CMIP5 Earth System Models with biogeochemistry: a Ross Sea assessment

GRAHAM RICKARD and ERIK BEHRENS

National Institute of Water and Atmospheric Research, PO Box 14-901, Kilbirnie, Wellington, New Zealand
graham.rickard@niwa.co.nz

Abstract: An assessment is made of the ability of the Coupled Model Intercomparison Project 5 (CMIP5) models to represent the seasonal cycles of biogeochemistry of the Ross Sea over the late twentieth century. In particular, sea surface temperature, sea ice concentration, surface chlorophyll *a*, nitrate, phosphate and silicate, and the depth of the seasonal thermocline (measuring vertical mixing) are examined to quantify the physical-biogeochemical capabilities of each model, and to provide for 'ranked' model ensembles. This permits critical assessment of modelled Ross Sea biogeochemical cycling, including less well observed variables such as iron and vertically integrated primary production. The assessment enables determination of model output confidence limits; these confidence limits are used to examine future model scenario projections for consideration of potential ecosystem changes. The future scenarios examined are the representative concentration pathways rcp4.5 and rcp8.5. Our study suggests that by the end of the twenty-first century under rcp4.5 and/or rcp8.5 that there will be average increases in sea surface temperature, surface chlorophyll *a*, integrated primary production and iron, average decreases in surface nitrate, phosphate and silicate, and relatively large decreases in the depth of the seasonal thermocline and percentage coverage by sea ice in the Ross Sea.

Introduction

Numerical model output for the Ross Sea can provide for an integrated picture of potential relationships between the oceanic flow and the distributions of species. As highlighted by Brandt *et al.* (2014), regular and routine measurement of both the physical and biological components of the Ross Sea remains an on-going challenge. This is particularly apparent for the deepest waters of the Ross Sea, not only in terms of establishing present day baselines but also in terms of the impact of likely future stresses (changes in temperature, acidification, phytoplankton deposition, etc.) on the deep sea communities.

In this context, Constable *et al.* (2014) used projections from the most recent generation of Earth System Models (ESMs) in the Coupled Model Intercomparison Project 5 (CMIP5) to look at potential future impacts on Southern Ocean biota. This is the first time that ESMs have been included in the CMIP archive, and clearly represents a first step for ESM development. Nevertheless, the ESMs in CMIP5 do include projections via representative concentration pathways (rcps), and the future projections rcp4.5 and rcp8.5 (Moss *et al.* 2010, van Vuuren *et al.* 2011) are used in particular to look at potential changes over the coming century.

Bowden *et al.* (2011) also used a combination of model output and satellite data to explain observed variations in benthic communities between isolated seamounts in the Ross Sea. Smith *et al.* (2014b) used output from a climate model (ECHAM5 in CMIP3, predecessor of the MPI models in CMIP5) to produce regional downscaled solutions for the present day and the future for the Ross Sea region. As shown later, the broad range of solutions of the climate scale models for the Ross Sea suggests care has to be taken in translating processes at the relatively large scale of these models to the regional scales. Nevertheless, it is apparent that the climate models are useful for filling in the observational gaps for the present day, to provide for hypothesis testing for species distributions, and then to extrapolate forward for future scenarios.

Part of the focus here on the Ross Sea is motivated by initiatives aimed at producing an 'end to end' trophic model of the region in order to try to identify potential impacts of, for example, fishing and/or future changes in climate on the Ross Sea ecosystem. Such a trophic/ foodweb model is reported by Pinkerton & Bradford-Grieve (2014) building on earlier developments by Pinkerton *et al.* (2010). The base of the trophic model is determined by the phytoplankton, and the integrated or net primary production (Intpp or NPP), i.e. the amount of

Table I. CMIP5 Earth System Models used in this study with their respective biogeochemical (BGC) models and the primary macro-nutrients included in each BGC model.

Model number	Model	BGC model	NO$_3$	PO$_4$	SiO$_3$	Fe
1	CANESM2	CMOC	✓	x	x	x
2	CESM1BGC	BEC	✓	✓	✓	✓
3	CMCC-CESM	PELAGOS	✓	✓	x	✓
4	CNRM-CM5	PISCES	✓	✓	✓	✓
5	GFDL-ESM2G	TOPAZ2	✓	✓	✓	✓
6	GFDL-ESM2M	TOPAZ2	✓	✓	✓	✓
7	GISS-E2-H-CC	NOBM	✓	x	✓	✓
8	GISS-E2-R-CC	NOBM	✓	x	✓	✓
9	HadGEM2-CC	Diat-HadOCC	✓	x	✓	✓
10	HadGEM2-ES	Diat-HadOCC	✓	x	✓	✓
11	IPSL-CM5A-LR	PISCES	✓	✓	✓	✓
12	IPSL-CM5A-MR	PISCES	✓	✓	✓	✓
13	IPSL-CM5B-LR	PISCES	✓	✓	✓	✓
14	MPI-ESM-LR	HAMOCC5.2	✓	✓	✓	✓
15	MPI-ESM-MR	HAMOCC5.2	✓	✓	✓	✓
16	MRI-ESM1	MRI.COM3 + NPZD	✓	✓	x	x

carbon produced per square metre per day over the Ross Sea.

As Pinkerton & Bradford-Grieve (2014) note, 'the importance of phytoplankton in the Ross Sea is clear and changes to the magnitude or characteristics (e.g. spatial patterns, seasonal progression and/or prymnesiophyte–diatom balance) of Ross Sea phytoplankton are likely to have considerable consequences for regional ecosystem structure and function'. Therefore, it is critical to provide best estimates of phytoplankton and NPP distributions (both seasonally and spatially) for the present day, and under future pathways. In this context, we seek to provide guidance on the best estimates of these parameters from the CMIP5 models.

The Ross Sea part of the Southern Ocean is also known to be the most productive. In their synthesis of Southern Ocean primary production from satellite data and *in situ* measurements, Arrigo *et al.* (2008) note values of between 13 and 237 mmol C m^{-2} d^{-1} from 95 samples over the Ross Sea. The synthesis is then used to build up annual primary production rates over the 1997–2006 satellite period, the main summaries given in tables 4 & 5 in Arrigo *et al.* (2008). The results show that of the total mean production over the Southern Ocean shelves of 66.1 Tg C yr^{-1}, by far the largest proportion, at 35% of that total, comes from the Ross Sea shelf.

In terms of model assessment, Bopp *et al.* (2013) looked at ecosystem stressors in global CMIP5 projections, and tried to place a measure of 'robustness' on the model predictions. From their fig. 14 it is apparent that the Ross Sea region is not assigned robust indicators for changes in sea surface temperature (SST), surface pH, Intpp and sub-surface O$_2$. Anav *et al.* (2013) also analysed CMIP5 ESMs to assess their representation of the present day land and ocean carbon cycles, and proposed a set of metrics to determine a ranked set of the models. They noted that their analysis focused on variables central to

carbon fluxes, and that other variables (such as chlorophyll *a* (hereafter chlorophyll) concentration) could be used in such an assessment. They also noted that 'there is a level of subjectivity in all such assessments, partly due to the variables chosen and partly from the choice of observational data' and suggest that 'users of the CMIP5 models need to assess each model independently for their regions of interest, against those variables that are important for their specific subject of research'. Their analysis is based on a pdf-derived skill score, which they suggest enables a quantification of model ability to simulate ranges of behaviour including the mean, interannual variability and trend, but that these measures are not necessarily 'definitive'.

Cabré *et al.* (2014) provide for a complementary analysis to those of Anav *et al.* (2013) and Bopp *et al.* (2013) by looking at the full suite of CMIP5 ESMs in the context of quantifying model predictions regarding changes to global primary and export production. Their analysis considers a number of physically based ecological biomes to divide up the globe, and a multi-model averaged biome map at the bottom of their fig. 2 shows the Ross Sea included within a 'marginal sea ice biome' (light green). Their analysis uses a bootstrap statistical technique to assess the significance of future trends in the multi-model mean fields, and hence get a measure of the most consistent signal from the model output. They also note (parallel to Anav *et al.* 2013) that further work is needed to obtain improved metrics for robustness in such multi-model assessments.

Boyd *et al.* (2015) use rotated factor analysis to examine projections from one of the CMIP5 ESMs (model 2, CESM1BGC, in Table I) in order to expose regional variations across the globe of the potential impact of biologically important multi-stressors. The analysis shows complex regional responses, thus providing not only for important interpretation of multi-stressor impacts within the models themselves, but also providing guidance toward field and laboratory experimental design. Clearly, the advent of the CMIP5 ESMs presents a challenge to attribution and verification in the marine biological context.

The main aim of this study was to try to place some confidence limits on the CMIP5 ESMs by examining their representation of the present day state of the Ross Sea compared to climatological datasets and *in situ* observations. These confidence limits can provide advice on the probable changes to the Ross Sea induced by the rcps. Present day and scenario model output can then be used to drive the lower levels of the foodweb models for more complete trophic assessments of the present day state, and further potential future impacts. Our analysis complements that of Anav *et al.* (2013), Bopp *et al.* (2013), Boyd *et al.* (2015) and Cabré *et al.* (2014) by attempting to provide a ranking for the model biogeochemical outcomes
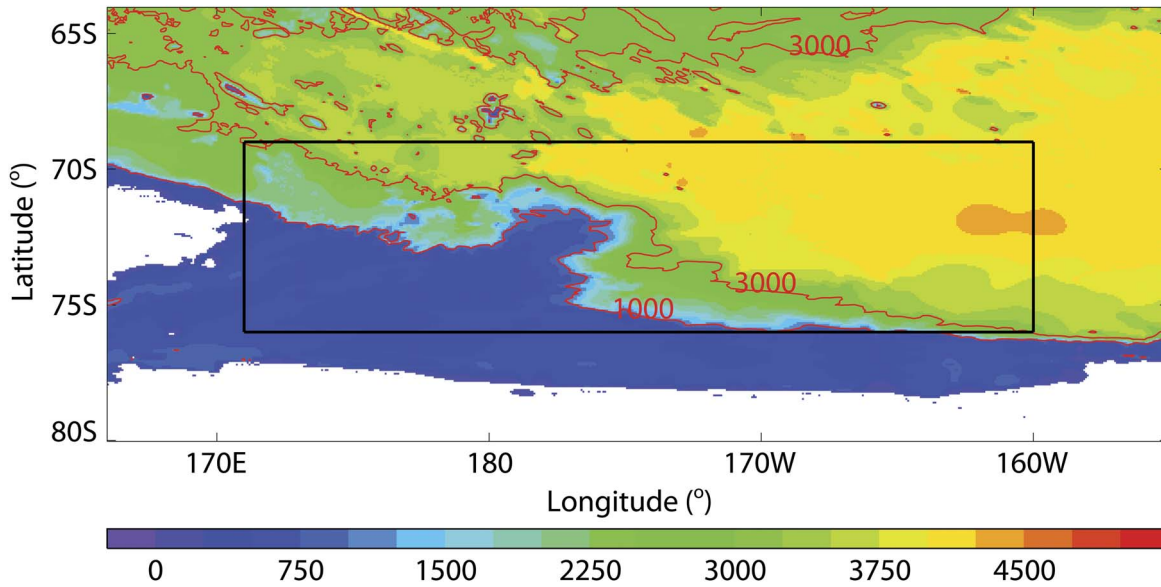
Fig. 1. Colour contour plot of the bathymetry (in metres) of the Ross Sea region (from GEBCO, www.ngdc.noaa.gov/mgg/gebco). The area within the black outline, referred to as region ROSS, is used for the data and model analysis, and spans 171°E to 160°W in longitude and 76°S to 69°S in latitude. Region ROSS samples the Ross Sea continental shelf (< 1000 m) and off-shelf waters (> 1000 m) of the Ross Sea environs. Red contours show isobaths at 1000 and 3000 m.
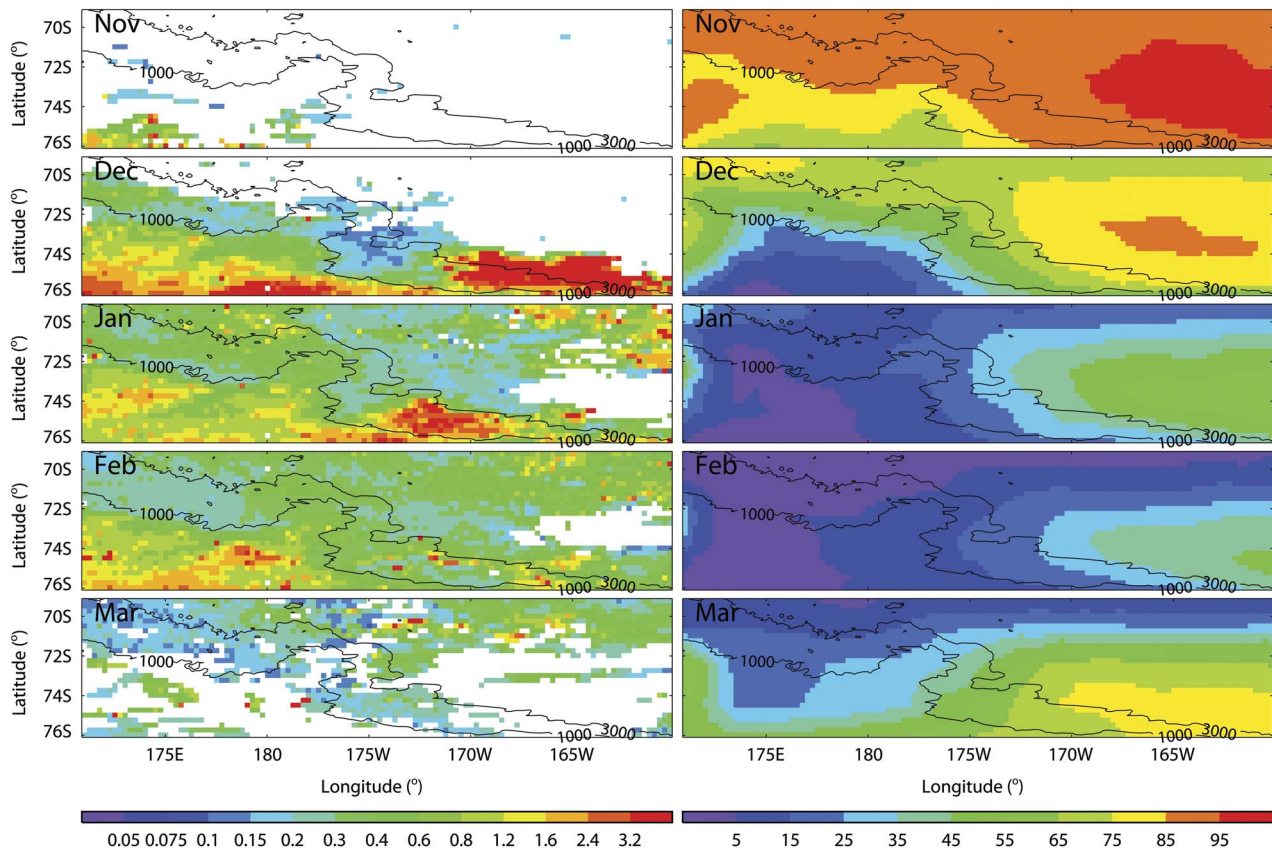

Fig. 2. Monthly mean surface chlorophyll from SeaWiFS 1997–2005 (mg Chl m^{-3}, left column) and sea ice concentration (% conc., right column) from HadISST (Rayner *et al.* 2003) for 1976–2005 for November to March over region ROSS (see Fig. 1). Note the non-uniform scale for chlorophyll. Black contour lines show bathymetry at 1000 and 3000 m.
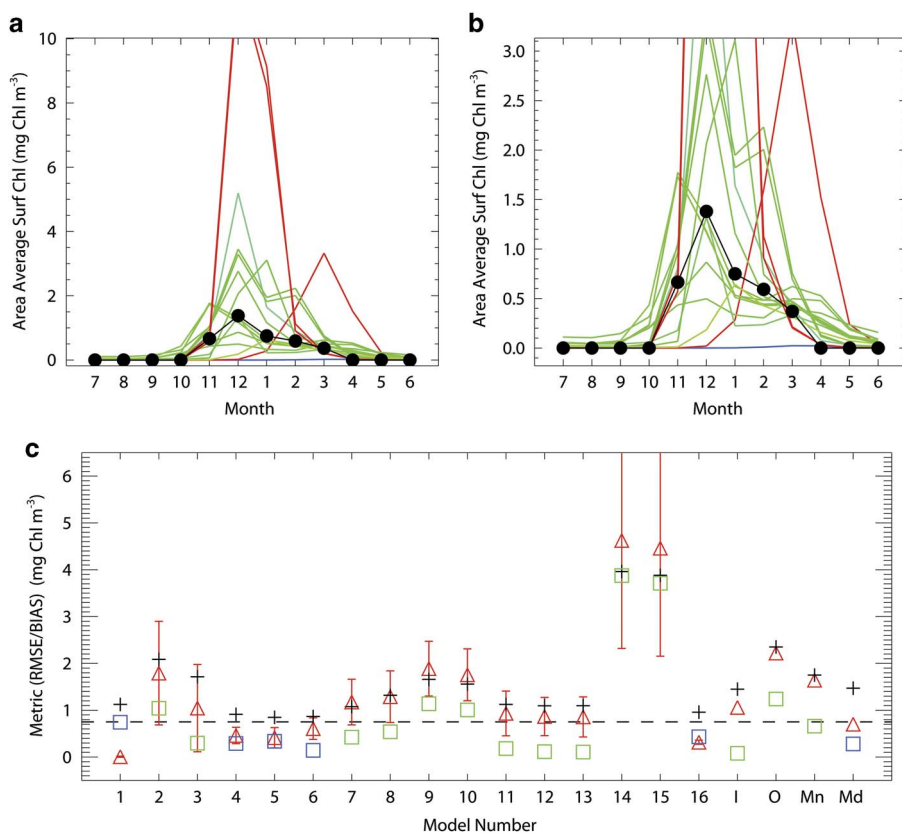
Fig. 3a. Satellite (black circles) and individual model monthly mean area averaged surface chlorophyll (mg Chl m^{-3}). **b.** Monthly mean area averaged surface chlorophyll (mg Chl m^{-3}) expanded around satellite data. **c.** Averaged November to March historical period root-mean-square error (RMSE) (+) and bias (BIAS) (model data) (□) metrics of area averaged surface chlorophyll per model. BIAS greater/less than zero plotted green/blue. The red triangle shows mean model area average surface chlorophyll, the red vertical line shows one standard deviation about the mean. The horizontal dashed line shows the November to March satellite average (0.75 mg Chl m^{-3}). Inner/outer ensemble and multi-model mean/median ensemble metrics (see text) are labelled I/O and Mn/Md, respectively.

for the Ross Sea region alone, and thus limits on the future predictions implied by the rcps critical to the kind of assessments reported by Constable *et al.* (2014).

The observational datasets and the numerical models to be analysed, along with the specific metrics, are detailed in the next section. The model analysis is divided into present day comparisons and model projections, followed by a comprehensive summary.

Methods: models and observational data

Here 16 full ESMs from a total of ∼ 60 models from 28 modelling centres whose output has been archived for CMIP5 are analysed. These 16 ESMs span a range of biogeochemical components and pathways. Table I lists the models used in this study, with a reference to their biogeochemical model and an indication of the primary macronutrients that each model simulates. Note that not only do the biogeochemical models vary in their levels of sophistication, but also in the number of nutrients they contain.

As biogeochemistry is a focus here, it is relevant to list the biogeochemical models for reference. There are nine biogeochemical models sampled since CNRM-CM5 (model 4 in Table I) and the IPSL models (11, 12 and 13) use PISCES (Aumont & Bopp 2006, Séférian *et al.*

2013). CANESM2 (1) uses CMOC (Zahariev *et al.* 2008, Christian *et al.* 2010), CESM1BGC (2) uses BEC (see Moore *et al.* 2013 and references therein), CMCC-CESM (3) uses PELAGOS (Vichi *et al.* 2007), the GFDL models (5 and 6) use TOPAZ2 (see the technical description provided in the supplement to Dunne *et al.* 2013), the GISS models (7 and 8) use NOBM (Gregg 2008), the Hadley Centre models (9 and 10) Diat-HadOCC (Palmer & Totterdell 2001) and the MPI models (14 and 15) HAMOCC5.2 (Ilyina *et al.* 2013). The biogeochemical model used in MRI-ESM1 (model 16) is outlined in Adachi *et al.* (2013), combining their dynamical component MRI.COM3 with a NPZD ecosystem model based on Oschlies (2001).

The 16 models have been assessed in terms of their representation of 'present day climate' or 'historical period', taken to be the period 1976 to 2005 of the model run. This assessment is used to give likelihood estimates on the model future projections under rcp4.5 and rcp8.5, and as a way to construct boundary conditions for downscaling of future states in a high resolution numerical model. At the time of writing the 16 models used had archived most (but not quite all) of the relevant physical and biogeochemical variables for the historical period and for both the rcps that we were able to post-process and analyse. Gaps in model output (in particular
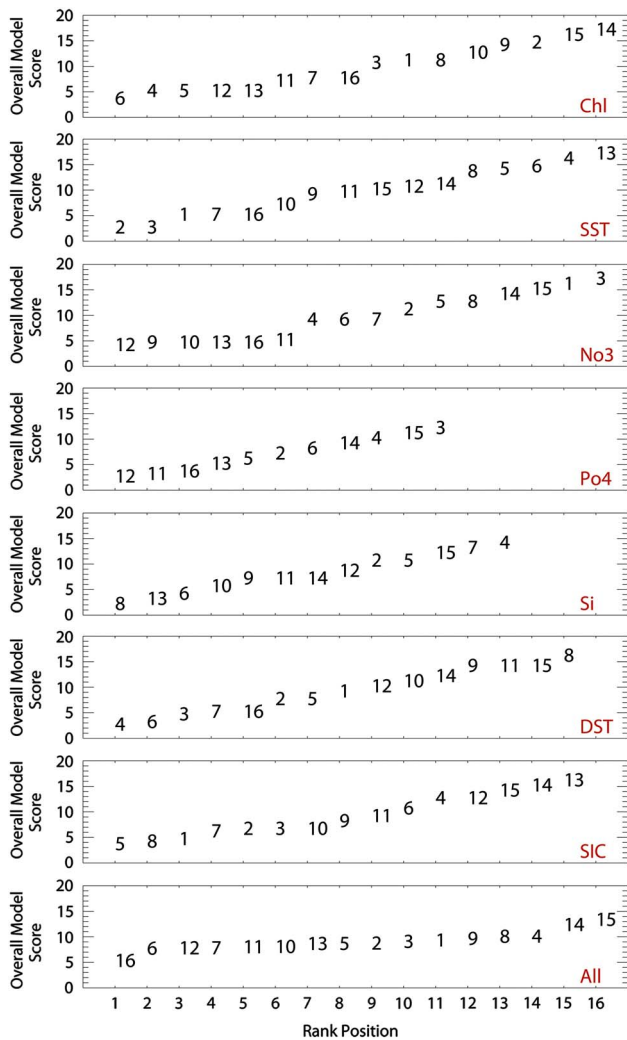
Fig. 4. Score and ranking per variable for all models in region ROSS (see Table I). The numbers plotted are the model numbers. Models are ranked by score from lowest (the 'best') to highest ('worst'). 'All' combines the scores from all of the other variables to obtain an estimate of an overall ranking.

associated with the rcps) arise because either the run was never performed, or the relevant model output was never archived.

Figure 1 contours the bathymetry of the Ross Sea region. Within this region a common domain for analysis of observations and model output is shown as the black bounding box. This bounding box has been chosen to capture some of the shelf and deeper waters associated with the Ross Sea gyre, and to avoid inter-model coastline inconsistencies associated with differences in the horizontal spatial resolution between the CMIP5 models, particularly along the front of the Ross Ice Shelf (noting that the present generation of CMIP5 models do not explicitly represent ice shelf processes). We refer to this common domain as region ROSS.

Figure 2 shows monthly mean averages for November to March over region ROSS of surface chlorophyll (left column) derived from the satellite data averaged over 1997–2005 from the NASA Sea-viewing Wide Field-of-view Sensor (SeaWiFS) instrument (with the white regions missing data), and sea ice concentration (SIC) (right column) from HadISST (Rayner *et al.* 2003) averaged over 1976–2005. According to the supplementary evidence of Pinkerton & Bradford-Grieve (2010) to the Pinkerton *et al.* (2010) paper, there is evidence of satellite underestimate of the *in situ* chlorophyll in the Ross Sea by ~ 13%, and this is accounted for in the satellite data used here. Further, the range of interannual variation is between 0.4 and 1.6 of each monthly mean estimate, and these values are used as a variance measure around the mean. The ROSS region overlaps with that used by Pinkerton & Bradford-Grieve (2014), and the present estimates of average surface chlorophyll over the respective domains are similar to those reported in the Pinkerton & Bradford-Grieve (2010) supplementary evidence.

Compared to the surface chlorophyll images from the wider domain in fig. 2 in Smith *et al.* (2012), region ROSS misses some of the high intensity signals on the shelf close to the Ross Ice Shelf, but does capture the large amplitude chlorophyll over the shelf break to the east of the domain in December and January. Monthly area averages (not shown) reveal that ROSS matches the timing of the Smith *et al.* (2012) domain seasonal cycle, except perhaps underestimating early growth in November, and captures ~ 60% of the total integrated surface chlorophyll.

The HadISST monthly averages for SIC show the emergence of the Ross Sea polynya from November into January, followed by the wider expanse of open water formation in February and the subsequent regrowth from the east and south of region ROSS during March. Compared to fig. 2 in Smith *et al.* (2012) it is again apparent that region ROSS is capturing most of the important SIC seasonal signal.

Data representing the historical period are from the World Ocean Atlas (WOA) 2009 database for salinity (Antonov *et al.* 2010), temperature (Locarnini *et al.* 2010), and the nutrients nitrate, phosphate and silicate (Garcia *et al.* 2010), and SIC from HadISST (Rayner *et al.* 2003). These sources provide spatial and temporal best fits to *in situ* observations, and are, therefore, best suited for assessment of climate model output in which we are looking for a statistical comparison of spatial and seasonal patterns averaged over the historical period, rather than a direct match to the individual observations themselves.

For the major circulation factors in the Southern Ocean we consider the Drake Passage transport, and the peak circulation in the sub-polar gyres. Values for the present
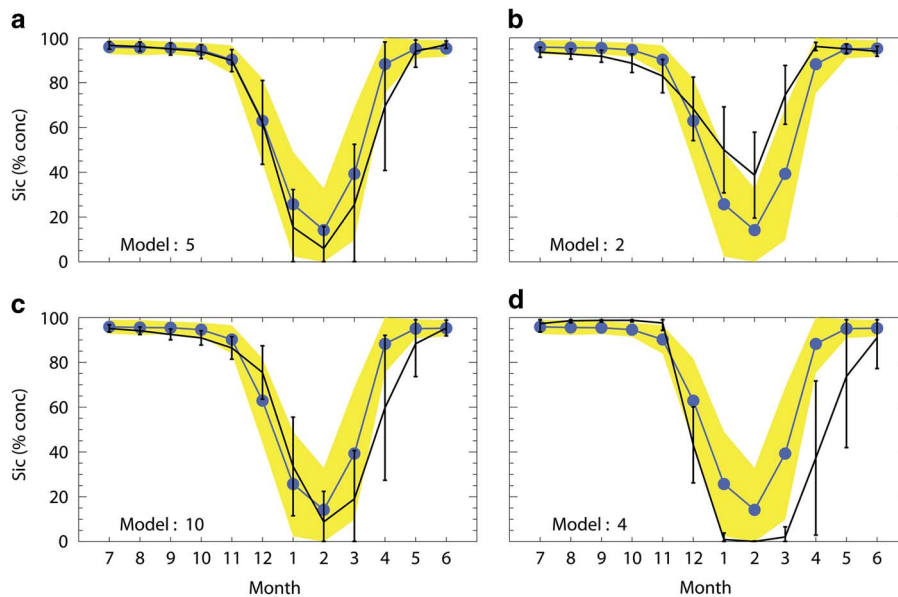
Fig. 5. Examples of model monthly mean sea ice concentration (SIC) for comparison with HadISST for models **a.** 5, **b.** 2, **c.** 10 and **d.** 4. The black and blue/yellow curves are for model and HadISST monthly mean and variance over the period 1976–2005, respectively.

day Drake Passage transport are obtained from long-term measurements (as indicated by Meijers *et al.* (2012) in their analysis of CMIP5 models in reference to Cunningham *et al.* (2003) and Griesel *et al.* (2012)) spanning a range of 134–164 Sv. The peak sub-polar Ross Sea and Weddell Sea gyre transports are from the Southern Ocean State Estimate (SOSE; Mazloff *et al.* 2010) numerical model re-analysis, estimated to be 20 and 40 Sv, respectively.

For an assessment of Intpp there are sparse (in space and time) *in situ* measurements of primary production in the Ross Sea (see e.g. Pinkerton & Bradford-Grieve 2010 and references therein), but these have typically been used to derive products for global estimates of Intpp. One such product is the so-called vertically generalized production model (VGPM; Behrenfeld & Falkowski 1997), and output from the VGPM has been used here to construct monthly means and standard deviations over region ROSS. Following Pinkerton & Bradford-Grieve (2010) we adjust the VGPM values by calibration of area averaged surface chlorophyll values, and independent estimates of annual production rates. These adjustments account for the different areas of the Ross Sea over which the estimates are made. In particular Pinkerton & Bradford-Grieve (2010) use the annual production rate of $140\,g\,C\,m^{-2}\,y^{-1}$ from Arrigo & van Dijken (2004) for the latter's study area (region AVD, which is largely over the continental shelf). Over the growing season, the satellite data suggests that the ratio of average chlorophyll concentration between regions ROSS and AVD is 0.47, thus we estimate a calibrated annual production over region ROSS to be $66\,g\,C\,m^{-2}\,y^{-1}$. Over region ROSS, VGPM predicts the annual production to be $\sim 55\,g\,C\,m^{-2}\,y^{-1}$, hence we scale the VGPM values by 1.2 (VGPM*).

Throughout this paper the spatial grids associated with the reference datasets are used as templates for the assessment. Model output is interpolated to the spatial grid associated with the data, and then the relevant statistics are computed. This interpolation can lead to some misrepresentation of the true model coastline, especially where there are relatively large geometrical and spatial differences between the data and model grids. For example, the native horizontal resolution of WOA 2009 is 1° on a regular longitude-latitude grid, while the model grids can have local resolutions smaller or larger, and can be on less standard projections (such as rotated poles etc.). Also, the satellite observations of surface chlorophyll over the Ross Sea contain missing data due to a mixture of low solar elevation, cloud and sea ice (as is apparent in Fig. 2). Any of the computed statistics will be based on the target grid and missing data distribution of the respective reference dataset.

Our statistics are based on area averages, such that the area average A_{ave} of function A_i, where i labels the grid points within the area, is:

$$A_{ave} = \sum A_i dS_i \,/\, \sum dS_i, \qquad (1)$$

where dS_i is the grid cell area, and the summation is over all grid cells with non-missing data. The results will be presented in terms of area averages of the observed and model values of interest (e.g. temperature, salinity, etc.), the standard deviation of the values of interest (if available), a root-mean-square error (RMSE) measure of the difference between the observed and model values, and a bias (BIAS) taken as (model– observations).
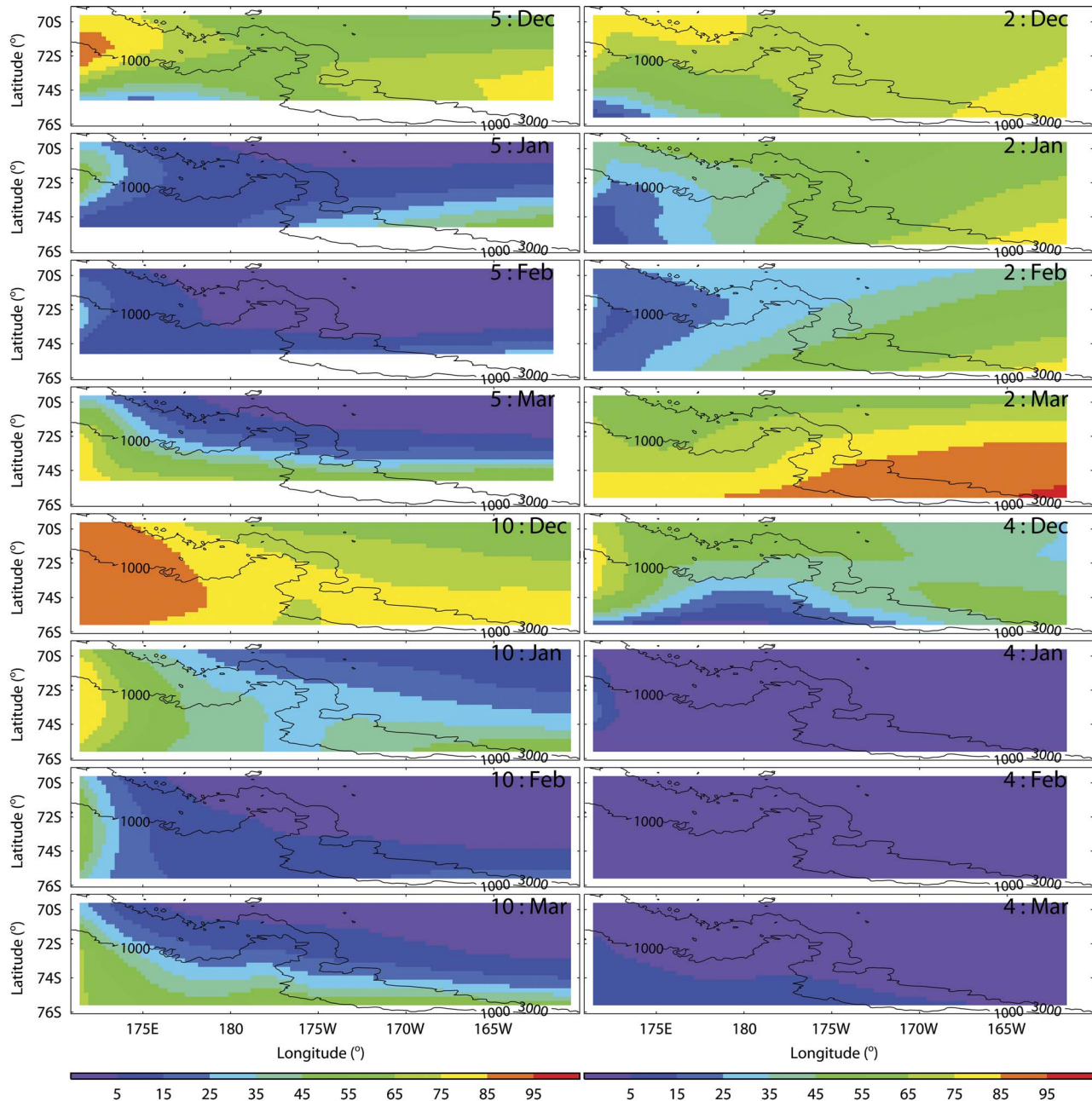
Fig. 6. Spatial pattern of model monthly mean sea ice concentration for December to March for comparison with HadISST for models 5, 2, 10 and 4.

Present day comparisons

From the satellite data, monthly mean area average surface chlorophyll values have been calculated for region ROSS and are plotted on Fig. 3a & b as black circles. Note that the month ordering has been rotated to put the Ross Sea summer months at the middle of the axis. The monthly mean values over the historical period from each of the models are also plotted as the coloured lines. Models 3, 14 and 15 have been plotted in red (model 3

distinguished by a peak shifted to March), the remainder in blue to green. An anomalously low solution associated with model 1 is the dark blue curve. In Fig. 3a the scale accommodates the excess amplitudes predicted by models 14 and 15. The scale is reduced to focus on most of the model solutions in Fig. 3b. Most of the models predict a peak in surface chlorophyll between November and January, with evidence of chlorophyll into April and May beyond that suggested by the light-limited satellite data.
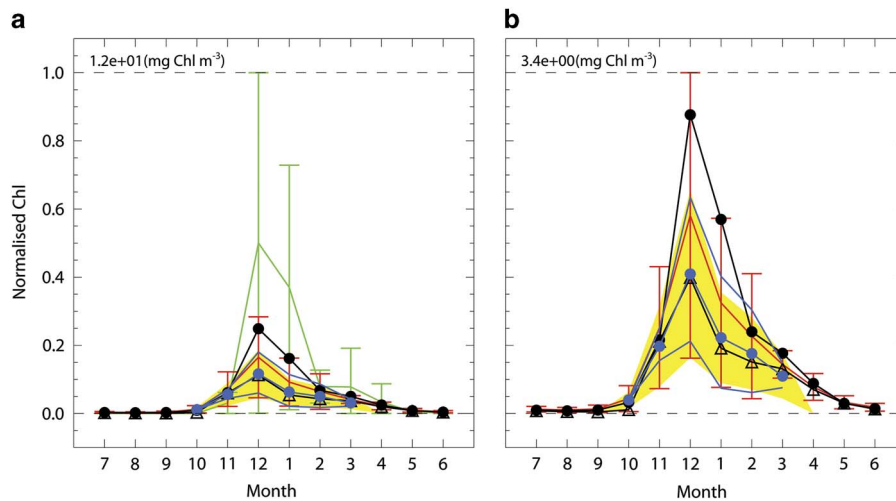
Fig. 7. Normalized satellite (blue-yellow), inner ensemble (red), outer ensemble (green), and Mn (black circles) and Md (black triangles) monthly mean area averaged surface chlorophyll (mg Chl m⁻³) for region ROSS for **a.** all models and **b.** excluding outer ensemble models 3, 14 and 15, and anomalously low model 1. Normalization factor is shown at the top left. The yellow band shows satellite variance about the mean using 0.4 and 1.6 estimates, and the blue lines show satellite variance based on satellite spatial standard deviations per month. Coloured bands for the model ensembles indicate ensemble spread (see text) about ensemble mean.

Figure 3c plots the RMSE and BIAS per model (numbered 1–16) averaged over November to March compared to the satellite data. Our assessment of (relative) model skill will consider spatial (over region ROSS) and temporal (over a season) components combined, and so estimation of the time dependence of the spatial errors in terms of RMSE and BIAS represents a potential set of metrics (e.g. see the discussion in Jolliff *et al.* (2009) about alternatives). In this instance, despite the poor solution of model 1, the RMSE and BIAS metrics are not the worst; however, the actual solution of model 1 is the worst as the average value (the red triangle) is well below the data average (the horizontal dashed line). Models 14 and 15 are clear outliers, with excess RMSE and BIAS. Model 3 has a relatively low overall BIAS but not so favourable RMSE, despite its misplaced peak.

Using the RMSE and BIAS from Fig. 3c, we can rank all models for surface chlorophyll. For example, the RMSE value is used to rank the models from 1 to 16, and similarly for the BIAS; the 'score' for each model is then simply taken to be the average of its RMSE and BIAS positions, and the lower the overall score the better ranked the model is taken to be. The result of this process is shown in the top frame (labelled 'Chl') in Fig. 4. As expected, models 14 and 15 rank the worst, while the relatively anomalous models 1 and 3 rank tenth and ninth, respectively. The next six frames repeat the exercise for SST, the surface nutrients nitrate (NO_3), phosphate (PO_4) and silicate (Si), an estimate of a mixed-layer depth (or depth of the seasonal thermocline, DST), and SIC. A mixed-layer depth is defined to be the depth at which the potential density changes by $0.125\,kg\,m^{-3}$ from its value at the surface (the potential density reconstructed from the monthly mean model values of potential temperature and salinity) and shares properties with the 'mixed-layer depth', although given the monthly mean model output timescales is perhaps better associated with a seasonal scale

than the relatively rapid dynamics of a 'mixed-layer depth' (hence our reference to it as DST). As an indicator of overall performance using the chosen metrics, the bottom frame ('All') combines the previous seven rankings by simply summing and averaging each model's ranking per variable, noting that this mixes physical with biogeochemical variables and models do not all sample the same number of parameters (see the comparably short PO_4 list, for example). This combined ranking shows that models 14 and 15 rank the worst. Therefore, we have defined an initial set of 'worst fit' models to form an 'outer ensemble' which are models 14 and 15, to which models 1 and 3 are added due to their poor surface chlorophyll representations. The remaining models define an 'inner ensemble' of the better ranked models.

Examination of the SIC ranking in Fig. 4 reveals that the models split into four approximate groupings. The first group (for the best models) comprises models 5, 8, 1 and 7; these models capture the annual seasonal cycle well, and of these, model 5 captures the February minimum amplitude the best. The annual cycle of area averaged SIC over region ROSS for model 5 is shown in Fig. 5a as the black curve, with the vertical lines indicating the model standard deviation. The HadISST data is shown as the blue circles, with the yellow band the data standard deviation. Apart from slight underestimates in March and April, model 5 matches the data well. Models 2 and 3 form the next group, with good timing of the annual cycle compared to the data (see Fig. 5b for model 2), but neither of these models capture the summer minimum values. Models 10, 9, 11 and 6 form the next group, each typically capturing the February minimum well, but tending to grow too slowly over the autumn period (see Fig. 5c for model 10). The final (worst) group comprises models 4, 12, 15, 14 and 13; these models have too little sea ice over the summer period, and also underestimate the autumn regrowth, typically not matching the data until around July (see Fig. 5d for model 4).
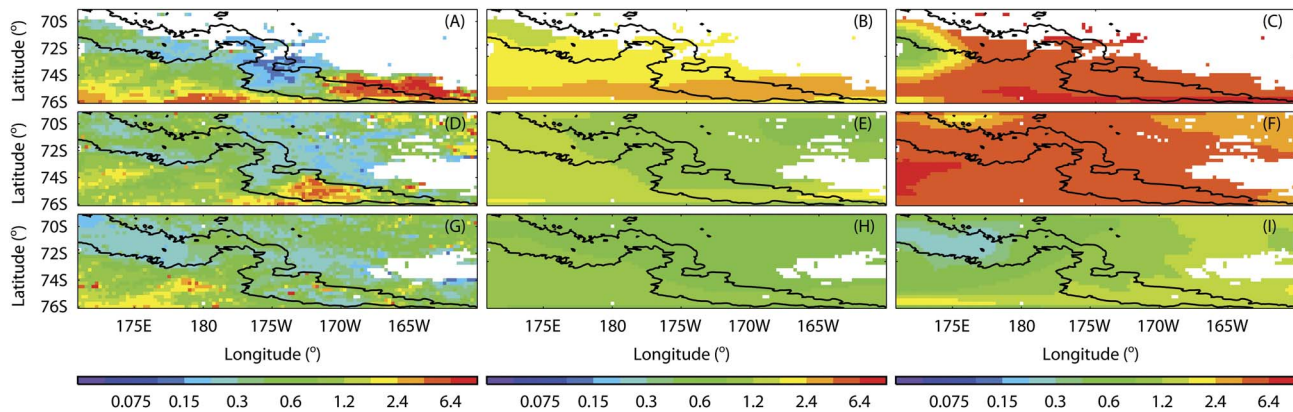
Fig. 8. Reconstructed ensembles from Fig. 7 for December, January and February (rows top to bottom). **a.**, **d.** & **g.** Satellite mean, **b.**, **e.** & **h.** inner ensemble and **c.**, **f.** & **i.** outer ensemble for surface chlorophyll (mg Chl m^{-3}). Note that the colour scale is extended to 6.4 mg Chl m^{-3} compared to the scale in Fig. 2.

The monthly mean spatial patterns of SIC for models 5, 2, 10 and 4 are shown in Fig. 6 for comparison with the HadISST data in Fig. 2. In December, model 5 shows a relatively weak Ross Sea polynya emergence in about the right location. In January the opening continues, but tends to spread too far to the east compared to the data, resulting in the relative underestimate in February. The regrowth arises in March with development along the southern part of the domain. For model 2, the Ross Sea polynya emergence and subsequent development matches well the data spatial patterns (in particular the east to west gradient over region ROSS), but as expected the area averaged model amplitude tends to be too high. For model 10, the Ross Sea polynya over region ROSS is not apparent, with the seasonal progression suggesting sea ice thinning at the northern part of the domain. In February and March, model 10 resembles model 5. Finally for model 4, its Ross Sea polynya compares well with the data in December (but perhaps a little underestimated to the east), but clearly has too little ice over region ROSS in January and February, and only a very weak recovery in March.

Note that the rankings for SIC in Fig. 4 reflect both the model temporal and spatial patterns (in Figs 5 & 6, respectively). Consistent with our 'outer ensemble' models, 14 and 15 fall into the last grouping for SIC, whereas models 1 and 3 rank relatively well.

The reconstructed surface chlorophyll monthly means from the inner (red) and outer (green) ensembles are shown in Fig. 7. Figure 7a includes the outer ensemble solution, while Fig. 7b does not. For each ensemble the spread about the mean is shown, where the spread is taken to be the average root-mean-square distance between the ensemble members and the ensemble mean. The blue curve with circles shows the mean satellite data, with the yellow band the satellite variance about the mean based on the 0.4 and 1.6 estimates. For comparison, the blue lines are area averaged values obtained from subtracting or adding the monthly mean standard deviation patterns from the SeaWiFS data over the period 1997–2005 to the mean, giving surface chlorophyll lower and upper bound estimates, respectively. The yellow band spans the blue lines, suggesting a consistent measure of the surface chlorophyll interannual variability. In Fig. 7a, the impact of models 14 and 15 to the outer ensemble peak and the shift due to model 3 are clear. In Fig. 7b, it is apparent that the inner ensemble solution matches the data well in terms of the timing with the model spread spanning the data variance, but the ensemble mean is typically in excess of the data mean, at least based on these area average metrics.

Also shown in Fig. 7 are the multi-model (using all 16 models) mean ('Mn', black circles) and median ('Md', black triangles) values. For surface chlorophyll, Mn is typically in excess of the data mean, especially in December and January, whereas Md tracks the data mean very well except for a slight underestimate in October. The respective RMSE and BIAS scores for the inner, outer, and Mn and Md are included in Fig. 3 for comparison with the individual models. The outer and Mn metrics are in excess of the inner and Md equivalents, and as expected Md represents the mean the best (the red triangle), whereas the inner ensemble produces the lowest BIAS score and a relatively lower RMSE value.

The low sample size of 16 models, coupled with the uncertainty regarding the statistical distribution of the model solutions, makes choosing the relevant parametric (or loosely 'mean-based') or non-parametric (or loosely 'median-based') statistical sampling problematic. The assessment and ranking process for the inner and outer ensembles is both for estimation of the 'best' solution and for trying to pick apart why we might need to exclude the outer models from such estimation. For surface chlorophyll the inner solution provides for a working 'best' solution (based on our metrics) compared to Mn and Md, and we shall use the inner solution to base
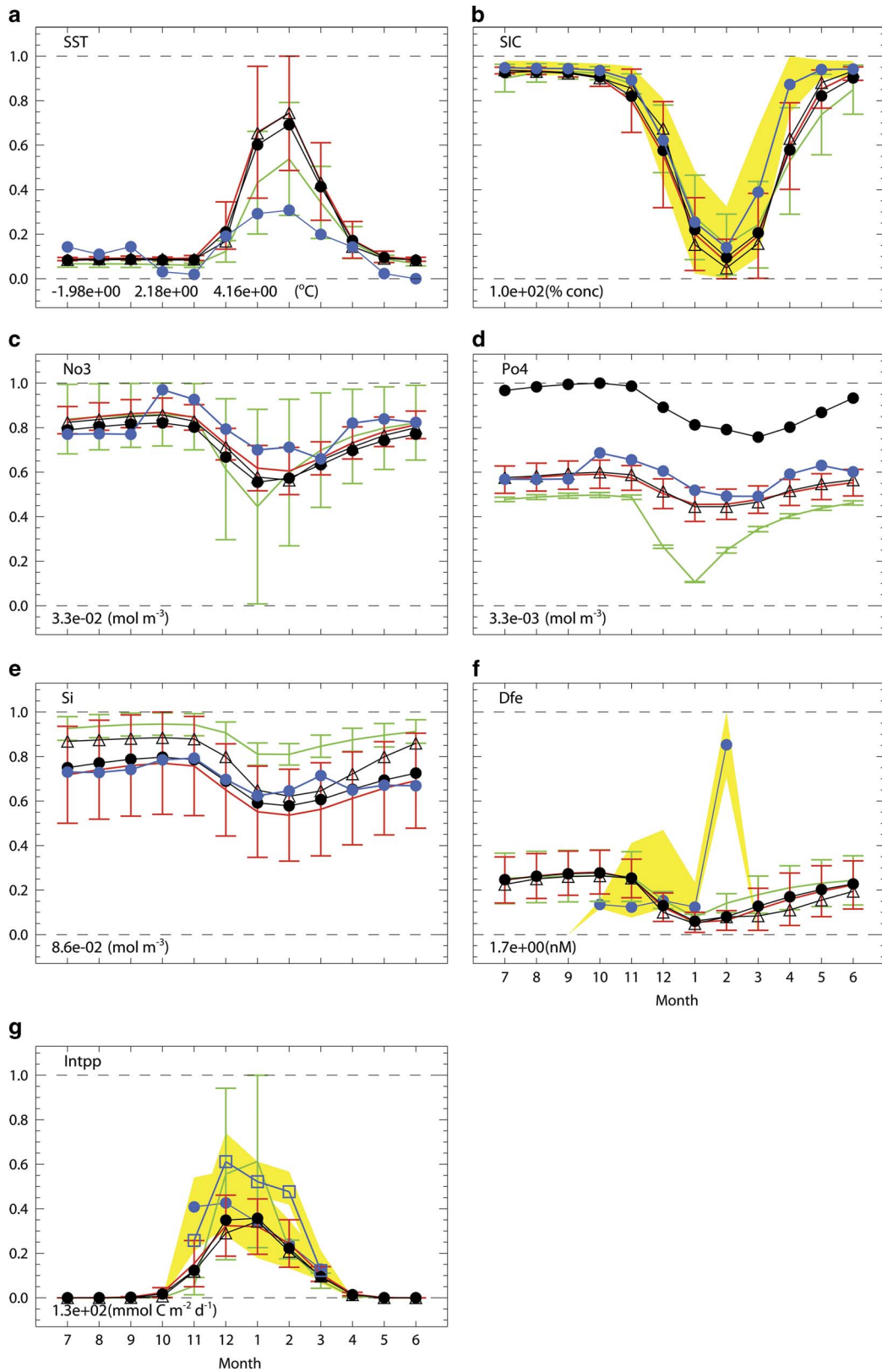
assessments of the other variables. However, for reference we shall also provide the Mn and Md values for each variable in the historical data comparisons and also for consideration of model future projections.

Examples of monthly mean surface chlorophyll patterns are presented in Fig. 8 for December to February (from top to bottom). The left column is the satellite data from Fig. 2, the middle column is the inner ensemble, and the right the outer ensemble (noting that the colour scale has been extended to 6.4 mg Chl m^{-3} to accommodate the larger outer ensemble amplitudes). For each month the missing data (white) are from the satellite record. For December the inner ensemble (Fig. 8b) captures the elevated band of surface chlorophyll around 75°S in the data (Fig. 8a), with the band extending from the Ross Sea continental shelf, across the shelf break and into the deeper water (the black lines showing the 1000 and 3000 m isobaths). The inner ensemble also captures the northward reduction seen in the data across the continental shelf, with the lower amplitudes arising on the shelf break around 173°E, 71°S. The mean is typically in excess of the data mean but within the variance (see Fig. 7) comprising underestimates along 75°S compared to the data, and overestimates elsewhere, but the spatial pattern is consistent with the data. In comparison, the outer ensemble (Fig. 8c) shows the enhanced chlorophyll signature across most of the shelf, which is even more extensive by January (Fig. 8f) compared to the data (Fig. 8d) and the inner ensemble (Fig. 8e). As before, the inner ensemble tends to underestimate the chlorophyll signature to the south, and overestimate elsewhere, but overall the spatial pattern is consistent with the data and results in the reasonable fit in Fig. 7. By February all three frames (Fig. 8g–i) show a similar average surface chlorophyll amplitude across the entire domain (both on and off the shelf). In fact it might be argued that the outer ensemble (Fig. 8i) is capturing some of the higher chlorophyll signatures in the data across the shelf slope for February better than the inner ensemble, but the metric in Fig. 7 cannot readily distinguish the solutions.

Figure 8 provides evidence that the inner models are able to represent the seasonal and spatial surface chlorophyll variations and patterns seen in the data. This suggests that our discrimination on the basis of the metrics used here is enabling a satisfactory fit from the available models. The data contains fine scale features not captured by the models (for example, Fig. 8g shows a number of off-shelf high intensity chlorophyll regions not seen in either of the ensembles), and this is perhaps not surprising given the relatively coarse spatial resolution of the models. Nevertheless, the evidence suggests the inner models are representing some of the important broader scale signatures seen in the data.

Figure 9 compares monthly means, medians and variances for SST (Fig. 9a), SIC (Fig. 9b), the surface macronutrients nitrate (Fig. 9c), phosphate (Fig. 9d), silicate (Fig. 9e) and iron (Fig. 9f), and the integrated (or net) primary production (Fig. 9g). The reference area averages are plotted as blue circles, with yellow bands for the temporal variance where available. As before the inner and outer ensemble means are red and green, respectively, and Mn and Md are the black circles and triangles, respectively. Each graph is normalized to the value at the bottom left of each frame. For SST, the scaling factors are the minimum, maximum and range.

Figure 9a for SST shows that all the ensembles overlap fairly closely over most of the year, except from January to March where the outer ensemble (green) suggests relatively cooler SSTs. Compared to the SST data, the model solutions represent the seasonal cycle well, but overestimate the summer SST peak by ~ 1.5°C (a signature perhaps of the tendency toward a warm bias in the Southern Ocean in CMIP5 coupled models as highlighted in, for example, Wang *et al.* (2014)). The lower bounds of the outer ensembles do a better job of approximating the summer SST data, and this is reflected in the good ranking scores for models 1 and 3 in Fig. 4. All ensembles tend not to capture some of the coldest average temperatures seen in the data.

For SIC in Fig. 9b all the ensembles capture the timing of the summer minimum in February, but have the tendency to underestimate the regrowth from March to June seen in the data. The SIC rankings in Fig. 4 show that the outer ensemble models are distributed across the positions and when combined with the different responses illustrated in Fig. 5 result in the range of solutions being blended together.

In Fig. 9c for surface nitrate the anomalously low model 3 solution has been omitted from the outer ensemble. The mean values of each ensemble follow the seasonal cycle and amplitude of the data relatively well, capturing the dip in the December to March period, and returning to similar winter values. However, the spread in the outer ensemble solution is clearly much broader than that of the inner model, perhaps reflecting the

---

Fig. 9. Normalized area averaged monthly mean plots over region ROSS for **a.** sea surface temperature (SST), **b.** sea ice concentration (SIC), **c.** nitrate (NO_3), **d.** phosphate (PO_4), **e.** silicate (Si), **f.** dissolved iron (DFe) and **g.** integrated primary production (Intpp) for the historical period. Data are indicated by the blue circles or squares, and the estimated data variance is represented by the yellow band (where available). Inner, outer, Mn and Md ensemble values are shown by red, green, black circles and black triangles, respectively. Normalization factors are shown at the bottom left. For SST the factors are minimum, maximum and range, respectively. Outlier model 3 was removed from the NO_3 and PO_4 outer ensembles, and similarly model 1 for Intpp.
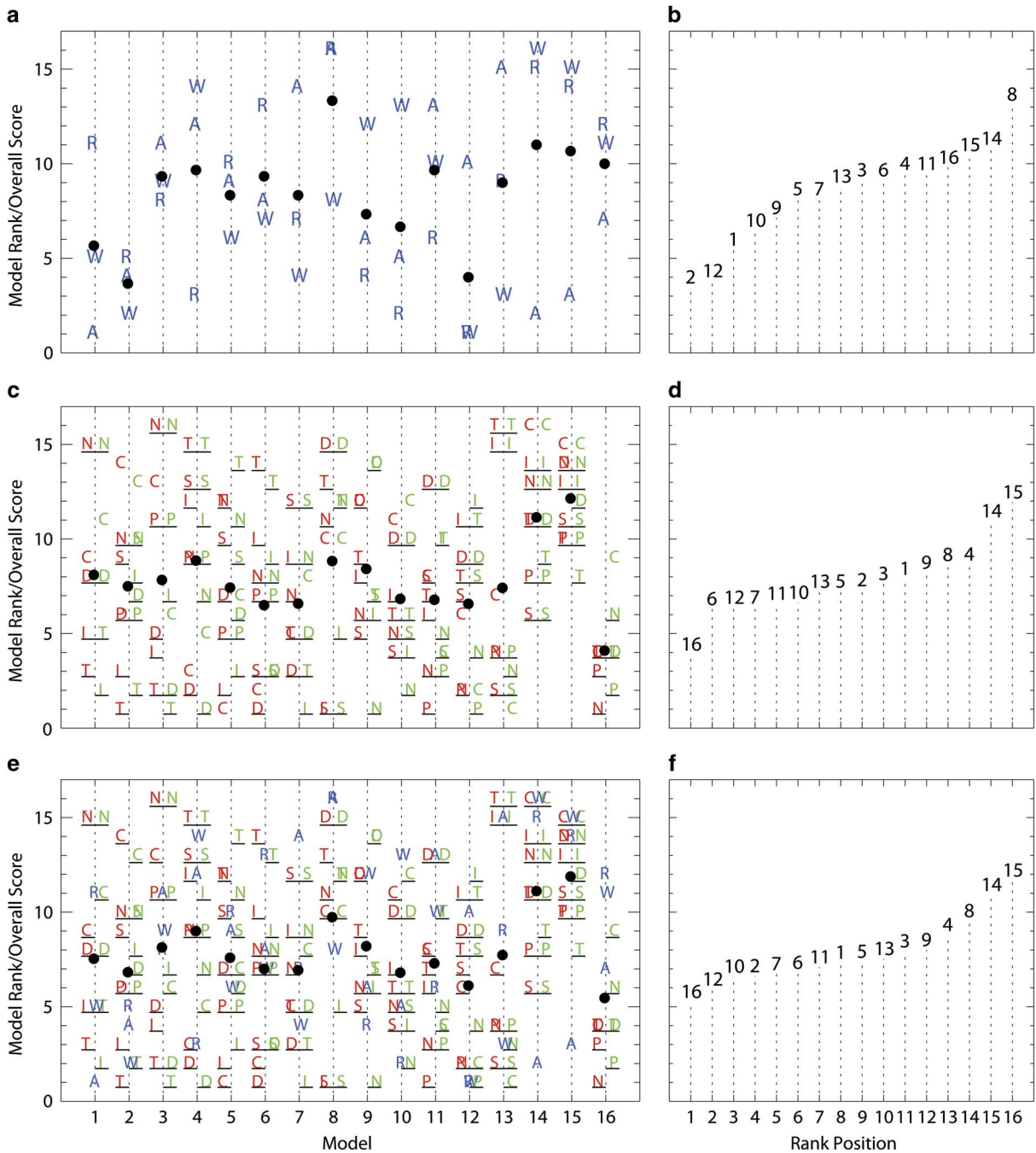
Fig. 10a., c. & e. Variable rankings per model and overall score (black circles), with respective ordered overall ranking in **b.**, **d.** & **f.** using 1976–2005 averages. Model number is plotted on **b.**, **d.** & **f.** Symbols A, R and W show Drake Passage transport, peak Ross Sea gyre transport and peak Weddell Sea gyre transport rankings, respectively. Symbols C, T, N, P, S, D and I label surface chlorophyll, sea surface temperature, nitrate, phosphate and silicate, depth of the seasonal thermocline, and sea ice concentration, respectively, for root-mean-square error (RMSE) (red) and bias (BIAS) (green) rankings.

comparative spreads in the surface chlorophyll solutions in Fig. 7a, and that the outer ensemble models are ranked the worst for nitrate (see Fig. 4). Further, the inner ensemble spread is relatively narrow, and its mean matches Mn and Md across the seasonal cycle.

For surface phosphate in Fig. 9d the anomalously large model 3 solution has been excluded from the outer ensemble, and also note (see Table I) that model 1 does not include phosphate in its biogeochemical model. The result is a distinct separation between the inner and outer

ensembles, with the outer ensemble amplitude lower than the inner ensemble, especially over the summer months. Note that the relatively narrow outer ensemble band presumably results from the outer ensemble models 14 and 15 being from the same modelling centre. Compared to surface nitrate, the phosphate solution for the inner and Md ensembles overlap each other and the data fairly well over the seasonal cycle, but Mn is too large (reflecting the influence of anomalously large model 3 phosphate).

Figure 9e is for surface silicate. Again the inner and outer ensembles are distinct for most of the year, but this time the inner ensemble lies below the outer band, and there is slightly more of an overlap between August to November. As before for surface nitrate, the inner, Md and Mn ensembles match the datapoints remarkably well across the seasonal cycle. Note that the outer ensemble is again only composed of models 14 and 15, as models 1 and 3 do not cycle silicate explicitly in their biogeochemical models.

Note that Smith *et al*. (2014a) observe that the surface macronutrients nitrate, phosphate and silicate are generally not depleted during the phytoplankton growing season, and the inner model solutions are consistent with this picture with proportionately small drops in their values over November to March compared to the rest of the year (see Fig. 9c–e).

The models also predict other less well observed variables of interest biogeochemically. For example, Fig. 9f shows a reconstruction of the monthly mean area average surface iron concentration. Unlike the other nutrients, all the iron ensembles strongly overlap, each ensemble having a seasonal maximum around October and a minimum in January. In the absence of an iron climatology for the oceans, comparison can be made with the *in situ* data compilation of Tagliabue *et al*. (2012), in particular their fig. 4c for the upper 100 m iron seasonal cycle for their Ross Sea region (a broad zone spanning 155°W to 180°W and south of 45°S) based on 240 observations over 1990–2008. The median values from fig. 4c are plotted here as the blue circles, and the first and third quartiles provide the yellow band about each value. Figure 9f shows that the models have a summer minimum in January of $\sim 2 \times 10^{-1}$ nM, and rising over the year to mean peak values of $\sim 7 \times 10^{-1}$ nM in September to October. Apart from February, the data compilation and model amplitudes match reasonably well, even though the respective regions do not overlap too well. However, the models are insistent on a drop in values over the summer, at variance with the observations. Tagliabue *et al*. (2012) suggest that their elevated February values might be associated with sampling near fast ice melt as opposed to indicating a seasonal trend; they also note that the 'low throughout the period October to January' might well signal utilization of winter reserves, which is consistent with the model realization. Also, the model iron is only

for the model surface layer (typically ~ 10 m) compared to the upper 100 m values from Tagliabue *et al*. (2012).

For Intpp (mmol C m^{-2} d^{-1}) in Fig. 9g the anomalously low model 1 has been omitted from the outer ensemble. The blue circles with the yellow variance are the averages from VGPM*, and the blue squares with yellow variance are the values from Pinkerton & Bradford-Grieve (2010) for their larger domain that also includes the relatively less productive northern waters. In January, the outer ensemble solution (green) has the highest production rate, consistent with the elevated surface chlorophyll in Fig. 7a. Compared to VGPM* all ensemble solutions fall below the November data estimates, but for December to March the inner, Mn and Md solutions compare well. For comparison, the inner ensemble has an average annual production rate for region ROSS of 60.7 g C m^{-2} y^{-1} underestimating the 66 and 85 g C m^{-2} y^{-1} from VGPM* and Pinkerton & Bradford-Grieve (2010), respectively.

If the conditions determining the wider circulation in the ROSS region are relevant to the biogeochemical outcomes (and presumably they are) then Fig. 10a & b rank the models based on major circulation characteristics in the Southern Ocean, *viz* the transport in the Antarctic Circumpolar Current (ACC), and the peak transports in the Ross and Weddell Sea sub-polar gyres, with the individual ranked transports plotted as the blue symbols A, R and W, respectively. Best observational estimates of the Drake Passage (ACC) transport, and the peak Ross and Weddell gyre transports are centred around (149, 20, 40) Sv, respectively. Figure 10a shows the ranking of each model per transport measure, and the black circle indicates an overall score based on the radial distance from the observed values. Figure 10b takes the overall scores and now plots the individual models in rank order. The higher the location of a symbol or the black circle, the lower the confidence in the model estimate relative to the other models in the sample. In this case, Fig. 10b shows that model 2 provides the relatively best combined estimates of the transport values, and model 8 relatively the worst; the A/R/W rankings are clustered around low ranking values for model 2, whereas model 8 has a moderate ranking for W, but equal worst values for A and R. Model 1 from the outer ensemble actually ranks well, whereas models 14 and 15 are at the lower end.

Figure 10c shows the rankings based on the RMSE (in red) and BIAS (in green) scores for the surface chlorophyll (C), temperature (T), nitrate (N), phosphate (P), silicate (S), DST (D) and SIC (I). It is these individual rankings per variable that were used to obtain Fig. 4. The RMSE ranking per variable is plotted to the left of the vertical dotted line locating each individual model, the BIAS ranking to the right, with the black circle on the line the overall model score as before. The overall score is based on taking equal contributions from each variable. Based on this combination of metrics, model 16 ranks
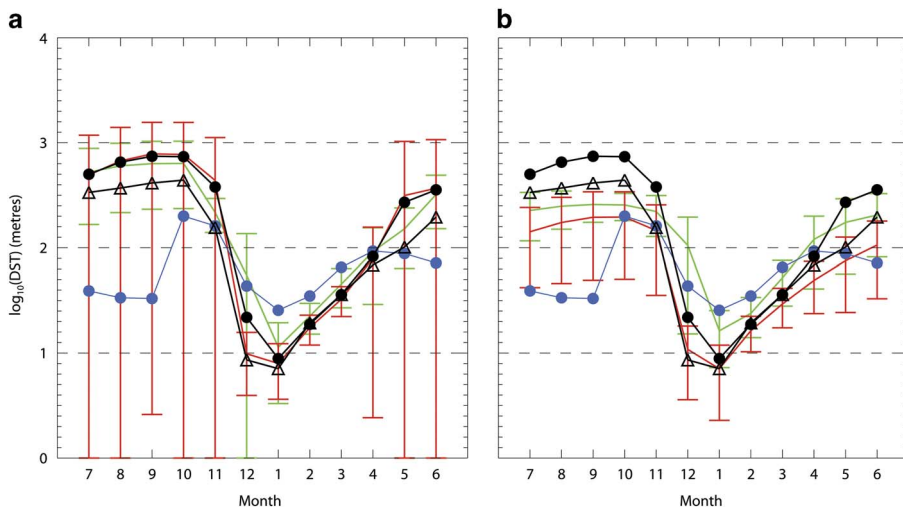
Fig. 11. Ensemble reconstruction for inner (red), outer (green), and Mn (black circles) and Md (black triangles) of monthly mean area averaged depth of the seasonal thermocline (DST) in metres (plotted logarithmically) for region ROSS. Data values are shown as blue circles. **a.** Includes all models, while **b.** excludes models 14 and 15 from the outer ensemble, and models 8, 9, 10, 11 and 12 from the inner ensemble.

well for nitrate, phosphate and DST, relatively poorly for the chlorophyll bias (in green), but overall ranks the best (although note that this model does not explicitly cycle silicate or iron biogeochemically). In contrast, model 1, that only cycles nitrate, scores poorly for the biogeochemical components chlorophyll and nitrate, and thus ranks as one of the weaker overall models. Models 14 and 15 rank the lowest in Fig. 10d, and model 3 is in the outer ensemble as a result of its poor seasonal solution for surface chlorophyll (see Fig. 3), even though its ranking scores for SST and DST are relatively good.

Comparing Fig. 10b and 10d a degree of overlap between the relative ranking of each model is apparent, e.g. models 8, 14 and 15 rank poorly in both and suggest a potential link between the wider dynamics and the processes in the Ross Sea. However, this does not explain all the rankings, e.g. models 1 and 16 swap their relative rankings between these frames. The transport estimates in Fig. 10a can be combined with the variables in Fig. 10c to produce the ranking shown in Fig. 10e. As before, each variable is assumed to contribute equally to the overall score. This combined ranking confirms that indeed the biogeochemical outcomes of models 14 and 15 are probably influenced by their relatively poor representation of the large scale circulation, and as Fig. 10a shows it is not their ACC transport that scales poorly but rather the individual strengths of each of the sub-polar gyres relative to the other models. On the other hand, the dynamical solution of model 1 is relatively good; therefore, perhaps in this case it is the biogeochemical model that leads to relatively poor outcomes for region ROSS. Comparing Fig. 10b and 10d it is apparent that the model ranking based on the transport estimates alone (Fig. 10b) separates the model ranking scores more widely than the relatively mixed physical and biogeochemical rankings (Fig. 10d). In

terms of the combined estimate in Fig. 10f, it is perhaps not surprising that models 8, 14 and 15 are the least well ranked models. Models 1 and 3 end up with combined rankings that place them toward the weaker end of the ranking. Note that the equal weighting per variable means that the wider spread in Fig. 10b is flattened somewhat by the relatively more even spread of Fig. 10d.

Based on Fig. 10 and the preceding analyses, the separation of the CMIP5 models into inner and outer ensembles seems to have a degree of consistency. In developing potential future physical and biogeochemical downscaling for region ROSS, we would consider excluding the outer models 3, 14 and 15, and model 1 due to its anomalously low surface chlorophyll, and use the remaining models comprising the inner ensemble. The metrics of RMSE and BIAS applied to area averaged satellite chlorophyll estimates and surface nutrients from WOA 2009 show that the inner ensemble is able to capture the seasonal cycle rather well in both phase and amplitude. This consistency also applies to the Intpp values compared to the derived primary production estimates that capture the phase well but perhaps underestimate the total production. Surface iron and SIC both show a degree of overlap between their respective ensembles. Such an assessment is useful as it allows a measure of confidence to be applied to the model results, especially when considering model reconstructions of relatively unobserved parameters (such as Intpp and surface iron).

Another dynamical issue associated with CMIP5 model representation of the Antarctic gyres has been recently documented by Heuzé *et al.* (2013). Bottom water formation around Antarctica for the present day is associated with dense water creation on the continental shelves that is then exported via shelf overflows to the lowest levels of the Southern Ocean. The analysis of Heuzé
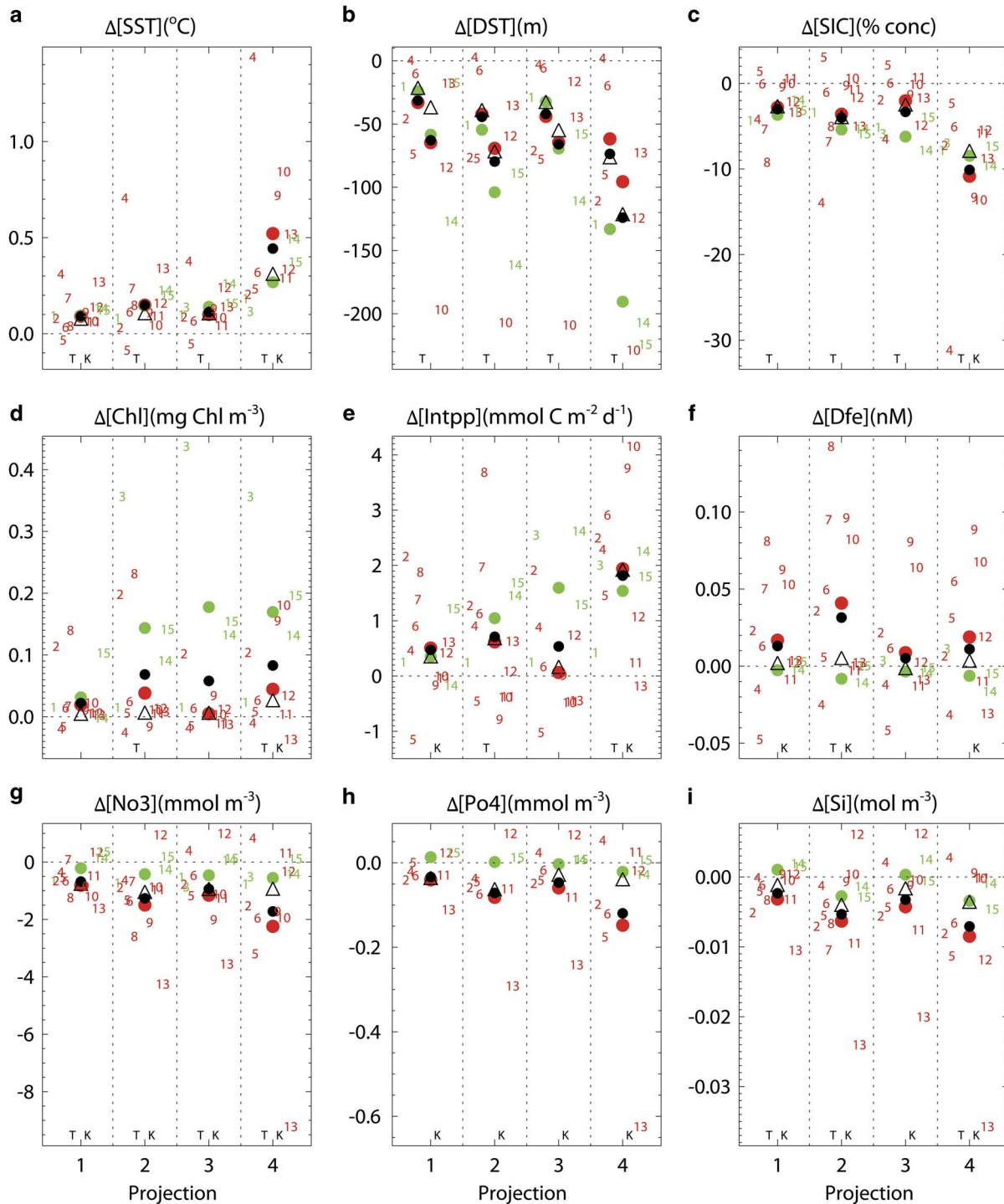
Fig. 12. Model projections 1, 2, 3 and 4 (see text) for region ROSS for changes in **a.** sea surface temperature (SST), **b.** depth of the seasonal thermocline (DST), **c.** sea ice concentration (SIC), **d.** surface chlorophyll, **e.** integrated primary production (Intpp), **f.** dissolved iron (DFe), **g.** nitrate (NO_3), **h.** phosphate (PO_4), **i.** silicate (Si) relative to model historical average. Red/green indicates inner/outer models, the symbols represent individual model values and the circles represent ensemble averages. Black circles/triangles show Mn/Md changes, respectively. Symbols left of each DST centreline show reduced model sets (see text). Vertical dashes separate projections. Horizontal dashes locate zero change. T and K show the 95% confidence level significance for Student's t-test and Kolmogorov-Smirnov test, respectively.

et al. (2013) finds that 10 of the 15 CMIP5 models do indeed generate dense water on the shelf but that this water is not exported to depth, i.e. none of the models analysed generate Southern Ocean bottom water in the expected manner. Around half of the models actually generate bottom water via *in situ* convection events rather than the advective pathway occurring in nature. Furthermore, some of these models have relatively extensive convection zones which seem to result in strong sea ice seasonality, so that there appear to be dynamical ramifications of this anomalous process. Any process that exaggerates the seasonality is likely to impact the biogeochemical response; in that context we examine the rankings we have obtained in the light of the Heuzé *et al.* (2013) results.

From figs 1 and 2 in Heuzé *et al.* (2013), the models that overlap with our study are 1, 4, 5, 6, 8, 10, 11 and 14. Using their metrics, models 1, 5, 10, 11 and 14 are found to have an acceptable bottom temperature representation compared to present day climatology, while models 1, 5, 8, 10 and 14 do likewise for bottom density. However, Heuzé *et al.* (2013) note that models 5, 6 and 8 have extensive areas of deep convection, while models 10, 11 and 14 exhibit this convection only within the sub-polar gyres. Model 4 has no apparent deep convection resulting in too light dense water; model 1 behaves similarly, but nevertheless seems to support deep water of about the right properties. It is apparent that our choice of inner and outer ensembles cannot simply be explained on the basis of the dynamical properties detailed by Heuzé *et al.* (2013), mirroring the earlier consideration of the ACC and gyre transports.

As an example of the potential impact, Fig. 11 plots the logarithm of the monthly mean area averaged DST for region ROSS. In comparison to the data (in blue), all ensembles have significantly larger depths from June to November, and underestimates compared to the data from December to March, but each accurately captures the timing of the summer minimum in January. The spread in model solutions is also large, particularly for the inner ensemble (in red) from April to November. The individual models 8, 9, 10, 11 and 12 from the inner ensemble, and models 14 and 15 from the outer ensemble have maximum DSTs in excess of 800 m; if we consider these models as DST outliers and exclude each from their respective ensembles we get the solutions shown in Fig. 11b. Now the model variance in both the inner and outer ensembles is reduced over winter, and their mean values are much closer to the data in October to November and May to June. In the summer, the reduced outer ensemble now more accurately tracks the data compared to the inner, Mn and Md which remain relatively low over this period.

In comparison with the properties discussed by Heuzé *et al.* (2013), our DST outlier models 10, 11 and 14 overlap in terms of their DST perhaps capturing deep convection within the sub-polar gyres. Models 5 and 6 have extensive convection regions, but not over region ROSS, whereas model 8 does, hence a DST outlier. Also, models 1 and 4 are not excluded, consistent with not being DST outliers. On this basis, models 14 and 15 might point to their inclusion in the outer ensemble because of these convective events; however, this is clearly not the complete explanation for model 1. Therefore, a metric based on the Heuzé *et al.* (2013) results does not uniquely determine a best-fit biogeochemical ensemble for region ROSS. The reason may simply be that potentially anomalous DST dynamics in the winter do not have a detrimental impact on the summer biogeochemical growth, and are perhaps associated with the relatively narrow range of DST solutions obtained by all of the models in summer.

Model projections

Given some confidence intervals on the model solutions for the 'present day', it is possible to look at future predictions based on the so-called rcp scenarios associated with the IPCC 5th Assessment Round (AR5). Here rcp4.5 and rcp8.5 are examined for their changes relative to the present day means for averages over the periods 2036–2055 and 2081–2100, such that projections 1 and 2 are for rcp4.5 for 2036–2055 and 2081–2100, and projections 3 and 4 are for rcp8.5 for 2036–2055 and 2081–2100, respectively. Examples of such future predictions are plotted in Fig. 12 for the area average changes over region ROSS relative to the present day. Inner and outer models are shown in red and green, respectively, with each individual model number labelled. The inner and outer ensemble average changes are shown by the red and green circles, respectively, and the black circles and triangles show the respective changes in Mn and Md. Individual model numbers have been spread horizontally to make them more readable, and vertical dotted lines separate the different projections.

Along the bottom of each frame in Fig. 12 there are symbols 'T' and 'K' plotted for each projection. The presence of T shows that the distribution of all of the model changes for that projection lead to a mean that is different from zero at the 95% confidence level based on the two-sided Student's *t*-test. The presence of K shows that > 50% of the individual model interannual 20 year time series over the projection period differs from that model's interannual 30 year time series over the historical period at the 95% confidence level based on the two-sample Kolmogorov-Smirnov (K-S) statistical test. The K-S test is used here to give a measure of how distinct the respective interannual time series are, and 50% is arbitrarily chosen to flag that there are indeed a proportion of the samples contributing to the overall mean that are measurably distinct.

For the SST changes in Fig. 12a it is clear that (apart from model 4 that suggests quite large SST changes, and model 5

which has reductions in SST for projections 1 to 3) the ensemble averages actually predict quite similar outcomes, in particular something like a 0.15°C annual average SST rise by the end of the century for projections 1 to 3, with 0.3 and 0.4°C rises for projection 4 for Md and Mn, respectively. All projections are labelled with T, and 1 and 4 with K, showing a good degree of statistical significance.

For DST in Fig. 12b the mean annual values down the projection centreline use the full ensembles, while those to the left of each centreline are based on reduced ensembles from eliminating models with excessive convection. In all cases there is considerable shallowing of DST. For the full ensemble there are decreases of ~ 50–70 m over projections 1 to 3, with decreases of between 90 and 190 m for projection 4. For the reduced ensembles the changes remain significant but smaller, ranging between 70 and 130 m of shallowing. Note that all projections pass the Student's *t*-test (and are labelled with T). Note that for DST we have not presently reconstructed the full annual time series using the annual average three dimensional values of temperature and salinity, hence cannot yet ascribe a K confidence level.

The SIC changes in Fig. 12c have a similar pattern to DST, with projections 1 to 3 showing annual average reductions of 2–7%, with the reduction rising to 8–11% for the ensemble means for projection 4. Again all projections are labelled with T and projection 4 with K. Note that model 4 has excessive reductions in concentration consistent with the projected changes in SST.

In comparison, Fig. 12d for surface chlorophyll shows a relatively distinct separation between the inner and outer ensemble projections, with the outer ensemble suggesting increases of ~ 0.18 mg Chl m^{-3} by the end of the century compared to much smaller increases from the inner ensemble of ~ 0.04 mg Chl m^{-3} (with some of the inner ensemble models predicting a potential fall). These contrast with the annual average value of ~ 0.31 mg Chl m^{-3} for the present day (see Fig. 7). The overall average suggests increases in surface chlorophyll in the future, most significantly for projection 4.

Figure 12e for changes in Intpp shows another broad spread in model solutions, with both sets of ensemble members generally showing increases of Intpp for each projection, although there are some predicting future decreases. By the end of the century both averages of the ensembles predict annual Intpp increases of ~ 1.7 mmol C m^{-2} d^{-1} compared to the annual average derived for the present day (see Fig. 9g) of ~ 16.0 mmol C m^{-2} d^{-1}. Consistent with the surface chlorophyll changes, Intpp is most significant for projection 4.

The individual model solutions for surface iron in Fig. 12f show a broad spread in predicted changes. As a result, only projection 2 passes the *t*-test to suggest changes statistically different from zero, and in this case the average increase is 0.03 nM compared to the present day model

Table II. Model projection 4 changes per variable presented in terms of the annual mean (the black circle in Fig. 12) and a seasonal mean (November to March). Integrated primary production data values from VGPM* (see text).

Variable	Projection 4 change		Data mean		Percent change	
	Annual	Seasonal	Annual	Seasonal	Annual	Seasonal
SST (°C)	0.4	0.85	-1.42	-1.14	28.0%	75.0%
DST (m)	-70.0	-34.0	74.0	66.0	-95.0%	-52.0%
SIC (% conc.)	-10.0	-10.7	74.3	46.5	-13.5%	-23.0%
Chl (mg Chl m^{-3})	0.08	0.37	0.31	0.75	26.0%	49.3%
Intpp (mmol C m^{-2} d^{-1})	2.0	4.3	16.0	38.4	12.5%	11.2%
DFe (nM)	0.01	0.0024	0.2	0.43	5.0%	0.6%
NO$_3$ (mmol m^{-3})	-2.0	-2.4	26.4	24.8	-7.6%	-9.7%
PO$_4$ (mmol m^{-3})	-0.16	-0.16	1.92	1.83	-8.3%	-8.7%
Si (mmol m^{-3})	-7.0	-7.7	60.7	59.9	-11.5%	-12.9%

DFe: dissolved iron, DST: depth of the seasonal thermocline, Intpp: integrated primary production, SIC: sea ice concentration, SST: sea surface temperature.

average of ~ 0.325 nM. In comparison, the changes in the surface macronutrients nitrate, phosphate and silicate in Fig. 12g, h & i, respectively, show a general decline toward the end of the century, with the falls in the inner ensemble average values larger than those of the outer ensemble (especially for model 13). In particular for projection 4, the surface nitrate, phosphate and silicate fall by ~ 2, 0.1 and 5 mmol m^{-3} compared to present day annual averages of 26.4, 1.8 and 60.2 mmol m^{-3} (from Fig. 9c–e), respectively. Note that phosphate passes the K-S test for each projection, but not the Student's *t*-test, whereas the nitrate and silicate changes are significant over nearly all projections.

The projection 4 changes based on the ensemble mean (the black circles in Fig. 12) are summarized in Table II as the annual means. The seasonal means are based on November to March over the growing season. Also listed are the respective means from the data, and the percentage change over projection 4 compared to the data mean. The DST changes are from the reduced ensembles that eliminate the excessive convection models, since the latter models are generally found to stop convecting over this period (see De Lavergne *et al.* 2014) and provide exaggerated changes relative to present day. The overall picture is of significant area averaged warming over region ROSS, consistent with decreases in sea ice cover and DST (hence increased stratification). The biological response is an area averaged increase in surface chlorophyll and Intpp, accompanied by reductions in surface nitrate, phosphate and silicate, and relatively smaller increases in surface iron.

Summary

Using metrics based on area averages over a region covering the Ross Sea sector of the Southern Ocean

(referred to as region ROSS), a consideration of the biogeochemical outcomes of 16 CMIP5 models has been made compared to present day observations. The consideration suggests that we can split the models into an inner and outer ensemble based on their general representation of the observations (where the inner ensemble models provide for a measurably better fit to the data, in this case with a focus on the surface chlorophyll, compared to the outer ensemble models). Models MPI-ESM-LR (14) and MPI-ESM-MR (15) perform relatively poorly overall compared to all the other models for region ROSS, model CANESM2 (1) is the next poorest in the overall ranking (with an anomalously low amplitude of surface chlorophyll and consequently low Intpp), while model CMCC-CESM (3) generates anomalously large surface phosphate, low nitrate and has a shifted peak in surface chlorophyll. These models 1, 3, 14 and 15 are, therefore, referred to as our outer ensemble.

The selection of these ensembles clearly separates the solutions for the surface chlorophyll and the surface nutrients nitrate, phosphate and silicate; the inner ensembles overlap the observations in amplitude and phase, whereas the outer ensembles tend to be clearly either too high or too low. For surface iron, the ensemble solutions tend to overlap each other with comparable seasonal cycles, whereas for the Intpp the outer ensemble again overestimates in comparison to the derived estimates (perhaps consistent with the model surface chlorophyll solutions). For many of the variables analysed, the mean inner model solution overlaps the multi-model median solution; however, this is not as clear for surface chlorophyll, and times in the seasonal cycle where the solutions diverge for the DST and Intpp.

Comparison with larger scale dynamical properties of the sub-polar gyres suggests that there is no single answer that explains the split into our inner and outer ensembles. Certainly models MPI-ESM-LR (14) and MPI-ESM-MR (15) produce sub-polar gyral transports in excess of many of the other models, which may explain their relatively poor overall performance. However, this is not apparent for models CANESM2 (1) and CMCC-CESM (3), so it may fall to the nature of their specific biogeochemical models rather than purely dynamics. This is also true for the recent analysis by Heuzé et al. (2013) looking at the model generation of bottom water; poor performance can partly explain our ensemble choice, but not completely. Indeed, our analysis of DST in region ROSS shows that anomalous DST models overlap the inner and outer ensembles, so again it would appear no single factor is at play in discriminating the models.

This is also true for the SIC comparisons, with members of the outer ensemble distributed right across the rankings. Recent papers have examined the broader Antarctic sea ice representation in the CMIP5 models, in particular Turner et al. (2013) and Shu et al. (2015). The general conclusion is that the models do a generally poor job of realizing the satellite observations of SIC over the historical period, with nearly all the models showing a decreasing trend in SIC at variance with the observations. Shu et al. (2015) expand the 18 models studied by Turner et al. (2013) to encompass 49 of the CMIP5 models, and come to much the same conclusions regarding SIC, but do find that some of the models are able to capture observed trends in sea ice extent (SIE) and volume (SIV) over the satellite period. The models that Shu et al. (2015) found to perform best from our sample are 3, 8, 12, 13 and 15 for SIE, and 3, 12, 13 and 15 for SIV. A comparison with our SIC regional rankings in Fig. 4 shows models 3 and 8 performing relatively well, but models 12, 13 and 15 not so. On this basis, the dynamical link between the broader Antarctic sea ice representation and the other variables is not immediately apparent. This does not mean, of course, that there is not a link, rather the present analysis and set of metrics is not able to reveal it.

On average the ensembles predict an increase in surface temperature over region ROSS by the end of the century, with smaller relative increases in surface chlorophyll and potentially larger relative increases in Intpp based on rcp4.5 and rcp8.5. For surface chlorophyll and Intpp, the average ensemble solutions do tend to separate, with the outer ensemble projected changes typically larger than those of the inner ensemble. There are also significant reductions in SIC and DST over this period.

However, the model generation of open ocean convection in the Southern Ocean models described by Heuzé et al. (2013) certainly complicates the analysis of predictive skill from the CMIP5 models. As has recently been highlighted by Stössel et al. (2015), models are sensitive to the distribution of freshwater over the Southern Ocean (in particular in enabling model behaviour to transition from convecting to non-convecting), consolidating similar results found in the studies of Martin et al. (2013) and Pierce et al. (1995), for example. Stössel et al. (2015) found that improved ACC transport in their model was also probably linked to changes in the Southern Ocean density and meridional density gradients (again with reference to the work of Martin et al. (2013) and Pierce et al. (1995)). Furthermore, the analysis of future CMIP5 model pathways by De Lavergne et al. (2014) suggests that models described by Heuzé et al. (2013) that are deemed convecting for the historical period will actually transition to non-convecting mode under rcp8.5. Our results, while not attributing any specific cause, certainly show how such competition between modelled Southern Ocean processes is consistent with a wide spectrum of CMIP5 model responses.

An assessment of this sort is generally focused around a particular outcome, in this case the seasonal biogeochemical response. However, the complete

solution is of course 'coupled' via all the dynamical variables that drive the system (and more so for a polar region such as the Ross Sea because of the direct cryospheric influences such as sea ice, ice shelves, etc.). Thus in reality the production of a 'best' ensemble solution will require a broader inclusion of forcing influences. Nevertheless, the assessment here shows that separation into inner and outer biogeochemical ensembles can result in a consistent solution compared to the observations within the biogeochemical context (and some of the dynamical constraints). This discrimination is then useful in assessing both less well observed present day variables (e.g. *in situ* iron and Intpp), and placing confidence levels on future projections. Bopp *et al.* (2013) in their conclusions urge continuation of inter-model comparisons such as these due to the high uncertainties in model projections, as well as signalling caution when attempting downscaling with single CMIP5 model outputs. This study is intended to progress such work, and although we may not be able to explain the processes resulting in our choices and conclusions, the results as we have them are useful in providing guidance on the quality of the regional solutions from the CMIP5 models for the Ross Sea.

Acknowledgements

Author contribution

Graham Rickard: CMIP5 model analysis, visualization of results, writing of manuscript. Erik Behrens: support of CMIP5 model sea ice–ocean analysis, substantial support of manuscript.

References

Adachi, Y., Yukimoto, S., Deushi, M., Obata, A., Nakano, H., Tanaka, T.Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Shindo, E., Tsujino, H., Mizuta, R., Yabu, S., Koshiro, T., Ose, T. & Kitoh, A. 2013. Basic performance of a new earth system model of the Meteorological Research Institute (MRI-ESM1). *Papers in Meteorology and Geophysics*, **64**, 10.2467/mripapers.64.1.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. & Zhu, Z. 2013. Evaluating the land and ocean components of the global carbon cycle in the CMIP5 earth system models. *Journal of Climate*, **26**, 6801–6843.

Antonov, J.I., Seidov, D., Boyer, T.P., Locarnini, R.A., Mishonov, A.V., Garcia, H.E., Baranova, O.K., Zweng, M.M. & Johnson, D.R. 2010. *World Ocean Atlas 2009. Volume 2: Salinity. In* Levitus, S., *ed. NOAA Atlas NESDIS 69*. Washington, DC: US Government Printing Office, 184 pp.

Arrigo, K.R. & van Dijken, G.L. 2004. Annual changes in sea-ice, chlorophyll *a*, and primary production in the Ross Sea, Antarctica. *Deep-Sea Research II - Topical Studies in Oceanography*, **51**, 117–138.

Arrigo, K.R., van Dijken, G.L. & Bushinsky, S. 2008. Primary production in the Southern Ocean, 1997–2006. *Journal of Geophysical Research - Oceans*, **113**, 10.1029/2007JC004551.

Aumont, O. & Bopp, L. 2006. Globalizing results from ocean *in situ* iron fertilization studies. *Global Biogeochemical Cycles*, **20**, 10.1029/2005GB002591.

Behrenfeld, M.J. & Falkowski, P.G. 1997. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography*, **42**, 1–20.

Bopp, L., Resplandy, L., Orr, J.C., Doney, S.C., Dunne, J.P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Seferian, R., Tjiputra, J. & Vichi, M. 2013. Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences*, **10**, 10.5194/bg-10-6225-2013.

Bowden, D.A., Schiaparelli, S., Clark, M.R. & Rickard, G.J. 2011. A lost world? Archaic crinoid-dominated assemblages on an Antarctic seamount. *Deep-Sea Research II - Topical Studies in Oceanography*, **58**, 119–127.

Boyd, P.W., Lennartz, S.T., Glover, D.M. & Doney, S.C. 2015. Biological ramifications of climate-change-mediated oceanic multi-stressors. *Nature Climate Change*, **5**, 71–79.

Brandt, A., Griffiths, H., Gutt, J., Linse, K., Schiaparelli, S., Ballerini, T., Danis, B. & Pfannkuche, O. 2014. Challenges of deep-sea biodiversity assessments in the Southern Ocean. *Advances in Polar Science*, **25**, 204–212.

Cabré, A., Marinov, I. & Leung, S. 2014. Consistent global responses of marine ecosystems to future climate change across the IPCC AR5 earth system models. *Climate Dynamics*, **45**, 10.1007/s00382-014-2374-3.

Christian, J.R., Arora, V.K., Boer, G.J., Curry, C.L., Zahariev, K., Denman, K.L., Flato, G.M., Lee, W.G., Merryfield, W.J., Roulet, N.T. & Scinocca, J.F. 2010. The global carbon cycle in the Canadian earth system model (CanESM1): preindustrial control simulation. *Journal of Geophysical Research - Biogeosciences*, **115**, 10.1029/2008JG000920.

Constable, A.J., Melbourne-Thomas, J., Corney, S.P. *et al.* 2014. Climate change and Southern Ocean ecosystems I: how changes in physical habitats directly affect marine biota. *Global Change Biology*, **20**, 10.1111/gcb.12623.

Cunningham, S.A., Alderson, S.G., King, B.A. & Brandon, M.A. 2003. Transport and variability of the Antarctic Circumpolar Current in Drake Passage. *Journal of Geophysical Research - Oceans*, **108**, 10.1029/2001JC001147.

De Lavergne, C., Palter, J.B., Galbraith, E.D., Bernardello, R. & Marinov, I. 2014. Cessation of deep convection in the open Southern Ocean under anthropogenic climate change. *Nature Climate Change*, **4**, 278–282.

Dunne, J.P., John, J.G., Shevliakova, E., Stouffer, R.J., Krasting, J.P., Malyshev, S.L., Milly, P.C.D., Sentman, L.T., Adcroft, A.J., Cooke, W., Dunne, K.A., Griffies, S.M., Hallberg, R.W., Harrison, M.J., Levy, H., Wittenberg, A.T., Phillips, P.J. & Zadeh, N. 2013. GFDL's ESM2 global coupled climate-carbon earth system models. Part II: carbon system formulation and baseline simulation characteristics. *Journal of Climate*, **26**, 2247–2267.

GARCIA, H.E., LOCARNINI, R.A., BOYER, T.P., ANTONOV, J.I., ZWENG, M.M., BARANOVA, O.K. & JOHNSON, D.R. 2010. World Ocean Atlas 2009. Volume 4: Nutrients (phosphate, nitrate, silicate). *In* LEVITUS, S., *ed. NOAA Atlas NESDIS 71.* Washington, DC: US Government Printing Office, 398 pp.

GREGG, W.W. 2008. Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model. *Journal of Marine Systems*, **69**, 205–225.

GRIESEL, A., MAZLOFF, M.R. & GILLE, S.T. 2012. Mean dynamic topography in the Southern Ocean: evaluating Antarctic Circumpolar Current transport. *Journal of Geophysical Research - Oceans*, **117**, 10.1029/2011JC007573.

HEUZÉ, C., HEYWOOD, K.J., STEVENS, D.P. & RIDLEY, J.K. 2013. Southern Ocean bottom water characteristics in CMIP5 models. *Geophysical Research Letters*, **40**, 10.1002/grl.50287.

ILYINA, T., SIX, K.D., SEGSCHNEIDER, J., MAIER-REIMER, E., LI, H. & NEZ-RIBONI, I. 2013. Global ocean biogeochemistry model HAMOCC: model architecture and performance as component of the MPI-earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modeling Earth Systems*, **5**, 287–315.

JOLLIFF, J.K., KINDLE, J.C., SHULMAN, I., PENTA, B., FRIEDRICHS, M.A.M., HELBER, R. & ARNONE, R.A. 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, **76**, 64–82.

LOCARNINI, R.A., MISHONOV, A.V., ANTONOV, J.I., BOYER, T.P., GARCIA, H.E., BARANOVA, O.K., ZWENG, M.M. & JOHNSON, D.R. 2010. World Ocean Atlas 2009. Volume 1: Temperature. *In* LEVITUS, S., *ed. NOAA Atlas NESDIS 71.* Washington, DC: US Government Printing Office, 184 pp.

MARTIN, T., PARK, W. & LATIF, M. 2013. Multi-centennial variability controlled by Southern Ocean convection in the Kiel Climate Model. *Climate dynamics*, **40**, 2005–2022.

MAZLOFF, M.R., HEIMBACH, P. & WUNSCH, C. 2010. An eddy-permitting Southern Ocean State Estimate. *Journal of Physical Oceanography*, **40**, 10.1175/2009JPO4236.1.

MEIJERS, A.J.S., SHUCKBURGH, E., BRUNEAU, N., SALLEE, J.-B., BRACEGIRDLE, T.J. & WANG, Z. 2012. Representation of the Antarctic Circumpolar Current in the CMIP5 climate models and future changes under warming scenarios. *Journal of Geophysical Research - Oceans*, **117**, 10.1029/2012JC008412.

MOORE, J.K., LINDSAY, K., DONEY, S.C., LONG, M.C. & MISUMI, K. 2013. Marine ecosystem dynamics and biogeochemical cycling in the Community Earth System Model [CESM1(BGC)]: comparison of the 1990s with the 2090s under the rcp4.5 and rcp8.5 scenarios. *Journal of Climate*, **26**, 10.1175/JCLI-D-12-00566.1.

MOSS, R.H., EDMONDS, J.A., HIBBARD, K.A., MANNING, M.R., ROSE, S.K., VAN VUUREN, D.P., CARTER, T.R., EMORI, S., KAINUMA, M., KRAM, T., MEEHL, G.A., MITCHELL, J.F.B., NAKICENOVIC, N., RIAHI, K., SMITH, S.J., STOUFFER, R.J., THOMSON, A.M., WEYANT, J.P. & WILBANKS, T.J. 2010. The next generation of scenarios for climate change research and assessment. *Nature*, **463**, 747–756.

OSCHLIES, A. 2001. Model-derived estimates of new production: new results point towards lower values. *Deep-Sea Research II - Topical Studies in Oceanography*, **48**, 2173–2197.

PALMER, J.R. & TOTTERDELL, I.J. 2001. Production and export in a global ocean ecosystem model. *Deep-Sea Research I - Oceanographic Research Papers*, **48**, 1169–1198.

PIERCE, D.W., BARNETT, T.P. & MIKOLAJEWICZ, U. 1995. Competing roles of heat and freshwater flux in forcing thermohaline oscillations. *Journal of Physical Oceanography*, **25**, 2046–2064.

PINKERTON, M.H. & BRADFORD-GRIEVE, J. 2010. Phytoplankton: trophic modelling of the Ross Sea. Available at: https://www.ccamlr.org/en/system/files/08_Phytoplankton.pdf.

PINKERTON, M.H. & BRADFORD-GRIEVE, J.M. 2014. Characterizing foodweb structure to identify potential ecosystem effects of fishing in the Ross Sea, Antarctica. *ICES Journal of Marine Science*, **71**, 1542–1553.

PINKERTON, M.H., BRADFORD-GRIEVE, J.M. & HANCHET, S.M. 2010. A balanced model of the food web of the Ross Sea, Antarctica. *CCAMLR Science*, **17**, 1–31.

RAYNER, N.A., PARKER, D.E., HORTON, E.B., FOLLAND, C.K., ALEXANDER, L.V., ROWELL, D.P., KENT, E.C. & KAPLAN, A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research - Atmospheres*, **108**, 10.1029/2002JD002670

SÉFÉRIAN, R., BOPP, L., GEHLEN, M., ORR, J.C., ETHÉ, C., CADULE, P., AUMONT, O., MELIA, D.S.Y., VOLDOIRE, A. & MADEC, G. 2013. Skill assessment of three earth system models with common marine biogeochemistry. *Climate Dynamics*, **40**, 2549–2573.

SHU, Q., SONG, Z. & QIAO, F. 2015. Assessment of sea ice simulations in the CMIP5 models. *Cryosphere*, **9**, 399–409.

SMITH, W.O. JR., SEDWICK, P.N., ARRIGO, K.R., AINLEY, D.G. & ORSI, A.H. 2012. The Ross Sea in a sea of change. *Oceanography*, **25**, 90–103.

SMITH, W.O. JR., AINLEY, D.G., ARRIGO, K.R. & DINNIMAN, M.S. 2014a. The oceanography and ecology of the Ross Sea. *Annual Review of Marine Science*, **6**, 469–487.

SMITH, W.O. JR., DINNIMAN, M.S., HOFMANN, E.E. & KLINCK, J.M. 2014b. The effects of changing winds and temperatures on the oceanography of the Ross Sea in the 21st century. *Geophysical Research Letters*, **41**, 10.1002/2014GL059311.

STÖSSEL, A., NOTZ, D., HAUMANN, F.A., HAAK, H., JUNGCLAUS, J. & MIKOLAJEWICZ, U. 2015. Controlling high-latitude Southern Ocean convection in climate models. *Ocean Modelling*, **86**, 58–75.

TAGLIABUE, A., MTSHALI, T., AUMONT, O., BOWIE, A.R., KLUNDER, M.B., ROYCHOUDHURY, A.N. & SWART, S. 2012. A global compilation of dissolved iron measurements: focus on distributions and processes in the Southern Ocean. *Biogeosciences*, **9**, 10.5194/bg-9-2333-2012.

TURNER, J., BRACEGIRDLE, T.J., PHILLIPS, T., MARSHALL, G.J. & HOSKING, J.S. 2013. An initial assessment of Antarctic sea ice extent in the CMIP5 models. *Journal of Climate*, **26**, 1473–1484.

VAN VUUREN, D.P., EDMONDS, J., KAINUMA, M., RIAHI, K., THOMSON, A., HIBBARD, K., HURTT, G.C., KRAM, T., KREY, V., LAMARQUE, J.F., MASUI, T., MEINSHAUSEN, M., NAKICENOVIC, N., SMITH, S.J. & ROSE, S.K. 2011. The representative concentration pathways: an overview. *Climatic change*, **109**, 5–31.

VICHI, M., PINARDI, N. & MASINA, S. 2007. A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part I: Theory. *Journal of Marine Systems*, **64**, 89–109.

WANG, C.Z., ZHANG, L.P., LEE, S.K., WU, L.X. & MECHOSO, C.R. 2014. A global perspective on CMIP5 climate model biases. *Nature Climate Change*, **4**, 201–205.

ZAHARIEV, K., CHRISTIAN, J.R. & DENMAN, K.L. 2008. Preindustrial, historical, and fertilization simulations using a global ocean carbon model with new parameterizations of iron limitation, calcification, and N_2 fixation. *Progress in Oceanography*, **77**, 56–82.