Equation Chapter 1 Section 1Machine learning-based modeling of vegetation leaf area index and gross primary productivity across North America and comparison with a process-based model

Zhicheng Zhang [a], Qinchuan Xin [a*], Wanjing Li [a]

[a] Guangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China

[*] *Correspondence to*: Qinchuan Xin (xinqinchuan@mail.sysu.edu.cn)

Telephone Number: (86)1-881-025-3088

Mailing Address:

Sun Yat-Sen University

Earth and Environment Building D101

Guangzhou 510275, China

Abstract

Vegetation plays a key role in regulating the material and energy exchanges among the biosphere, the atmosphere, and the pedosphere. Modeling and predicting vegetation key variables such as leaf area index (LAI) and gross primary productivity (GPP) is crucial to understand and project the processes of vegetation growth in response to climate change. While a number of studies developed models to simulate vegetation GPP using satellite-derived LAI, the requirement of satellite-based model inputs largely limits the predicting power of these developed models. This study developed the machine learning models, including both support vector regression (SVR) and random forests (RF), which are capable of modeling LAI and GPP time series using only meteorological variables. We first simulated the LAI time series directly using meteorological variables as inputs to the machine learning models and then buffered its unrealistic day-to-day fluctuation, and further modeled the GPP time series using meteorological variables and modeled LAI time series. We tested our methods for four main plant functional types across North America and evaluated the models using both satellite-based and flux tower data. The results demonstrate that the machine learning models perform well on simulating the time series of both LAI and GPP. We identified that there is a need to improve the phenology representation in the Biome-BGC model. The machine learning models provide an alternative way to predict time series of LAI and GPP using only meteorological variables across large geographic regions, and also provide benchmarking accuracies for future developments of the process-based models.

Keywords: leaf area index; machine learning; gross primary productivity; terrestrial ecosystem models

1. Introduction

Terrestrial vegetation, through a series of physiological and ecological processes such as radiative transfer, photosynthesis, respiration, and evapotranspiration, exchanges material and energy with the atmosphere, pedosphere, and the other spheres, and affects the interaction between the land surface and the atmosphere. Studying and modeling of the physiological and ecological processes of vegetation in response to climate and environmental changes help understand the interaction and degree of global climate change and terrestrial surface processes. Ecological metrics such as leaf area index (LAI), gross primary productivity (GPP), net primary productivity (NPP) and evapotranspiration (ET) are quantitative indicators for studying the ecosystem processes, and are key variables in terrestrial ecosystem models. The simulation accuracies of these ecological metrics largely mark the robustness and reliability of the terrestrial ecosystem models, and therefore, they have become important and difficult problems in many ecosystem modeling studies.

Vegetation GPP, defined as the amount of organic carbon fixed by green plants through photosynthesis per unit area per unit time, is the energy basis for vegetation to perform the other physiological and ecological activities, and is a key indicator of carbon flux exchange between terrestrial ecosystems and the atmosphere. Thanks to experimental studies at the leaf level, our understanding on leaf photosynthesis is greatly improved and scientific researchers have developed methods, such as the light use efficiency (LUE) models, process-based terrestrial ecosystem models and machine learning models, to simulate GPP under given climate conditions and LAI (defined as green leaf area per unit ground area). The LUE model assumes that light use efficiency that plant leaves absorb and transform solar radiation for photosynthesis is reduced under environmental constraints. The LUE models often have simple forms and are suitable for large-scale modelling with satellite-derived LAI data. The widely used LUE models include CASA [*C B Field et al.*, 1995; *C S Potter et al.*, 1993], MOD17 [*S Running et al.*, 2000], Eddy Covariance-Light Use Efficiency (ECLUE) [*W Yuan et al.*, 2007], and two-leaf light use efficiency (TL-LUE) [*M He et al.*, 2013] model. The

process-based terrestrial ecosystem models depict the physiological and ecological mechanisms associated with vegetation growth and development in details, including radiative transfer, photosynthesis, respiration, evapotranspiration, and soil processes. The process-based terrestrial ecosystem models mark the state-of-the-art scientific understanding and achievement in numerical modeling of the ecosystem processes. Terrestrial ecosystem models provide a number of output variables associated with energy and material fluxes between the biosphere and the atmosphere. The widely used terrestrial ecosystem models and land surface models include Biome-BGC [*M A White, Thornton, P. E. , Running, S. W. , & Nemani, R. R. .* , 2000], SiB2 [*Piers J. Sellers et al.*, 1996; *P. J. Sellers et al.*, 1996], TEM [*Q Zhuang et al.*, 2011], JULES [*M J Best et al.*, 2011; *D B Clark et al.*, 2011], CLM [*K W Oleson et al.*, 2013], and CoLM [*Y Dai et al.*, 2003]. These models commonly adopt the photosynthesis model proposed by [*G D Farquhar et al.*, 1980] and [*G J Collatz et al.*, 1991] to simulate leaf-scale photosynthesis rates and further upscale leaf photosynthesis to the canopy level. The Terrestrial ecosystem models have sound scientific basis but rely on climate forcing variables and model parameterization. Accompanying the development of the computer science, machine learning models, such as artificial neural network (ANN) and support vector machine (SVM), have been applied in modeling ecosystem processes. Machine learning models are data-driven models based on mathematical and statistical principles and establish the non-linear relationships between input and target features by minimizing the loss function through an iterative training process [*M F McCabe et al.*, 2017; *J Verrelst et al.*, 2015]. A number of studies had proven that the performance of machine learning models on simulating vegetation GPP and its time series. For example, [*F Yang et al.*, 2007] trained SVM to predict vegetation GPP using remote sensing variables, such as land surface temperature, enhanced vegetation index (EVI), land cover, and ground-measured climate variables. [*M Schlund et al.*, 2020] proposed a new two-step approach that apply an existing emergent constraint on CO_2 fertilization in combination with a supervised machine learning model to constrain uncertainties in multi-model predictions of GPP. The machine learning models once developed and trained have fast computing capabilities, and can be conveniently applied to studies on the continental or

85 global scale. However, it is nearly impossible to interpret the processes in the model and gain knowledge on
86 the physiological mechanisms of vegetation processes.
87
88 Vegetation LAI, the metric that reflects the amount of vegetation leaves, is the core input or intermediate
89 variable in the above-mentioned models for GPP simulation. In essence, our ability to simulate LAI largely
90 determines the accuracy of the modeled GPP as well as related ecosystem processes. The current models,
91 however, still have limited accuracies on simulating vegetation LAI time series. The LUE models and the
92 machine learning-based models normally adopt observational LAI data obtained from field measurements or
93 remote sensing images, and thus have limited predicting powers when observational LAI data are not
94 available. LAI used in terrestrial ecosystem models or land surface models can be static (i.e., using time series
95 or temporally averaged LAI derived from satellite data on a pixel basis) or dynamic (i.e., predicting the
96 seasonality of LAI based on climate drivers). A number of numeric models, such as Ecosys [*R F Grant et al.*,
97 2009], Biome-BGC [*M A White, Thornton, P. E. , Running, S. W. , & Nemani, R. R. .* , 2000], IBIS [*J A Foley*
98 *et al.*, 1996], and ORCHIDEE [*G Krinner et al.*, 2005], enclose a phenology sub-model to simulate LAI time
99 series via meteorological variables. The land surface models such as CLM simulates the timing of key
100 phenological phases, such as spring onset and autumn offset, using a combination of empirical equations. The
101 dynamic global vegetation models such as LPJ-ML use equations that directly predict LAI time series from
102 meteorological variables. [*A D Richardson et al.*, 2012] highlighted considerable uncertainties associated with
103 vegetation phenology modeling in 14 state-of the-art terrestrial ecosystem models.
104
105 To our best knowledge, nearly few studies to date attempt to develop machine learning models for simulating
106 LAI time series or seasonal cycles using only meteorological variables as model inputs. There are a number of
107 studies that adopt machine learning approaches to retrieve LAI time series or key phenophases from remote
108 sensing data. For example, [*T Wang et al.*, 2017] compared machine learning models that retrieve LAI from

several Moderate-resolution Imaging Spectroradiometer (MODIS) products. [*R Houborg and M McCabe*, 2018] found that combining random forest and the Cubist model is able to derive LAI and key phenophases well from satellite-derived indexes. While satellite-based studies provide valuable time series of LAI data for retrospective analysis, it is pivot to improve our ability on predicting the LAI time series and seasonal cycles via meteorological variables so as to subsequently improve terrestrial ecosystem models and land surface models. Note that there are challenges when using machine learning models to predict LAI time series based on daily or weekly meteorological variables directly, because day-to-day variation of vegetation LAI is much lower than of meteorological variables and because vegetation has lagged responses to environmental conditions.

The main goals of this research are to: 1) develop a machining learning-based scheme to simulate LAI and GPP time series via meteorological variables; 2) compare the simulated results between machine learning models and a process-based model, so as to advance our understanding on the uncertainties of the current terrestrial ecosystem model and discover the potential of improvements. In this research, we implement two machine learning models, including support vector regression (SVR) and random forest (RF), and the process-based model of Biome-BGC to simulate both LAI and GPP time series across North America.

6

2. Study materials and preprocessing

The study area covers the entire continent of North America. Due to large latitude gradients across North America, vegetation types are rich and diverse under varied natural factors, such as solar radiation, temperature, precipitation, and topography (Fig.1). Owing to the availability of the flux tower data, vegetation that we studied includes four natural plant functional types, i.e., deciduous broadleaf forest (DBF), evergreen needleleaf forest (ENF), grassland (GRA), and mixed forest (MIF), according to the International Geosphere-Biosphere Programme (IGBP) classification scheme.

To develop and evaluate the machine learning models and the process-based model, we used large-scale climate data, satellite-based products, and the flux tower dataset. The Daymet dataset [*P E Thornton et al.*, 2016] is a 1-km gridded daily meteorological data product derived from a collection of algorithms designed to interpolate and extrapolate daily meteorological observations. We downloaded the third version Daymet dataset from the Oak Ridge National Laboratory Distributed Active Archive Center website (http://daymet.ornl.gov/). The 1 km Daymet data were re-projected to the sinusoidal projection. As the Daymet dataset does not contain the variable of vapor pressure deficit (VPD), it is calculated as follows:

$$VPD = \frac{1}{2} 0.6108 * \left(e^{\frac{17.269 * Tmax}{237.3 + Tmax}} + e^{\frac{17.269 * Tmin}{237.3 + Tmin}} \right) - VP \quad (1)$$

where VPD and VP denote vapor pressure deficit and vapor pressure, respectively (kPa); and T_{max} and T_{min} denote daily maximum and minimum air temperature, respectively (℃).

7

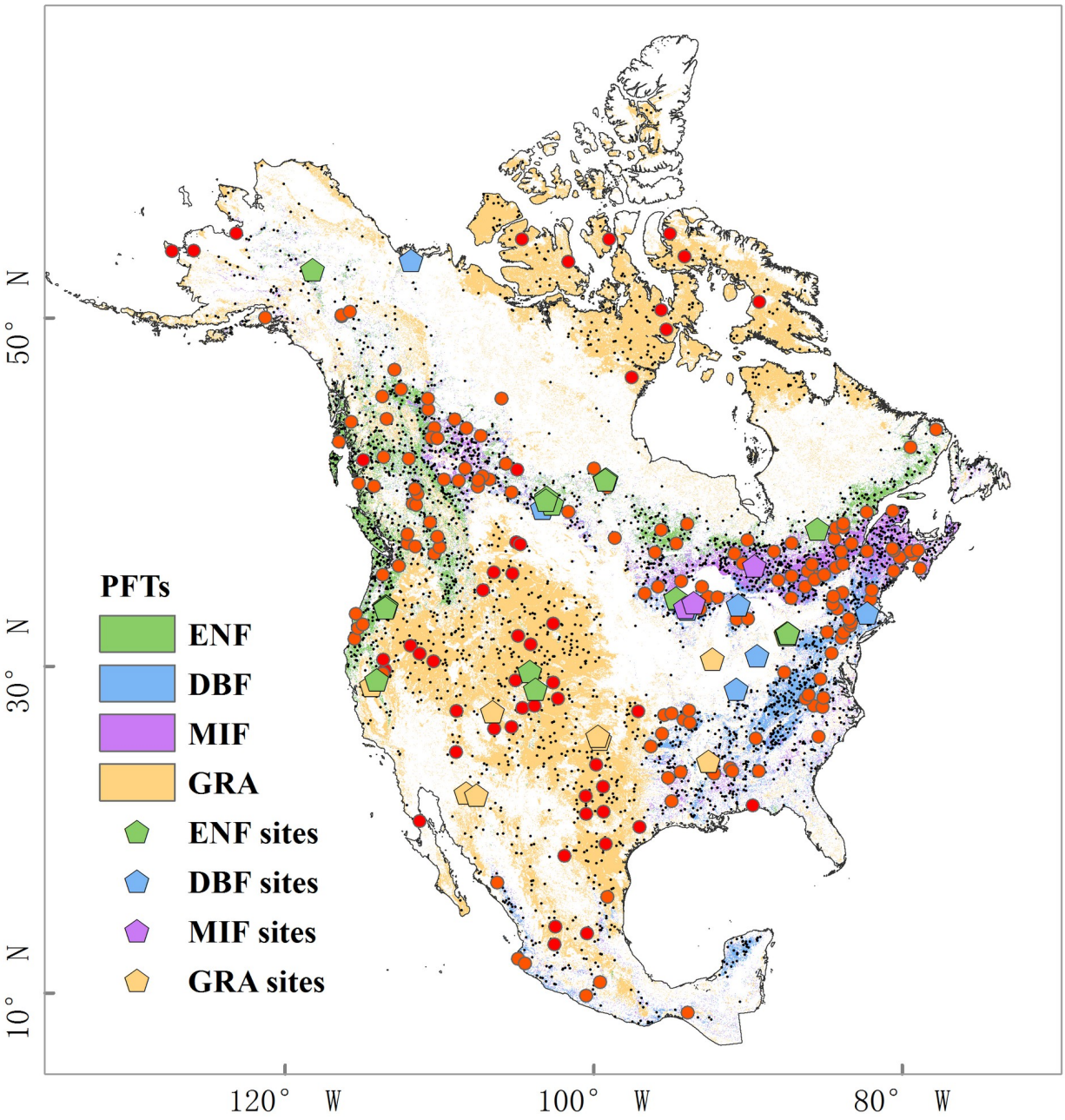


Fig. 1. The spatial distributions of modeling sample points and flux tower sites on a backdrop of the land cover map across North America. Colorful pentagon points denote flux tower sites in different plant functional types, and black and red points denote randomly selected training and validation samples, respectively.

The satellite-based products we used include the 8-day composite MODIS LAI product at 500-m resolution (MOD15A2H), the 8-day composite MODIS GPP product at 500-m resolution (MOD17A2H), and the annual MODIS land cover product at 500-m resolution (MCD12Q1). As remote sensing data often have data quality

8

issues such as cloud cover and aerosol contamination, we used the quality control layer in the MODIS LAI data product to filter pixels with poor qualities and then interpolated the missing values using the three-window moving median filtering method. We screened out outliers in the time series with the Hampel filter and obtained 8-day smoothed LAI time series by applying the Savitzky-Golay fitting method. Because the MODIS GPP data cannot be simply smoothed, we only used the quality control data to filter pixels with poor qualities and then interpolated the missing values using the three-window moving median filtering method. The MCD12Q1 data were used as the land cover mask to derive the spatial distribution of four studied plant functional types, namely DBF, ENF, GRA, and MIF. The processed 500 m MODIS data were resampled to 1 km in the sinusoidal projection.

We used the elevation data from Global Multi-resolution Terrain Elevation Data (GMTED) [*J J Danielson and D B Gesch*, 2011], which is a global DEM data and has three resolution levels: 30-arc-second (1 kilometer), 15-arc-second (450 meters), and 7.5-arc-second (225 meters). The soil particle-size data is derived from the Global Soil Dataset for use in Earth System Models (GSDE) [*W Shangguan et al.*, 2014] that provides global gridded soil information, such as soil particle-size distribution, organic carbon, and nutrients, and quality control information, with a 30-arc-second (1 kilometer) resolution and for eight vertical layers to a depth of 2.3m. Both products on a global scale we used are in 1 km spatial resolution and were clipped to the region of North American in the sinusoidal projection.

The FLUXNET2015 Tier 1 FULLSET dataset (https://fluxnet.org/data/fluxnet2015-dataset/) was used for model development and evaluation. The used data include 8 DBF sites, 22 ENF sites, 7 GRA sites, and 3 MIF sites (in total 71, 118, 43, and 32 site-years data, respectively) in North America. The FLUXNET2015 dataset contains daily and half-hour or hourly meteorological and canopy flux data. Daily maximum and minimum temperatures are derived from half-hour or hourly temperature data. The latitude, longitude, elevation, and

9

soil particle-size information related to each site is obtained from site description in the FLUXNET2015 dataset. The LAI time series are extracted from the corresponding pixel in MOD15A2H according to the latitude and longitude of the study site. The extracted LAI time series were pre-processed similar to that for the entire North America.

3. Methods

We test two machine learning models, i.e., SVR and RF, on simulating vegetation LAI and GPP. The basic idea is that we firstly treat meteorological variables as input features and satellite-derived LAI as target for model training and validating. However, due to large fluctuations in daily or 8-day meteorological variables, direct training and prediction using the machine learning models would lead to unrealistic variation in the predicted LAI time series. We adopt a time smoothing method that has proven effective in reducing noises and buffering fluctuation to process daily vegetation LAI time series predicted directly by the machine learning models. Secondly, we treat meteorological variables and the simulated LAI as input features and field-measured GPP as target for model training and validating, and use the trained model for simulating the GPP time series. The details regarding to our method are described in the following sections.

3.1 Brief introduction of the machine learning methods

Support Vector Regression (SVR), a branch of the regression fitting method in support vector machine, is a machine learning model based on mathematical theories. The principle of SVR is to search the optimal hyperplane that minimizes the geometric distance of samples furthest from the hyperplane. SVR combines geometric distance minimization and tube regression constraint condition, and introduces slack variables to construct a convex quadratic programming problem with global optimal solutions, and the quadratic programming problems can be solved by the Lagrange multiplier method and its duality.

10

The core of the SVR algorithm is to introduce the kernel function $K(x_i, x)$ to replace the inner product operation in a low-dimensional space, which implicitly map the linear inseparable problem in the low-dimensional feature space to the high-dimensional feature space. The fitting function of SVR can be described as follows:

$$f_{SVR}(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)K(x_i,x)+b \quad (2)$$

where $f_{SVR}(x)$ denotes the fitting function of SVR, m denotes the number of training samples, x_i denotes the i^{th} support vector sample, α_i and $\hat{\alpha}_i$ respectively denote the upper and lower limits of the Lagrange multipliers, $K(x_i, x)$ denotes kernel function, b denotes the intercept of the fitting function.

As SVR can use a small number of sparse samples for model training and handle high-dimensional feature data, it can avoid the dimensional curse problem while computing in high-dimensional space. SVR has convenient operability and powerful nonlinear fitting capability. For more details about SVR, please refer to [*N Cristianini and J Shawe-Taylor*, 2000] and [*F Yang et al.*, 2006].

RF is a machine learning method based on the ensemble learning idea and it integrates multiple independent weak learners to improve the overall fitting and classification capability. RF utilizes the Bagging strategy to randomly select training samples and features while constructing the fitting decision trees and iteratively finds the optimal segmentation feature and segmentation point from the selected dataset at each regression decision tree. RF adopts the average prediction of all regression decision trees and verifies the trained model using the unselected samples as out-of-bag samples. The function used in RF to find the optimal segmentation feature and point is as follows:

$$f_{RF}(x)=\min_{j,s}\left[\min_{c1}\sum_{x_i\in R_1(j,s)}(x_i-c_1)^2+\min_{c2}\sum_{x_i\in R_2(j,s)}(x_i-c_2)^2\right] \quad (3)$$

where $f_{RF}(x)$ denotes the fitting function of RF, x_i denotes the i^{th} training sample, j and s denote the selected

optimal segmented feature and position, respectively, $R_1(j,s)=(x|x^{(j)}\le s)$ and $R_2(j,s)=(x\vee x^{(j)}\ge s)$, and c_1 and c_2 denote the input sample mean of the R_1 and R_2 datasets, respectively.

RF uses an integrated strategy and it is suitable for processing high-dimensional feature space data. RF utilizes independent out-of-bag samples to evaluate models during the training phase and provides unbiased estimations. RF is widely applied in the regression fitting studies because the model has the anti-noise and anti-overfitting ability and is insensitive to parameter settings [*Chen et al.*, 2017; *L Mareike et al.*, 2012; *R A V Rossel and T Behrens*, 2010].

3.2 Implementation of the machine learning methods on modeling LAI and GPP

There have tremendous efforts to develop numerical models and machine learning models that simulate vegetation GPP using LAI as the variable representing vegetation cover and meteorological variables at the corresponding time. Note that vegetation LAI changes in response to meteorological conditions and the current phenology models still have considerable uncertainties, one key issue is how to accurately simulate LAI time series via daily meteorological data. Once robust models are developed to predict vegetation LAI using meteorological variables, we can incorporate the predicted LAI as well as the meteorological variables into the vegetation photosynthesis model, such that terrestrial ecosystem models no longer require field-measured or satellite-derived LAI data and our abilities on modeling the vegetation processes are greatly improved.

Our idea to simulate LAI time series is to treat meteorological variables as the input features and satellite-derived LAI as the target to train machine learning models such as SVR and RF. Inputs and targets to the machine learning models need to be normalized for model training and the modeled data from the trained machine learning models are denormalized to obtain the predicted LAI. Note that vegetation LAI predicted from the above method has lags and unrealistic fluctuations in the time series as compared with the observed

ones. In the study, we use the simple moving average method to buffer the fluctuations and account for the time lagging effect in the modeled LAI time series. The equation of the simple moving average method is as follows:

$$LAI = SMA\left(LAI_s, n_{day}\right) \quad (4)$$

where LAI_s denotes LAI modeled by the machine learning models and n_{day} denotes the number of days in the moving windows. Here we set n_{day} as 25 for the daily time scale based on previous studies.

Eventually, the LAI time series obtained from Equation 7 are considered as the predicted LAI in our study. We use normalized LAI and meteorological variables as input features and normalized GPP as the target to train the machine learning models, and then use the trained machine learning models to make predictions of vegetation GPP.

In the continental studies, we randomly selected 1000 pixels for each of the four plant functional types and using the time series data in the entire year as the training and validation samples. As each plant functional type in each pixel contains 46 points per year in the 8-day composite data, there are 46000 samples for each plant functional type. We used 70% samples for model training and 30% samples for model validation. The testing samples used to test the model accuracy independently comes from randomly selected 50 pixels for each plant functional type for each 8-day composite data, resulting in 2300 testing samples in total. In the site scale studies, we selected 2/3 of the site-year flux tower data as training and validation samples, and the remaining as the testing samples. The number of decision trees in the RF model was set as 500. We adopted radial basis function regression in the SVR model. We used the grid 8-fold cross-validation method to optimize the penalty factor and the parameter in gamma function of kernel function so as to minimize the root mean square error (RMSE) between modeled results and validation samples.

3.3 Comparative studies using the Biome-BGC model

The Biome-BGC model is a process-based terrestrial ecosystem model based on the sun-shade two-leaf model that simulates the states and fluxes of carbon, nitrogen, and water, in the composition of vegetation and soil within terrestrial ecosystems [*P Thornton and B Law*, 2002; *Q Wang et al.*, 2005]. Biome-BGC is composed of a variety of physical and biological process modules, where the physical modules mainly include solar radiation, precipitation and water cycle processes and the biological process modules mainly include autotrophic and heterotrophic respiration, photosynthesis, microbial decomposition, carbon and nitrogen cycles, etc. In the phenology module, Biome-BGC first simulates the timing of vegetation key phenophases, such as spring onset and autumn offset, based on the climate forcing data, and then uses the carbon and nitrogen cycle module to simulate LAI time series for a specific vegetation type. The input climate forcing data consist of daily maximum and minimum temperature, daylight average temperature, daily total precipitation, daylight average partial pressure of water vapor, daylight average shortwave radiant flux density, and day length. Biome-BGC has a number of physical and eco-physiological parameters, where the eco-physiological parameters mainly depend on the plant functional type and the physical parameters contain latitude, elevation, soil particle-size, and so on. We use the empirical physio-ecological parameterization schemes in [*M A White, Thornton, P. E. , Running, S. W. , & Nemani, R. R. .* , 2000] and [*F A Tatarinov and E Cienciala*, 2006] for different plant functional types. As we found no physio-ecological parameterization schemes for the MIF type in literatures, we utilize the grid optimization algorithm based on flux tower data to optimize key eco-physiological parameters for MIF, including maximum stomatal conductance, the ratio of new fine root C to new leaf C, canopy average specific leaf area, fraction of leaf nitrogen in Rubisco, the ratio of carbon to nitrogen in leaves, annual leaf and fine root turnover fraction. The 4.2 version of the Biome-BGC model is unable to perform regional grid simulation directly. We extracted 100 × 100 pixels of the climate forcing datasets in each MODIS tile based on equal interval sampling, and then interpolated the modeled results into each MODIS tile using the spline function interpolation method as the continental results modeled

297 by Biome-BGC.

298

299 **4. Results**

300 **4.1 Continental-scale modeling of leaf area index**

301 Fig. 2 compares the spatial distribution of modeled and satellite-derived annual average LAI in 2010. Both the
302 SVR and RF models predicted annual average LAI consistent with satellite observations. Annual average LAI
303 modeled by Biome-BGC are largely different from satellite data. Biome-BGC generally overestimates LAI in
304 the NEF and GRA types and underestimates LAI in both the DBF and MIF types. Fig. 2e indicates that
305 latitudinal average of annual average LAI derived from both the machine learning models and remote sensing
306 data gradually increase from north to south and Biome-BGC has large overestimation around about 50 to 60
307 degrees north latitude. The latitudinal average of annual average LAI modeled by the SVR and RF models
308 have strong correlations with the MODIS data ($R = 0.985$ and 0.987, respectively), and the corresponding
309 correlation between Biome-BGC and the MODIS data is much lower ($R=0.703$).

310

15

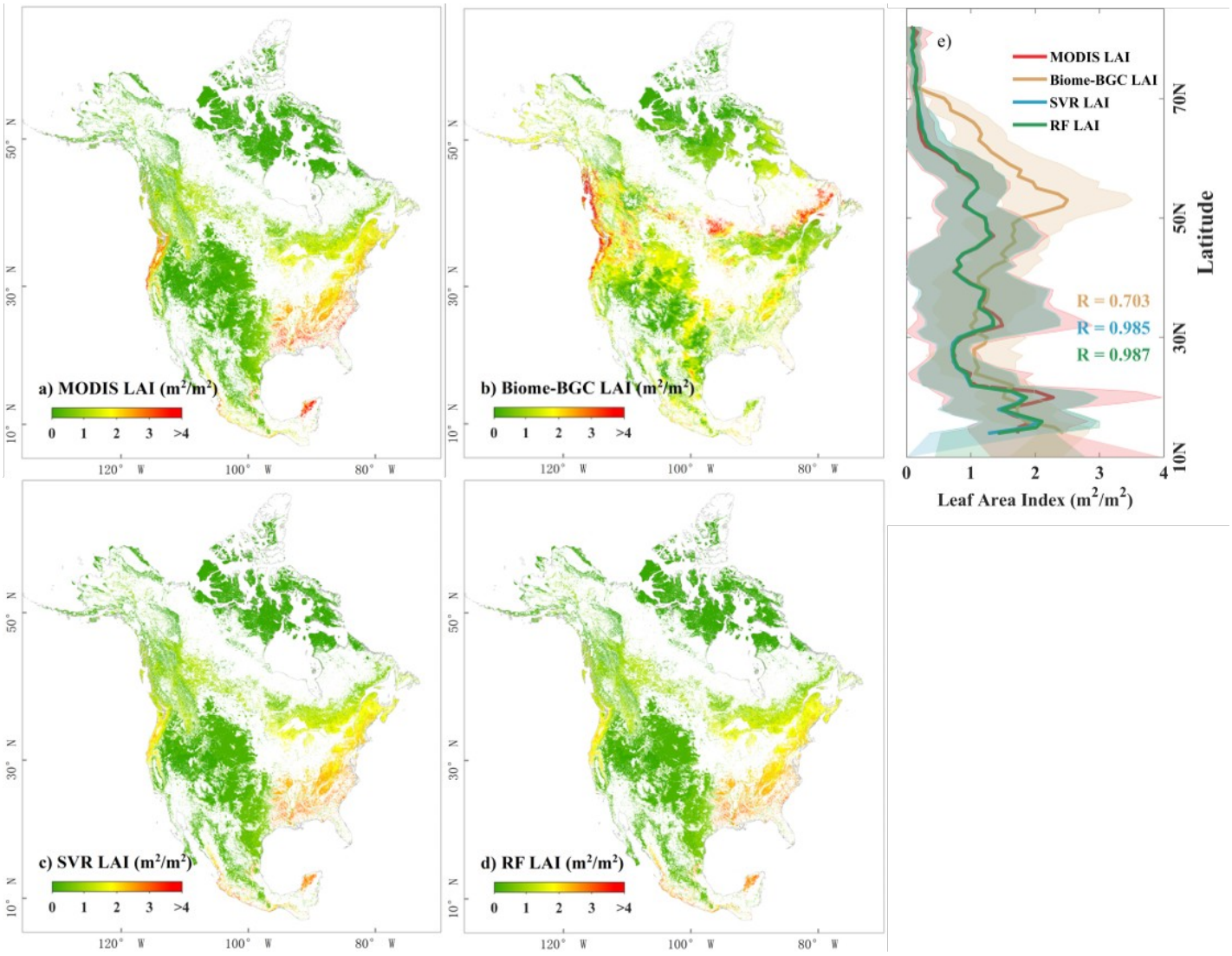


Fig. 2. The spatial distributions of annual average LAI derived from a) the MODIS data, b) the Biome-BGC model, c) the support vector regression (SVR) model, and d) the random forests (RF) model, and e) their latitudinal averages in 2010 across North America. The units for LAI are m^2 (leaf area) per m^2 (ground area).

Fig. 3 shows the error analysis between modeled and satellite-derived LAI time series. The LAI time series modeled by both SVR and RF are consistent with the MODIS data with only slight underestimates in the south. RMSE between machine learning-modeled and satellite-derived annual LAI are generally lower than 1.5 m^2/m^2 and the correlation coefficient is greater than 0.9 for most of the areas except the GRA areas in the north and the ENF areas near the coast of northwestern United States. Compared with satellite data, Biome-BGC underestimates annual average LAI in many DBF areas with negative Bias lower than -1.0 m^2/m^2 and

16

RMSE greater than 1.5 m^2/m^2, and largely underestimates in most MIF areas with negative Bias lower than -1.5 m^2/m^2 and RMSE greater than 2.0 m^2/m^2. For the ENF type, Biome-BGC largely overestimates annual average LAI compared with satellite data. For the GRA type, LAI time series modeled by Biome-BGC have low correlation with satellite data in most areas ($R < 0.6$) as well as positive Bias greater than 1 m^2/m^2). In general, the machine learning-based approaches make accurate predictions of the LAI time series in most of the areas with small errors.

17

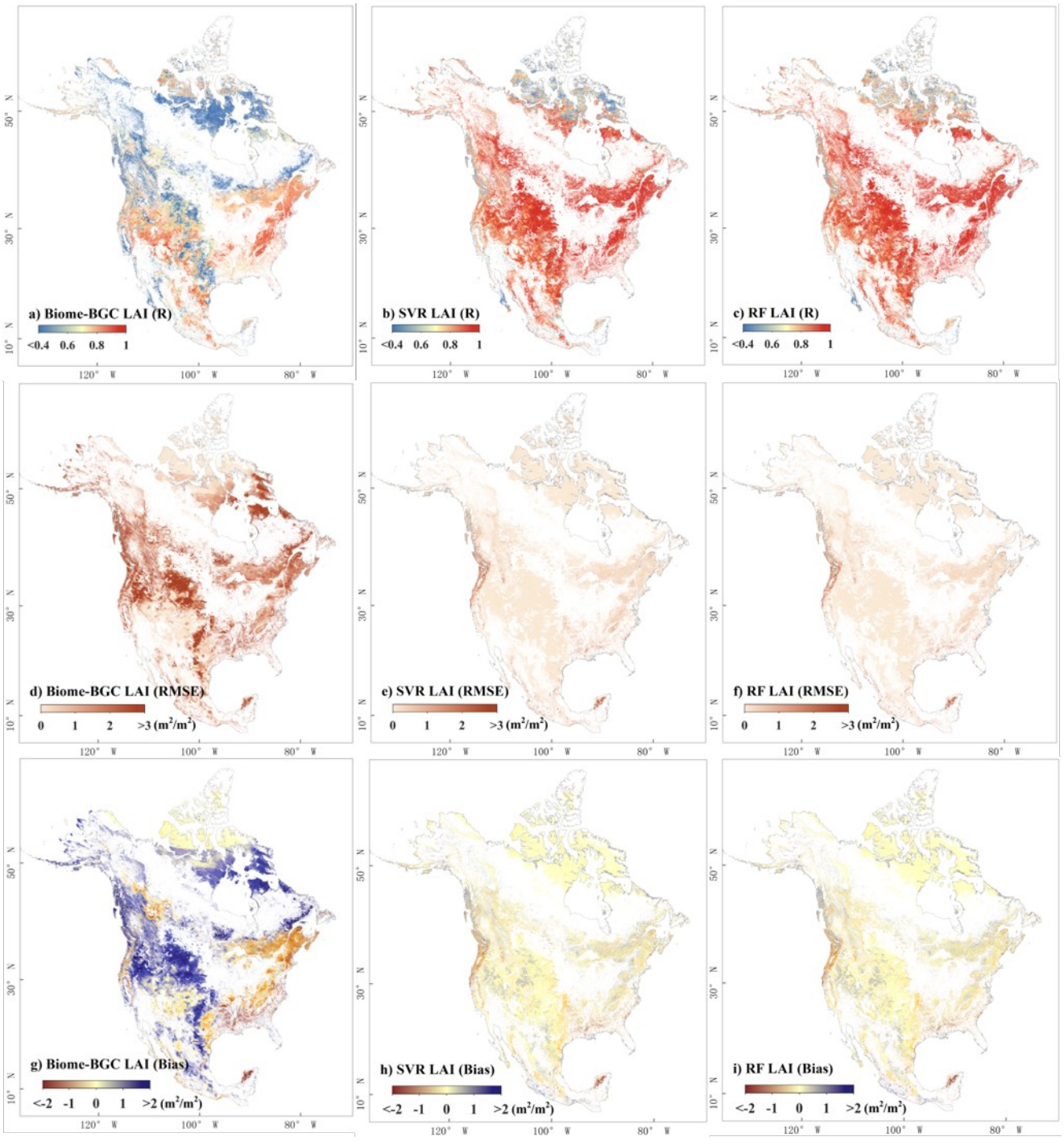


Fig. 3. The spatial distributions of the metrics for the assessment of model performance, including the correlation coefficient (top row), RMSE (middle row), and Bias (top row) of 8-day LAI time series in 2010 as derived between the Biome-BGC model and MODIS (left column), between the SVR model and MODIS (middle column), and between the RF model and MODIS (right column), respectively, across North America.

Different pixels are shown as examples (Fig.4) to illustrate the model performance on simulating the LAI time series. The machine learning-based approaches are able to capture the seasonal dynamics of LAI accurately,

and the Biome-BGC model produces large errors. The Biome-BGC model is able to make predictions on key phenophases, such as onset and offset, in the DBF and MIF types, but the amplitudes of the simulated seasonal LAI time series do not match the observed ones. In addition, the Biome-BGC model assumes zero LAI in the non-growing period but the assumption is inconsistent with satellite observation. Because Biome-BGC does not distinguish seasonal cycles in a year when simulating the phenology of evergreen forests, the simulated LAI time series in ENF do not have obvious seasonal cycles and do not match the satellite data. For the GRA type, Biome-BGC does not accurately capture the timing of key phenophases during a vegetation growing season, resulting in large overestimation or underestimation of LAI in the time series.

19

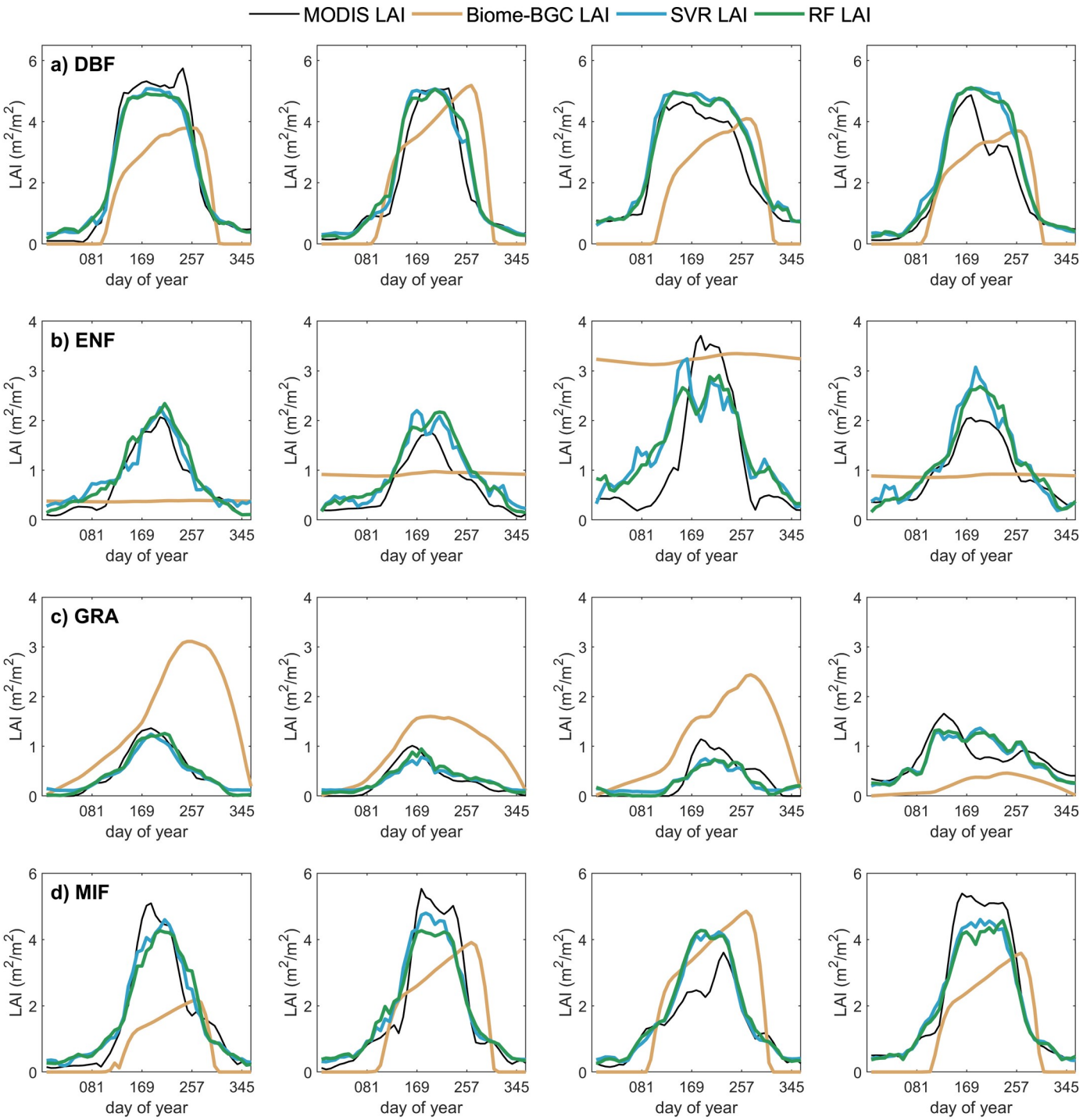


Fig. 4. Comparisons between modeled and satellite-derived 8-day LAI time series for four different example pixels (each column) in deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row), grasslands (the third row), and mixed forests (bottom row).

Fig.5 shows the comparisons between model simulation and satellite observations for the independent test dataset. The regression analysis suggests that the simulation results of the RF model, among three methods,

20

have the highest positive linear correlation (R = 0.944, 0.819, 0.972, and 0.933 for DBF, ENF, GRA, MIF, respectively), and the smallest RMSE (RMSE = 0.639, 0.604, 0.289, and 0.658 m^2/m^2 for DBF, ENF, GRA, MIF, respectively) with the MODIS observations. The model performance of RF is close to that of SVR. The correlation of between Biome-BGC modeled and satellite-derived LAI is insufficient (R=0.742, 0.091, 0.474, and 0.583 for DBF, ENF, GRA, and MIF, respectively) and the RMSEs are larger than 1 m^2/m^2.

21

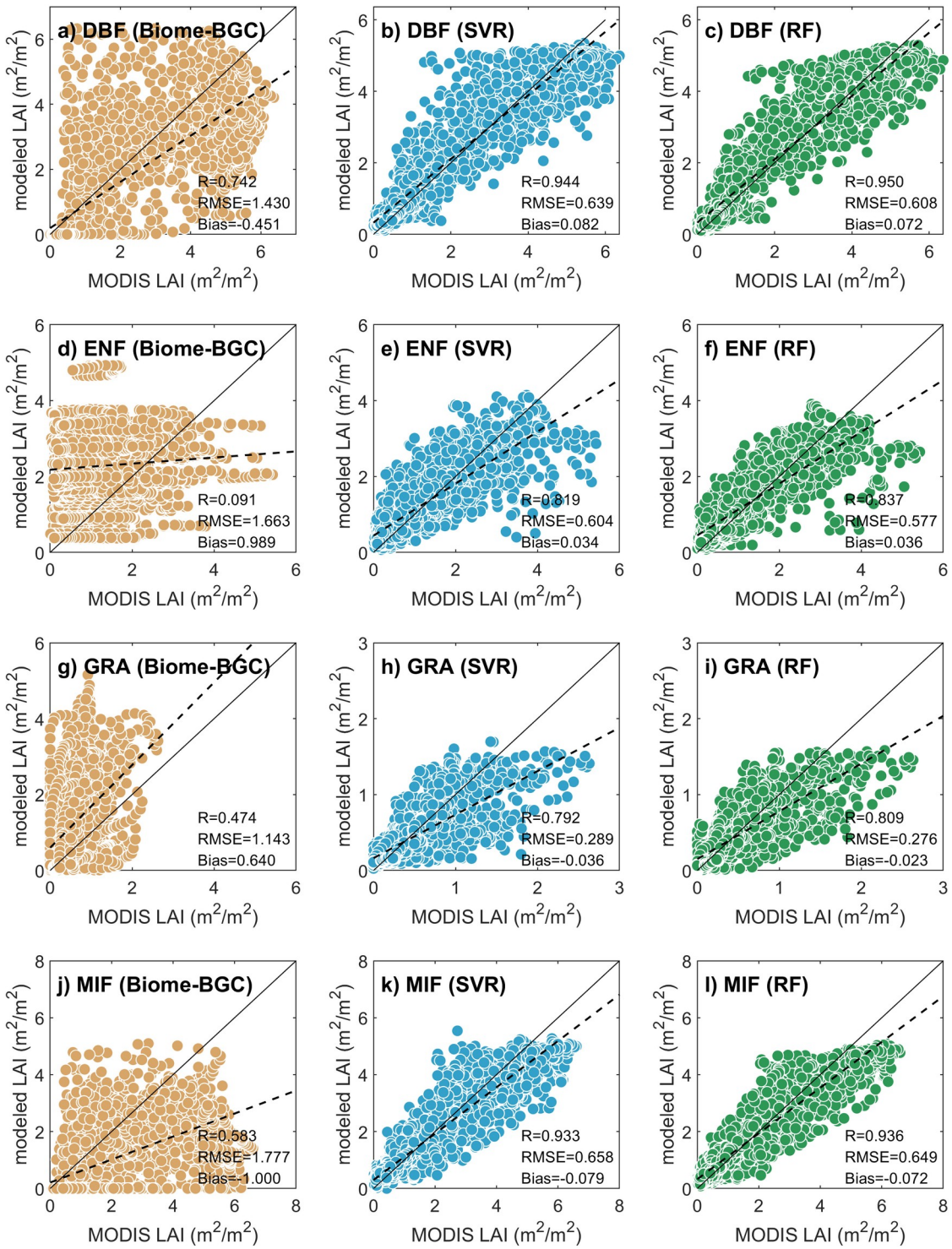


Fig. 5. Scatter plots are shown for comparisons between Biome-BGC-modeled and MODIS-derived LAI (left column), comparisons between SVR-modeled and MODIS-derived LAI (middle column), and comparisons between RF-modeled and MODIS-derived LAI (right column) for the plant functional types of deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row), grasslands (the third row), and mixed forests (bottom row), respectively. Each subplot contains the time series data from 50 randomly selected pixels in each plant functional type. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

22

4.2 Continental-scale modeling of gross primary productivity

Figure 6 shows the spatial distribution of annual total GPP in 2010 derived from the MODIS GPP product and different models across North America. When compared with the MODIS GPP product, the machine learning-based approaches outperform the Biome-BGC model. For the DBF type, the spatial distributions of GPP derived from both the RF model and the SVR model are consistent with derived from MODIS, whereas Biome-BGC has underestimates in the north and overestimates in the south. For the MIF type, the RF model performs better than both the SVR model and the Biome-BGC model and Biome-BGC shows large underestimates as compared with the MODIS GPP product. For the ENF type, both the machine learning-based models have underestimates of annual total GPP along the coastal area in northwestern United States. Compared with the MODIS product, the Biome-BGC model nearly overestimates annual total GPP of the ENF type across the study region and underestimates annual total GPP of the GRA type in southwestern United States. The latitudinal average of annual total GPP derived from both machine learning-based approaches have high correlation with the MODIS observations (R=0.961 and 0.987 for SVR and RF, respectively). The correlation coefficient between latitudinal average of annual total GPP derived from the Biome-BGC model and that derived from the MODIS data is 0.830 (Figure 6e).

23

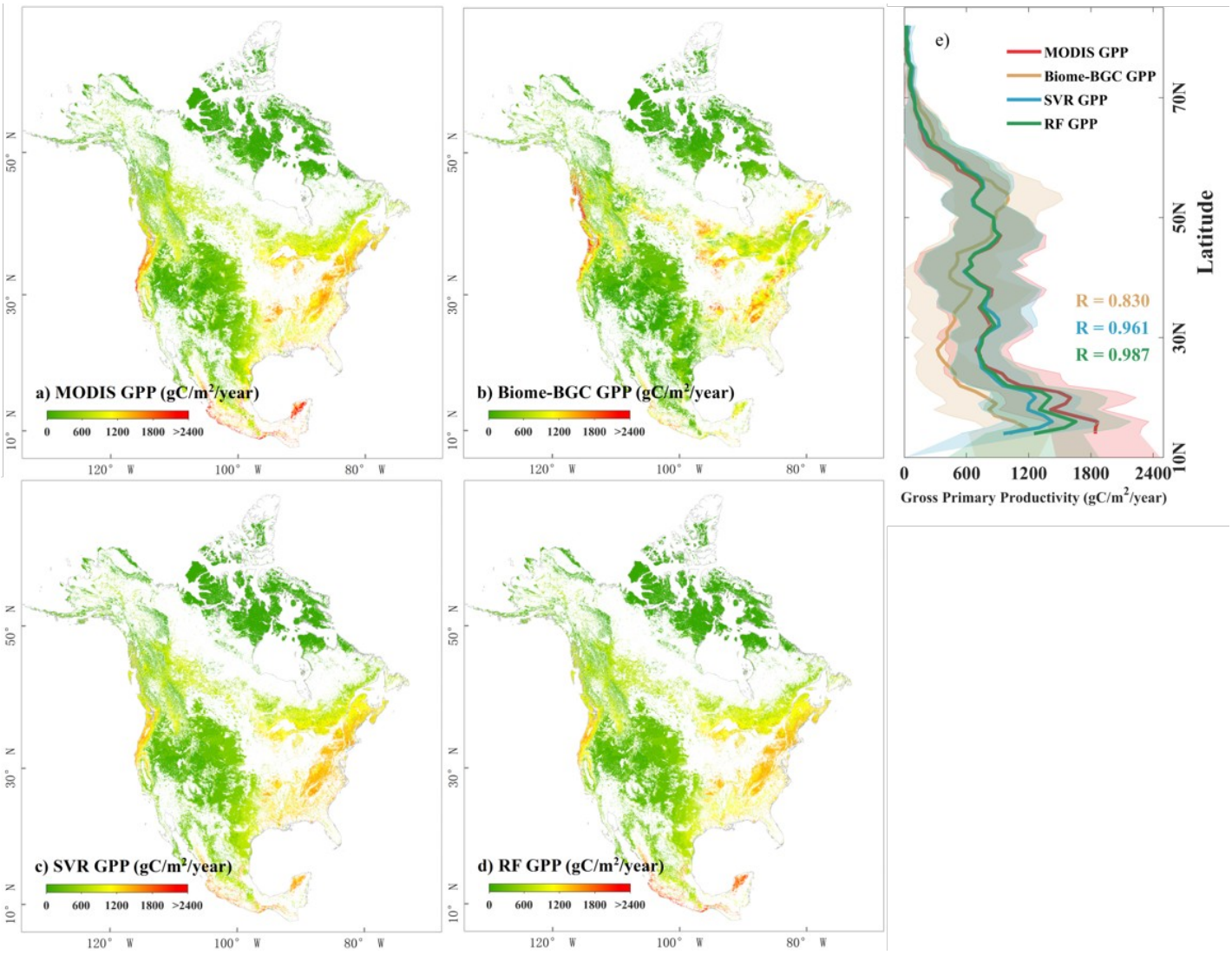
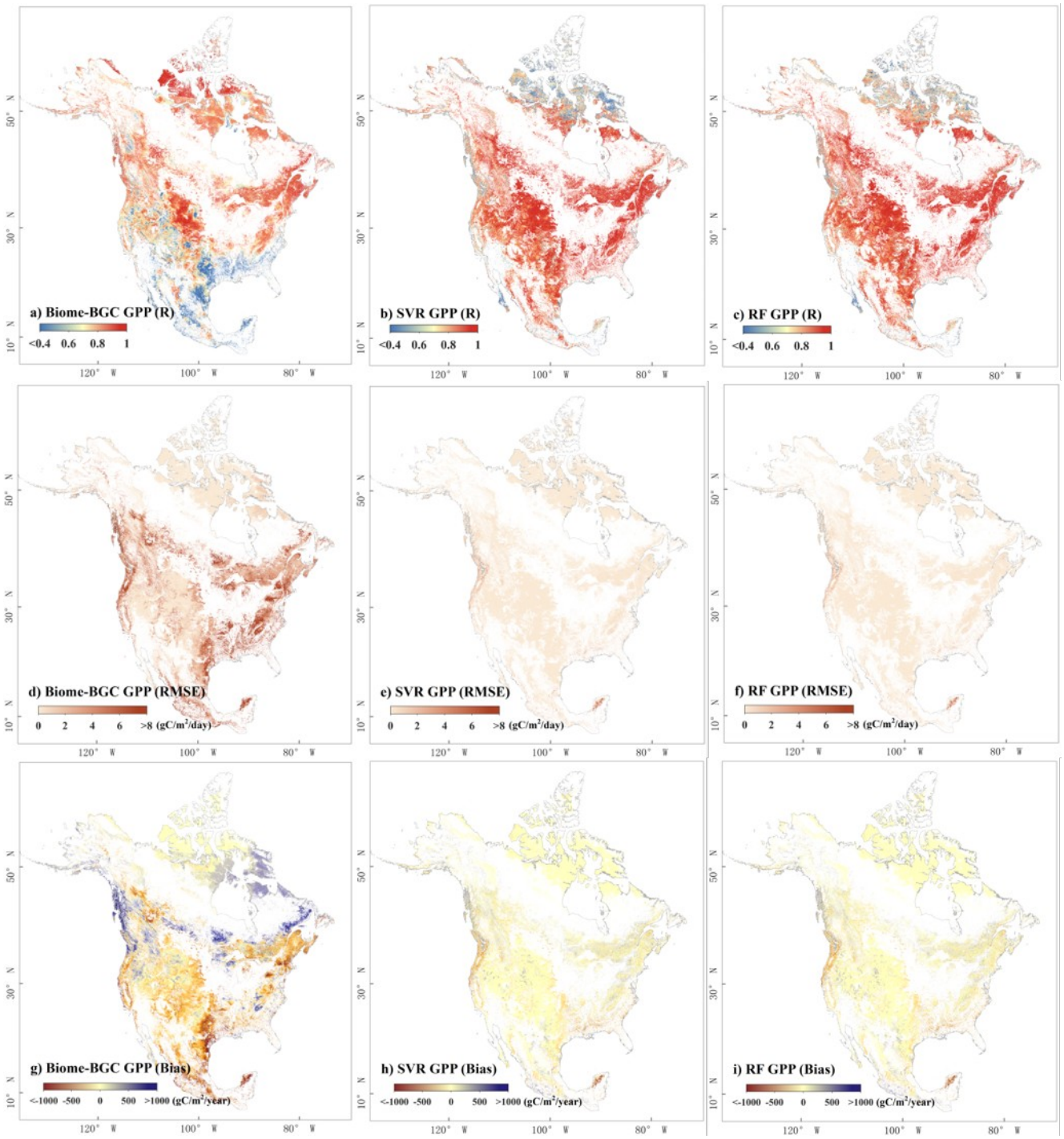


Fig. 6. The spatial distributions of annual total GPP derived from a) the MODIS product, b) the Biome-BGC model, c) the support vector regression (SVR) model, and d) the random forests (RF) model, and e) their latitudinal averages in 2010 across North America. The units for GPP are gC per m^2 per year.

Figure 7 shows the spatial distribution of error metrics for the model assessment on modeling 8-day GPP time series in 2010. The machine learning-based approaches produce results consistent with the MODIS GPP product in most areas with high correlation (R > 0.9) and low errors (RMSE < 1 gC/m^2/year). Compared with the MODIS 8-day GPP product, Biome-BGC has correlation coefficient less than 0.6 and RMSE greater than 3 gC/m^2/year for many areas. Both machine learning models produce satisfactory simulation results for the DBF, MIF, and GRA types across the study region but have underestimation in ENF along the coasts of

24

394 northwestern United States. The Biome-BGC model generally has large overestimation in ENF with Bias
395 greater than 500 gC/m^2/year and underestimation in MIF and GRA.
396

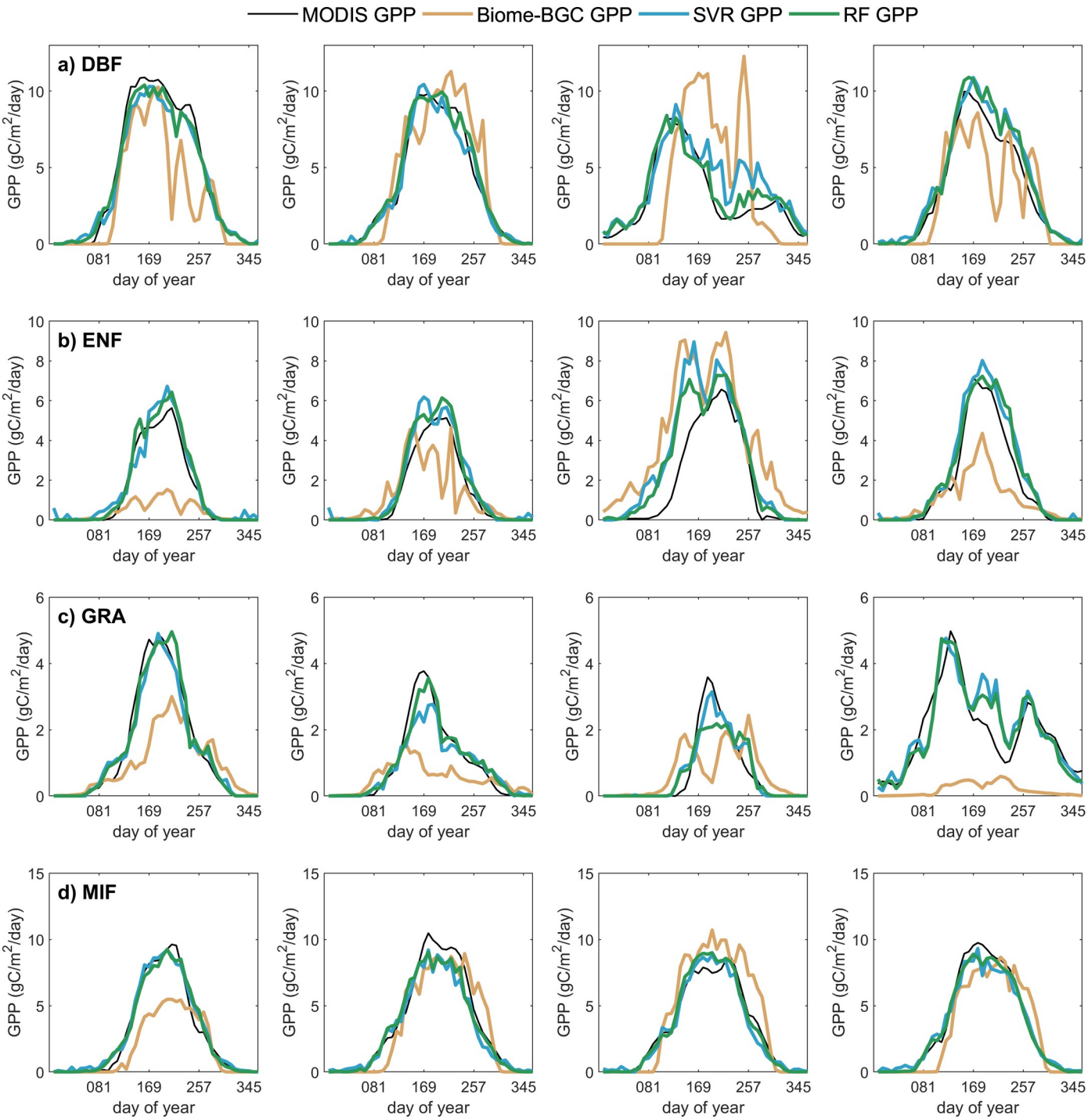


397

398 Fig. 7. The spatial distributions of the metrics for the assessment of model performance, including the
399 correlation coefficient (top row), RMSE (middle row), and Bias (top row) of 8-day LAI time series in 2010 as
400 derived between the Biome-BGC model and MODIS (left column), between the SVR model and MODIS
401 (middle column), and between the RF model and MODIS (right column), respectively, across North America.

Figure 8 shows the 8-day time series of both modeled and satellite-derived GPP for four example pixels for each plant functional type. The machine learning models can well capture the seasonal cycles of GPP. Compared with the remote sensing product, the Biome-BGC model has the best performance in the DBF type, but unstable performance in the ENF type and large underestimation in the GRA type. Figure 9 shows the regression analysis of modeled and satellite-derived GPP for different plant functional types. The SVR model has strong correlation with the MODIS GPP data for four studied vegetation types (R = 0.936, 0.927, 0.878, and 0.960, for DBF, ENF, GRA, and MIF, respectively)and so does the RF model (R = 0.967, 0.952, 0.885, and 0.982, for DBF, ENF, GRA, and MIF, respectively). The RMSE values between SVR-modeled and satellite-derived GPP are 1.215, 1.038, 0.671, and 0.913 $gC/m^2/day$, for DBF, ENF, GRA, and MIF, respectively. By comparison, the RMSE values between RF-modeled and satellite-derived GPP are 0.882, 0.840, 0.649, and 0.621$gC/m^2/day$, for DBF, ENF, GRA, and MIF, respectively. The Biome-BGC model has considerable errors on modeling GPP in different plant functional types. There are large underestimation of Biome-BGC in the GRA and MIF types (Bias = -0.497 and -0.771 $gC/m^2/day$ for GRA and MIF, respectively). In general, Biome-BGC underperforms the machine learning-based models, probably because Biome-BGC needs better representation of the phenology model for modeling the LAI time series accurately.

26



418

419 Fig. 8. Comparisons between modeled and satellite-derived 8-day GPP time series for four different example
420 pixels (each column) in deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row),
421 grasslands (the third row), and mixed forests (bottom row).
422

27

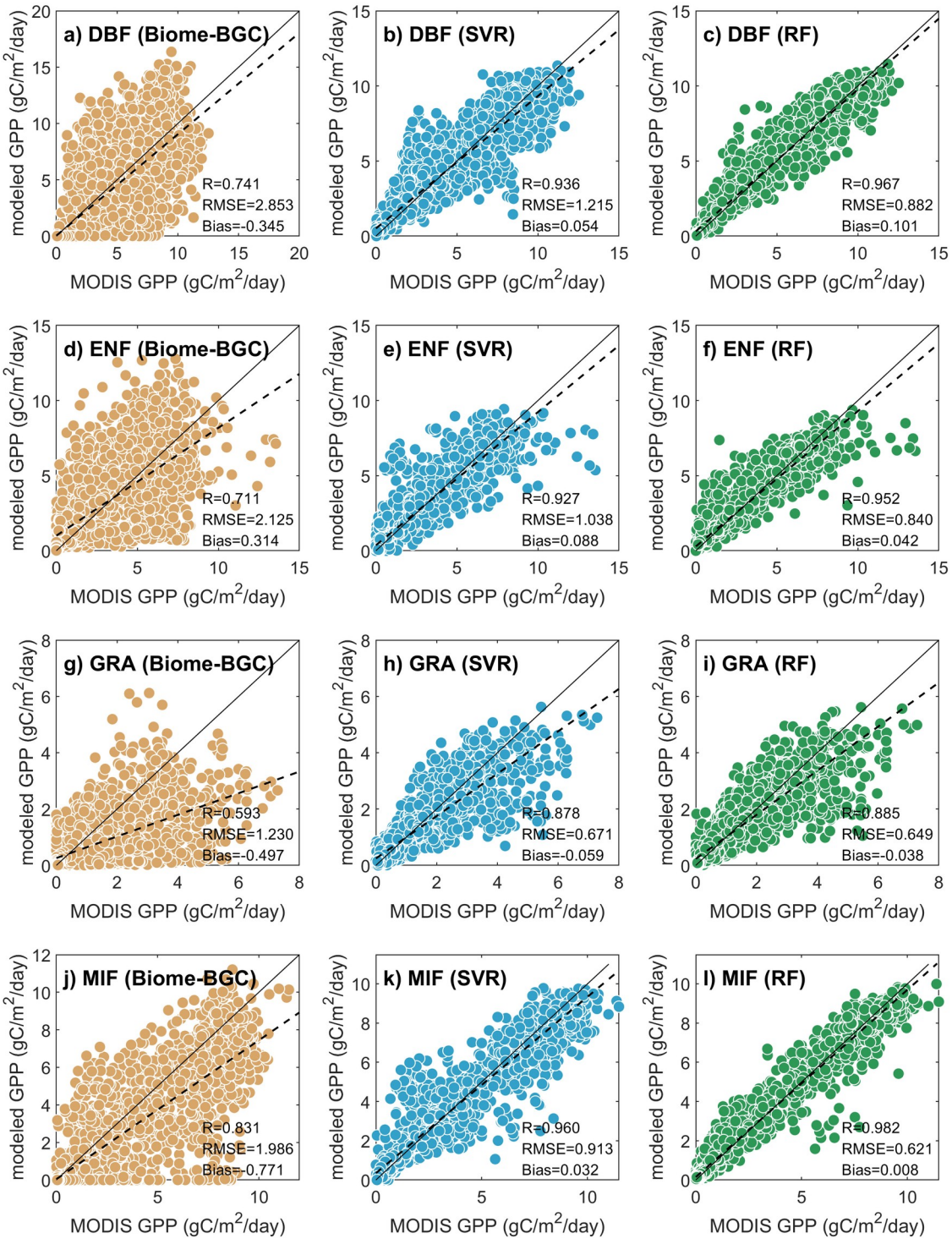


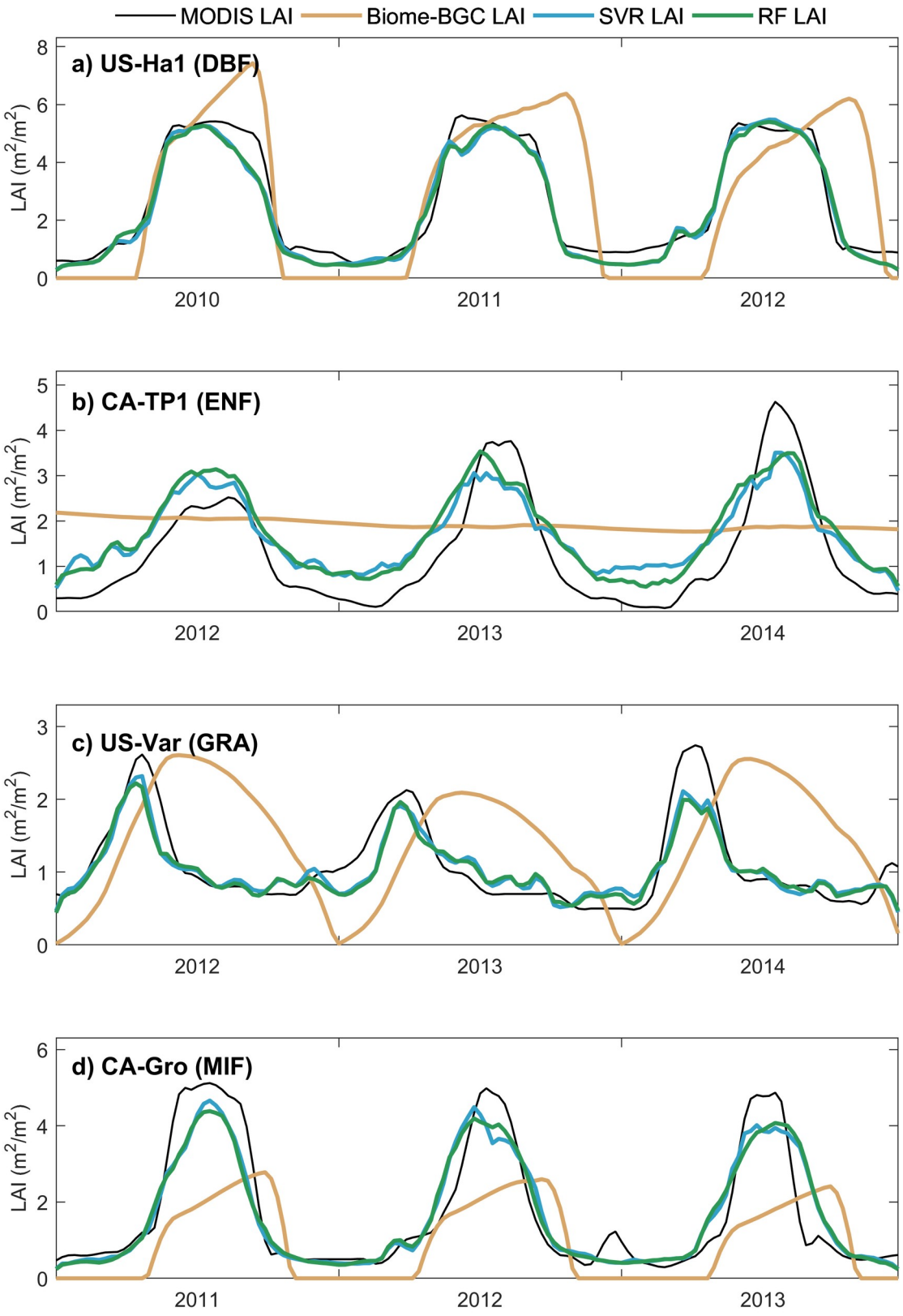
Fig. 9. Scatter plots are shown for comparisons between Biome-BGC-modeled and MODIS-derived GPP (left column), comparisons between SVR-modeled and MODIS-derived GPP (middle column), and comparisons between RF-modeled and MODIS-derived GPP (right column) for the plant functional types of deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row), grasslands (the third row), and mixed forests (bottom row), respectively. Each subplot contains the time series data from 50 randomly selected pixels in each plant functional type. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

28

4.3 Site-scale modeling of leaf area index

Figure 10 shows both modeled and satellite-derived LAI time series at four different flux tower sites. Compared with remote sensing data, the machine learning models have accurate simulation and Biome-BGC still has room for improvements, especially in the ENF and GRA types. In the ENF type, Biome-BGC assumes no apparent growing season and produces nearly stable LAI in the time series, while both remote sensing data and the machine learning models have distinctive seasonal cycles. In the GRA type, the machine learning-based approaches can well capture irregular intra-year variation in LAI derived from the MODIS observations. The timings of phenology metrics such as onset and offset modeled by Biome-BGC do not match satellite data, resulting large differences in seasonal LAI dynamics between Biome-BGC and MODIS data. For both the DBF and MIF types, the machine learning models are able to make accurate predictions on the time series of LAI. Biome-BGC could predict accurate estimates on the onset of the growing season but lagged estimates on the offset of the growing season. In addition, maximum LAI predicted by Biome-BGC is largely different from that derived from satellite data.

29



445

446 Fig. 10. Comparisons between modeled and satellite-derived 8-day LAI time series for the flux tower sites of
447 US-Ha1 (deciduous broadleaf forest), CA-TP1 (evergreen needleleaf forest), US-Var (grasslands), and CA-
448 Gro (mixed forests).

Figure 11 shows the regression analysis between modeled and satellite-derived 8-day LAI for sites of different vegetation types. Based on all the test dataset, the results predicted by SVR have strong correlation (R = 0.961, 0.809, 0.900, and 0.928, for DBF, ENF, GRA, and MIF, respectively) and low errors (RMSE = 0.594, 0.697, 0.445, and 0.678 m^2/m^2, for DBF, ENF, GRA, and MIF, respectively) as compared with satellite observations. RF has the model performance similar to SVR and is able to simulate the seasonal cycles of LAI accurately with low biases (Bias = 0.101, 0.042, 0.038, and 0.008 m^2/m^2, for DBF, ENF, GRA, and MIF, respectively) . Both SVR and RF have underestimation in the ENF and GRA types when LAI is large. The LAI time series modeled by Biome-BGC have low correlation coefficients (R=0.681, 0.105, 0.474, and 0.657, for DBF, ENF, GRA, and MIF, respectively) and large errors (RMSE = 1.681, 1.256, 0.990, and 1.417 m^2/m^2, for DBF, ENF, GRA, and MIF, respectively) as compared with satellite data.

31

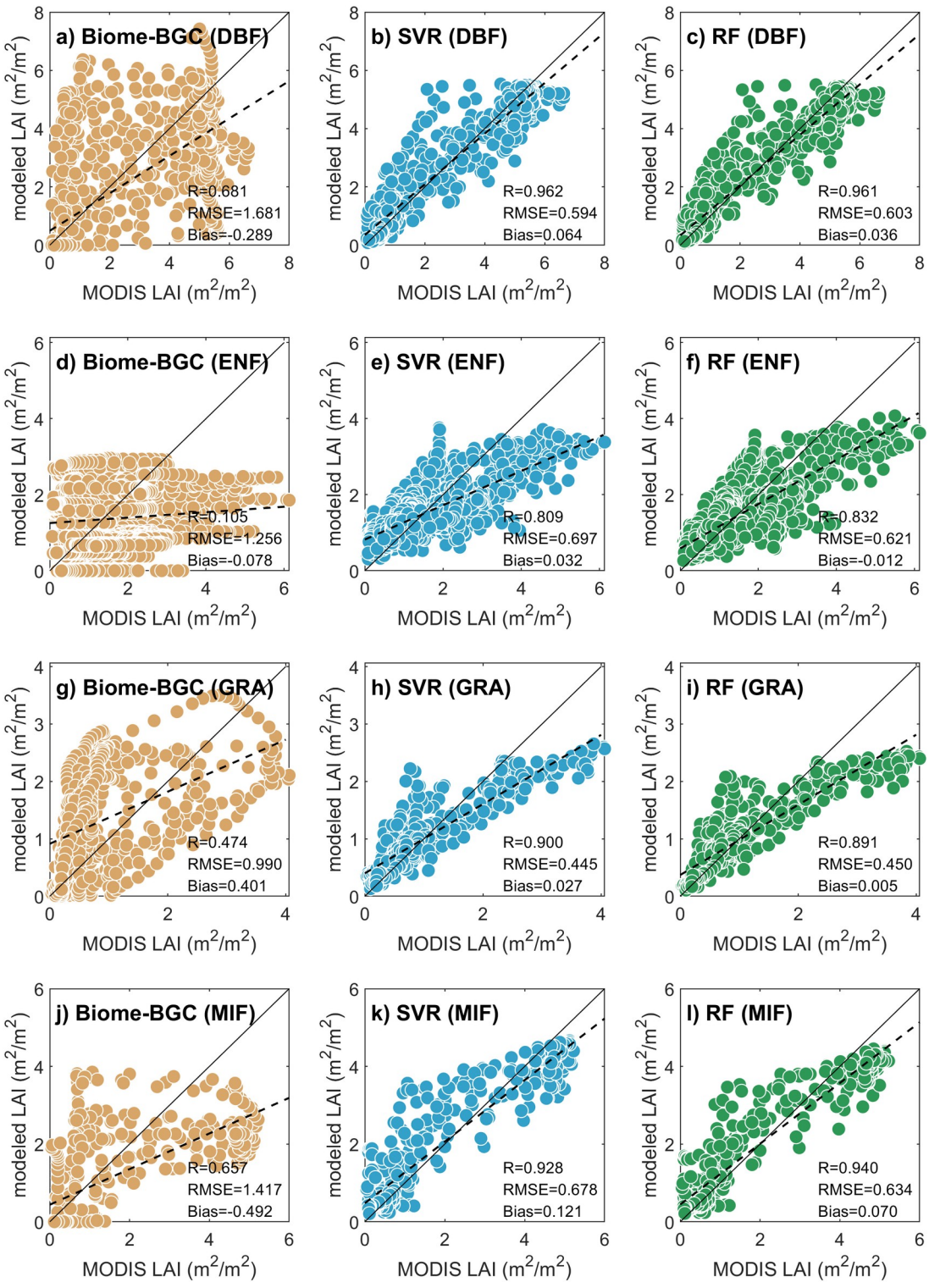


Fig. 11. Scatter plots are shown for comparisons between Biome-BGC-modeled and MODIS-derived 8-day LAI (left column), comparisons between SVR-modeled and MODIS-derived 8-day LAI (middle column), and comparisons between RF-modeled and MODIS-derived 8-day LAI (right column) for the plant functional types of deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row), grasslands (the third row), and mixed forests (bottom row), respectively. Each subplot contains the entire time series data from flux towers. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

32

The modeled and satellite-derived annual average LAI are compared for each individual site-year data in Figure 12. The regression analysis indicates that both machine learning models could well predict has annual average LAI with high correlation coefficients (R = 0.811 and 0.839 for SVR and RF, respectively) and low errors (RMSE = 0.425 and 0.392 m^2/m^2 for SVR and RF, respectively) as compared with satellite data. Biome-BGC has considerable errors on predicting annual average LAI where RMSE is 0.803 m^2/m^2 and Bias is -0.060 m^2/m^2 as compared with the MODIS data.

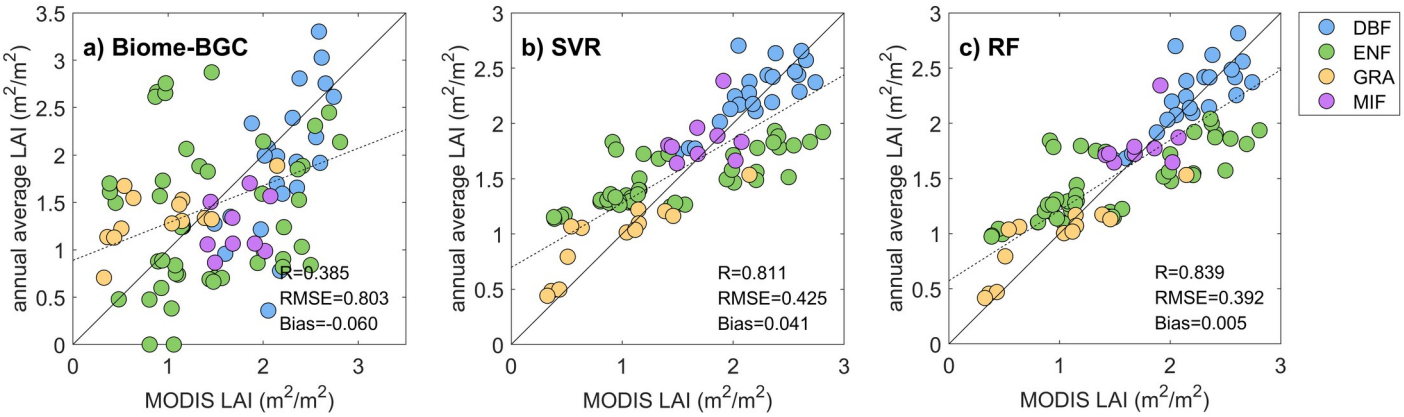


Fig. 12. Scatter plots are shown for comparisons a) between Biome-BGC-modeled and MODIS-derived annual average LAI, b) between SVR-modeled and MODIS-derived annual average LAI, and c) between RF-modeled and MODIS-derived annual average LAI using all site-year flux tower data. The blue, green, orange and purple points denote the site-year data from the DBF, ENF, GRA and MIF types, respectively. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

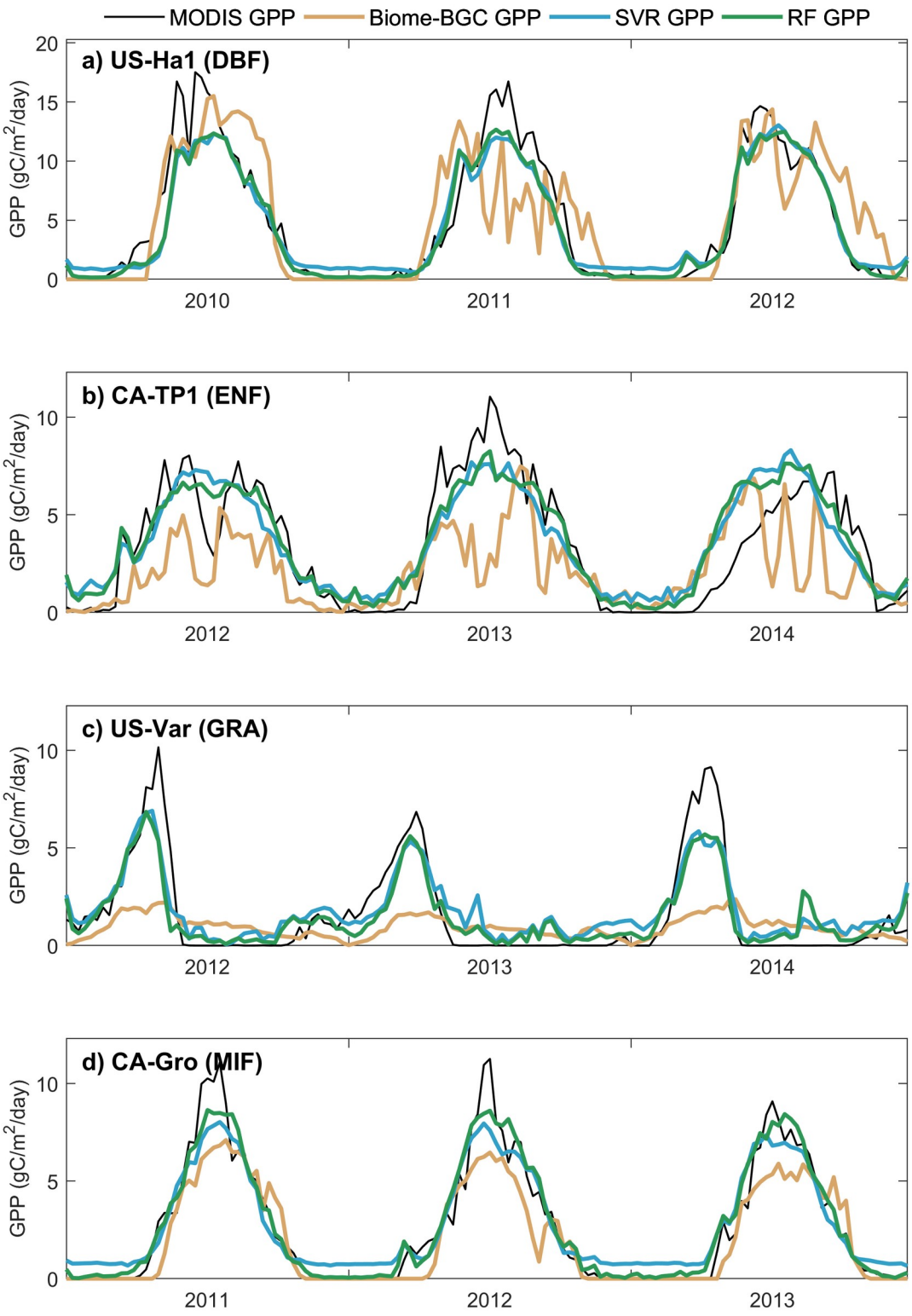
4.4 Site-scale modeling of gross primary productivity

Figure 13 shows the 8-day GPP time series simulated by different models and measured from the eddy-covariance flux tower data at four different sites. The machine learning approaches perform well on modeling 8-day GPP time series across sites. For the GRA sites, the machine learning models are capable of capturing seasonal GPP fluctuations with slight underestimation near the peaks and the Biome-BGC model has large underestimation on GPP as compared with the flux tower data. For the ENF sites, Biome-BGC has underestimation during the summer growing period. For the DBF and MIF types, the machine learning

33

models are able to simulate both the amplitude and the phase of the 8-day GPP time series accurately, and the Biome-BGC model has large fluctuation and overestimation in DBF as compared to the flux tower measurements.

34



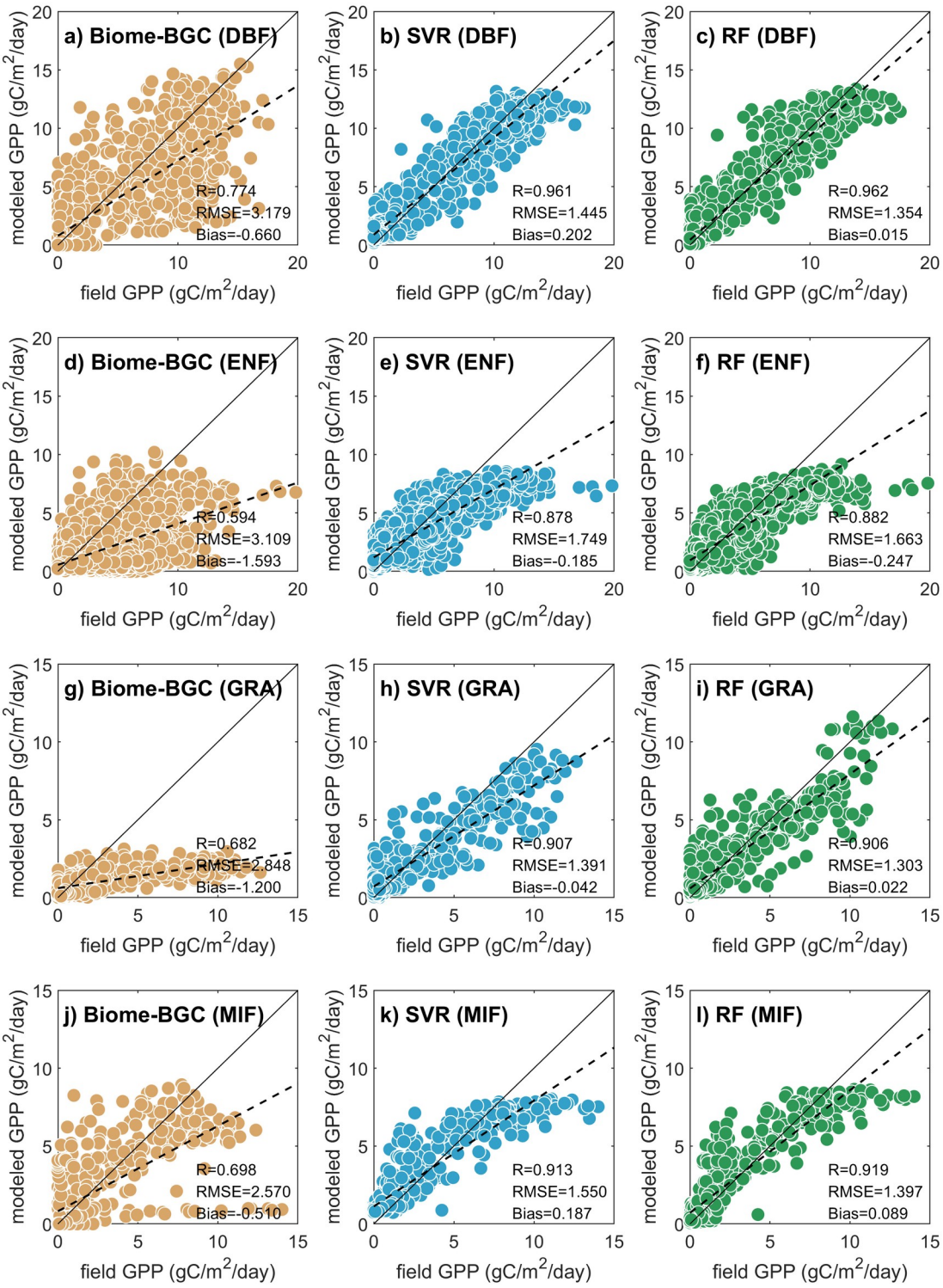
492

493 Fig. 13. Site 8-day time series comparison between modeled and measured of GPP in each PFTs, respectively.
494 The reference GPP time series (black line) are derived from the MODIS data. a) line denotes modeled and
495 MODIS GPP in US-Ha1 site (DBF), b) line denotes modeled and MODIS GPP in CA-TP1 site (ENF), c) line
496 denotes modeled and MODIS GPP in US-Var site (GRA), d) line denotes modeled and MODIS GPP in CA-
497 Gro site (MIF).

35

498 Figure 14 compares the modeled and field-measured GPP data for different plant functional types using all the
499 8-day test data. RF gives results with high correlation coefficients (R = 0.962, 0.882, 0.906, and 0.919 for
500 DBF, ENF, GRA, and MIF, respectively) and low errors (RMSE = 1.354, 1.663, 1.303, and 1.397 $gC/m^2/day$
501 for DBF, ENF, GRA, and MIF, respectively) with flux tower observation. SVR has the model performance
502 similar to RF as indicated by the assessment metrics. Both RF and SVR are able to accurately simulate the 8-
503 day GPP data but have underestimation when GPP is high during the summer growing period, particular for
504 the ENF and MIF sites. Biome-BGC has considerable negative biases (Bias = -0.660, -1.593, -1.200, and -
505 0.510$gC/m^2/day$, for DBF, ENF, GRA, MIF, respectively) and large RMSE values (RMSE = 3.179, 3.109,
506 2.848, and 2.570 $gC/m^2/day$) on simulating GPP as compared with flux tower measurements. Note that
507 Biome-BGC has large underestimates in the GRA type.
508

36



509

510 Fig. 14. Scatter plots are shown for comparisons between Biome-BGC-modeled and field-measured 8-day
511 GPP (left column), comparisons between SVR-modeled and field-measured 8-day GPP (middle column), and
512 comparisons between RF-modeled and field-measured 8-day GPP (right column) for the plant functional
513 types of deciduous broadleaf forests (top row), evergreen needleleaf forests (the second row), grasslands (the
514 third row), and mixed forests (bottom row), respectively. Each subplot contains the entire time series data
515 from flux towers. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

Figure 15 compares the modeled and field-measured annual total GPP using all available site-year data. The regression analysis suggests that both SVR and RF outperform Biome-BGC. Compared with flux tower observations, results modeled by RF have high correlations (R = 0.805), low errors (RMSE = 327.835 $gC/m^2/year$), and low biases (Bias = -50.378 $gC/m^2/year$). The model performance of SVR is comparable to but slightly lower than that of RF. The correlation coefficient between Biome-BGC-modeled and field-measured annual total GPP is only 0.378 with the RMSE value of 837.7 $gC/m^2/year$.

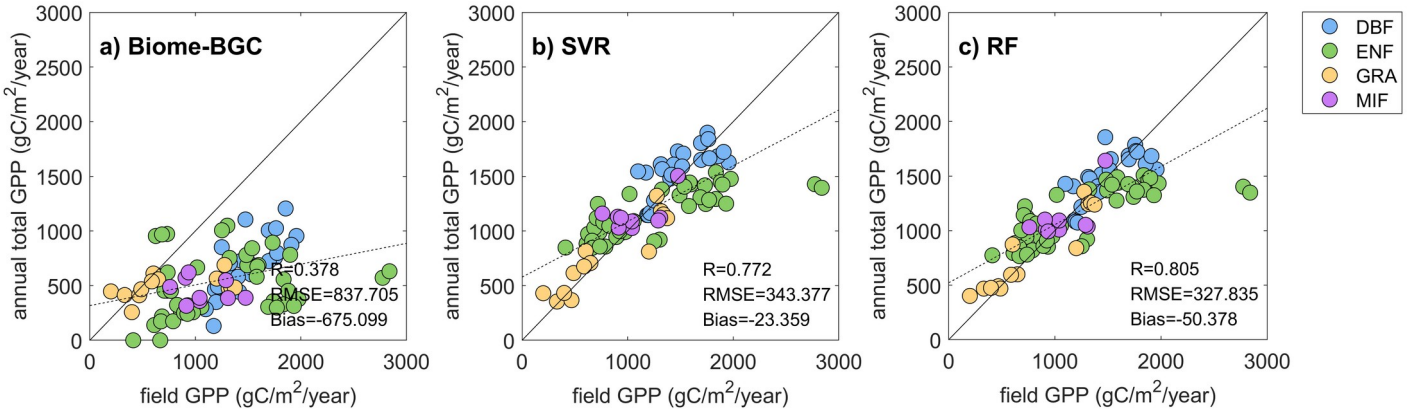


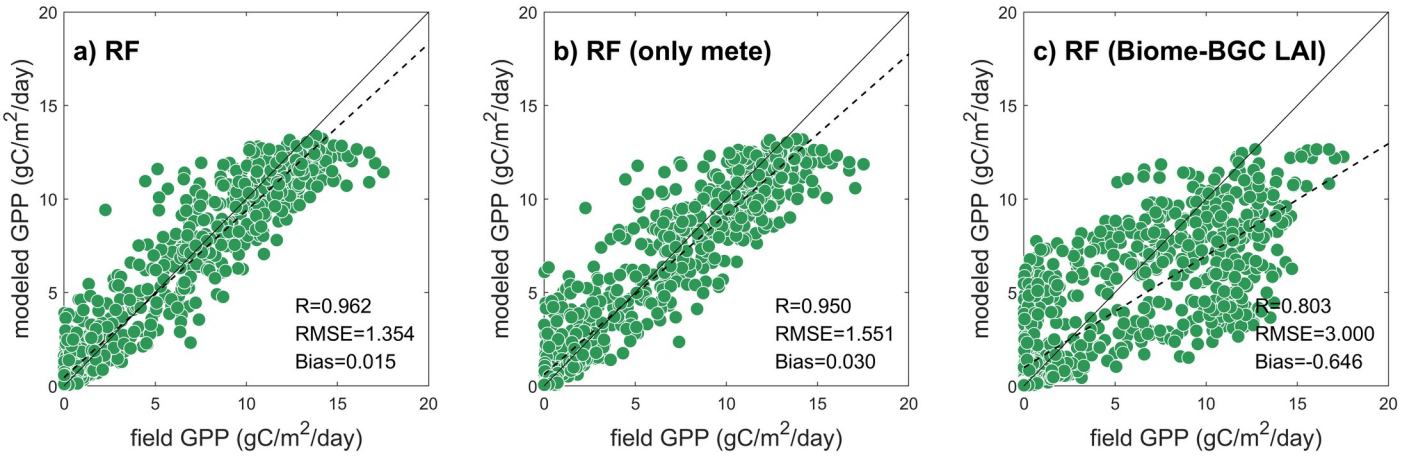
Fig. 15. Scatter plots are shown for comparisons a) between Biome-BGC-modeled and flux tower annual total GPP, b) between SVR-modeled and flux tower annual total GPP, and c) between RF-modeled and flux tower annual total GPP using all site-year flux tower data. The blue, green, orange, and purple points denote the DBF, ENF, GRA and MIF types, respectively. The solid lines denote the 1:1 lines and the dotted lines denote the regression lines.

38

5. Discussion

Existing studies have used the machine learning approaches to simulate GPP with remote sensing data, such as vegetation index, reflectance, and LAI, as input variables, and these studies are not capable of predicting vegetation GPP under future scenarios. This research explores the possibilities of using the machine learning models on predicting vegetation GPP with only meteorological variables as inputs. Our approach first simulates time series of vegetation LAI using meteorological variables and then uses the modeled time series of vegetation LAI as well as meteorological variables as model inputs to simulate time series of vegetation GPP. One question is how the modeling of the LAI time series affects subsequent GPP time series. In Figure 16b, we conduct comparative studies and develop the machine learning models that directly predict vegetation GPP via meteorological variables without predicting intermediate LAI variables. Compared with our developed approach (Fig. 16a), the machine learning models that directly predict vegetation GPP via meteorological variables have reduced correlation coefficient (R = 0.950) and increased errors (RMSE = 1.551 $gC/m^2/day$) as evaluated using the flux tower data. It implies that accurate modeling of the LAI time series helps improve the model ability on simulating GPP. Another implication is that canopy LAI tend to have a seasonal cycle similar to the carbon sequestration process. Fig. 16c evaluates the RF model that takes Biome-BGC-modeled LAI as inputs using flux tower data. It is found that the performance of the machine learning model on predicting GPP is dramatically reduced when using inaccurate LAI time series in the model, resulting in low correlation coefficient (R = 0.803) and high errors (RMSE = 3.000 $gC/m^2/day$) between modeled and observed GPP. It indicates that accurate simulation of LAI is crucial to the modeling of vegetation GPP.

39



551

552 Fig. 16. Scatter plots for the comparisons a) between RF-modeled GPP and flux tower GPP, b) between RF-
553 modeled GPP using meteorological data without predicting intermediate LAI variables and flux tower GPP,
554 and c) between RF-modeled GPP using Biome-BGC-modeled LAI as inputs and flux tower GPP. The solid
555 lines denote the 1:1 lines and the dotted lines denote the regression lines.
556

557 Note that all models, particularly the Biome-BGC model, have better performance on modeling GPP than LAI
558 when considering the metrics of correlation coefficients between modeled and observed data. The day-to-day
559 variation of vegetation LAI is much smaller than that of the meteorological variables and vegetation LAI has
560 lagged responses up to months to meteorological variation. By comparisons, vegetation GPP responds rapidly
561 to daily fluctuations in the meteorological conditions. It is therefore challenging to simulate LAI directly on
562 the daily or weekly basis using the machine learning approaches via meteorological variables. We hence adopt
563 the simple moving average method to buffer the impacts of day-to-day variation in meteorological variables
564 on modeling LAI. Biome-BGC first uses meteorological variables to predict the timing of vegetation key
565 phenophases such as the onset and offset of the growing season and then simulates the LAI time series and
566 canopy photosynthesis by modeling the allocation processes of chemical materials such as carbon and
567 nitrogen. Given LAI is a crucial factor that affects vegetation carbon fluxes, the Biome-BGC model still
568 requires substantial improvements on the phenology sub-models based on our tests. Our findings are in line
569 with [*A D Richardson et al.*, 2012] and suggest that better understanding on vegetation phenology is necessary
570 in future researches.

571

572 Note that the machine learning models are essentially data-driven approaches and it is difficult to interpret the
573 underlying physiological and ecological mechanisms in the models. Hence, the machine learning models
574 should not be viewed as surrogates to the process-based models. Instead, the machine learning models provide
575 benchmarking accuracies that we aim to achieve when developing the process-based models. The machine
576 learning approaches also provide intermediate solutions when our scientific knowledge on a particular process
577 like vegetation phenology is still limited. Another potential use of the machine learning models is to identify
578 the most relevant meteorological variables that affect vegetation phenology and provide references when
579 developing the process-based models.
580

41

6. Conclusions

Vegetation plays a key role in regulating the material and energy exchanges among the biosphere, the atmosphere, and the pedosphere. Predicting vegetation-related variables such as LAI and GPP under a changing climate is a core task in understanding the processes of vegetation growth and its responses to external environmental changes. While a number of existing studies has developed models to simulate vegetation GPP using satellite-derived LAI, the requirement of satellite-based model inputs largely limits the predicting power of these developed models. In this study, we attempt to develop the machine learning models, including both support vector regression (SVR) and random forests (RF), which are capable of modeling LAI and GPP time series using only meteorological variables. We tested our methods for four main plant functional types across North America. The results demonstrate that the machine learning models perform well on simulating the time series of both LAI and GPP. The spatial distributions of both LAI and GPP modeled using the developed machine learning models are consistent to those derived from satellite data. We also found that there is a need to improve the phenology representation in the Biome-BGC model for improving the modeling of vegetation LAI and GPP. Our modeling approaches provide an alternative way to predict time series of vegetation LAI and GPP using only meteorological variables across large geographic regions.

42

Acknowledgments

The authors declare no conflicts of interest. We thank the researchers and investigators who are involved in collecting and sharing the FLUXNET2015, Daymet and MODIS dataset. Biome-BGC version 4.2 was provided by Peter Thornton at the National Center for Atmospheric Research (NCAR), and by the Numerical Terradynamic Simulation Group (NTSG) at the University of Montana. This research is supported by National Key R&D Program of China (Grants 2017YFA0604300 and 2017YFA0604400), National Natural Science Foundation of China (Grant 41875122), Western Talents (Grant 2018XBYJRC004), Guangdong Top Young Talents (Grant 2017TQ04Z359), and One Hundred Talents Program of the Chinese Academy of Science (Grant Y674141001). We also thank anonymous reviewers for their constructive comments.

Data Availability Statement

The Daymet dataset proposed by [*P E Thornton et al.*, 2016] are available for download through the following link: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1328. The MODIS datasets are available from https://ladsweb.modaps.eosdis.nasa.gov/search/order. The GMTED data produced by [J J Danielson and D B Gesch, 2011] can be accessed via the website: https://yceo.yale.edu/gmted2010-global-multi-resolution-terrain-elevation-data. The GSDE data produced by [W Shangguan et al., 2014] is available for download through the following link: http://globalchange.bnu.edu.cn/research/soilw#download. The FLUXNET2015 Tier 1 FULLSET dataset are available via the website: https://fluxnet.org/data/fluxnet2015-dataset/. The SVR and RF model codes are available respectively from link: https://github.com/cjlin1/libsvm and https://code.google.com/archive/p/randomforest-matlab/. And we also thank [*M A White, Thornton, P. E. , Running, S. W. , & Nemani, R. R. . ,* 2000] for providing Biome-BGC model code through the site: https://github.com/bpbond/Biome-BGC.

43

621 **References**

622 Best, M. J., et al. (2011), The Joint UK Land Environment Simulator (JULES), model description – Part 1:
623 Energy and water fluxes, *Geosci. Model Dev.*, *4*(3), 677-699.

624 Chen, Wei, Xie, Xiaoshen, Wang, Jiale, Pradhan, Biswajeet, Hong, and Haoyuan (2017), A comparative study
625 of logistic model tree, random forest, and classification and regression tree models for spatial prediction of
626 landslide susceptibility, *Catena Giessen Then Amsterdam*.

627 Clark, D. B., et al. (2011), The Joint UK Land Environment Simulator (JULES), model description - Part 2:
628 Carbon fluxes and vegetation dynamics, *Geoscientific Model Development*, *4*, 701-722.

629 Collatz, G. J., J. T. Ball, C. Grivet, and J. A. Berry (1991), Physiological and environmental regulation of
630 stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer,
631 *Agricultural and Forest Meteorology*, *54*(2), 107-136.

632 Cristianini, N., and J. Shawe-Taylor (2000), *An introduction to support Vector Machines: and other kernel-*
633 *based learning methods*, Cambridge University Press.

634 Dai, Y., et al. (2003), The Common Land Model, *Bulletin of the American Meteorological Society*, *84*(8),
635 1013-1024.

636 Danielson, J. J., and D. B. Gesch (2011), Global multi-resolution terrain elevation data 2010 (GMTED2010),
637 *Report Rep. 2011-1073*.

638 Farquhar, G. D., S. von Caemmerer, and J. A. Berry (1980), A biochemical model of photosynthetic CO2
639 assimilation in leaves of C3 species, *Planta*, *149*(1), 78-90.

640 Field, C. B., J. T. Randerson, and C. M. Malmström (1995), Global net primary production: Combining
641 ecology and remote sensing, *Remote Sensing of Environment*, *51*(1), 74-88.

642 Foley, J. A., I. C. Prentice, N. Ramankutty, S. Levis, D. Pollard, S. Sitch, and A. Haxeltine (1996), An
643 integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics,
644 *Global Biogeochemical Cycles*, *10*(4), 603-628.

645 Grant, R. F., A. G. Barr, T. A. Black, H. Margolis, A. Dunn, J. Metsaranta, S. Wang, H. McCaughey, and C. A.
646 Bourque (2009), Interannual variation in net ecosystem productivity of Canadian forests as affected by
647 regional weather patterns-A Fluxnet-Canada synthesis, *Agricultural and Forest Meteorology*, *149*, 2022-2039.

648 He, M., et al. (2013), Development of a two-leaf light use efficiency model for improving the calculation of
649 terrestrial gross primary productivity, *Agricultural and Forest Meteorology*, *173*, 28-39.

650 Houborg, R., and M. McCabe (2018), A hybrid training approach for leaf area index estimation via Cubist and
651 random forests machine-learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, *135*, 173-188.

652 Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I.
653 C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere
654 system, *Global Biogeochemical Cycles*, *19*(1).

655 Mareike, L., B. Glaser, and B. Huwe (2012), Uncertainty in the spatial prediction of soil texture: Comparison

656 of regression tree and Random Forest models, *Geoderma*, *170*(none), 0-79.

657 McCabe, M. F., et al. (2017), The future of Earth observation in hydrology, *Hydrol. Earth Syst. Sci.*, *21*(7),
658 3879-3914.

659 Oleson, K. W., D. M. Lawrence, and G. B. Bonan (2013), Technical description of version 4.5 of the
660 Community Land Model (CLM). Ncar Tech. Note NCAR/TN-503+STR. National Center for Atmospheric
661 Research, Boulder.

662 Potter, C. S., J. T. Randerson, C. B. Field, P. A. Matson, P. M. Vitousek, H. A. Mooney, and S. A. Klooster
663 (1993), Terrestrial ecosystem production: A process model based on global satellite and surface data, *Global*
664 *Biogeochemical Cycles*, *7*(4), 811-841.

665 Richardson, A. D., R. S. Anderson, M. A. Arain, A. G. Barr, G. Bohrer, G. Chen, J. M. Chen, P. Ciais, K. J.
666 Davis, and A. R. Desai (2012), Terrestrial biosphere models need better representation of vegetation
667 phenology: results from the North American Carbon Program Site Synthesis, *Global Change Biology*, *18*(2),
668 566-584.

669 Rossel, R. A. V., and T. Behrens (2010), Using data mining to model and interpret soil diffuse reflectance
670 spectra, *Geoderma*, *158*(1), 46-54.

671 Running, S., P. Thornton, R. Nemani, and J. Glassy (2000), Global Terrestrial Gross and Net Primary
672 Productivity from the Earth Observing System, *Methods in ecosystem science*.

673 Schlund, M., V. Eyring, G. Camps-Valls, P. Friedlingstein, P. Gentine, and M. Reichstein (2020), Constraining
674 Uncertainty in Projected Gross Primary Production With Machine Learning, *Journal of Geophysical*
675 *Research: Biogeosciences*, *125*(11), e2019JG005619.

676 Sellers, P. J., C. J. Tucker, G. J. Collatz, S. O. Los, C. O. Justice, D. A. Dazlich, and D. A. Randall (1996), A
677 Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part II: The Generation of Global
678 Fields of Terrestrial Biophysical Parameters from Satellite Data, *Journal of Climate*, *9*(4), 706-737.

679 Sellers, P. J., D. A. Randall, G. J. Collatz, J. A. Berry, C. B. Field, D. A. Dazlich, C. Zhang, G. D. Collelo, and
680 L. Bounoua (1996), A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part I: Model
681 Formulation, *Journal of Climate*, *9*(4), 676-705.

682 Shangguan, W., Y. Dai, Q. Duan, B. Liu, and H. Yuan (2014), A global soil data set for earth system modeling,
683 *Journal of Advances in Modeling Earth Systems*, *6*.

684 Tatarinov, F. A., and E. Cienciala (2006), Application of BIOME-BGC model to managed forests: 1.
685 Sensitivity analysis, *Forest Ecology and Management*, *237*(1), 267-279.

686 Thornton, P., and B. Law (2002), Modeling the effects of disturbance history and climate on carbon and water
687 budgets in evergreen needleleaf forests, *AGU Fall Meeting Abstracts*, 03.

688 Thornton, P. E., M. M. Thornton, B. W. Mayer, Y. Wei, R. Devarakonda, R. S. Vose, and R. B. Cook (2016),
689 Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3, edited, ORNL Distributed
690 Active Archive Center.

691 Verrelst, J., G. Camps-Valls, J. Muñoz, J. Rivera Caicedo, F. Veroustraete, J. G. P. W. Clevers, and J. Moreno

(2015), Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties – A review, *ISPRS Journal of Photogrammetry and Remote Sensing*.

Wang, Q., M. Watanabe, and O. Zhu (2005), Simulation of water and carbon fluxes using BIOME-BGC model over crops in China, *Agricultural & Forest Meteorology*, *131*(3-4), 0-224.

Wang, T., Z. Xiao, and Z. Liu (2017), Performance Evaluation of Machine Learning Methods for Leaf Area Index Retrieval from Time-Series MODIS Reflectance Data, *Sensors*, *17*, 81.

White, M. A., Thornton, P. E. , Running, S. W. , & Nemani, R. R. . (2000), Parameterization and Sensitivity Analysis of the BIOME-BGC Terrestrial Ecosystem Model: Net Primary Production Controls, *Earth Interactions*, *4*(3), 1--84.

Yang, F., M. White, A. Michaelis, K. Ichii, H. Hashimoto, P. Votava, A. X. Zhu, and R. Nemani (2006), Prediction of Continental-Scale Evapotranspiration by Combining MODIS and AmeriFlux Data Through Support Vector Machine, *Geoscience and Remote Sensing, IEEE Transactions on*, *44*, 3452-3461.

Yang, F., K. Ichii, M. A. White, H. Hashimoto, A. R. Michaelis, P. Votava, A. X. Zhu, A. Huete, S. W. Running, and R. R. Nemani (2007), Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach, *Remote Sensing of Environment*, *110*(1), 109-122.

Yuan, W., et al. (2007), Deriving a light use efficiency model from eddy covariance flux data for predicting daily gross primary production across biomes, *Agricultural & Forest Meteorology*, *143*(3-4), 0-207.

Zhuang, Q., et al. (2011), Carbon cycling in extratropical terrestrial ecosystems of the Northern Hemisphere during the 20th century: a modeling analysis of the influences of soil thermal dynamics, *Tellus B: Chemical and Physical Meteorology*, *55*(3), 751-776.