrument noise (e.g. band-difference, see Hu
694 et al., 2012).

695 With regard to chlorophyll derived by the semi-analytical algorithms, Mod-
696 els A, G, H and I have a higher score when compared with Models B, D, E and
697 F. However, overlapping error bars from the bootstrap ensemble run, particularly
698 with regard to Model D and E, clearly indicate the difficulty in ranking the per-
699 formance of many of these semi-analytical models objectively. For Models A,
700 G, H and I, error bars from the bootstrap ensemble overlap with the empirical
701 models, suggesting that the performance of these semi-analytical algorithms are
702 comparable with the empirical algorithms in certain conditions. Models C and
703 K perform with low scores, indicating that these semi-analytical models perform
704 less accurately at deriving chlorophyll when compared with the other models in
705 the comparison (Fig. 4).

706 *5.2. $K_d$(489) comparison*

707 Figure 5 shows results of the quantitative comparison on $K_d$(489). All models
708 are seen to capture a high amount of the variability in the $K_d$(489) *in situ* data ($r$
709 >0.93). The bar chart indicates empirical Model Q performs with a high points
710 score in the $K_d$(489) comparison, followed by semi-analytical Models D and E.
711 Models F, I, J and K are shown to perform similarly (slightly above average with
712 scores >1), followed by Models G, H and C. Models A and B have low scores.

Model A shows a systematic overestimation in $K_d(489)$. Considering $a(489)$ and $b_b(489)$ are used as inputs to Eq. (7), this overestimation in $K_d(489)$ associated with Model A can be linked to an overestimation in $b_b(489)$ for this model (see Figure 10) as opposed to the influence of $a(489)$ (see Figure 7).

### *5.3. The total absorption coefficient ($a(\lambda)$) comparison*

Figures 6 and 7 show results of the quantitative intercomparison on $a(\lambda)$. Assessing the scatter plots (Fig. 7), all models capture a high amount of the variability in the *in situ* data at blue and green wavelengths (412-510 nm, $r$ >0.87); at longer wavelengths (e.g. 665 nm), Models A, B, D, and E (all algebraic approaches) have a low score in comparison with the other IOP models in the points classification (Fig. 6). When summing scores over all the wavelengths ($a(\lambda)$ Fig. 6), results from the points classification indicate that, with the exception of Model F which has the highest score, Models C through to K perform with similar scores, as indexed by overlapping error bars. Model A and B have a slightly lower score, which can be attributed to lower scores at longer wavelengths (e.g. Model A and B have a similar score to some models at shorter wavelengths (411, 443 and 489 nm, note the overlapping error bars), but lower scores at longer wavelengths (>510 nm) in Fig. 6). Models A, B, D and E retrieve $a(665)$ directly from $R_{rs}(665)$, consequently when $R_{rs}(665)$ is very low and has a high signal-to-noise ratio (common in oceanic waters), this will result in low quality $a(665)$. However, in such cases, semi-analytical optimisation models (e.g. Models C, F, G, H, I and K) have less dependence on the quality of $R_{rs}(665)$, as $a(665)$ is inferred using a bio-optical model that operates a minimi-

sation using wavelengths in blue, green and red regions of the spectrum, often with fixed spectral shapes for the IOPs.

### *5.4. The absorption coefficient of phytoplankton ($a_{ph}(\lambda)$) comparison*

Figures 6 and 8 show results of the quantitative intercomparison on $a_{ph}(\lambda)$. The results indicate a large range of variability between semi-analytical models. Models A, B, D, and E (algebraic approaches) perform reasonably well at shorter wavelengths (411-489nm), as indexed by a higher points score, but perform less accurately at longer wavelengths (555-665 nm), as indexed by a lower points score. Models C and F through to J alternatively have a higher points score at longer wavelengths (510-665 nm) and lower points score at shorter wavelengths, likely a result of the algebraic approaches performing less accurately at longer wavelengths (555-665 nm). When summing the points across all wavelengths ($a_{ph}(\lambda)$ Fig. 6), Models I and J have the highest scores followed by Models C, G, and H. Model J computes $a_{ph}(\lambda)$ assuming relationships between the chlorophyll concentration of three size-classes of phytoplankton (micro-, nano- and pico-phytoplankton), and their associated specific absorption coefficient ($a^*_{ph}$), as does Model C during a first iteration to compute $b_{bp}$ and $a_{dg}$. Models G and H estimate $a_{ph}(\lambda)$ as a linear function of chlorophyll and Model I relates changes in the spectral shape of $a^*_{ph}$ with changes in chlorophyll. Models A and F have an average score ($\sim$1), in comparison with the other models, with Model K having the lowest score when summing the points across all wavelengths.

Figures 6 and 11 show results of the quantitative intercomparison on $a_{ph}(555)/a_{ph}(443)$. Models A and B are seen to perform less accurately at es-

timating $a_{ph}(555)/a_{ph}(443)$, as indexed by a low points score. This can be attributed to the fact that $a_{ph}(555)$ is strongly overestimated by Models A and B despite performing well at retrieving $a_{ph}(443)$ (Fig. 8), causing an overestimation of $a_{ph}(555)/a_{ph}(443)$ (Fig. 11). Models C, F, I, and J have the highest scores for $a_{ph}(555)/a_{ph}(443)$, and it is worth noting that these models tie the spectral shape of $a_{ph}$ to either the chlorophyll concentration or $a_{ph}(443)$ (Model C only during a first iteration). Models D, E and K have intermediate scores, as do Models G and H which assume a fixed spectral shape for $a_{ph}$ (scores of ~1). Overlapping error bars indicate the scores of some of these models are statistically similar.

### *5.5. The absorption coefficient by detrital and dissolved matter ($a_{dg}(\lambda)$) comparison*

Figures 6 and 9 show results of the quantitative intercomparison on $a_{dg}(\lambda)$. In comparison with $a(\lambda)$ (Fig. 7), the majority of semi-analytical models are seen to capture a lower amount of the variability in *in situ* $a_{dg}(\lambda)$ ($r \leq 0.88$), indicating lower performance in retrieving $a_{dg}(\lambda)$ in comparison with $a(\lambda)$, at least for blue and green wavelengths. Slight variations in the performance of the algorithms for each wavelength are observed over the visible spectrum, which is likely caused by variations in $S_{dg}$ and the spectral shape of $a_{ph}$ between models. Despite these variations, the points score of all algorithms when summed across all wavelengths ($a_{dg}(\lambda)$ Fig. 6), is strikingly similar to the performance of the models at a single wavelength (e.g. $a_{dg}(443)$), highlighting the importance of correctly estimating the magnitude of $a_{dg}$ at a reference wavelength. However, it is worth noting that the NOMAD $a_d(\lambda)$ and $a_g(\lambda)$ multi-spectral data were de-

782 veloped by fitting an exponential slope to original data on a sample-by-sample
783 basis, to remove moderate noise often resulting from instrument artifacts or poor
784 sample baselines (Werdell, 2005). When summing scores across all wavelengths
785 ($a_{dg}(\lambda)$ Fig. 6), Models D and F have slightly higher scores, followed by Mod-
786 els H, G, E, B, J, A and I. However, with the exception of Models C and K,
787 which have consistently low scores, many models have overlapping error bars
788 indicating statistically similar results.

789 Figures 6 and 11 show results of the quantitative intercomparison on $S_{dg}$. To
790 compute $S_{dg}$ for each semi-analytical model and *in situ* sample, the spectral $a_{dg}$
791 results were fitted using an exponential equation between 411-665 nm. What is
792 clear from the scatter plots is that none of the models capture well the variability
793 in $S_{dg}$ ($r < 0.15$, Fig. 11). Models C to F and Model J have a slightly higher
794 score in the points classification when compared with Models A, B, G, H, I and
795 K. The higher points score for Models C to F and J are related to a lower $\Psi$, $\Delta$
796 and $\delta$ for these models (Fig. 11). It is worth noting that Models G, H, and I, have
797 higher $S_{dg}$ (0.018 to 0.0206) than the other models in the comparison.

798 *5.6. The total backscattering coefficient ($b_b(\lambda)$) comparison*

799 Figures 6 and 10 show results of the quantitative intercomparison on $b_b(\lambda)$.
800 Results indicate that it is difficult to separate the performance of the semi-
801 analytical models at determining $b_b(\lambda)$, as indexed by large error bars on the
802 mean score of the bootstrap distribution. These larger error bars are in part a con-
803 sequence of a lower number of *in situ* samples in the $b_b(\lambda)$ dataset, as compared
804 with the other IOPs. Models A and B display a positive bias (Fig. 10), indicating

an overestimation of $b_b(\lambda)$, and Model J appears to underestimate $b_b(\lambda)$ at larger values (Fig. 10). When summing scores across all wavelengths ($b_b(\lambda)$, Fig. 6), Models A, C and K have lower scores and Models D, G, H and J slightly higher scores, when compared with the majority of models.

Figures 6 and 11 show results of the quantitative intercomparison on $\gamma$. To compute $\gamma$ for each semi-analytical model, and for the *in situ* data, the spectral $b_b$ results were fitted using a power-law equation between 411-665 nm. As with the $b_b(\lambda)$ points classification, it is difficult to separate the performance of some of the algorithms (overlapping error bars). Models D and E have a higher points scores in the $\gamma$ test (note for these models the slope of $b_{bp}$ was parameterised using some of the data in NOMAD), followed by Models B, C and F through to J. Models D, E, F, I and J all vary the spectral dependency of particulate backscattering ($b_{bp}$) as a function of a blue-green ratio, Model J indirectly through chlorophyll which is first estimated using a blue-green ratio from Model L. Models G and H assume a constant spectral dependency of particulate backscattering ($b_{bp}$). Models A and K have a lower score when compared with the other semi-analytical models.

### *5.7. Chi-square tests*

Figure 12 shows the results from the chi-square ($\chi^2$) test for the non-algebraic semi-analytical models (Models C, F, G, H, I, J, and K). Results indicate that the models with the lowest chi-square are Models I and F, followed by Model K then Models G, H and C. Model J has a higher chi-square when compared to the other models, indicating the agreement between $R_{rs}$ *in situ* and model is lower for this

model. For the algorithms that use non-linear optimisation (Models C, F, G, H, I and K) the chi-square results are influenced by both the convergence criteria of the optimisation scheme and the degrees of freedom in the bio-optical model. A more stringent convergence criterion can result in a lower chi-square, but only to an extent that is constrained by the freedom of the model to reproduce observed $R_{rs}$. The chi-square is also dependent upon the optimisation scheme itself (e.g. Levenberg-Marquardt, Gradient descent, Nelder-Mead method, Quasi-Newton, Trust region), each of which has its advantages and disadvantages (see Mu et al., 2011), how each approach minimises the $\chi^2$ (minimising to the absolute values of $R_{rs}$, relative values, or even logarithmically transformed values), and the number of wavelengths used in the minimisation.

### *5.8. Overarching comparison of semi-analytical models*

Figure 13 shows results for the quantitative intercomparison when combining the points score for all variables for each semi-analytical model, then normalising with respect to the mean score. This was conducted in four ways: (i) all points for spectral IOPs ($a(\lambda)$, $a_{dg}(\lambda)$, $a_{ph}(\lambda)$, $b_b(\lambda)$, $\gamma$, $a_{ph}(555)/a_{ph}(443)$ and $S_{dg}$), chlorophyll ($C$) and $K_d(489)$; (ii) all points for all spectral IOPs and $K_d(489)$; (iii) all points for all spectral IOPs; (iv) and all points for IOPs from wavelengths 411-555 nm. The later was conducted as some algorithms perform poorly at retrieving some IOPs at 665 nm (e.g. Model A, B, D, and E) which could have repercussions on the points score for other models (see discussion on this aspect in Section 6.1.2).

When combining the scores of all these variables, regardless of approach (i-

iv above), it is evident that Models D to J have higher scores than Models A, B, C and K. It is important to note that despite this, Models A, B, C and K do, in some cases, have higher or comparable scores to Models D to J for particular variables (Fig. 6). Regarding Models D to J, it is very difficult to objectively rank their performance with respect to each other, considering overlapping error bars. Models H and J have a higher points score than Model E in all cases except when summing points for IOPs from wavelengths 411-555 nm. However, in all cases Model E has a statistically similar score to Models D, F, G and I, as indexed by overlapping error bars, and Models F and G have statistically similar scores to Models H and I. Models D to J all have statistically similar scores for IOPs from wavelengths 411-555 nm. Therefore, results from the objective classification indicate that Models D to J perform similarly, when the ensemble of variables are considered. However, as highlighted in Fig 6, the scores of these models vary depending on product and wavelength.

## 6. Discussion

### *6.1. Methodological Uncertainties*

#### *6.1.1. Data*

This paper focuses on the development of a methodology to classify and rank objectively the performance of a variety of in-water bio-optical algorithms. The classification has been applied to a selection of in-water algorithms and the NOMAD *in situ* dataset. We have used the NOMAD dataset as, to our knowledge, it is the most extensive globally-representative dataset of co-located measurements

of *in situ* $R_{rs}(\lambda)$ and in-water variables (IOPs, $C$ and $K_d(489)$). To implement the classification requires a large database. Ideally an inter-comparison of this nature should be performed using a database entirely independent of any data used to parameterise the models. In the intercomparison carried out here, it has been difficult to evaluate the impact of the NOMAD dataset on algorithm performance, because most algorithms are influenced to some degree by the dataset (see Table 3). The limited availability of *in situ* observations on $R_{rs}(\lambda)$ and in-water variables, coupled with the need for a large database to implement our objective classification has meant that some data used in the comparison are not independent of those used to parameterise many of the models. This was partly addressed using the bootstrap method which allowed for some investigation into the performance of the algorithms in the context of the range of variability in the dataset. However, the work highlights the need for an independent dataset to be developed and used to evaluate algorithms further, to ascertain the extent to which the results are influenced by this issue.

Whereas NOMAD is the most extensive global database of *in situ* $R_{rs}(\lambda)$ and in-water variables (IOPs, $C$ and $K_d(489)$), the distribution of measurements in NOMAD is not equivalent to the distribution in the global ocean. Eutrophic waters are over-represented in NOMAD and oligotrophic waters under-represented (Werdell and Bailey, 2005). Ideally, when comparing global bio-optical algorithms, a dataset should be used that corresponds approximately to the distribution of measurements in the global ocean, highlighting the need for continued on-going *in situ* campaigns that focus on the areas of the ocean that are under-

represented in *in situ* databases, such as the oligotrophic gyres.

In the objective classification, the *in situ* datum is essentially deemed to be the truth, whereas, in reality *in situ* data also have associated errors. Measurement outliers were minimised using robust quality control procedures adopted in NOMAD (Werdell and Bailey, 2005). However, quantifying these errors is a very difficult task and some variables have a higher level of uncertainty than others. For some of the statistical tests, the measurement errors were partly accounted for (e.g. Type-2 regression). Nonetheless, it is recommended that future efforts include uncertainty indices for *in situ* observations.

In this study, *in situ* observations of $R_{rs}$ were used as input to the models. It can be assumed that errors in the *in situ* $R_{rs}$ values are small in comparison to satellite-derived $R_{rs}$. The performance of the algorithms tested may differ when used with data containing higher levels of noise. The tolerance of the bio-optical models to errors in $R_{rs}$ will need to be evaluated further to reflect realistic satellite measurement conditions. This could be done using simulated datasets (e.g. Lee et al., 2010) or satellite and *in situ* match-ups (e.g. Mélin et al., 2005; Bailey and Werdell, 2006; Maritorena et al., 2010). A global database of satellite and *in situ* match-ups would also allow for a thorough investigation into the suitability of coupling different in-water bio-optical models with atmospheric correction models. For example, atmospheric-correction models that focus on estimating the spectral-shape of $R_{rs}$ accurately, with low bias, maybe better suited to band-ratio in-water models. Hu et al. (2012) found that band-difference chlorophyll algorithms are less sensitive than band-ratio algorithms to various errors induced

by instrument noise and imperfect atmospheric correction in low chlorophyll waters. It is recommended that future efforts investigate potential synergistic benefits of combining different in-water and atmospheric correction models.

### *6.1.2. Objective classification*

The objective classification developed here is a step toward a fully-automated tool for the comparison and development of emerging bio-optical algorithms. The strategy for algorithm selection has to be open to the possibility that better algorithms will emerge in the future, requiring periodic re-evaluations of algorithms, adoptions of new algorithms and re-processing of data archives, as and when necessary. The objective classification developed here can aid the quantitative comparison between emerging and existing algorithms. However, the classification itself may undergo refinement with use and with changing user requirements.

There are issues with using the average performance of all models as a baseline from which to compare algorithm performance. If some algorithms perform very poorly this can significantly influence the average performance of all models, to the extent that it becomes difficult to differentiate between the higher performing models. This happened for $a(665)$ and $a_{ph}(665)$ (see Fig. 6). Models A, B, D, and E performed poorly, with high $\Psi$, $\Delta$ and $\delta$ in comparison with the other models (Fig. 7 and 8) resulting in minimal points for Models A, B, D, and E and maximum points for all other algorithms. Supplementary Fig. S1 shows the $a(665)$ results with and without the inclusion of Models A, B, D and E. When these models are removed from the comparison, it becomes apparent that Model

942 G, H and J have a higher point score than Model C. This issue is to some extent
943 dependent on the number of algorithms being tested. For instance, if one algo-
944 rithm performs poorly it will have a larger effect on the mean of all models when
945 only a small number of algorithms are being compared.

946 It is also important to note that the objective classification was conducted on
947 a variable-by-variable basis. For example, there is no reason why the scores of
948 the individual absorptions ($a_{ph}$ and $a_{dg}$) should be related to total absorption ($a$).
949 In Fig 6, Model K has an average score for $a(443)$ but low score for $a_{ph}(443)$ and
950 $a_{dg}(443)$. The performance of Model K impacts the average performance of all
951 models, such that Models G and H have a higher score for $a_{ph}(443)$ and $a_{dg}(443)$
952 than they do for $a(443)$.

953 Another disadvantage of using the average performance of all models as a
954 baseline from which to compare algorithm performance, is that it gives an in-
955 dication only as to the relative performance of each model with respect to the
956 others, and not in absolute terms. For instance, it is clear from the scatter plots
957 (Fig. 5) that $K_d(489)$ is retrieved better by all models than $S_{dg}$ (Fig. 11), yet it is
958 not clear from the scores in the objective classification (Fig. 6). The univariate
959 statistical tests were chosen in the objective classification as they are commonly
960 used in comparisons between modelled and *in situ* data. However, varying the
961 number of statistical tests in the comparison is likely to influence results. Future
962 refinement of the classification may include incorporating additional statistics,
963 or refining the number of statistical tests used, or even weighing the score of the
964 statistics, should one statistic be deemed more important than others.

An additional uncertainty is the challenging issue of how to filter the influence of spurious inversion results. Here, we used extreme upper and lower boundaries for each variable to avoid the influence of spurious results on the statistical tests, filtering results if they fall outside the boundaries. For some optimisation models, inversion results are constrained by positive boundaries which differ among approaches and with those used here to filter results. When the boundaries are hit should we consider the results valid or invalid? One may argue that such results are not valid as they are likely to change if the boundaries assigned by the optimisation scheme change. Setting the boundaries to the same values for all optimisation models, consistent with those used to filter results from other models, could minimise some differences. However, these boundaries are often chosen according to range of data used for parameterisation, which vary among models. There appears to be some subjectivity in the selection of a suitable criterion for filtering spurious inversion results, yet the decision may have a large influence on the results of the classification. For future model comparisons, it is recommended that significant efforts be focused toward the development of an objective filter for spurious inversion results.

The models tested here differ implicitly in their treatment of uncertainties in the measured $R_{rs}$ values. Band-ratio algorithm assume negligible uncertainties in the blue to green ratios of $R_{rs}$. Optimisation methods, that neglect certain bands (e.g. Model C), are effectively assuming very large uncertainties in these neglected bands. These differences impose some unavoidable limits on the comparison. As progress is made in the quantification of uncertainty in $R_{rs}$ (e.g.

Moore et al., 2009) treatment of uncertainties in the various models should become less diverse.

To account for methodological uncertainties in the classification, bootstrapping was introduced. This Monte-Carlo approach not only provides a simple method to check the stability of the results, but also offers a straightforward way to derive confidence estimates on the resulting classification (Efron, 1979; Efron and Tibshirani, 1993), which is useful when comparing model performance. However, bootstrapping can be computationally expensive and cannot offer insight beyond the range of data to which it is applied.

### *6.2. Implications for algorithm performance and development*

What is clear from the results of the comparison is that the performance of each model varied depending on the product and wavelength being tested. Based on the results in Figures 4, 5, 6 and 12, Table 4 highlights the variables in which each semi-analytical model (A-L) performed well and less well in the classification. This information may be of use to algorithm developers and to users who are potentially interested in a specific property, as it highlights components in these models that may require improvement.

Aside from the individual performance of the models, there are variables for which all models perform reasonably well or less well at retrieval. From the scatter plots (Fig. 4 to 11) in general, it is apparent that most models perform well at retrieving $K_d$(489), $a$(411-555) and $a_{ph}$(443). Some algorithms also retrieve $b_b$ reasonably well. Decomposing $a$ into $a_{ph}$ and $a_{dg}$ is a problem with some models. An increase in performance of $a_{ph}$ often results in a reduction in performance of

$a_{dg}$ and vice-versa (e.g. see Fig. 6 Models A and B, and Models D and E). In general, all models struggle to retrieve $a_{dg}(\lambda)$, as seen in a higher dispersion in the $a_{dg}(\lambda)$ scatter plots (Fig. 9) compared with other variables, confirming other studies (e.g. Mélin et al., 2007). Many of the models also struggle at retrieving $a_{ph}(555)/a_{ph}(443)$ and $S_{dg}$, since they assume fixed values for these variables despite clear variability in the *in situ* data (Fig. 11). As previously highlighted, some of these *in situ* variables may have a higher level of measurement error than others, which is also dependent on the signal-to-noise ratio of the measurement at the wavelength of interest.

Algebraic approaches (Models A, B, D and E) struggle to retrieve reasonable results for $a$ and $a_{ph}$ at 665 nm. These algebraic approaches derive the absorption coefficients at a specific wavelength directly from measured $R_{rs}$ at that wavelength. Typically, for most Case-1 global waters $R_{rs}(665)$ approaches zero, due to the dominating effect of water absorption at this wavelength. Therefore, direct retrievals of $a_{ph}$ at 665 nm, when there is little $a_{ph}$ signal, are particularly challenging using these algebraic approaches. This is further complicated by additional inelastic processes (e.g. Raman scattering) that become increasingly important at longer wavelengths. Alternatively, many of the optimisation approaches operate a minimisation with respect to the to absolute magnitude of $R_{rs}$. For most Case-1 global waters, where $R_{rs}(665)$ approaches zero, $R_{rs}(665)$ has lower weight in the optimisation than $R_{rs}$ at shorter wavelengths, meaning that retrievals, such as $a_{ph}(665)$, are actually inferred primarily from $R_{rs}$ at shorter wavelengths. Under phytoplankton bloom conditions or turbid waters, where

there is a higher signal in $R_{rs}(665)$, it is a different story. Under such conditions, variables such as $a_{ph}(665)$ could be derived from the measured $R_{rs}(665)$ using the algebraic approaches (possibly by shifting the reference wavelength further into the red or near-infrared). It is also likely that optimisation approaches, that operate a minimisation with respect to the absolute magnitude of $R_{rs}$, will give more weight to $R_{rs}(665)$ when deriving $a_{ph}(665)$ in bloom conditions, despite not deriving $a_{ph}(665)$ directly from $R_{rs}(665)$.

In this comparison, models were tested against a suite of IOPs, $K_d(489)$ and chlorophyll. It is important to note that many of these models are not designed for retrieving all these variables. The algebraic QAA model is not intended to derive IOPs at wavelengths longer than the reference wavelength, and many of the optimisation algorithms are typically designed to retrieve IOPs at specific wavelengths assuming an underlying bio-optical model. The advantages and disadvantages of each approach are, to a certain degree, characteristic of model design, making built-in biases difficult to avoid in this comparison. Nonetheless, this comparison has demonstrated that all the algorithms compared have certain desirable features. Further algorithm improvements could be explored by combining the best features of various algorithms. The NASA GIOP framework is an ideal platform for such algorithm development, offering users freedom to specify and compare various optimisation approaches and parameterisations. Alternatively, algorithm improvements may also come from looking outside the current set of approaches (e.g. Morel and Gentili, 2009; Shanmugam, 2011).

When using semi-analytical approaches to estimate IOPs, it is generally as-

sumed that there is a good closure between the Apparent Optical Properties (AOPs) (or quasi-Inherent Optical Properties, such as $R_{rs}$) and the IOPs themselves. Figure 14 shows a comparison between measured $R_{rs}$ and modelled $R_{rs}$ for 87 samples in NOMAD with corresponding $R_{rs}$, $a$ and $b_b$ at wavelengths from 411-555 nm. Modelled $R_{rs}$ in Fig. 14 was reconstructed using *in situ* $a$ and $b_b$ and the approximation of Gordon et al. (1988). What is clear from Fig. 14, is that for the 87 samples used there is an imperfect closure between $R_{rs}$ and modelled $R_{rs}$ reconstructed from the *in situ* IOPs. Interestingly, there appears to be better closure when reconstructing the shape of $R_{rs}$ from the IOPs. The reasons for this lack of closure are likely related to (i) uncertainty or errors in the *in situ* measurements themselves (both IOP and $R_{rs}$) and (ii) errors in the model, both of which require further investigation and have implications for algorithm development.

### *6.3. Algorithm selection for climate studies*

Figure 13 indicates that when combining results from all variables, semi-analytical Models D through to J have higher scores than Models A, B, C and K. Depending on the combination of variables (Fig. 13), it is difficult to rank the performance of these algorithms, as many of the models have overlapping error bars. The selection of suitable algorithms for any project depends not only on the quantitative performance of these algorithms, but also their suitability for the applications envisaged and the user requirements.

Algorithm selection for climate-change studies should take into consideration also the development of future ocean-colour products. The detection of

phytoplankton functional types is an emerging area of research (Nair et al., 2008) particularly relevant in the context of a changing climate. The spectral shape of the phytoplankton absorption coefficient provides an indication of the community structure of phytoplankton (Sathyendranath et al., 2001, 2004; Ciotti et al., 2002). To estimate the particle size distribution from satellite data requires measurements of the spectral slope of particle backscattering (Loisel et al., 2006; Kostadinov et al., 2009). The exponential slope of the CDOM coefficient can potentially provide information on the proportions of humic and fulvic acids, the semi-labile and refractory fractions, photo-degradation status, and the relative contribution of $a_d$ to $a_{dg}$. Bio-optical algorithms that do not allow for variations in the spectral shape of these IOPs are unsuitable for development of such products (nor are they designed with such applications in mind). Accurate retrievals of the phytoplankton absorption coefficient at 670 nm have the potential to improve chlorophyll estimates, considering that absorption at this wavelength is less affected by absorption from accessory pigments, and allow for estimates of the average size of the phytoplankton (Roy et al., 2010). Algorithms that fail to detect $a_{ph}(670)$ will be unsuitable for such purposes. Furthermore, algorithms that infer $a_{ph}(670)$ from other wavelengths, or from chlorophyll, are not providing the independent information required for such purposes.

Algorithms for climate change studies need to be robust in a changing environment. For example, if the phytoplankton community structure changes, the alteration in community structure should not interfere with the performance of the algorithm at retrieving chlorophyll. Empirical relationships that tie one

property to the next need to be minimised in models, since correlations between elements of the ecosystem may not be stable in a changing climate. Empirical relationships are based on observations in the past, often pooling data from multiple years, which may not be a faithful guide to the future state of the ocean. If empirical relationships are unavoidable, on-going re-calibration is required to reduce ambiguity in interpretation of results. A theoretical underpinning of the empirical models should be established to ascertain sensitivity to possible climate-related scenarios. Algorithms should also be robust against potential modifications in relationships between optically-significant constituents, meaning that retrievals of the different contributors to ocean colour should ideally be independent of each another. This would also facilitate seamless merging of Case-1 and Case-2 algorithms, considering both water-types are vulnerable to climate-related change. The different ocean-colour products have to be consistent with each other, in the sense that they close the radiation budget with minimal error. For instance, the empirical nature of Model J was such that when combining the individual products the radiation budget was not closed with minimal error (Fig. 12).

## 7. Summary

An objective classification has been presented designed to rank the quantitative performance of a suite of bio-optical models based on a variety of univariate statistics. Eleven semi-analytical models, as well as five empirical chlorophyll algorithms and an empirical diffuse attenuation coefficient algorithm, were ranked for some 29 variables using the NASA NOMAD dataset. Uncertainty in the

ranking, and sensitivity of the objective classification to the test dataset, were addressed using a bootstrapping (Monte-Carlo) approach. Results from the classification suggest that algorithm performance varies depending on the product and wavelength of interest, and that empirical algorithms in general performed better in the classification than semi-analytical models at retrieving chlorophyll, either due to their immunity to scale errors or instrument noise in $R_{rs}$ data, or simply that data used for model parameterisation were not independent of NOMAD. However, uncertainty in the classification suggest some semi-analytical algorithms performed comparably to the empirical algorithms at retrieving chlorophyll. Methodological uncertainties in the approach were discussed, and indicate the need for an independent *in situ* dataset for testing models, the need for additional data in undersampled water types, particularly in oligotrophic waters, and error quantification of *in situ* data. In addition to testing the quantitative performance, algorithm selection for climate change studies need also to consider the suitability of the algorithm for the purpose and the development of future ocean-colour products. The objective classification developed here has the potential to be routinely implemented, for testing the performance of emerging ocean-colour algorithms and aiding their development.

## 8. Acknowledgments

NOMAD data were contributed by participants in the NASA SIMBIOS Program (NRA-96-MTPE-04 and NRA-99-OES-09 ) and by voluntary contributors. A cruise name accompanies each data record. Cruise details, including contributors' names, are available online (http://seabass.gsfc.nasa.gov/seabasscgi/nomad.cgi)

using the General Search and Cruise Search utilities to facilitate communication, collaboration, and acknowledgement. NASA should be commended for the development of the NOMAD database and for on-going *in situ* activities. This work is a contribution to the Ocean Colour Climate Change Initiative of the European Space Agency and was supported by the UK National Centre for Earth Observation.

## References

Aiken, J., Pradhan, Y., Barlow, R., Lavender, S., Poulton, A., Holligan, P., Hardman-Mountford, N. J., 2009. Phytoplankton pigments and functional types in the Atlantic Ocean: A decadal assessment, 1995-2005. Deep Sea Research I 56 (15), 899–917.

Bailey, S. W., Werdell, P. J., 2006. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. Remote Sensing Environment 102, 12–23.

Brewin, R. J. W., Devred, E., Sathyendranath, S., Hardman-Mountford, N. J., Lavender, S. J., 2011. Model of phytoplankton absorption based on three size classes. Applied Optics 50 (2), 4535–4549.

Bricaud, A., Babin, M., Claustre, H., Ras, J., Tiéche, F., 2010. Light absorption properties and absorption budget of Southeast Pacific waters. Journal of Geophysical Research 115, C08009.

Bricaud, A., Babin, M., Morel, A., Claustre, H., 1995. Variability in the chlorophyll specific absorption coefficients of natural phytoplankton: Analysis and parameterization. Journal of Geophysical Research 100, 13,321–13,332.

1172 Buiteveld, H., Hakvoort, J. H. M., Donze, M., 1994. The optical properties of
1173 pure water. In: Ocean Optics XII, Proc. SPIE, Vol. 2258, edited by: Jaffe, J.
1174 S., SPIE, Bellingham, Washington, 174-183.
1175 Ciotti, A. M., Lewis, M. R., Cullen, J. J., 2002. Assessment of the relation-
1176 ships between dominant cell size in natural phytoplankton communities and
1177 the spectral shape of the absorption coefficient. Limnology and Oceanography
1178 47 (2), 404–417.
1179 Cohen, J., Cohen, P., 1983. Applied multiple regression/correlation analysis for
1180 the behavioral sciences. L. Erlbaum Associates.
1181 Devred, E., Sathyendranath, S., Stuart, V., Platt, T., 2011. A three component
1182 classification of phytoplankton absorption spectra: Applications to ocean-
1183 colour data. Remote Sensing of Environment 115 (9), 2255–2266.
1184 Doerffer, R., Heymann, K., Schiller, H., 2002. Case 2 water algorithm for the
1185 medium resolution imaging spectrometer (MERIS) on ENVISAT. In: Pro-
1186 ceedings of the ENVISAT validation workshop, 9-13 December 2002, ESA
1187 report.
1188 Doerffer, R., Schiller, H., 2000. Neural Network for retrieval of concentrations
1189 of water constituents with the possibility of detecting exceptional out of scope
1190 spectra. In: IEEE 2000 International Geoscience and Remote Sensing Sym-
1191 posium, Honolulu, Hawaii USA, p. 714-717.
1192 Doerffer, R., Schiller, H., 2006. The MERIS Case 2 water algorithm. Interna-
1193 tional Journal of Remote Sensing 28 (3-4), 517–535.
1194 Doney, S. C., Lima, I. D., Moore, J. K., Lindsay, K., Behrenfeld, M. J., West-

berry, T. K., Mahowald, N., M., G. D., Takahashi, T., 2009. Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data. Journal of Marine Systems 76, 95–112.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Franz, B. A., Werdell, P. J., 2010. A Generalized Framework for Modeling of Inherent Optical Properties in Ocean Remote Sensing Applications. In: Ocean Optics XX, Anchorage, Alaska, 27$^{th}$ Sept.-1$^{st}$ Oct. 2010.

Friedrichs, M. A. M., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. A., Asanuma, I., Behrenfeld, M., Buitenhuis, E. T., Chai, F., Christian, J. R., Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J., Gentili, B., Gregg, W. W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Mélin, F., Moore, J. K., Morel, A., O'Malley, R. T. O., O'Reilly, J. E., Saba, V. S., Schmeltz, M., Smyth, T. J., Tjiputraw, J., Waters, K., Westberry, T. K., Winguth, A., 2009. Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean. Journal of Marine Systems 76 (1-2), 113–133.

Garver, S. A., Siegel, D. A., 1997. Inherent optical property inversion of ocean color spectra and its biogeochemical interpretation: 1. Time series from the Sargasso Sea. Journal of Geophysical Research 102, 18,607–18,625.

GCOS, 2011. Systematic observation requirements from satellite-based data

products for climate. Tech. rep., World Meteorological Organisation (WMO), 7 bis, avenue de la Paix, CH-1211 Geneva 2, Switzerland.

Glover, D. M., Jenkins, W. J., Doney, S. C., 2011. Modeling Methods for Marine Science. Cambridge Univeristy Press.

Gordon, H. R., Brown, O. B., Evans, R. H., Brown, J., Smith, R. C., Baker, K. S., Clark, S. K., 1988. A semianalytic radiance model of ocean color. Journal of Geophysical Research 93, 10,909–10,924.

Gordon, H. R., Clark, D. K., Brown, J. W., Brown, O. B., Evans, R. H., Broenkow, W. W., 1983. Phytoplankton pigment concentrations in the Middle Atlantic Bight: Comparison of ship determinations and CZCS estimates. Applied Optics 22, 20–36.

Guo, L., Hunt, B., Santschi, P., 2001. Effect of dissolved organic matter on the uptake of trace metals by American oysters. Environmental Science and Technology 35 (5), 885–893.

Hu, C., Lee, Z., Franz, B., 2012. Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. Journal of Geophysical Research 117, C01011.

Huot, Y., Morel, A., Twardowski, M. S., Stramski, D., Reynolds, R. A., 2008. Particle optical backscattering along a chlorophyll gradient in the upper layer of the eastern South Pacific Ocean. Biogeosciences 5, 495–507.

IOCCG, 2000. Remote Sensing of Ocean Colour in Coastal, and Other Optically Complex waters. Tech. rep., Sathyendranath, S. (e.d.), Reports of the International Ocean-Colour Coordinating Group, No. 3, IOCCG, Dartmouth,

1241 Canada.
1242 IOCCG, 2006. Remote Sensing of Inherent Optical Properties: Fundamentals,
1243 Tests of Algorithms, and Applications. Tech. rep., Lee, Z. P. (e.d.), Reports
1244 of the International Ocean-Colour Coordinating Group, No. 5, IOCCG, Dart-
1245 mouth, Canada.
1246 Kostadinov, T. S., Siegel, D. A., Maritorena, S., 2009. Retrieval of the particle
1247 size distribution from satellite ocean color observations. Journal of Geophysi-
1248 cal Research 114, C09015.
1249 Lavender, S., Pinkerton, M. H., Morales, J. F., Aiken, J., Moore, G. F., 2004. Sea-
1250 WiFS validation in European coastal waters using optical and bio-geochemical
1251 measurements. International Journal of Remote Sensing 25 (7-8), 1481–1488.
1252 Lee, Z., Arnone, R., Hu, C., Werdell, P., Lubac, B., 2010. Uncertainties of op-
1253 tical parameters and their propagations in an analytical ocean color inversion
1254 algorithm. Applied Optics 49 (3), 369–381.
1255 Lee, Z., Carder, K. L., Arnone, R. A., 2002. Deriving inherent optical properties
1256 from water color: a multiband quasi-analytical algorithm for optically deep
1257 waters. Applied Optics 41 (27), 5755–5772.
1258 Lee, Z., Carder, K. L., Mobley, C. D., Steward, R. G., Patch, J. S., 1998. Hy-
1259 perspectral remote sensing for shallow waters. 1. A semianalytical model. Ap-
1260 plied Optics 37 (27), 6329–6338.
1261 Lee, Z., Carder, K. L., Mobley, C. D., Steward, R. G., Patch, J. S., 1999. Hy-
1262 perspectral remote sensing for shallow waters. 2. Deriving bottom depths and
1263 water properties by optimization. Applied Optics 38, 3831–3843.

1264 Lee, Z., Du, K., Arnone, R., 2005. A model for the diffuse attenuation coefficient
1265 of downwelling irradiance. Journal of Geophysical Research 110, C02016.
1266 Lee, Z., Lubac, B., Werdell, P. J., Arnone, R., 2009. An Up-
1267 date of the Quasi-Analytical Algorithm (QAA_v5). Tech. rep., In-
1268 ternational Ocean Colour Coordinating Group (IOCCG) Online:
1269 http://www.ioccg.org/groups/software.html (assessed 10/02/2012).
1270 Loisel, H., Nicolas, J.-M., Sciandra, A., Stramski, D., A., P., 2006. Spectral
1271 dependency of optical backscattering by marine particles from satellite remote
1272 sensing of the global ocean. Journal of Geophysical Research 111, C09024.
1273 Maritorena, S., Fanton d'Andon, O. H., Mangin, A., Siegel, D. A., 2010. Merged
1274 satellite ocean color data products using a bio-optical model: Characteristics,
1275 benefits and issues. Remote Sensing Environment 114, 1791–1804.
1276 Maritorena, S., Siegel, D. A., Peterson, A. R., 2002. Optimization of a semi-
1277 analytical ocean color model for global-scale applications. Applied Optics
1278 41 (15), 2705–2714.
1279 Mélin, F., Berthon, J.-F., Zibordi, G., 2005. Assessment of apparent and inher-
1280 ent optical properties derived from SeaWiFS with field data. Remote Sensing
1281 Environment 97, 540–553.
1282 Mélin, F., Zibordi, G., Berthon, J.-F., 2007. Assessment of satellite ocean color
1283 products at a coastal site. Remote Sensing Environment 110, 192–215.
1284 Moore, T. S., Campbell, J. W., Dowell, M. D., 2009. A class-based approach to
1285 characterizing and mapping the uncertainty of the MODIS ocean chlorophyll
1286 product. Remote Sensing Environment 113, 2424–2430.

1287 Morel, A., 1974. Optical properties of pure water and pure seawater. In: Jerlov,
1288 N. G., Steemann Nielsen, E. (Eds.), Optical Aspects of Oceanogrpahy. Aca-
1289 demic, San Diego, California, pp. 1–24.
1290 Morel, A., 1980. In-water and remote measurements of ocean color. Boundary
1291 Layer Meteorology 18, 177–201.
1292 Morel, A., 2009. Are the empirical relationships describing the bio-optical prop-
1293 erties of case 1 waters consistent and internally compatible? Journal of Geo-
1294 physical Research 114, C01016.
1295 Morel, A., Antoine, D., 2011. MERIS Algorithm Theoretical Basis Documents
1296 (ATBD 2.9) - Pigment Index Retrieval in CASE 1 Waters (PO-TN-MEL-GS-
1297 0005), Issue 4, July 2011. Tech. rep., Laboratoire d'Océanographie de Ville-
1298 franche (LOV), MERIS ESL, ACRI-ST.
1299 Morel, A., Antoine, D., Gentili, B., 2002. Bidirectional reflectance of oceanic
1300 waters: accounting for Raman emission and varying particle scattering phase
1301 function. Applied Optics 41, 6289–6306.
1302 Morel, A., Gentili, B., 2009. A simple band ratio technique to quantify the col-
1303 ored dissolved and detrital organic material from ocean color remotely sensed
1304 data. Remote Sensing Environment 113, 998–1011.
1305 Morel, A., Huot, Y., Gentili, B., Werdell, P. J., Hooker, S. B., Franz, B. A.,
1306 2007. Examining the consistency of products derived from various ocean color
1307 sensors in open ocean (case 1) waters in the perspective of a multi-sensor
1308 approach. Remote Sensing of Environment 111, 69–88.
1309 Morel, A., Maritorena, S., 2001. Bio-optical properties of oceanic waters: A

reappraisal. Journal of Geophysical Research 106 (C4), 7163–7180.

Morel, A., Prieur, L., 1977. Analysis of variations in ocean color. Limnology and Oceanography 22, 709–722.

Mu, X., Shen, Q and, L. Z.-L., Yan, G., & Sobrino, J. A., 2011. A comparison of different optimization algorithms for retrieving aerosol optical depths from satellite data: an example of using a dual-angle algorithm. International Journal of Remote Sensing 32 (24), 8949–8968.

Müller, D., Krasemann, H., Brewin, R. J. W., Brockmann, C., Deschamps, P.-Y., Doerffer, R., Fomferra, N., Franz, B. A., Grant, M. G., Groom, S., Mélin, F., Platt, T., Regner, P., Sathyendranath, S., Steinmetz, F., Swinton, J., Submitted. The Ocean Colour Climate Chnage Initiative: I. A methodology for assessing atmospheric correction processors based on in-situ measurements. Remote Sensing Environment.

Nair, A., Sathyendranath, S., Platt, T., Morales, J., Stuart, V., Forget, M.-H., Devred, E., Bouman, H., 2008. Remote sensing of phytoplankton functional types. Remote Sensing of Environment 112 (8), 3366–3375.

NASA, June 2009. Diffuse attenuation coefficient (KD) for downwelling irradiance at 490-nm.
URL http://oceancolor.gsfc.nasa.gov/REPROCESSING/R2009/kdv4/

NASA, March 2010. Ocean Color Chlorophyll (OC) v6.
URL http://oceancolor.gsfc.nasa.gov/REPROCESSING/R2009/ocv6/

Nelson, N. B., Siegel, D. A., 2013. The Global Distribution and Dynamics of Chromophoric Dissolved Organic Matter. Annual Review of Marine Science

5, 20.1–20.3.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., McClain, C., 1998. Ocean chlorophyll algorithms for SeaWiFS. Journal of Geophysical Research 103 (C11), 24,937–24,953.

O'Reilly, J. E., Maritorena, S., Siegel, D.and O'Brien, M. C., Toole, D.and Mitchell, B. G., Kahru, M., Chavez, F. P., Strutton, P., Cota, G., Hooker, S. B., McClain, C. R., Carder, K. L., Muller-Karger, F., Harding, L., Magnuson, A., Phinney, D., Moore, G. F., Aiken, J., Arrigo, K. R., Letelier, R., Culver, M., 2000. Ocean color chlorophyll a algorithms for SeaWiFs, OC2, and OC4:. Tech. rep., In: Hooker, S. B., Firestone, E. R. (Eds.), SeaWiFS Postlaunch Technical Report Series. Vol. 11. SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3. NASA, Goddard Space Flight Center, Greenbelt, Maryland. 9-23.

Pope, R., Fry, E., 1997. Absorption spectrum (380-700 nm) of pure water. II. Integrating cavity measurements. Applied Optics 36 (33), 8710–8723.

Roy, S., Sathyendranath, S., Platt, T., 2010. Retrieval of phytoplankton size from bio-optical measurements: theory and applications. Journal of the Royal Society Interface 8 (58), 650–660.

Santschi, P., Lenhart, J., Honeyman, B., 1997. Heterogeneous processes affecting trace contaminant distribution in estuaries: the role of natural organic matter. Marine Chemistry 58 (1-2), 99–125.

Sathyendranath, S., Platt, T., 1997. Analytic model of ocean color. Applied Optics 36, 2620–2629.

1356 Sathyendranath, S., Stuart, V., Cota, G., Maas, H., Platt, T., 2001. Remote sens-
1357 ing of phytoplankton pigments: a comparison of empirical and theoretical
1358 approaches. International Journal of Remote Sensing 22, 249–273.
1359 Sathyendranath, S., Watts, L., Devred, E., Platt, T., Caverhill, C., Maass, H.,
1360 2004. Discrimination of diatoms from other phytoplankton using ocean-colour
1361 data. Marine Ecological Progress Series 272, 59–68.
1362 Shanmugam, P., 2011. New models for retrieving and partitioning the colored
1363 dissolved organic matter in the global ocean: Implications for remote sensing.
1364 Remote Sensing Environment 115, 1501–1521.
1365 Smyth, T. J., Moore, G. F., Hirata, T., Aiken, J., 2006. Semianalytical model for
1366 the derivation of ocean color inherent optical properties: description, imple-
1367 mentation, and performance assessment. Applied Optics 45 (31), 8116–8131.
1368 Werdell, P. J., 2005. An evaluation of inherent optical property data
1369 for inclusion in the NASA bio-Optical Marine Algorithm Data
1370 set," NASA Ocean Biology Processing Group Science Systems and
1371 Applications, Inc. Document Version 1.1, corresponding to NO-
1372 MAD Version 1.3 19th September 2005. Tech. rep., NASA: Online
1373 http://seabass.gsfc.nasa.gov/seabass/data/werdell_nomad_iop_qc.pdf (2005).
1374 Werdell, P. J., 2009. Global bio-optical algorithms for ocean color satellite ap-
1375 plications. Eos Trans. AGU 90 (1), 4.
1376 Werdell, P. J., Bailey, S. W., 2005. An improved in-situ bio-optical data set for
1377 ocean colour algorithm development and satellite data production validation.
1378 Remote Sensing Environment 98, 122–140.

1379 Werdell, P. J., Franz, B. A., Bailey, S. W., Feldman, G. C., Boss, E., Brando,
1380 V. E., Dowell, M., Hirata, T., Lavender, S. J., Lee, Z., Loisel, H., Maritorena,
1381 S., Mélin, F., Moore, T. S., Smyth, T. J., Antoine, D., Devred, E., d'Andon, O.
1382 H. F., Mangin, A., 2013. Generalized ocean color inversion model for retriev-
1383 ing marine inherent optical properties. Applied Optics 52, 2019–2037.
1384 WET-Labs, 2012. An introduction to in-situ absorption and attenuation meters.
1385 URL http://www.wetlabs.com/technicalnotes/technoteindex.htm
1386 Zhang, X., Hu, L., He, M.-X., 2009. Scattering by pure seawater: Effect of
1387 salinity. Optics Express 17, 5698–5710.

Table 1: Variables tested in the in-water comparison.

| Abbreviation | Variable | Usage | Unit |
|---|---|---|---|
| $L_w(\lambda)$ | Spectral water-leaving radiance | Input | uW cm$^{-2}$ nm$^{-1}$ sr$^{-1}$ |
| $E_s(\lambda)$ | Spectral surface irradiance | Input | uW cm$^{-2}$ nm$^{-1}$ sr$^{-1}$ |
| $R_{rs}(\lambda)$ | Remote sensing reflectance ($L_w(\lambda)/E_s(\lambda)$) | Input | sr$^{-1}$ |
| $\theta$ | Solar sun-zenith angle | Input[#] | Degrees |
| $C$ | HPLC chlorophyll-a concentration | Output | mg m$^{-3}$ |
| $K_d(489)$ | Diffuse downwelling irradiance coefficient at 489 nm | Output | m$^{-1}$ |
| $a(\lambda)$ | Total absorption coefficient | Output | m$^{-1}$ |
| $a_{ph}(\lambda)$ | Phytoplankton absorption coefficient | Output | m$^{-1}$ |
| $a_{dg}(\lambda)$ | Dissolved (gelbstoff) and detrital (non-algal) absorption coefficient | Output | m$^{-1}$ |
| $S_{dg}$ | Exponential slope of $a_{dg}$ with wavelength [$] | Output | nm$^{-1}$ |
| $a_{ph}(555)/a_{ph}(443)$ | Index of spectral shape in $a_{ph}$ | Output | Dimensionless |
| $b_b(\lambda)$ | Total backscattering coefficient | Output | m$^{-1}$ |
| $\gamma$ | Power slope of $b_b$ with wavelength [&] | Output | Dimensionless |

[$] Computed from fitting an exponential function to *in situ data* and model.
[&] Computed from fitting a power function to *in situ* data and model.
[#] Solar sun-zenith angle was used as input to some of the semi-analytical models and for estimating $K_d(489)$ (see Eq. (6)).
$\lambda$ = wavelength.

Table 2: Model output variables.

| Model | Output variable | | | | | | | | | Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| | $K_d(489)$ | $C$ | $a(\lambda)$ | $a_{ph}(\lambda)$ | $a_{dg}(\lambda)$ | $b_b(\lambda)$ | $\gamma$ | $S_{dg}$ | $a_{ph}(555)/a_{ph}(443)$ | |
| A | ×* | ×$ | × | × | × | × | × | × | × | Smyth et al. (2006) |
| B | ×* | ×$ | × | × | × | × | × | × | × | Smyth et al. (2006) |
| C | ×* | × | × | × | × | × | × | × | × | Devred et al. (2011) |
| D | ×* | ×$ | × | × | × | × | × | × | × | Lee et al. (2002) |
| E | ×* | ×$ | × | × | × | × | × | × | × | Lee et al. (2009) |
| F | ×* | ×$ | × | × | × | × | × | × | × | Lee et al. (1998, 1999) |
| G | ×* | × | × | × | × | × | × | × | × | Maritorena et al. (2002) |
| H | ×* | × | × | × | × | × | × | × | × | Maritorena et al. (2002) |
| I | ×* | × | × | × | × | × | × | × | × | Werdell et al. (2013) |
| J | ×* | | × | × | × | × | × | × | × | see# |
| K | ×* | × | × | × | × | × | × | × | × | Doerffer and Schiller (2000) |
| L | | × | | | | | | | | O'Reilly et al. (2000) |
| M | | × | | | | | | | | O'Reilly et al. (2000) |
| N | | × | | | | | | | | O'Reilly et al. (2000) |
| O | | × | | | | | | | | Morel et al. (2007) |
| P | | × | | | | | | | | Hu et al. (2012) |
| Q | × | | | | | | | | | NASA (2009) |

* Computed following Eq. (6) with $\theta$, $a(489)$ and $b_b(489)$ as input from the model.
$ Computed following Eq. (1) with $a_{ph}(443)$ as input from the model.
# This model represents a Case-1 approach that uses Model L as input. The model computes IOPs as a function of $C$ through combining relationships proposed by: Gordon et al. (1983); Buiteveld et al. (1994); Pope and Fry (1997); Huot et al. (2008); Morel (2009); Bricaud et al. (2010); Brewin et al. (2011).

Table 3: Summary of models used in the comparison.

| Model | Approach | Method | Input $R_{rs}$ wavelengths[#] | NOMAD Independence[$] |
|---|---|---|---|---|
| A | Semi-analytical | Algebraic | 411, 443, 489, 510, 555, 665 | 1 |
| B | Semi-analytical | Algebraic | 411, 443, 489, 510, 555, 665 | 2 |
| C | Semi-analytical | Optimisation | 443, 489, 510, 555 | 2 |
| D | Semi-analytical | Algebraic | 411, 443, 489, 510, 555, 665 | 1 |
| E | Semi-analytical | Algebraic | 411, 443, 489, 510, 555, 665 | 2 |
| F | Semi-analytical | Optimisation | 411, 443, 489, 510, 555, 665 | 1 |
| G | Semi-analytical | Optimisation | 411, 443, 489, 510, 555, 665 | 1 |
| H | Semi-analytical | Optimisation | 411, 443, 489, 510, 555, 665 | 1 |
| I | Semi-analytical | Optimisation | 411, 443, 489, 510, 555, 665 | 1 |
| J | Semi-analytical | Band-ratio | 443, 489, 510, 555 | 2 |
| K | Semi-analytical | Optimisation | 411, 443, 489, 510, 555, 665 | 1 |
| L | Empirical | Band-ratio | 443, 489, 510, 555 | 3 |
| M | Empirical | Band-ratio | 443, 489, 555 | 3 |
| N | Empirical | Band-ratio | 489, 555 | 3 |
| O | Empirical | Band-ratio | 443, 489, 510, 555 | 1 |
| P | Empirical | Band-ratio / CI[*] | 443, 489, 510, 555, 665 | 3 |
| Q | Empirical | Band-ratio | 489, 555 | 3 |

[#] Wavelengths used that are available in the comparison.

[$] Qualitative assessment of algorithm independence to NOMAD: 1 = NOMAD dataset has a small influence on model parameterisation; 2 = NOMAD dataset has some influence on model parameterisation; 3 = NOMAD dataset has a large influence on model parameterisation.

[*] CI refers to a colour index defined as the difference between $R_{rs}$ in the green region of the visible spectrum and a reference formed linearly between $R_{rs}$ in the blue and red region of the visible spectrum.

Table 4: Performance of semi-analytical models in the objective classification.

| Model | Higher performance | Lower performance |
|---|---|---|
| A | $a_{ph}$(411-510), $b_b$(411), $C$ | $a(\lambda)$, $a_{ph}$(555-665), $a_{ph}(555)/a_{ph}(443)$, $b_b(510-665)$, $\gamma$, $K_d$(489) |
| B | $a_{ph}$(411-443), $a_{dg}$(665) | $a$(489-665), $a_{ph}$(555-665), $a_{ph}(555)/a_{ph}(443)$, $K_d$(489) |
| C | $a_{ph}$(510-555), $a_{ph}(555)/a_{ph}(443)$, $S_{dg}$ | $a$(411-443), $a_{dg}(\lambda)$, $b_b(\lambda)$, $C$ |
| D | $a$(411-555), $a_{dg}(\lambda)$, $b_b(\lambda)$, $S_{dg}$, $\gamma$, $K_d$(489) | $a$(665), $a_{ph}$(510-665) |
| E | $a$(411-555), $a_{ph}$(411-443), $\gamma$, $K_d$(489) | $a$(665), $a_{ph}$(510-665) |
| F | $a(\lambda)$, $a_{dg}(\lambda)$, $S_{dg}$, $a_{ph}(555)/a_{ph}(443)$, $\chi^2$ | $a_{ph}$(411), $b_b$(665) |
| G | $a$(665), $a_{ph}$(510-665), $b_b(\lambda)$, $C$ | $a$(443-489) |
| H | $a$(510-665), $a_{ph}$(510-665), $b_b(\lambda)$, $C$ | $a$(443) |
| I | $a$(665), $a_{ph}(\lambda)$, $a_{ph}(555)/a_{ph}(443)$, $C$, $\chi^2$ | $a_{dg}$(555-665) |
| J | $a$(665), $a_{ph}(\lambda)$, $a_{ph}(555)/a_{ph}(443)$, $S_{dg}$ | $\chi^2$ |
| K | $a$(665), $a_{ph}(555)/a_{ph}(443)$ | $a$(489-510), $a_{ph}(\lambda)$, $a_{dg}(\lambda)$, $b_b$(411-510), $S_{dg}$, $\gamma$, $C$ |

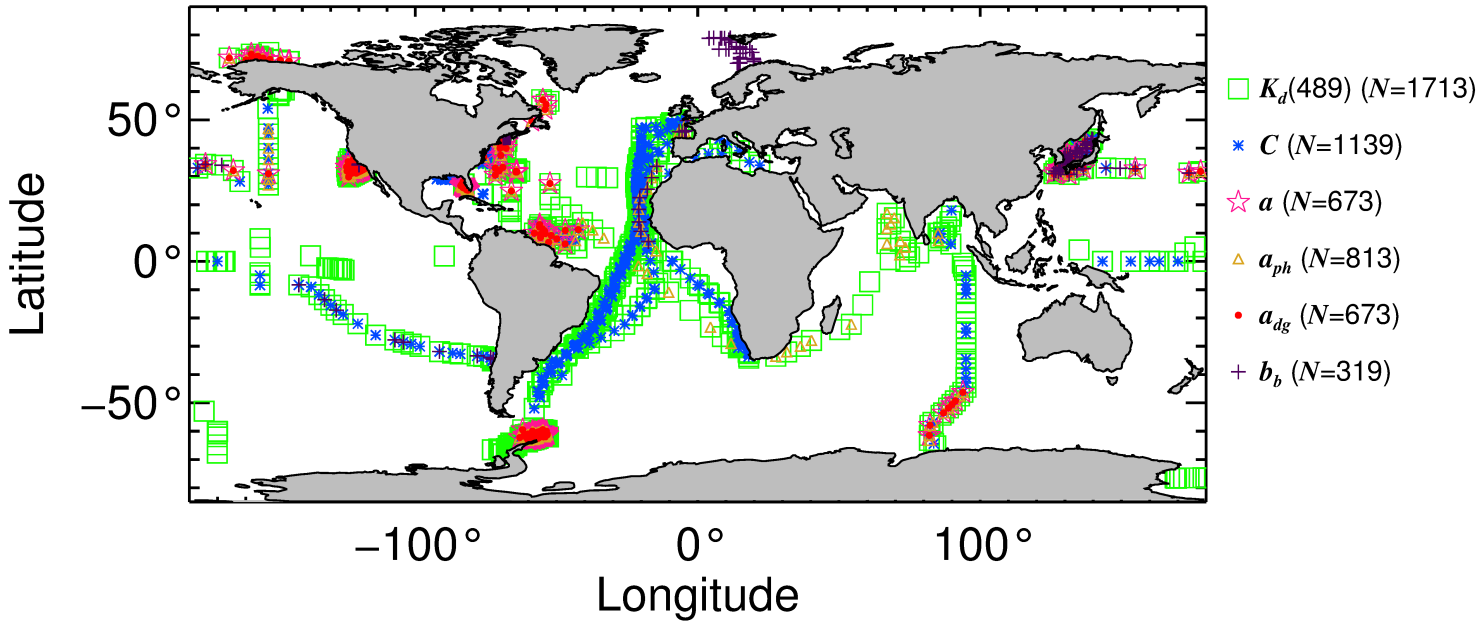


Figure 1: NOMAD *in situ* data used in the study ($N$ = number of samples).

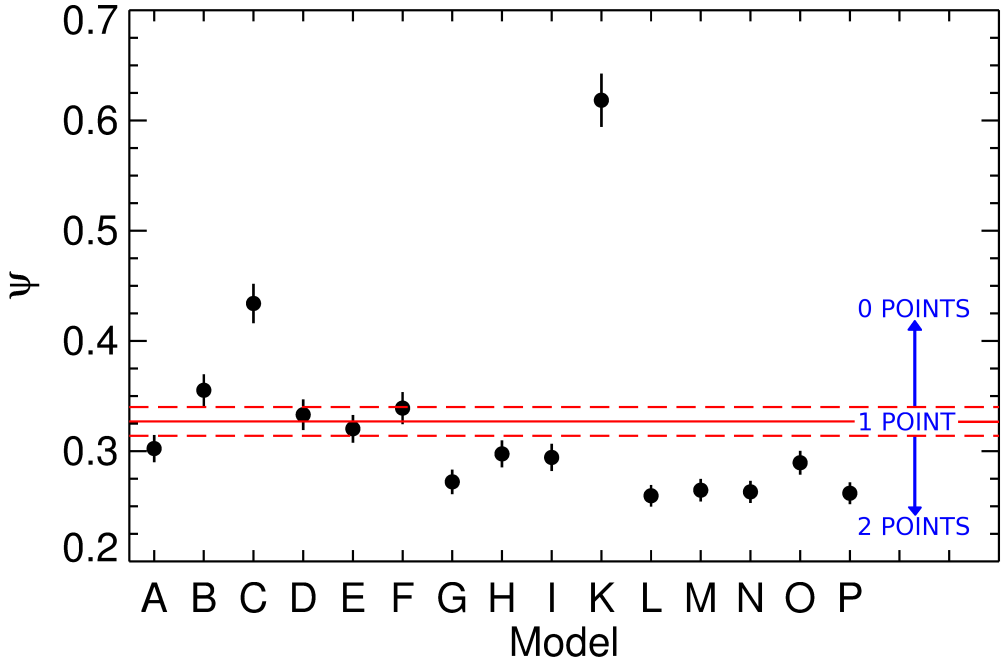


Figure 2: An example of the points classification for a number of models tested in the chlorophyll comparison using the Root Mean Square Error (Ψ). Red solid line represents the mean Ψ for all models and dashed red lines represent the mean Ψ ± mean 95% confidence intervals. The Ψ of each model is shown by the filled black circle and the black lines represent the Ψ of each model ±95% confidence intervals.

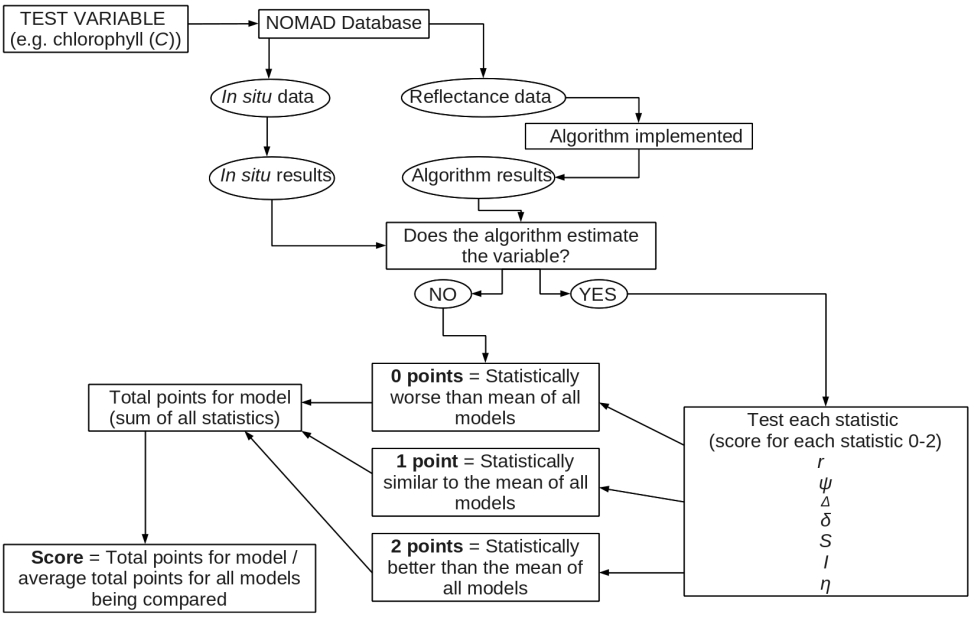


Figure 3: Flow chart illustrating the methodology of the scoring system.

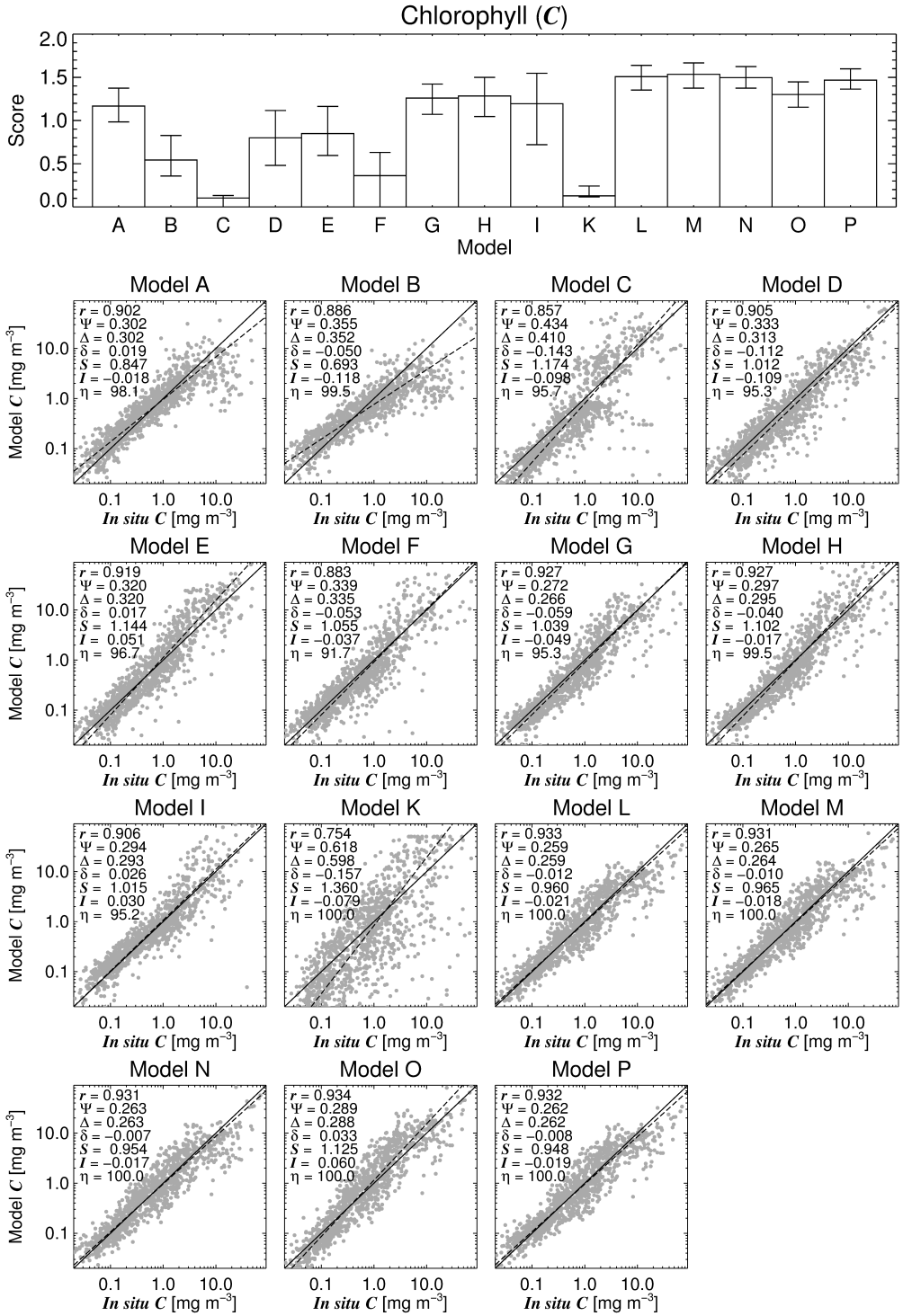




Figure 4: Results from the chlorophyll ($C$) model comparison.

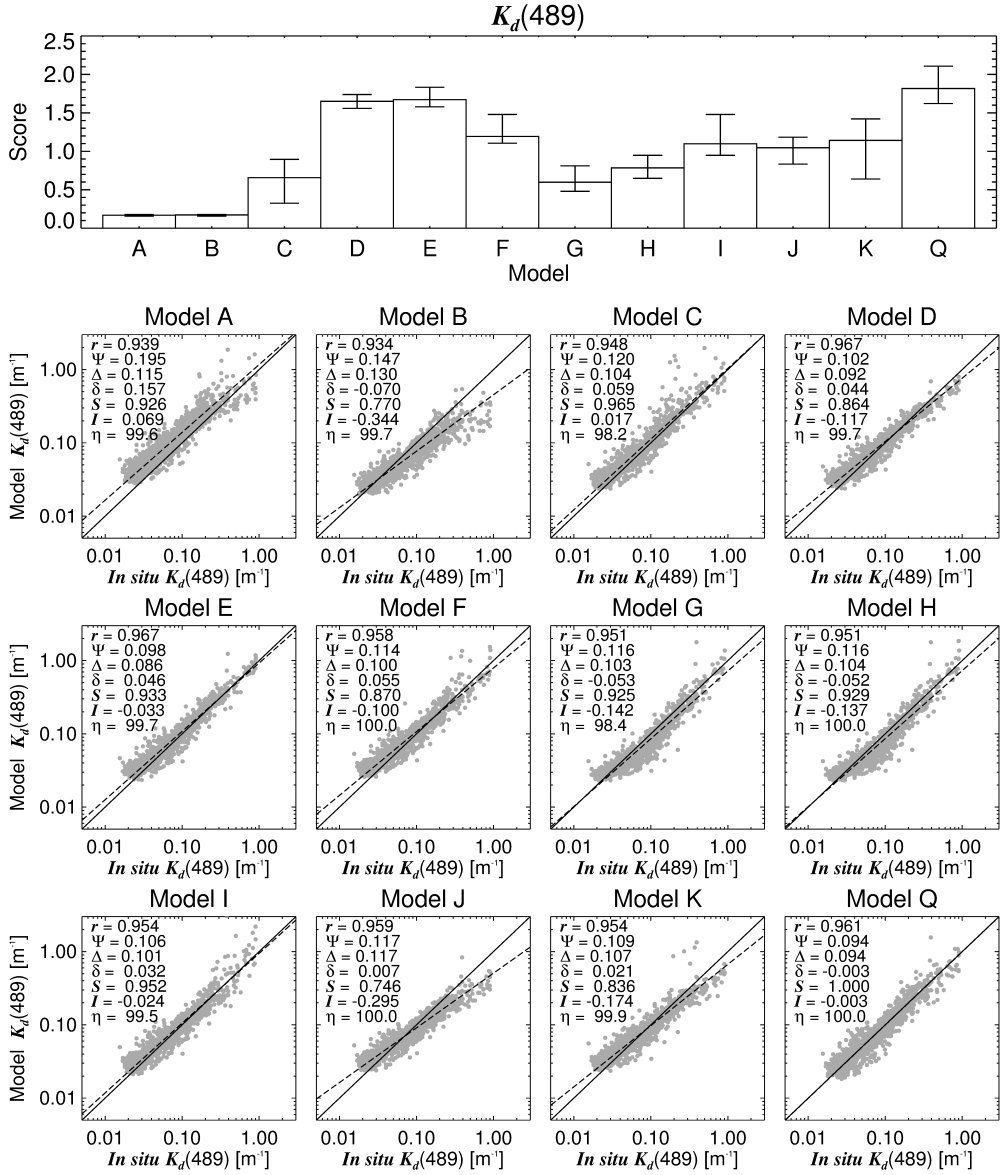


Figure 5: Results from the diffuse attenuation coefficient at 489 nm ($K_d$(489)) model comparison.

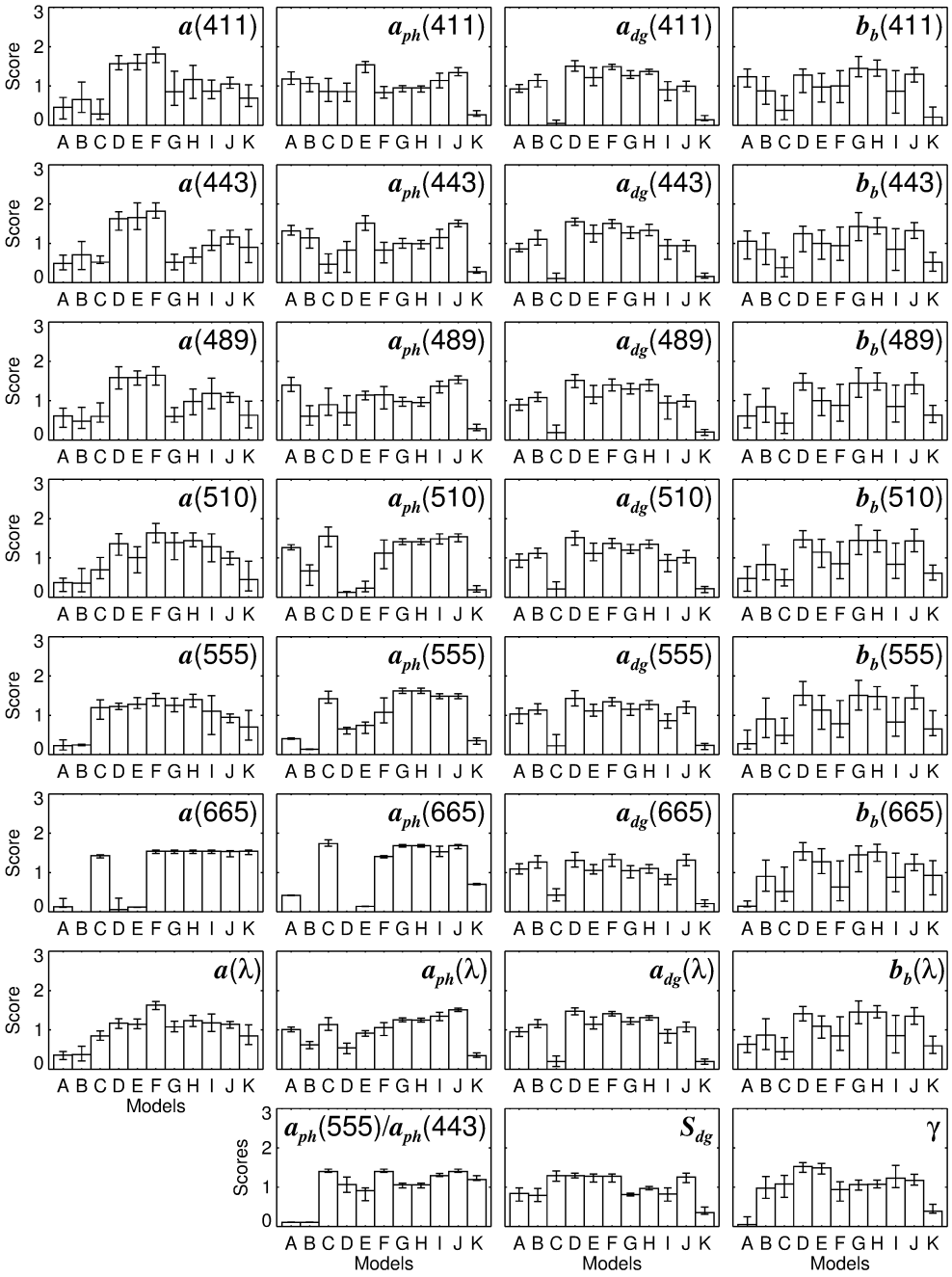


Figure 6: Results of the semi-analytical models at retrieving Inherent Optical Properties (IOP) according to the points classification.

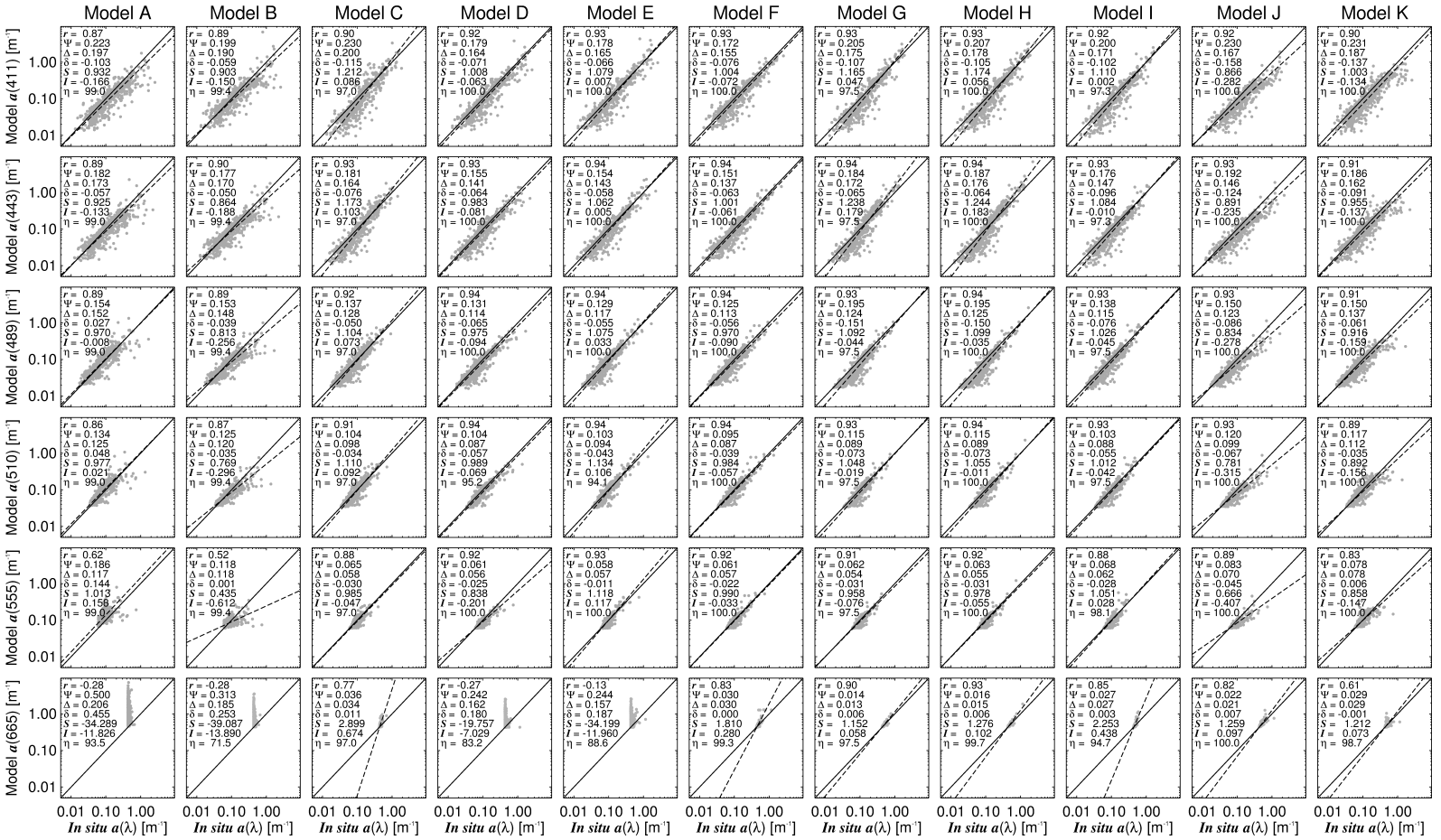


Figure 7: Scatter plots of the comparsion between model and *in situ* $a(\lambda)$

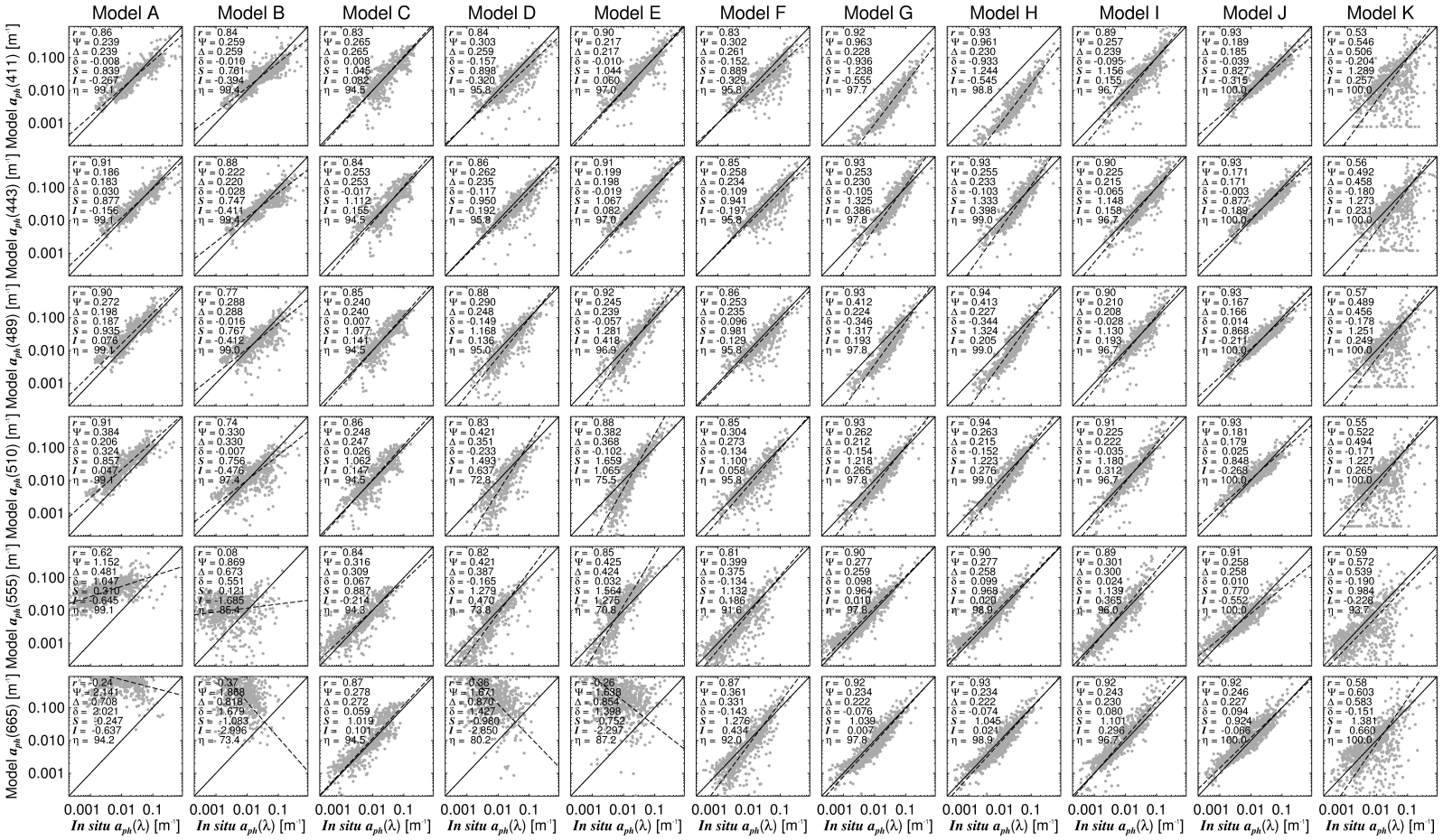


Figure 8: Scatter plots of the comparsion between model and *in situ* $a_{ph}(\lambda)$

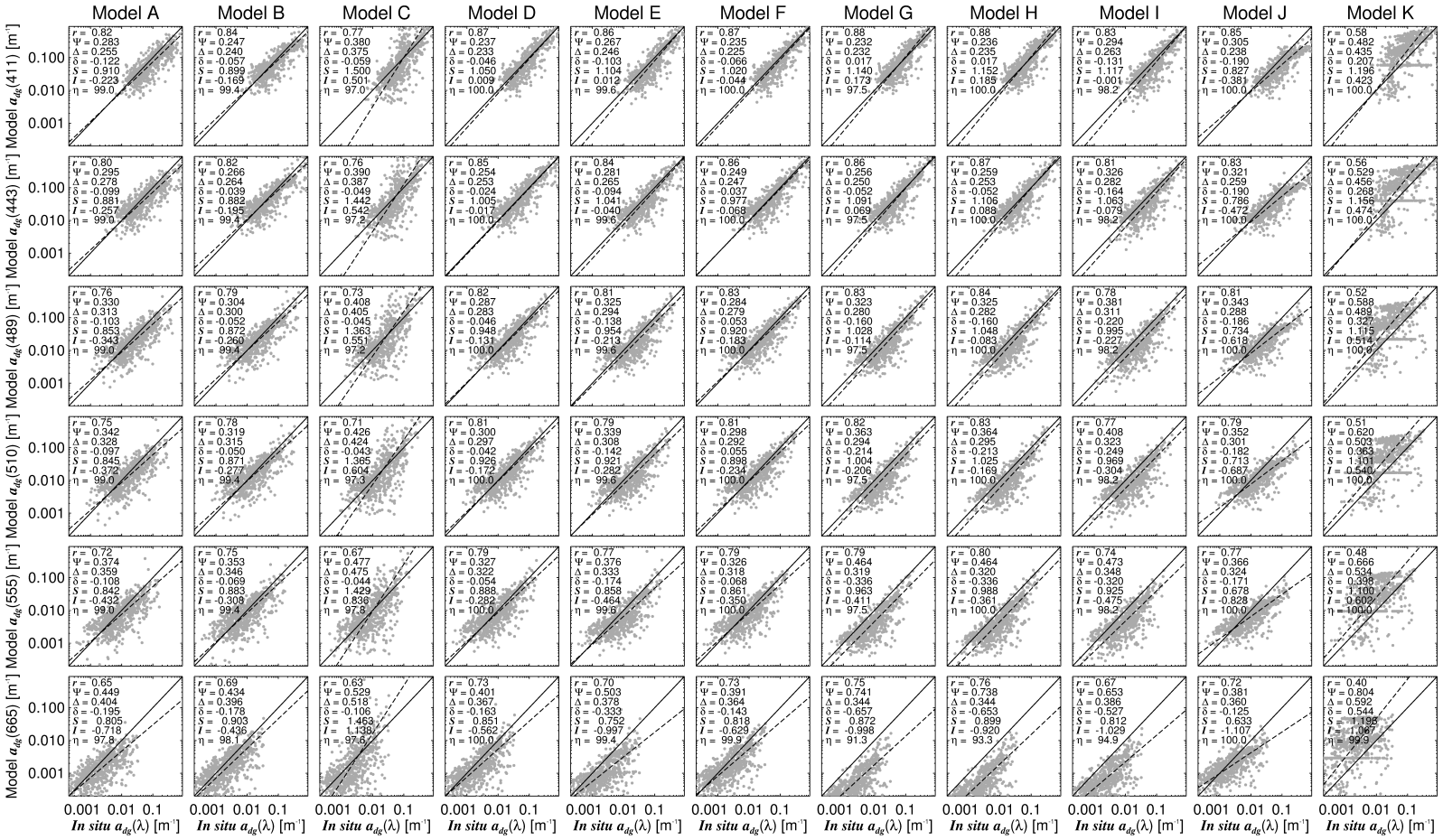


Figure 9: Scatter plots of the comparsion between model and *in situ* $a_{dg}(\lambda)$


Figure 10: Scatter plots of the comparsion between model and *in situ* $b_b(\lambda)$

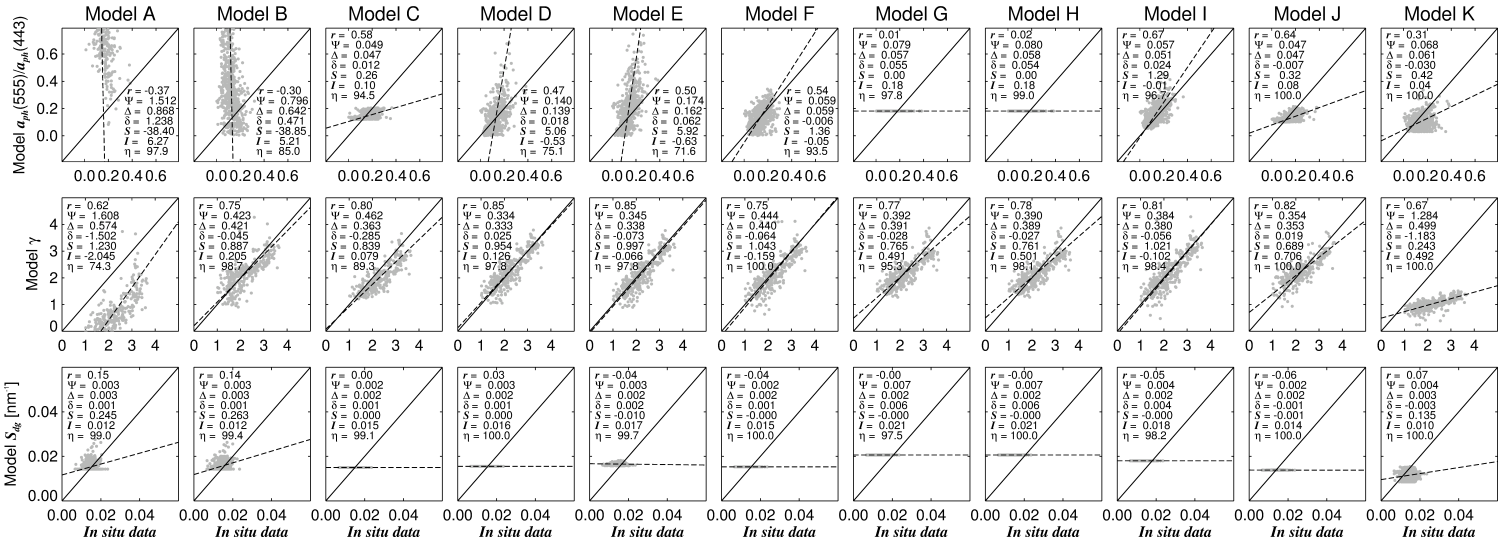


Figure 11: Scatter plots of the comparsion between model and *in situ* $a_{ph}(555)/a_{ph}(443)$, $\gamma$ and $S_{dg}$.

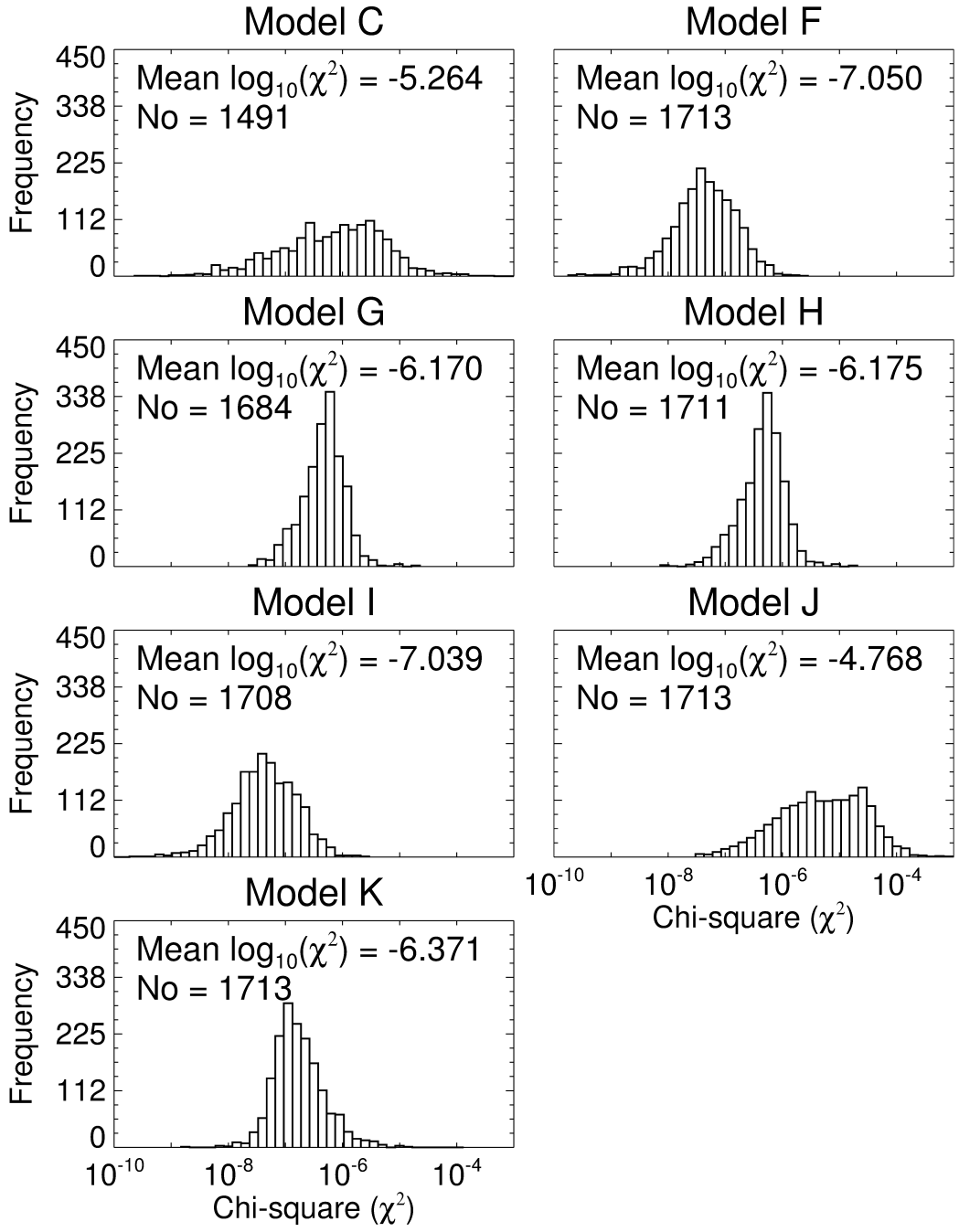


Figure 12: Results from the chi-square test.

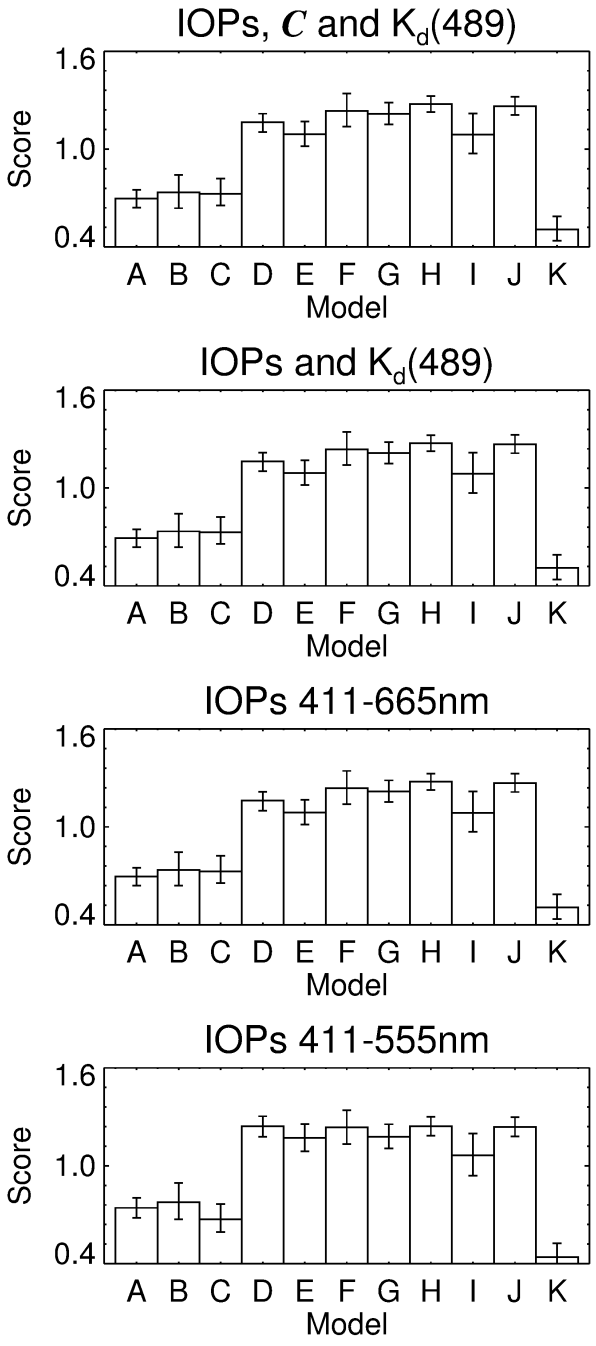


Figure 13: Results for semi-analytical models when summing all points in the classification.

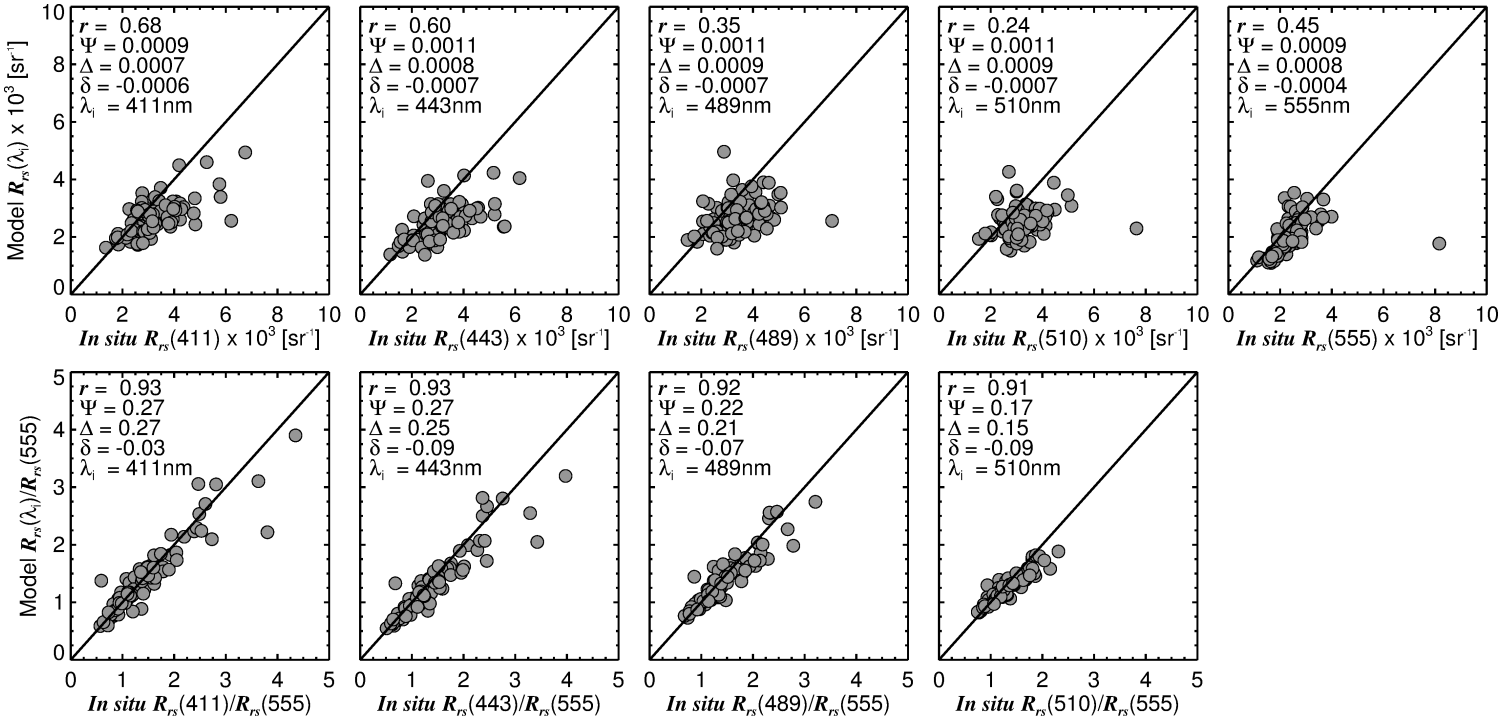


Figure 14: A comparison between measured $R_{rs}$ and modelled $R_{rs}$ at wavelengths from 411-555 nm for 87 samples in NOMAD with corresponding $R_{rs}$, $a$ and $b_b$. Modelled $R_{rs}$ has been reconstructucted using *in situ* $a$ and $b_b$ and the approximation of Gordon et al. (1988).