Journal of Geophysical Research: Biogeosciences

Supporting Information for

Constraining uncertainty in projected gross primary production with machine learning

Manuel Schlund[1], Veronika Eyring[1,2], Gustau Camps-Valls[3], Pierre Friedlingstein[4,5], Pierre Gentine[6,7], and Markus Reichstein[8,9]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.

[2]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.

[3]Image Processing Laboratory (IPL), University of València, Valencia, Spain.

[4]University of Exeter, College of Engineering, Mathematics and Physical Sciences, Exeter, UK.

[5]LMD/IPSL, ENS, PSL Université, École Polytechnique, Institut Polytechnique de Paris, Sorbonne Université, CNRS, Paris, France.

[6]Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027.

[7]Earth Institute and Data Science Institute, Columbia University, New York, NY 10027.

[8]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany.

[9]Michael-Stifel-Center Jena for Data-driven and Simulation Science, Jena, Germany.

Contents of this file

Text S1. Data preprocessing

The raw monthly mean output of every participating climate model and observation-driven dataset is regridded to a 2° by 2° grid and masked with a common mask to remove all oceans and Antarctica. For Step 2a (target variable: absolute GPP at the end of the 21st century), monthly climatologies are calculated for every dataset by averaging over all available years for every month. For Step 2b (target variable: fractional GPP change over the 21st century), temporal means are calculated for every dataset by averaging over the full time dimension. In addition, values greater than 300% in the target variable in Step 2b (fractional GPP change over the 21st century) are masked to avoid numerical inconsistencies caused by the division of small numbers in the derivation of the target variable. In the next step, the multidimensional data is flattened and all the training data from the different climate models is stacked into a single large training array. Finally, to account for the varying magnitudes of the different features, all of them are linearly scaled by their respective means and standard deviations so that they have a mean of zero and unit variance. In total, 237'852 (16'503) training data points, 79'284 (5'501) hold-out test data points, and 46'344 (3'727) points for the prediction are used in the machine learning model for Step 2a (Step 2b).

Text S2. Gradient Boosted Regression Trees (GBRT)

The basic elements of GBRT are decision trees. These models create decision rules based on binary splits to predict a target variable y ("label") from a set of predictors $\boldsymbol{x} = (x^{(1)}, x^{(2)}, \dots, x^{(K)})$ (K is the number of "features"). These features do not need to be of the same type: GBRT allows the simultaneous input of numerical and categorical features, which is a great advantage for our use case. There is no need to encode the categorical variables in any way. Due to their simple nature, machine learning models based on decision trees are easy to interpret and explain but cannot be used to create satisfying predictions for complex datasets. This issue can be overcome by a technique called "boosting". Boosting improves the performance of "weak learners" (in our case decision trees) by combining a large number of them (Freund & Schapire, 1996). The regression function used to predict $\hat{y} = F(\boldsymbol{x})$ can be written as a linear combination of simple decision trees $h(\boldsymbol{x})$

$$F(\boldsymbol{x}) = \sum_{m=0}^{M} \beta_m h(\boldsymbol{x}; \alpha_m), \tag{1}$$

where M is the total number of decision trees, β_m expansion coefficients and α_m parameters of the trees. Using all N training data points $(\boldsymbol{x}_i, y_i)$ ("classic" gradient boosting), the expansion coefficients and parameters are jointly fitted by minimizing a loss function $L\big(y, F(\boldsymbol{x})\big)$ in a forward iteration:

$$(\beta_m, \alpha_m) = \underset{\beta,\alpha}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i; \alpha)) \tag{2}$$

In practice, this iteration step only uses a randomly selected subsample of the training data (drawn without replacement), i.e. the sum does not cover all N training points. Starting with an initial guess $F_0(\boldsymbol{x})$, the model is recursively built by

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \beta_m h(\boldsymbol{x}; \alpha_m). \tag{3}$$

The minimization procedure of the loss function (regular least squares function with additional sample weights determined by the grid cell areas) is called "stochastic gradient boosting" and is explained in detail by Friedman (2001); (Friedman, 2002). Fitting the GBRT model involves building the decision trees by splitting the data at points with maximum information gain. Boosting those simple trees greatly improves the overall predictive power of the machine learning algorithm: poorly modeled training points in the early stages of the algorithm will gradually improve throughout the training process.

A crucial criterion for the successful application of any GBRT algorithm is the choice of several hyperparameters. The three main control parameters of the learning procedure are the total number of decision trees M, the complexity of the individual trees (for example measured by the maximum tree depth) and the learning rate $\nu \ll 1$. The latter parameter is used for regularization and dramatically reduces the risk of overfitting by scaling down the contribution of each added weak learner (De'ath, 2007; Elith et al., 2008; Friedman, 2001). A common way to optimize the algorithm is K-fold cross-validation (Bishop, 2006): The data is randomly divided into a training and a validation dataset and the GBRT model is fitted on the training data only. After that, the performance of this model can be evaluated on the validation dataset by a suitable metric (e.g. the mean squared error). This process is repeated K times so that every input point is part of the validation set at least once. The optimal hyperparameters are the set of hyperparameters with optimal performance on the validation datasets (e.g. (Elith et al., 2008)).

Text S3. Evaluation of prediction uncertainty

We estimate the standard prediction error (SPE) of the GBRT model itself as the root mean squared error (RMSE) of the predicted $\hat{y}'$ and true values y' of a hold-out test dataset, the so-called root mean square error of prediction (RMSEP) (Bishop, 2006) (assumed to be constant for all prediction input points):

$$\sigma_{GBRT} = RMSE(\hat{y}', y') \quad (4)$$

For this, we randomly selected 25% of the input data prior to training, so that this part of the data neither enters the training of the GBRT model nor the hyperparameter optimization process. Moreover, the test dataset allows an assessment of the prediction residuals, which is useful to detect overfitting, see Figure S2.

A second source of uncertainty is the error in the re-scaling of the target variable in Step 1 of our approach. Analogous to Equation (2) in the main paper, we estimate this error as

$$\sigma_{j,RESC} = \bar{y}_j \cdot \frac{\sigma_{f'}}{\bar{f}}, \quad (5)$$

for each prediction input point j (j runs over all grid cells and months). $\bar{y}_j$ is the CMIP5 multi-model mean of the target variable (Step 2a: absolute GPP at the end of the 21st century; Step 2b: fractional GPP change over the 21st century), σ_f, the standard error in the global GPP fractional change given by the emergent constraint from Step 1 and $\bar{f}$ the CMIP5 multi-model mean global fractional change in GPP over the 21st century.

The final source of uncertainty is the error of the prediction input data $\sigma_{j,k}$ (k corresponds to the feature and j again to the prediction input point). These are only available for the FLUXNET-

MTE product (Jung et al., 2011). To account for this, we use the LIME technique (Ribeiro et al. (2016); see Section 2.2. in the main paper) to build a local linear model for every sample point, which yields the linear coefficients $b_{j,k}$. Using error propagation for all predictors and assuming independence of all individual errors, the SPE due to observational uncertainty $\sigma_{j,OBS}$ can then be calculated as

$$\sigma_{j,OBS}^2 = \sum_{k=1}^{F} b_{j,k}^2 \sigma_{j,k}^2. \quad (6)$$

The total SPE at a prediction input point j is the sum of the squared errors presented above (assuming all of them are independent):

$$\sigma_j^2 = \sigma_{GBRT}^2 + \sigma_{j,RESC}^2 + \sigma_{j,OBS}^2 \quad (7)$$

The specified error ranges for the multi-model mean approaches are calculated in a similar way: the constant SPE per grid cell is estimated by the mean RMSEP given in the pseudo-reality experiment. For the plain multi-model mean, this is the only source of uncertainty. For the re-scaled multi-model mean, the total error can be calculated similarly to Equation (7) without the last term (observational uncertainty).

Text S4. Evaluation of residuals

A convenient way to gain information about statistical models is to analyze the residuals ε_i which are defined as the difference between the true values of the target variable y_i and the predicted value $\hat{y}_i$ at a sample point i with known ground truth:

$$\varepsilon_i = y_i - \hat{y}_i \quad (8)$$

A common way to visualize the residuals is to plot their probability distribution (see Figure S2). The two panels (for Steps 2a and 2b) show that the machine learning model is not overfitting in both cases: the distributions of the training data and the independent hold-out data are very similar. Moreover, the distributions do not show significant biases, as the residuals are approximately unbiased (zero mean) for the training and the test dataset. This justifies the use of the RMSEP to estimate the SPE, since for unbiased residuals the RMSEP is equal to the standard deviation of the residuals.

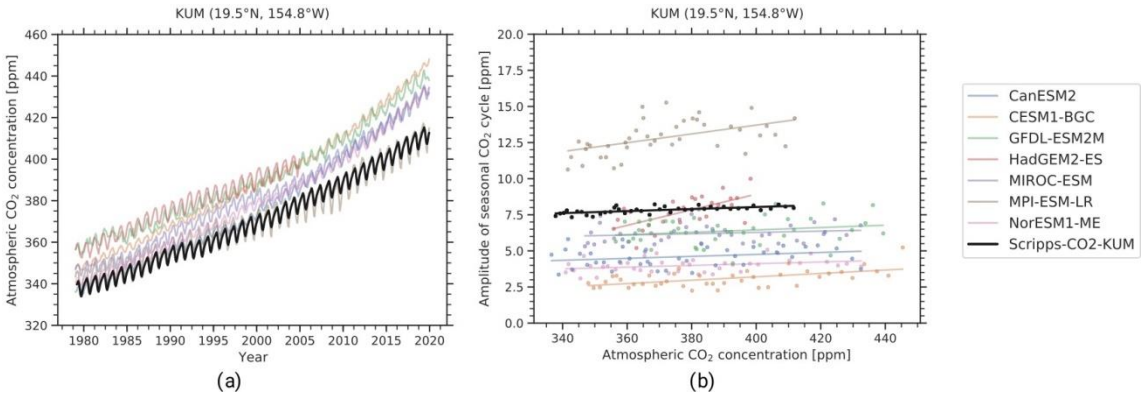

Figure S1. (a) Monthly-mean atmospheric CO_2 concentrations at Cape Kumukahi, Hawaii (KUM; 19.5 °N, 154.8 °W) from 1979 to 2019. The thin colored lines show the individual CMIP5 models (emission-driven historical simulations for the years 1979–2005 and emission-driven RCP 8.5 simulations for the years 2006–2019; the latter is not available for HadGEM2-ES). The thick black line shows the observations. For the CMIP5 models, the grid cell closest to KUM is considered. The curves show an increase of the atmospheric CO_2 concentration superimposed by a pronounced seasonal cycle. (b) Annual amplitude of the seasonal cycle of CO_2 (defined as the difference between the maximum and the minimum monthly mean atmospheric CO_2 concentration for each year) against the annual mean atmospheric CO_2 concentration at KUM. Colored dots show the CMIP5 models (similar time ranges as in (a)); thick black dots the observations. The lines show the corresponding linear regression fits for each dataset. The slopes of these linear fits define the sensitivity of the seasonal CO_2 cycle amplitude to atmospheric CO_2 concentrations, which is used as predictor for the emergent constraint step of our approach.

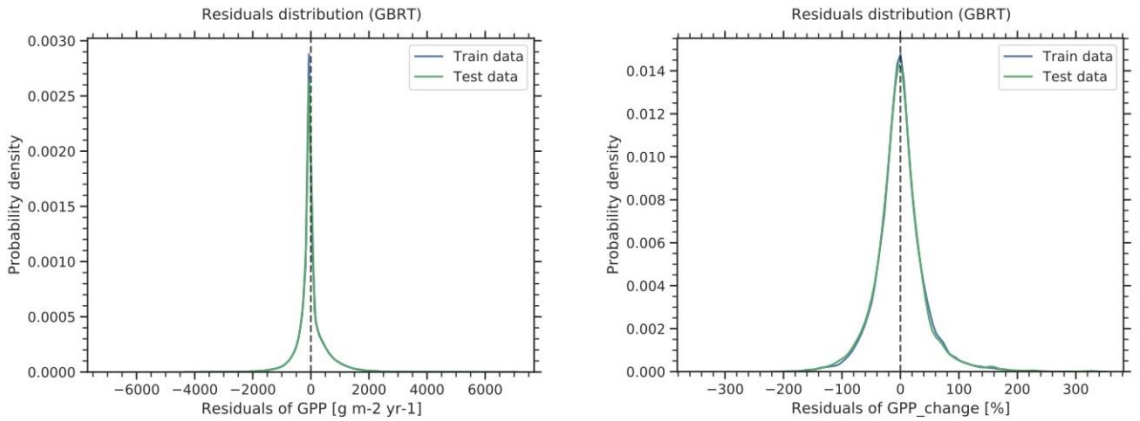

Figure S2. Distribution of the residuals for the two different target variables used in Step 2a (absolute GPP at the end of the 21st century) and Step 2b (fractional GPP change over the 21st century). The distributions are derived by Kernel Density Estimation (KDE) using training (blue) and test (green) data. The plots show approximately unbiased distributions for the training and the test datasets, which are very similar to each other. This indicates that the machine learning model does not overfit the data.

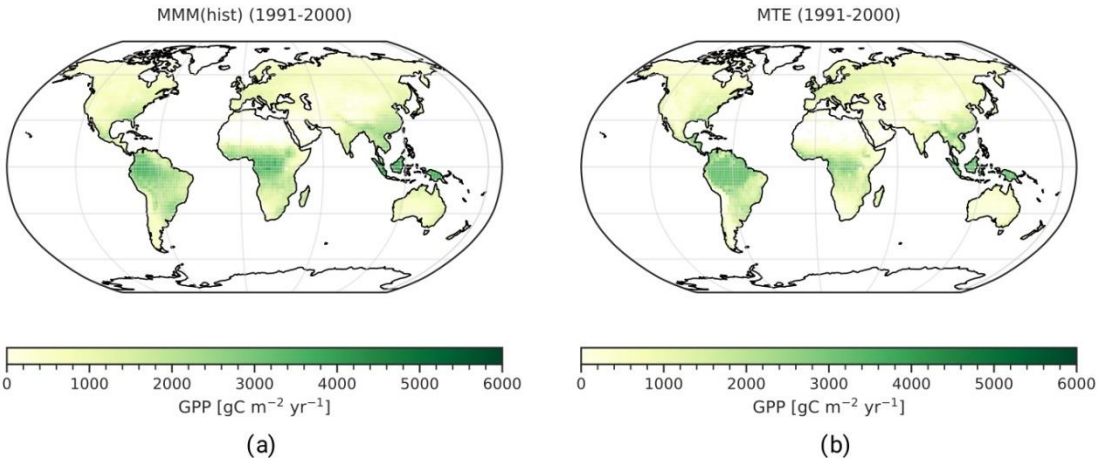

Figure S3. Geographical distributions of the historical GPP averaged between 1991 and 2000. (a) CMIP5 multi-model mean. (b) FLUXNET-MTE product (Jung et al., 2011).

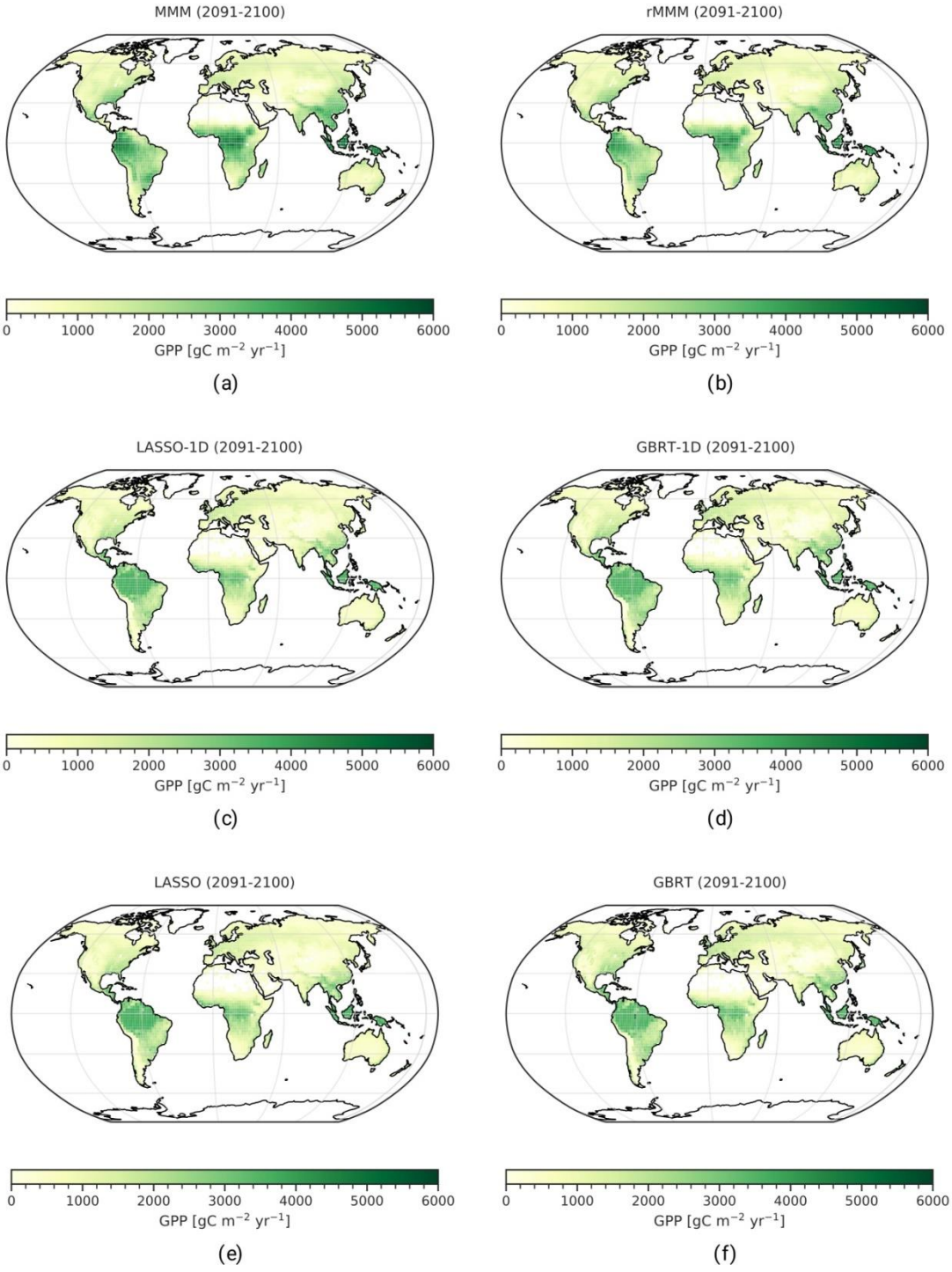

Figure S4. Geographical distributions of the absolute GPP at the end of the 21st century in the RCP 8.5 scenario (Step 2a) for different statistical models. (a) CMIP5 multi-model mean. (b) Re-scaled CMIP5 multi-model mean using Equation (2) from the main paper. (c) LASSO model using only the historical GPP as single predictor. (d) GBRT model using only the historical GPP as single predictor. (e) LASSO model using all predictors. (f) GBRT model using all predictors.

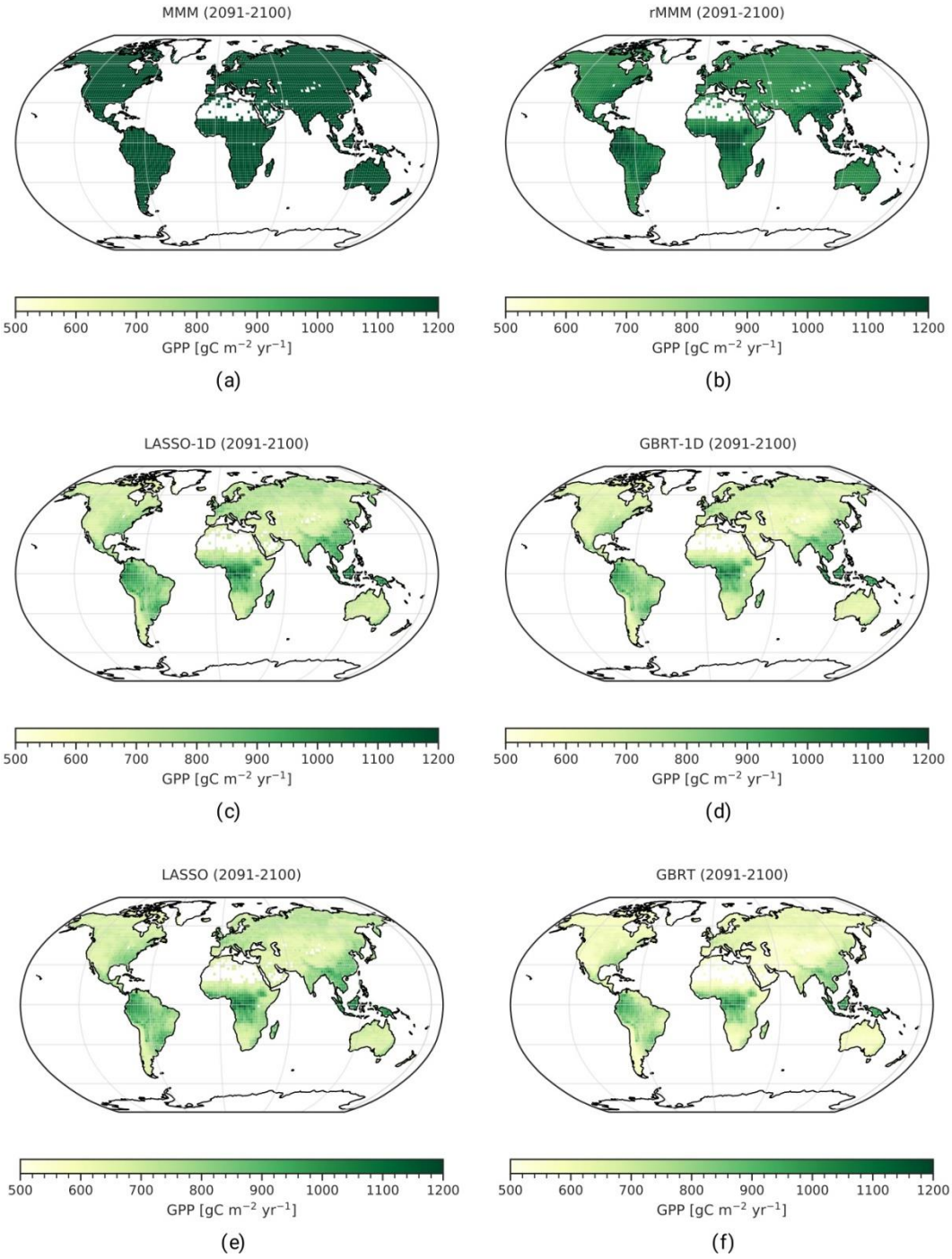

Figure S5. Geographical distributions of the standard prediction errors (SPEs) of the absolute GPP at the end of the 21st century in the RCP 8.5 scenario (Step 2a) for different statistical models. Details on the calculation of the SPE are given in Text S3. (a) CMIP5 multi-model mean. (b) Re-scaled CMIP5 multi-model mean using Equation (2) from the main paper. (c) LASSO model using only the historical GPP as single predictor. (d) GBRT model using only the historical GPP as single predictor. (e) LASSO model using all predictors. (f) GBRT model using all predictors. The SPE is minimal for the GBRT model using all predictors.

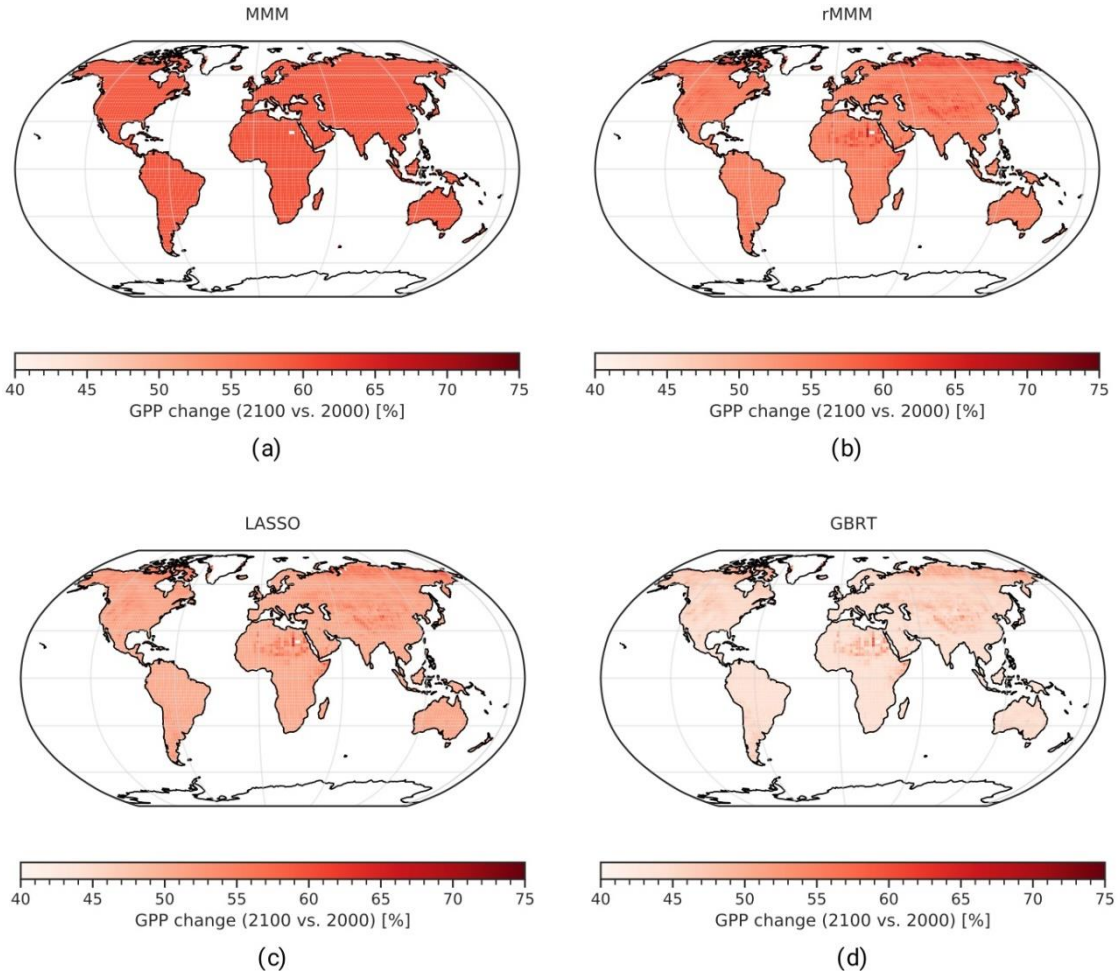

Figure S6. Geographical distributions of the standard prediction errors (SPEs) of the fractional GPP change over the 21st century in the RCP 8.5 scenario (Step 2b) for different statistical models. Details on the calculation of the SPE are given in Text S3. (a) CMIP5 multi-model mean. (b) Re-scaled CMIP5 multi-model mean using Equation (2) from the main paper. (c) LASSO model. (d) GBRT model. The SPE is minimal for the GBRT model.

Climate model	Land model	Main reference
CanESM2	CLASS2.7 + CTEM1	(Arora et al., 2011)
CESM1-BGC	CLM4	(Gent et al., 2011)
GFDL-ESM2M	LM3	(Dunne et al., 2012)
HadGEM2-ES	JULES + TRIFFID	(Collins et al., 2011)
MIROC-ESM	MATSIRO + SEIB-DGVM	(Watanabe et al., 2011)
MPI-ESM-LR	JSBACH + BETHY	(Giorgetta et al., 2013)
NorESM1-ME	CLM4	(Iversen et al., 2013)

Table S1. Overview over all seven CMIP5 models used in this study. More details are given by Anav et al. (2013). We chose all CMIP5 models which provide all necessary variables (*co2*, *gpp*, *lai*, *pr*, *rsds* and *tas*) for all used experiments (esmHistorical, *esmrcp85* and *esmFixClim1*). For all models, we only used the first ensemble member available.