The DOI for this manuscript is doi: 10.1175/JCLI-D-12-00417.1

Anav, A., P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu, 2013: EVALUATING THE LAND AND OCEAN COMPONENTS OF THE GLOBAL CARBON CYCLE IN THE CMIP5 EARTH SYSTEM MODELS. J. Climate. doi:10.1175/JCLI-D-12-00417.1, in press.

EVALUATING THE LAND AND OCEAN COMPONENTS OF THE GLOBAL CARBON CYCLE IN THE CMIP5 EARTH SYSTEM MODELS

A. Anav [1*], P. Friedlingstein [1], M. Kidston [2], L. Bopp [2], P. Ciais [2], P. Cox [1], C. Jones [3], M. Jung [4], R. Myneni [5], Z. Zhu [5]

[1]University of Exeter, College of Engineering, Mathematics and Physical Sciences, Exeter, England

[2]Laboratoire des Sciences du Climat et de l'Environnement, LSCE, Gif sur Yvette, France

[3]Met Office Hadley Centre, Exeter, UK

[4]Max Planck Institute for Biogeochemistry, MPI, Jena, Germany

[5]Boston University, Department of Geography & Environment, Boston, USA

****Corresponding author***: Alessandro Anav, A.Anav@exeter.ac.uk, College of Engineering, Mathematics & Physical Sciences, Harrison Building, North Park Road, Exeter EX4 4QF, UK

ABSTRACT

We assess the ability of 18 Earth System Models to simulate the land and ocean carbon cycle for the present climate. These models will be used in the next Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) for climate projections, and such evaluation allows identification of the strengths and weaknesses of individual coupled carbon-climate models as well as identification of systematic biases of the models.

Results show that models correctly reproduce the main climatic variables controlling the spatial and temporal characteristics of the carbon cycle. The seasonal evolution of the variables under examination is well captured. However, weaknesses appear when reproducing specific fields: in particular, considering the land carbon cycle, a general overestimation of photosynthesis and leaf area index is found for most of the models, while the ocean evaluation shows that quite a few models underestimate the primary production.

We also propose climate and carbon cycle performance metrics in order to assess whether there is a set of consistently better models for reproducing the carbon cycle. Averaged seasonal cycles and probability density functions (PDFs) calculated from model simulations are compared with the corresponding seasonal cycles and PDFs from different observed datasets.

Although the metrics used in this study allow identification of some models as better or worse than the average, our ranking is partially subjective due to the choice of the variables under examination, and can be also sensitive to the choice of reference data. In addition, we found that the model performances show significant regional variations.

1. INTRODUCTION

Earth System Models (ESMs) are complex numerical tools designed to simulate physical, chemical and biological processes taking place on Earth between the atmosphere, the land and the ocean. Worldwide, only a few research institutions have developed such models and used them to carry out historical and future simulations in order to project future climate change.

ESMs, and numerical models in general, are never perfect. Consequently, before using their results to make future projection of climate change, an assessment of their accuracy reproducing several variables for the present climate is required. In fact, the ability of a climate model to reproduce the present-day mean climate and its variation adds confidence to projections of future climate change (Reifen and Toumi 2009). Nevertheless, good skills reproducing the present climate do not necessarily guarantee that the selected model is going to generate a reliable prediction of future climate (Reichler and Kim 2008).

ESMs are routinely subjected to a variety of tests to assess their capabilities, and several papers provide extensive model evaluation (e.g. Tebaldi et al. 2006; Lin et al. 2007; Lucarini et al. 2007; Santer et al. 2007; Gillett et al. 2008; Gleckler et al. 2008; Reichler and Kim 2008; Schneider et al. 2008; Santer et al. 2009; Tjiputra et al 2009; Knutti et al. 2010; Steinacher et al. 2010; Radić and Clarke 2011; Scherrer 2011; Chou et al. 2012; Séférian et al. 2012; Yin et al. 2012). In these papers, the authors describe the performance of climate models by measuring their ability to simulate today's climate at various scales from global to regional. Results reported in these papers indicate that not all models simulate the present climate with similar accuracy. Furthermore, it should be noted that these papers also highlighted that the best models for a particular region of the Earth do not always achieve the same degree of performance in other regions. Additionally, the skill of the models is different according to the meteorological variables examined.

Within this context, the aim of this paper is twofold. The first aim is to quantify how well the CMIP5 (Coupled Model Intercomparison Project phase-5, Taylor et al. 2011) models represent the 20th century carbon cycle over the land and ocean, as well as the main climatic variables that influence the carbon cycle.

Traditional model evaluation, or diagnostics (e.g. Collins et al. 2006; Delworth et al. 2006; Johns et al. 2006; Zhou and Yu 2006; Waliser et al. 2007; Lin et al. 2008; Volodin et al. 2009; Marti et al. 2010; Xavier et al. 2010; Arora et al. 2011; Chylek et al. 2011; Collins et al. 2011; Radić and Clarke 2011; Watanabe et al. 2011), provide detailed assessments of the strengths and weaknesses of individual climate models based principally on seasonal and annual timescales, as well as on anomaly maps and zonal means.

Our model evaluation is performed at three different time scales: first, we analyze the long-term trend, which provides information on the model capability to simulate the temporal evolution over the 20th century, given GHG and aerosol radiative forcing. Second, we analyze the interannual variability (IAV) of physical variables as a constraint on the model capability to simulate realistic climate patterns that influence both ocean and continental carbon fluxes (Rayner et al 2008). Third, we evaluate the modelled seasonal cycle which, particularly in the Northern Hemisphere, constrains the model's simulation of the continental fluxes.

The second aim of the paper is to assess whether there is a set of consistently better models reproducing the carbon cycle and the main physical variables controlling the carbon cycle. One of the scientific motivations is that modellers commonly make use of large climate model projections to underpin impact assessments. So far, IPCC assumed that all climate models are equally good and they are equally weighted in future climate projections (Meehl et al. 2007). If an impacts modeller wants to choose the best models for a particular region however, assuming all models are equally good is not a requirement and models could be ranked, weighted or omitted based on performance.

Contrasting with diagnostics, metrics could be developed and used for such purposes (Gleckler et al. 2008; Maximo et al. 2008; Cadule et al. 2010; Räisänen et al. 2010; Chen et al. 2011; Errasti et al. 2011; Moise et al. 2011; Radić and Clarke 2011).

2. MODELS, REFERENCE DATA SETS, AND ASSESMENT OF PERFOMANCES

2.1 CMIP5 simulations

In this study we analyze outputs from 18 coupled carbon-climate models that are based on the set of new global model simulations planned in support of the IPCC Fifth Assessment Report (AR5). These simulations are referred to as CMIP5 (Coupled Model Intercomparison Project phase-5). This set of simulations comprises a large number of model experiments, including historical simulations, new scenarios for the 21st century, decadal prediction experiments, experiments including the carbon cycle and experiments aimed at investigating individual feedback mechanisms (Taylor et al. 2011). The CMIP5 multi-model data set has been archived by PCMDI and has been made available to the climate research community (http://cmip-pcmdi.llnl.gov/cmip5/).

Here we summarize the physical and biogeochemical model's performances for the historical experiment only (i.e. ESMs driven by CO_2 concentration). Among all the available CMIP5 ESMs, we selected the only models simulating both the land and ocean carbon fluxes and reporting enough variables for our analysis.

The models used in this study, as well as their atmospheric and ocean grids, are listed in **Table 1**; note that all the diagnostics and statistics are computed after regridding each model's output, and reference datasets, to a common 2x2 degrees grid. In case of carbon fluxes, our regridding approach assumed conservation of mass, while for the physical fields as well as for the LAI, we used a bilinear interpolation.

Table 2 reports the land and ocean biogeochemical models used by ESMs, while **Table 3** lists the variables considered in this study with the number of independent realizations (or ensemble member) for each model/variable. In fact, some models have only one run (realization), but other models have up to five runs (**Table 3)**. These realizations are climate simulations with different initial conditions. In the next section, we present results only from the first realization for each individual climate model, while for the final ranking we use the realization with the highest score for each individual model. In general it is expected that the ensemble of runs associated with a particular model with the same external forcing will reproduce very similar seasonal cycle and range of climate variability, irrespective of the initial conditions (Errasti et al. 2011). However because of each ensemble member having its own internal variability (largely unforced), the interannual variability of the ensemble average is expected to be reduced with respect to one individual simulation; for such reason we decided to use results from only the first realization, rather than the ensemble mean over the available realizations.

Our analysis focuses on the historical period (20th century simulations; historical experiment, CO_2 concentration driven), which was forced by a variety of externally imposed changes such as increasing greenhouse gas and sulfate aerosol concentrations, change in solar radiation, and forcing by volcanic eruptions. Considering the land surface, except for BCC-CSM1, BCC-CSM1-M and INMCM4 all models account for land use change (Table 2); likewise, except BCC models, NorESM1-ME, and CESM1-BGC none of the models have an interactive land nitrogen cycle (Table 2).

Since considerable uncertainty as to the true forcing remains, the forcing used and its implementation in the climate model is not exactly the same for all models (Jones et al. 2011). Rather, these runs represent each group's best effort to simulate the 20th century climate. The models were spun up under conditions representative of the pre-industrial climate (generally 1850 for almost all models, see **Table 2**). From this point, external time varying forcing, consistent with the historical period, was introduced, and the simulations were extended through to year 2005.

Although the CMIP5 archive includes daily means for a few variables, we focus here only on the monthly mean model output, since this temporal frequency is high enough to provide a reasonably comprehensive picture of model performance both in terms of mean state of the system, its seasonal and interannual variability, and trends.

In this study we focus mostly on the last 20 years of the 20th century simulations (1986–2005). During this period, in fact, the observational record is most reliable and complete, largely due to the expansion and advances in space-based remote sensing of vegetation greenness.

2.2 Reference data

The main focus of this paper is the evaluation of the land and ocean carbon fluxes. However, climatic factors exert a direct control on the terrestrial and ocean carbon exchange with the atmosphere (Houghton 2000; Schaefer et al. 2002), therefore we also provide an evaluation of the physical variables. The main physical factors controlling the land carbon balance are the surface temperature and precipitation (Piao et al. 2009), but also the cloud cover through its control on incoming radiation is important for the land carbon balance; however we decided to consider only the two most important variables influencing the land carbon cycle (Piao et al. 2009). In the ocean, physical fields include sea surface temperature (SST), which is important for biological growth and respiration rates as well as air–sea gas exchange, and mixing layer depth (MLD), which influences nutrient entrainment and the average light field observed by the phytoplankton (Martinez et al. 2002).

Considering the land and ocean carbon fluxes, some of the available datasets used for the comparison come from atmospheric inversion (discussed in section 2.2.6). To avoid pitfalls arising from weak data constraints, most inversion studies have relied on regularization techniques that include the aggregation of estimate fluxes over large regions (Engelen et al. 2002); as matter of fact, aggregating the observed regional fluxes in space is one way to lower the uncertainty due to the limited observational constraint (Kaminski et al. 2001; Engelen et al. 2002). Therefore, we only evaluate the net CO_2 fluxes simulated by models at global scale or over large latitudinal bands (see below). For all other model variables, the evaluation is performed at the grid level, conserving the spatial information. However, when presenting the results, all model performances are averaged over the following domains for land variables: Global (90S-90N), Southern Hemisphere (20S-90S), Northern Hemisphere (20N-90N), and Tropics (20S-20N). Considering the ocean carbon, according to Gruber et al. (2009) we aggregate results over 6 large regions: Globe (90S-90N), Southern Ocean (90S-44S), temperate Southern Ocean (44S-18S), Tropics (18S-18N), temperate Northern Ocean (18N-49N) and Northern Ocean (49N-90N).

In the following sub-sections we describe the different dataset used for the model comparison (see also **Table 3**).

2.2.1 Land temperature and precipitation

Monthly gridded surface temperature and precipitation were constructed from statistical interpolation of station observations by the Climatic Research Unit (CRU) of the University of East Anglia (New et al. 2002; Mitchell and Jones 2005). CRU provides a global coverage only for land points between 1901 and 2006 with a spatial resolution of 0.5° (**Table 3**). Most of previous model-data comparison studies use ERA40 (or other reanalysis) instead of the CRU dataset, due to the complete global land and ocean coverage, and the way these reanalysis are built. Specifically, the reanalysis are a combination of weather model output and a large amount of assimilated different observational data. Therefore, unlike CRU that is built on statistical principles, the reanalysis are based on physical principles (Scherrer 2011). Also comparison of the ERA40 dataset with the CRU land temperature shows good agreement for most regions and the differences are comparatively small in comparison to the model differences (Scherrer 2011). However, CRU provides data for the entire 20th century allowing the evaluation of the simulated temperature and precipitation trends.

2.2.2 Sea Surface Temperature

For the Sea Surface Temperature (SST) evaluation we use the HadISST (Rayner et al. 2003), a combination of monthly global SST and sea ice fractional coverage on a 1°x1° spatial grid from 1870 to date.

The SST data are taken from the Met Office Marine Data Bank (MDB), which from 1982 onward also includes data received through the Global Telecommunications System. To enhance data coverage, monthly median SSTs for 1871–1995 from the Comprehensive Ocean–Atmosphere Data Set (COADS) were also used where there were no MDB data. HadISST temperatures are reconstructed using a two-stage reduced-space optimal interpolation procedure, followed by superposition of quality-improved gridded observations onto the reconstructions to restore local detail (Dima and Lohmann 2010). SSTs near sea ice are estimated using statistical relationships between SST and sea ice concentration (Rayner et al. 2003).

2.2.3 Mixed Layer Depth

The ocean Mixed Layer Depth (MLD) can be defined in different ways, according to the dataset used. In this paper, MLD data are from the Ocean Mixed Layer Depth Climatology Dataset as described in de Boyer Montégut et al. (2004). Data are available in monthly format on a 2°×2° latitude–longitude mesh and were derived from more than five million individual vertical profiles measured between 1941 and 2008, including data from Argo profilers, as archived by the National Oceanographic Data Centre (NODC) and the World Ocean Circulation Experiment (WOCE). In order to solve the MLD overestimation due to salinity stratification, in this dataset the depth of the mixed layer is defined as the uppermost depth at which temperature differs from the temperature at 10 m by 0.2°C. A validation of the temperature criterion on moored time series data shows that this method is successful at following the base of the mixed layer (de Boyer Montégut et al. 2004).

2.2.4 Terrestrial Gross Primary Production

Gross Primary Production (GPP) represents the uptake of atmospheric CO_2 during photosynthesis and is influenced by light availability, atmospheric CO_2 concentration, temperature, availability of water and nitrogen, and several interacting factors (e.g. atmospheric pollution, harvesting, insect attacks).

Direct GPP observations at global scale and for our reference period (1986-2005) do not exist, since in the 1980s no measurement sites existed, and satellite observations of GPP were not yet available. Recently, satellite derived GPP products have been developed (e.g Mao et al. 2012) but do not cover the reference period.

Here we use GPP estimates derived from the upscaling of data from the FLUXNET network of eddy covariance towers (Beer et al. 2010). The global FLUXNET upscaling uses data oriented diagnostic models trained with eddy covariance flux data to provide empirically derived, spatially gridded fluxes (Beer et al. 2010). In this study, we use the global FLUXNET upscaling of GPP based on the model tree ensembles (MTE) approach, described by Jung et al. (2009, 2011). The upscaling relies on remotely sensed estimates of the fraction of absorbed photosynthetically active radiation (fAPAR), climate fields, and land cover data. The spatial variation of mean annual GPP as well as the mean seasonal course of GPP are the most robust features of the MTE-GPP product, while there is less confidence on its interannual variability and trends (Jung et al 2011). MTE-GPP estimates are provided as monthly fluxes covering the period 1982-2008 with a spatial resolution of 0.5° (**Table 3**).

2.2.5 LAI

Leaf area index (LAI) is defined as the one-sided green leaf area per unit ground area in broadleaf canopies and as one-half the total needle surface area per unit ground area in coniferous canopies (Myneni et al. 2002). The LAI data set used in this study (LAI3g) was generated using an Artificial Neural Network (ANN) from the latest version (third generation) of GIMMS AVHRR NDVI data for the period July 1981 to December 2010 at 15-day frequency (Zhu et al. 2013). The ANN was trained with best-quality Collection 5 MODIS LAI product and corresponding GIMMS NDVI data for an overlapping period of 5 years (2000 to 2004) and then tested for its predictive capability over another five year period (2005 to 2009). The accuracy of the MODIS LAI product is estimated to be 0.66 LAI units (Yang et al. 2006); further details are provided in Zhu et al. (2012).

2.2.6 Land-atmosphere and ocean-atmosphere CO_2 fluxes

The net land-atmosphere (NBP) and ocean-atmosphere (fgCO_2) CO_2 exchange estimated by CMIP5 models are compared with results from atmospheric inversions of the Transcom 3 project (Gurney et al. 2004; Baker et al. 2006), an intercomparison study of inversions (Gurney et al. 2002, 2003, 2004, 2008). Within this project a series of experiments were conducted in which several atmospheric tracer transport models were used to calculate the global carbon budget of the atmosphere.

Transcom 3 results represent the a posteriori surface CO_2 fluxes inferred from monthly atmospheric CO_2 observations at a set of GLOBALVIEW stations after accounting for the effects of atmospheric transport on a prescribed a priori surface flux, which is corrected during the atmospheric inversion (Gurney et al., 2003). In other words, the goal of the atmospheric inversion process is to find the most likely combination of regional surface net carbon fluxes that best matches observed CO_2 within their error, given values of prior fluxes and errors, after those fluxes have been transported through a given atmospheric model (Gurney et al., 2003, 2008).

Flux estimates from atmospheric inverse models are comprehensive, in the sense that all ecosystem sources and sinks, fossil fuel emissions, and any other processes emitting or absorbing CO_2 (e.g. aquatic CO_2 fluxes, decomposition of harvested wood and food products at the surface of the Earth) are, in principle, captured by the inversion CO_2 fluxes results.

Transcom 3 also provides an ensemble mean computed over 13 available atmospheric models in the period 1996-2005 at a spatial resolution of 0.5°. The use of several models was motivated because large differences in modelled CO_2 were found between models using the same set of prescribed fluxes (Gurney et al. 2004). However it is argued that an average of multiple models may show characteristics that do not resemble those of any single model, and some characteristics may be physically implausible (Knutti et al. 2010). In absence of any other information to select the most realistic transport models, Gurney et al. (2002) used the “between-model” standard deviation to assess the error of inversions induced by the transport model errors. In addition, Stephens et al. (2007) suggest that an average taken across all models does not provide the most robust estimate of northern versus tropical flux partitioning. Additionally, they point to three different models as best representing observed vertical profiles of [CO2] in the Northern Hemisphere (Stephens et al. 2007). For such reasons, instead of using the Transcom 3 ensemble mean and the “between-model” standard deviation, we used results from the only JMA model (Gurney et al. 2003), being one of the three models suggested by Stephens et al. (2007) and the only one available in our reference period 1986-2005.

We also use results from the Global Carbon Project (GCP, http://www.tyndall.ac.uk/global-carbon-budget-2010), which estimates, using several models and observations, the ocean-atmosphere and land-atmosphere CO_2 exchange (Le Quéré et al. 2009). These results are the most recent estimates of global CO_2 fluxes for the period 1959-2008. Within this project, the global ocean uptake of anthropogenic carbon was estimated using the average of four global ocean biogeochemistry models forced by observed atmospheric conditions of weather and CO_2 concentration (Le Quéré et al. 2009). The global residual land carbon sink was estimated from the residual of the other terms involved in the carbon budget, namely the residual land sink is equal to the sum of fossil fuel emissions and land use change less the atmospheric CO_2 growth and the ocean sink (Le Quéré et al. 2009). From the GCP analysis, the NBP can easily be computed as the difference between the residual sink and the land use change.

Finally, in addition to the inversion and GCP data, for the ocean-atmosphere flux we also use results from Takahashi et al. (2002, 2009). This product contains a climatological mean distribution of the partial pressure of CO_2 in seawater (pCO_2) over the global oceans with a spatial resolution of 4° (latitude) x 5° (longitude) for the reference year 2000 based upon about 3 million measurements of surface water pCO_2 obtained from 1970 to 2007 (Takahashi et al. 2009). It should be noted that Takahashi et al. (2002) data are used as prior knowledge in many atmospheric inversions, suggesting that the two datasets are not completely independent.

Although the difference between the partial pressure of CO_2 in seawater and that in the overlying air (ΔpCO_2) would be a better reference data set for the oceanic uptake of CO_2, in this study we have used the net sea-air CO_2 flux ($fgCO_2$) to be consistent with the land flux component of this paper. The net air-sea CO_2 flux is estimated using the sea-air pCO_2 difference and the air-sea gas transfer rate that is parameterized as a function of wind speed (Takahashi et al. 2009).

2.2.7 Vegetation and soil carbon content

Heterotrophic organisms in the soil respire dead organic carbon, the largest carbon pool in the terrestrial biosphere (Jobbagy and Jackson 2000); therefore the soil carbon, through the heterotrophic respiration, represents a critical components of the global carbon cycle.

There are several global datasets that include estimates of soil carbon to a depth of 1 m. Generally, there are two different approaches to creating such datasets: (1) estimation of carbon stocks under natural, or mostly undisturbed, vegetation using climate and ecological life zones (2) extrapolation of soil carbon data from measurement in soil profiles using soil type (Smith et al., 2012).

The Harmonized World Soil Database (HWSD) developed by Food and Agriculture Organization of the United Nations (FAO 2012) and International Institute for Applied Systems Analysis (IIASA) is the most recent, highest resolution global soils dataset available. It uses vast volumes of recently collected regional and national soil information to supplement the 1:5000000 scale FAO-UNESCO Digital Soil Map of the World. It is an empirical dataset and it provides soil parameter estimates for topsoil (0–30 cm) and subsoil (30–100 cm), at 30 arc-second resolution (about 1 km).

The CMIP5 ESMs do not report the depth of carbon in the soil profile, making direct comparison with empirical estimates of soil carbon difficult. For our analysis, we assumed that all soil carbon was contained with the top 1 meter. Litter carbon was a small fraction of soil carbon for the models that reported litter pools; thus, we combined litter and soil carbon for this analysis and refer to the sum as soil carbon.

For the HWSD, the major sources of error are related to analytical measurement of soil carbon, variation in carbon content within a soil type, and assumption that soil types can be used to extrapolate the soil carbon data. Analytical measurements of soil carbon concentrations are generally precise, but measurements of soil bulk density are more uncertain (Todd-Brown et al. 2012).

In addition to the soil carbon, also the vegetation carbon is a key variable in the global carbon cycle. In the 1980s, Olson et al. (1985) developed a global ecosystem-complex carbon stocks map of above and below ground biomass following more than 20 years of field investigations, consultations, and analyses of the published literature. Gibbs (2006) extended Olson et al.'s methodology to more contemporary land cover conditions using remotely sensed imagery and the Global Land Cover Database (GLC, 2000). For this analysis we used the data created by Gibbs (2006), with a spatial resolution of 0.5 degree.

2.2.8 Oceanic Net Primary Production

Oceanic integrated net primary production (NPP or intPP) is the gross photosynthetic carbon fixation (photosynthesis), minus the carbon used in phytoplankton respiration. NPP is regulated by the availability of light, nutrients and temperature and affects the magnitude of the biological carbon pump. Oceanic export production (EP) exerts a more direct control on air-sea CO_2 fluxes, however due to limited EP data we assess models compared to NPP estimates. In addition, we used the NPP to be consistent with the use of GPP in the land section of the study, however often it is argued that a proper validation of biological oceanic models should be based on the comparison of surface chlorophyll concentration rather than phytoplankton primary production.

We used NPP estimated from satellite chlorophyll by the Vertically Generalised Production Model (VGPM) (Behrenfeld and Falkowski 1997). The VGPM computes marine NPP as a function of chlorophyll, available light, and temperature dependent photosynthetic efficiency. The NPP, estimated with the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) from 1997-2007, is a monthly dataset with a spatial resolution of about 6 km.

As well as previous datasets (GPP-MTE, LAI, Transcom 3 and GCP data derived CO_2 fluxes), it should be noted that although this is one of the best available global NPP products it is not actually data, rather a model estimate dependent on parameterisations (the temperature dependent assimilation efficiency for carbon fixation and an empirically determined light dependency term).

2.2.9 Uncertainty in the observed dataset

One limitation of most of the above chosen reference datasets is that it is in general difficult to estimate their observational errors (except for Bayesian inversions that explicitly come with uncertainty estimates). Sources of uncertainty include random and bias errors in the measurements themselves, sampling errors, and analysis error when the observational data are processed through models or otherwise altered. In short, the quality of observational measurements varies considerably from one variable to the next (Gleckler et al. 2008) and is often not reported.

Errors in the reference data are frequently ignored in the evaluation of models. It is often argued that this is acceptable as long as these errors remain much smaller than the errors in the models (Gleckler et al. 2008). A full quantitative assessment of observational errors by the estimation of its impact on the model ranking is however beyond the scope of this study.

Nevertheless, we would report that some of the reference data used for model validation show relevant problems. For instance, the ocean NPP is calculated from SeaWiFS satellite chlorophyll data which contains a significant uncertainty of ~30% (Gregg and Casey, 2004).

The MLD and SST data sets have a lack of observations in the Southern Ocean compared to other regions, hence the uncertainty in these data sets is greatest in the Southern Ocean (De Boyer Montégut et al. 2004).

It is also argued that CRU has been designed to provide best estimates of interannual variations rather than detection of long-term trends and (Mitchell and Jones 2005).

Finally, the soil databases are based on a limited number of soil profiles and extrapolated to other areas according to soil type. Climate or land cover and management are usually not considered so that these data have high associated uncertainty.

2.3 Assessment of model performances

A series of measures of analysis are employed here for model evaluation and ranking; the model performances are evaluated at every grid point and then aggregated over the different land and ocean sub-domains. However, as previously described in section 2.2 the atmospheric inversion estimates do not provide any reliable information at grid cell level, therefore for land-atmosphere and ocean-atmosphere CO_2 fluxes only the evaluation is performed using regional averages of the CO_2 fluxes. In the following we describe the diagnostics used for model evaluation and the metrics used for model ranking.

2.3.1 Diagnostics definition

Climatic trends for land surface temperature, land precipitation and SST are estimated by the linear trend value obtained from a least square fit line computed for the full period 1901-2005 of data, while for the LAI, and GPP due to the unavailability of data before 1982, the trends are computed in the same way but for the reference period 1986-2005.

Looking at simulated interannual variability, the root-mean square error (RMSE) is not an appropriate measure for characterizing this aspect of model performance because there is no reason to expect models and observations to agree on the phasing of internal (natural unforced) interannual variations (e.g., the timing of El Niño events) (Lin 2007; Gleckler et al. 2008). Standard measures of model mean variability such as the ratio of the standard deviation of the model means divided by the standard deviation of the means in the reference data set suffer from the serious problem that regions with too large/small IAV can cancel out and therefore give a too optimistic picture of model performance (Gleckler et al. 2008; Scherrer, 2011). To avoid these cancellation effects the Model Variability Index (MVI) as introduced by Gleckler et al. (2008) and Scherrer (2011) is used here to analyze the performance for each model, as given by:

$$MVI_{x,y}^{M} = \left(\frac{s_{x,y}^{M}}{s_{x,y}^{O}} - \frac{s_{x,y}^{O}}{s_{x,y}^{M}} \right)^{2} \qquad (1)$$

where $s_{x,y}^{M}$ and $s_{x,y}^{O}$ are the standard deviations of the annual time series of models and observation for a given variable, at each grid-point (x, y). Using this simple index of performance, we compare each model's variability at every grid cell and then average over the different sub-domains in the period 1986-2005. Perfect model–reference agreement would result in a MVI value of 0. The MVI provides a good measure to assess differences between model and reference data standard deviations and allow us to identify consistent biases in the standard deviations of single models. The definition of a MVI threshold value that discriminates between 'good' and 'bad' is somewhat arbitrary. Scherrer (2011), in his CMIP3 validation paper, defined a MVI < 0.5 as a good representation of IAV. In this paper we use the same threshold, although in case of biological variables the MVI could be much larger than 0.5.

Often it is also argued that a 20-year window could be not long enough for characterizing the long time-scale variance of a model (Wittenberg 2009; Johnson et al. 2011). This means that when the MVI is being computed over the last 20 years there is an implicit assumption that the variability is representative of the full length of the simulation. To test whether this is the case, we also have accounted the MVI for the physical variables over the period 1901-2005, and we found a relevant reduction in the MVI of global surface temperature, precipitation and SST compared to the MVI computed in the period 1986-2005 (not shown). This confirms that a 20-year windows is pretty marginal in characterizing what the actual variability of the model is. However, considering this work, while for climate variables it is possible to compute the MVI from the beginning of last century, in case of all the other variables the data are limited to the only last 20 years, therefore we decided to analyze the MVI over the period 1986-2005 to be consistent between physical and biological variables.

2.3.2 Metrics definition

Two different skill scores are used for the model ranking. In the case of mean annual cycle we check the ability of models to reproduce both the phase and amplitude of the observations during the period 1986-2005. Starting for monthly mean climatological data, we use the centered root-mean square (RMS) error statistic to account for errors in both the spatial pattern and the annual cycle. Given a model (M) at the grid-point (x, y) and the reference dataset at the same location ($O_{x,y}$), the errors of the model *m* ($E_{x,y}^{m^2}$) is calculated as follow:

$$E_{x,y}^{m^2} = \frac{1}{N}\sum_{t=1}^{N}\left[\left(M_t^{x,y} - \overline{M}^{x,y}\right) - \left(O_t^{x,y} - \overline{O}^{x,y}\right)\right]^2 \qquad (2)$$

where *t* corresponds to the temporal dimension, N is the number of months (i.e. 12), and $\overline{M}^{x,y}$ and $\overline{O}^{x,y}$ are the mean values of the model and reference data, respectively, at the grid point (x,y).

In order to get an error between 0 and 1 (where 0 corresponds to poor skill and 1 perfect skill), we normalize the error of the model *m* dividing it by the maximum error computed considering all the models at the grid point (x,y). Therefore the relative error (Re) of a single model *m* becomes:

$$\mathrm{Re}_{x,y}^{m} = 1 - \frac{E_{x,y}^{m^2}}{\max(E_{x,y}^{2})} \quad (3)$$

Unlike Gleckler et al. (2008) that normalized their seasonal skill score by the median of the RMS errors computed considering all the models, here we decided to divide by the maximum RMS error in order to have a skill score ranging between 0 and 1.

The second skill score used for model ranking is based on the comparison of Epanechnikov kernel-based probability density functions (PDFs; Silverman 1986) of models with observations (Perkins et al., 2007). This skill score provides a very simple but powerful measure of similarity between data and observations since it allows to compare both the mean state and the interannual variability of a given variable by calculation of the common area under the two PDFs (Maximo et al. 2008). If models perfectly reproduce the observed condition, the skill score would equal 1, which is the total area under a given PDF. On the contrary, if a model simulates the observed PDF poorly, it will have a skill score close to 0, namely there is not any overlap between the observed and modelled PDF. Note that despite this seeming to be similar to the Kolmogorov–Smirnov test for the similarity of PDFs, there is a fundamental difference between them: the Kolmogorov–Smirnov test is based on the maximum difference between cumulative PDFs, whilst the skill score is based on the common area under the PDF curves (Errasti et al. 2011). Starting from yearly data, and given $Z_{x,y}$ the common area under the observed PDF ($z_{x,y}^{O}$) and the simulated PDF ($z_{x,y}^{M}$) at the grid point (x,y):

$$Z_{x,y} = \min(z_{x,y}^{O}, z_{x,y}^{M}) \quad (4)$$

the skill score at a given geographical location is computed in the following way:

$$s_{x,y} = w * \int_{1}^{N} Z_{x,y} \quad (5)$$

where $s_{x,y}$ is the numerical value of the skill score ($0 \leq s_{x,y} \leq 1$), N is the number of intervals used to discretize the PDF estimated by means of the Epanechnikov kernels (in this study, N=100), and w is a weight (**Table 4**) introduced in order to give lower weight at the grid points where models are expected to poorly reproduce the observations. In fact, models are expected not to faithfully reproduce the observation in some specific regions such as in area of complex topography (i.e. in mountainous regions the coarse resolution of models does not allow to correctly reproduce the right temperature pattern) or over specific surface cover (ex. costal regions, ice-covered area, sparse vegetated points).

This measure is however imperfect: a model that is able to simulate the tails of a distribution well (i.e. extreme events like heat waves or cold spells, drought or heavy rain) would be very valuable, but if it simulates the more common regions of the PDF poorly it could score badly overall. Conversely, a model could appear skilful by simulating all the probabilities one or two standard deviations from the mean while being poor towards the tails (Maximo et al. 2008).

In general, models that properly simulate the observed mean value of a given variable, namely they fall into the range of $\pm 1\sigma$ of the observed PDF, are able to reproduce at least the 68.2% of the reference data. Maximo et al. (2008) defined as 'adequate' those models with a skill score greater than 0.9; this value was chosen since it allows identification of not only models that correctly capture the mean value, but also those models that capture a considerable amount of the interannual variability.

However, a threshold of 0.9 is too large when aggregating the skills over sub-regions, therefore in this study we consider a model as having relevant skill when it simulates at least 1σ of the observed PDF.

This method has already been used for AR4 ranking over Australia (Perkins et al. 2007; Maxino et al. 2008), Spain (Errasti et al. 2011) and CORDEX regions (Jacob et al. 2012). In their study, Errasti et al. (2011) removed all the points below a threshold value of 0.7 to avoid models characterized by very poor values affecting the overall score. However, this latter procedure is questionable since over large sub-regions removing the points with a skill lower than 0.7 will favour only the points with good agreement to observations and any poor performance of models related to severe bias will not be regarded. Additionally, removing all the points below a particular low threshold (e.g. 0.05) can lead to an overestimation of a model's skill. For this reason, in order to compute the regional skill score we apply a weighted mean, giving relatively large weights to points where the skill score exceed 0.75 and low importance to points where the score is poor (**Table 4**). We also have computed the ranking without weighting the skill scores (not shown) and we found that the weights only change the models skill values, leaving unchanged the overall ranking.

In addition, for those variables we are unable to build the PDFs due to the lack of yearly data (e.g.soil carbon, vegetation carbon and MLD) the skill score is computed using the bias between a given model (M) and the reference data (O). Given the bias (B) of the model *m* at the grid point (x,y):

$$B_{x,y}^{m} = \left| M_{x,y} - O_{x,y} \right| \qquad (6)$$

the skill score is computed following the equation 3. It should also be noted that normalizing the skill score calculations in this way yields a measure of how well a given model (with respect to a particular reference data set) compares with the typical model error, namely it leads to a more optimistic skill compared to the PDF-based skill score.

3. CMIP5 MODELS PERFORMANCES DURING THE 20th CENTURY

Since the simulation of physical variables will affect the simulation of the carbon cycle, we first briefly show how CMIP5 models reproduce these variables and then we focus on the carbon cycle performances. In particular, the evaluation of climatic variables is needed to assess whether any bias in the simulated carbon variables can be related to poor performances of the ESMs reproducing physical variables or is mainly due to the poor representation of some biogeochemical processes into the biological components of ESMs.

3.1 Land surface temperature, land precipitation, SST and MLD evaluation

The temporal evolution of global mean surface temperature, for the land points only (without Antarctica), is shown in **Figure 1** (upper panel) for the CMIP5 simulation as well as for the observations derived data-product (CRU).

Like for the AR4 results (Solomon et al. 2007), the CMIP5 simulations of 20th century that incorporate anthropogenic forcing (including increasing greenhouse gas concentrations and aerosols concentrations), as well as natural external forcing (volcanoes, change in solar radiation) are able to correctly reproduce the observed temperature anomaly, the observed data being systematically within the grey shading representing the range of variability of CMIP5 models. Plotting the CMIP5 temperature time series as anomalies with respect to the base period 1901–1930, all the models exhibit a general upward temperature trend (**Figure 1**); the net temperature increase over the historical period is determined primarily by a balance between the warming caused by increased GHGs and the cooling over some regions associated with increasing aerosols.

The ensemble mean suggests that CMIP5 models correctly reproduce the transient drop in global mean temperatures owing to main volcanic eruptions followed by gradual recovery over several years (**Figure 1**). Larger interannual variations are seen in the observations than in the ensemble mean, consequently, mainly during the first 50 years, the observed evolution lies outside the 90% confidence limits diagnosed from the CMIP5 ensemble spread (red shading). This result is related with the multi-model ensemble mean that filters out much of the natural variability (unforced and forced, i.e. volcanic, solar, and aerosols) simulated by each of the CMIP5 models. In addition, the ensemble spread (i.e. range of model variability) shows an increase with lead time, reflecting the loss of predictability associated with the different climate sensitivities, i.e. with the different model responses to forcing (Solomon et al. 2007; Hawkins and Sutton 2009).

In **Figure 1** (lower panels) we present, for each model, the mean surface temperature over the period 1986-2005, the MVI computed in the same temporal period, and the trend during 1901-2005. On the x-axis, models falling at the left (right) of observations indicate a cold (warm) bias, while on the y-axis models above (below) the observations have a stronger (lower) trend than observations.

The comparison with CRU data shows that in general few models have a warm bias (within 1 °C), while most of the models have a cold bias (**Figure 1**). Poor performances have been found for the INMCM4 model: specifically, its global cold bias is around 2.3 °C, with the minimum found in northern hemisphere (1.8 °C), and a maximum in the tropics (3.2 °C). Conversely, the best performances have been found in IPSL-CM5A-MR, MPI-ESM-LR, MPI-ESM-MR and GFDL-ESM2M models that are consistently closer to CRU data. Looking at the trends, however, IPSL-CM5A-MR and GFDL-ESM2M generally seem to be closer to the observations than MPI-ESM-LR and MPI-ESM-MR.

On the other hand, GFDL-ESM2M shows the poorest performances reproducing the observed IAV, having a MVI larger than 1.4 at global scale, while only few models show a MVI lower than 0.5 (indicating a good representation of the simulated IAV). The best results in terms of simulated IAV are found in the Northern Hemisphere, where several models show a MVI lower than 0.5; conversely, in the tropics most of models have a MVI larger than 1.

In **Figure 2 (**upper panel) we compare precipitation changes during the 20th century over land surfaces as reconstructed from station data (CRU) and simulated by individual CMIP5 models; shown are annual anomalies with respect to the period 1901-1930.

The CMIP5 models correctly reproduce the precipitation variability: specifically, for most of the time the reference data falls inside the range of variability of models, identified by the grey shading. Explosive volcanoes eruptions prescribed to models introduce anomalies in the simulated historical precipitation as seen by temperature; a clear precipitation reductions around the year 1991 associated with the Pinatubo eruptions is found in both CRU data and CMIP5 simulations.

Looking at the multi-model ensemble mean, it does not reproduce the amplitude of temporal evolution in 20th century terrestrial precipitation (see also Allan and Soden 2007; John et al. 2009; Liepert and Previdi 2009), being the observations larger than the 90% confidence limits diagnosed from the ensemble spread (blue shading). As already described for the temperature, the averaging process partially filters out the IAV.

The evaluation of precipitation for every model is given in **Figure 2** (lower panels). The best performances reproducing global precipitation are found in IPSL-CM5B-LR, BCC-CSM1-M and MPI models. BCC-CSM1, HadGEM2-ES, and HadGEM2-CC models show a slight wet bias (less than 40 mm/y), while CanESM2, IPSL-CM5A-LR and IPSL-CM5A-MR have a dry bias of about 80 mm/y. All the other models overestimate global precipitation with a bias of about 100 mm/y. In the Southern

Hemisphere several models match the CRU data well, while IPSL-CM5A-LR and IPSL-CM5A-MR showing a dry bias, and NorESM1-ME and CESM1-BGC have a strong wet bias. In the tropical region, quite a few models are able to reproduce the mean precipitation, while in the Northern Hemisphere, except CanESM2, all the models show a wet bias.

Looking at the IAV none of the models has a MVI close to the threshold of 0.5; the best results are found in the Southern Hemisphere for the Hadley models. As expected, the worst performances reproducing the precipitation IAV occur in the tropical region, reflecting the inability of these models in reproducing the interannual variations in the hydrological cycle (Lin 2007; Scherrer 2011); as already suggested by Wild and Liepert (2010) inadequacies in the simulation of surface radiation balance may contribute to the poor simulation of IAV during the 20th century**.** In addition, shortcomings in the representation of the natural variability in atmosphere/ocean exchanges of energy and water that result in variations of convection and consequently in cloudiness and humidity can contribute to a poor representation of precipitation IAV in CMPI5 models (Lin 2007; Wild and Liepert 2010).

The evaluation of the trend show that at global scale and in the tropical region several models are close to CRU, while in the Southern and Northern Hemisphere in general the models are not capable to capture the observed wettening trend. This is particularly evident in the Southern Hemisphere where the CMIP5 models show an ensemble trend around zero, while the CRU data gives a positive trend of 5.5 mm/decade over the period 1901-2005.

In order to understand the source of this mismatch between CMIP5 models and CRU data, we also use precipitation data from the Global Precipitation Climatology Project (GPCP) (Adler et al., 2003) for a further comparison. The GPCP trend in the Southern Hemisphere during the period 1979-2005 is -0.4±9.5 mm/decade, while CRU shows a strong positive trend of 13±10mm/decade over the same period; this suggests that the two datasets show a completely different trend. Although these results are affected by a large uncertainty, it is often argued on the reliability of CRU for the long term trends (Mitchell and Jones 2005).

Figure 3 (upper panel) shows the temporal evolution of global mean SST. Unlike the observed surface temperature that is scatted around the CMIP5 ensemble mean and falls in the middle of the gray shading, the observed SST is markedly above the ensemble mean, particularly during the period 1940-1970.

The CMIP5 ensemble mean shows an increasing trend, with declining periods in the early 1960's and 1990's as a consequence of the cooling due to the Agung and Pinapubo eruptions, and a sharper rise in the post 1960 period. The HadISST data shows an overall more linear increase than the CMIP5 model ensemble mean. Similar to the land temperature trend, the SST trend is primarily a balance between warming caused by GHG concentrations in the atmosphere and cooling resulting from aerosol emissions, modulated by the heat uptake by the ocean. Thus, factors regulating the heat uptake by the ocean such as changes in the thermohaline circulation, and upwelling have an effect on SST.

Aerosols from volcanic eruptions can lower SST at the time of the eruption and for a few years following the eruption. The CMIP5 models simulate a drop in SST as a result of the main volcanic eruptions, as can be seen in **Figure 3** (upper panel).

Figure 3 (lower panels) shows that the increasing trend in SST is evident in all regions for all the CMIP5 models except in the high latitude Southern Hemisphere where GFDL-ESM2M shows a cooling and the high latitude Northern Hemisphere where GFDL-ESM2G displays a cooling. It should also be noted that the trend for BNU-ESM has been computed over the period 1950-2005, rather than in the period 1901-2005, and it explains why this model exhibits this large trend compared to both observations and other CMIP5 models.

Most of the models show a cold bias, particularly in the Northern Hemisphere, and a lower trend than the observations, particularly in the Southern Hemisphere. At the global scale most of the models display a cold bias, with IPSL-CM5A-LR having the largest cold bias (1 °C). All models except IPSL-CM5A-LR, IPSL-CM5A-MR, MPI-ESM-LR and BCC-CSM1 show a lower trend than observations, with the lowest trend being in HadGEM2-ES, which has an increase of 0.4 °C/decade less than is seen in observations. The interannual variability is fairly well simulated by CMIP5 models, with a MVI lower than 1.5 in most of the sub-domains and for most of the models; however, severe problems reproducing the IAV are found in the high latitude Northern Hemisphere where most of models generally show a MVI larger than 2. Since we also found poor performances for a few models in reproducing the IAV in the Southern Hemisphere, the poor skill could be related to sea ice cover that affects both measured and modelled SST.

As already described in section 2.2.3 the reference MLD dataset is a climatology, therefore it is not possible to provide the same evaluation used for the other physical variables. However, the MLD seasonal cycle allows identification of some importance differences between models, and also allows the identification of possible bias when compared to observations. **Figure 4** shows the seasonal performance of each of the models in comparison to observed MLD (De Boyer Monégut et al., 2004). In general all the models simulate the basic seasonal cycle. However, in all the models (except the Hadley models) there is a consistent slight deep bias at the global scale, with a strong bias found in MPI-ESM-LR and MPI-ESM-MR.

The large global bias found in MPI models is related to a very deep mixed layer in the Weddell gyre, the aggregation of regions means that the entire Southern Ocean MLD is over estimated during austral winter. However it must also be considered that deep mixed layers of up to 800m are indeed observed in this region (Rintoul and Trull 2001). In addition, there is a lack of observations in the Southern Ocean compared to other regions and therefore there are biases in the data, which is based on individual profiles of temperature and salinity.

The biases are less pronounced in the Northern Hemisphere, however several models display a deep bias, particularly in winter. Most of the models show a shift in the timing of the maximum and minimum MLD compared to the observations, with the maximum occurring 1 month later. This would have a knock on effect on other components of the model, such as the timing of the spring bloom. Summer MLDs are better simulated as there is less variability at this time, with summer depths between approximately 10 and 50m in all sub-regions.

It should also be noted that some inconsistencies between CMIP5 models might arise due to differing definitions of mixed layer depth between the CMIP5 modelling groups.

3.2 CMIP5 land carbon

The land-atmosphere CO_2 flux, or net exchange of carbon between the terrestrial biosphere and the atmosphere (NBP), represents the difference between carbon uptake by photosynthesis and release by plant respiration, soil respiration and disturbance processes (fire, windthrow, insects attack and herbivory in unmanaged systems, together with deforestation, afforestation, land management and harvest in managed systems) (Denman et al. 2007). In **Figure 5** we compare the temporal evolution of simulated global land-atmosphere CO_2 flux with the GCP global carbon budget estimates (Le Quéré et al. 2009). Mainly thanks to CO_2 fertilization effect, the CMIP5 ensemble mean shows increasing global land CO_2 uptake between 1960 and 2005 with large year-to-year variability. The temporal variability of the land carbon is primarily driven by variability in precipitation, surface temperature and radiation, largely caused by ENSO variability (Zeng et al. 2005). Specifically, the observed land carbon sink decreases during warm climate El Niño events and increases during cold climate La Niña and volcanic eruption events (Sarmiento et al. 2009). Consistent with surface temperature results (**Figure 1**), CMIP5 models do capture the right NBP response after volcanic eruptions, but are not meant to reproduce the observed phase of ENSO variability (**Figure 5**).

The CMIP5 multi-model ensemble land-atmosphere flux (± standard deviation of the multi-model ensemble) evolved from a small source of -0.31±0.52 PgC/y over the period 1901-1930 (with a mean year-to-year variability of ±0.33 PgC/y) to a sink of 0.7±0.6 PgC/y in the period 1960-2005 (with a mean yearly variability of ±0.69 PgC/y), while GCP estimates show a weaker land sink of 0.36±1 PgC/y during the latter period. As already shown for the physical variable, the GCP IAV (±1 PgC/y) is larger than the IAV of multi-model ensemble (±0.6 PgC/y) owing to the averaging process that partially filters out the IAV.

At the regional level, the evaluation is performed against the atmospheric inversions, the GCP estimate being only global. Individual model performances reproducing the land-atmosphere CO_2 fluxes over different regions are given in **Figure 6.** The global value of land-atmosphere flux from JMA atmospheric CO_2 inversion in the period 1986-2005 is 1.17±1.06 PgC/y, with GCP showing a slightly lower global mean (0.75±1.30 PgC/y).

As shown in **Figure 6** quite a few models correctly reproduce the global land sink: in particular, MIROC-ESM (0.91±1.20 PgC/y) IPSL-CM5A-LR (0.99±1.18 PgC/y), IPSL-CM5A-MR (1.27±1.54 PgC/y), HadGEM2-CC (1.33±1.44 PgC/y), MIROC-ESM-CHEM (1.45±1.21 PgC/y), and BNU-ESM (1.55±1.37 PgC/y) simulate global NBP within the range of reference datasets. CanESM2 (0.31±2.32 PgC) underestimates the land sink, as does NorESM1-ME (-0.09±1.03 PgC/y) and CESM1-BGC (-0.23±0.78 PgC/y), these latter models showing a global carbon source in our reference period, in contradiction with the atmospheric inversion and GCP estimates. Despite showing a realistic mean uptake, GFDL-ESM2M (0.67±4.53 PgC/y) has severe problems reproducing the IAV, GFDL-ESM2G (0.72±2.58 PgC/y) showing a strong reduction in IAV compared to GFDL-ESM2M.

In the Transcom 3 inversions the Southern Hemisphere land is found to be either carbon neutral or a slight source region of CO_2 (-0.25±0.23 PgC/y) potentially due to deforestation; CMIP5 results in general put a slight carbon sink in this region and only a few of the models (IPSL-CM5A-MR, IPSL-CM5A-LR, CESM1-BGC, and MIROC-ESM) agree with observations (**Figure 6**).

Inversions place a substantial land carbon sink in the Northern Hemisphere (2.22±0.43 PgC/y), while tropical lands are a net source of carbon (-0.8±0.75 PgC/y) due to deforestation.

Looking at the Northern Hemisphere all CMIP5 models predict a CO_2 sink, despite an overall underestimation. Possible reasons for this underestimation could be the poor representation of forest regrowth from abandoned crops fields (Shevliakova et al. 2009), as well as the absence of sinks due to nitrogen deposition for most models (Dezi et al. 2010). It should also be noted that Stephens et al. (2007) found JMA having a weaker sink in the Northern Hemisphere compared to the other inversion datasets, therefore using an other inversion model from TRANSCOM would further increase the mismatch between CMIP5 models and the inversion estimates over this sub-domain.

Over the tropical region several models simulate a carbon source, i.e. CESM1-BGC (-0.24±0.55 PgC/y), MIROC-ESM (-0.24±0.79 PgC/y), NorESM1-ME (-0.11±0.74 PgC/y), and GFDL-ESM2G (-0.03±1.52 PgC/y), the rest of the ESM simulating a tropical sink, with IPSL-CM5B_LR (0.97±1.30 PgC/y) simulating the strongest carbon sink.

In **Figure 7** the seasonal evolution of simulated land-atmosphere CO_2 fluxes is compared against the JMA atmospheric inversion estimates. While at global scale and in the Northern Hemisphere only CanESM2 has serious problems reproducing the net uptake of carbon during spring and summer months due to increasing GPP over respirations and the release of carbon during autumn and winter months owing to respiration processes, in the Southern Hemisphere and in the tropics some models do not capture the right seasonal cycle. The performances of CMIP5 models are particularly poor in the tropics, where most of the models are shifted by a few months or are even anti-correlated with observations. Looking at surface climate, quite a few models do correctly reproduce the right phase of temperature and precipitation in the tropics, therefore this suggests that the poor performances reproducing the right NBP phase are not directly related with bad skills simulating surface climate. Among other possibilities, missing or coarse parameterization of harvesting, fires and LUC might helps to explain the seasonal cycle discrepancy between models and data, as well as the well known problems related to tree rooting depth (Saleska et al. 2003; Baker et al. 2008). Additionally, it should also be noted that there are no CO_2 station data in the tropics, and consequently the seasonal cycle estimates might suffer from large uncertainty (Gurney et al. 2004). It is also remarkable that in the tropics the amplitude of the NBP seasonal cycle is small, therefore it is partially expected that models do not perfectly reproduce the flat temporal evolution.

In the following, we try to identify the causes that might lead to wrong land-atmosphere CO_2 fluxes, namely we check how CMIP5 models reproduce the GPP, the LAI, and soil and vegetation carbon pools. Note that like GPP, the heterotrophic respiration (RH) is a key variables affecting NBP; however, owing to the lack of global datasets, the RH evaluation is not performed in this study.

The comparison of GPP simulated by CMIP5 models with estimates derived from FLUXNET site-level observations using a multiple tree ensemble (MTE) upscaling approach (Jung et al. 2009, 2011) shows that all the models overestimate the GPP over the period 1986-2005 (**Figure 8**). In general we can identify two groups of models: the first group has a mean global GPP value ranging from 106 to 140 PgC/y, which despite an overall overestimation is reasonably similar to the value of 119±6 PgC/y found in MTE (where 6 PgC/y is the uncertainty due to the different approaches used to estimate the MTE-GPP), and a second group that has a mean global GPP value greater than 150 PgC/y.

Using eddy covariance flux data and various diagnostic models (a similar approach used by Jung et al. 2009), Beer et al. (2010) provide an observation-based estimate of this flux at 123±8 PgC/y in the period 1998-2005 consistent with result of Jung et al. (2009), while MODIS GPP estimates (Mao et al. 2012) indicate a mean value of 114 PgC/y over the period 2000-2005. These results suggest that IPSL, GFDL and MPI models strongly overestimate the global GPP (**Figure 8**). We note that recent studies suggest that current estimates of global GPP of 120 PgC/y may be too low, and that a best guess of 150–175 PgC/y (Welp et al. 2011) or 146±19 PgC/y (Koffi et al. 2012) better reflects the observed rapid cycling of CO_2. In light of these recent results, one could suggest that the best CMIP5 models are those having a global GPP value greater than 150 PgC/y. However it is argued that Welp et al. (2011) have used only a limited number of observations and a very simple model for their studies, while Koffi et al. (2012) cannot distinguish the best estimate of 146±19 PgC/y from a different assimilation experiment yielding a terrestrial global GPP of 117 PgC/y. For such reasons our reference dataset for GPP still remains the MTE-GPP of Jung et al. (2011).

With the clear exception of high latitudes, annual GPP or LAI zonal means follow precipitation zonal distributions, i.e. more productive ecosystems are found in correspondence of precipitation maxima. Therefore, as a first approximation, the precipitation is the main limiting factor for the photosynthesis across the globe, temperature being mainly limiting at high latitudes (Piao et al. 2009). In fact too high temperatures could produce a negative effect on GPP, while a wet bias would generally be a benefit for the GPP. Looking at **Figure 2**, we can exclude that the bias in GPP is caused by a wet bias in precipitation, since the models that systematically overestimate the GPP are in fact the closer to the observed precipitation. Therefore there are other reasons explaining the systematic overestimation of global mean GPP in all the CMIP5 models. Firstly, most of these models do not consider nutrient limitation on GPP (Zaehle et al. 2010; Goll et al. 2012); it should be noted that the few models simulating the N cycling are the closer to the reference data. Second, the parameterization of the impact of tropospheric ozone on reducing GPP is not implemented yet in the models; Sitch et al.

(2007) and Wittig et al. (2009) quantified that ozone leads to a mean global GPP reduction of about 20% during the historical period as compared with a simulation without elevated tropospheric ozone.

Finally the original FLUXNET stations data sets used in the MTE approach are affected by uncertainties originating from u* filtering (Papale et al. 2006), gap-filling (Moffat et al. 2008), and flux partitioning (Reichstein et al. 2005; Lasslop et al. 2009). In addition, uncertainties increase when extrapolating to the globe, which also carries uncertainties related to the accuracy and spatial-temporal consistency of global forcing data (Jung et al. 2011).

A further comparison with results from different process-based terrestrial carbon cycle models forced offline by observed climate (i.e. CRU) shows that the land surface components of the CMIP5 ESMs still overestimate the GPP when forced by observations. Specifically, Piao et al (2013) found that the global terrestrial GPP averaged across 10 models forced by observed climate is 133±15 PgC/y, with ORCHIDEE and CLM4 having a mean global GPP of 151±4 PgC/y over the period 1982-2008, and TRIFFID showing a global GPP of about 140 PgC/y, consistent with our results from the IPSL-CM5 models, CESM1-BGC and the HadGEM2 models respectively. Since TRIFFID does not show any relevant bias reduction between the online and offline version and although the bias in ORCHIDEE is slightly lowered when forced by observed climate, we can exclude that the coupling generates this large bias in GPP.

Looking at the interannual variability of GPP, in the tropics and in the Northern Hemisphere no model captures the IAV of the observation based product, all models simulating larger GPP IAV that the one given by the MTE-GPP. Several models show relatively good performances in the Southern Hemisphere despite none of these models show a MVI value close to the good performance threshold of 0.5 defined by Scherrer (2011). The poor performances found in the tropics and in the Northern Hemisphere affect the global MVI and all the models show a MVI larger than 3.

However, it is worth seriously questioning the realism of the MTE-GPP product regarding its magnitude of interannual variability and in particular in the tropics (Zhao and Running 2010). Most of the MTE GPP sensitivity to temperature and precipitation is learned from the spatial variability of the FLUXNET data, not its interannual variability. Also, there are virtually no FLUXNET sites in the tropics to train the MTE product. The MTE tropical temporal variability is hence derived from the spatial variability of temperate ecosystems. Hence, we prefer not to use the MTE-GPP IAV as a target for CMIP5 models' evaluation.

All models predict a significant increase in vegetation productivity at global scale from 1986 to 2005, although the magnitude of the trend from all the CMIP5 models (ranging from 0.2 PgC/y^2 to 0.66 PgC/y^2) is significantly larger than MTE estimates (0.09 PgC/y^2). Again, one could question the MTE-GPP trend as atmospheric CO_2 fertilization was not explicitly accounted for in MTE-GPP framework. Also, the MTE-GPP trend may be affected by changing satellite products of vegetation activity before and after 1998. Hence, we prefer not to use the MTE-GPP trend as a target for CMIP5 models' evaluation.

In the Southern Hemisphere almost all CMIP5 models do not show any relevant increase in vegetation productivity, being the trend scattered around zero, while over the Northern Hemisphere and tropics all the models exhibit a positive trend in GPP.

In **Figure 9** we compare the phase of the mean annual cycle of CMIP5 models with the GPP from the MTE dataset. At global scale, all the CMIP5 models correctly reproduce the phase of the seasonal cycle of GPP. In particular, over the globe and Northern Hemisphere the CMIP5 models capture the GPP minimum during winter and fall and the summer GPP maximum related to the spring leaf out and maximum growing season, while in the Southern Hemisphere, the models reproduce the phase of the winter GPP minimum. Several problems are found in the tropical regions and only a few of the models (BCC-CSM1, INMCM4, HadGEM2-ES, and NorESM1-ME) are able to accurately reproduce the phase of the GPP seasonal cycle in this region. IPSL-CM5A-LR and IPSL-CM5A-MR models, indeed, show in the Northern Hemisphere (and a global scale as well) a strong positive bias of GPP during JJA. Since the evaluation of precipitation does not show a coincident wet bias, this suggest that the land surface component of the IPSL models overestimates the GPP in summer, maybe because this model does not have N-limitations or because the water stress is not strong enough during the peak growing season.

The comparison of simulated LAI with a global data set derived from satellite data is presented in **Figure 10**. However, before describing model's deficiencies we would highlight that there are several limitations in the satellite observations that could explain the mismatch between the LAI data set and CMIP5 results.

The remote sensing LAI products are estimates derived from top-of-the-atmosphere reflectances, and use different sensors and algorithms (Los et al. 2000; Myneni et al. 2002). Therefore, the quality of LAI retrievals is limited by the intrinsic characteristics of the sensor systems, the dynamic of the signal received at the satellite level, and the physical properties of the target (Gibelin et al. 2006). For instance, cloud cover hides the surface and produces discontinuities in time series. In addition, the layers of a vegetation canopy cast shadow and LAI of lower layers near the ground may not be well documented. This may yield a 30% underestimation in the case of clumped canopies (Roujean and Lacaze 2002). This occurs mostly for dense forested areas and fully developed crops. On the other hand, over semiarid ecosystems, soil brightness contaminates sufficiently the signal to restrict its sensitive response to LAI increase. Similarly, high reflectance of snow may hamper an accurate LAI retrieval at high latitudes at springtime (Gibelin et al. 2006).

Similarly to the temperature, precipitation, and GPP evaluation, the overall behaviour of CMIP5 models reproducing the LAI is analyzed by comparing the yearly mean simulated value with the satellite-derived data set. In **Figure 10** we present, for each model, the mean LAI, the trend, and the MVI computed in the period 1986-2005 for different sub-domains.

Looking at the mean global value, only INMCM4 and CanESM2 models capture the main features of the global pattern, while all the remaining models overestimate the global LAI. Serious problems have been found in BNU-ESM and GFDL models, all showing a global LAI above 2.4, while the reference values is much lower (1.45). We found BNU-ESM having severe problems in reproducing the right amplitude of LAI in the tropics (**Figure 10**) and the GFDL models completely unable to reproduce the eastward gradient over Europe and Asia, as well as overestimating the LAI in North America (Anav et al 2013). Consequently as shown in **Figure 10** in the Northern Hemisphere GFDL-ESM2G and GFDL-ESM2M are far outliers and the global result is affected by this erroneous pattern. This problem is likely due to the initialization of the vegetation during the spin up phase: in fact the GFDL land model only allows coniferous trees to grow in cold climates, i.e. deciduous trees and grass do not grow in these cold regions. As a result, coniferous trees are established in areas where there should be tundra or cold deciduous trees (Anav et al 2013). Additionally, since all CMIP5 models were spun up for many thousands of years, in case of GFDL models the coniferous vegetation eventually builds up high LAI. It is also noteworthy that this positive bias in LAI does not significantly affect the GPP in the Northern Hemisphere (**Figure 8**).

Over the Southern and Northern Hemispheres as well as in the tropical bounds we found a general tendency by CMIP5 models to overestimate the LAI and only a few models are close to the observation.

There are several reasons to explain the large overestimation of LAI by CMIP5 models. First, the high GPP could lead to a surplus of biomass stored into the leaves. Also the missing parameterization of ozone partially explains the LAI overestimation due to the GPP: specifically Wittig et al. (2009) and Anav et al. (2011) found that ozone leads to a mean global LAI reduction of about 10-20% during the historical period as compared with a simulation without elevated tropospheric ozone. Finally, as the LAI dataset does not come out from true observations we cannot exclude that it is affected by a significant bias. However, compared to other LAI datasets our reference data shows a good agreement: in particular, considering the period 2000-2005, the mean global LAI of our dataset is 1.46, while MODIS LAI (Yuan et al. 2011) shows a value of 1.49 and CYCLOPES LAI (Baret et al. 2007; Weiss et al. 2007) has a global mean slightly lower at 1.27. However, this latter dataset has some low values in dense canopies, especially evergreen broadleaf forests, which results in a lower value for the whole Earth (Zhu et al. 2013).

Considering the interannual variability, none of models are close to the good performance threshold of 0.5, the MVI being systematically larger than 2 in all the domains. On the other side, the LAI trend is well simulated by all models except BNU-ESM that largely overestimates the greening in the Northern Hemisphere and tropics, as well as by GFDL-ESM2M and IPSL-CM5A-LR which show a browning in Southern Hemisphere. Looking at global scale, most of the models do reproduce a slight greening of the same magnitude than the observed data.

The comparison of LAI seasonal cycle is given in **Figure 11**. At the global scale and in the Northern Hemisphere all the models (except GFDL) correctly reproduce the seasonal variability, namely CMIP5 models reproduce the right timing of bud-burst and leaf-out, as well as the weak leaf coverage during fall and winter. Some problems are found in the tropics and Southern Hemisphere, where some models are anti-correlated to observations. Despite that the MIROC models show a good phase of LAI compared to observations, they also show a strong positive bias during JJA in both the Hemispheres and at the global scale.

The mean global soil carbon (± ensemble standard deviation) reported across all ESMs is 1502±798 PgC, whereas the global soil carbon in the reference dataset is 1343 PgC (**Figure 12**). CESM1-BGC has the lowest total at 512 PgC and MPI-ESM-MR the highest at 3091 PgC. Looking at the global mean, most of the ESMs are clustered around the HWSD reference data (Todd-Brown et al 2012). It is also interesting to note that both CESM1-BGC and NorESM1-ME models show the lowest totals and these models both use CLM4 as land surface model (Table 2). This severe global underestimation is due by the lower carbon soil simulated in the Northern Hemisphere. On the other side, MIROC and MPI models strongly overestimate the soil carbon in all the sub-regions.

Similarly to the soil carbon results, the vegetation carbon evaluation shows that ESMs are also clustered around the reference value (**Figure 12**). The multi-model mean of global vegetation carbon (± ensemble standard deviation) reported across all ESMs is 522±162 PgC, value close to the reference data (556 PgC). At global scale MIROC and MPI models underestimate the reference value, whereas BNU-ESM reported the highest total at 927 PgC, compared to the reference data. It is also interesting to note that in the Northern Hemisphere GFDL-ESM2M shows the highest value; as already observed for the LAI, the overestimation of vegetation carbon by GFDL-ESM2M is related to the substitution of tundra with coniferous forest in the cold regions of North Hemisphere.

These results also show that CESM1-BGC and the NorESM1-ME models have a realistic vegetation carbon, indicating that the large underestimation of their soil carbon content most probably comes from an overestimation of the soil carbon decomposition rate. This might also contribute to explain the low than average NBP simulated by these two models (**Figure 6**).

3.3 CMIP5 ocean carbon

The simulated evolution of ocean-atmosphere CO_2 flux is compared with GCP estimates in **Figure 13**. Analogous to the land-atmosphere CO_2 flux (**Figure 5**), the CMIP5 models show increasing global ocean CO_2 uptake, evident from the 1940's-2005. The CMIP5 ensemble air-sea flux increased from a sink of 0.56±0.13 PgC/y (with a mean yearly variability of ±0.07 PgC/y) over the period 1901-1930 to 1.6±0.2 PgC/y in the period 1960-2005 (with a mean yearly variability of ±0.4 PgC/y). This multi-model mean is slightly lower than GCP estimates, which show an ocean sink of 1.92±0.3 PgC/y for the period 1960-2005.

During El Niño events there is a suppression of the normally strong outgassing of CO_2 in the Equatorial Pacific, and hence a larger than average global ocean sink. Keeling et al. (1995) show a much smaller effect on the atmospheric CO_2 variability from the ocean than the biosphere, however observational based estimates show contrasting results in terms of timing and magnitude of the variations in net air-sea CO_2 fluxes (Francey et al. 1995; Rayner et al. 1999). The CMIP5 ensemble mean shows a smaller variability in the ocean CO_2 uptake than in the biosphere (i.e. models agree on the sign and magnitude of ocean CO_2 fluxes), as well as it has a lower year-to-year variability than GCP estimates, partly because the interannual variability is somewhat smoothed out due to the model averaging.

The mean ocean-atmosphere CO_2 fluxes for any individual model and in each ocean sub-domain are shown in **Figure 14**. The global estimate of oceanic uptake of CO_2 from JMA inversion over the period 1986-2005 is 1.73±0.33 PgC/y, which is significantly lower than GCP estimate (2.19±0.17 PgC/y) and Takahashi estimate (2.33 PgC/y), however similar to the estimates made in the IPCC 4th assessment report (Denman et al. 2007).

At the global scale all CMIP5 models, except INMCM4, which overestimates the ocean sink with a 1986-2005 average of 2.65±0.37 PgC/y, are in the range of observational uncertainty. In particular, IPSL-CM5A-MR (2.22±0.11 PgC/y), IPSL-CM5A-LR (2.17±0.21 PgC/y), BCC-CSM1-M (2.09±0.18

PgC/y), GFDL-ESM2M (2.04±0.3 PgC/y), HadGEM2-ES (2.01±0.12 PgC/y), HadGEM2-CC (2.00±0.19 PgC/y) and MPI-ESM-LR (1.96±0.17 PgC/y) simulate values of both the global mean and interannual variability close to the observational values, while CanESM2 (1.64±0.25 PgC/y) shows the weaker CO_2 sink, and NorESM1-ME (2.32±0.15 PgC/y) well matches Takahashi estimate.

The fact that the CMIP5 models lack processes associated to the river loop of the carbon cycle, might explain why the JMA inversions give a slightly lower CO_2 uptake than the models. Although carbon fluxes from rivers are small compared to natural fluxes, they have the potential to contribute substantially to the net air-sea fluxes of CO_2 (Aumont et al. 2001)

Using oceanic inversion methods it is possible to separately estimate the natural and anthropogenic components of the air-sea CO_2 fluxes (Gruber et al. 2009). Here we consider the CMIP5 historical simulations only, and therefore all regional patterns described are largely characteristic of natural air-sea CO_2 exchanges and do not elucidate anthropogenic CO_2 uptake patterns.

At the regional scale the CMIP5 models demonstrate the expected pattern of outgassing of CO_2 in the tropics and an uptake of CO_2 in the mid and high latitudes, with comparatively small fluxes in the high latitudes. The exceptions are INMCM4, which shows an outgassing of CO_2 in the high latitude Northern Hemisphere, and CanESM2, which shows an outgassing in the high latitude Southern Hemisphere.

Inversion and Takahashi estimates show the mid-latitude Southern Ocean is a large sink of atmospheric CO_2 (Takahashi et al. 2002). Its magnitude has been estimated over the period 1986-2005 to be about 0.73±0.19 PgC/y from JMA inversion and 1.28 PgC/y from the Takahashi product (**Figure 14**). All the CMIP5 models simulate a similar magnitude sink in this region except CanESM2, which overestimates the sink (1.59±0.05 PgC/y).

The mid latitude Northern Hemisphere Ocean is also a net sink for CO_2 (Denman et al. 2007), with a magnitude of the order of 0.77±0.08 PgC/y from JMA, and 1.15 PgC/y from Takahashi over the period 1986-2005 (**Figure 14**). All the CMIP5 models, simulate a net sink, with values comparable to the JMA inversion results.

The tropical oceans outgassing of CO_2 to the atmosphere has a mean flux of the order of -0.73±0.14

PgC/y in the period 1986-2005 (**Figure 14**), estimated from JMA inversions, and a value of -1.25 PgC/y estimated from Takahashi. We find INMCM4 (1.10±0.17 PgC/y) the only model unable to reproduce the tropical source of carbon.

The seasonal air-sea CO_2 fluxes are compared against the JMA inversion estimates and the Takahashi product in **Figure 15**. All the models except INMCM4 accurately reproduce the observational based estimates in the mid latitudes. The model estimates for the tropics and high latitudes show greater ambiguity. This is attributed to large uncertainties in modelled SST, MLD and ocean NPP in the high latitude Southern Ocean, while in the equatorial region uncertainties can arise due to the lack of mesoscale processes simulated by the models. At the global scale all of the models are out of phase with the observations, and the MPI models as well as INMCM4 show a larger seasonal variation than observations. In the MPI models this is a result of the poor performance in the high latitude Southern Hemisphere where they strongly overestimate the CO_2 sink in austral summer and underestimate during austral winter.

The air-sea CO_2 flux is driven in part by the biological pump. **Figure 16** shows individual model performances at reproducing SeaWiFS based estimates of oceanic NPP in the reference ocean sub-domains. The mean global NPP estimate based on the SeaWiFS data used here during the period 1998–2005 is 52.2 PgC/y. Using CZCS chlorophyll fields Longhurst et al. (1995) estimated global NPP to be between 45-50 PgC/y, and Behrenfeld and Falkowski (1997) estimated a global rate of 43.5 PgC/y.

Globally quite a few models, except GFDLs, underestimate SeaWiFS NPP. Most of the models predict a global average of ~30-40 PgC/y. This is reasonable when compared with published chlorophyll based estimates, and considering the large uncertainty in the observational based datasets. The significant under estimation of ocean NPP by most of the CMIP5 models could occur partly due to the lack of explicit representation of coastal processes. The coarse resolution of ocean models does not allow realistic simulation of the processes taking place in these shallow waters that are naturally eutrophic because of riverine discharge, coastal upwelling and a high recycling rate of organic nutrient matter.

On the other side, the strong positive bias found in the GFDL models for ocean NPP predominantly stems from an overestimation of phytoplankton activity in the Eastern Equatorial Pacific. The GFDL SST (**Figure 3**) and MLD do not show a larger deviation from observations than other models, therefore we can exclude these two variables as the cause of the bias in this region.

Conversely, MPI models and CESM1-BGC have a global mean marine NPP most similar to that of the SeaWiFS NPP, however in the case of MPI models this is a misleading result since the agreement arises from a large overestimation of NPP in the Southern Hemisphere and an underestimation in the Northern Hemisphere. Regionally all of the model biases take a different pattern to that of the global scale. In the northern high latitudes we see that all of the models under estimate NPP whereas in the Southern Hemisphere high latitudes all the models except CanESM2, IPSL-CM5A-LR and IPSL-CM5A-MR overestimate NPP.

In all the CMIP5 models, and the SeaWiFS based estimates, zonally summed NPP is greatest in the tropics. This is simply due to a larger ocean surface area, since on average NPP is lower in the tropics and highest in Northern Hemisphere high latitudes.

Looking at the interannual variability the models in general are clustered around the reference data, albeit in the two Northern Hemisphere sub-regions larger interannual variations are seen in the reference data than in the CMIP5 models.

In **Figure 17** we show the mean annual cycle of NPP as simulated by the CMIP5 models compared with the NPP estimated from SeaWiFS data. The largest seasonal variability in the SeaWiFS based NPP is seen the Northern Hemisphere high latitudes (49N–90N) with the peak in observations occurring in July. None of the CMIP5 models capture the magnitude or timing of this significant peak in productivity, with the majority of the models biased towards lower NPP and predicting the peak in productivity up to 2 months too early. Accurate model simulations of NPP are more difficult in this ocean sub-domain since it includes a mixture of several different regions and has a large proportion of coastal areas.

Many of the models show the largest seasonal peak in marine NPP in the Southern Ocean (90S-44S), which is not supported by SeaWiFS estimates. This is due to a combination of model and observational errors. SeaWiFS observations generally underestimate surface chlorophyll in the Southern Ocean (Moore et al. 1999) and contain the largest uncertainty in the Southern Ocean due to under sampling and frequent deep chlorophyll maxima that cannot be observed on satellites. The models tend to overestimate NPP in the Southern Ocean due to too shallow simulated mixed layers in summer months and uncertainty in light parameterisations (Séférian et al. 2012). The models with the greatest overestimation of springtime NPP in the high latitude Southern Ocean are MPI models and NorESM1-ME with peak values of ~3 PgC/y compared to ~ 0.75 PgC/y for SeaWiFS based NPP estimates. All these models use the same biogeochemical model HAMOCC5 (Table 2), although with different parameterisations. It should also be noted that these latter models show the largest bias in the MLD seasonal cycle and this can contribute to the poor representation of temporal evolution of primary production.

4. MODEL RANKING

Different diagnostics were used in section 3 to investigate the performances of CMIP5 Earth System Models during the 20th century at reproducing the mean value, IAV, trends and mean annual cycle for various different variables crucial to characterizing the global carbon cycle. These measures or “diagnostics” show that in general, the CMIP5 models simulate all the variables well when compared to the observations used here, although a few of the models do show notably poorer agreement than others and general problems exist for quite a few of the models. Specifically, all the variables in the tropical regions prove to be problematic for the models, reinforcing well-known deficiencies of models in reproducing the decadal variations in the ocean-atmosphere system, but also questioning the availability and quality of the data in the tropics.

However, the diagnostics presented in sections 3 are not sufficient to clearly identify the best models; for such a purpose we need to define specific metrics that allow a quantitative model ranking. Metrics can be contrasted with ‘diagnostics’, which may take many forms (e.g., maps, time series, power spectra, errorbars, zonal means, etc.) and may often reveal more about the causes of model errors and the processes responsible for those errors. Following Gleckler et al. (2008) the metrics used in this paper are designed to quantify how much the model simulations differ from observations.

4.1 Land carbon ranking

We used two different metrics to estimate the models' skills. In case of the mean annual cycle the skill score is computed following equation 3, and the model performances and ranking of the land variables are shown in **Figure 18.** Considering the mean annual cycle in addition to this skill score, in order to check how models reproduce only the phase of the observations, we also have computed the correlation coefficient (not shown). In fact, the correlation coefficient allows to identify models that are in phase with observations (r>0), and models that are out of phase (r<0). Correlation values close to 1 point out models that perfectly reproduce the seasonal phase of observations.

Looking at the land surface temperature, at global scale and in Southern and Northern Hemisphere the best performances reproducing the mean annual cycle have been found for MPI models, CESM1-BGC, and NorESM1-ME, whilst in the tropics BNU-ESM and BCC-CSM1 have the highest scores. All the models have a correlation coefficient greater that 0.9 at global scale and in the 2 Hemispheres, while in the tropics it ranges between 0.6 and 0.8.

The precipitation shows a similar pattern, with MPI models having the best performances in all the sub-domains, except the Southern Hemisphere, where BCC-CSM1 and IPSL-CM5A-MR have the best scores (**Figure 18**).

Unlike seasonal variation in temperature, which at large scales is strongly determined by the insolation pattern, seasonal precipitation variations are strongly influenced by vertical movement of air due to atmospheric instabilities of various kinds and by the flow of air over orographic features. For models to simulate accurately the seasonally varying pattern of precipitation, they must correctly simulate a number of processes (e.g. evapotranspiration, condensation, transport) that are difficult to evaluate at a global scale (Randall et al. 2007). The precipitation exhibits a correlation never exceeding a value of 0.8 in all the sub-domains and for all the models, with the lowest value (0.4) found in the Northern Hemisphere for the BNU-ESM model (not shown).

Looking at the GPP, at global scale CESM1-BGC shows the best performances, albeit its GPP decrease during fall does not match the phase of observation (**Figure 9**). In fact, for a given seasonal skill score it is impossible to determine how much of the error is due to a difference in structure and phase and how much is simply due to a difference in the amplitude of the variations. Also in the Southern Hemisphere and Tropics CESM1-BGC has the highest scores for the GPP, while in the Northern Hemisphere the best results are found in BCC-CSM1-M.

Looking at the phase of GPP there is a relevant agreement with the reference data, the correlation being systematically positive. This is particularly evident in the Northern Hemisphere where all the models have a correlation above 0.8 (not shown). Contrarily, in the Tropics there is a poorer agreement and some models (e.g. CanESM2, and IPSL-CM5B-LR) show a correlation around 0.4 (not shown).

The same considerations drawn for the GPP are also valid for the LAI, with CanESM2 showing the best skills at global scale, although it seems to be 2 months out of phase with respect to observations during the peak season (**Figure 11**). In addition, all the models show a correlation greater than 0.6 both at global scale and in the Northern Hemisphere, while in the Tropics we found the poorest results with some models (BNU-ESM, BCC-CSM1, and BCC-CSM1-M) having a correlation of about 0.2.

Considering the global NBP, consistent with results of **Figure 7,** MPI-ESM-LR and MIROC-ESM have the best performances, whilst CanESM2, BNU-ESM, MPI-ESM-MR, and CESM1-BGC show the poorest scores. Contrarily, in the Southern Hemisphere CESM1-BGC and CanESM2 have the highest scores, while in the Tropics the 2 Hadley models show the best results.

Several models show a negative correlation compared to inversion estimates in the Tropical region and in the Southern Hemisphere, while in the Northern Hemisphere quite a few models have a correlation above 0.9 (not shown).

The second skill score is computed following equation 5, and it essentially allows to asses the skills of models in reproducing the mean state of the system with its IAV. **Figure 19** shows an absolute measure of ESMs skill in simulating the observed PDFs of the variables under examination for the land carbon. There is no obvious way to define ‘good’ or ‘bad’ performance, or indeed, ‘adequate’ from the skill score, but identifying those models with a relatively better skill is straightforward.

According to the skill threshold defined in Section 2.3, looking at global temperature, only few models are close to the threshold value of 0.68. Consistent with **Figure 1**, the best performances have been found in the MPI models, while the poorest skills are found in INMCM4. The same considerations are valid also for the Southern and Northern Hemisphere. Looking at the Tropics, consistent with **Figure 1**, INMCM4 shows a very poor skill, related to the large cold bias previously described. Unlike **Figure 1,** the skill score shows that BCC-CSM1 is not the best model in the Tropical region. This results however is not surprising, the agreement in the mean tropical temperature shown in **Figure 1** could arise from a compensation between overestimation in some regions of the tropics and underestimation in other regions of the tropics, while the skill score does not lead to the same optimistic picture. In fact the overlapping of the PDFs allows equal weighting of all the points with a relevantly poor mismatch to the mean value. This suggests that the models we found using the previous diagnostics that have a bias in the mean values still score badly, but models with a good agreement with the mean do not necessarily score well.

The precipitation shows the same picture of temperature with a general good agreement in the Southern and Northern Hemisphere and poorer skills in the Tropical region, likely related to the poor skill reproducing the IAV (**Figure 2**). Relevant skills are found in the Southern Hemisphere for the Hadley models, where the overall score is greater than 0.7.

Contrarily, very poor skills are found for GPP and LAI, both a global scale and in all the sub-domains. In **Figure 8** and **Figure 10**, respectively, we show how almost all CMIP5 models overestimate these two variables, possibly because these models do not have nutrient limitations and any ozone impact on carbon assimilation. Consequently none of models achieve a relevant score, and for quite a few models the skill score is less than 0.3. As pointed out before, we cannot exclude risks of significant bias in the GPP and LAI evaluation datasets as these are not true observations.

Unlike other variables related to the land carbon cycle, good scores are found for the NBP. As already shown in **Figure 6** most of the models match both the mean value and the IAV, therefore, except

GFDL-ESM2M that significantly overestimates the IAV, at global scale we found a score above 0.5 for all the models, with the best result found in IPSL-CM5A-LR that simulates more than 2σ of the reference PDF. Conversely, none of the models are able to simulate the observed PDF for the NBP in the Northern Hemisphere, and this is consistent with the negative bias already shown in **Figure 6**. However it should also be noted that the NBP PDFs are build from regional averages, while other variables are based on the comparisons of skills at each grid point, then averaged over large sub-regions; this explains why the NBP skill scores are consistently better than the scores of the other variables.

In case of soil and vegetation carbon the skill scores reported in **Figure 19** are not based on the PDF overlapping, but they have been computed as a relative bias. Results in general agree with finding of **Figure 12**, namely the best results for the soil carbon are found in BCC models, while MIROC and MPI models show the poorest performances due to the large positive bias. Considering the vegetation carbon, INMCM4 has the best skill score, while BNU-ESM and GFDL-ESM2M show the poorest performances. The only exception is the Tropical region, where the best model reproducing the vegetation carbon is MPI-ESM-MR, with BNU-ESM still showing the poorest results.

4.2 Ocean carbon ranking

The skills of CMIP5 models at reproducing the mean annual cycle of relevant variables for the ocean carbon cycle are shown in **Figure 20.**

Considering the SST, there is a large variability in the skill score of models between the different sub-domains; in general, the best results are found for CanESM2, CESM1-BGC and MPI models, while BNU-ESM and GFDL models show the poorest skills. Consistent with results of **Figure 4,** the Hadley models show the best performances at reproducing the mean annual cycle of the MLD, with the MPI models having the poorest skill scores (**Figure 20**).

We also have found excellent performances of CMIP5 models in reproducing the only phase of the mean annual cycle of physical variables (i.e. SST and MLD), with correlations above 0.85 for all the models and sub-domains (not shown).

As discussed previously, the poor performances of the MPI models in reproducing the seasonal evolution of the MLD also affect the overall skill score of the ocean-atmosphere CO_2 fluxes; in particular, we found the MPI models having the worst performances at global scale, as a consequence of the poor results found in the extreme Southern Ocean, whilst in the tropical bound and in the 2 Northern Hemisphere sub-domains the MPI models show a relevant skill in reproducing the CO_2 fluxes (**Figure 20**).

Nevertheless, severe problems exist in reproducing the only phase of global seasonal cycle of CO_2 fluxes, where several models are anti-correlated with observations. The poor performances in the global values are caused by the inability of models in simulating the correct seasonal cycle in the tropical sub-domain as well as in the high-latitude Southern and Northern Oceans. Conversely, in the mid-latitude Southern and Northern Oceans, except INMCM4, all the models are positively correlated with JMA inversions and the correlation coefficient is generally higher than 0.7 (not shown).

Considering the ocean primary production the best performances have been found for CESM1-BGC and IPSL models, while the worst results are found for the MPI models and NorESM1-ME. It should be noted that all these models use the same ocean biogeochemical model (**Table 2**). Conversely, with the only exception of CanESM2, all the models show a relevant correlation with SeaWIFS data in all the sub-domains (not shown).

Considering the PDF-based skill score, consistent with land surface temperature and precipitation results, the SST skill score for several models is above the threshold of 1σ, with some models having a score above 0.8 (**Figure 21**). This is particularly evident in the temperate Southern and Northern Oceans as well as in the tropics. Although the models exhibit relevant skills at reproducing the SST in some basins, in the Northern and Southern Ocean none of the model is able to reproduce at least 1σ of the reference dataset.

Since the observed MLD is a climatology, the ranking is tricky and the values shown in **Figure 21** do not represent the skill score defined in section 2. Therefore, for this variable only the ranking is based on the bias rather than on the overlapping of the PDFs. Globally, we found HadGEM2-ES and HadGEM2-CC the best models at reproducing the MLD, and NorESM1-ME is found to have the largest bias in all the sub-domains, except in the Southern Ocean where MPI models show the worst agreement to the observations.

The ocean-atmosphere CO_2 flux shows an acceptable skill score for most of the models; however it should be noted that likewise the NBP also the ocean-atmosphere CO_2 flux PDFs are based on regional comparisons. Globally several models have a score higher than 0.7, and only IPSL-CM5A-MR, INMCM4, and NorESM1-ME show poor performances. As already seen in **Figure 14**, the poor skill found in INMCM4 at global scale is due to the poor performances of this model to correctly reproduce the fluxes in the tropical regions (18S-18N) and in the Northern Hemisphere. Therefore, consistent with results of **Figure 14** INMCM4 shows the poorest performances in these sub-domains. Conversely, INMCM4 has the best performances in the temperate Southern Hemisphere where it is able to reproduce almost 2σ of the observed PDF.

As we previously discussed, the simulated global ocean primary production is affected by a negative (or positive for GFDL models and MPI-ESM-LR) bias, consequently the skill score does not exceed a value of 0.4. The same considerations are also valid for the other sub-domains, and the only relevant performances are found in the Southern Hemisphere where several models show a skill score above 0.6. In previous sections we speculated that the ocean primary production underestimation by models is likely due to a coarse resolution of the ocean grids that does not allow to properly simulate the dynamics in the shallow waters; the good performances found in the Southern Ocean would support this assumption.

5. CONCLUSION

In this study the evaluation of the CMIP5 ESMs focused on the ability of the models to reproduce the seasonal cycle, the mean state with its interannual variability, and trends of land and ocean variables related to the carbon cycle. This task allows the identification of the strengths and weaknesses of individual coupled carbon-climate models as well as identification of systematic biases of the models.

We have highlighted that the evaluation is partly subjective due to the choice of the variables. In this paper we focused only on the validation of carbon fluxes and main variables affecting the fluxes, however many more data (e.g. DIC, pCO_2, chlorophyll concentration) could be used to evaluate the ESMs

Multi-model databases offer both scientific opportunities and challenges. One challenge is to determine whether the information from each individual model in the database is equally reliable, and should be given equal ‘‘weight’’ in a multi-model detection and attribution study (Santer et al. 2009).

We used a skill score based on the overlapping of PDFs, and the centered RMS error for the model ranking. In general we found that the ranking is sensitive to the large latitudinal bounds and the variable under examination, i.e. models that poorly perform in some sub-domains could have relevant skills in other sub-domains.

Although both the skill scores identify some models as having the best global performances, several criticisms must be noted.

Firstly, the evaluation presented here is partly subjective due to the choice of the variables, and these are sensitive to the choice of reference data. In other words, the best models for our reference variables might have poor performances reproducing other variables of interest. This suggests, therefore, 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.

Secondly, we did not account for the uncertainty in the reference data; in general for the physical variables it is expected that errors remain much smaller than the errors in the models, but in case of biological variables this is not true. However, we believe that considering the uncertainties in the observed datasets does not significantly change our model ranking, except for land GPP interannual variability and ocean NPP that might suffer large uncertainty in the mean value. For instance, Gregg and Casey (2004) report an uncertainty in the ocean primary production of about 30%, and considering this uncertainty the model ranking could significantly differ from our results.

In addition the observations used in this study do not always come from direct measurements, and in the case of biological variables some models or algorithms have been used to retrieve the values used in this study. This suggests that additional uncertainty should be added to the reference data, or in some case (e.g GPP trend) the data should simply not be used in the model evaluation.

Thirdly, the aggregation of regions can give distorted results. The choice of regions in itself affects the outcome of the regional metrics calculated, but also affects the global result through neutralising or enhancing regional outcomes when Northern and Southern hemispheres are combined.

In addition, the skill scores could be sensitive to the spatial scale. Considering 22 coupled ocean-atmosphere general circulation models (OAGCMs), Gleckler et al (2008) have evaluated the impact of alternative reference data set, other available realizations, and different resolution grids to the final ranking, finding that ‘*‘in some cases these variations on our analysis choices lead to small differences in a model’s relative ranking, whereas in others the differences can be quite large. Rarely, however, would the model rank position change by more than 5 or 6”.*

In order to cross check the sensitivity of the skill score to resolution, we regridded the surface temperature to 4 different resolutions (i.e. 0.5, 1, 1.5, and 2 degrees), finding that the resolution does not significantly affect the ranking. Best models and poor models are always the same for all the resolutions, and in general the model rank position does not change by more than 4 (not shown).

Fourthly, considering the model ranking, one could argue that choosing the highest score would favour models with more than one realization. However we also produced alternative rankings using either only the first realization from all the models, or computing the mean skill score averaged over the available realizations. We found no relevant differences in the model ranking between the three different methods (not shown).

Lastly, a PDF-derived skill-score is a useful means of evaluating models since skill in this measure implies an ability to simulate a range of behaviour (e.g., mean, IAV, trend), however, we do not argue that the skill metrics used in this paper are definitive nor do these identify models that are more predictive. We believe that it is a substantial advance on the assessment of climate and carbon cycle models skill, but as with all statistics, must be interpreted with a degree of caution so as to avoid misleading assertions.

Acknowledgements

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

We also thank G. Maze for providing blended monthly time series of ocean NPP (http://data.guillaumemaze.org/ocean_productivity), the three anonymous reviewers and Dr. Anand Gnanadesikan for their helpful comments. A special thank to Tomohiro Hajima, Dr. Hideki Okajima, and Spencer Liddicoat for providing additional datasets not available on PDMDI server, to Philippe Peylin for sharing TRANSCOM3 inversion results, and Markus Reichstein for the several clarifications. This work was supported by the European Commission's 7th Framework Programme under grant agreement 238366 (GREENCYCLESII) and 282672 (EMBRACE), while Dr. Jones was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

REFERENCES

Adler, R.F., and Coauthors, 2003: The Version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeor.*, **4**, 1147-1167.

Allan, R. P., and B. J. Soden, 2007: Large discrepancy between observed and simulated precipitation trends in the ascending and descending branches of the tropical circulation. *Geophys. Res. Lett.*, **34**, L18705, doi:10.1029/ 2007GL031460.

Anav, A., L. Menut, D. Khvorostyanov, and N. Viovy, 2011: Impact of tropospheric ozone on the Euro-Mediterranean vegetation. *Global Change Biol.*, **17**, 2342–2359.

Anav, A., G. Murray-Tortarolo, P. Friedlingstein, S. Sitch, S. Piao, Z. Zhu, 2013: Evaluation of DGVMs in Reproducing Satellite Derived LAI over Northern Hemisphere. Part II: Earth System Models. *Remote Sens*., in prep.

Arora, V. K., J. F. Scinocca, G. J. Boer, J. R. Christian, K. L. Denman, G. M. Flato, V. V. Kharin, W. G. Lee, and W. J. Merryfield, 2011: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys. Res. Lett.*, **38**, L05805, doi:10.1029/2010GL046270.

Aumont, O., J. C. Orr, P. Monfray, W. Ludwig, P. Amiotte-Suchet, J.-L. Probst, 2001: Riverine-driven interhemispheric transport of carbon. *Global Biogeochem. Cycles*, **15**, 393–405.

Baker, D. F., and Coauthors, 2006: Transcom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003. *Global Biogeochem. Cycles*, **20**, GB1002, doi:10.1029/2004GB002439.

Baker, I. T., L. Prihodko, A. S. Denning, M. Goulden, S. Miller, and H. R. da Rocha, 2008: Seasonal drought stress in the Amazon: Reconciling models and observations. *J. Geophys. Res.*, **113**, G00B01, doi:10.1029/2007JG000644.

Baret, F., and Coauthors, 2007: LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION: Part 1: Principles of the algorithm. *Remote Sens. Environ.*, **110**, 275-286.

Beer, C., and Coauthors, 2010: Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science*, **329**, 834–838.

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

Cadule, P., P. Friedlingstein, L. Bopp, S. Sitch, C. D. Jones, P. Ciais, S. L. Piao, and P. Peylin, 2010: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Global Biogeochem. Cycles*, **24**, GB2016, doi:10.1029/2009GB003556.

Chen, W., Z. Jiang, and L. Li, 2011: Probabilistic projections of climate change over China under the SRES A1B scenario using 28 AOGCMs. *J. Climate*, **24**, 4741-4756.

Chou, C., and C.-W. Lan, 2012: Changes in the annual range of precipitation under global warming. *J. Climate*, **25**, 222-235.

Chylek, P., Li, J., Dubey, M. K., Wang, M., and Lesins, G., 2011: Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmos. Chem. Phys. Discuss.*, **11**, 22893-22907.

Collins, W., and Coauthors, 2006: The community climate system model version 3 (CCSM3). *J. Climate*, **19**, 2122–2143.

Collins, W., and Coauthors, 2011: Development and evaluation of an earth-system model-HadGEM2. *Geosci. Model Dev.*, **4**, 1051–1075.

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

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: formulation and simulation characteristics. *J. Climate*, **19**, 643–674.

Denman, K. L., and Coauthors, 2007: Couplings between changes in the climate system and biogeochemistry. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 589–662.

Dezi, S., B. E. Medlyn, G. Tonon, and F. Magnani, 2010: The effect of nitrogen deposition on forest carbon sequestration: a model-based analysis. *Global Change Biol.*, **16**, 1470–1486.

Dima, M., and G. Lohmann, 2010: Evidence for two distinct modes of large-scale ocean circulation changes over the last century. *J. Climate*, **23**, 5–16.

Dufresne, J.-L., and Coauthors, 2012: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Clim. Dyn.*, under review.

Dunne, J. P., and Coauthors, 2012: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics. *J. Climate*, **25**, 6646–6665.

Engelen, R. J., A. S. Denning, and K. R. Gurney, 2002: On error estimation in atmospheric CO2 inversions. *J. Geophys. Res.*, **107**, 4635, doi:10.1029/2002JD002195.

Errasti, I., A. Ezcurra, J. Sáenz, and G. Ibarra-Berastegi, 2011: Validation of IPCC AR4 models over the Iberian Peninsula. *Theor. Appl. Climatol.*, **103**, 61–79.

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012: Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Francey, R. J., P. P. Tans, C. E. Allison, I. G. Enting, J. W. C. White, and M. Trolier, 1995: Changes in oceanic and terrestrial carbon uptake since 1982. *Nature*, **373**, 326-330.

Gibbs, H. K: 2006: Olson's Major World Ecosystem Complexes Ranked by Carbon in Live Vegetation: An Updated Database Using the GLC2000 Land Cover Product (available at http://cdiac.ornl.gov/epubs/ndp/ndp017/ndp017b.html).

Gibelin, A.-L., J.-C. Calvet, J.-L. Roujean, L. Jarlan, and S. O. Los, 2006: Ability of the land surface model ISBA-A-gs to simulate leaf area index at the global scale: Comparison with satellites products. *J. Geophys. Res.*, **111**, D18102, doi:10.1029/2005JD006691.

Gillett, N. P., P. A. Stott, and B. D. Santer, 2008: Attribution of cyclogenesis region sea surface temperature change to anthropogenic influence. *Geophys. Res. Lett.,* **35**, L09707, doi: 10.1029/2008GL033670.

GLC2000. Global Land Cover 2000 database. European Commission, Joint Research Centre, 2003 (available at http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php).

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models, *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

Goll, D. S., V. Brovkin, B. R. Parida, C. H. Reick, J. Kattge, P. B. Reich, P. M. van Bodegom, and Ü. Niinemets, 2012: Nutrient limitation reduces land carbon uptake in simulations with a model of combined carbon, nitrogen and phosphorus cycling. *Biogeosciences Discuss.*, **9**, 3173–3232.

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

Gruber, N., and Coauthors, 2009: Oceanic sources, sinks, and transport of atmospheric CO2. *Global Biogeochem. Cycles*, **23**, GB1005, doi:10.1029/2008GB003349.

Gurney, K. R., and Coauthors, 2002: Towards robust regional estimates of CO2 sources and sinks using atmospheric transport models. *Nature* **415**, 626–630.

Gurney, K. R., and Coauthors, 2003: Transcom 3 CO2 Inversion Intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information. *Tellus*, **55B**, 555–579.

Gurney, K. R., and Coauthors, 2004: Transcom 3 inversion intercomparison: model mean results for the estimation of seasonal carbon sources and sinks. *Global Biogeochem. Cycles*, **18**, GB1010, doi:10.1029/2003GB002111.

Gurney, K. R., D. Baker, P. Rayner, and S. Denning, 2008: Interannual variations in continental-scale net carbon exchange and sensitivity to observing networks estimated from atmospheric CO2 inversions for the period 1980 to 2005. *Global Biogeochem. Cycles,* **22**, GB3025, doi:10.1029/2007GB003082.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1107.

Houghton, R. A., 2000: Interannual variability in the global carbon cycle. *J. Geophys. Res.*, **105**, 20121–20130.

Kaminski, T., P. J. Rayner, M. Heimann, and I. G. Enting, 2001: On aggregation errors in atmospheric transport inversions. *J. Geophys. Res.*, **106**, 4703–4715.

Keeling, C.D., T. P. Whorf, M. Wahlen, and J. van der Plicht, 1995: Interannual extremes in the rate of rise of atmospheric carbon dioxide since 1980. *Nature*, **375**, 666-670.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl, 2010: Challenges in Combining Projections from Multiple Climate Models. *J. Climate*, **23**, 2739–2758.

Koffi, E. N., P. J. Rayner, M. Scholze, and C. Beer, 2012: Atmospheric constraints on gross primary productivity and net ecosystem productivity: Results from a carbon-cycle data assimilation system. *Global Biogeochem. Cycles*, **26**, GB1024, doi:10.1029/2010GB003900.

Jacob D., A. Elizalde, A. Haensler, S. Hagemann, P. Kumar, R. Podzun, D. Rechid, A. R. Remedio, F. Saeed, K. Sieck, C. Teichmann, C. Wilhelm, 2012: Assessing the Transferability of the Regional Climate Model REMO to Different COordinated Regional Climate Downscaling EXperiment (CORDEX) Regions. *Atmosphere*, **3**, 181-199.

Jobbagy, E. G., and R. B. Jackson, 2000: The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications*, **10**, 423–436.

John, V. O., R. P. Allan, and B. J. Soden, 2009: How robust are observed and simulated precipitation responses to tropical ocean warming? *Geophys. Res. Lett.*, **36**, L14702, doi:10.1029/2009GL038276.

Johns, T. C., and Coauthors, 2006: The new Hadley centre climate model (HadGEM1): evaluation of coupled simulations. *J. Climate*, **19**, 1327–1353.

Johnson, F., S. Westra, A. Sharma, and A. J. Pitman, 2011: An Assessment of GCM Skill in Simulating Persistence across Multiple Time Scales. *J. Climate*, **24**, 3609–3623.

Jones, C. D., and Coauthors, 2011: The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geosci. Model Dev.*, **4**, 543-570

Jung, M., M. Reichstein, and A. Bondeau, 2009: Towards global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, **6**, 2001–2013.

Jung, M., and Coauthors, 2011: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.*, **116**, G00J07, doi:10.1029/2010JG001566.

Lasslop, G., M. Reichstein, D. Papale, A. Richardson, A. Arneth, A. Barr, P. Stoy, and G. Wohlfahrt, 2009: Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. *Global Change Biol.*, **16**, 187–208.

Le Quere, C., and Coauthors, 2009: Trends in the sources and sinks of carbon dioxide. *Nat. Geosci.*, **2**, 831–836.

Liepert, B. G., and M. Previdi, 2009: Do Models and Observations Disagree on the Rainfall Response to Global Warming? *J. Climate*, **22**, 3156–3166.

Lin, J. L., 2007: Interdecadal variability of ENSO in 21 IPCC AR4 coupled GCMs. *Geophys. Res. Lett.*, **34**, L12702, doi:10.1029/2006GL028937.

Lin, J. L., K. M. Weickmann, G. N. Kiladis, B. E. Mapes, S. D. Schubert, M. J. Suarez, J. T. Bacmeister, and M. I. Lee, 2008: Subseasonal variability associated with Asian Summer Monsoon simulated by 14 IPCC AR4 coupled GCMs. *J. Climate*, **21**, 4541–4567.

Los, S. O., P. J. Sellers, G. J. Collatz, R. S. DeFries, C. J. Tucker, N. H. Pollack, D. A. Dazlich, and L. Bounoua, 2000: A global 9-year biophysical land surface dataset from NOAA AVHRR data. *J. Hydrometeor.*, **1**, 183–199.

Lucarini, V., S. Calmanti, A. Della Aquila, P. M. Ruti, and A. Speranza, 2007: Intercomparison of the northern hemisphere winter mid-latitude atmospheric variability of the IPCC models. *Climate Dyn.*, **28**, 829-848.

Mao, J., P. Thornton, X. Shi, M. Zhao, and W. Post, 2012: Remote sensing evaluation of CLM4 GPP for the period 2000 to 2009. *J. Climate*, **25**, 5327–5342.

Marti, O., and Coauthors, 2010: Key features of the IPSL ocean atmosphere model and its sensitivity to atmospheric resolution. *Climate Dyn.*, **34**, 1–26.

Martin, G. M., and Coauthors, 2011: The HadGEM2 family of Met Office Unified Model climate configurations. *Geosci. Model Dev.*, **4**, 723-757.

Martinez, E., D. Antoine, F. D’Ortenzio, and B. Gentili, 2009: Climate-driven basin-scale decadal oscillations of oceanic phytoplankton. *Science*, *326*, 1253–1256.

Maxino, C. C., B. J. McAvaney, A. J. Pitman, and S. E. Perkins, 2008: Ranking the AR4 climate models over the Murray Darling Basin using simulated maximum temperature, minimum temperature and precipitation. *Int. J. Climatol.*, **28**, 1097–1112.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon, Eds., Cambridge University Press.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, **25**, 693–712.

Moffat, A. M., and Coauthors, 2007: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agric. For. Meteorol.*, **147**, 209–232.

Moise, A. F., and F. P. Delage, 2011: New climate model metrics based on object orientated pattern matching of rainfall. *J. Geophys. Res.*, **116**, D12108, doi:10.1029/2010JD015318.

Moore, J. K., and Coauthors, 1999: SeaWiFS satellite ocean color data from the Southern Ocean. *Geophys. Res. Lett.*, **26**, 1465−1468.

Myneni, R., S. Hoffman, J. Glassy, Y. Zhang, P. Votava, R. Nemani, S. Running, and J. Privette, 2002: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sens. Environ.*, **83**, 214–231.

New, M., D. Lister, M. Hulme, and I. Makin, 2002: A high-resolution data set of surface climate over global land areas. *Climate Res.*, **21**, 1–25.

Olson, J. S., J. A. Watts, and L. J. Allison, 1985: Major World Ecosystem Complexes Ranked by Carbon in Live Vegetation (NDP-017), available at: (http://cdiac.ornl.gov/ndp017.html) from the Carbon Dioxide Information Center, U.S. Department of Energy, Oak Ridge National Laboratory, Oak Ridge TN.

Papale, D., M. Reichstein, M. Aubinet, E. Canfora, C. Bernhofer, W. Kutsch, B. Longdoz, S. Rambal, R. Valentini, T. Vesala, and D. Yakir, 2006: Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. *Biogeosciences*, **3**, 1–13.

Perkins, S. E., A. J. Pitman, N. J. Holbrook, J. McAneney, 2007: Evaluation of the AR4 Climate Models’ Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. *J. Climate*, **20**, 4356–4376.

Piao, S., P. Ciais, P. Friedlingstein, N. de Noblet-Ducoudre, P. Cadule, N. Viovy, and T. Wang, 2009: Spatiotemporal patterns of terrestrial carbon cycle during the 20th century. *Global Biogeochem. Cycles*, **23**, GB4026, doi:10.1029/2008GB003339.

Piao, S., and Coauthors, 2013: Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends. *Global Change Biol.*, In press.

Radić, V., and G. K. C. Clarke, 2011: Evaluation of IPCC Models' Performance in Simulating Late-Twentieth-Century Climatologies and Weather Patterns over North America. *J. Climate*, **24**, 5257–5274.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi, 2010: Weighting of model results for improving best estimates of climate change. *Climate Dyn.*, **35**, 407–422.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, doi: 10.1029/2002JD002670.

Rayner, P., I. Enting, R. Francey, and R. Langenfelds, 1999: Reconstructing the recent carbon cycle from atmospheric CO2, d13C, and O2/N2 observations. *Tellus*, **51**, 213–232.

Rayner, P., R. M. Law, C. E. Allison, R. J. Francey, C. M. Trudinger, and C. Pickett-Heaps, 2008: Interannual variability of the global carbon cycle (1992 – 2005) inferred by inversion of atmospheric CO2 and d13CO2 measurements. *Global Biogeochem. Cycles*, **22**, GB3008, doi:10.1029/2007GB003068.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, **89**, 303–311.

Reichstein, M., and Coauthors, 2005: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biol.*, **11**, 1424–1439.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704, doi:10.1029/2009GL038082.

Rintoul, S. R., and T. W. Trull, 2001: Seasonal evolution of the mixed layer in the Subantarctic Zone south of Australia. *J. Geophys. Res.*, **106**, 31447–31462.

Roujean, J.-L., and R. Lacaze, 2002: Global mapping of vegetation parameters from POLDER multi-angular measurements for studies of surface-atmosphere interactions: A pragmatic method and its validation. *J. Geophys. Res.*, **107**, doi:10.1029/2001JD000751.

Saleska, S. R., and Coauthors, 2003: Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. *Science*, **302**, 1554-1557.

Santer, B. D., and Coauthors, 2007: Identification of human-induced changes in atmospheric moisture content. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 15248–15253.

Santer, B. D., and Coauthors, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14778–14783.

Sarmiento, J. L., M. Gloor, N. Gruber, C. Beaulieu, A. R. Jacobson, S. M. Fletcher, S. Pacala, and K. Rodgers, 2009: Trends and regional distributions of land and ocean carbon sinks. *Biogeosciences,* **7**, 2351–2367.

Schaefer, K., A. S. Denning, N. Suits, J. Kaduk, I. Baker, S. Los, and L. Prihodko, 2002: Effect of climate on interannual variability of terrestrial CO2 fluxes. *Global Biogeochem. Cycles*, **16**, 1102–1029.

Scherrer, S. C., 2010: Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming. *Int. J. Climatol.*, **31**, 1518–1529

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

Séférian, R., L. Bopp, M. Gehlen, J. Orr, C. Éthé, P. Cadule, O. Aumont, D. Salas-y-Mélia, A. Voldoire, and G. Madec, 2012: Skill Assessment of Three Earth System Models with Common Marine Biogeochemistry. *Climate Dyn,* **1-25**, doi:10.1007/s00382-012-1362-8

Shevliakova, E., S. W. Pacala, S. Malyshev, G. C. Hurtt, P. C. D. Milly, J. P. Caspersen, L. T. Sentman, J. P. Fisk, C. Wirth, and C. Crevoisier, 2009: Carbon cycling under 300 years of land use change: Importance of the secondary vegetation sink. *Global Biogeochem. Cycles*, **23**, GB2022, doi:10.1029/2007GB003176.

Silverman, B. W., 1986: Density Estimation for Statistics and Data Analysis. Chapman and Hall.

Sitch, S., P. Cox, W. J. Collins, and C. Huntingford, 2007: Indirect radiative forcing of climate change through ozone effects on the land–carbon sink. *Nature,* **448**, 791–794.

Smith, P., and Coauthors, 2012: Towards an integrated global framework to assess the impacts of land use and management change on soil carbon: current capability and future vision. *Global Change Biol.,*, **18**, 2089–2101.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 996 pp.

Steinacher, M., F. Joos, T. L. Frölicher, L. Bopp, P. Cadule, V. Cocco, S. C. Doney, M. Gehlen, K. Lindsay, J. K. Moore, B. Schneider, J. Segschneider, 2010: Projected 21st century decrease in marine productivity: a multi-model analysis. *Biogeosciences*, **7**, 979-1005.

Stephens, B. B., and Coauthors, 2007: Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO2. *Science,* **316**, 1732–1735.

Takahashi, T., and Coauthors, 2002: Global sea–air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects. *Deep-Sea Res.*, **49B**, 1601–1622.

Takahashi, T., and Coauthors, 2009: Climatological mean and decadal change in surface ocean pCO2, and net sea–air CO2 flux over the global oceans. *Deep-Sea Res*., **56B**, 554–577.

Taylor, K. E., R. J. Stouffer, and G. Meehl, 2011: An Overview of CMIP5 and the Experiment Design. *Bull. Am. Meteorol. Soc.*, **93**, 485–498.

Tebaldi, C, K. Hayhoe, J. Arblaster, and G. A. Meehl, 2006: Going to the extremes. An intercomparison of model-simulated historical and future changes in extreme events. *Clim. Change,* **79**, 185– 211.

Tjiputra, J. F., K. Assmann, M. Bentsen, I. Bethke, O. H. Otterå, C. Sturm, and C. Heinze, 2009: Bergen earth system model (BCM-C): model description and regional climate-carbon cycle feedbacks assessment. *Geosci. Model Dev. Discuss.*, **2**, 845-887

Todd-Brown, K. E. O., J. T. Randerson, W. M. Post, F. M. Hoffman, C. Tarnocai, E. A. Schuur, and S. D. Allison, 2012: Causes of variation in soil carbon predictions from CMIP5 Earth system models and comparison with observations. *Biogeosciences Discuss.*, **9**, 14437-14473.

Volodin, E. M., N. A. Dianskii, and A. V. Gusev, 2010: Simulating Present Day Climate with the INMCM4.0 Coupled Model of the Atmospheric and Oceanic General Circulations. *Izv. Ocean. Atmos. Phys.*, **46**, 414–431.

Waliser, D., K.-W. Seo, S. Schubert, and E. Njoku, 2007: Global water cycle agreement in the climate models assessed in the IPCC AR4. *Geophys. Res. Lett.*, **34**, L16705, doi:10.1029/2007GL030675.

Watanabe, S., and Coauthors, 2011: MIROC-ESM, 2010: model description and basic results of CMIP5-20c3m experiments. *Geosci. Model Dev.*, **4**, 845-872.

Weiss, M., F. Baret, S. Garrigues, and R. Lacaze, 2007: LAI and fAPAR CYCLOPES global products derived from VEGETATION. Part 2: validation and comparison with MODIS collection 4 products. *Remote Sens. Environ*, **110**, 317-331.

Welp, L.R., R. F. Keeling, H. A. J. Meijer, A. F. Bollenbacher, S. C. Piper, K. Yoshimura, R. J. Francey, C. E. Allison, and M. Wahlen,, 2001: Interannual variability in the oxygen isotopes of atmospheric CO2 driven by El Nino. *Nature,* **477**, 579-582.

Wild, M., and B. Liepert, 2010: The Earth radiation balance as driver of the global hydrological cycle. *Environ. Res. Lett.*, **5**, 025203, doi:10.1088/1748-9326/5/2/025203.

Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.*, **36**, L12702, doi:10.1029/2009GL038710.

Wittig, V. E., E. A. Ainsworth, S. L. Naidu, D. F. Karnosky, and S. P Long, 2009: Quantifying the impact of current and future tropospheric ozone on tree biomass, growth, physiology and biochemistry: a quantitative meta-analysis. *Global Change Biol.*, **15**, 396–424.

Xavier, P. K., J. P. Duvel, P. Braconnot, and F. J. Doblas-Reyes, 2010: An evaluation metric for interannual variability and its application to CMIP3 twentieth-century simulations. *J. Climate* **23**, 3497–3508.

Yang, W., and Coauthors, 2006: MODIS leaf area index products: from validation to algorithm improvement. *IEEE Trans. Geosci. Remote Sens.*, **44**, 1885-1898.

Yin, L., R. Fu, E. Shevliakova, and R. Dickinson, 2012: How well can CMIP5 simulate precipitation and its controlling processes over tropical South America? *Clim. Dyn.*, doi:10.1007/s00382-012-1582-y

Yuan, H., Y. Dai, Z. Xiao, D. Ji, and S. Wei, 2011: Reprocessing the MODIS leaf area index products for land surface and climate modelling. *Remote Sens. Environ*, **115**, 1171–1187.

Zaehle, S., A. D. Friend, P. Friedlingstein, F. Dentener, P. Peylin, and M. Schulz, 2010: Carbon and nitrogen cycle dynamics in the O-CN land surface model: 2. Role of the nitrogen cycle in the historical terrestrial carbon balance. *Global Biogeochem. Cycles*, **24**, GB1006, doi:10.1029/2009GB003522.

Zeng, N., A. Mariotti, and P. Wetzel, 2005: Terrestrial mechanisms of interannual CO2 variability, *Global Biogeochem. Cycles*, **19**, GB1016, doi:10.1029/2004GB002273.

Zhao, M., and S. W. Running, 2010: Drought-Induced Reduction in Global Terrestrial Net Primary Production from 2000 Through 2009. *Science*, **329**, 940–943.

Zhou, T., and R. Yu, 2006: Twentieth-Century Surface Air Temperature over China and the Globe Simulated by Coupled Climate Models. *J. Climate*, **19**, 5843–5858.

Zhu, Z., and Coauthors, 2013: Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011. *Remote Sens.*, **5**, 927-948.

Table 1. *CMIP5 models used with the associated atmospheric and ocean grids, with the number of vertical levels.*

MODELS	SOURCE	ATMOSPHERIC RESOLUTION (lon x lat, levels)	OCEAN RESOLUTION (lon x lat, levels)
BCC-CSM1.1	Beijing Climate Center, China Meteorological Administration, China	2.8125°x~2.8125°, L26	1°x(1-1/3)°, L40
BCC-CSM1.1-M	Beijing Climate Center, China Meteorological Administration, China	1.1°x~1.1°, L26	1°x(1-1/3)°, L40
BNU-ESM	Beijing Normal University	2.8125°x~2.8125°, L26	~1°x~0.6, L50
CanESM2	Canadian Centre for Climate Modelling and Analysis, Canada	2.8125°x~2.8125°, L35	1.40625°x~0.9375°, L40
CESM1-BGC	National Center for Atmospheric Research, United States	0.9°x1.25°, L26	384x320 points (gx1v3),L60
GFDL-ESM2G[x]	Geophysical Fluid Dynamics Laboratory, United States	2.5°x2°, L24	1°x~0.6, L63
GFDL-ESM2M[x]	Geophysical Fluid Dynamics Laboratory, United States	2.5°x2°, L24	1°x~0.6, L50
HadGEM2-CC[y]	Met Office Hadley Centre, UK	1.875°x1.25°, L60	1°x(1-0.3)°, L40
HadGEM2-ES[y]	Met Office Hadley Centre, UK	1.875°x1.25°, L38	1°x(1-0.3)°, L40
INMCM4	Institute for Numerical Mathematics, Russia	2°x1.5°, L21	1°×0.5°, L40
IPSL-CM5A-LR[*]	Institut Pierre Simon Laplace, France	3.75°x~1.875°, L39	~2°x~2°, L31
IPSL-CM5A-MR[*]	Institut Pierre Simon Laplace, France	2.5°x1.25°, L39	~2°x~2°, L31
IPSL-CM5B-LR[*]	Institut Pierre Simon Laplace, France	3.75°x1.875°, L39	~2°x~2°, L31

MIROC-ESM-CHEM[z]	Japan Agency for Marine-Earth Science and Technology, Japan;
	Atmosphere and Ocean Research Institute, Japan;
	National Institute for Environmental Studies, Japan
		2.8125°x2.8125°, L80	1.40625°x~0.9375°, L44

MIROC-ESM[z]	Japan Agency for Marine-Earth Science and Technology, Japan;
	Atmosphere and Ocean Research Institute, Japan;
	National Institute for Environmental Studies, Japan
		2.8125°x2.8125°, L80	1.40625°x~0.9375°, L44

MPI-ESM-LR	Max Planck Institute for Meteorology, Germany	1.875°x1.875°, L47	1.5°x~1.5°, L40
MPI-ESM-MR	Max Planck Institute for Meteorology, Germany	1.875°x1.875°, L47	~0.4°x~0.4°, L40
NorESM1-ME	Norwegian Climate Centre, Norway	2.5°x1.9°, L26	~1°x~0.5°, L53

[x] *The two GFDL models differ almost exclusively in the physical ocean component; ESM2M uses Modular Ocean Model version 4.1 with vertical pressure layers, while ESM2G uses Generalized Ocean Layer Dynamics with a bulk mixed layer and interior isopycnal layers (Dunne et al. 2012).*

[y] *HadGEM2 models differ for the number of vertical levels in the atmospheric component and for different representation of processes (HadGEM2-ES also reproduce the atmospheric chemistry, Martin et al. 2011).*

[*] *IPSL-CM5A-LR and IPSL-CM5A-MR models differ for the resolution of the atmospheric component, while IPSL-CM5A-LR and IPSL-CM5B-LR differ only for some parameterizations in the atmospheric model (Dufresne et al. 2012).*

[z] *The difference between MIROC-ESM and MIROC-ESM-CHEM is that this latter simulates the atmospheric chemistry (Watanabe et al. 2011).*

Table 2. *Summary of land and ocean biogeochemistry models used by ESMs and comparison of the selected processes (dynamic vegetation, nitrogen cycling and land use change) for the only terrestrial modules.*

MODELS	LAND MODELS	DYNAMIC VEGETATION	N CYCLE	LUC	OCEAN MODELS
BCC-CSM1	BCC_AVIM1.0	Y	Y	N	Simple model into MOM4
BCC-CSM1-M	BCC_AVIM1.0	Y	Y	N	Simple model into MOM4
BNU-ESM	CoLM + BNU-DGVM	Y	N	Y	iBGC
CanESM2	CLASS2.7 + CTEM1	N	N	Y	CMOC
CESM1-BGC	CLM4	N	Y	Y	BEC
GFDL-ESM2G	LM3	Y	N	Y	TOPAZ2
GFDL-ESM2M	LM3	Y	N	Y	TOPAZ2
HadGEM2-CC	JULES + TRIFFID	Y	N	Y	Diat-HadOCC
HadGEM2-ES	JULES + TRIFFID	Y	N	Y	Diat-HadOCC
INMCM4	Simple model into INMCM4 atmospheric component	N	N	Y*	Simple model into INMCM4 ocean component
IPSL-CM5A-LR	ORCHIDEE	N	N	Y	PISCES
IPSL-CM5A-MR	ORCHIDEE	N	N	Y	PISCES
IPSL-CM5B-LR	ORCHIDEE	N	N	Y	PISCES
MIROC-ESM-CHEM	MATSIRO + SEIB-DGVM	Y	N	Y	NPZD
MIROC-ESM	MATSIRO + SEIB-DGVM	Y	N	Y	NPZD
MPI-ESM-LR	JSBACH + BETHY	Y	N	Y	HAMOCC5
MPI-ESM-MR	JSBACH + BETHY	Y	N	Y	HAMOCC5
NorESM1-ME	CLM4	N	Y	Y	HAMOCC5

* *In INMCM4 land use change was prescribed at low preindustrial level.*

Table 3. *Temporal range of available data for historical simulation, and variable used in this study, with associated the number of independent realization for each variable. Note that not all the variables for all the ensembles are available on PDMDI server.*

MODELS	**PHYSICAL VARIABLES**				**BIOLOGICAL VARIABLES**						
	LAND		**OCEAN**		**LAND**					**OCEAN**	
	Surface Temperature	**Precipitation**	**SST**	**MLD**	**GPP**	**LAI**	**NBP**	**SoilC**	**VegC**	**$fgCO_2$**	**PP**
BCC-CSM1-1	3	3	3	n/a	3	3	n/a	3	3	3	n/a
BCC-CSM1-1-M	3	3	3	n/a	3	3	n/a	3	3	3	n/a
BNU-ESM	1	1	1[*]	n/a	1	1	1	1	1	1	n/a
CanESM2	5	5	5	1	5	5	5	5	5	5	5
CESM1-BGC	1	1	1	1	1	1	1	1	1	1	1
GFDL-ESM2G	1	1	1	1	1	1	1	1	1	1	1
GFDL-ESM2M	1	1	1	1	1	1	1	1	1	1	1
HadGEM2-CC	1	1	1	1	1	1	1	1	1	1	1
HadGEM2-ES	4	4	4	1	4	4	4	4	4	4	4
INMCM4	1	1	1	n/a	1	1	1[y]	1	1	1	n/a
IPSL-CM5A-LR	5	5	5	5	5	5	5	5	5	5	5
IPSL-CM5A-MR	1	1	1	1	1	1	1	1	1	1	1
IPSL-CM5B-LR	1	1	1	1	1	1	1	1	1	1	1
MIROC-ESM-CHEM	1	1	1	1[x]	1	1	1	1	1	1	1
MIROC-ESM	3	3	1	1[x]	3	3	3	3	3	3	1
MPI-ESM-LR	3	3	3	3	3	3	3	3	3	3	3
MPI-ESM-MR	3	3	3	3	3	3	3	3	3	3	3
NorESM1-ME	1	1	1	1	1	1	1	1	1	1	1

[x] *MLD from MIROC models was not directly provided as output, but it has been estimated from potential temperature, potential density and salinity.*

[*] *Monthly SST were not available on the server; we used daily SST in the reference period 1950-2005 to compute the monthly SST.*

[y] In INMCM4 *the land use was prescribed at preindustial level and kept constant during the whole simulation; this means that the provided NBP does not include the LUC term and therefore it should be considered as NEP rather NBP. For this reason we decided to exclude the INMCM4 NBP from our analysis.*

Table 4. *Observationally-based data sets used to validate models. The spatial resolution is given as latitude x longitude.*

VARIABLES	REFERENCE	TEMPORAL WINDOW	SPATIAL RESOLUTION	TEMPORAL RESOLUTION
Temperature	CRU (Mitchell and Jones 2005)	1901-2006	Global (land), 0.5°x0.5°	Monthly
Precipitation	CRU (Mitchell and Jones 2005)	1901-2006	Global (land), 0.5°x0.5°	Monthly
SST	HadISST (Rayner et al. 2003)	1870-2011	Global, 1°x1°	Monthly
MLD	de Boyer Montégut et al. (2004)	1941-2008	Global, 2°×2°	Climatology
GPP	MTE (Jung et al. 2009)	1982-2008	Global, 0.5°x0.5°	Monthly
LAI	LAI3g (Zhu et al. 2013)	1981-2011	Global, ~0.08°x ~0.08°	15 Days
NBP	Inversion (Gurney et al. 2004)	1995-2008	Global, 0.5°x0.5°	Monthly
	GCP (Le Quéré et al. 2009)	1959-2008	Global, spatial average	Yearly
Soil Carbon	HSWD, (FAO 2012)	n/a	Global, 1 km x1 km	Annual Value
Vegetation Carbon	NDP-017b (Gibbs 2006)	n/a	Global, 0.5x0.5	Annual Value
fgCO2	Inversion (Gurney et al. 2004)	1995-2008	Global, 0.5°x0.5°	Monthly
	GCP (Le Quéré et al. 2009)	1959-2008	Global, spatial average	Yearly
	Takahashi (Takahashi et al. 2009)	2000	Global, 4°x5°	Climatology
NPP	SeaWIFS. (Behrenfeld and Falkowski, 1997)	1998-2007	Global, 6x6 km	Monthly

Table 5. *Skill score values with the corresponding weights used to compute regional estimates.*

SKILL SCORE	***WEIGHT***
$\int Z_{x,y}<0.05$	*0.05*
$0.05\leq\int Z_{x,y} <0.25$	*0.1*
$0.25\leq\int Z_{x,y} <0.5$	*0.15*
$0.5\leq\int Z_{x,y} <0.75$	*0.25*
$\int Z_{x,y} \geq 0.75$	*0.45*

FIGURE CAPTIONS

Figure 1. *Globally averaged surface air temperature (only land points, without Antarctica) from observations (CRU), and as simulated by CMIP5 models in response to major forcings, natural and anthropogenic (upper panel). The anomaly has been computed with respect to the reference period 1901-1930.*

Vertical grey lines indicate the timing of major volcanic eruptions, while orange line shows the most intense El-Niño event occurred in the 20th century. The grey shaded area represents range of variability of the 18 CMIP5 models, i.e. the envelope of positive and negative temperature extremes based on multi-model mean, while the red shading shows the confidence interval diagnosed from the ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve).

Lower panels show inter-comparison of surface temperature over land estimated by 18 different CMIP5 models (circles) with reference temperature estimated by CRU dataset (triangles) for the whole Globe, Southern Hemisphere (20°S-90°S, without Antarctica), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). Scatter plot shows multi-year average temperature in x-axis computed during the period 1986-2005, its linear trend in y-axis over the full period 1901-2005, and the Model Variability Index (MVI).

Figure 2. *As Figure 1 but for land precipitation.*

Figure 3. *As Figure 1 but for SST. The regional SST are computed over the ocean sub-regions rather than over the land sub-domains. The reference SST dataset is HadISST. Note that BNU-ESM trend has been computed over the period 1950-2005 due to the unavailability of data on PCMDI server; in addition, in the upper panel BNU-ESM has been excluded by the analysis.*

Figure 4. *Simulated and observed climatological seasonal cycle of MLD (meters) for each ocean sub-domain.*

Figure 5. *Temporal variability of CMIP5 global land-atmosphere CO2 flux compared to Global Carbon Project (GCP) estimates (black line). Green shading shows the confidence interval diagnosed from the CMIP5 ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve), while the grey shading represents the range of variability of CMIP5 models. Positive values correspond to land uptake.*

Figure 6. *Error-bar plot showing the 1986-2005 CMIP5 integrated NBP over the land sub-domains. Positive values correspond to land uptake, and vertical bars are computed considering the interannual variation. At global scale CMIP5 models are compared also with GCP estimates, while in all the other sub-regions the reference observations are inversion estimates (triangles).*

Figure 7. *Comparison of mean annual cycle of NBP (PgC/y) as simulated by CMIP5 models and JMA inversion in the 20-year period 1986-2005.*

Figure 8. *Integrated GPP over the land sub-domains. The linear trend has been computed over the period 1986-2005, and the reference dataset is MTE-GPP.*

Figure 9. *Comparison of mean annual cycle of GPP (PgC/y) as simulated by CMIP5 models with MTE-GPP data over the 20-year period 1986-2005.*

Figure 10. *Mean annual LAI as simulated by CMIP5 models and the reference LAI3g data (black triangle) over the land sub-domains.*

Figure 11. *Mean annual cycle of LAI over the period 1986-2005.*

Figure 12. *Simulated CMIP5 soil and vegetation carbon content over the period 1986-2005 compared against the Harmonized World Soil Database (HWSD) and the NDP-017 vegetation data.*

Figure 13. *Temporal variability of CMIP5 global ocean-atmosphere CO2 flux compared to Global Carbon Project (GCP) estimates (black line). Blue shading shows the confidence interval diagnosed from the CMIP5 ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve), while the grey shading represents the range of variability of CMIP5 models. Positive values correspond to ocean uptake.*

Figure 14. *Error-bar plot showing the 1986-2005 CMIP5 means and standard deviations of ocean-atmosphere carbon fluxes (fgCO2) in the chosen ocean sub-domains. Positive values correspond to ocean uptake, while vertical bars are computed considering the interannual variation. At global scale CMIP5 models are compared also with GCP estimates, while in all the other sub-regions the reference observations are JMA inversion estimates and Takahashi data (triangles).*

Figure 15. *Comparison of mean annual cycle of fgCO2 (PgC/y) as simulated by CMIP5 models with JMA inversion and Takahashi data in the 20-year period 1986-2005.*

Figure 16. *Ocean primary production integrated over the ocean sub-domains as simulated by CMIP5 models and observed (SeaWIFS) in the period 1998-2005.*

Figure 17. *Comparison of ocean primary production (PgC/y) mean annual cycle as simulated by CMIP5 models and SeaWIFS observations in the period 1998-2005.*

Figure 18. *Seasonal skill score matrix as computed according to Equation 3 for the whole Globe, Southern Hemisphere (20°S-90°S), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). A score of 0 indicates poor performance of models reproducing the phase and amplitude of the reference mean annual cycle, while a perfect score is equal to 1.*

Figure 19. *PDF-based skill scores for temperature, precipitation, LAI, and NBP for the*

whole Globe, Southern Hemisphere (20°S-90°S), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). A perfect score is 1.

Note that since the reference data for the soil and vegetation carbon pools are a single annual data, we were unable to build the PDF, therefore the skill scores for these variables are based on the normalized mean bias between the model and the reference data (see equation 6).

Figure 20. *As Figure 18 but for the ocean variables.*

Figure 21. *As Figure 19 but for the ocean variables. Note that since the MLD dataset is a climatology we were unable to compute the PDF, consequently the skill scores have been computed according to equation 6.*

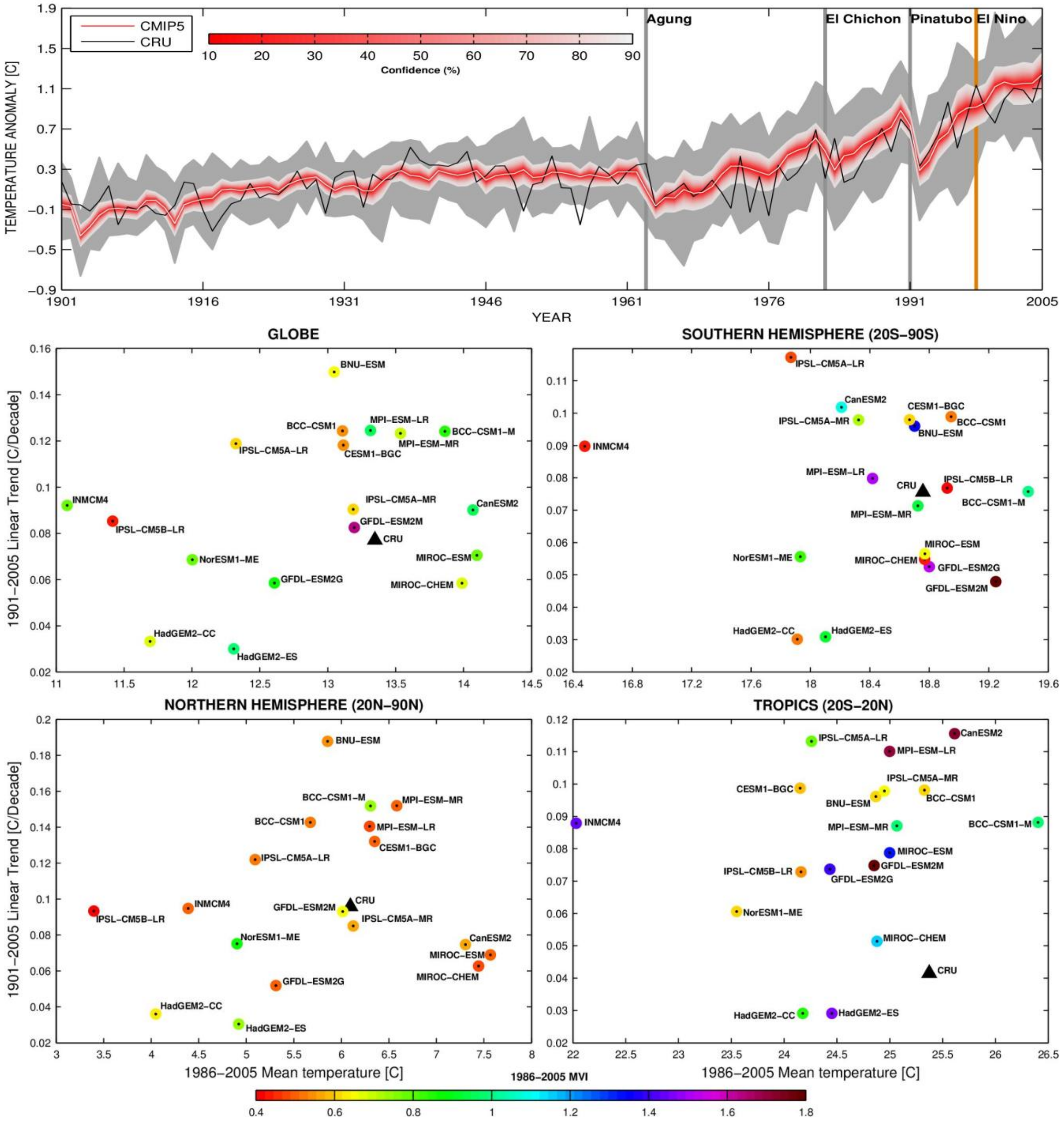

Figure 1. Globally averaged surface air temperature (only land points, without Antarctica) from observations (CRU), and as simulated by CMIP5 models in response to major forcings, natural and anthropogenic (upper panel). The anomaly has been computed with respect to the reference period 1901-1930.

Vertical grey lines indicate the timing of major volcanic eruptions, while orange line shows the most intense El-Niño event occurred in the 20th century. The grey shaded area represents range of variability of the 18 CMIP5 models, i.e. the envelope of positive and negative temperature extremes based on multi-model mean, while the red shading shows the confidence interval diagnosed from the ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve).

Lower panels show inter-comparison of surface temperature over land estimated by 18 different CMIP5 models (circles) with reference temperature estimated by CRU dataset (triangles) for the whole Globe, Southern Hemisphere (20°S-90°S, without Antarctica), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). Scatter plot shows multi-year average temperature in x-axis computed during the period 1986-2005, its linear trend in y-axis over the full period 1901-2005, and the Model Variability Index (MVI).

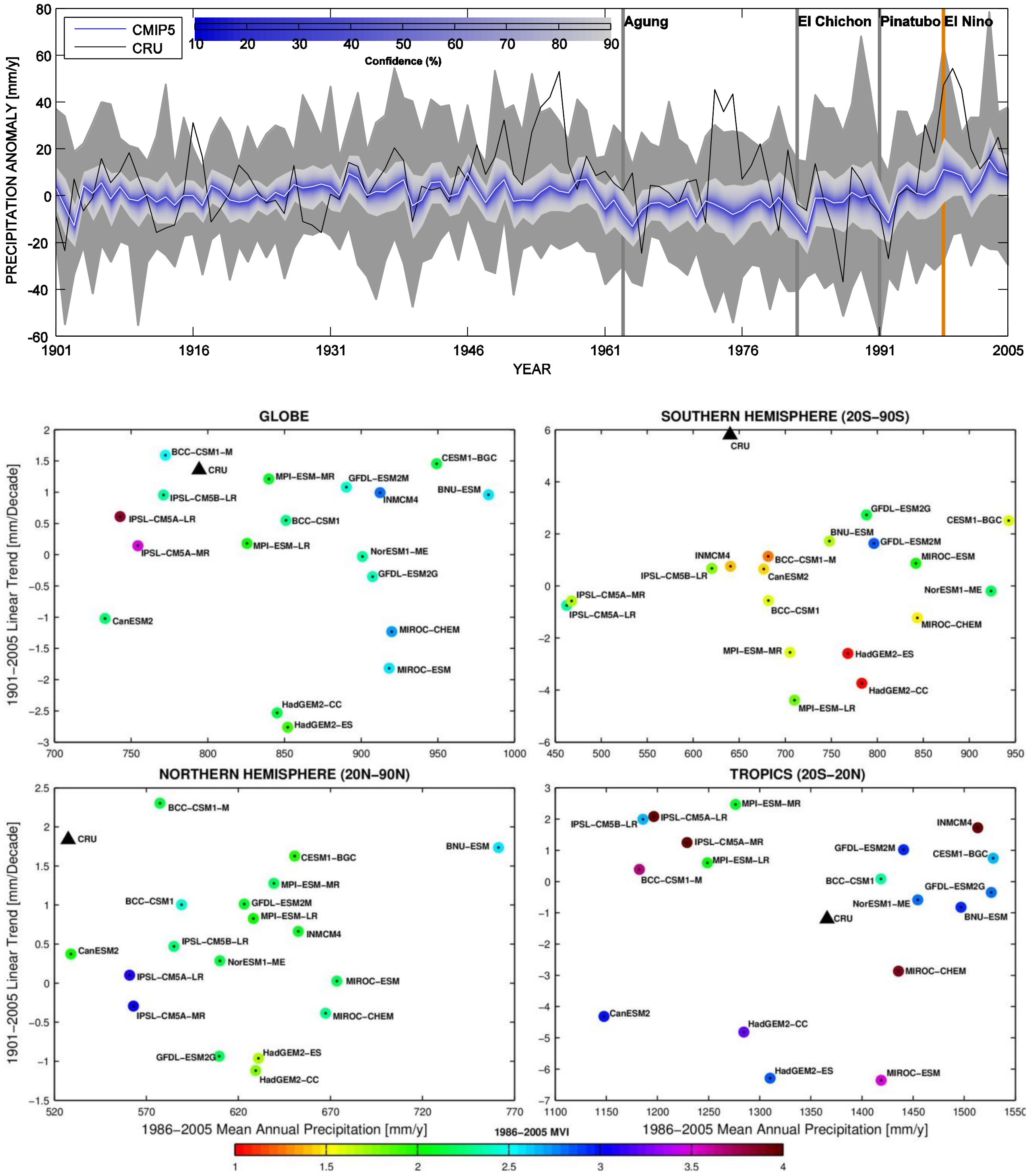

Figure 2. *As **Figure 1** but for land precipitation.*

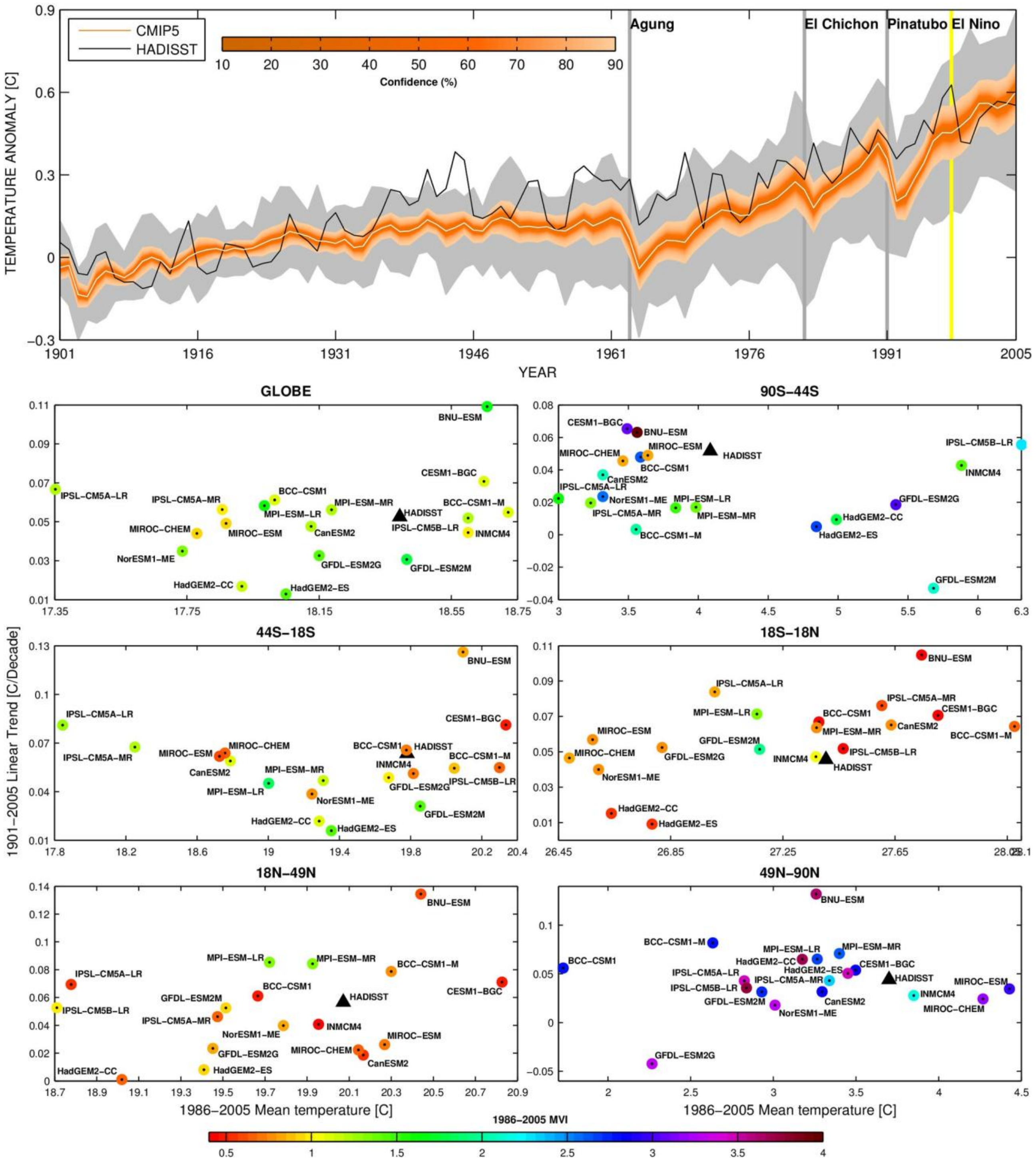

Figure 3. As *Figure 1* but for SST. The regional SST are computed over the ocean sub-regions rather than over the land sub-domains. The reference SST dataset is HadISST. Note that BNU-ESM trend has been computed over the period 1950-2005 due to the unavailability of data on PCMDI server; in addition, in the upper panel BNU-ESM has been excluded by the analysis.

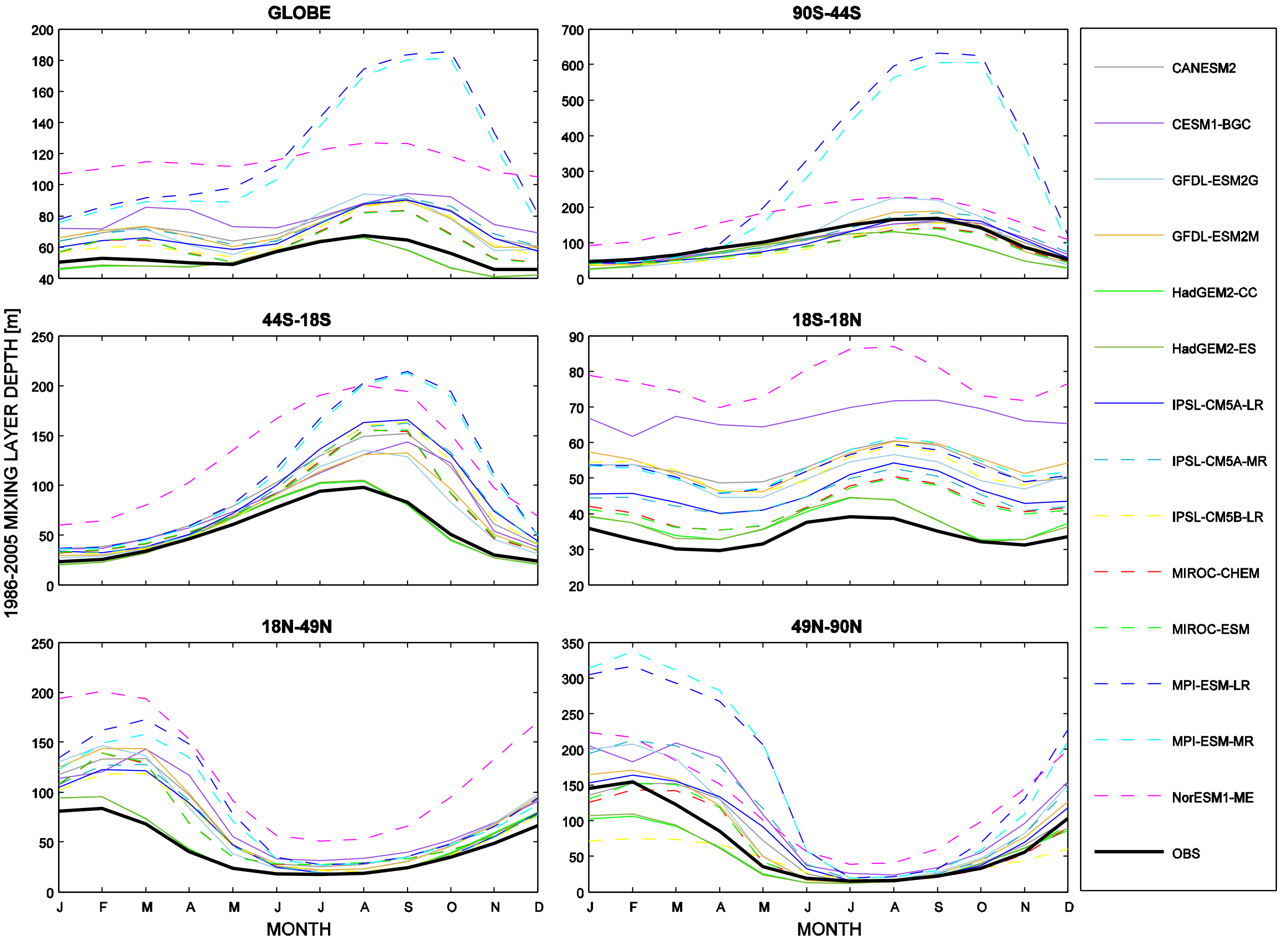

Figure 4. *Simulated and observed climatological seasonal cycle of MLD (meters) for each ocean sub-domain.*

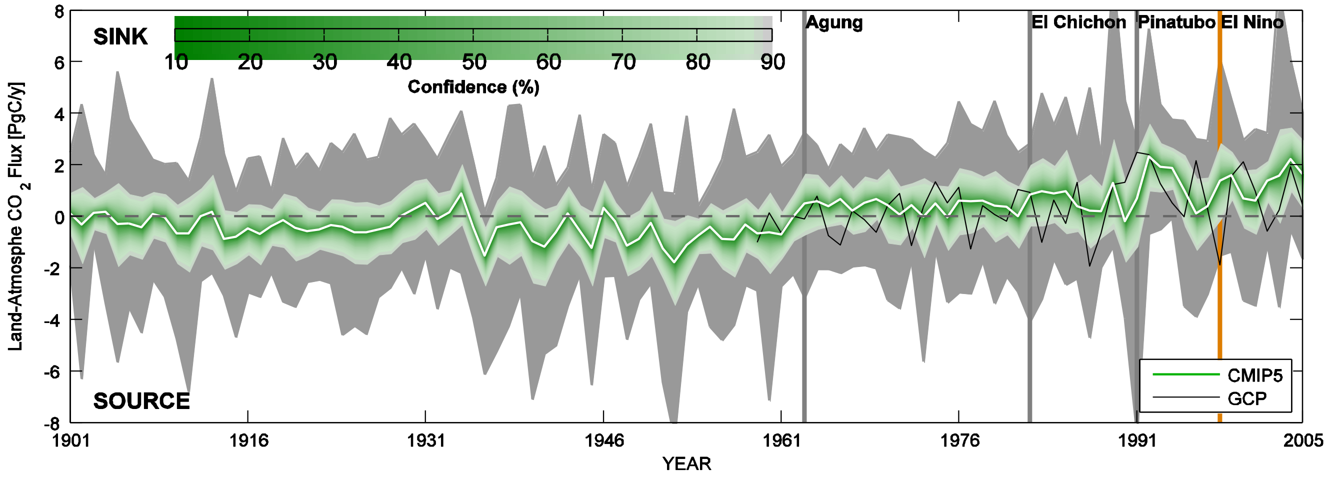

Figure 5. *Temporal variability of CMIP5 global land-atmosphere CO2 flux compared to Global Carbon Project (GCP) estimates (black line). Green shading shows the confidence interval diagnosed from the CMIP5 ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve), while the grey shading represents the range of variability of CMIP5 models. Positive values correspond to land uptake.*

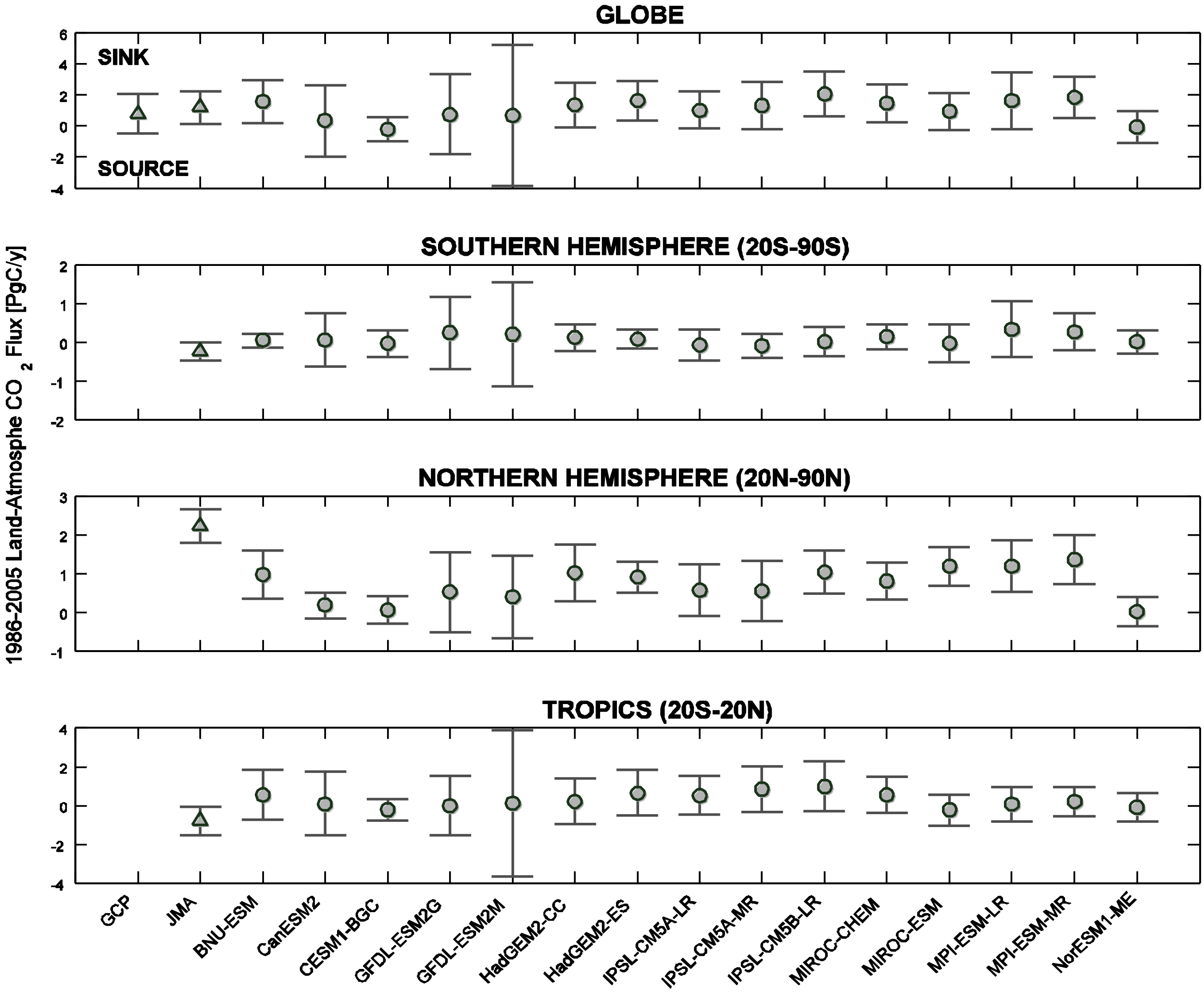

Figure 6. *Error-bar plot showing the 1986-2005 CMIP5 integrated NBP over the land sub-domains. Positive values correspond to land uptake, and vertical bars are computed considering the interannual variation. At global scale CMIP5 models are compared also with GCP estimates, while in all the other sub-regions the reference observations are inversion estimates (triangles).*

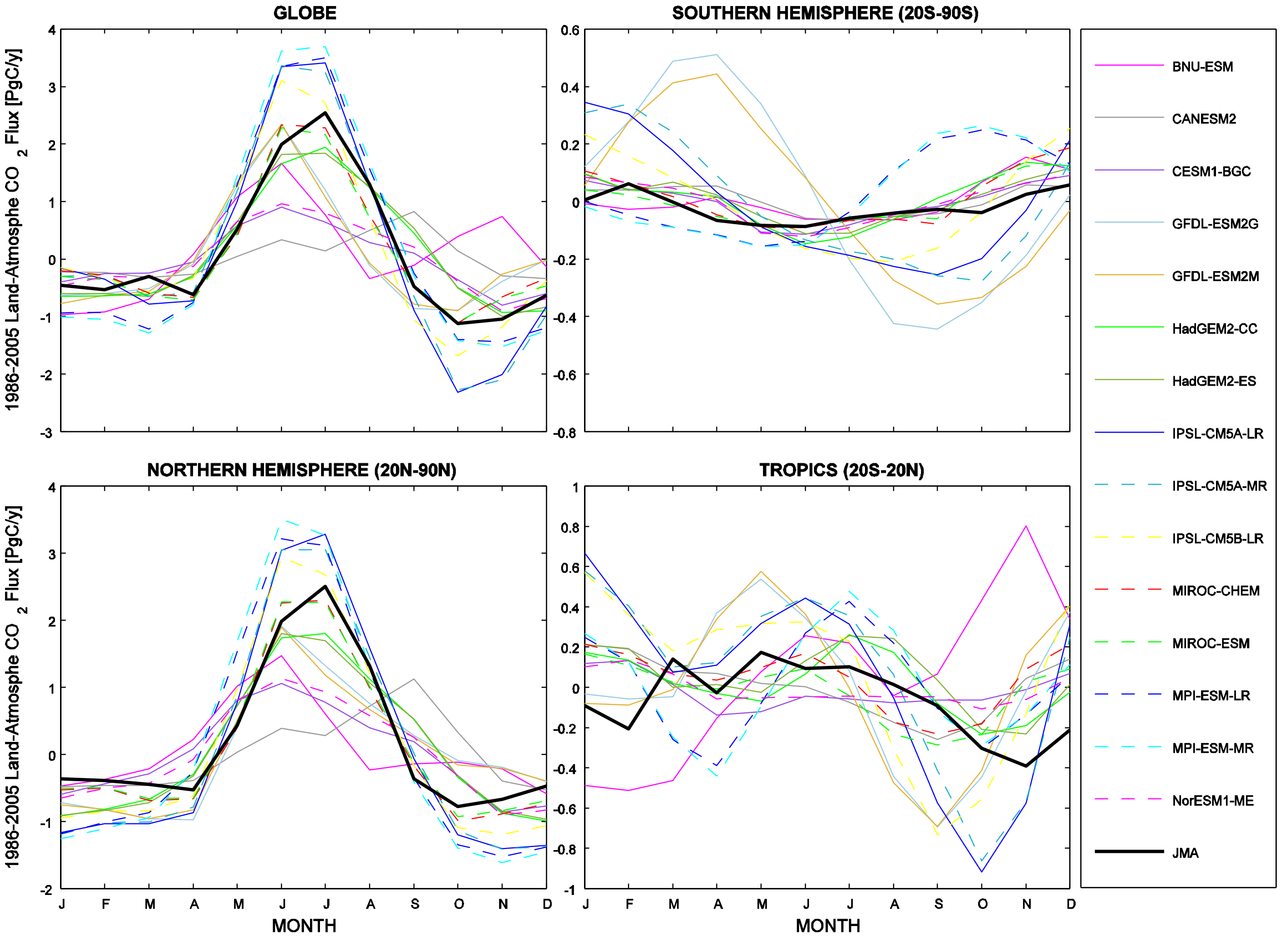

Figure 7. *Comparison of mean annual cycle of NBP (PgC/y) as simulated by CMIP5 models and JMA inversion in the 20-year period 1986-2005.*

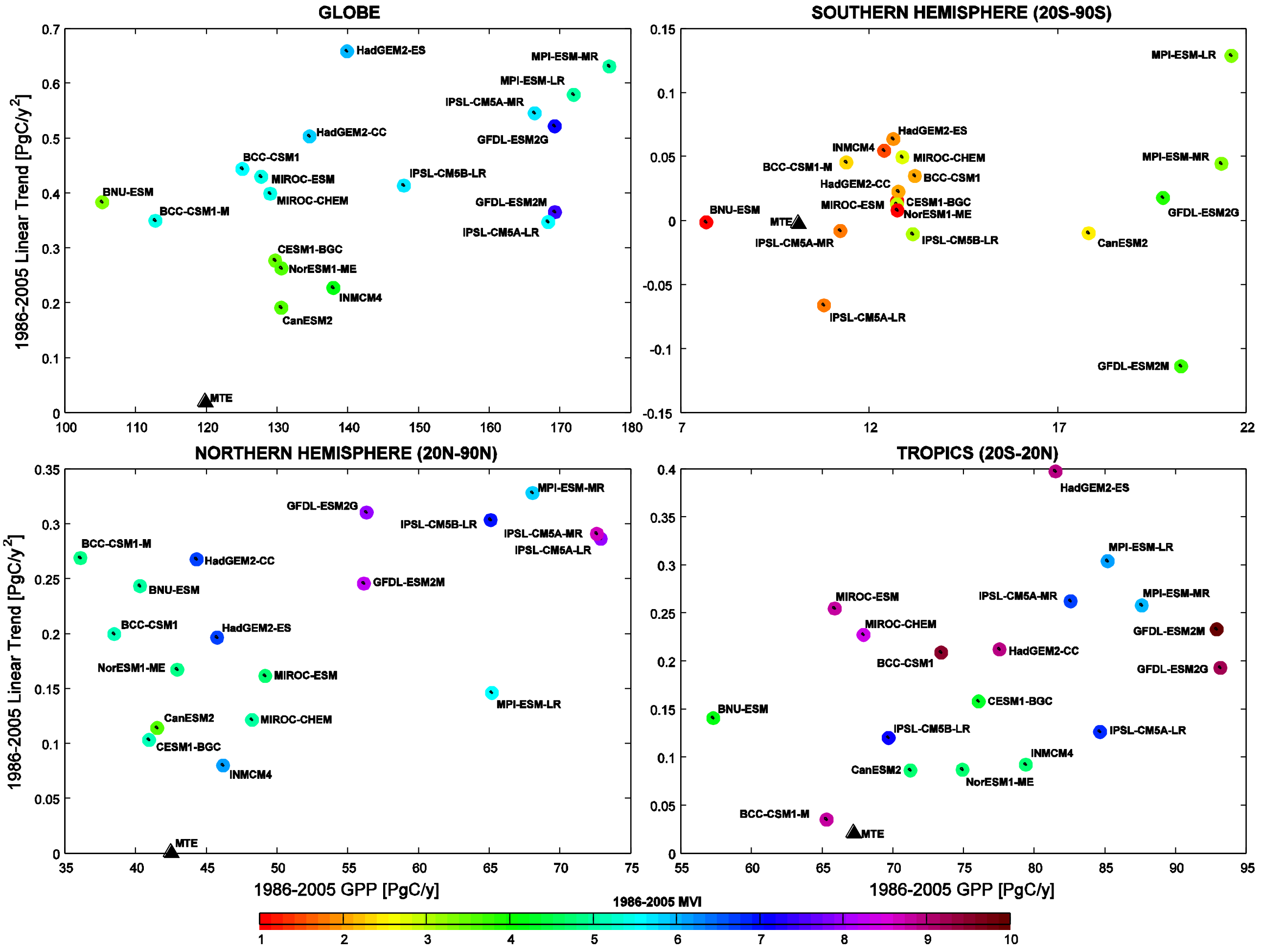

Figure 8. *Integrated GPP over the land sub-domains. The linear trend has been computed over the period 1986-2005, and the reference dataset is MTE-GPP.*

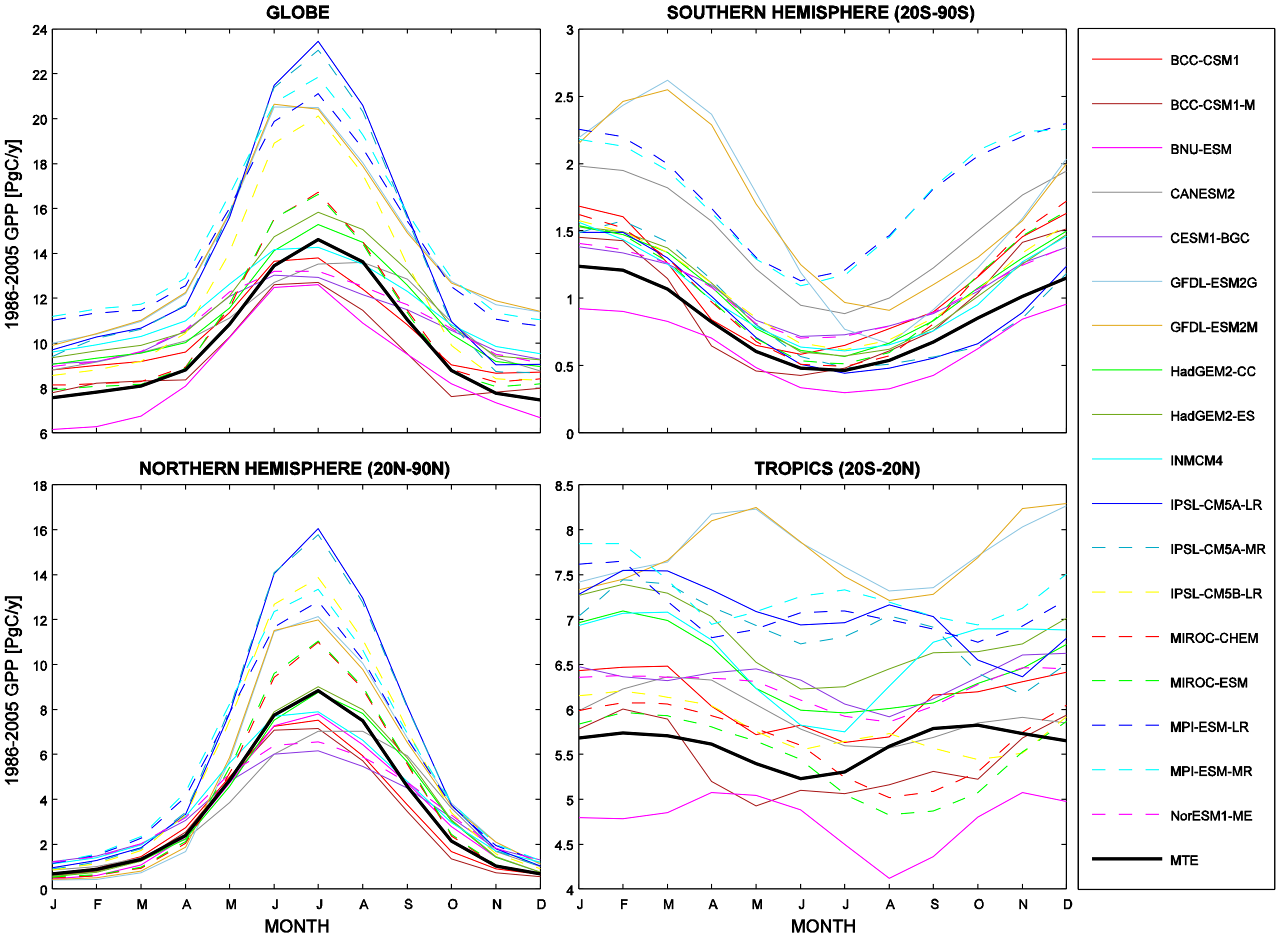

Figure 9. *Comparison of mean annual cycle of GPP (PgC/y) as simulated by CMIP5 models with MTE-GPP data over the 20-year period 1986-2005.*

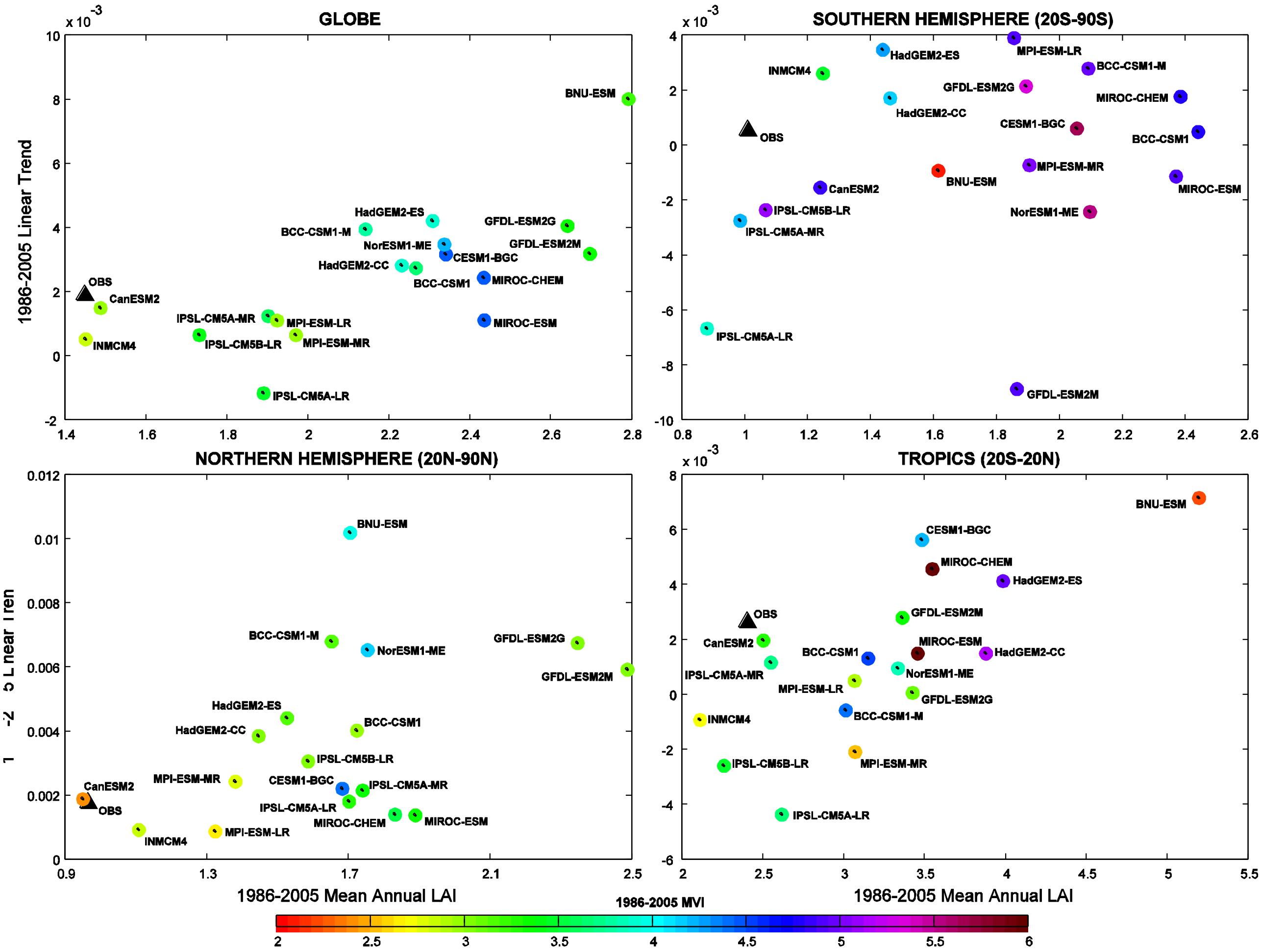

Figure 10. *Mean annual LAI as simulated by CMIP5 models and reference LAI3g data (black triangle) over the land sub-domains.*

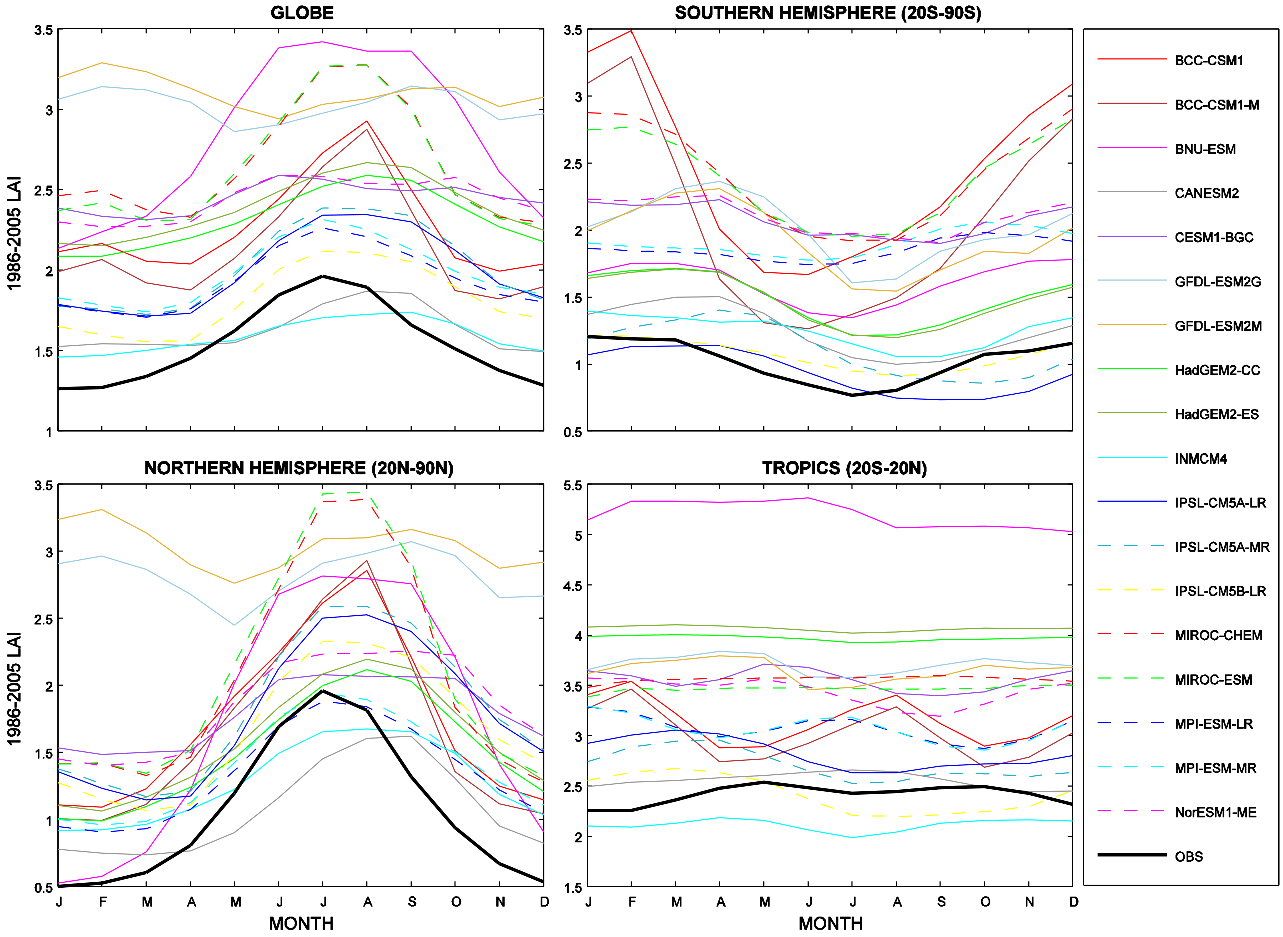

Figure 11. *Mean annual cycle of LAI over the period 1986-2005.*

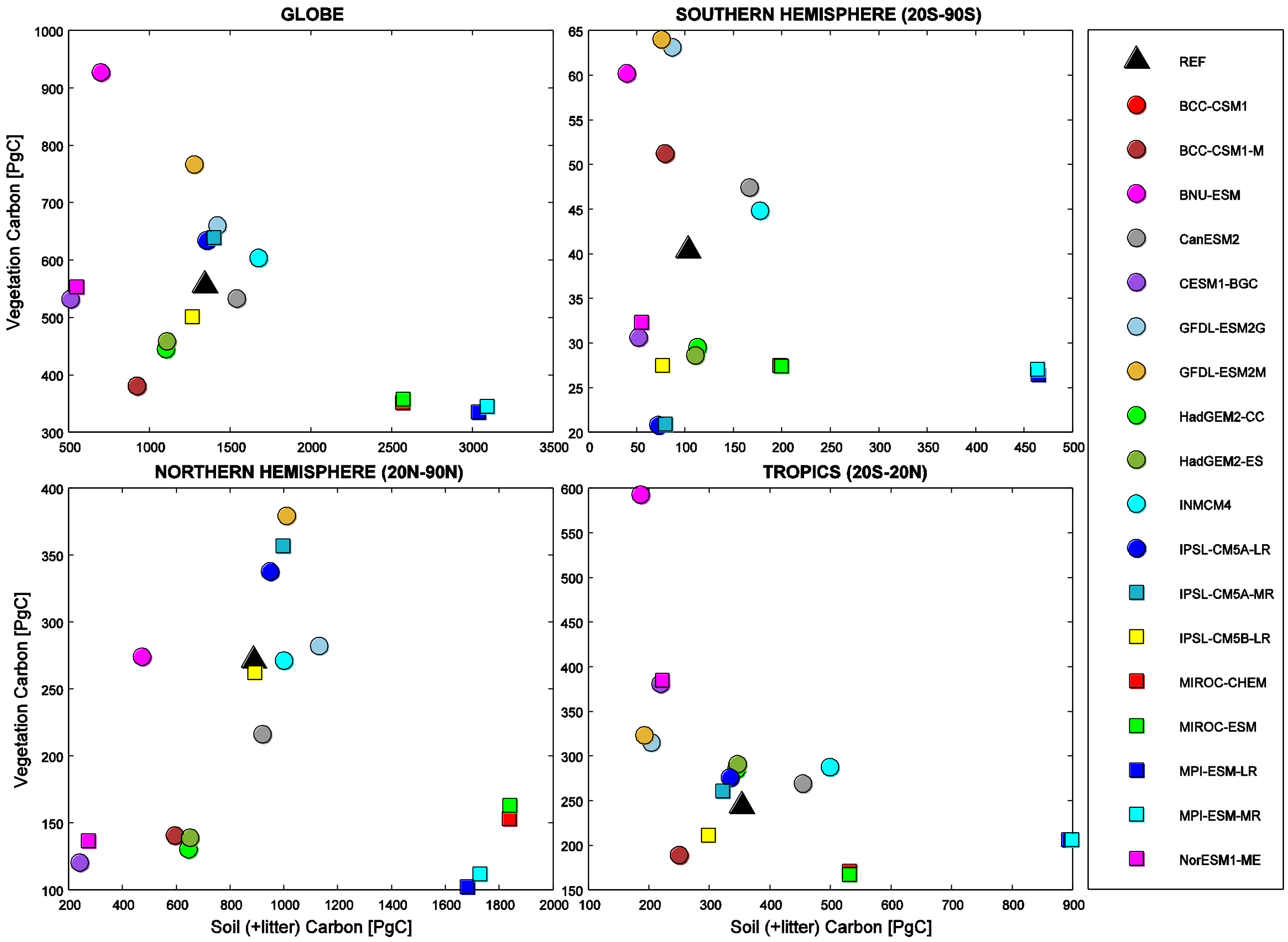

Figure 12. *Simulated CMIP5 soil and vegetation carbon content over the period 1986-2005 compared against the Harmonized World Soil Database (HWSD) and the NDP-017 vegetation data.*

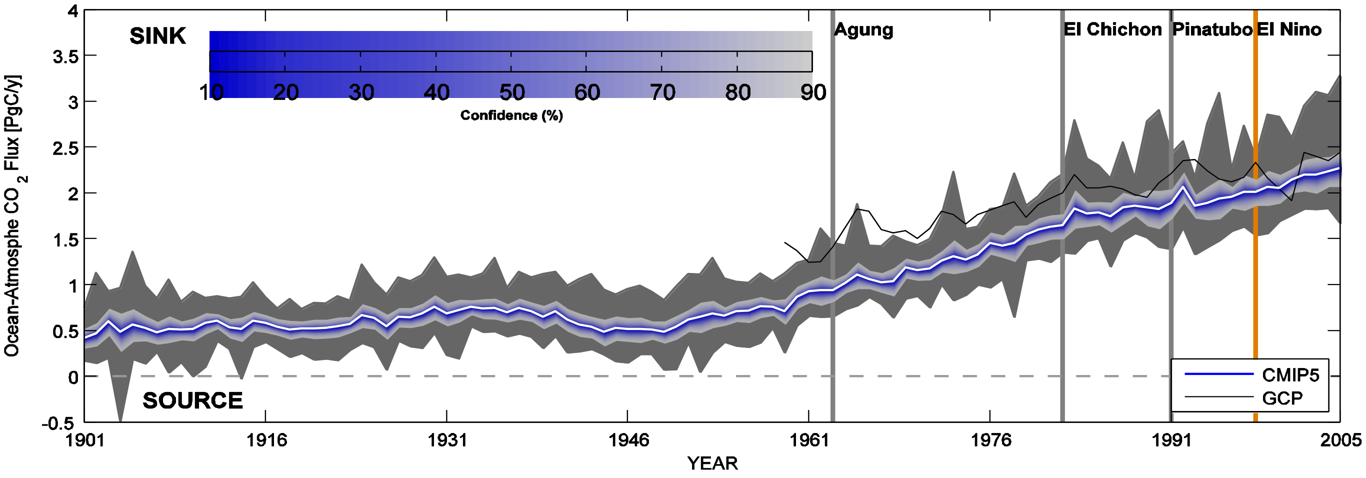

Figure 13. *Temporal variability of CMIP5 global ocean-atmosphere CO2 flux compared to Global Carbon Project (GCP) estimates (black line). Blue shading shows the confidence interval diagnosed from the CMIP5 ensemble standard deviation assuming a t-distribution centred on the ensemble mean (white curve), while the grey shading represents the range of variability of CMIP5 models. Positive values correspond to ocean uptake.*

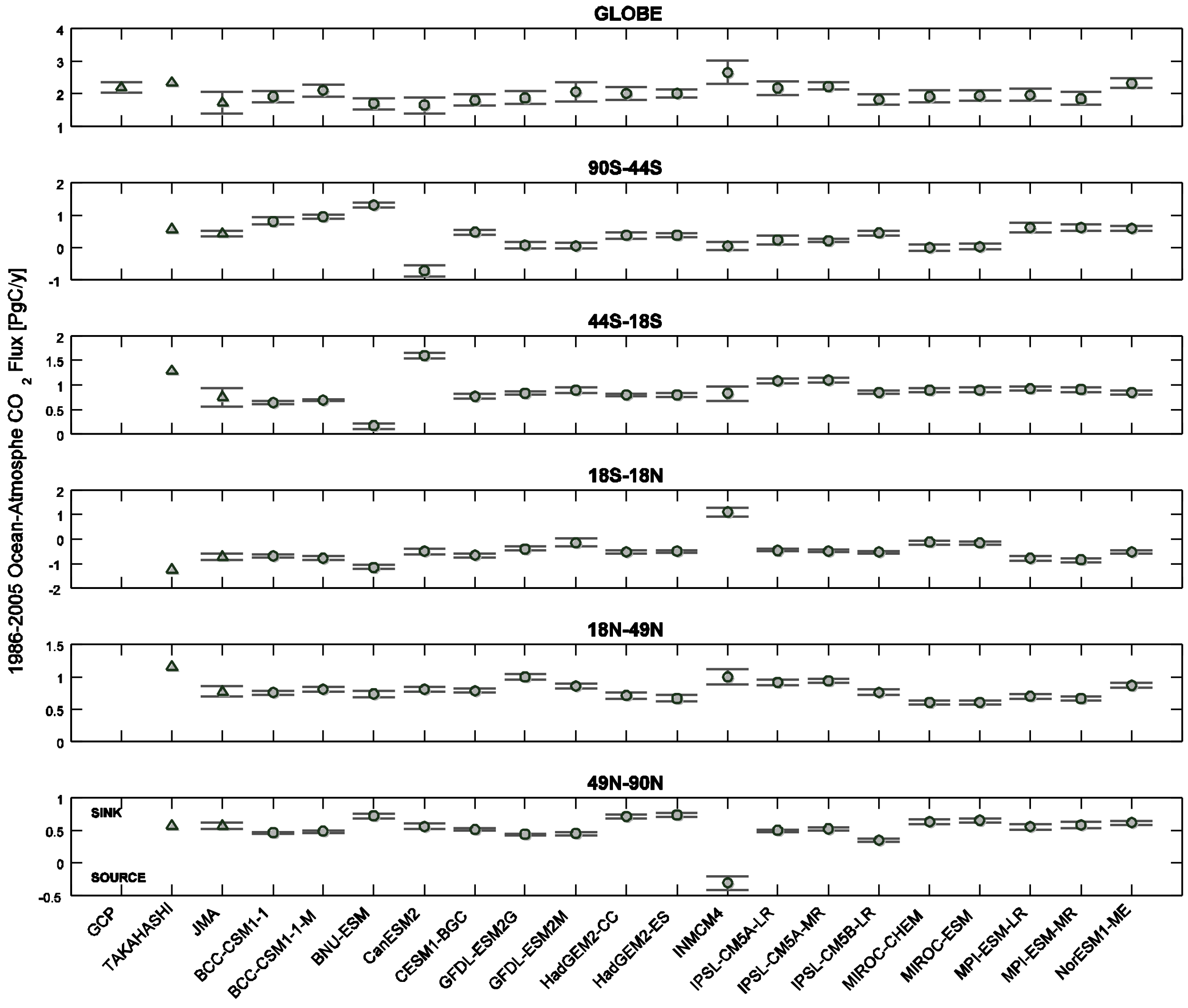

Figure 14. *Error-bar plot showing the 1986-2005 CMIP5 means and standard deviations of ocean-atmosphere carbon fluxes (fgCO2) in the chosen ocean sub-domains. Positive values correspond to ocean uptake, while vertical bars are computed considering the interannual variation. At global scale CMIP5 models are compared also with GCP estimates, while in all the other sub-regions the reference observations are JMA inversion estimates and Takahashi data (triangles).*

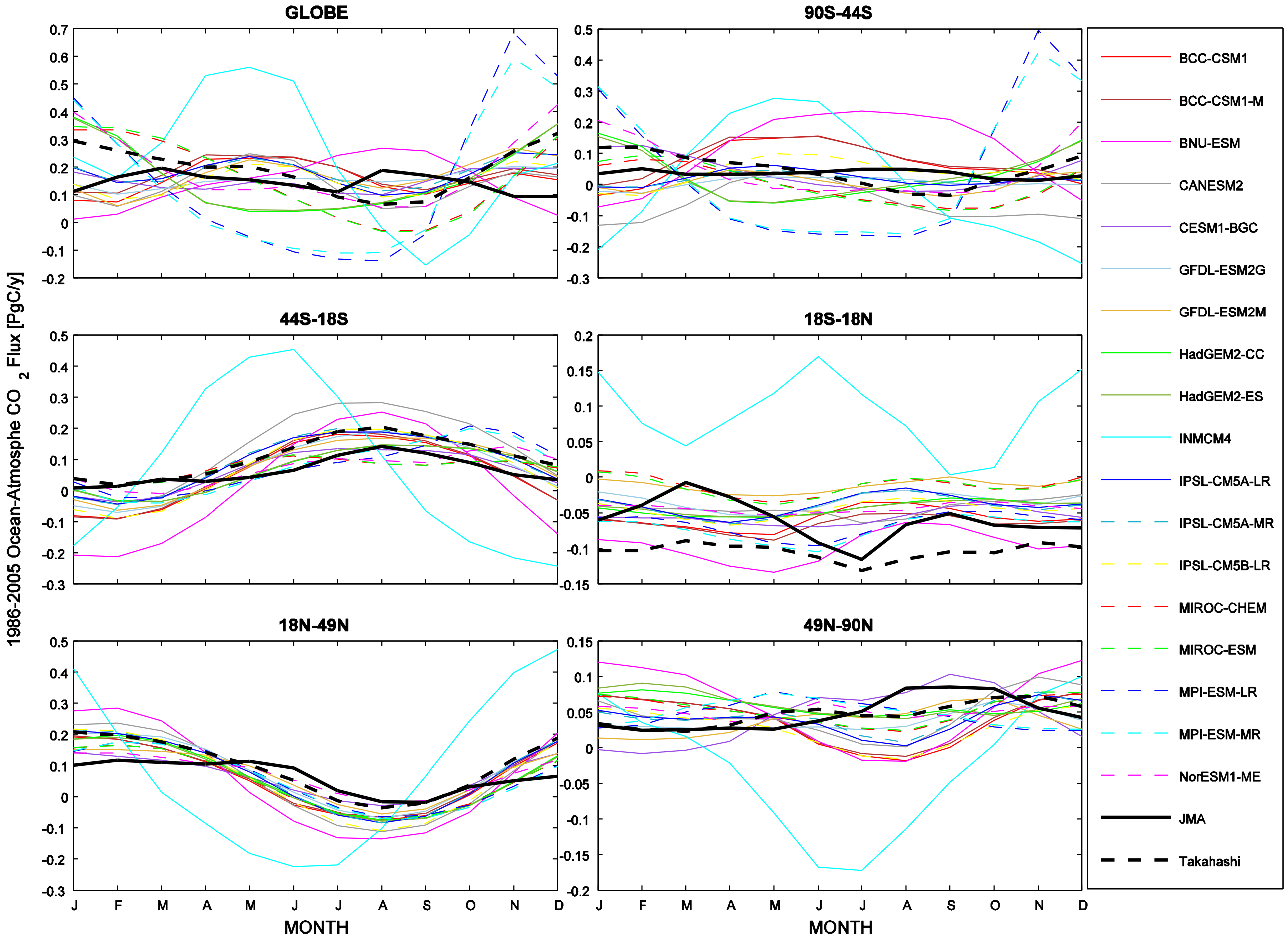

Figure 15. *Comparison of mean annual cycle of fgCO2 (PgC/y) as simulated by CMIP5 models with JMA inversion and Takahashi data in the 20-year period 1986-2005.*

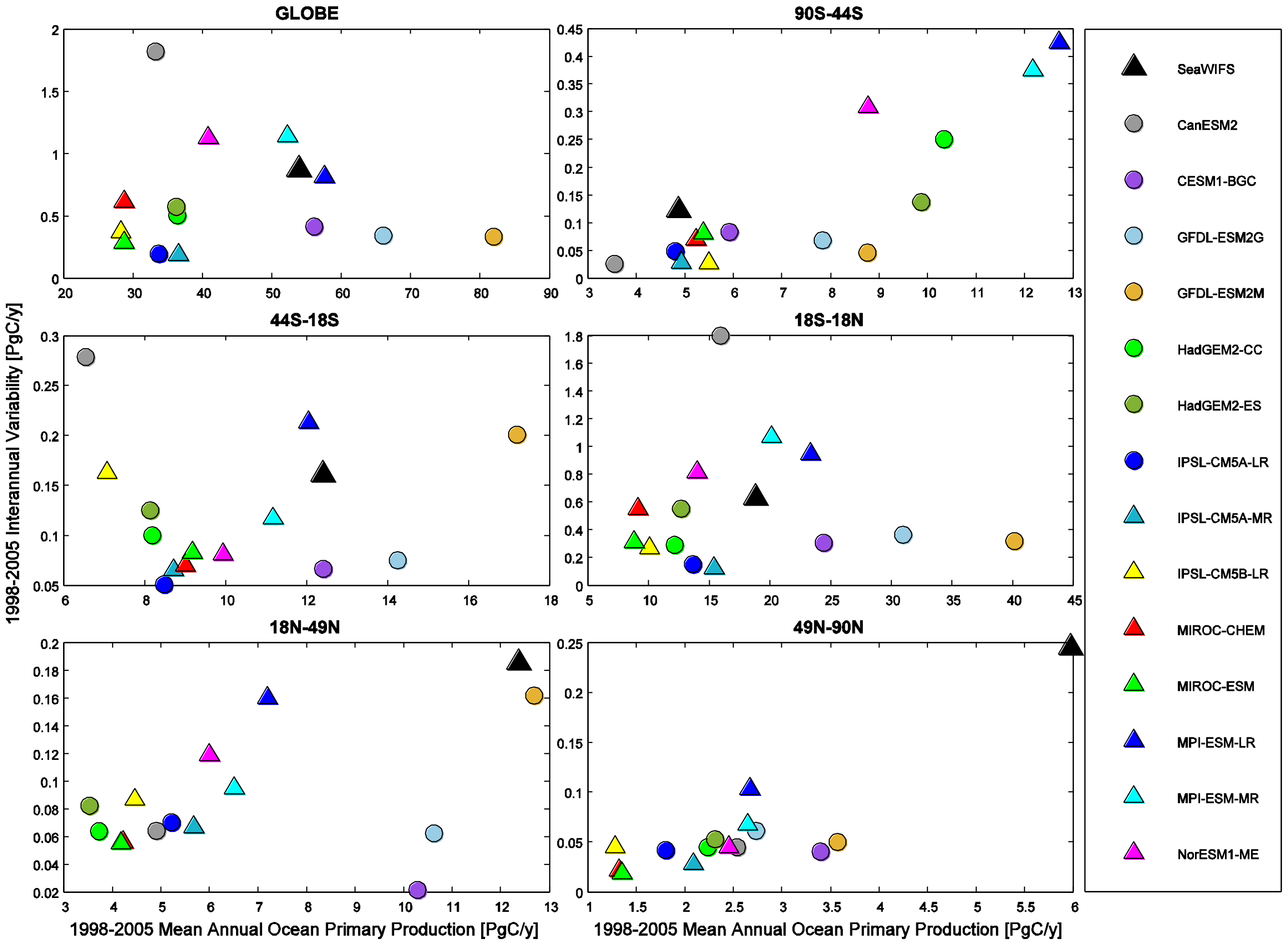

Figure 16. *Ocean primary production integrated over the ocean sub-domains as simulated by CMIP5 models and observed (SeaWIFS) in the period 1998-2005.*

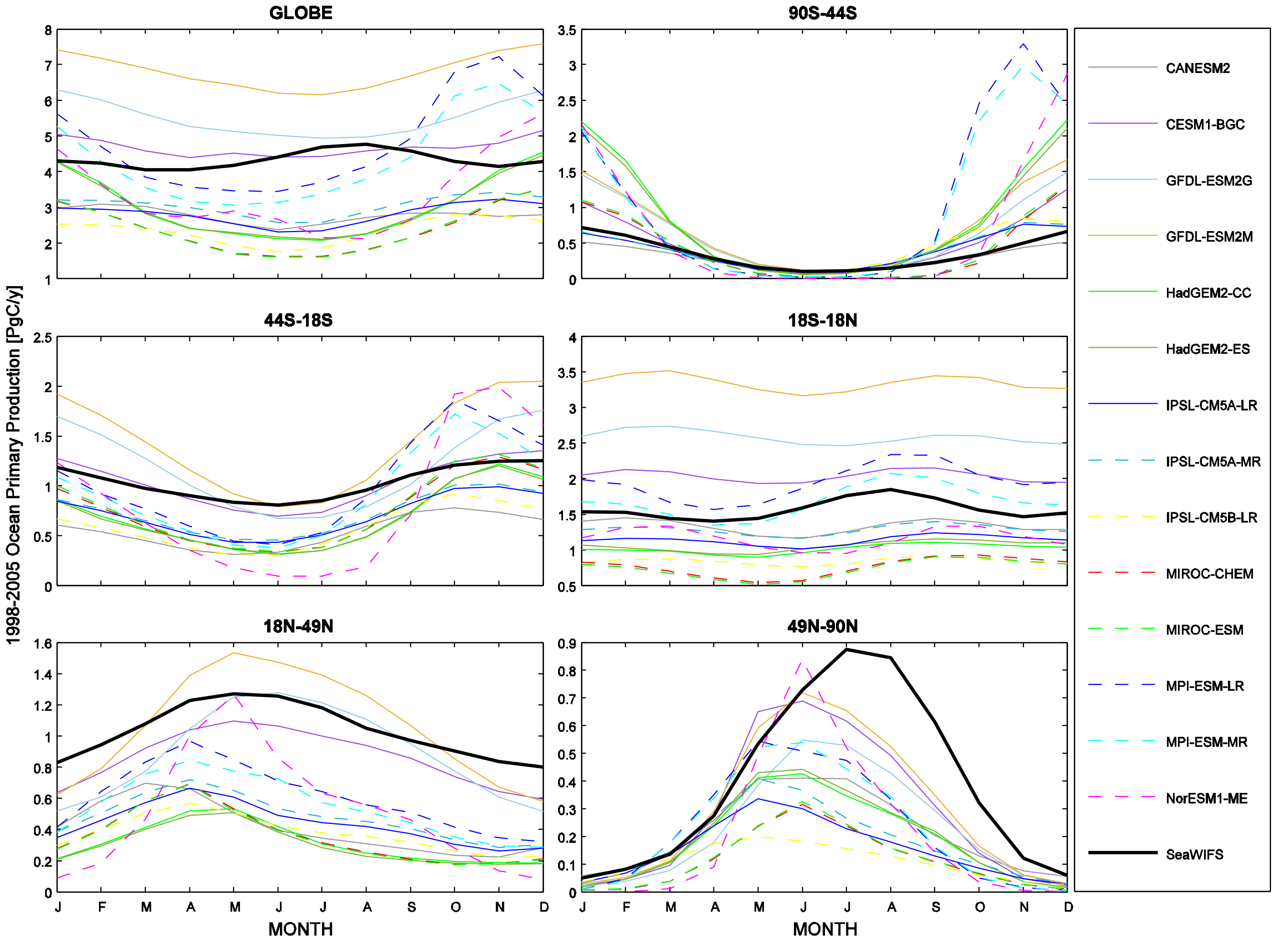

Figure 17. *Comparison of ocean primary production (PgC/y) mean annual cycle as simulated by CMIP5 models and SeaWIFS observations in the period 1998-2005.*

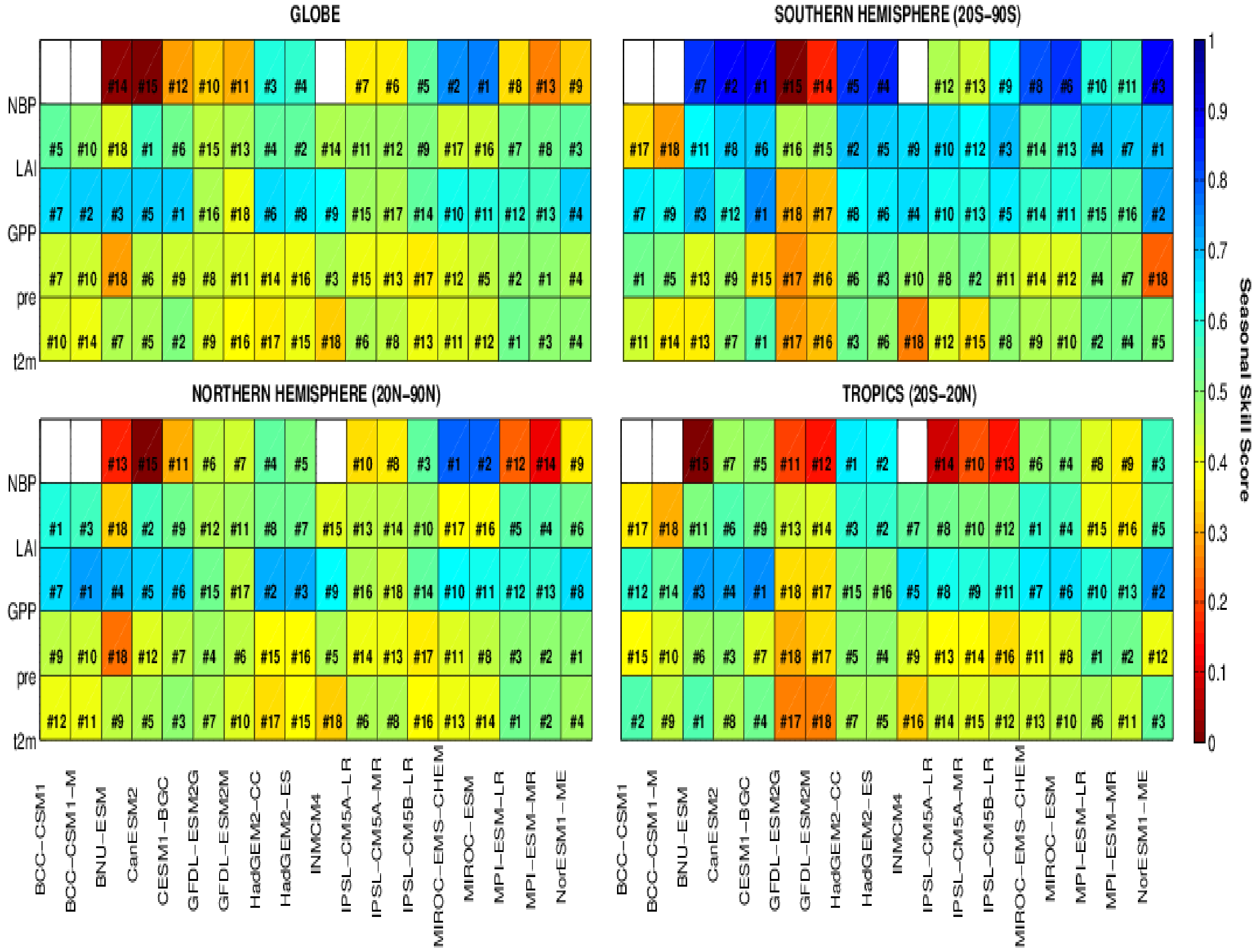

Figure 18. Seasonal skill score matrix as computed according to Equation 3 for the whole Globe, Southern Hemisphere (20°S-90°S), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). A score of 0 indicates poor performance of models reproducing the phase and amplitude of the reference mean annual cycle, while a perfect score is equal to 1.

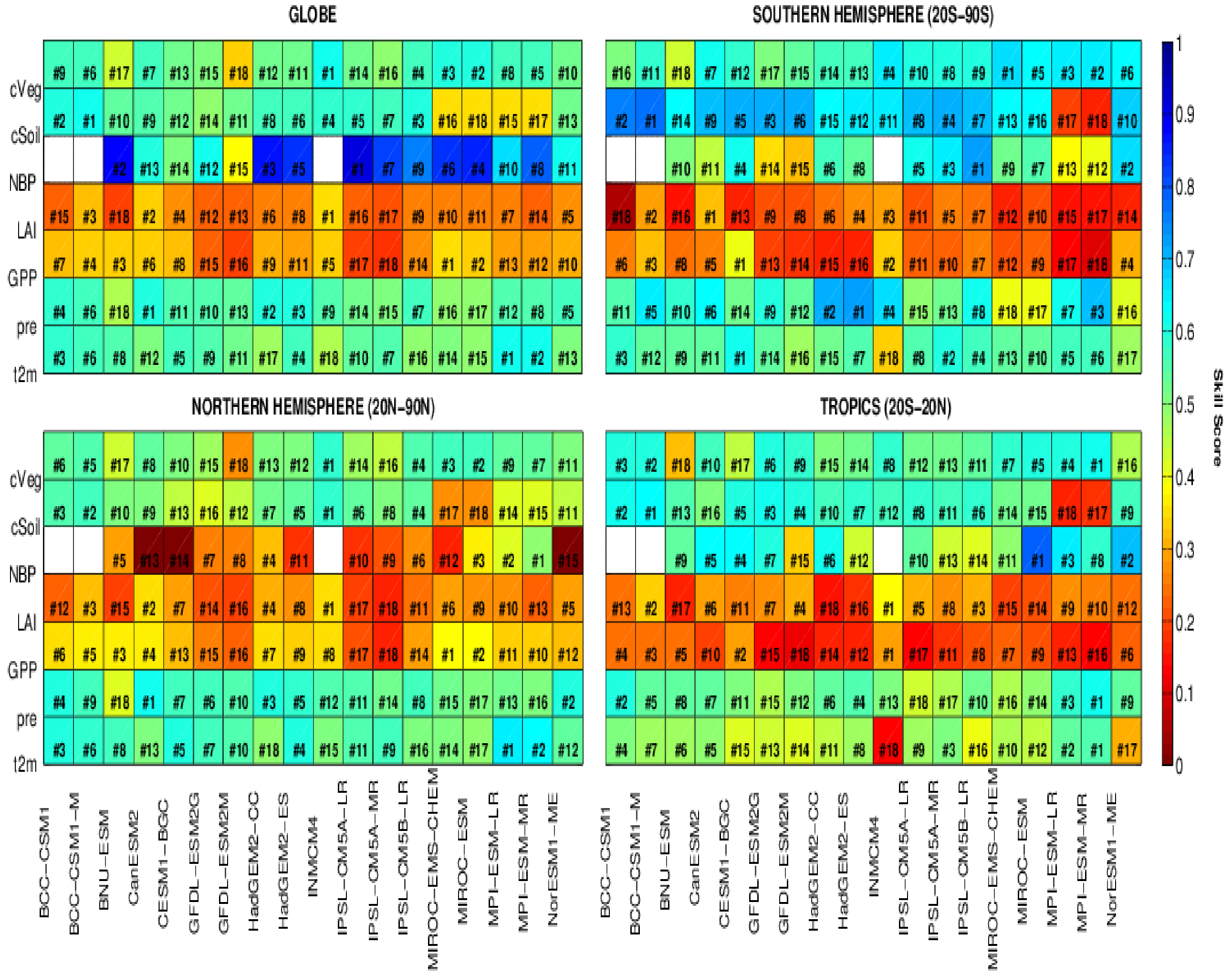

Figure 19. *PDF-based skill scores for temperature, precipitation, LAI, and NBP for the whole Globe, Southern Hemisphere (20°S-90°S), Northern Hemisphere (20°N-90°N), and Tropic (20°S-20°N). A perfect score is 1.*

Note that since the reference data for the soil and vegetation carbon pools are a single annual data, we were unable to build the PDF, therefore the skill scores for these variables are based on the normalized mean bias between the model and the reference data (see equation 6).

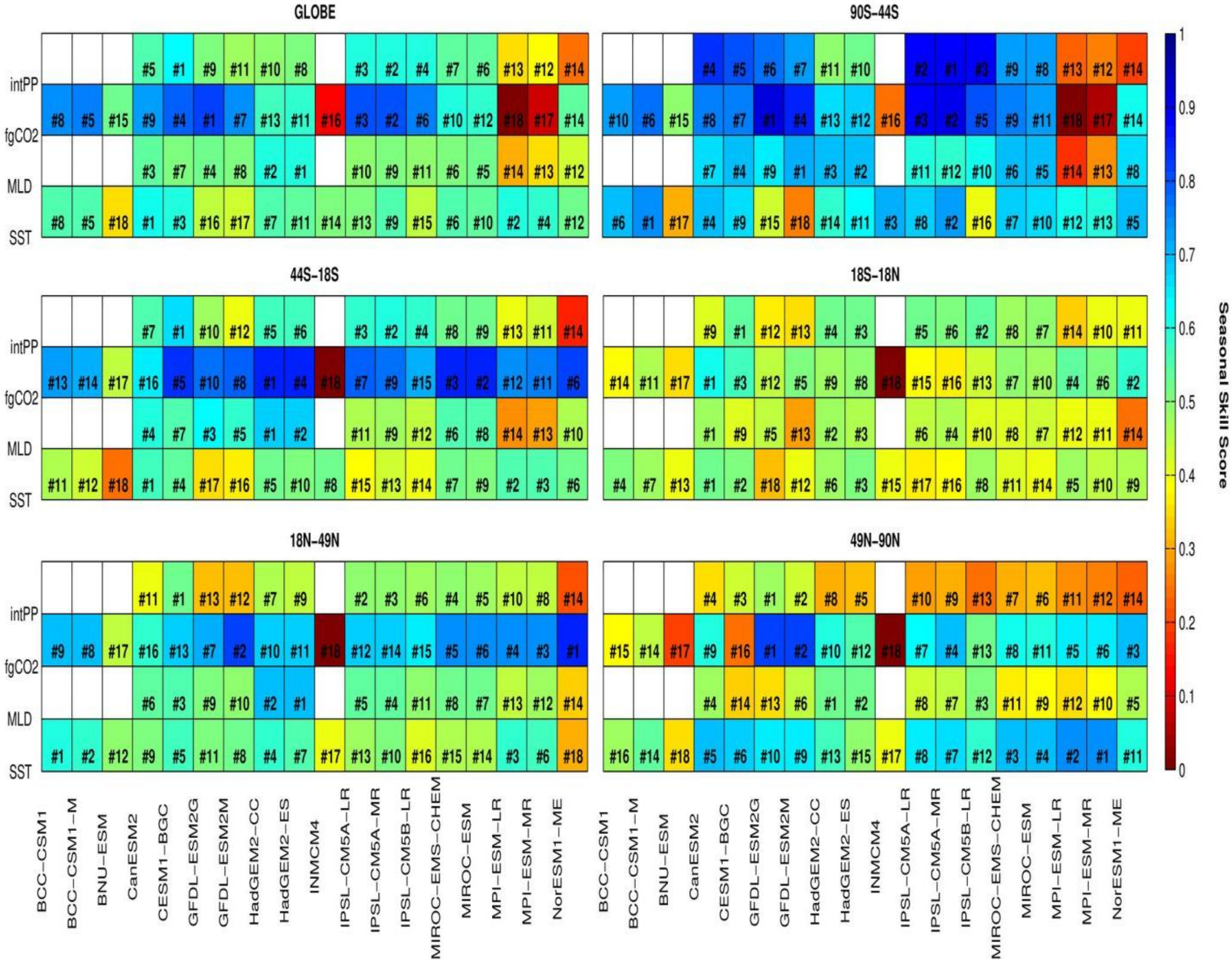

Figure 20. *As Figure 18 but for the ocean variables.*

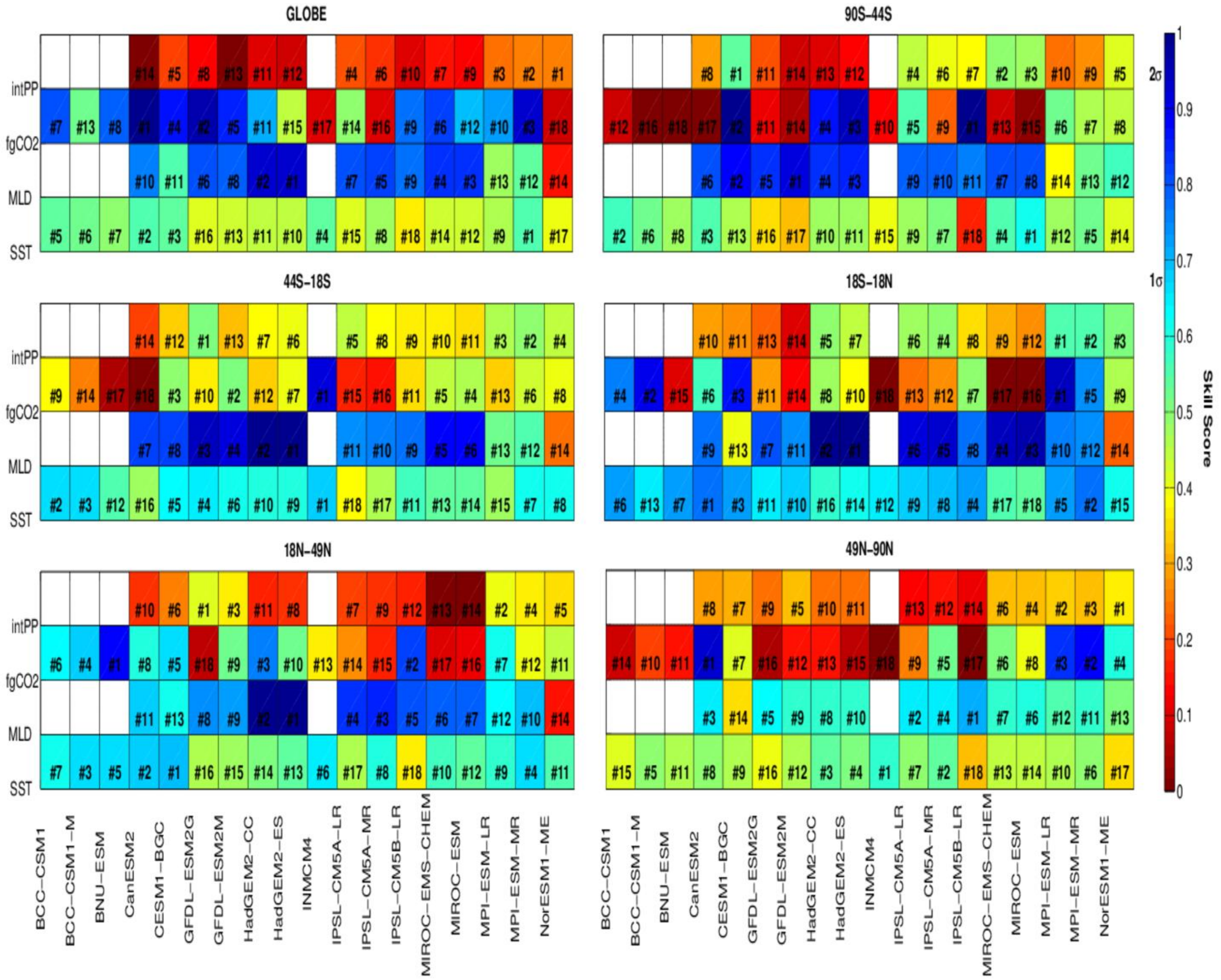

Figure 21. *As Figure 19 but for the ocean variables. Note that since the MLD dataset is a climatology we were unable to compute the PDF, consequently the skill scores have been computed according to equation 6.*