Supporting Information: Description of Scoring Metrics and Data Processing

Part One Scoring Metrics

A. Global Bias Metric

For different variables, we use 2 different methods to calculate their global bias scores. For above ground biomass (biomass), burned area (burntarea), gross primary production (gpp), lead area index (lai), latent heat (le), net ecosystem exchange (nee), precipitation (pr), ecosystem respiration (reco), sensible heat (sh) and soil carbon (soilc), we use Method 1. For other variables, we use Method 2.

Method 1:

$$M_i = 1 - \left| \frac{AM_{\mathrm{mod},i} - AM_{obs,i}}{AM_{obs,i}} \right| \quad \text{(A1)}$$

$$M_i^{'} = e^{M_i} / e \quad \text{(A2)}$$

$$M = \frac{\sum_{i=1}^{ncells} M_i^{'} \times A_i \times AM_{obs,i}}{\sum_{i=1}^{ncells} A_i \times AM_{obs,i}} \quad \text{(A3)}$$

We use Eqs. A1-2 and Eq. 3 to calculate global bias metric score M_i at grid cell i and its global mean M, respectively. $AM_{obs,i}$ and $AM_{mod,i}$ are annual mean of the observation and the model at grid cell i, separately. A_i is the are for grid cell or site i. If the observation is site data, we set A_i equal to 1.

Method 2:

$$AM_{obs} = \frac{\sum_{i=1}^{ncells} AM_{obs,i} \times A_i}{TotalArea} \quad \text{(A1)}$$

$$M_i = 1 - \left| \frac{AM_{\mathrm{mod},i} - AM_{obs,i}}{AM_{obs}} \right| \quad \text{(A2)}$$

$$M_i^{'} = e^{M_i} / e \quad \text{(A3)}$$

$$M = \frac{\sum_{i=1}^{ncells} M_i^{'} \times A_i}{TotalArea} \quad \text{(A4)}$$

We use Eqs. A1-3 and Eq. 4 to calculate global bias metric score M_i at grid cell i and its global mean M, respectively. $AM_{obs,i}$ and $AM_{mod,i}$ are annual mean of the observation and the model at grid cell i, separately. AM_{obs} is the global mean of observation annual mean over land where data are available. A_i is the area for grid cell or site i. *TotalArea* is sum of the area A_i for all land grid cells or sites (*ncells*) where observation data is available. If the observation is site data, we set A_i equal to 1.

B. Root Mean Square Error Metric

$$\Phi_{obs} = \frac{\sum_{i=1}^{ncells} \Phi_{obs,i} \times A_i}{TotalArea} \quad \text{(B1)}$$

$$M_i = 1 - \frac{RMSE_i}{\Phi_{obs}} \quad \text{(B2)}$$

$$M_i^{'} = e^{M_i} / e \quad \text{(B3)}$$

$$M = \frac{\sum_{i=1}^{ncells} M_i^{'} \times A_i}{TotalArea} \quad \text{(B4)}$$

We use Eqs. B1-3 to calculate root mean square error metric score M_i at grid cell or site i and its global mean M, respectively. Where $\Phi_{obs,i}$ is the root mean square for monthly mean annual cycle of the observation at grid cell i (for grid data) or site i (for site observation), and $RMSE_i$ is the root mean square error between model and observation. Φ_{obs} is the global mean of observation root mean square over land where data are available. A_i is the area for grid cell or site i. *TotalArea* is sum of the area A_i for all land grid cells or sites (*ncells*) where observation data is available. If the observation is site data, we set A_i equal to 1 (*Ref: David Lawrence's personal Communication*). This metric is used to compare magnitude and phase difference of the monthly mean annual cycle between the model and the observation.

C. Spatial Distribution Metric

$$M = \frac{4(1+R)}{(\sigma_f + 1/\sigma_f)^2 (1+R_0)} \quad \text{(C)}$$

We use Eq. C to calculate spatial distribution metric score M. R is the spatial correlation coefficient of the annual mean between model and observation. R_o is their ideal maximum correlation. Here, we set R_o equal to 1 for all models. σ_f is ratio for standard

deviation of model to that of observation (*Ref: Taylor, J. Geophys. Res., 106, 2001*). This metric is used to compare magnitude and spatial pattern of annual mean of model with observation.

D. Seasonal Cycle Phase Metric

$$M_i = (1 + \cos \vartheta_i) / 2 \tag{D1}$$

$$M = \frac{\sum_{i=1}^{ncells} M_i \times A_i}{TotalArea} \tag{D2}$$

We use Eqs. D1 and D2 to calculate seasonal cycle phase metric score M_i at grid cell or site i and its global mean M, respectively. ϑ_i is the difference of the angle between the month of the maximum value for the model and that for the observation at grid cell i (for the grid data) or site i (for the site data). A_i is the area for grid cell or site i. *TotalArea* is sum of the area A_i for all land grid cells or sites (*ncells*) where observation data is available. If the observation is site data, we set A_i equal to 1 (*Ref: Prentice, et al., GBC, 25, 2011*). This metric is used to compare phase difference of the monthly mean annual cycle between the model and the observation.

E. Interannual Variability Metric

$$\sigma_{obs} = \frac{\sum_{i=1}^{ncells} \sigma_{obs,i} \times A_i}{TotalArea} \tag{E1}$$

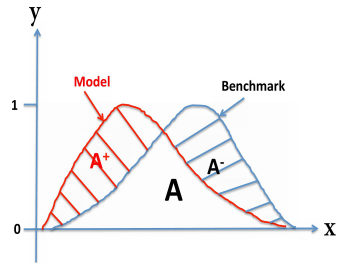
$$M_i = 1 - \left| \frac{\sigma_{\mathrm{mod},i} - \sigma_{obs,i}}{\sigma_{obs}} \right| \tag{E2}$$

$$M_i^{'} = e^{M_i} / e \tag{E3}$$

$$M = \frac{\sum_{i=1}^{ncells} M_i^{'} \times A_i}{TotalArea} \tag{E4}$$

We use Eqs. E1-4 to calculate interannual variability metric score M_i at grid cell or site i and its global mean M, respectively. Where $\sigma_{obs,i}$ and $\sigma_{mod,i}$ is standard deviation at grid cell i (for grid data) or site i (for site data) for observation and model simulations. σ_{obs} is the global mean of observation standard deviation over land where data are available. A_i is the area for grid cell or site i. *TotalArea* is sum of the area A_i for all land grid cells or sites (*ncells*) where observation data is available. If the observation is site data, we set A_i equal to 1 (*Ref: Randerson, et al., GCB, 15, 2009*).

F. Variable to Variable Relationship Metric



$$M = 1 - (A^{+} + A^{-})/A$$

A: Total Area of Benchmark

The curves in the above figure show relationship for annual mean of one variable (y) is a function of another variable (x). Black and red curves are for observation and model, respectively. The red area (A^{+}) shows the fractions for the part of model over observation, and the blue area (A^{-}) shows the part of observation over model. The sum of the fractions (A^{+}+A^{-}) shows the total area for inconsistence of model with observation. This metric measures the similarity of 2 variables relationship between model and benchmark.

G. Overall Score Metric

We calculate a couple of sets of overall scores in this diagnostic package, one for individual variable (G1), one for all variables mean (G2), one for all variable to variable relationships mean (G3), and the last one for the overall score combined both G2 and G3 (G4).

G1. Overall score for individual variable

To obtain an overall score for individual variable, we combine component score metrics (A to E) with unequally weighting functions. The weighting is the function of the source and component metrics. For example, considering importance of root mean square error, we give double weighting from this metric compared to all others. For the above ground live biomass, we also give more weighting for the source from the Pan Tropical Forest than the other 2 sources (Contiguous US and Contiguous US +Alaska) because of the more data coverage from the Pan Tropical Forest. Here is the approach:

$$M = \frac{\sum_{i=1,j=1}^{nmet,nsur} M_{i,j} \times A_{i,j}}{SUM(A)}, \quad SUM(A) = \sum_{i=1,j=1}^{nmet,nsur} A_{i,j} \qquad \text{(G1)}$$

Where $M_{i,j}$ is the component metric (A to E), and $A_{i,j}$ is the contribution for each component metric and source. *nmet* and *nsur* are total numbers of the component metrics and sources, respectively.

G2. Overall score for all variables mean

To obtain overall score for all variables, we combine overall score for each variable from G1 with unequally weighting functions. The weighting is the function of categories (carbon cycle, hydrology cycle, energy cycle, forcing or others). Considering importance of the global carbon cycle, we give double weighting from variables belong to this category compared with other categories. Here is the equation we use for this metric calculation:

$$M = \frac{\sum_{i=1}^{n\text{var}} M_i \times A_i}{SUM(A)},\ SUM(A) = \sum_{i=1}^{n\text{var}} A_i \qquad \text{(G2)}$$

Where M_i is the overall score for individual variable from G1, and A_i is the contribution from each variable which is the function of categories. *nvar* is total numbers of the variables.

G3. Overall score for all variable to variable relationships mean

We calculate variable to variable relationship metric score for each pair of variables using the approach (F), then we calculate the mean score for all pairs of variables by simply straight averaging.

G4. Overall score for each model

This score is simply averaging of the overall scores for both all variables mean (G2) and all variable to variable relationships mean (G3). This is the final score for each model.

References

1. Prentice I.C. et al., Modeling fire and the terrestrial carbon balance, *Global Biogeochemical Cycles*, 25, doi.10.1029/2010GB003906, 2011.
2. Randerson, J.T., et al., Global burned area and biomass burning emissions from small fires. *J. Geophys. Res*., 106, DOI: 10.1029/2012JG002128, 2012.
3. Randerson, J.T., et al., Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. Global Change Biology 15: 2462–2484. doi: 10.1111/j.1365-2486.2009.01912.x, 2009.
4. Taylor, K. E., Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res*., 106, D7, 7183-7192, 2001.

Part Two Data Sources and Processing

1. Burned area

Data Source: GFED3.1 monthly burned area from:

Giglio, L., J.T. Randerson, G.R. van der Werf, P.S. Kasibhatla, G.J. Collatz, D.C. Morton, and R.S. DeFries. 2010. Assessing variability and long-term trends in burned area by merging multiple satellite fire products. Biogeosciences. 6: 1171–1186.

First Downloaded from the following server on 04/20/2010, and updated on 06/22/2014

- ftp fuoco.geog.umd.edu
- login: fire
- pword: burnt
- cd gfed/monthly

Data processing: The GFED3.1 monthly burned area observations were available from 1997-2011. The original observations had units of hectares. These were converted to fractional burned area using the land area of each 0.5° grid cell.

Model processing: We extracted the variable with the standard name burntArea from the models which had units of fractional burned area in each grid cell, in %. We directly compared this variable to the transformed GFED3.1 observations described above.

Recommended Metrics for Overall Score:

Global bias (A), root mean square error (B), spatial distribution (C), seasonal cycle phase (D) and interannual variability (E) score metrics were recommended to use for computing overall score for this variable.

2. Biomass

Data Source:

I. Contiguous US above ground live biomass (US.FOREST)

Blackard et al., 2008. Mapping U.S. Forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sens. Environ.*, 112, 1658-1677.

Downloaded from the following website on 04//09/2012.

- http://fsgeodata.fs.fed.us/rastergateway/biomass/

II. Contiguous US + Alaska above ground live biomass (NBCD2000)

Kellndorfer, J., Walker, W., LaPoint, E., Cormier, T., Bishop, J., Fiske, G., & Kirsch, K.. Vegetation height, biomass, and carbon stock for the conterminous United States: A high-resolution dataset from Landsat ETM+, SRTM-InSAR, National Land Cover Database, and Forest Inventory and Analysis data fusion, in review.

Original downloaded from the following website on 04//09/2012, and updated on 06/21/2013

- http://www.whrc.org/mapping/nbcd/index.html

III. Pan Tropical Forest biomass (GLOBAL.CARBON)

Saatch, Sassan S., et al., 2011. Benchmark map of forest carbon stocks in tropical regions across three continents, *Proc. Natl. Acad. Sci.,* 108 (24), 9899-9904.

Downloaded from the following website on 10//17/2012.

- ftp://www-radar.jpl.nasa.gov/projects/carbon/datasets/

Data processing: Both US Forest Biomass and NBCD2000 datasets have a 250-m horizontal resolution, and Global Tropical Forest Biomass dataset has 1km resolution. All three datasets have an original unit of tons per hectare at each grid cell. US Forest Biomass dataset also has observations in Alaska and Puerto Rico, but we only used data in the 48 lower US states and Alaska to compare with models. All three datasets were regridded to 0.5° grid cell.

Model processing: We extracted the variable with the standard name cVeg (carbon mass in vegetation) from the models and convert from the original unit of Kg per m2 to ton per hectare as the observations. We also transformed the model data to 0.5° resolution at grid cells where observations were available. We compared means of the models in the year of 2000-2005 with US Forest, NBCD2000 and global tropical forest Biomass observations, respectively.

Recommended Metrics for Overall Score:

Global bias (A) and spatial distribution (C) score metrics were recommended to use for computing overall score for this variable.

3. Soil Carbon

Data Source:

I. Harmonized World Soil Database (HWSD) v1.2

Todd-Brown, T.E.O, J.T. Randerson, W. M. Post, F.M. Hoffman, C. Tarnocai, E.A.G. Schuur, and S.D. Allison, Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. Biogeosciences, 10, 1717-1736, 2013.

Downloaded from the following website on 03//05/2014.

- Personal communication

II. Top 3m Northern Circumpolar Soil Carbon Database v2 (NCSCDv2)

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G., and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Glob. Biogeochem. Cy., 23, GB2023, doi:10.1029/2008GB003327, 2009.

Downloaded from the following website on 03//04/2014

- http://bilin.su.se/data/ncscd

Data processing: Both HWSD and NCSCDv2 datasets have 0.5°×0.5° horizontal resolution. Both units were also converted to KgC/m^2.

Model processing: We extracted the variable with the standard name cSoil (carbon mass in soil) from the models. We also transformed the model data to 0.5° resolution at grid cells where observations were available. We compared means of the models in the year of 1996-2005 with HWSD and NCSCDv2 benchmarks, respectively.

Recommended Metrics for Overall Score:

Global bias (A) and spatial distribution (C) score metrics were recommended to use for overall score calculation for this variable.

4. Global GPP

Data Source:

I. AmeriFlux L4 site observattions (AMERICFLUX)

Gower et al., 1999. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote Sens. Environ.,* 70:29-51.

Downloaded from the following website on 11/22/2011.

- http://public.ornl.gov/ameriflux/site-select.cfm

II. FluxNet L4 site observations (FLUXNET)

Lasslop G, Reichstein M, Papale D et al., Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Global Change Biology, 16, 187â208, 2010.

Downloaded from the following website on 12/11/2012.

- https://www.bgc-jena.mpg.de/bgc-mdi/index.php/Services/Services

Data processing: We extracted variable GPP_f directly from monthly AmeriFlux and FluxNet L4 observations, and select the sites with valid datasets at least for 24 months from January 1996 through December 2005.

Model processing: We extracted model variables with the CMIP5 standard names gpp. Then we sampled models data at AmeriFlux and FluxNet sites and times when observations were available. Meanwhile we convert their units to those corresponding to observations.

III. Fluxnet Multi-Tree-Ensemble (FLUXNET-MTE)

Jung, Martin et al., Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature, 467, 951-954, doi:10.1038/nature09396, 2010.

Downloaded from the following website on 12/2/2013.

- *personal exchange from NCAR group*

Data processing: We extracted global variables GPP and LE directly from monthly Fluxnet-MTE L4 dataset in the period of 1982-2005.

Model processing: We extracted model variables with the CMIP5 standard names gpp. Meanwhile we convert their units to those corresponding to observations.

Recommended Metrics for Overall Score:

Global bias (A), root mean square error (B), spatial distribution (C), seasonal cycle phase (D) and interannual variability (E) score metrics were recommended to use for computing overall score for this variable.

5. Latent heat (LE)

Data Source:

I. AmeriFlux L4 site observattions (AMERICFLUX)

Gower et al., 1999. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote Sens. Environ.,* 70:29-51.

Downloaded from the following website on 11/22/2011.

- http://public.ornl.gov/ameriflux/site-select.cfm

II. FluxNet L4 site observations (FLUXNET)

Lasslop G, Reichstein M, Papale D et al., Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Global Change Biology, 16, 187â208, 2010.

Downloaded from the following website on 12/11/2012.

- https://www.bgc-jena.mpg.de/bgc-mdi/index.php/Services/Services

Data processing: We extracted variables LE_f directly from monthly AmeriFlux and FluxNet L4 observations, and select the sites with valid datasets at least for 24 months from January 1996 through December 2005.

Model processing: We extracted model variable with the CMIP5 standard names hfls (latent heat flux). Then we sampled models data at AmeriFlux and FluxNet sites and times when observations were available. Meanwhile we convert their units to those corresponding to observations.

III. Fluxnet Multi-Tree-Ensemble (FLUXNET-MTE)

Jung, Martin et al., Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature, 467, 951-954, doi:10.1038/nature09396, 2010.

Downloaded from the following website on 12/2/2013.

- *personal exchange from NCAR group*

Data processing: We extracted global variables GPP and LE directly from monthly Fluxnet-MTE L4 dataset in the period of 1982-2005.

Model processing: We extracted model variable with the CMIP5 standard names hfls (latent heat flux). Meanwhile we convert their units to those corresponding to observations.

Recommended Metrics for Overall Score:

Global bias (A), Root mean square error (B), spatial distribution (C), seasonal cycle phase (D), and interannual variability (E) score metrics were recommended to use for overall score calculation for this variable.

6. NEE, ecosystem respiration (reco) and sensible heat (SH)

Data Source:

I. AmeriFlux L4 site observattions

Gower et al., 1999. Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote Sens. Environ.,* 70:29-51.

Downloaded from the following website on 11/22/2011.

- http://public.ornl.gov/ameriflux/site-select.cfm

II. FluxNet L4 site observations

Lasslop G, Reichstein M, Papale D et al., Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Global Change Biology, 16, 187â208, 2010.

Downloaded from the following website on 12/11/2012.

- https://www.bgc-jena.mpg.de/bgc-mdi/index.php/Services/Services

Data processing: We extracted variables H_f, Reco_or, GPP_f and NEE_f directly from monthly AmeriFlux and FluxNet L4 observations, and select the sites with valid datasets at least for 24 months from January 1996 through December 2005.

Model processing: We extracted model variables with the CMIP5 standard names hfss (sensible heat flux), gpp, ra and rh. We added ra and rh to obtain reco (total ecosystem respiration). We calculated nee by using gpp minus reco. Then we sampled models data at AmeriFlux and FluxNet sites and times when observations were available. Meanwhile we convert their units to those corresponding to observations.

Recommended Metrics for Overall Score:

Global bias (A), Root mean square error (B), spatial distribution (C), seasonal cycle phase (D), and interannual variability (E) score metrics were recommended to use for overall score calculation for this variable.

7. Albedo

Data Source:

I. CERES (The Clouds and the Earth Radiant Energy System)

Young, D. F., P. Minnis, D. R. Doelling, G. G. Gibson, T. Wong, 1998. Temporal Interpolation Methods for the Clouds and the Earth Radiant Energy System (CERES) Experiment. Journal of Applied Meteorology, Vol 37(6), 572-590.

Downloaded from the following server on 10/22/2011.

- http://ceres.larc.nasa.gov/products.php?product=SRBAVG

II. MODIS MCD43C3 16-day 0.05 degree CMG L3, version 5

Schaaf, C. B., W. Lucht, T. Tsang, F. Gao, N. Strugnell, L. Chen, Y. Liu, and A.H. Strahler, 1999. Prototyping the MODerate Resolution Imaging Spectroradiometer (MODIS) BRDF and Albedo Product, Proc. Int. Geosci. Remote Sens. Symp. (IGARSS'99), Hamburg, Germany, 28 June - 2 July, 1506-1508.

Downloaded from the following server on 10/22/2011.

- https://lpdaac.usgs.gov/products/modis_products_table/brdf_albedo_model_parameters/16_day_l3_0_05deg_cmg/mcd43c1

Data processing: Both CERES and MODIS observations were available from 2000-2005. The original CERES observations are only available for radiation, thus we calculated albedo by using all sky net surface shortwave and all sky surface downward shortwave. We regridded from the original 1° resolution to the final 0.5° resolution. We also regridded MODIS albedo from the original 0.05° resolution to the final 0.5° grid cell.

Model processing: We extracted the variable with the standard name rsds (surface downwelling shortwave) and rsus (surface upwelling shortwave) from the models and used them to calculate albedo. We compared models albedo to the CERES and MODIS observations described above.

Recommended Metrics for Overall Score:

Global bias (A), Root mean square error (B), spatial distribution (C), seasonal cycle phase (D), and interannual variability (E) score metrics were recommended to use for overall score calculation for this variable.

8. Precipitation

Data Source: GPCP Version 2.2 Combined Precipitation Data Set

Adler, R.F., G. Gu, G.J. Huffman, Estimating Climatological Bias Errors for the Global Precipitation Climatology Project (GPCP). J. Appl. Meteor. and Climatol., 51(1), doi:10.1175/JAMC-D-11-052.1, 84-99, 2012.

Downloaded from the following website on 10/31/2011.

- *ftp://precip.gsfc.nasa.gov/pub/gpcp-v2.2/psg*

Data processing: We extracted variable pr directly from monthly GPCP v2.2 dataset from 1979 till 2005, and regrided the data from 2.5×2.5 to 0.5×0.5 resolution.

Model processing: We extracted model variable with the CMIP5 standard name pr (precipitation). Meanwhile we convert its unit from $Kg/m^2/s$ to mm/day in the benchmark.

Recommended Metrics for Overall Score:

Global bias (A), Root mean square error (B), spatial distribution (C), seasonal cycle phase (D), and interannual variability (E) score metrics were recommended to use for overall score calculation for this variable.

9. Surface air temperature

Data Source: High resolution CRU mean temperature

Harris, I., Jones, P.D., Osborn, T.J., and Lister, D.H., Updated high-resolution grids of monthly climatic observations. Int. J. Climatol., Doi: 10.1002/joc.3711, 2013.

Downloaded from the following website on 7/24/2013.

- *http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_1256223773328276*

Data processing: We extracted variable tmp (surface 2m air temperature) from1979 through 2005.

Model processing: We extracted model variables with the CMIP5 standard names tas. Meanwhile we convert its unit from K to C in consistence with the observation.

Recommended Metrics for Overall Score:

Global bias (A), Root mean square error (B), spatial distribution (C), seasonal cycle phase (D), and interannual variability (E) score metrics were recommended to use for overall score calculation for this variable.