Evaluation of CMIP5 Earth System Models in Reproducing Leaf Area Index and Vegetation Cover over the Tibetan Plateau

BAO Yan[1,2] (鲍 艳), GAO Yanhong[2*] (高艳红), LÜ Shihua[2] (吕世华), WANG Qingxia[2] (王青霞), ZHANG Shaobo[2] (张少波), XU Jianwei[2] (许建伟), LI Ruiqing[2] (李瑞青), LI Suosuo[2] (李锁锁), MA Di[2] (马 迪), MENG Xianhong[2] (孟宪红), CHEN Hao[2] (陈 昊), and CHANG Yan[2] (常 燕)

1 *Earth System Modeling Center, Nanjing University of Information Science & Technology, Nanjing* 210044

2 *Key Laboratory of Land Surface Process and Climate Change in Cold and Arid Regions, Chinese Academy of Sciences, Lanzhou* 730000

(Received March 19, 2014; in final form July 21, 2014)

ABSTRACT

The abilities of 12 earth system models (ESMs) from the Coupled Model Intercomparison Project Phase 5 (CMIP5) to reproduce satellite-derived vegetation biological variables over the Tibetan Plateau (TP) were examined. The results show that most of the models tend to overestimate the observed leaf area index (LAI) and vegetation carbon above the ground, with the possible reasons being overestimation of photosynthesis and precipitation. The model simulations show a consistent increasing trend with observed LAI over most of the TP during the reference period of 1986–2005, while they fail to reproduce the downward trend around the headstream of the Yellow River shown in the observation due to their coarse resolutions. Three of the models: CCSM4, CESM1-BGC, and NorESM1-ME, which share the same vegetation model, show some common strengths and weaknesses in their simulations according to our analysis. The model ensemble indicates a reasonable spatial distribution but overestimated land coverage, with a significant decreasing trend (–1.48% per decade) for tree coverage and a slight increasing trend (0.58% per decade) for bare ground during the period 1950–2005. No significant sign of variation is found for grass. To quantify the relative performance of the models in representing the observed mean state, seasonal cycle, and interannual variability, a model ranking method was performed with respect to simulated LAI. INMCM4, bcc-csm-1.1m, MPI-ESM-LR, IPSL CM5A-LR, HadGEM2-ES, and CCSM4 were ranked as the best six models in reproducing vegetation dynamics among the 12 models.

Key words: Coupled Model Intercomparison Project Phase 5 (CMIP5), vegetation cover, earth system model (ESM), dynamic global vegetation model (DGVM), Tibetan Plateau

Citation: Bao Yan, Gao Yanhong, Lü Shihua, et al., 2014: Evaluation of CMIP5 earth system models in reproducing leaf area index and vegetation cover over the Tibetan Plateau. *J. Meteor. Res.*, **28**(6), 1041–1060, doi: 10.1007/s13351-014-4023-5.

1. Introduction

Terrestrial ecosystems substantially affect near-surface thermal and hydrological fluxes, as well as the greenhouse gas exchange between the land surface and atmosphere. The vegetation cover effects can be biophysical and biochemical. On one hand, changes in vegetation biomass and coverage between vegetated and bare land can affect the land surface albedo and evapotranspiration, which, in turn, modify near-ground climatic characteristics, such as temperature and precipitation; on the other hand, changes in tree cover strongly affect the amount of carbon stored in biomass and the soil, which alters the atmospheric CO_2 concentration and operates as a biogeochemical feedback mechanism between vegetation dynamics and the climate (Arneth et al., 2010; Bathiany et al., 2010; Port et al., 2012).

Supported by the National Basic Research and Development (973) Program of China (2010CB950503 and 2013CB956004) and Research Fund for Climate Change of the China Meteorological Administration (CCSF201403).

*Corresponding authors: gaoyh@lzb.ac.cn.

In a modeling system, changes in vegetation biomass and coverage affect the simulated climate in future climate projections through biophysical and biogeochemical effects. In the earth system models (ESMs) as part of the Coupled Model Intercomparison Project Phase 5 (CMIP5), a dynamic global vegetation model (DGVM) is often included, which calculates interactive vegetation variation (biomass and coverage) due to climate change simulated by the atmospheric model component (Collins et al., 2011; Watanabe et al., 2011). Since DGVMs are driven by atmospheric models, and simulated biases inevitably exist in these atmospheric components, the subsequent simulations of vegetation dynamics are also far from perfect.

These biases need to be quantified. Model quantification with several metrics is a good option to express the simulation quality from a variety of different aspects. Recently, a number of studies have been carried out for assessing land surface models with several metrics involving carbon and hydrological characteristics (Abramowitz et al., 2008; Randerson et al., 2009; Cadule et al., 2010; Blyth et al., 2011; Anav et al., 2013a). Although these studies are good examples of multi-model assessment, their quantification approaches cannot clearly identify the best and worst models for reproducing vegetation biological features. A ranking approach based on different variables could solve this problem. Brunke et al. (2003) developed a ranking scheme to score the multi-bulk aerodynamic algorithms in computing ocean surface turbulent fluxes. Decker (2012) applied this approach to rank the bias and standard deviation of errors between reanalysis products and flux tower measurements. Wang and Zeng (2012) extended this ranking approach to all four statistical quantities, i.e., correlation coefficient (ρ), ratio of standard derivations ($\sigma_{\rm r}/\sigma_{\rm obs}$), standard deviation of differences ($\sigma_{\rm d}$), and mean bias (BIAS), computed from surface meteorological variables, and then ranked the six reanalysis datasets in reproducing climate features over the Tibetan Plateau (TP). Anav et al. (2013b) ranked 18 models from CMIP5 with averaged seasonal cycles and probability density functions (PDFs) of ocean carbon and land carbon. These studies provide good examples of model performance identification.

The TP is one of the highest plateaus on the earth. It has a unique alpine vegetation composition and climatic features, along with a low intensity of human disturbance, which makes the TP an ideal place to study the response of vegetation variation to climate change. Previous analysis of climate records has shown that the TP has experienced very substantial climate change in recent decades, and temperatures are projected to continue increasing throughout the remainder of the present century according to global climate models (Piao et al., 2010). According to the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4), by the end of the 21st century, the temperature of this highland region will have risen by at least 1.8–4.0℃ (IPCC, 2007). Under this warming scenario, the vegetation of the TP is expected to actively respond (Kato et al., 2004; Piao et al., 2006). Piao et al. (2009) and Zhang et al. (2009) pointed out that terrestrial ecosystems over the TP act as a small carbon sink. These ecosystems are highly sensitive to temperature, and any rise in temperature will cause loss of biomass in the dominant biomes of the TP (e.g., alpine steppe and alpine meadow; Tan et al., 2010). Therefore, it is important to explore potential future changes in vegetation variation on the TP in response to climate change in order to develop appropriate countermeasures. The ESMs coupled with DGVMs in CMIP5 can produce the climate and vegetation distribution, and provide evolutionary information for both the historical period and the future. These simulations and projections can be used to understand the potential vegetation response to climate change. However, to improve the reliability of such predictions, model evaluations are essential.

This study aims to assess 12 ESMs from CMIP5 in reproducing the vegetation dynamics and biological properties over the TP. In Section 2, the DGVMs coupled with ESMs and the datasets used for the model validation, as well as the evaluation approaches applied, are described. The performance of the 12 ESMs is then reported in Section 3. The model ranking results for leaf area index (LAI) with respect to three skill score metrics are presented in Section 4. In Section 5, a summary and discussion of the key findings is provided.

2. Data and methods

2.1 Vegetation biological variables of ESMs

Twelve ESMs with vegetation dynamic distribution (including both DGVM outputs and non-DGVM outputs) from CMIP5 were selected in this study. Table 1 lists the model names and summarizes the components and characteristics of each ESM. In terms of the land surface, apart from bcc-csm1.1-m and INMCM4, all of the models account for land use change; likewise, apart from BNU-ESM, NorESM1-ME and CESM1-BGC, none of the models have an interactive land nitrogen cycle. Several biological variables such as LAI, plant functional types (PFTs), net primary productivity (NPP), gross primary productivity (GPP), and the climatic fields related to vegetation growth and terrestrial carbon budget such as precipitation ($P_{\rm r}$) and surface air temperature ($T_{\rm as}$), are provided in the models' outputs. These variables are not available for every model, e.g., NPP for INMCM4 is unavailable, GPP for INMCM4 is available but with unrealistic values (most GPP values are around zero), and only 6 of the 12 ESMs provide outputs of PFTs. Table 2 lists the variables used in our analysis (marked by asterisks).

Although the vegetation models in this study differ in their representations of vegetation types, soil

Table 1. Details of the models and model components used in this study

Model acronym	Full model name	Land/vegetation model	Land horizontal resolution	Dynamic vegetation
bcc-csm1.1-m	Beijing Climate Center Climate System Model	BCC-AVIM1.1	1° × 1°	N
BNU-ESM	Beijing Normal University Earth System Model	CoLM/BUN-DGVM(CN)	2.8° × 2.8125°	Y
CanESM2	The second generation Canadian Earth System Model	CLASS2.7	2.8° × 2.8125°	N
CCSM4	Community Climate System Model, version 4.1	CLM4/CLM4CN	0.9° × 1.25°	N
CESM1-BGC	Community Earth System Model- geochemistry		0.9° × 1.25°	N
GFDL-ESM2G	Geophysical Fluid Dynamics Laboratory Earth System Model, version 2, with Generalized Ocean Layer Dynamics (GOLD) model as ocean component (ESM2G)	LM3/LM3V	2.5° × 2°	Y
HadGEM2-ESM	Hadley Centre Global Environment Model version 2 - Earth System	JULES/TRIFFID	1.875° × 1.25°	Y
IPSL-CM5A-LR	Climate model of Laboratory of Meteorological Dynamic with NEMO for the ocean, Institute Pierre Simon Laplace	SVAT/ORCHIDEE	3.75° × 1.875°	Y
MIROC-ESM	Earth System Model of Model for Interdisciplinary Research on Climate	MATSIRO/SEIB-DGVM	2.8° × 2.8125°	Y
MPI-ESM-MR	The Max Planck Institute for Meteorology	JSBACH/BETHY	2.8° × 2.8125°	Y
NorESM1-ME	The Norwegian Earth System Model that includes prognostic biogeochemical cycling	CLM4/CLM4CN	2.5° × 1.875°	N
INMCM4	Institute for Numerical Mathematics	No name	1.5° × 2°	N

Note: "Y" means "yes"; "N" means "no".

Table 2. Variables provided by the 12 CMIP5 ESMs

Model	P_r	T_{as}	LAI	NPP/ GPP	Run numbers	Vegetation fraction		
						Bare ground	Tree	Grass
bcc-csm1.1-m	*	*	*	*	3			
BNU-ESM	*	*	*	*	1	*	*	*
CanESM2	*	*	*	*	5			
CCSM4	*	*	*	*	6			
CESM1-BGC	*	*	*	*	1			
GFDL-ESM2G	*	*	*	*	1		*	*
HadGEM2-ESM	*	*	*	*	5	*	*	*
IPSL-CM5A-LR	*	*	*	*	4	*	*	*
MIROC-ESM	*	*	*	*	3	*	*	*
MPI-ESM	*	*	*	*	3	*	*	*
INMCM4	*	*	*		1	*	*	*
NorESM1-ME	*	*	*	*	1			

Grid cells marked by an asterisk indicate that the data were available for the model. For INMCM4, GPP was provided but with unreasonable values of around zero, so these were neglected in our analysis.

properties, carbon, and nitrogen pools, as well as their horizontal and vertical resolutions at the surface and in the atmosphere and ocean, there are some similarities in their treatment of vegetation cover and the terrestrial carbon cycle: plants are categorized into several different PFTs, and the parameterizations for leaf photosynthesis, autotrophic respiration, carbon allocation, and phenology are similar across the models, although their specific parameters and limiting conditions are different. For instance, in most of the ESMs, the leaf carbon pool (C_{leaf}) is related to LAI via the equation

$$\text{LAI} \propto C_{\text{leaf}} \times \text{SLA}, \tag{1}$$

where SLA is the specific leaf area, which is either a PFT-specific constant or a value that varies along the vertical gradient in the canopy (Thornton and Zimmermann, 2007).

Our analysis focuses on the period 1950–2005 (the concentration-driven historical period) referred to as the "20th-century simulation" period. The last 20 years of this period (1986–2005) form a particular focus due to the reliable and complete observational record during this time, which is suitable for comparison purposes. Although the CMIP5 archive includes daily means for a selection of variables, only the monthly-mean output is used in our analysis, since this temporal frequency is high enough to provide a reasonably comprehensive picture of model performance both in terms of the mean state of the system, the trend, and its seasonal and interannual variability.

Before performing the evaluation, the model outputs and observations were preprocessed, including the elimination of snow-cover effects by identifying and filtering the snow-mask points in both the model simulations and observations, and regridding them into a common resolution of 1° ×1°, with only fractional land in grid cells considered. It should be noted that some of our selected models performed more than one experiment and generated several groups of outputs (see Table 2). To ensure the reliability of our analysis, all the data available from the CMIP5 data portal were collected and integrated into the model ensemble.

2.2 Validation data

Data from a variety of sources are used in our model evaluation. Table 3 lists the data sources, temporal and spatial resolutions, and regional mean values over the TP and standard deviations (SD) provided as the reference uncertainty (the actual data uncertainties are generally larger than the standard deviations provided here).

2.2.1 Precipitation and surface air temperature

The monthly precipitation (P_r) and surface air temperature T_{as} data of 2400 meteorological stations from 1961 to 2010 covering the whole of the Chinese mainland (Wu and Gao, 2013) are used for the model evaluation. It is found that precipitation over the TP has an annual mean value of 1.13 mm day^{-1} with an SD of ±0.12 mm day^{-1} (11%), while surface air tem-

Table 3. Observed data used for the model evaluation

Variable	Type	Temporal coverage	Spatial resolution	Mean (SD)
$P_{\rm r}$ (mm day^{-1})	Station	Monthly (1960–2011)	0.5° × 0.5°	1.13 (±0.12)
$T_{\rm as}$ (°C)				–1.3 (±0.51)
LAI	GLASS	8 days (1982–2011)	5 km (AVHRR)/	Annual 0.44 (±0.06)
			1 km (MODIS)	GS 0.66 (±0.09)
NPP	IGBP	Climatology (1986–1995)	0.5° × 0.5°	150.25
(gC m^{-2} yr^{-1})	MODIS	Annual (2000–2005)	0.5° × 0.5°	116.25 (±8.06)
GPP	MTE	Monthly (1982–2011)	0.5° × 0.5°	143.69 (±9.49)
(gC m^{-2} yr^{-1})	MODIS	Annual (2000–2005)	0.5° × 0.5°	246.25 (±14.18)
Vegetation	MODIS/CLM4	Climatology	0.5° × 0.5°	BGD 58.95
fraction (%)				TRE 8.96
				GAS 21.89
	MODIS/VCF	Monthly (2000–2001)	8 km	TRE 6.72 (±1.1)

Standard deviation (SD) is computed from the interannual variations of the regional average over the TP. BGD, TRE, and GAS indicate bare ground, tree, and grass, respectively. GS indicates growing season (April–October).

perature has an annual mean of –1.3°C with a larger SD range than precipitation of ±0.51°C (39%).

2.2.2 LAI, NPP, and GPP

The LAI product of the Global Land Surface Satellite (GLASS) dataset is generated from the Advanced Very High Resolution Radiometer (AVHRR) (1982–1999) and the Moderate Resolution Imaging Spectroradiometer (MODIS) reflectance data (2000–2011) using general regression neural networks (GRNNs) (Liang et al., 2013; Xiao et al., 2014). Different from the existing neural network methods that use only remote sensing data acquired at a specific time to retrieve LAI, the reprocessed MODIS reflectance data for an entire year were input into the GRNNs to estimate 1-yr LAI profiles. The MODIS reflectance product (MOD09A1) provides surface reflectance for each of the MODIS land spectral bands with a 500-m spatial resolution and an 8-day temporal sampling period. The AVHRR reflectance data are from NASA's Land Long Term Data Record (LTDR) project, which reprocessed Global Area Coverage (GAC) data from AVHRR sensors onboard NOAA satellites and created a daily surface reflectance product on a 0.05° spatial resolution. The maximum value composite (MVC) approach is used to composite the daily surface reflectance data into composites of 8-day intervals in order to maintain a consistent time resolution with MODIS surface reflectance data. The time series of red and near-infrared (NIR) reflectance data of AVHRR and MODIS are used to generate the GLASS LAI product. The data quality of the MODIS and AVHRR images is greatly influenced by clouds, cloud shadows, snow, and other abnormal climate conditions, which hinder the surface reflectance inversion and further impact the quality of the GLASS products. Some data, such as AVHRR, MOD09A1, MOD09GA, MCD43B3, and MOD02, are preprocessed before being used to produce the GLASS products. To improve the data quality, the existing MODIS snow and cloud mask and the reflectance characteristics of the non-snow/cloud pixels are used in combination to identify pixels of snow, clouds and abnormal values. All of the identified values are filled by the clear pixel. GLASS LAI on the TP shows an annual mean of 0.44 with an SD of ±0.06 (13.6%), and a mean value for the growing season (GS; April–October) of 0.66 with an SD of ±0.09 (13.6%).

The International Geosphere Biosphere Programme (IGBP) Global NPP Model Intercomparison Data were used to compare with the simulated NPP. The IGBP NPP data were obtained from the website of the International Satellite Land Surface Climatology Project, Initiative II (ISLSCPII) with a resolution of 0.5°. The IGBP NPP data were derived from an original dataset containing both the gridded average NPP values from 17 global models of biogeochemistry (Cramer et al., 1999; Cramer, 2011) for 1986–1995 and their climatological average (http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1027). This dataset is popular for model validation (Dan et al., 2007; Fisher et al., 2008), although it is a simulation product. The IGBP NPP data show an annual mean NPP value of 150.25 g C m^{-2} yr^{-1} with a wide uncertainty range of 6%–100% according to the officially provided SD. To make our evaluation more precise, the annual mean NPP and GPP data from MODIS (Zhao et al., 2005, 2006, 2010) with a 1-km resolution covering the period 2000–2012 are also used. The data quality is affected by the uncertainties in descriptions of the biome type and meteorological input data, as well as in the algorithm that translates measured parameters into inferred process rates. It has also been indicated that these uncertainties may be large in some regions or during certain seasons (Zhao et al., 2005). It is found that MODIS NPP and GPP have mean values of 116.25 and 143.69 g C m^{-2} yr^{-1} respectively, with relatively small SD values of ± 8.06 g C m^{-2} yr^{-1} (7%) and ± 9.49 g C m^{-2} yr^{-1} (6.6%), suggesting that the MODIS products are of high quality. The GPP data derived from the global upscaling flux tower measurements on the global scale based on the model tree ensemble (MTE) approach with the relatively long time series (1982–2012) described by Jung et al. (2009, 2011) are also used for the validation. Jung et al. (2011) estimated the uncertainty of globally averaged GPP to be ± 6 kg C m^{-2} yr^{-1} ($< 15\%$). In the TP region, the GPP data show an annual mean value of 246.25 g C m^{-2} yr^{-1}, with a small and ideal SD range of ± 14.18 g C m^{-2} yr^{-1} (5.8%).

2.2.3 Vegetation cover

The Community Land Model version 4 (CLM4) vegetation fraction data (Lawrence et al., 2007, 2011) derived from MODIS (hereafter MODIS/CLM4), which cover the whole global land surface with a high resolution of $0.5° \times 0.5°$, are used to evaluate the fractional distributions of bare ground and two vegetation cover types: trees and grass. The MODIS Vegetation Continuous Fields (VCF) dataset for vegetation cover (hereafter MODIS/VCF) derived from Hansen et al. (2003) is also used for evaluation of tree and bare ground simulations in this study. These two datasets describe the land surface in fractions of vegetation cover types: woody vegetation (trees and shrubs), herbaceous vegetation (grasses and crops), and bare (non-vegetated) ground. This is similar to the way in which DGVMs describe the vegetation cover (or PFTs), which is why they are chosen for our evaluation. For comparison, the MODIS-derived and model-simulated PFTs are placed into three broad vegetation classes: bare ground, tree, and grass PFTs. Table 4 provides the details of these classifications. The tree fraction of MODIS/VCF shows a mean value of 6.72% with an uncertainty of 1.1% (14.8%). Note that in most of the ESMs, the anthropogenic land use is predetermined; in particular, the extent of pasture and cropland is prescribed, and the dynamic vegetation models used only affect the natural vegetation distribution. Therefore, only the evaluation of the coverage

Table 4. Correspondence between model and MODIS vegetation PFT classifications

MODIS-derived PFT Type	Acronym	Model type
Needleleaf evergreen tree–temperate	NET temperate	Tree
Needleleaf evergreen tree–boreal	NET boreal	
Needleleaf deciduous tree–boreal	NDT boreal	
Broadleaf evergreen tree–tropical	BET tropical	
Broadleaf evergreen tree–temperate	BET temperate	
Broadleaf deciduous tree–tropical	BDT tropical	
Broadleaf deciduous tree–temperate	BDT temperate	
Broadleaf deciduous tree–boreal	BDT boreal	
C3 arctic grass		Grass
C3 grass		
C4 grass		
Pasture		
Crop1		
Crop2		
Bare soil		Bare ground

of natural vegetation is performed here, and the land cover variation due to transitions from natural to anthropogenic vegetation, and vice versa, is not considered.

2.3 Evaluation approach

A series of analyses were conducted for evaluating and ranking the models. In the following, we describe the diagnostics used for model evaluation and the metrics used for model ranking.

2.3.1 Evaluation metrics

There are multiple metrics that can be used for evaluating the agreement between simulated and observed LAI and vegetation cover. Here, the square of the Pearson correlation coefficient (r^2) is used to quantify the spatial correlation between the vegetation distribution in the model and observation. In a linear approximation, this metric quantifies a fraction of variation explained by the model as

$$r^2=\frac{\left[\sum\limits_{i}^{N}W_i(M_i-\overline{M})(O_i-\overline{O})\right]^2}{\left[\sum\limits_{i=1}^{N}W_i(M_i-\overline{M})\sum\limits_{i=1}^{N}W_i(O_i-\overline{O})\right]^2},\qquad(2)$$

where M_i and O_i are the variables simulated by the model or observed in the grid cell i, and W_i is an areal weight of grid cell i ($\sum\limits_{i=1}^{N}W_i=1$). Here, we calculated W_i in the Pearson correlation coefficient (CC) equation based on the area of each grid associated with the central geographic latitude of each grid. In our case, the study region (i.e., the TP) covers 25°–40°N, where the values of W_i do not vary much and can almost be neglected. N is the total number of grid cells under evaluation.

The amplitude of the difference between two datasets is measured by using the root-mean-square error (RMSE):

$$\mathrm{RMSE}=\sqrt{W_i\sum_{i=1}^{N}(M_i-O_i)^2}.\qquad(3)$$

The two metrics, r^2 and RMSE, are calculated separately for LAI and each vegetation class. During this process, the model is evaluated at every grid point and then aggregated over the entire land surface of the TP.

2.3.2 Ranking metrics

The method described in Section 2.3.1 cannot fully identify the best and the worst models among the 12 ESMs with respect to TP vegetation simulation. Accordingly, a ranking method is used to assess the model performance with three ranking metrics. One metric is an upgraded version of Eq. (3), with the aim to check the annual seasonal cycle (in terms of monthly data) of vegetation features:

$$\mathrm{RMSE}^2_{m,i}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(M_{i,t}-\overline{M}_i\right)-\left(O_{i,t}-\overline{O}_t\right)\right]^2,\quad(4)$$

where t corresponds to the temporal dimension, and N is the number of months. Equation (4) can be normalized by the maximum to obtain the relative error (Re) as follows:

$$R\mathrm{e}=1-\frac{\mathrm{RMSE}^2_{m,i}}{\max(\mathrm{RMSE}^2_{m,i})}.\qquad(5)$$

The other metric, the model variability index (MVI), as introduced by Gleckler et al. (2008) and Scherrer (2011), designed to check the model's representation of the interannual features of the observation, is equated by the ratio of the standard deviation of the model means divided by the standard deviation of the observed means (see Eq. (6)), where $\sigma_{o,i}$ and $\sigma_{m,i}$ are the standard deviations of the annual time series of the model and observation for a given variable at each grid cell i, respectively. Perfect model-reference agreement would result in an MVI value approaching zero. This approach avoids the cancellation effects of a model when experiencing problems related to excessively large or small interannual variability (IAV) (Gleckler et al., 2008; Scherrer, 2011). It is a good method for assessing the difference between a single model and observation, providing a consistent standard for identifying the standard deviation of a single model.

$$\mathrm{MVI}=\left(\frac{\sigma_{m,i}}{\sigma_{o,i}}-\frac{\sigma_{o,i}}{\sigma_{m,i}}\right).\qquad(6)$$

In some model evaluation studies, an MVI value equal to 0.5 is considered as a good representation of IAV (Scherrer, 2011; Anav et al., 2013b). However, the MVI threshold can vary markedly, changing in a wide range due to differences in physical variables and study regions, especially biological variables. For example, Anav et al. (2013b) showed that, on the global scale, calculated MVI of LAI and GPP from 18 models of CMIP5 had ranges of 2–6 and 1–10, respectively, and the ranges were even wider in the Southern Hemisphere. Previous studies have shown that numerical models generally perform poorly over the TP than in other places of China due to the TP snow coverage, which may have caused large uncertainty. In this case, a much higher and more variable MVI is expected. To solve this problem, we normalize the MVI by the maximum to obtain the relative MVI (RMVI) (Eq. (7)). A value approaching 1 is considered to denote the best level of agreement between model and observation.

$$\mathrm{RMVI} = 1 - \frac{\mathrm{MVI}_{m,i}}{\max(\mathrm{MVI}_i)}. \tag{7}$$

It should be noted that normalizing the skill score calculations in this way only yields a measure of how good a given model is with respect to a particular reference dataset, and does not have any real meaning.

In addition, the bias between a given model (M) and the reference data (O) is also computed as the third skill score to check the main bias between ESM simulations and the observation (Eq. (8)).

$$B_i^m = |M_i - O_i|. \tag{8}$$

Similar to the other metrics, models are evaluated at every grid point and then aggregated over the entire land area of the TP.

3. CMIP5 model performance

LAI is defined as one side of the green leaf area per unit ground area in broad leaf canopies and as one half of the total needle surface area per unit ground area in coniferous canopies (Watson, 1947). It is an important indicator of vegetation state because it affects the radiative transfer process within the canopy, as well as evapotranspiration from the surface, and consequently modulates near-surface climate and atmospheric circulations (Kang et al., 2007). The mean values of vegetation distribution are useful for simply quantifying vegetation ecosystem—climate interactions, which may provide insight into model performance, as emphasized in some model intercomparison studies. Therefore, we begin by providing general information on the LAI simulations of the 12 ESMs.

Figure 1a shows the simulated LAI compared with the observed LAI derived from satellite data (GLASS). Large differences are found among the models. Apart from CanESM2 and INMCM4, the LAI in the growing season over 1986–2005 in the remaining 10 models is overestimated, with values scattered around the CMIP5 ensemble mean and ranging from 0.44 to 3.6. Unrealistically high LAI is found in BNU-ESM and GFDL-ESM2G, with average values greater than 3.0 over the TP. This result is consistent with previous LAI evaluation on the global scale (Anav et al., 2013b; Shao et al., 2013). The overestimation in BNU-ESM is universal around the globe, and the overestimation in GFDL-ESM2G may be caused by a flaw in the land surface model physics, which allows only coniferous trees to grow in cold climate in case large LAI contributed by coniferous trees establishes in areas where there should be tundra or cold deciduous trees (Anav et al., 2013a). CanESM2 shows a very small box range (or slight difference between minimum and maximum), suggesting weak interannual variability in the CanESM2 simulation.

For vegetation carbon flux above the ground (or NPP above the ground) during 1986–1995, Fig. 1b shows much more scattered distributions, with magnitudes varying from 125.8 to 554.86 g C m^{-2} yr^{-1}, implying large uncertainty among the models. With the exception of CanESM2 and NorESM1-ME, 9 of the 11 models (Note that the NPP of INMCM4 was not available in the CMIP5 data portal) tend to overestimate the IGBP NPP magnitudes (150.25 g C m^{-2} yr^{-1}). The systematic bias in NPP generally reflects the accuracy of the simulated LAI shown in Fig. 1a. Following the erroneous pattern of LAI, GFDL-ESM2G, and CanESM2 respectively show the largest and least bias, when compared with the IGBP NPP. For CCSM4 and CESM1-BGC that use CLM4 extended with a carbon-nitrogen (CN) biogeochemical model (hereafter CLM4CN) as their vegetation model, very similar overestimated NPP is simulated. How-

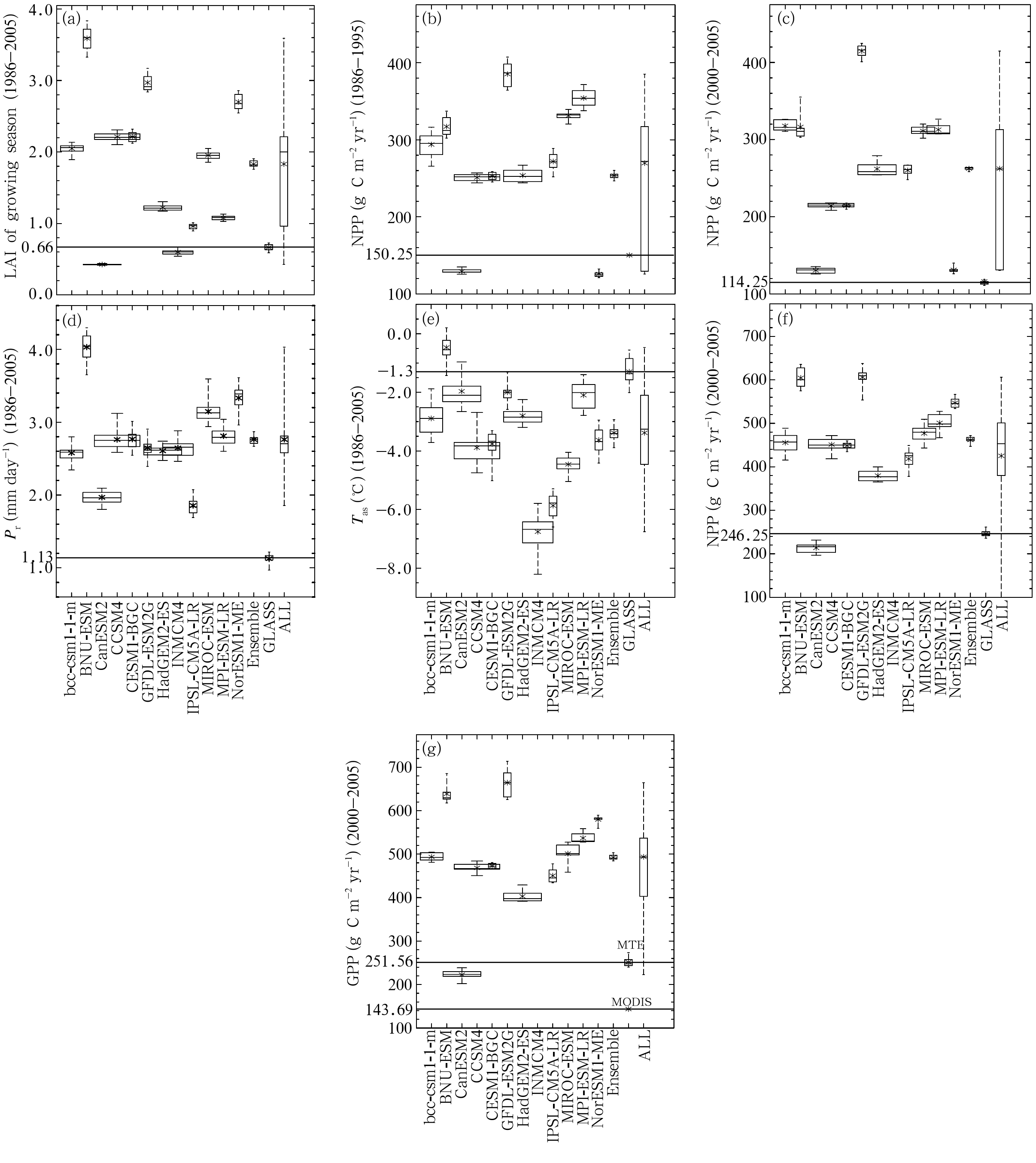

Fig. 1. Model statistics of LAI in the growing season (April–October), annual mean net primary productivity (NPP), precipitation (P_r), surface air temperature (T_as), and gross primary productivity (GPP). (a) LAI in the growing season over the reference period of 1986–2005; (b) annual mean NPP during 1986–1995; (c) NPP during 2000–2005; (d) P_r; (e) T_as; (f) GPP during 1986–1995; and (g) GPP during 2000–2005. Values from top to bottom of each box inside each panel are the maximum, 75 percentile, median, 25 percentile, and minimum of the model values during the evaluation period. The x-axis identifies the 12 models and the observation dataset. The box values of "ALL" are calculated based on the model sequences. Each box is marked (*) with the mean value of the individual model. The width of each box indicates the run numbers for every ESM. In (b, c) and (f, g), NPP and GPP for INMCM4 are missing because they were unavailable in the CMIP5 data portal. The black lines in each of the panels indicate the mean value of observations.

ever, NorESM1-ME, which also uses CLM4/CLM4CN as its land surface/vegetation model but forced by the revised Community Atmosphere Model version 4 (CAM4) (Neale et al., 2010) with CCSM4, an abnormally low NPP mean value is found, suggesting that the relationship between LAI and biomass production in this model is unrealistic. NPP derived from MODIS during 2000–2005 shows a much lower annual mean value (116.25 g C m^{-2} yr^{-1}) than the IGBP NPP mean during this period, which is severely overestimated by all of the 11 ESMs, and with a much larger range from 131.29 to 588.68 g C m^{-2} yr^{-1}—even for CanESM2, which is one of the two ESM models that underestimate the observed LAI, as mentioned above. The simulated NPP during 2000–2005 shows a 4.2–65.3 g C m^{-2} yr^{-1} increase with a generally smaller interannual variability (see the much narrower box range) compared with that during 1986–1995, which is reasonable and consistent with the LAI variations.

The overestimation of LAI may be associated with an overestimation of observed precipitation. It is found that the erroneous pattern of simulated LAI conforms well with the simulated precipitation (Fig. 1d), with the exception of IPSL-CM5A-LR, GFDL-ESM2G, and INMCM4. The large bias found in BNU-ESM could be in some way due to the serious wet bias that this model has in reproducing the observed precipitation. The mean annual precipitation as reported by the station data is 1.12 mm day^{-1}, while BNU-ESM produces a value about 4 times as large (4.03 mm day^{-1}). CanESM2 shows a small and underestimated LAI, which is consistent with the relatively small wet bias in the model. The simulated surface air temperature does not contribute much to the overestimated LAI, since all the models apart from BNU-ESM generate a colder than observed atmosphere near the surface of the TP, which is not conducive for plant growth, vegetation photosynthesis, and carbon exchange between vegetation and the atmosphere.

GPP represents the uptake of atmospheric CO_2 during photosynthesis. Anav et al. (2013b) attributed the overestimated LAI in 18 ESM models from CMIP5 to two reasons. One is associated with the overestimated photosynthesis (or GPP), which could lead to a surplus of biomass stored into the leaves, and the missing parameterization of ozone also partially explains the LAI overestimation due to the high GPP, with the proof 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 (Sitch et al., 2007; Wittig et al., 2009). In our case, the 12 ESMs are found to seriously overestimate the MODIS (Fig. 1f) and MTE (Fig. 1g) GPP, with a similar erroneous pattern to LAI. Therefore, the overestimation of photosynthesis could be one of the possible reasons for the LAI overestimation. This highlights the wet bias of models since precipitation is another main limiting factor for plant photosynthesis across the globe besides temperature. The level of uncertainty in satellite-derived LAI is also another possible reason for the unrealistically overestimated LAI over the TP. This is because remote sensing data cannot represent the real LAI distribution spatially and temporally, although it has been widely recognized as a valuable tool for detection and analysis of LAI.

Figure 2 shows spatial distributions of the mean LAI of each model and their ensemble, as well as the GLASS observation in the growing season over the reference period. In general, each model reproduces the observed LAI pattern, with LAI decreasing from the southeast border where forests dominate, to the northwestern TP mainly covered by bare land (Yu et al., 2010). With the exception of CanESM2 and INMCM4, LAI in the southeastern TP is severely overestimated in the remaining models, especially BNU-ESM and GFDL-ESM2G, as mentioned when analyzing Fig. 1a. In addition, CCSM4, CESM1-BGC, and NorESM1-ME produce similar overprediction patterns in southeastern TP due to their use of the same land/vegetation model. The dark crosses in Fig. 2 mark the areas where the simulated interannual variability of LAI is reliably consistent with the observation ($p < 0.05$). It can be seen that the simulated interannual variability in most of the models has higher reliability in the eastern and southwestern TP compared to elsewhere. BNU-ESM and GFDL-ESM2G generate increasing LAI in agreement with the obser-

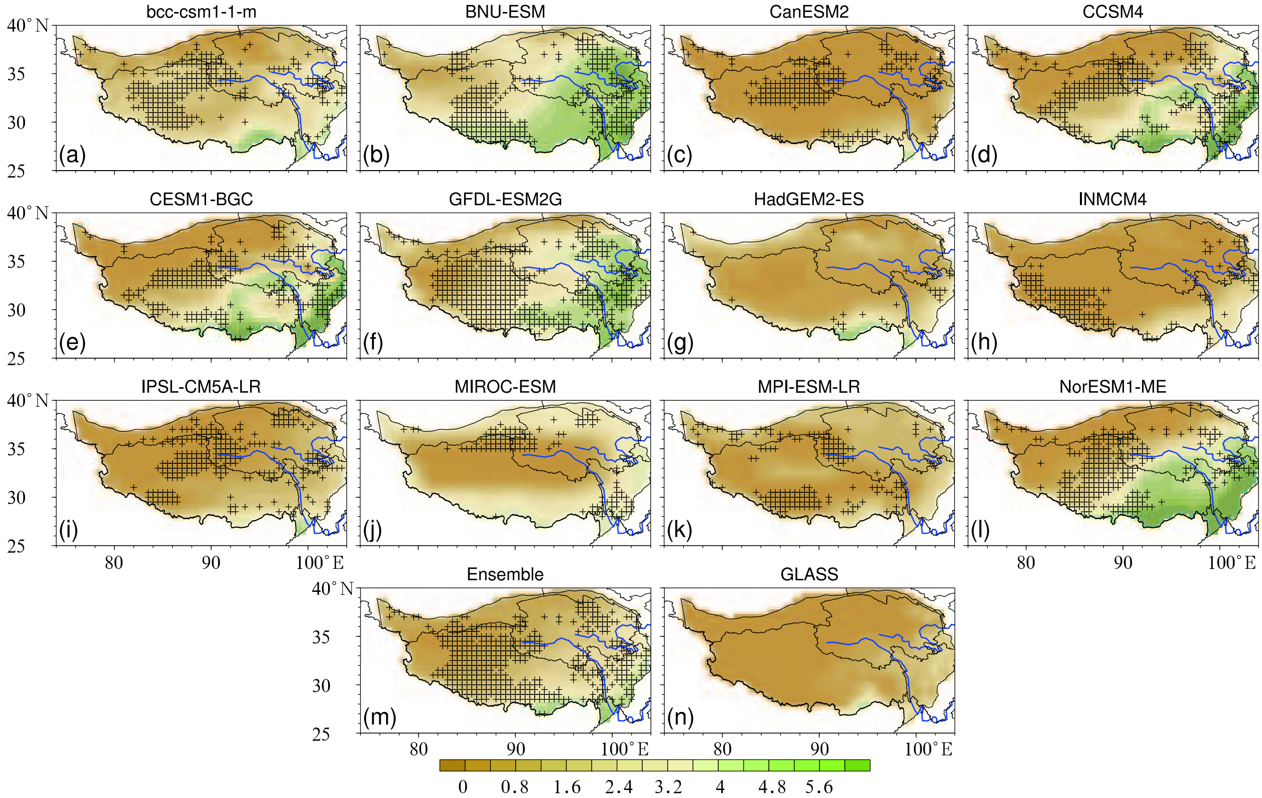

Fig. 2. Spatial distributions of simulated and observed LAI in the growing season. Figures (a–l) and (m) respectively indicate the simulated LAI for the 12 individual models and the model ensemble, while (n) indicates the observed LAI of GLASS. The hatched areas in (a–m) indicate the grids with statistically significant interannual change ($p < 0.05$). The dark blue lines in each panel are the Yellow River (top) and the Yangtze River (bottom).

vation, albeit with unrealistically high values. However, HadGEM2-ES shows continuously descending LAI, which is opposite to the observed interannual variability (figure omitted), implying that there is something wrong in the vegetation dynamics of the HadGEM2-ES model.

Figure 3 compares the spatial distribution of the linear trend between the GLASS observation (Figs. 3a–c) and the 12-model ensemble (Figs. 3d–f). The observation shows a significant increasing tendency in 46% of the area of the TP ($p < 0.05$), with the trend < 0.15 per decade (bar plot in the bottom left) (Figs. 3a and 3b). The most noticeable increase can be seen at the eastern and southern borders, where the coverage is mainly forest (Yu et al., 2010) and where LAI increases by more than 0.15 per decade (Fig. 3a). However, 2.2% of the area with the significant decreasing tendency (between –0.1 and –0.15 per decade) can also be found in the upper reaches (or headstream) of the Yellow River and the edges of the northern and western TP, suggesting a degraded vegetation status there in the past 20 years.

The model ensemble shows a distinct increasing trend of LAI, with an expanded area of 82% over the TP and a comparable trend of 0.05–0.15 per decade in the observation (Figs. 3d and 3e). At the southern and eastern borders of the TP, the ensemble simulation shows a slightly small increasing trend (0.10 to –0.15 per decade) compared with the observation. However, the evident decreasing trend found in the upper reaches of the Yellow River is not clearly shown in the model ensemble due to the low resolution of the models.

The consistency of the model simulations with the

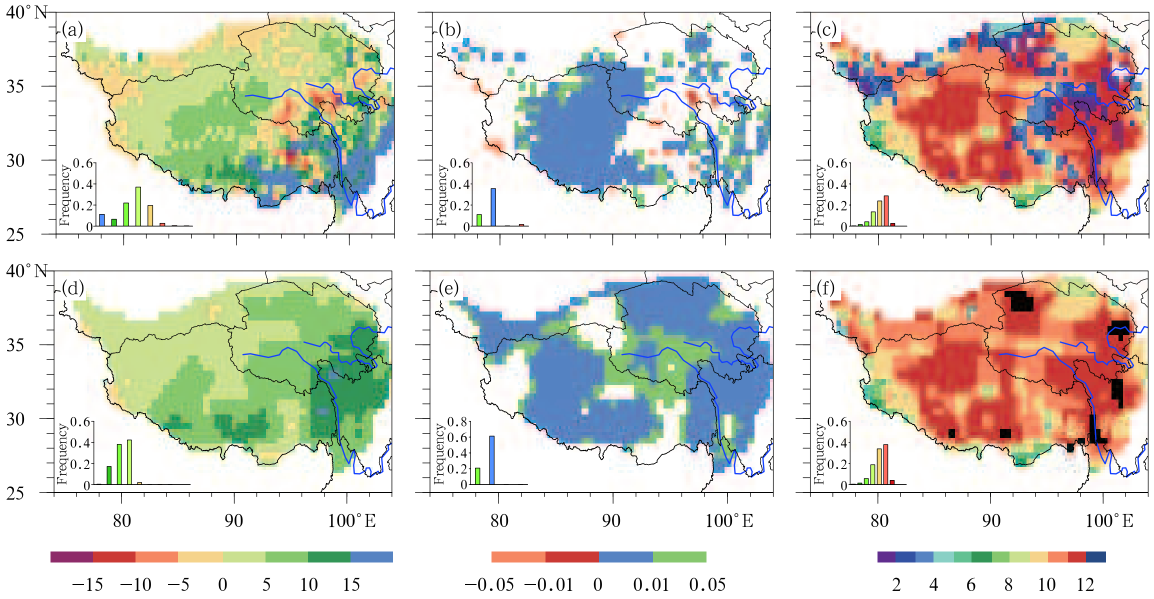

Fig. 3. Spatial distributions of the LAI linear trend, significance level, and consistency during the growing season from 1986 to 2005 in (a, b, c) the observation and (d, e, f) simulation: (a, d) LAI linear trend (10^{-3} $(10\ \mathrm{yr})^{-1}$); (b, e) significance level of the linear trend, in which the increasing/decreasing significance (p) is expressed quantitatively, with colored areas indicating two divided levels ($p < 0.01$ and $0.01 < p < 0.05$); (c) consistency of single models with the observation (GLASS); and (f) consistency of single models with the model ensemble. Filled colors in (c) and (f) indicate the cumulative number of models showing a trend (increasing or decreasing) that is consistent with the observation or the 12-ESM ensemble. The insets show the (a, d) frequency distributions of the corresponding trends, (b, e) different significance levels, and (e, f) model numbers greater than 6.

observation reflects the reliability of the simulations to a certain degree. It is indicated that the models can generally reflect the variation of observed LAI in the reference period, with 8 of the 12 models illustrating consistent increasing trends with the observation in more than 80% of the area (Fig. 3c). As mentioned above, the apparent disagreement exists at the northern border and headstream of the Yellow River, where the simulated LAI shows a strong increasing trend that is out of phase with the observation. The model simulations show better agreement with the model ensemble than that with the observation, with approximately 80% of the area possessing a coherent increasing tendency, simulated by more than 10 models.

Figure 4 shows the seasonal evolution of observed and simulated LAI for the entire TP. The GLASS observation indicates a clear interannual cycle, with the LAI magnitude showing a remarkable increase in May when the vegetation of the TP starts to green up, reaching its peak in June and July, before recovering to a low in the dormancy period after October. Most models show a seasonal cycle with its phase in agreement with the observation, except IPSL-CM5A-LR, which does not present a complete seasonal variation as the other models do, although the start of the growing season is not accurately simulated by the other models (e.g., HadGEM-ESM2 shows a one-month advance, and CanESM2 and INMCM4 a one-month delay). For GFDL-ESM2G and BNU-ESM, the simulated LAIs show unrealistically high values for the entire year, even during the dormancy seasons, with the former possessing a very small seasonal variability and the latter an extremely high LAI (maximum of 4.4) during summer (June–August). The extensive coverage of evergreen vegetation (trees, shrubs, or tundra) and the seasonal herbaceous vegetation or deciduous trees, are respectively considered to contribute to their overestimated LAI. Our assumption does not conflict

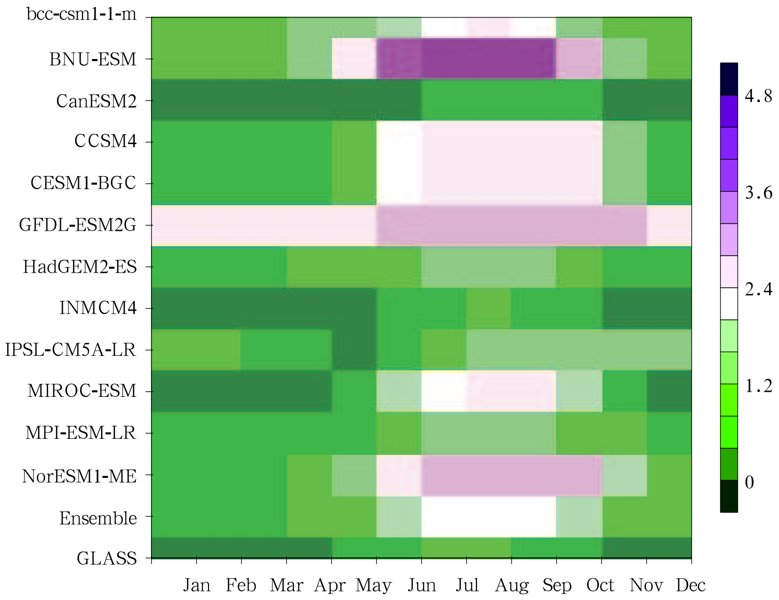

Fig. 4. Observed and simulated climatological monthly mean LAI during the reference period 1986–2005 for each calendar month. Color intensities reflect the magnitude of the climatological LAI mean. The x-axis corresponds to calendar months, and the y-axis indicates the 12 ESMs, its ensemble, and the observation (GLASS).

with the previous explanation for overestimated LAI in GFDL-ESM2G, since coniferous trees are also part of the evergreen vegetation. The underestimation of LAI in CanESM2 is probably related to the delayed start of the growing season. A similar pattern of seasonal variation is shown in CCSM4, CESM1-BGC, and NorESM1-ME simulations, which stresses the importance of vegetation model performance to simulated vegetation dynamics.

For vegetation cover, only three classes (bare ground, trees, and grass) are focused upon, since they have the widest coverage on the TP and the strongest effect on biophysical properties of the land surface (Brovkin et al., 2013). Figure 5a shows the area-averaged PFTs of bare ground, trees, and grass from model simulations during 1986–2005 compared with the observations of MODIS/CLM4. The models generally underestimate bare ground (except GFDL-ESM2G) and overestimate tree coverage (except IPSL-CM5A-LR), with very scattered grass coverage, although individual models capture well the observed PFTs (e.g., both HadGEM2-ES and MPI-ESM-LR simulate comparable coverages of bare ground and trees with the observation).

During the period from the middle of the 20th to the early 21st century, the model ensemble shows a slight increase in bare ground and decrease in tree coverage, with respective trends of 0.58% and –1.48% per decade (Fig. 5b). The simulated grass coverage does not show significant variations during this period. In the last two decades (1986–2005), the simulated tree coverage shows a continuous but much more moderate descending trend (–0.18% per decade), while the bare ground area turns to a slight decrease (–0.08% per decade); the sign of grass variation is still too weak to be discerned.

Figure 6 shows the MODIS/VCF and simulated spatial coverage for bare ground (Figs. 6a and 6d), trees (Figs. 6b and 6e), and grass (Figs. 6c and 6f). It can be seen from the observation that bare ground is mainly located in the northwestern TP, and there is almost no bare ground coverage at the southern and eastern borders (Fig. 6a). Tree coverage exists mainly at the southeastern edge of the TP, with a high fraction of over 80% in most of the area. In the southeastern TP, grass is the dominant biome type, with an area of coverage of 40.82%.

The simulated distribution of bare ground from

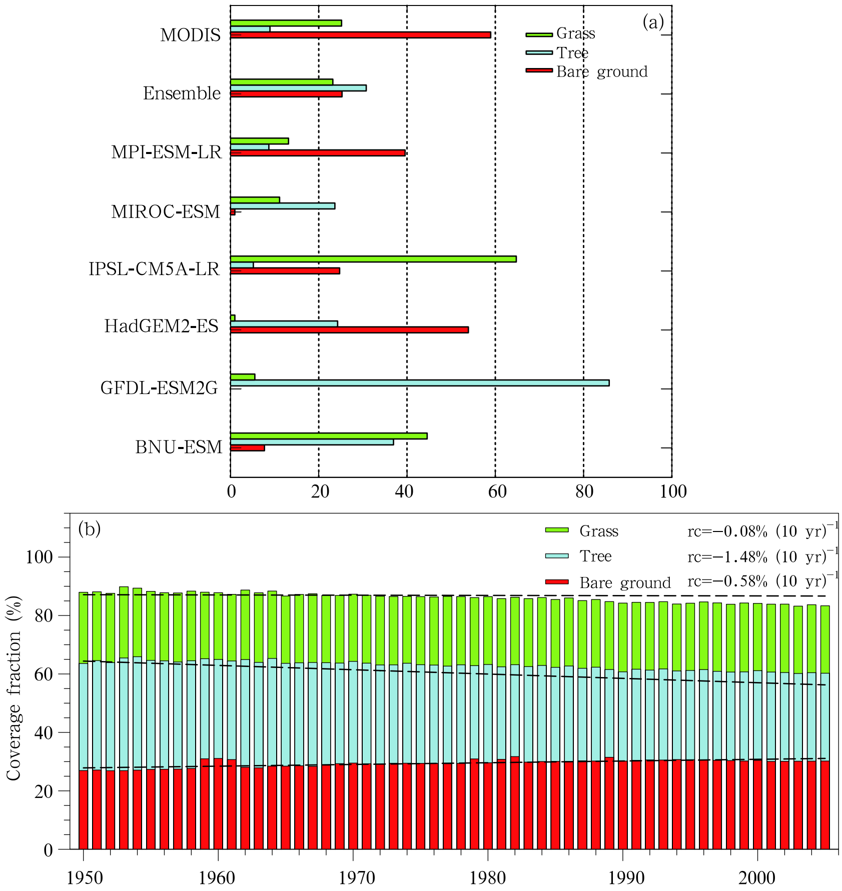

Fig. 5. Observed and simulated vegetation coverage. (a) Three PFTs over the TP. The tree fraction in HadGEM2-ES and bare ground fraction in MIROC-ESM are not shown since they are within 5%, which is considered to have insufficient accuracy. (b) Temporal variation of PFT coverage for the model ensemble over the period 1950–2005. The black dashed lines in (b) indicate the linear trends of the three PFTs, and "rc" indicates the trends of the PFTs.

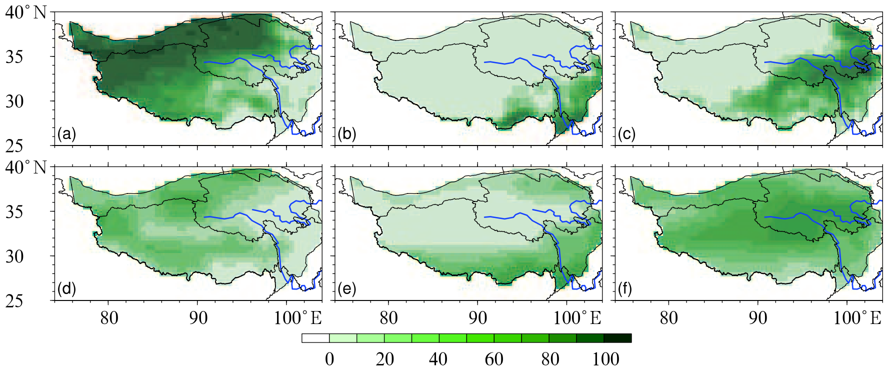

Fig. 6. Spatial distributions of (a, b, c) observed and (d, e, f) simulated vegetation coverage (%) during 2000–2001 for (a, d) bare ground, (b, e) trees, and (c, f) grass.

the model ensemble agrees well with the MODIS-derived coverage (Fig. 6d), albeit the magnitudes in most parts of the TP are underestimated by around 50%. The model ensemble reproduces the spatial distribution of observed tree coverage at the southern and eastern borders of the TP, with an extra northextending tree band at the northern edge of the TP (Fig. 3e) due to the overestimation of GFDL-ESM2G as described in Section 3.2. The observed grass fraction is not captured well by the model ensemble. Unlike the observed grass fraction, which is mainly located in the southeastern TP, the simulated grass fraction from the model ensemble evenly covers the entire TP, with a PFT fraction two times of that of the observation.

It is clear that CMIP5 ESMs cannot well reproduce the observed PFTs. Besides the systemic bias of PFTs in the models, the large differences between CMIP5 ESM outputs and observations may also be due to the ESMs' coarse spatial resolutions, which does not adequately represent the control that the complex topography has on the vegetation distribution of the TP, even when they are downscaled to a relatively high resolution of 1° × 1° as in our analysis. As a result, the heterogeneous spatial distributions of surface air temperature and precipitation are smoothed compared to observations, which then influence the resulting vegetation distribution.

To facilitate comparisons in a concise way, the square of the Pearson correlation coefficient (r^2) and the root-mean-square error (RMSE), as well as a Taylor diagram, are used to quantify the performance of the model simulations. Table 5 lists the values of r^2 and RMSE separately for LAI and the three PFTs mentioned above. Over the TP, r^2 and RMSE of LAI are equal to 0.69 and 1.25, respectively, which are comparable with the result of Brovkin et al. (2013) based on CMIP5 ESM simulations on the global scale. The model ensemble shows a relatively higher r^2 and lower RMSE values (0.59 and 14.48%) for tree fraction, when compared with bare ground and grass. The grass fraction, as an intermediate class between tree and bare ground coverage, is reproduced less reliably over the TP, with r^2 of 0.19 and RMSE of 31.61%.

Figure 7 shows the Taylor diagram (Taylor, 2001) derived from the standard deviations and correlation coefficients of LAI in the growing season. The distance from point to point (1.00, 1.00) in the Taylor diagram indicates the relative skill of the model. It can be seen that the normalized standard deviations spread out over a large range from 0.75 to 2.79. Most of the models show larger interannual variability than the observation, with the ratio of the standard deviation to the observation being more than 1, except for CanESM2 and MPI-ESM-LR, which also have smaller relative bias (< 50%) than other models, suggesting that these models perform well in reproducing the observed mean state. Of the 12 models, INMCM4 shows the closest value with the observation, while NorESM1-ME shows the most dispersed standard deviation from the observation.

The correlation coefficients reflect agreement between the model simulations and the observation in terms of spatial distribution. It is shown in Fig. 7 that the correlation coefficient values spread within the range of 0.65–0.90, with most models having a

Table 5. Evaluation of simulated LAI and vegetation coverage in terms of r^2 and RMSE

Variable	RMSE	r^2
LAI	1.25	0.69
Bare ground	43.86%	0.56
Tree	14.84%	0.59
Grass	31.61%	0.19

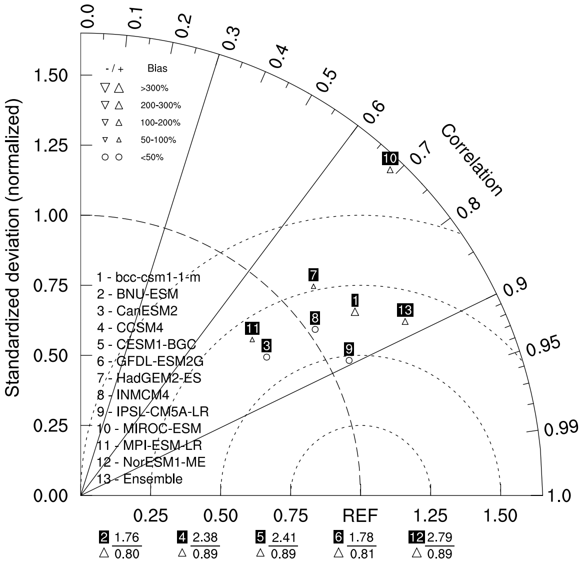

Fig. 7. Taylor diagram of LAI during the growing season for the reference period (1986–2005). The correlations and ratios of standard deviations among model simulations and the observation (GLASS) are calculated spatially.

high value of more than 0.8, except HadGEM2-ES and MIROC-ESM. Models CCSM4, CESM1-BGC, and NorESM1-M show relatively high correlation coefficient values of 0.89 in their simulations. This suggests good ability of CLM4/CLM4CN to represent observed spatial distributions of LAI.

4. Model ranking

The diagnostics in Section 3 indicate that, generally, the CMIP5 ESMs can adequately reproduce the observed biological characteristics of vegetation, although a few of the models do show notably poorer agreement than others, and general problems exist for quite a few of the models. The measures provide the basic information of model performance, which is crucial for identifying model differences in a model evaluation system. However, the diagnostics are not sufficient to clearly identify the best and worst models in model ensemble members. To achieve this objective, specific metrics were defined (see Eqs. (4)–(8)) and calculated to produce a quantitative ranking of the models.

Figure 8 shows the skill scores of simulated LAI in terms of the metrics defined in Eqs. (4)–(8). It is indicated that CanESM2, MPI-ESM-LR, and INMCM4 posses the best skills for reproducing observed amplitude and seasonal evolution (see "R_e" of Figs. 8a and 1a), although it seems to be two months out of phase with the observation during the peak season for CanESM2 (Fig. 4). The poorest performance is found in BUN-ESM and NorESM1-ME, mainly due to their great discrepancies in LAI magnitude with the observation of GLASS (Fig. 1a). The relatively low skill score for IPSL-CM5A-LR (ranking number of 7) is related to its bad representation of seasonal evolution, since the simulation is totally out of phase compared with the observation (Fig. 4). Our results are consistent with another study based on CMIP5 biological variables on the global scale (Anav et al., 2013b).

The models show very different skill scores when ranked with respect to interannual variability, e.g., bcc-csm1.1-m and MPI-ESM-LR achieve the top two scores among all 12 models (see "RMVI" in Fig. 8a). The simulations of CCSM4, CESM1-BGC, as well as NorESM1-ME, all achieve high scores, indicating good ability of CLM4 to represent interannual variability. Among the 12 models, CanESM2 and GFDL-ESM show the worst simulation skills in reproducing interannual variability of LAI.

Figure 8b shows an absolute measure of the ESMs' skills in reproducing the mean state of observed LAI. As in Figs. 1a and 7, CanESM2 and INMCM4 present the smallest bias of the 12 ESMs, when compared with the observed LAI. It is not surprising that GFDL-ESM and BNU-ESM show very poor skill, which in both cases is related to the improper description of land surface model physics and the large wet bias, as previously mentioned.

Finally, we calculated the arithmetic product of weights according to the ranking orders for Re, RMVI, and BIAS, with the ranking number greater than 10 (here larger numbers mean poorer simulation skill) out of the lists, and then obtained the ranking sequence for the model group. It was found that INMCM4, bcc-csm1-m, MPI-ESM-LR, IPSL-CM5A-LR, HadGEM2-

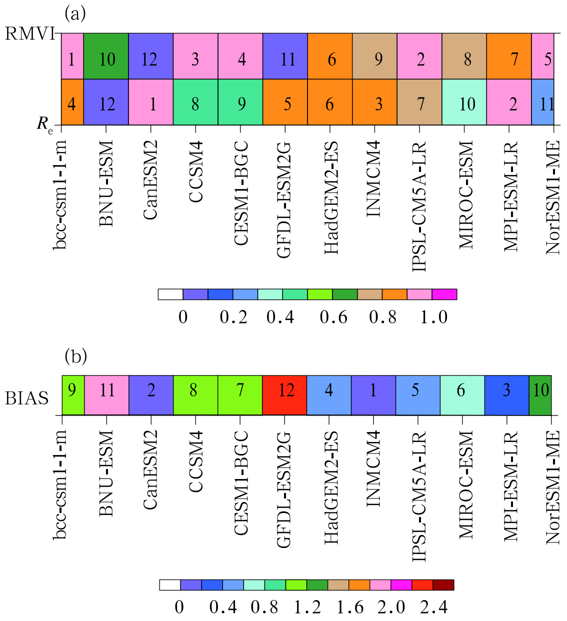

Fig. 8. Model ranking with respect to LAI: (a) model ranking results based on relative error (R_e) calculated by Eqs. (4) and (5) and relative model variability index (RMVI) calculated by Eqs. (6) and (7); (b) model ranking result based on absolute bias (BIAS) between simulations and the observation calculated by Eq. (8).

ES, and CCSM4 perform the best among the 12 models. No model ranking was performed for vegetation cover due to the deficiency of monthly vegetation cover. It is important to note that the model ranking results are dependent on the selection of variables and skill score metrics applied, as well as the study region. Therefore, this result is limited to the conditions of the present analysis only.

5. Conclusions and discussion

In this study, the abilities of 12 CMIP5 ESMs to reproduce the mean state, trends, seasonal cycle, and interannual variability of vegetation dynamics for the present day over the TP have been evaluated by comparing against the remotely sensed biological vegetation products. Several metrics, including three ranking metrics, were applied to identify the strengths and weaknesses of the individual models, as well as their systematic biases.

The LAI patterns generated by the models agree well with the observed data, though most of the models tend to overestimate the satellite LAI magnitude, with r^2 and RMSE values of 0.69 and 1.25 respectively for the model ensemble. The simulated NPP is generally overestimated when compared with the IGBP NPP (except for CanESM2 and NorESM1-ME) and MODIS NPP during different periods. The wet bias found in most models and overestimation of photosynthesis as well as the bias of satellite data, are considered to be plausible reasons for the overestimation of simulated LAI in most of the models. The model simulations capture the observed increasing trend of LAI over most of the TP during the period 1986–2005; however, the decreasing trend around the headstream of the Yellow River is not detected due to the coarse resolution of the ESMs. The model ensemble produces overestimated bare ground and underestimated tree fraction, with r^2 values of 0.56 and 0.59 for bare soil and tree fraction respectively. Grass coverage shows the poorest performance. During 1950–2005, bare ground over the TP shows a slight increasing trend of 0.58% per decade, while forest is decreasing at 1.48% per decade. Grass coverage does not show any significant variation. The models show very different skill scores in their simulations of the seasonal evolution and interannual variability. By synthetically considering the model performance in terms of the mean state, seasonal and interannual variability compared with observed LAI, INMCM4, bcc-csm1-1m, MPI-ESM-LR, IPSL-ESM-LR, HadGEM2-ES, and CCSM4 are ranked the best models in representing the vegetation characteristics of the TP.

In our study, LAI and vegetation cover have been evaluated by using two metrics (r^2 and RMSE) to show the general performance of model simulations, and we carried out vegetation cover validation separately on three classes. There are a number of other metrics used for vegetation biophysical variables. For example, Monserud and Leemans (1992) used j statistics to evaluate discrete vegetation, Poulter et al. (2011) attempted to evaluate vegetation cover simultaneously in more than two classes based on the b-diversity metric (mean Euclidean distance). The chosen metrics depend mainly on the research objective. Similarly, the relatively simple skill score metrics of Re, RMVI, and BIAS were adopted for the model ranking in this study, but there are more complicated and mature metrics (Brunke et al., 2003; Decker et al., 2012; Wang and Zeng, 2012) that could also be used for ranking the models.

The CCSM4, CESM1-BGC, and NorESM1-ME, which share CLM4/CLM4CN as their land surface model and vegetation model, show some common weaknesses and strengths in their simulations, such as good performance in representing the observed spatial distribution, seasonal cycle, and interannual variability, and bad performance in reproducing the mean values of observed LAI and NPP. This suggests the importance of land surface and vegetation physics for the successful description of vegetation dynamics. It was also noted in our analysis that the simulated PFT fractions show much poorer performance compared with LAI, for both the mean state and the spatial distribution, which highlights a major weakness for model developers to work on for future model improvements.

Acknowledgments. We thank the Beijing Normal University for providing the Global Products of

Essential Land Variables (GLASS). Dr. Zhao Maosheng, Martin Jung, and Victoria Brovkin are thanked for providing the MODIS NPP, GPP, model tree ensembles (MTE) GPP and MODIS/VCF vegetation cover datasets. We thank Dr. Gao Xuejie for providing the 2400 meteorological station data of China. The two anonymous reviewers are also thanked for their insightful comments and suggestions, which helped to improve the quality of the manuscript. We acknowledge the computing resources and time of the Supercomputing Center of Cold and Arid Regions Environment and Engineering Research Institute of the Chinese Academy of Sciences. We also acknowledge the modeling groups from BCC, BNU, CCCma, GFDL, INM, IPSL, MIROC, MOHC, MPI-M, NCAR, and NCC, and the World Climate Research Programme's (WCRP) Working Group on Coupled Modeling (WGCM) for making the model output available on the U.S. Department of Energy PCMDI server.

REFERENCES

Abramowitz, G., R. Leuning, M. Clark, et al., 2008: Evaluating the performance of land surface models. *J. Climate*, **21**, 5468–5481.

Anav, A., G. Murray-Tortarolo, P. Friedlingstein, et al., 2013a: Evaluation of land surface models in reproducing satellite derived leaf area index over the high-latitude Northern Hemisphere. Part II: Earth system models. *Remote Sens.*, **5**, 3637–3661.

——, P. Friedlingstein, M. Kidston, et al., 2013b: Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models. *J. Climate*, **26**, 6801–6843.

Arneth, A., S. P. Harrison, S. Zaehle, et al., 2010: Terrestrial biogeochemical feedbacks in the climate system. *Nature Geosci.*, **3**, 525–532.

Bathiany, S., M. Claussen, V. Brovkin, et al., 2010: Combined biogeophysical and biogeochemical effects of large-scale forest cover changes in the MPI earth system model. *Biogeosci. Discuss.*, **7**, 1383–1399.

Blyth, E., D. Clark, R. Ellis, et al., 2011: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale. *Geosci. Model Dev.*, **4**, 255–269.

Brovkin, V., L. Boysen, T. Raddatz, et al., 2013: Evaluation of vegetation cover and land-surface albedo in MPI-ESM CMIP5 simulations. *J. Adv. Modeling Earth Syst.*, **5**, 48–57.

Brunke, M. A., C. W. Fairall, X. B. Zeng, et al., 2003: Which bulk aerodynamic algorithms are least problematic in computing ocean surface turbulent fluxes? *J. Climate*, **16**, 619–635.

Cadule, P., P. Friedlingstein, L. Bopp, et al., 2010: Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements. *Global Biogeochem. Cy.*, **24**, doi: 10.1029/2009GB003556.

Collins, W., N. Bellouin, M. Doutriaux-Boucher, et al., 2011: Development and evaluation of an earth-system model—HadGEM2. *Geosci. Model Dev. Discuss.*, **4**, 997–1062.

Cramer, W., 2011: ISLSCP II IGBP NPP output from terrestrial biogeochemistry models. *ISLSCP Initiative II Collection. Data Set.* Hall, F. G., G. Collatz, B. Meeson, et al., Eds., Oak Ridge National Laboratory Distributed Active Center, Oak Ridge, Tennessee, U. S. A., doi: 10.3334/ORNLDAAC/1027.

——, D. W. Kicklighter, A. Bondeau, et al., 1999: Comparing global models of terrestrial net primary productivity (NPP): Overview and key results. *Global Change Biology*, **5**(S1), doi: 10.1046/j.1365-2486.1999.00009.x.

Dan, L., J. J. Ji, and Y. He, 2007: Use of ISLSCP II data to intercompare and validate the terrestrial net primary production in a land surface model coupled to a general circulation model. *J. Geophys. Res. Atmos. (1984–2012)*, **112**, doi: 10.1029/2006JD007721.

Decker, M., M. A. Brunke, Z. Wang, et al., 2012: Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations. *J. Climate*, **25**, 1916–1944.

Fisher, J. B., K. P. Tu, and D. D. Baldocchi, 2008: Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sens. Environ.*, **112**, 901–919.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res. Atmos. (1984–2012)*, **113**, doi: 10.1029/2007JD008972.

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, et al., 2003: Global percent tree cover at a spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm. *Earth Interact.*, **7**, 1–15.

IPCC, 2007: Summary for Policymakers. *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Solomon, S., et al., Eds., Cambridge University Press, 996 pp.

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. *Biogeosci.*, **6**, 2001–2013.

—–, —–, H. A. Margolis, et al., 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. Biogeosci. (2005–2012)*, **116**, doi: 10.1029/2010JG001566.

Kang, H.-S., Y. K. Xue, and G. J. Collatz, 2007: Impact assessment of satellite-derived leaf area index datasets using a general circulation model. *J. Climate*, **20**, 993–1015.

Kato, T., Y. H. Tang, S. Gu, et al., 2004: Carbon dioxide exchange between the atmosphere and an alpine meadow ecosystem on the Qinghai-Tibetan Plateau, China. *Agric. Forest Meteor.*, **124**, 121–134.

Lawrence, D. M., K. W. Oleson, M. G. Flanner, et al., 2011: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *J. Adv. Modeling Earth Syst.*, **3**, doi: 10.1029/2011MS00045.

Lawrence, P. J., and T. N. Chase, 2007: Representing a new MODIS consistent land surface in the Community Land Model (CLM 3.0). *J. Geophys. Res. Biogeosci. (2005–2012).* **112**, doi: 10.1029/2006JG000168.

Liang, S. L., X. Zhao, S. H. Liu, et al., 2013: A long-term global land surface satellite (GLASS) dataset for environmental studies. *Int. J. Digital Earth*, **6**(Sup1), 5–33.

Monserud, R. A., and R. Leemans, 1992: Comparing global vegetation maps with the Kappa statistic. *Ecological Modeling*, **62**, 275–293.

Neale, R. B., C. Chen, A. Gettelman, et al., 2010: Description of the NCAR Community Atmosphere Model (CAM 5.0). NCAR Tech. Note NCAR/TN-486+STR. 274 pp.

Piao, S. L., J. Y. Fang, and J. S. He, 2006: Variations in vegetation net primary production in the Qinghai-Xizang Plateau, China, from 1982 to 1999. *Climatic Change*, **74**, 253–267.

—–, —–, P. Ciais, et al., 2009: The carbon balance of terrestrial ecosystems in China. *Nature*, **458**, 1009–1013.

—–, P. Ciais, Y. Huang, et al., 2010: The impacts of climate change on water resources and agriculture in China. *Nature*, **467**, 43–51.

Port, U., V. Brovkin, and M. Claussen, 2012: The role of dynamic vegetation cover in future climate change. *Earth Syst. Dyn. Discuss.*, **3**, 485–522.

Poulter, B., P. Ciais, E. Hodson, et al., 2011: Plant functional type mapping for earth system models. *Geosci. Model Dev. Discuss.*, **4**, 2081–2121.

Randerson, J. T., F. M. Hoffman, P. E. Thornton, et al., 2009: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Global Change Biology*, **15**, 2462–2484.

Scherrer, S. C., 2011: Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming. *Int. J. Climatol.*, **31**, 1518–1529.

Shao, P., X. P. Zeng, K. Sakaguchi, et al., 2013: Terrestrial carbon cycle: Climate relations in eight CMIP5 earth system models. *J. Climate*, **26**, 8744–8764.

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

Tan, K., P. Ciais, S. Piao, et al., 2010: Application of the ORCHIDEE global vegetation model to evaluate biomass and soil carbon stocks of Qinghai-Tibetan grasslands. *Global Biogeochem. Cy.*, **24**, doi: 10.1029/2009GB003530.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos. (1984–2012)*, **106**, 7183–7192.

Thornton, P. E., and N. E. Zimmermann, 2007: An improved canopy integration scheme for a land surface model with prognostic canopy structure. *J. Climate*, **20**, 3902–3923.

Wang, A. H., and X. B. Zeng, 2012: Evaluation of multireanalysis products with in situ observations over the Tibetan Plateau. *J. Geophys. Res. Atmos. (1984–2012)*, **117**, doi: 10.1029/2011JD016553.

Watanabe, S., T. Hajima, K. Sudo, et al., 2011: MIROC-ESM: Model description and basic results of CMIP5-20c3m experiments. *Geosci. Model Dev. Discuss.*, **4**, 1063–1128.

Watson, D. J., 1947: Comparative physiological studies on the growth of field crops. I: Variation in net assimilation rate and leaf area between species and varieties, and within and between years. *Ann. Bot.*, **11**, 41–76.

Wittig, V. E., E. A. Ainsworth, S. L. Naidu, et al., 2009: Quantifying the impact of current and future tropospheric ozone on tree biomass, growth, physiology and biochemistry: A quantitative meta-analysis. *Global Change Biology*, **15**, 396–424.

Wu Jia and Gao Xuejie, 2013: A gridded daily observation dataset over China region and comparison with the other datasets. *Chinese Geophys.*, **56**, 1102–1111. (in Chinese)

Xiao, Z. Q., S. L. Liang, J. D. Wang, et al., 2014: Use of general regression neural networks for generating the GLASS leaf area index product from time-series MODIS surface reflectance. *Geosci. Remote Sens.*, **52**, 209–223, doi: 10.1109/TGRS.2013.2237780.

Yu, H., E. Luedeling, J. Xu, 2010: Winter and spring warming result in delayed spring phenology on the Tibetan Plateau. *Proc. Natl. Acad. Sci. USA*, **107**, 22151–22156.

Zhang, P., M. Hirota, H. Shen, et al., 2009: Characterization of CO_2 flux in three Kobresia meadows differing in dominant species. *J. Plant Ecol.*, **2**, 187–196.

Zhao, M. S., F. A. Heinsch, R. R. Nemani, et al., 2005: Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sens. Environ.*, **95**, 164–176.

——, S. W. Running, and R. R. Nemani, 2006: Sensitivity of Moderate Resolution Imaging Spectroradiometer (MODIS) terrestrial primary production to the accuracy of meteorological reanalyses. *J. Geophys. Res. Biogeosci. (2005–2012)*, **111**, doi: 10.1029/2004JG000004.

——, and S. W. Running, 2010: Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science*, **329**, 940–943.