Evaluating Vegetation Modeling in Earth System Models with Machine Learning Approaches

Ranjini Swaminathan[1], Tristan Quaife[1], and Richard Philip Allan[2]

[1]University of Reading
[2]Reading University

November 20, 2023

Abstract

Vegetation Gross Primary Productivity (GPP) is the single largest carbon flux of the terrestrial biosphere which, in turn, is responsible for sequestering $25-30\%$ of anthropogenic carbon dioxide emissions. The ability to model GPP is therefore critical for calculating carbon budgets as well as understanding climate feedbacks. Earth System Models (ESMs) have the capability to simulate GPP but vary greatly in their individual estimates, resulting in large uncertainties. We describe a Machine Learning (ML) approach to investigate two key factors responsible for differences in simulated GPP quantities from ESMs: the relative importance of different atmospheric drivers and differences in the representation of land surface processes. We describe the different steps in the development of our interpretable Machine Learning (ML) framework including the choice of algorithms, parameter tuning, training and evaluation. Our results show that ESMs largely agree on the physical climate drivers responsible for GPP as seen in the literature, for instance drought variables in the Mediterranean region or radiation and temperature in the Arctic region. However differences do exist since models don't necessarily agree on which individual variable is most relevant for GPP. We also explore a distance measure to attribute GPP differences to climate influences versus process differences and provide examples for where our methods work (South Asia, Mediterranean)and where they are inconclusive (Eastern North America).

Evaluating Vegetation Modeling in Earth System Models with Machine Learning Approaches

Ranjini Swaminathan[1,2], **Tristan Quaife**[1,2] **and Richard Allan**[1,2]

[1]University of Reading
[2]National Centre for Earth Observation

Key Points:

- A Machine Learning framework to advance our understanding of the terrestrial carbon cycle in Earth System Models or ESMs is proposed
- Differences in the relative importance of atmospheric drivers of gross primary productivity highlights differences across models
- A method to attribute differences in productivity estimates from ESMs due to process representation versus atmospheric forcing is demonstrated

Corresponding author: Ranjini Swaminathan, `r.swaminathan@reading.ac.uk`

Abstract

Vegetation Gross Primary Productivity (GPP) is the single largest carbon flux of the terrestrial biosphere which, in turn, is responsible for sequestering 25−30% of anthropogenic carbon dioxide emissions. The ability to model GPP is therefore critical for calculating carbon budgets as well as understanding climate feedbacks. Earth System Models (ESMs) have the capability to simulate GPP but vary greatly in their individual estimates, resulting in large uncertainties. We describe a Machine Learning (ML) approach to investigate two key factors responsible for differences in simulated GPP quantities from ESMs: the relative importance of different atmospheric drivers and differences in the representation of land surface processes. We describe the different steps in the development of our interpretable Machine Learning (ML) framework including the choice of algorithms, parameter tuning, training and evaluation. Our results show that ESMs largely agree on the physical climate drivers responsible for GPP as seen in the literature, for instance drought variables in the Mediterranean region or radiation and temperature in the Arctic region. However differences do exist since models don't necessarily agree on which individual variable is most relevant for GPP. We also explore a distance measure to attribute GPP differences to climate influences versus process differences and provide examples for where our methods work (South Asia, Mediterranean)and where they are inconclusive (Eastern North America).

Plain Language Summary

Gross Primary Productivity (GPP) is the rate at which plants remove carbon dioxide from the atmosphere during photosynthesis. Carbon dioxide is a greenhouse gas and excess in the atmosphere causes global warming and climate change. Changes in the amounts of atmospheric carbon dioxide will impact the entire Earth System. We therefore need the ability to accurately calculate GPP, especially for different possible carbon usage pathways in the future. Earth System Models or ESMs allow us to simulate various processes happening in the earth's atmosphere and biosphere including photosynthesis and can help us estimate GPP changes for such different pathways. However, ESMs can vary significantly in their simulated GPP estimates making it difficult to have confidence in using these estimates. We describe a Machine Learning (ML) framework to better understand where ESMs differ in calculating GPP so that we can address knowledge gaps in models. This approach allows us to understand the processes involved without having to run computationally expensive simulations. With improved models, we can also improve our ability to predict climate change outcomes for the future.

---

1 Introduction

Terrestrial Gross Primary Production (GPP) is the flux of carbon into the land surface driven by photosynthesis.

It is estimated that terrestrial GPP is in the order of $\sim 132PgC$ and it is the single largest annual flux of the global carbon cycle. It plays a key role in determining atmospheric carbon dioxide, since approximately a quarter to a third of anthropogenic emissions are sequestered by the land surface (on Climate Change, 2023; Schimel et al., 2001; Schwalm et al., 2020). GPP is influenced by natural climate variability as well as anthropogenic factors associated with global warming (Santini et al., 2014; Zampieri et al., 2021). Our ability to estimate GPP, its spatio-temporal patterns and the factors influencing GPP is therefore essential to understanding and forecasting global carbon budgets with greater reliability. GPP is not a directly measurable quantity at spatial scales of interest for carbon budget calculations (global or regional), so we rely on indirect measurements with inevitable assumptions, for example about the partitioning of fluxes at eddy covariance sites (Jung et al., 2019) or from satellite observations of quantities such as Solar Induced Fluorescense (SIF) (Sun et al., 2017; Y. Zhang et al., 2018), which are not direct measures of the carbon flux.

Earth System Models (ESMs) provide the capability to simulate GPP by modelling the various interactions between the atmosphere and biosphere including under different climate change scenarios in the future (Fisher et al., 2018; Levis, 2010). However, there is not only a large spread in GPP estimates from different ESMs but there are also large uncertainties in observational products that could be used to evaluate these estimates (Z. Wu et al., 2017; Anav et al., 2015). Therefore, there is a real need for evaluation methods that will help us understand better the possible reasons for such a large spread in GPP simulations, both in terms of the influence of atmospheric variables driving GPP as well as in the representation of the processes involved in simulating GPP. Identifying these differences can further help us address key gaps in modeling the terrestrial carbon cycle and will make for more reliable simulations from ESMs.

Machine Learning (ML) approaches have recently been used extensively in the study as well as generation of more accurate GPP data sets. Examples are seen work done in simulating GPP using observations of meteorological data or satellite data (Z. Zhang et al., 2021; Sarkar et al., 2022), upscaling GPP estimates from eddy covariance sites (Yu et al., 2021), to constrain uncertainty in GPP projections from models (Schlund et al., 2020) and for evaluating GPP representation in models (Z. Zhang et al., 2021; Dunkl et al., 2023). Our goal in this study is to use interpretable Machine Learning approaches (Molnar, 2020; Doshi-Velez & Kim, 2017) to better understand the sources of differences in GPP estimates between ESMs. Such an ML based evaluation framework can serve as a basis for process based improvements to ESMs, complementary to existing strategies, and can help reduce process uncertainty in modelled GPP estimates leading to more reliable simulations.

In previous studies, differences in GPP estimates from ESMs have been attributed to differences in the simulations of climate projections, modeling of complex terrestrial processes such as dynamic vegetation modeling, as well as atmospheric $\mathrm{CO_2}$ concentrations for given emission scenarios (Nishina et al., 2015; Schwalm et al., 2020; Fisher & Koven, 2020; Kim et al., 2018; Koch et al., 2021). In this work, we focus on two key attributes responsible for variability in GPP across ESMs - (a) the differences in climate simulations or input atmospheric forcing influencing GPP in individual models and (b) differences arising from vegetation process representation in these models. While we acknowledge that GPP is dependent on several land and atmospheric variables, in keeping with other similar studies such as Churkina and Running (1998); Schwalm et al. (2020); Anav et al. (2015), we evaluate the influence of three atmospheric variables as primary determinants of photosynthesis – precipitation, air temperature and downwelling shortwave radiation.

Our framework uses simulations from the CMIP pre-industrial Control (pi-Control) experiments that simulate climate before industrialization and the addition of anthropogenic CO_2 to the atmosphere. These simulations do not have the effects of elevated CO_2 that could lead to vegetation feedbacks or of any warming signal due to climate change. This allows us to better isolate the direct influence of the input climate variables on GPP without these factors. ESM simulations from pi-Control runs are also run for longer time periods, typically a few hundred years as opposed to a few decades from the historical experiment simulations and so this gives us a larger data set to learn from.

The methods used in this framework are based on Information Theory and Machine Learning, and compare the differences in input atmospheric forcings and vegetation process modeling associated with simulating GPP, across different ESMs from the Sixth Phase of the Coupled Model Intercomparison Project (CMIP6) (Eyring et al., 2016). These methods are directed towards formulating informed hypotheses for investigating the underlying factors influencing GPP estimates from ESMs. Specifically, the methods described target the following questions:

1. How do CMIP6 models differ in the input atmospheric forcings they consider most relevant for GPP? This will help us understand potential differences in how climate variables may influence GPP across models.
2. Can we compare differences in input forcings across ESMs with their process based differences? This will guide us towards attributing differences in GPP to the appropriate underlying factors.

We address the above questions by building ML based emulators of CMIP6 models that estimate GPP with input climate data. We query these emulators using robust Feature Selection methods to determine the relevance of individual atmospheric variables with respect to GPP. We also compare the differences in input forcing vs GPP by using a distance metric called the Jensen-Shannon distance measure. This is a novel approach that allows a comparison of two different attributory factors responsible for GPP and to the best of our knowledge is not previously seen in the literature.

We find that while the CMIP6 models considered largely agree on the variables considered relevant for GPP, there are regions of uncertainty such as the tropics. We are also able to show that models with similar input forcings do not always show similar estimates in GPP, indicating differences in process representation possibly due to parameterization. The remainder of the paper is organized as follows – Section 2 describes the ML framework including the parameter tuning process and algorithmic description of the learning and Feature Selection approaches. In Section 3, we discuss results where the ML framework identifies differences in climate variables influencing GPP across ESMs. In Section 4, we discuss the interpretability of the ML framework described, how this framework can be used for evaluation and some of the challenges involved. Finally we present our conclusions and planned future work using for this framework in 5.

2 Data and Methods

2.1 Data and Pre-processing

Our experimental experimental input data consists of five ESMs (UKESM1-0-LL, IPSl-CM6A-LR, CanESM5, CNRM-ESM2-1 and GISS-E2-1-G) from the CMIP6 project, all with different vegetation and land surface models as shown in Table2.1. The criteria applied for selection was to pick a small set of models with diversity in their vegetation modeling schemes, permitting exploration of various aspects of GPP simulation through our ML framework.

Seasonal means were calculated from monthly means of the data for two seasons, the boreal summer season of June-July-August (JJA) and austral summer season of December-January-February (DJF). All data considered is from the pre-industrial control (pi-Control) experiments which do not have an anthropogenic warming signal and for which a few hundred years of data are available from every model. Analysis is done for regions defined in the Intergovernmental Panel on Climate Change's Sixth Assessment Report (IPCC AR6), (Gutiérrez et al., 2021). Data was downloaded and pre-processed from the Earth System Grid Federation servers (Cinquini et al., 2014) using the open source evaluation tool, ESMValTool (Righi et al., 2020). We removed all non-land grid cells of a model in a selected region to focus on terrestrial GPP and then sampled data uniformly across time and space. Every grid cell and every time instance constitutes a sample data point and for each data point, we have one value each for the three atmospheric variables as well as for GPP. We then use this pre-processed data for further analysis. A pictorial description of our ML framework is shown in Figure 1.

Earth System Model	Land Surface Model	Reference	Dynamic Vegetation
UKESM1-0-LL	Joint UK Land Environment Simulator (JULES)	(Sellar et al., 2019; Clark et al., 2011)	Yes
IPSL-CM6A-LR	Organising Carbon and Hydrology In Dynamic Ecosystems (ORCHIDEE)	(Boucher et al., 2020; Krinner et al., 2005)	No
CanESM5	The Canadian Land Surface Scheme (CLASS)	(Swart et al., 2019; Verseghy, 2012)	No
CNRM-ESM2-1	Interaction Soil-Biosphere-Atmosphere (ISBA)	(Séférian et al., 2019; Delire et al., 2020)	No
GISS-E2-1-G	ENT Terrestrial Biosphere Model	(Kelley et al., 2020; Kiang, 2012)	No

Table 1. The CMIP6 models evaluated with our framework and their corresponding vegetation models. Data on dynamicity of vegetation obtained from the Earth System Documentation Project (Greenslade et al., 2014) and (Zarakas et al., 2020)

.

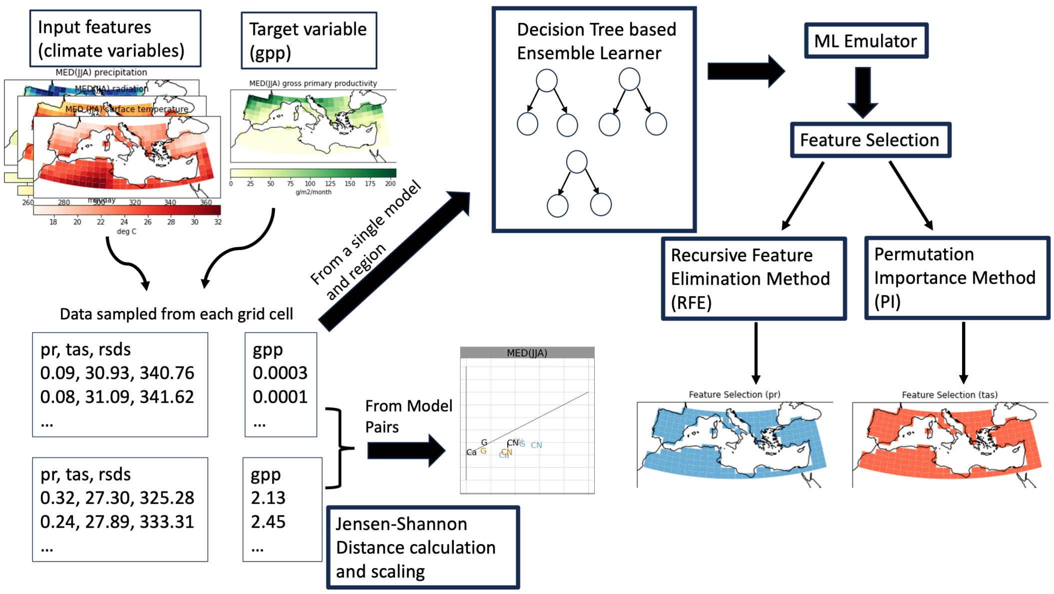


Figure 1. A description of the Machine Learning framework for evaluating GPP in CMIP6 models: Data from atmospheric variables and GPP for a given region, season and ESM is used to train an ensemble learner which serves as the ML emulator. The ML emulator is then queried using two different Feature Ranking algorithms (RFE or Recursive feature Elimination and PI or Permutation Importance) to find the most relevant features or atmospheric variables for GPP in that region. Data from pairs of ESMs is also used to calculate the Jensen-Shannon Distance (JSD) metric to compare distances measured in the input variable space with distances measured in the GPP distributions across regions.

2.2 ML Emulators with Ensemble Learning

Our requirement for an ML based emulator was one that would effectively model the relationship between input atmospheric forcing variables (and any other similar GPP influencing variables to be included as needed) and GPP; and one that would allow us to interpret or make inferences on the modeled relationships to answer questions on the relative importance or sensitivity to the climate variables. An additional goal was to develop a flexible framework that could be applied to observed data to better facilitate model evaluation. For this reason, we designed the core of the emulator to be a multivariate regression model and one that can be interpreted or queried on the decisions made for regression. In this, the climate forcing variables are the input features or predictors and GPP is the predictand. The ML emulator is trained for every region, season and ESM in our experimental setup. We use a regression model with Boosting called Adaptive Boosting or AdaBoost (Mendes-Moreira et al., 2012; Schapire, 2013) for our framework. Boosting is a well established ML approach that works towards developing a highly accurate prediction rule by repeatedly combining several weaker predictors or learners (Drucker, 1997) which in this case would be regressors. In Boosting, the first weak predictor is trained with a subset of samples uniformly sampled from the training data set with replacement permitted, meaning a training sample can be used again to build a different predictor. Once a predictor is built, all the training samples are passed through the predictor and the samples with the largest prediction errors are identified. The sampling probabilities of the samples with the most error are adjusted so that they are more likely to get picked as training samples for the next weak learner to be built. As this process repeats, harder to learn patterns get picked more often to build subsequent predictors. This means that some predictors will do better than others in a given subspace of the input feature space. The predictors are further assigned weights of the form, $\bar{\beta} = \frac{\bar{L}}{1-\bar{L}}$ where $\bar{L}$ is a calculated loss function. Cumulative predictions are calculated as a weighted median of all the predictors. The algorithm terminates when the average loss across all weak learners is below a certain threshold. The weak learners or regressors in this boosting algorithm can be any one of a wide array of regression methods. We calculated the Root Mean Square Error scores on held out test data sets and determined that the Decision Tree algorithm described in Breiman et al. (1984); Quinlan (1986); Breiman (1996) was best suited for our task after experimenting with different ML regression algorithms such as Linear Regression (James et al., 2021) and Support Vector Machines (Smola & Schölkopf, 2004). We therefore use an Ensemble Tree Learner with Boosting for our ML emulators.

As shown in Fig 1, CMIP6 data in the form of gridded data sets was used to train the ML emulators by treating each grid cell at every time step as an individual sample for learning. However, ESMs differ in grid resolution and in the length or number of years of the pi-Control experiment runs. So, for a given region, the number of training samples can be different across ESMs. In order to avoid biases resulting from differences in the number of samples, we randomly sampled a minimal sample set from every model such that the number of samples to train an emulator is the same across all ESMs. This sample set is then used to tune the parameters and build the Decision Trees in the ML emulator.

2.3 Parameter Tuning

In applied Machine Learning, parameter tuning is considered an important step in developing ML models that best capture patterns in the training data without overfitting (Yang & Shami, 2020). Overfitting occurs when we train the ML model to fit the training data too well which could result in a loss of generality. In other words, the ML model performs exceedingly well on the data it is trained with but fails to perform well on a new test set of samples even if from the same or similar distribution. We employ the Adaboost algorithm with an ensemble of Decision Tree regressors from the open source Python Scikit-learn package (Pedregosa et al., 2011) to build our ML emulators. A built in mechanism for pruning the ensemble learner exists for removing learners in a way that diversity is maximized. This essentially means that learners are selected such that a wide range of associations or rules are learnt and duplication of rules learnt is minimized by pruning. This helps to avoid overfitting by balancing the need to add more rules in the predictor with the ability to generalize well. In our experiments we tune for the depth parameter in the Decision Tree for optimal performance of the emulator, determined as the best fit to the data as evaluated by the Root Mean Squared Error (RMSE) in the predictions. The depth of the Decision Tree is the number of levels at which decision nodes are split in the tree. For example, a decision could be $tas > 20$ which would split training samples into those where the surface temperature is greater than 20°C (condition is true) and those where the temperature is less than 20°C (condition is false) and so on. For every region-season-ESM combination, we split the samples available into a training set and a held out test set. The ML emulator (AdaBoost with Decision Tree regressor) is learnt using the training samples and tested on the held out samples. RMSE scores are calculated for both training and held out test sets. For a given value of the depth parameter, this process is repeated by splitting the data n times and the average training and test RMSE scores over the n splits is calculated. This is how *n-fold* cross-validation (where n=*6* in this case) is performed. The depth parameter that has the lowest RMSE score on the held out test data, with cross–validation is then chosen as the most optimal parameter for the task and a final ML emulator is built using that depth parameter and all the samples available for that region. This builds robustness against overfitting, and sampling multiple times during cross validation further makes the model more reliable ensuring that the final emulator has seen a good representation of the available data. ML emulator estimates of GPP for a selection of regions are shown as an illustration of the results from this process in Supplementary Figure S1.

2.4 Feature Selection Methods

After the ML emulators were constructed to specification and sufficiently satisfied requirements, meaning the final emulator had the lowest possible RMSE scores for held out test data in cross validation experiments as described, we focused on querying or interpreting these emulators to better understand the relationship between the different input climate variables and GPP. Feature Selection or Feature Importance Ranking is the process of selecting or ranking features (input variables or predictors) that are most relevant to the predictand as evaluated by some chosen measurement or metric (Kumar & Minz, 2014; Guyon & Elisseeff, 2003). It is a process that is often used to prune the number of input features required for accurate predictions but in our case, with just three features, we use feature ranks to find the input atmospheric forcing variable(s) that the ML emulators find most important for GPP. Two different feature selection methods were applied to the ML emulators - (a) Recursive Feature Elimination (RFE) and (b) Permutation Importance (PI). The two methods use slightly different criteria to evaluate feature importances as described below but both provide useful information regarding the relative importance of a climate variable for GPP and are complementary. In the Recursive Feature Elimination algorithm, the input features are recursively removed one at a time to find the feature that has the most influence on the predictand (Guyon et al., 2002). For our experiments, we used the RMSE values to quantify the influence of an input climate variable on GPP. So, if the RFE method determines precipitation to be the most important feature for GPP, this effectively means that removing precipitation from the set of input features would have the most impact on the emulator's ability to predict GPP well i.e increase the RMSE by the most compared to other variables. In the Permutation Importance method, the decrease in model score when an individual feature is randomly shuffled or permutated is the measure of how important that feature is to the emulator (Breiman, 2001). The model score here is the Regression coefficient of determination (R^2) and is a measure of how well the ML emulator fits the data. Thus, the PI method works well once a reliable ML emulator is developed and is a measure of sensitivity of GPP to an input variable given such an emulator. As in the case of developing the ML emulator, we performed 6-fold cross-validation for the feature selection process as well. We did this by devising a simple voting scheme with small differences based on the Feature Selection approach. In the case of the RFE method, we assigned a single vote to the feature(s) that was ranked highest in terms of influencing the prediction with the RMSE score. We then averaged the votes across all the input features to determine the actual ranks of these features. In the PI method, we calculated the contribution of each feature to the R^2 score (permutation importances) and granted a vote to an input feature if it contributed to more than half of the score, which is the fit of the model. As in the RFE method, the votes were once again averaged across the cross-validation subsets. This scheme allowed us to account for collinearity or multiple variables equally influencing GPP especially as these are physical climate variables which are very closely related to each other.

2.5 Distance measure for climate and GPP distribution comparisons

While the ML emulators and Feature Selection are used to understand differences in models, we also calculate using a relative measure, how close or similar models are in the input forcing space vs. how similar they are in their simulated GPP distributions. Essentially we evaluate whether models that are similar in input atmospheric forcing simulated by the ESMs are also similar in their GPP simulations. If we consider that every data sample is represented as an instance in a 3-Dimensional input climate parameter space, where each dimension corresponds to a climate feature, then for a given region-season-ESM, we have a distribution of these 3-Dimensional data points. A distance metric is applied to quantify how close climate distributions from two different ESMs are for a given region and season. The same distance metric is now used to measure similarity between the GPP distributions of models in the 1-Dimensional space of GPP values. The distance metric we use is the Jensen-Shannon distance, which is calculated as the square root of the Jensen-Shannon divergence between two distributions (Lin, 1991). This is a symmetric and smoothed version of the more commonly used Kullback-Divergence measure. This measure has been widely used in applications such as evaluating generative adversarial networks by measuring differences in distributions (Goodfellow et al., 2020), text classification with high dimensional feature sets (Dhillon et al., 2003) and in bioinformatics for mutation detection (Gültas et al., 2014). The Jensen Shannon Divergence itself is defined as :

$$JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M), M = \frac{1}{2}(P+Q), \tag{1}$$

where $D(P||Q)$ is the Kullback-Divergence (Csiszár, 1975) between two distributions P and Q. When a base-2 logarithm is used, the Jensen-Shannon divergence has an upper bound of 1 i.e, $0 \leq JSD(P\|Q) \leq 1$. The existence of upper and lower bounds and the fact that distances are symmetric, are important properties we take advantage of when comparing ESMs. We refer to JSD as the Jensen-Shannon Distance instead of divergence as they both hold the same meaning for our analysis. Using the JSD, we compare how much ESMs differ in their input forcing vs in the simulated GPP for a region and season. A JSD of 0 implies the distributions are identical and as the JSD increases going towards 1, it implies that distributions get more dissimilar. While it is not possible to directly compare distance values between pairs of ESMs across two different distribution spaces (as in the 3-D climate space and the 1-D GPP space), we can compare how ESM-pair distances are ordered in both distribution spaces. That is we can see how distances between pairs of models compare in the two different spaces. We further apply a simple scaling by a factor of the shortest distance among all pairs of models in the in-

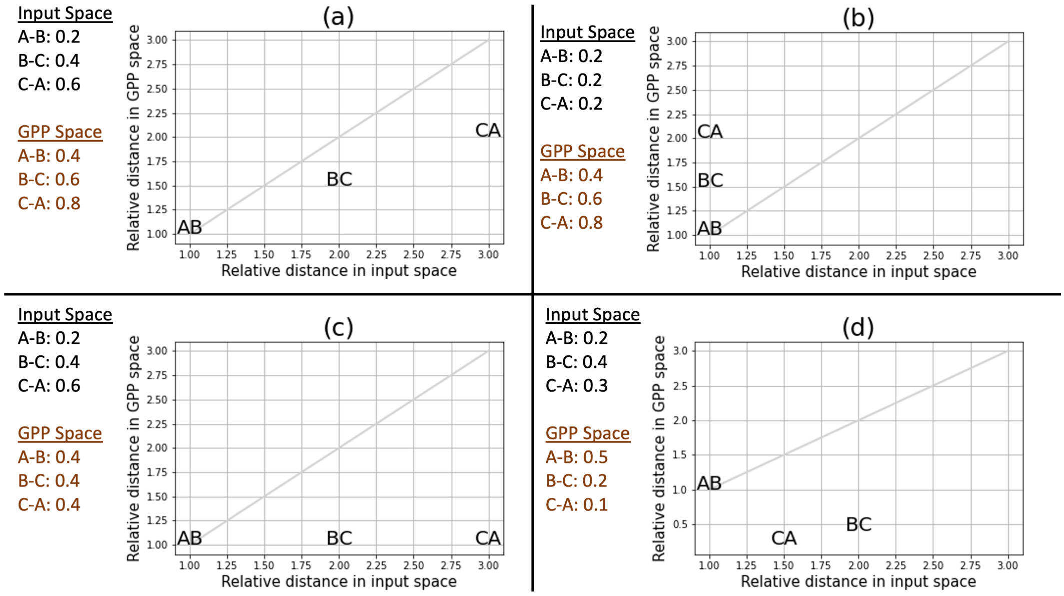


Figure 2. An illustration of how the Jensen Shannon distance metric is used to understand differences in input space (atmospheric forcings) and GPP space. In subplot (a) of the figure, we can make the inference that similarities in input forcing are consistent with similarities in GPP . Where that does not hold, we can start to explore the possibility that there might be larger differences in process representation or parameterization between pairs of ESMs which results in this difference in GPP as seen in subplots (b) and (c) and in the case of model pair A-B in (d). Thus the JSD scaled in this manner gives us a way to actually compare the differences in input forcings of ESMs relative to their simulated GPP.

put space so we can effectively make inferences about whether relative orderings in input climate variable space are reflected in the GPP space as well.

We illustrate analysis based on the JSD in Figure 2 with four different possible use cases and how inferences can be made from them. Each sub figure shows the actual JSD in input (on the x-axis) and GPP (y-axis) space between three hypothetical models - A, B and C. The distances are then scaled by dividing all the distances in input space by the smallest such distance among all pairs of models. The distance in GPP space between that same pair of models is then used to scale all model pair distances in GPP space. This scaling allows us to effectively compare distances in input space vs GPP space. In subplot (a), we see that the relative ordering of distances between pairs of models is the same on both axes, the model pair A-B has the smallest distance in input space as well as GPP space while the model pair C-A has the largest distance in both these spaces. This provides some evidence that similarities or differences between pairs of models in the atmospheric forcing is also reflected in their GPP simulations. In (b), the distances in the atmospheric forcing are the same for all pairs of models but that's not the case in their GPP simulations where the distance between C-A is larger than the other pairs indicating possible differences in process representation across the models. In (c), the model pairs show larger differences in their input forcing but not in the simulated GPP space, indicating that despite having different climate, the models end up simulating very similar GPP values potentially differing in the processes involved in calculating GPP from these climate variables. Finally, in (d) we see another example for where proximity in input forcing does not translate to similar GPP simulations. In model pair A-B, differences lie more in simulated GPP than in the atmospheric forcing while the opposite is the case for model pairs C-A and B-C. We can thus use this analysis to attribute reasons for differences in GPP simulations between pairs of models.

The JSD measure was also used to determine how well the ML emulators estimate GPP by comparing the emulator estimated values with ESM simulations and we found that these distances tended to zero (results not shown). This further gives us confidence in our deployment of these ML emulators.

The ML emulators with Feature Selection, Jensen-Shannon Distance metric comparisons and more traditional analysis involving univariate statistics are all combined in our analysis of differences across ESMs in how they simulate GPP. Results from the analysis and a discussion on where the ML methods work well and where they don't is discussed in the next sections.

3 Results

In this section, we look at two key sets of results coming from the ML framework proposed in section 2.4. We first look at regional feature importances, that is, what the ML emulators determine to be the most relevant climate variable for GPP in a given region. We discuss results for regions in the JJA and DJF seasons as seen in Figures 3 and 4 but also provide results from the annual mean analysis for a more general overview in Supplementary Figure S2. We study the differences and similarities in GPP representation across pi-Control simulations in ESMs but due to the lack of observational datasets for this period, we use the literature on historical observations to guide our evaluation.

Our second set of results is from the comparison of relative distances between ESMs in the input climate space vs the GPP distribution space as described in Subsection 2.5 and shown in Figure 5. In our current analysis, we provide examples for how the JSD based comparisons can be useful as a tool to identify potential sources of differences in ESMs but leave more detailed region by region analysis for future work.

3.1 Model differences in relevant climate variables for GPP

Figures 3 and 4 show the most relevant climate variables for predicting GPP from two feature selection methods – Recursive Feature Elimination (RFE) and Permutation Importance (PI) in the first and second columns respectively. The RFE method's selection of best feature is considered the most relevant variable for GPP by the ML emulator and means that this variable is primarily responsible for estimating GPP. The PI method's selection on the other hand is more a measure of GPP's sensitivity to climate variables given the ML emulator. The most important climate variable could also be the variable GPP is most sensitive to, as in both methods could agree on the choice of climate variable(s) but differences are possible since the metrics involved are slightly different (low error vs best fit). ESM differences in the top features from the methods are considered an appropriate potential starting point for investigating divergence in GPP estimates from ESMs. We refer to the regions by their acronyms as defined in Iturbide et al. (2022) and are shown in Supplementary Figure S3 for reference.

Overall, all ESMs considered agree that temperature followed by precipitation are key variables for GPP for most of Europe, N.America and Asia. Over Africa and S.America, there is less of a consensus across ESMs and methods in accordance with previous analysis (Churkina & Running, 1998). Temperature is considered the most important variable for GPP in the Russian-Arctic (RAR) and Northern Europe (NEU) regions in JJA for most ESMs. Conditions of almost constant sunlight and water availability make temperature the key driver for GPP here. The northern N.American regions are a combination of arctic tundra and boreal forests and similarly show temperature as the main

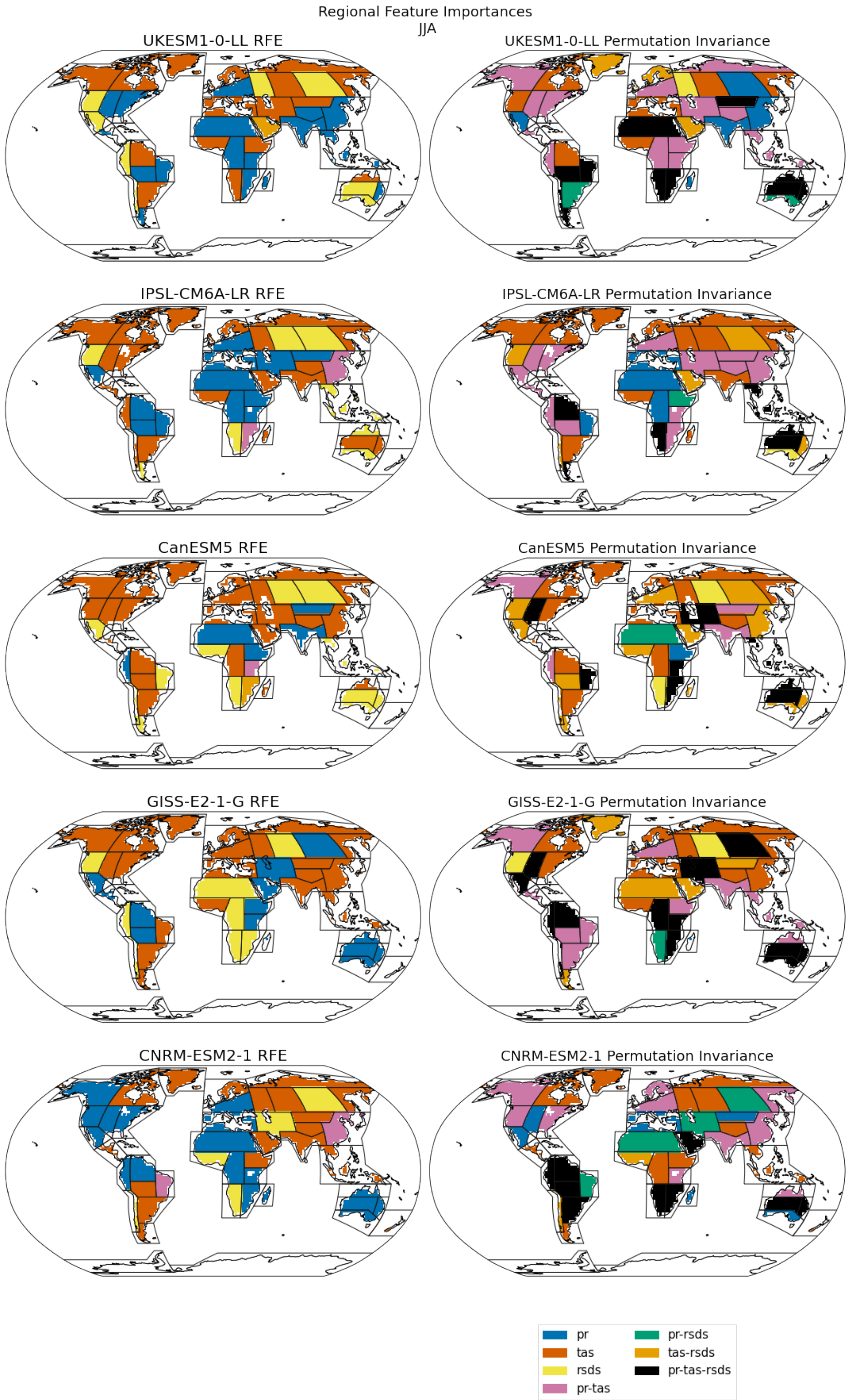


Figure 3. JJA feature importance from two methods - Recursive Feature elimination and Permutation Invariance for the IPCC regions defined in Iturbide et al. (2022).

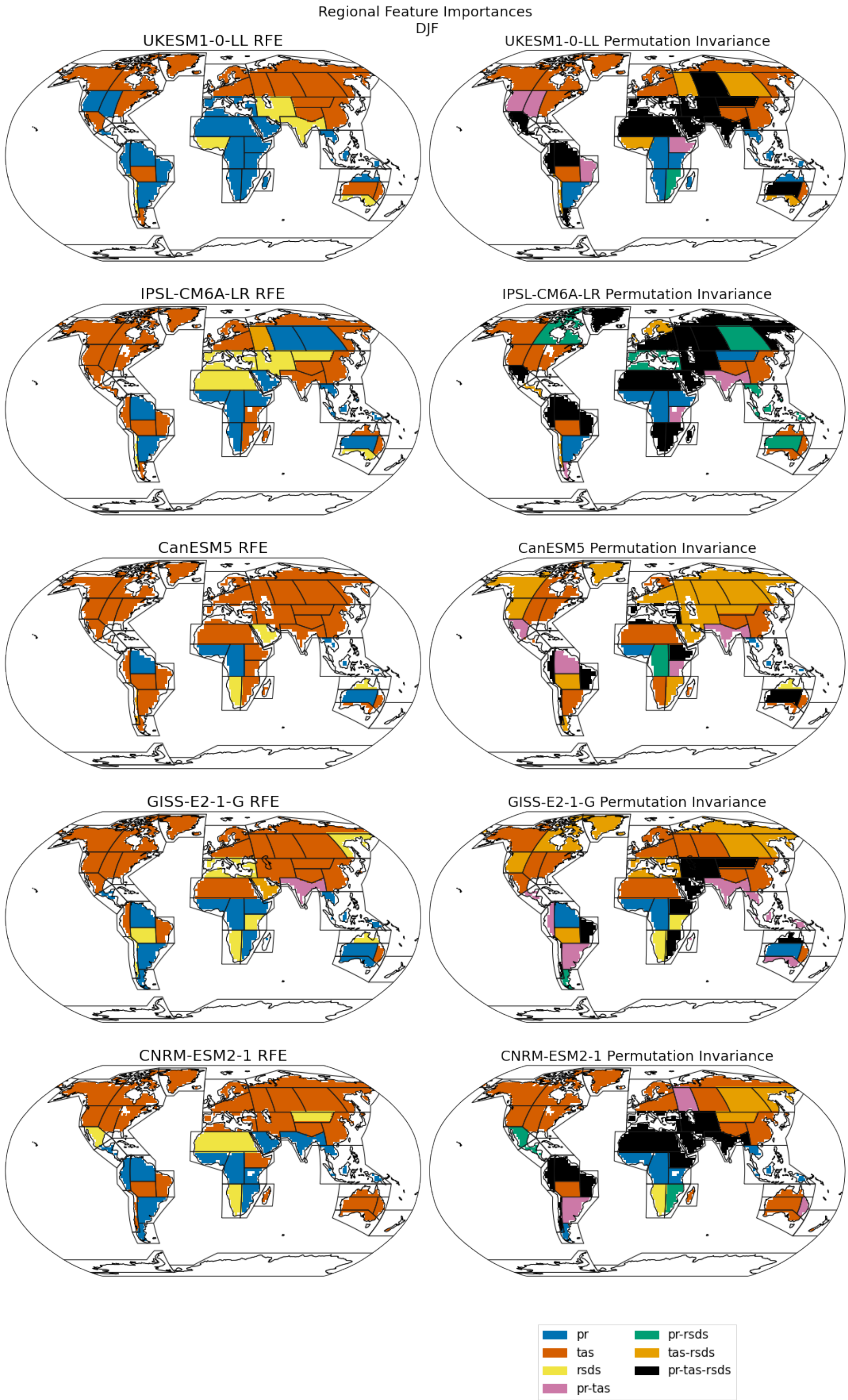


Figure 4. DJF feature importance from two methods - Recursive Feature elimination and Permutation Invariance for IPCC regions defined in Iturbide et al. (2022)

driving factor except for Northwestern North America (NWN) in CNRM-ESM2-1 where precipitation is determined as the key driver.

Boreal forest regions such as Eastern Europe (EEU), Western and Eastern Siberia (WSB, ESB) and the Russian Far East (RFE) show more divergence across ESMs with GPP being more dependent in both RFE and PI methods on temperature or radiation or both but in some instances (ESB for GISS-E2-1-G) on precipitation. In the central and eastern continental United States (CNA, ENA), UKESM1-0-LL and CNRM-ESM2-1 models consider precipitation to be most relevant for GPP while all other models find temperature more relevant. The variability in GPP is also dominated by a combination of these two variables as seen in the PI method. In the western north American region (WNA), radiation is seen as driving GPP except in CanESM5 (temperature) and CNRM-ESM2-1 (precipitation). In fact, precipitation seems to be most relevant for GPP in almost all N.American regions in the CNRM-ESM2-1 model and this can be considered as an indication that either the availability or the parameterization of this variable is important for GPP in this model more so than in others.

All ESMs in our study agree precipitation and temperature play a more important role than radiation in the Mediterranean region (MED), where radiation is largely available and a lack of rainfall or very high temperature is likely to influence vegetation more (Gea-Izquierdo et al., 2015). The CNRM1-ESM2-1 and IPSL-CM6A-LR are the two models that rank precipitation higher than temperature as an important feature. For the region covering the Indian subcontinent (SAS), precipitation is considered most important in the UKESM1-0-LL and CanESM5 models, consistent with previous studies (Varghese & Behera, 2019; Verma et al., 2022) while all three other models favor temperature as the key factor. In East Asia (EAS) temperature is considered the most important driver for GPP followed by precipitation and radiation in some regions (Yao et al., 2018; Bo et al., 2022) and all models except UKESM1-0-LL (precipitation) are in agreement.

In the DJF season, all models except CanESM5 consider precipitation most relevant for GPP in South East South America (SES) and all models agree that temperature is most relevant for Eastern Australia (EAU). We find the largest source of disagreement with regards to GPP drivers (looking at both DJF and JJA seasons) in regions where there is a significant presence of tropical forests such as Northern South America (NSA), Central-Africa (CAF), South-East Asia (SEA) and Northern Australia (NAU). We note radiation plays a role in some regions, possibly due to the lack of sufficient radiative energy available due to cloud cover which makes it hard to distinguish the relative importance between features. However almost all ESMs over a majority of these regions reference temperature and precipitation as key variables and from observational records we know that the two variables are strongly correlated in these regions (Nzabarinda et al., 2021; F. Zhang et al., 2022; Kanniah et al., 2011). Although precipitation appears most frequently as as the most important variable in determining GPP, especially using the RFE method of feature selection, in more than one instance all three features are considered relevant. This is consistent with results from previous studies using observations and non-ML approaches applied to finding GPP drivers (Churkina & Running, 1998; Kanniah et al., 2013; D. Wu et al., 2014). Another area where models lack consensus over the drivers is Southern Africa (ESAF and WSAF) for the DJF season. In reality, these areas are dominated by savannah, and are likely water limited but this is seen only in the UKESM1-0-LL model. Water limitaion effects on GPP in ESMs is typically modelled quite crudely, with uncertain parameterization (Harper et al., 2020) , and this is likely a significant source of disparity between the models.

3.2 Comparing differences in climate forcing vs GPP in model pairs

We compare ESM differences in the input feature space with their GPP distributions with the approach described in 2.5. In Figure 5 we show the comparative distances as a scatter plot to illustrate how we can potentially develop our hypotheses for quantifying and thus attributing differences in GPP to differences in climate forcing or process representation.

From the scatter plots in 5, we see differences across regions in how the pairwise model distances relate. If distances in input climate space between pairs of models translated to similar distances in GPP distributions, we would see the data points scattered along the diagonal unit slope line as seen in the NSA region. However this is not always the case, and we see more of a spread along the input space or x-axis (MED, RAR and somewhat also in SAS) where the plot indicates a spread in climate not quite seen in the simulated GPP and where relative differences in GPP are smaller than in input forcing . In other regions (SEA) however almost all pairs are above the unit slope line, which means that distances are larger in the GPP space.

We can use information from where there is a spread to investigate the likely causes underlying GPP divergence across models. In at least two regions (RAR and SAS), we notice that relative model distances with UKESM1-0-LL are greater in the y-axis even though such distances in the input space lie more or less in the middle range. This is an indication that the GPP simulated by UKESM1-0-LL is most different compared to other models even though not largely different in climate. In the SAS region for instance, the IPSL-CM6A-LR and UKESM1-0-LL models are closest in input space relative to other model pairs (seen as black colored letter I), and the CanESM5 model is identically distanced from both these models in the input space (seen as black and blue letters Ca). However, we see that in GPP space the UKESM1-0-LL distance with CanESM5 is more than the distance between CanESM5 and IPSL-CM6A-LR. Therefore one hypothesis worth investigating for this region is whether GPP process representation in IPSL-CM6A-LR and CanESM5 is similar in parameterization and different from UKESM1-0-LL. We would also include information from our feature importance results in 3 where we see that the two models differ in the variable considered most relevant for GPP (this is precipitation for UKESM1-0-LL, CanESM5 and temperature for IPSL-CM6A-LR). We argue that this type of analysis would be difficult to apply if we only consider univariate statistics as we show with examples in Supplementary Figure S4.

As a counter example, the ENA and to some extent the WSAF regions are examples of where it is not so clear how much of the difference in GPP to attribute to the influence of atmospheric forcing vs process representation from the scatter plot in Figure 5 due to close clustering in the relative distances.

4 Discussion

4.1 Choice of ML Approach for Evaluation

GPP is the largest individual carbon flux in the Earth System and changes to it have implications for the atmospheric carbon dioxide concentration, net carbon balance of the land surface and climate feedbacks (Friedlingstein et al., 2014). Interannual variability in GPP is influenced by changes in climate especially in hotspot regions such as tropical forests (O'Sullivan et al., 2020; Jung et al., 2011). Earth System Models provide the capability to simulate the Earth System's biogeochemical interactions and carbon cycle but global GPP estimates from ESMs vary greatly. For instance, in the five CMIP6 ESMs in our study, we found the global mean annual GPP to be in the range of 82-115 PgC year^{-1} for the pre-industrial period. The need to evaluate the carbon cycle in ESMs is thus critical for both better process representation and for modeling interactions with other components of the Earth System such as the atmosphere (Spafford & MacDougall, 2021; Reichler & Kim, 2008). Advances in Machine Learning and AI provides the algorithms that can help to facilitate evaluation of these complex interactions and uncover process based differences across ESMs (Huntingford et al., 2019). Our ap-

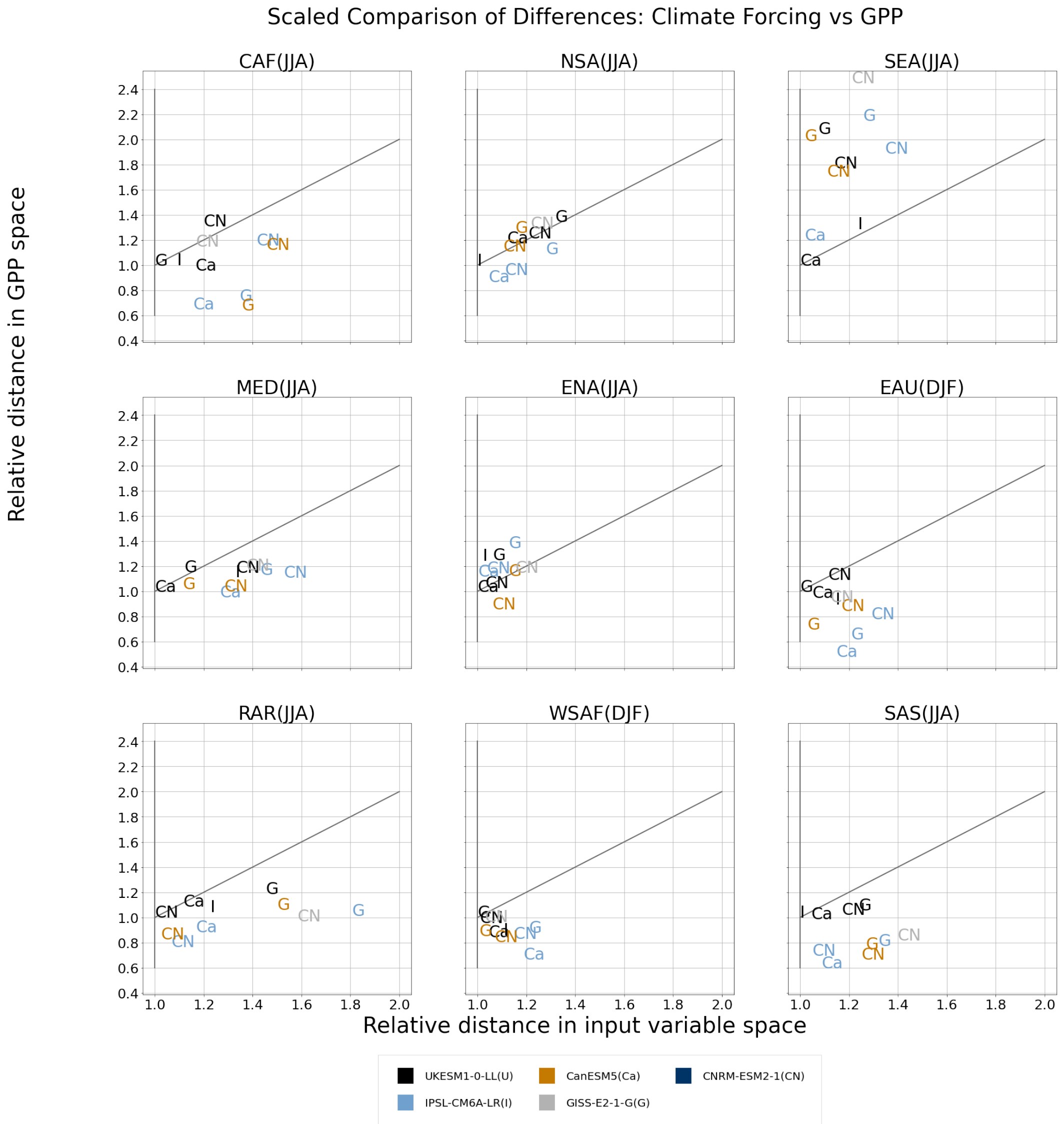


Figure 5. A comparison of relative distances in climate forcing and in GPP from different climate models is shown. Every model is referenced by both a color and an alphabet, the color and alphabet pairing tells us which pair of models are represented. Since the JSD is symmetric, there is only one colored symbol to show the distance between every pair of models. For this reason, there is no letter seen for the first model in the list, UKESM1-0-LL but its color (black) and letters for other models show the distance between UKESM1-0-LL and other models. For each region, the actual JSD values are scaled by factor that is the smallest distance in the input space across all pairs of models as seen in the x-axis and by the distance measure for that same pair in the GPP space as seen in the y-axis. This scaling follows from the description in Section 2 and Figure 2.

proach has been to start with the simplest ML models suited for our purpose. For this study, we build ML emulators with three input climate features to estimate GPP and for that emulator to be interpretable, which we demonstrate with our Feature Selection algorithms. Therefore, our ML emulators are not black boxes but can be interpreted in the context of physical and biogeochemical Earth System processes. We evaluated a choice of regression schemes before determining that Decision Trees best suited our task and further added better generalization capabilities with Boosting in the form of an Ensemble Learner with Adaboost. Such an emulator was capable of readily providing explanations on the modeled interactions between the atmospheric variables and GPP. At the same time, our framework is flexible enough for this emulator to be replaced with more complex ML algorithms such as Deep Architectures (LeCun et al., 2015) as we expand our suite of interacting variables for more nuanced evaluation of the carbon cycle. We further built robustness into our methods through rigorous cross validation and through the approaches outlined in Section 2.3 and demonstrate a reliable and adaptable framework that is also interpretable. With this framework, we were able to show regional similarities and differences in ESMs in the influence of key climate variables for GPP. Our emulator has the capability to capture non-linear relationships between the climate variables and GPP which can help to address limitations or complement more traditional approaches using correlations or calculated indices seen in the literature (O'Sullivan et al., 2020; Seddon et al., 2016).

The second component of our framework is a way to compare differences in climate variables influencing GPP with differences in processes estimating GPP in ESMs and we choose an algorithm based on the Jensen Shannon distance that is robust against small variations in distributions, allows a comparison of the joint input space with three variables and has bounds [0,1] to enable relative placement of distances. Also where a statistic such as a mean could be close for two different distributions, such as unimodal vs bimodal, the JSD will capture a difference in parameterization resulting in quite different distributions with similar means. Finally, our method enables a more flexible and less expensive way to perform this comparison where previously modeling experiments had to be conducted for similar analysis (Hardouin et al., 2022).

4.2 Application of ML framework for GPP Evaluation

The ML framework described in this paper can be used to identify areas of differences in GPP modeling in ESMs. For instance, from Figure 4 and Figure 3, we see that while models have overall agreement on what variables are important for certain regions (temperature and precipitation for the Mediterranean, South Asia, Eastern and Central North America; temperature and radiation in the tundra and boreal forest regions) differences exist in the which individual climate variable matters for a given ESM. Further the comparison using JSD gives us a starting point for whether these differences are more in the state of the climate influencing GPP or in the processing of these variables such as through parameterizations. This ML framework can serve as a guide to investigate probable reasons why differences in GPP modeling exist in ESMs in a computationally less expensive manner to actually running model simulations.

4.3 Limitations and Challenges

In our current study, we sample data uniformly from the spatio-temporal domain which does not capture sub-regional and sub-seasonal variability and trends. This limitation is mainly driven by the lack of availability of GPP data from CMIP6 ESMs at higher temporal resolutions for the pi-Control experiment. However, this is more a feature of the data used and our framework will allow us to experiment with different resolutions in data when available. The JSD approach provides a relatively inexpensive method, without actually having to run model simulations, to compare differences across models in GPP vs climate variables but in some regions such as Eastern North America (ENA) seen in Figure 5, it is harder to infer where the differences lie. Along with future work to develop this analysis, we also suggest that individual components of the ML framework as well as more traditionally considered descriptive statistics such as means and variability should all be used in a complementary fashion in the evaluation process so we can take insights from different modes of analysis. Finally, the three predictor variables were chosen because of their importance in determining the supply of water (precipitation), its loss through evapotranspiration (temperature) and the available energy for photosynthesis (shortwave radiation). We recognize the need to include a broader suite of variables for a more holistic evaluation of the carbon cycle which is possible to do with our framework.

5 Conclusions

This study demonstrates the potential of using interpretable ML approaches to investigate differences in GPP modeling across a selection of CMIP6 models and over land regions defined in the IPCC's Sixth Assessment Report and two seasons. In conclusion:

1. The relative importance of key climate drivers for GPP was identified across different regions and ESMs using Feature Selection Methods with interpretable ML emulators. We illustrate this with examples such as the Mediterranean region where all models agree that drought variables such as temperature or precipitation influence GPP more than radiation but models differ in which of the two variables is most relevant.
2. With a comparative distance metric based on the Jensen Shannon Distance, we are able to show that proximity or distance in climate between any two models does not necessarily translate to a similar proximity or distance in their estimated GPP distributions with the Russian Arctic (RAR) and Mediterranean regions (MED) as two such examples. We take this as evidence that process based differences exist across models and are at least partly responsible for differences in GPP estimates from ESMs.
3. Where the JSD method suggests divergence in GPP potentially due to process modeling, for instance in South Asia (SAS) between the UKESM1-0-LL, IPSL-CM6A-LR and CanESM5 models, the Feature Selection process can offer an explanation. In this case the UKESM1-0-LL and IPSL-CM6A-LR models differ in the key climate variable for GPP but the UKESM1-0-LL and CanESM5 models don't and a possible reason for this can be differences in parameterization or characteristics of this variable not considered in the input features.
4. There are some regions where models do not show a clear consensus on what climate variables matter the most or identify all three variables as equally important such as the tropics. Similarly our distance metric based comparison also presents cases where a direct inference on attributing GPP differences cannot be made, such as the Eastern North American (ENA) region. We identify these as regions of uncertainty to address in future work.

Data from the pre-industrial Control experiments served as a baseline for the development of this evaluation framework. In future work, additional climate drivers and characteristics such as sub-monthly variability will also be incorporated as possible causes for variations in GPP estimates from ESMs and analysis will be conducted with data from historical experiments and observations towards the goal of improving vegetation modeling in Earth System Models.

6 Open Research

Data from CMIP6 climate models is available for download on Earth System Grid Federation nodes and were downloaded and preprocessed using the open source software

ESMValTool v2.8.0 (doi:10.5281/zenodo.3401363) and ESMValCore v2.8.0 (doi:10.5281/zenodo.3387139). Code used to produce the results in this paper is available under the CC-BY license at the Github respository (https://github.com/rswamina/gpp-ml-eval-1-publish) which is currently private but will be made public once the manuscript has been accepted for publication.

Acknowledgments

RS and TQ are funded by the UK Research Infrastructure Natural Environment Research Council (UKRI-NERC) funded TerraFIRMA: Future Impacts, Risks and Mitigation Actions in a changing Earth system grant (NE/W004895/1). RA is funded by the National Centre for Earth Observation grant: NE/RO16518/1.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., ... others (2015). Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, *53*(3), 785–818. doi: 10.1002/2015RG000483

Bo, Y., Li, X., Liu, K., Wang, S., Zhang, H., Gao, X., & Zhang, X. (2022). Three decades of gross primary production (gpp) in china: Variations, trends, attributions, and prediction inferred from multiple datasets and time series modeling. *Remote Sensing*, *14*(11), 2564. doi: 10.3390/rs14112564

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., ... others (2020). Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. doi: 10.1029/2019MS002010

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140. doi: /10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi: doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Cart. *Classification and regression trees*.

Churkina, G., & Running, S. W. (1998). Contrasting climatic controls on the estimated productivity of global terrestrial biomes. *Ecosystems*, *1*(2), 206–215. doi: 10.1007/s100219900016

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., ... others (2014). The earth system grid federation: An open infrastructure for access to distributed geospatial data. *Future Generation Computer Systems*, *36*, 400–417. doi: 10.1016/j.future.2013.07.002

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M., ... others (2011). The joint uk land environment simulator (jules), model description–part 2: carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701–722. doi: 10.5194/gmd-4-701-2011

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, 146–158. doi: 0.1214/aop/1176996454

Delire, C., Séférian, R., Decharme, B., Alkama, R., Calvet, J.-C., Carrer, D., ... others (2020). The global land carbon cycle simulated with isba-ctrip: Improvements over the last decade. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS001886. doi: 10.1029/2019MS001886

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of machine learning research*, *3*, 1265–1287. doi: 10.5555/944919.944973

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. doi: 10.48550/arXiv.1702

.08608
Drucker, H. (1997). Improving regressors using boosting techniques. In *Icml* (Vol. 97, pp. 107–115). doi: 10.5555/645526.657132
Dunkl, I., Lovenduski, N., Collalti, A., Arora, V. K., Ilyina, T., & Brovkin, V. (2023). Gross primary productivity and the predictability of co 2: more uncertainty in what we predict than how well we predict it. *Biogeosciences*, *20*(16), 3523–3538. doi: 10.5194/bg-20-3523-2023
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. doi: 10.5194/gmd-9-1937-2016
Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2018MS001453. doi: 10.1029/2018MS001453
Fisher, R. A., Koven, C. D., Anderegg, W. R., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., … others (2018). Vegetation demographics in earth system models: A review of progress and priorities. *Global change biology*, *24*(1), 35–54. doi: 10.1111/gcb.13910
Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in cmip5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, *27*(2), 511–526. doi: /10.1175/JCLI-D-12-00579.1
Gea-Izquierdo, G., Guibal, F., Joffre, R., Ourcival, J., Simioni, G., & Guiot, J. (2015). Modelling the climatic drivers determining photosynthesis and carbon allocation in evergreen mediterranean forests using multiproxy long time series. *Biogeosciences*, *12*(12), 3695–3712. doi: 10.5194/bg-12-3695-2015
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. doi: 10.1145/3422622
Greenslade, M., Murphy, S., Treshansky, A., DeLuca, C., Guilyardi, E., & Denvil, S. (2014). The earth system (es-doc) project. In *Egu general assembly conference abstracts* (p. 12988).
Gültas, M., Düzgün, G., Herzog, S., Jäger, S. J., Meckbach, C., Wingender, E., & Waack, S. (2014). Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum jensen-shannon divergence and cuda programming. *BMC bioinformatics*, *15*(1), 1–17. doi: 10.1186/1471-2105-15-96
Gutiérrez, J. M., Jones, R., Narisma, G., et al. (2021). Ipcc interactive atlas. In *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change.* Cambridge University Press Cambridge. doi: 10.1175/BAMS-D-20-0256.1
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182. doi: 10.5555/944919.944968
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389–422. doi: 10.1023/A:1012487302797
Hardouin, L., Delire, C., Decharme, B., Lawrence, D. M., Nabel, J. E., Brovkin, V., … others (2022). Uncertainty in land carbon budget simulated by terrestrial biosphere models: the role of atmospheric forcing. *Environmental Research Letters*, *17*(9), 094033. doi: 10.1088/1748-9326/ac888d
Harper, A. B., Williams, K. E., McGuire, P. C., Duran Rojas, M. C., Hemming, D., Verhoef, A., … others (2020). Improvement of modelling plant responses

to low soil moisture in julesvn4. 9 and evaluation against flux tower measurements. *Geoscientific Model Development Discussions*, *2020*, 1–42. doi: gmd-14-3269-2021

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. doi: 10.1088/1748-9326/ab4e55

Iturbide, M., Fernández, J., Gutiérrez, J. M., Pirani, A., Huard, D., Al Khourdajie, A., ... others (2022). Implementation of fair principles in the ipcc: the wgi ar6 atlas repository. *Scientific data*, *9*(1), 629. doi: 0.5281/zenodo.3691645

James, G., Witten, D., Hastie, T., Tibshirani, R., James, G., Witten, D., ... Tibshirani, R. (2021). Linear regression. *An introduction to statistical learning: with applications in R*, 59–128. doi: 10.1007/978-1-0716-1418-1_3

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... Reichstein, M. (2019). The fluxcom ensemble of global land-atmosphere energy fluxes. *Scientific data*, *6*(1), 1–14. doi: 10.1038/s41597-019-0076-8

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., ... others (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, *116*(G3). doi: 10.1029/2010JG001566

Kanniah, K. D., Beringer, J., & Hutley, L. (2013). Exploring the link between clouds, radiation, and canopy productivity of tropical savannas. *Agricultural and Forest Meteorology*, *182*, 304–313. doi: 10.1016/j.agrformet.2013.06.010

Kanniah, K. D., Beringer, J., & Hutley, L. B. (2011). Environmental controls on the spatial variability of savanna productivity in the northern territory, australia. *Agricultural and Forest Meteorology*, *151*(11), 1429–1439. doi: 10.1016/j.agrformet.2011.06.009

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., ... others (2020). Giss-e2. 1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025. doi: 10.1029/2019MS002025

Kiang, N. (2012). *Description of the nasa giss vegetation dynamics model* (Tech. Rep.). Tech. rep., NASA.

Kim, D., Lee, M.-I., Jeong, S.-J., Im, J., Cha, D. H., & Lee, S. (2018). Intercomparison of terrestrial carbon fluxes and carbon use efficiency simulated by cmip5 earth system models. *Asia-Pacific Journal of Atmospheric Sciences*, *54*(2), 145–163. doi: 10.1007/s13143-017-0066-8

Koch, A., Hubau, W., & Lewis, S. L. (2021). Earth system models are not capturing present-day tropical forest carbon dynamics. *Earth's Future*, *9*(5), e2020EF001874. doi: /10.1029/2020EF001874

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, *19*(1). doi: 10.1029/2003GB002199

Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, *4*(3), 211–229. doi: 10.6029/smartcr.2014.03.007

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444. doi: doi.org/10.1038/nature14539

Levis, S. (2010). Modeling vegetation and land use in models of the earth system. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(6), 840–856. doi: 10.1002/wcc.83

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151. doi: 10.1109/18.61115

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, *45*(1), 1–40. doi: 10.1145/2379776.2379786

Molnar, C. (2020). *Interpretable machine learning.* Lulu. com.

Nishina, K., Ito, A., Falloon, P., Friend, A., Beerling, D., Ciais, P., ... others (2015). Decomposing uncertainties in the future terrestrial carbon budget associated with emission scenarios, climate projections, and ecosystem simulations using the isi-mip results. *Earth System Dynamics*, *6*(2), 435–445. doi: 10.5194/esd-6-435-2015

Nzabarinda, V., Bao, A., Xu, W., Uwamahoro, S., Jiang, L., Duan, Y., ... Long, G. (2021). Assessment and evaluation of the response of vegetation dynamics to climate variability in africa. *Sustainability*, *13*(3), 1234. doi: 10.3390/su13031234

on Climate Change, I. P. (2023). Global carbon and other biogeochemical cycles and feedbacks. In *Climate change 2021 – the physical science basis: Working group i contribution to the sixth assessment report of the intergovernmental panel on climate change* (p. 673–816). Cambridge University Press. doi: 10.1017/9781009157896.007

O'Sullivan, M., Smith, W. K., Sitch, S., Friedlingstein, P., Arora, V. K., Haverd, V., ... others (2020). Climate-driven variability and trends in plant productivity over recent decades based on three global products. *Global Biogeochemical Cycles*, *34*(12), e2020GB006613. doi: 10.1029/2020GB006613

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. doi: 10.5555/1953048.2078195

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106. doi: doi.org/10.1007/BF00116251

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–312. doi: 10.1175/BAMS-89-3-303

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., ... others (2020). Earth system model evaluation tool (esmvaltool) v2. 0–technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199.

Santini, M., Collalti, A., & Valentini, R. (2014). Climate change impacts on vegetation and water cycle in the euro-mediterranean region, studied by a likelihood approach. *Regional Environmental Change*, *14*(4), 1405–1418.

Sarkar, D. P., Shankar, B. U., & Parida, B. R. (2022). Machine learning approach to predict terrestrial gross primary productivity using topographical and remote sensing data. *Ecological Informatics*, *70*, 101697. doi: 10.1016/j.ecoinf.2022.101697

Schapire, R. E. (2013). Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37–52. doi: 10.1007/978-3-642-41136-6_5

Schimel, D. S., House, J. I., Hibbard, K. A., Bousquet, P., Ciais, P., Peylin, P., ... others (2001). Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, *414*(6860), 169–172. doi: 10.1038/35102500

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., & Reichstein, M. (2020). Constraining uncertainty in projected gross primary production with machine learning. *Journal of Geophysical Research: Biogeosciences*, *125*(11), e2019JG005619. doi: 10.1029/2019JG005619

Schwalm, C. R., Huntzinger, D. N., Michalak, A. M., Schaefer, K., Fisher, J. B., Fang, Y., & Wei, Y. (2020). Modeling suggests fossil fuel emissions have been driving increased land carbon uptake since the turn of the 20th century. *Scientific Reports*, *10*(1), 9059. doi: 10.1038/s41598-020-66103-9

Seddon, A. W., Macias-Fauria, M., Long, P. R., Benz, D., & Willis, K. J. (2016). Sensitivity of global terrestrial ecosystems to climate variability. *Nature*, *531*(7593), 229–232. doi: 10.1038/nature16986

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., ... others (2019). Evaluation of cnrm earth system model, cnrm-esm2-1: role of earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4182–4227. doi: 10.1029/2019MS001791

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., ... others (2019). Ukesm1: Description and evaluation of the uk earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. doi: 10.1029/2019MS001739

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Spafford, L., & MacDougall, A. H. (2021). Validation of terrestrial biogeochemistry in cmip6 earth system models: a review. *Geoscientific Model Development*, *14*(9), 5863–5889. doi: gmd-14-5863-2021

Sun, Y., Frankenberg, C., Wood, J. D., Schimel, D., Jung, M., Guanter, L., ... others (2017). Oco-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. *Science*, *358*(6360), eaam5747. doi: 10.1126/science.aam57

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., ... others (2019). The canadian earth system model version 5 (canesm5. 0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. doi: 10.5194/gmd-12-4823-2019

Varghese, R., & Behera, M. (2019). Annual and seasonal variations in gross primary productivity across the agro-climatic regions in india. *Environmental monitoring and assessment*, *191*(10), 631. doi: 10.1007/s10661-019-7796-2

Verma, A., Chandel, V., & Ghosh, S. (2022). Climate drivers of the variations of vegetation productivity in india. *Environmental Research Letters*, *17*(8), 084023. doi: 10.1088/1748-9326/ac7c7f

Verseghy, D. (2012). Class–the canadian land surface scheme (version 3.6). *Environment Canada Science and Technology Branch Tech. Rep*, *176*.

Wu, D., Zhao, X., Zhao, W., Tang, B., & Xu, W. (2014). Response of vegetation to temperature, precipitation and solar radiation time-scales: A case study over mainland australia. In *2014 ieee geoscience and remote sensing symposium* (pp. 855–858). doi: 10.1109/IGARSS.2014.6946559

Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R., & Lehsten, V. (2017). Climate data induced uncertainty in model-based estimations of terrestrial primary productivity. *Environmental Research Letters*, *12*(6), 064013. doi: 10.1088/1748-9326/aa6fd8

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. doi: /10.1016/j.neucom.2020.07.061

Yao, Y., Wang, X., Li, Y., Wang, T., Shen, M., Du, M., ... others (2018). Spatiotemporal pattern of gross primary productivity and its covariation with climate in china over the last thirty years. *Global Change Biology*, *24*(1), 184–196.

Yu, T., Zhang, Q., & Sun, R. (2021). Comparison of machine learning methods to up-scale gross primary production. *Remote Sensing*, *13*(13), 2448. doi: 10.3390/rs13132448

Zampieri, M., Grizzetti, B., Toreti, A., De Palma, P., & Collalti, A. (2021). Rise and fall of vegetation annual primary production resilience to climate variability projected by a large ensemble of earth system models' simulations. *Environmental Research Letters*, *16*(10), 105001. doi: 10.1088/1748-9326/ac2407

Zarakas, C. M., Swann, A. L., Laguë, M. M., Armour, K. C., & Randerson, J. T. (2020). Plant physiology increases the magnitude and spread of the transient climate response to co2 in cmip6 earth system models. *Journal of Climate*, *33*(19), 8561–8578. doi: 10.1175/JCLI-D-20-0078.1

Zhang, F., Lu, X., Huang, Q., & Jiang, F. (2022). Impact of different era reanalysis data on gpp simulation. *Ecological Informatics*, *68*, 101520. doi: 10.1016/j.ecoinf.2021.101520

Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S., & Gentine, P. (2018). A global spatially contiguous solar-induced fluorescence (csif) dataset using neural networks. *Biogeosciences*, *15*(19), 5779–5800. doi: 10.5194/bg-15-5779-2018

Zhang, Z., Xin, Q., & Li, W. (2021). Machine learning-based modeling of vegetation leaf area index and gross primary productivity across north america and comparison with a process-based model. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002802. doi: 10.1029/2021MS002802

Evaluating Vegetation Modeling in Earth System Models with Machine Learning Approaches

Ranjini Swaminathan[1,2], **Tristan Quaife**[1,2] **and Richard Allan**[1,2]

[1]University of Reading
[2]National Centre for Earth Observation

Key Points:

- A Machine Learning framework to advance our understanding of the terrestrial carbon cycle in Earth System Models or ESMs is proposed
- Differences in the relative importance of atmospheric drivers of gross primary productivity highlights differences across models
- A method to attribute differences in productivity estimates from ESMs due to process representation versus atmospheric forcing is demonstrated

Corresponding author: Ranjini Swaminathan, `r.swaminathan@reading.ac.uk`

Abstract

Vegetation Gross Primary Productivity (GPP) is the single largest carbon flux of the terrestrial biosphere which, in turn, is responsible for sequestering 25−30% of anthropogenic carbon dioxide emissions. The ability to model GPP is therefore critical for calculating carbon budgets as well as understanding climate feedbacks. Earth System Models (ESMs) have the capability to simulate GPP but vary greatly in their individual estimates, resulting in large uncertainties. We describe a Machine Learning (ML) approach to investigate two key factors responsible for differences in simulated GPP quantities from ESMs: the relative importance of different atmospheric drivers and differences in the representation of land surface processes. We describe the different steps in the development of our interpretable Machine Learning (ML) framework including the choice of algorithms, parameter tuning, training and evaluation. Our results show that ESMs largely agree on the physical climate drivers responsible for GPP as seen in the literature, for instance drought variables in the Mediterranean region or radiation and temperature in the Arctic region. However differences do exist since models don't necessarily agree on which individual variable is most relevant for GPP. We also explore a distance measure to attribute GPP differences to climate influences versus process differences and provide examples for where our methods work (South Asia, Mediterranean)and where they are inconclusive (Eastern North America).

Plain Language Summary

Gross Primary Productivity (GPP) is the rate at which plants remove carbon dioxide from the atmosphere during photosynthesis. Carbon dioxide is a greenhouse gas and excess in the atmosphere causes global warming and climate change. Changes in the amounts of atmospheric carbon dioxide will impact the entire Earth System. We therefore need the ability to accurately calculate GPP, especially for different possible carbon usage pathways in the future. Earth System Models or ESMs allow us to simulate various processes happening in the earth's atmosphere and biosphere including photosynthesis and can help us estimate GPP changes for such different pathways. However, ESMs can vary significantly in their simulated GPP estimates making it difficult to have confidence in using these estimates. We describe a Machine Learning (ML) framework to better understand where ESMs differ in calculating GPP so that we can address knowledge gaps in models. This approach allows us to understand the processes involved without having to run computationally expensive simulations. With improved models, we can also improve our ability to predict climate change outcomes for the future.

---

1 Introduction

Terrestrial Gross Primary Production (GPP) is the flux of carbon into the land surface driven by photosynthesis.

It is estimated that terrestrial GPP is in the order of $\sim 132PgC$ and it is the single largest annual flux of the global carbon cycle. It plays a key role in determining atmospheric carbon dioxide, since approximately a quarter to a third of anthropogenic emissions are sequestered by the land surface (on Climate Change, 2023; Schimel et al., 2001; Schwalm et al., 2020). GPP is influenced by natural climate variability as well as anthropogenic factors associated with global warming (Santini et al., 2014; Zampieri et al., 2021). Our ability to estimate GPP, its spatio-temporal patterns and the factors influencing GPP is therefore essential to understanding and forecasting global carbon budgets with greater reliability. GPP is not a directly measurable quantity at spatial scales of interest for carbon budget calculations (global or regional), so we rely on indirect measurements with inevitable assumptions, for example about the partitioning of fluxes at eddy covariance sites (Jung et al., 2019) or from satellite observations of quantities such as Solar Induced Fluorescense (SIF) (Sun et al., 2017; Y. Zhang et al., 2018), which are not direct measures of the carbon flux.

Earth System Models (ESMs) provide the capability to simulate GPP by modelling the various interactions between the atmosphere and biosphere including under different climate change scenarios in the future (Fisher et al., 2018; Levis, 2010). However, there is not only a large spread in GPP estimates from different ESMs but there are also large uncertainties in observational products that could be used to evaluate these estimates (Z. Wu et al., 2017; Anav et al., 2015). Therefore, there is a real need for evaluation methods that will help us understand better the possible reasons for such a large spread in GPP simulations, both in terms of the influence of atmospheric variables driving GPP as well as in the representation of the processes involved in simulating GPP. Identifying these differences can further help us address key gaps in modeling the terrestrial carbon cycle and will make for more reliable simulations from ESMs.

Machine Learning (ML) approaches have recently been used extensively in the study as well as generation of more accurate GPP data sets. Examples are seen work done in simulating GPP using observations of meteorological data or satellite data (Z. Zhang et al., 2021; Sarkar et al., 2022), upscaling GPP estimates from eddy covariance sites (Yu et al., 2021), to constrain uncertainty in GPP projections from models (Schlund et al., 2020) and for evaluating GPP representation in models (Z. Zhang et al., 2021; Dunkl et al., 2023). Our goal in this study is to use interpretable Machine Learning approaches (Molnar, 2020; Doshi-Velez & Kim, 2017) to better understand the sources of differences in GPP estimates between ESMs. Such an ML based evaluation framework can serve as a basis for process based improvements to ESMs, complementary to existing strategies, and can help reduce process uncertainty in modelled GPP estimates leading to more reliable simulations.

In previous studies, differences in GPP estimates from ESMs have been attributed to differences in the simulations of climate projections, modeling of complex terrestrial processes such as dynamic vegetation modeling, as well as atmospheric $\mathrm{CO_2}$ concentrations for given emission scenarios (Nishina et al., 2015; Schwalm et al., 2020; Fisher & Koven, 2020; Kim et al., 2018; Koch et al., 2021). In this work, we focus on two key attributes responsible for variability in GPP across ESMs - (a) the differences in climate simulations or input atmospheric forcing influencing GPP in individual models and (b) differences arising from vegetation process representation in these models. While we acknowledge that GPP is dependent on several land and atmospheric variables, in keeping with other similar studies such as Churkina and Running (1998); Schwalm et al. (2020); Anav et al. (2015), we evaluate the influence of three atmospheric variables as primary determinants of photosynthesis – precipitation, air temperature and downwelling shortwave radiation.

Our framework uses simulations from the CMIP pre-industrial Control (pi-Control) experiments that simulate climate before industrialization and the addition of anthropogenic CO_2 to the atmosphere. These simulations do not have the effects of elevated CO_2 that could lead to vegetation feedbacks or of any warming signal due to climate change. This allows us to better isolate the direct influence of the input climate variables on GPP without these factors. ESM simulations from pi-Control runs are also run for longer time periods, typically a few hundred years as opposed to a few decades from the historical experiment simulations and so this gives us a larger data set to learn from.

The methods used in this framework are based on Information Theory and Machine Learning, and compare the differences in input atmospheric forcings and vegetation process modeling associated with simulating GPP, across different ESMs from the Sixth Phase of the Coupled Model Intercomparison Project (CMIP6) (Eyring et al., 2016). These methods are directed towards formulating informed hypotheses for investigating the underlying factors influencing GPP estimates from ESMs. Specifically, the methods described target the following questions:

1. How do CMIP6 models differ in the input atmospheric forcings they consider most relevant for GPP? This will help us understand potential differences in how climate variables may influence GPP across models.
2. Can we compare differences in input forcings across ESMs with their process based differences? This will guide us towards attributing differences in GPP to the appropriate underlying factors.

We address the above questions by building ML based emulators of CMIP6 models that estimate GPP with input climate data. We query these emulators using robust Feature Selection methods to determine the relevance of individual atmospheric variables with respect to GPP. We also compare the differences in input forcing vs GPP by using a distance metric called the Jensen-Shannon distance measure. This is a novel approach that allows a comparison of two different attributory factors responsible for GPP and to the best of our knowledge is not previously seen in the literature.

We find that while the CMIP6 models considered largely agree on the variables considered relevant for GPP, there are regions of uncertainty such as the tropics. We are also able to show that models with similar input forcings do not always show similar estimates in GPP, indicating differences in process representation possibly due to parameterization. The remainder of the paper is organized as follows – Section 2 describes the ML framework including the parameter tuning process and algorithmic description of the learning and Feature Selection approaches. In Section 3, we discuss results where the ML framework identifies differences in climate variables influencing GPP across ESMs. In Section 4, we discuss the interpretability of the ML framework described, how this framework can be used for evaluation and some of the challenges involved. Finally we present our conclusions and planned future work using for this framework in 5.

2 Data and Methods

2.1 Data and Pre-processing

Our experimental experimental input data consists of five ESMs (UKESM1-0-LL, IPSl-CM6A-LR, CanESM5, CNRM-ESM2-1 and GISS-E2-1-G) from the CMIP6 project, all with different vegetation and land surface models as shown in Table2.1. The criteria applied for selection was to pick a small set of models with diversity in their vegetation modeling schemes, permitting exploration of various aspects of GPP simulation through our ML framework.

Seasonal means were calculated from monthly means of the data for two seasons, the boreal summer season of June-July-August (JJA) and austral summer season of December-January-February (DJF). All data considered is from the pre-industrial control (pi-Control) experiments which do not have an anthropogenic warming signal and for which a few hundred years of data are available from every model. Analysis is done for regions defined in the Intergovernmental Panel on Climate Change's Sixth Assessment Report (IPCC AR6), (Gutiérrez et al., 2021). Data was downloaded and pre-processed from the Earth System Grid Federation servers (Cinquini et al., 2014) using the open source evaluation tool, ESMValTool (Righi et al., 2020). We removed all non-land grid cells of a model in a selected region to focus on terrestrial GPP and then sampled data uniformly across time and space. Every grid cell and every time instance constitutes a sample data point and for each data point, we have one value each for the three atmospheric variables as well as for GPP. We then use this pre-processed data for further analysis. A pictorial description of our ML framework is shown in Figure 1.

Earth System Model	Land Surface Model	Reference	Dynamic Vegetation
UKESM1-0-LL	Joint UK Land Environment Simulator (JULES)	(Sellar et al., 2019; Clark et al., 2011)	Yes
IPSL-CM6A-LR	Organising Carbon and Hydrology In Dynamic Ecosystems (ORCHIDEE)	(Boucher et al., 2020; Krinner et al., 2005)	No
CanESM5	The Canadian Land Surface Scheme (CLASS)	(Swart et al., 2019; Verseghy, 2012)	No
CNRM-ESM2-1	Interaction Soil-Biosphere-Atmosphere (ISBA)	(Séférian et al., 2019; Delire et al., 2020)	No
GISS-E2-1-G	ENT Terrestrial Biosphere Model	(Kelley et al., 2020; Kiang, 2012)	No

Table 1. The CMIP6 models evaluated with our framework and their corresponding vegetation models. Data on dynamicity of vegetation obtained from the Earth System Documentation Project (Greenslade et al., 2014) and (Zarakas et al., 2020)

.

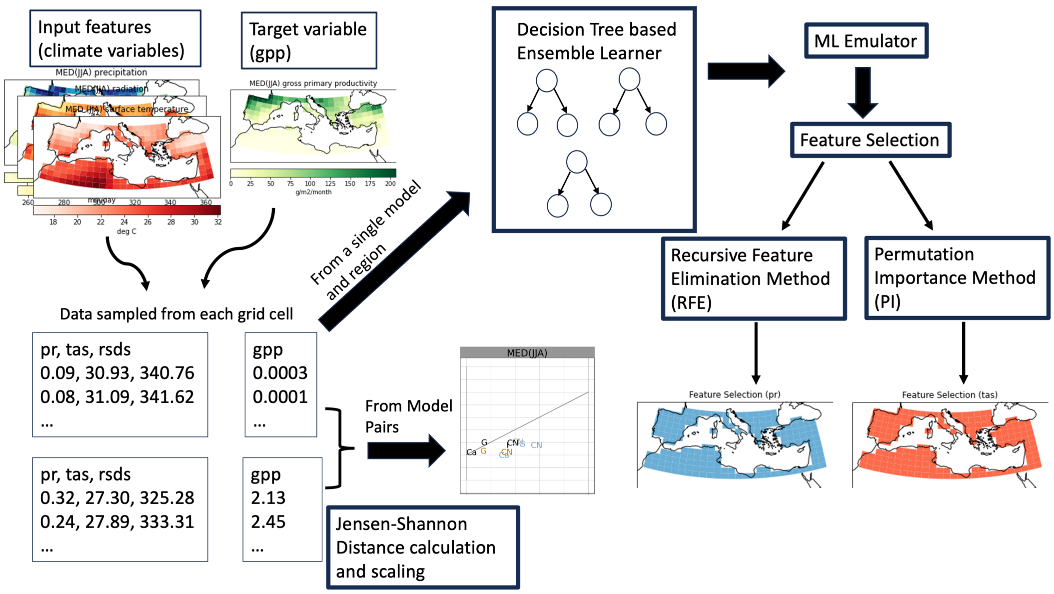


Figure 1. A description of the Machine Learning framework for evaluating GPP in CMIP6 models: Data from atmospheric variables and GPP for a given region, season and ESM is used to train an ensemble learner which serves as the ML emulator. The ML emulator is then queried using two different Feature Ranking algorithms (RFE or Recursive feature Elimination and PI or Permutation Importance) to find the most relevant features or atmospheric variables for GPP in that region. Data from pairs of ESMs is also used to calculate the Jensen-Shannon Distance (JSD) metric to compare distances measured in the input variable space with distances measured in the GPP distributions across regions.

2.2 ML Emulators with Ensemble Learning

Our requirement for an ML based emulator was one that would effectively model the relationship between input atmospheric forcing variables (and any other similar GPP influencing variables to be included as needed) and GPP; and one that would allow us to interpret or make inferences on the modeled relationships to answer questions on the relative importance or sensitivity to the climate variables. An additional goal was to develop a flexible framework that could be applied to observed data to better facilitate model evaluation. For this reason, we designed the core of the emulator to be a multivariate regression model and one that can be interpreted or queried on the decisions made for regression. In this, the climate forcing variables are the input features or predictors and GPP is the predictand. The ML emulator is trained for every region, season and ESM in our experimental setup. We use a regression model with Boosting called Adaptive Boosting or AdaBoost (Mendes-Moreira et al., 2012; Schapire, 2013) for our framework. Boosting is a well established ML approach that works towards developing a highly accurate prediction rule by repeatedly combining several weaker predictors or learners (Drucker, 1997) which in this case would be regressors. In Boosting, the first weak predictor is trained with a subset of samples uniformly sampled from the training data set with replacement permitted, meaning a training sample can be used again to build a different predictor. Once a predictor is built, all the training samples are passed through the predictor and the samples with the largest prediction errors are identified. The sampling probabilities of the samples with the most error are adjusted so that they are more likely to get picked as training samples for the next weak learner to be built. As this process repeats, harder to learn patterns get picked more often to build subsequent predictors. This means that some predictors will do better than others in a given subspace of the input feature space. The predictors are further assigned weights of the form, $\bar{\beta} = \frac{\bar{L}}{1-\bar{L}}$ where $\bar{L}$ is a calculated loss function. Cumulative predictions are calculated as a weighted median of all the predictors. The algorithm terminates when the average loss across all weak learners is below a certain threshold. The weak learners or regressors in this boosting algorithm can be any one of a wide array of regression methods. We calculated the Root Mean Square Error scores on held out test data sets and determined that the Decision Tree algorithm described in Breiman et al. (1984); Quinlan (1986); Breiman (1996) was best suited for our task after experimenting with different ML regression algorithms such as Linear Regression (James et al., 2021) and Support Vector Machines (Smola & Schölkopf, 2004). We therefore use an Ensemble Tree Learner with Boosting for our ML emulators.

As shown in Fig 1, CMIP6 data in the form of gridded data sets was used to train the ML emulators by treating each grid cell at every time step as an individual sample for learning. However, ESMs differ in grid resolution and in the length or number of years of the pi-Control experiment runs. So, for a given region, the number of training samples can be different across ESMs. In order to avoid biases resulting from differences in the number of samples, we randomly sampled a minimal sample set from every model such that the number of samples to train an emulator is the same across all ESMs. This sample set is then used to tune the parameters and build the Decision Trees in the ML emulator.

2.3 Parameter Tuning

In applied Machine Learning, parameter tuning is considered an important step in developing ML models that best capture patterns in the training data without overfitting (Yang & Shami, 2020). Overfitting occurs when we train the ML model to fit the training data too well which could result in a loss of generality. In other words, the ML model performs exceedingly well on the data it is trained with but fails to perform well on a new test set of samples even if from the same or similar distribution. We employ the Adaboost algorithm with an ensemble of Decision Tree regressors from the open source Python Scikit-learn package (Pedregosa et al., 2011) to build our ML emulators. A built in mechanism for pruning the ensemble learner exists for removing learners in a way that diversity is maximized. This essentially means that learners are selected such that a wide range of associations or rules are learnt and duplication of rules learnt is minimized by pruning. This helps to avoid overfitting by balancing the need to add more rules in the predictor with the ability to generalize well. In our experiments we tune for the depth parameter in the Decision Tree for optimal performance of the emulator, determined as the best fit to the data as evaluated by the Root Mean Squared Error (RMSE) in the predictions. The depth of the Decision Tree is the number of levels at which decision nodes are split in the tree. For example, a decision could be $tas > 20$ which would split training samples into those where the surface temperature is greater than 20°C (condition is true) and those where the temperature is less than 20°C (condition is false) and so on. For every region-season-ESM combination, we split the samples available into a training set and a held out test set. The ML emulator (AdaBoost with Decision Tree regressor) is learnt using the training samples and tested on the held out samples. RMSE scores are calculated for both training and held out test sets. For a given value of the depth parameter, this process is repeated by splitting the data n times and the average training and test RMSE scores over the n splits is calculated. This is how *n-fold* cross-validation (where n=*6* in this case) is performed. The depth parameter that has the lowest RMSE score on the held out test data, with cross–validation is then chosen as the most optimal parameter for the task and a final ML emulator is built using that depth parameter and all the samples available for that region. This builds robustness against overfitting, and sampling multiple times during cross validation further makes the model more reliable ensuring that the final emulator has seen a good representation of the available data. ML emulator estimates of GPP for a selection of regions are shown as an illustration of the results from this process in Supplementary Figure S1.

2.4 Feature Selection Methods

After the ML emulators were constructed to specification and sufficiently satisfied requirements, meaning the final emulator had the lowest possible RMSE scores for held out test data in cross validation experiments as described, we focused on querying or interpreting these emulators to better understand the relationship between the different input climate variables and GPP. Feature Selection or Feature Importance Ranking is the process of selecting or ranking features (input variables or predictors) that are most relevant to the predictand as evaluated by some chosen measurement or metric (Kumar & Minz, 2014; Guyon & Elisseeff, 2003). It is a process that is often used to prune the number of input features required for accurate predictions but in our case, with just three features, we use feature ranks to find the input atmospheric forcing variable(s) that the ML emulators find most important for GPP. Two different feature selection methods were applied to the ML emulators - (a) Recursive Feature Elimination (RFE) and (b) Permutation Importance (PI). The two methods use slightly different criteria to evaluate feature importances as described below but both provide useful information regarding the relative importance of a climate variable for GPP and are complementary. In the Recursive Feature Elimination algorithm, the input features are recursively removed one at a time to find the feature that has the most influence on the predictand (Guyon et al., 2002). For our experiments, we used the RMSE values to quantify the influence of an input climate variable on GPP. So, if the RFE method determines precipitation to be the most important feature for GPP, this effectively means that removing precipitation from the set of input features would have the most impact on the emulator's ability to predict GPP well i.e increase the RMSE by the most compared to other variables. In the Permutation Importance method, the decrease in model score when an individual feature is randomly shuffled or permutated is the measure of how important that feature is to the emulator (Breiman, 2001). The model score here is the Regression coefficient of determination (R^2) and is a measure of how well the ML emulator fits the data. Thus, the PI method works well once a reliable ML emulator is developed and is a measure of sensitivity of GPP to an input variable given such an emulator. As in the case of developing the ML emulator, we performed 6-fold cross-validation for the feature selection process as well. We did this by devising a simple voting scheme with small differences based on the Feature Selection approach. In the case of the RFE method, we assigned a single vote to the feature(s) that was ranked highest in terms of influencing the prediction with the RMSE score. We then averaged the votes across all the input features to determine the actual ranks of these features. In the PI method, we calculated the contribution of each feature to the R^2 score (permutation importances) and granted a vote to an input feature if it contributed to more than half of the score, which is the fit of the model. As in the RFE method, the votes were once again averaged across the cross-validation subsets. This scheme allowed us to account for collinearity or multiple variables equally influencing GPP especially as these are physical climate variables which are very closely related to each other.

2.5 Distance measure for climate and GPP distribution comparisons

While the ML emulators and Feature Selection are used to understand differences in models, we also calculate using a relative measure, how close or similar models are in the input forcing space vs. how similar they are in their simulated GPP distributions. Essentially we evaluate whether models that are similar in input atmospheric forcing simulated by the ESMs are also similar in their GPP simulations. If we consider that every data sample is represented as an instance in a 3-Dimensional input climate parameter space, where each dimension corresponds to a climate feature, then for a given region-season-ESM, we have a distribution of these 3-Dimensional data points. A distance metric is applied to quantify how close climate distributions from two different ESMs are for a given region and season. The same distance metric is now used to measure similarity between the GPP distributions of models in the 1-Dimensional space of GPP values. The distance metric we use is the Jensen-Shannon distance, which is calculated as the square root of the Jensen-Shannon divergence between two distributions (Lin, 1991). This is a symmetric and smoothed version of the more commonly used Kullback-Divergence measure. This measure has been widely used in applications such as evaluating generative adversarial networks by measuring differences in distributions (Goodfellow et al., 2020), text classification with high dimensional feature sets (Dhillon et al., 2003) and in bioinformatics for mutation detection (Gültas et al., 2014). The Jensen Shannon Divergence itself is defined as :

$$JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M), M = \frac{1}{2}(P+Q), \tag{1}$$

where $D(P||Q)$ is the Kullback-Divergence (Csiszár, 1975) between two distributions P and Q. When a base-2 logarithm is used, the Jensen-Shannon divergence has an upper bound of 1 i.e, $0 \ \leq \ JSD(P\|Q) \ \leq \ 1$. The existence of upper and lower bounds and the fact that distances are symmetric, are important properties we take advantage of when comparing ESMs. We refer to JSD as the Jensen-Shannon Distance instead of divergence as they both hold the same meaning for our analysis. Using the JSD, we compare how much ESMs differ in their input forcing vs in the simulated GPP for a region and season. A JSD of 0 implies the distributions are identical and as the JSD increases going towards 1, it implies that distributions get more dissimilar. While it is not possible to directly compare distance values between pairs of ESMs across two different distribution spaces (as in the 3-D climate space and the 1-D GPP space), we can compare how ESM-pair distances are ordered in both distribution spaces. That is we can see how distances between pairs of models compare in the two different spaces. We further apply a simple scaling by a factor of the shortest distance among all pairs of models in the in-

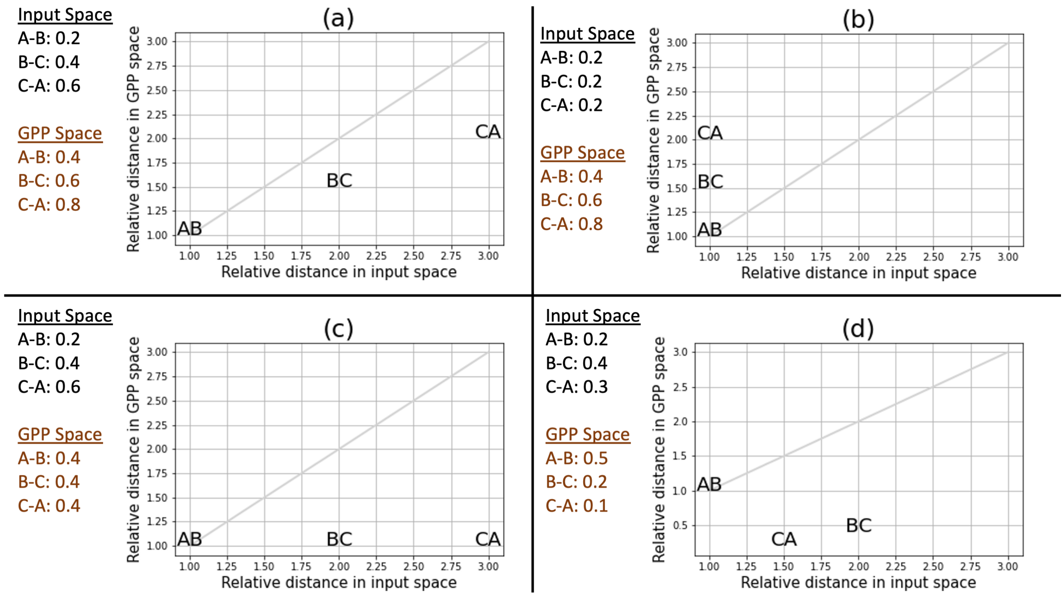


Figure 2. An illustration of how the Jensen Shannon distance metric is used to understand differences in input space (atmospheric forcings) and GPP space. In subplot (a) of the figure, we can make the inference that similarities in input forcing are consistent with similarities in GPP . Where that does not hold, we can start to explore the possibility that there might be larger differences in process representation or parameterization between pairs of ESMs which results in this difference in GPP as seen in subplots (b) and (c) and in the case of model pair A-B in (d). Thus the JSD scaled in this manner gives us a way to actually compare the differences in input forcings of ESMs relative to their simulated GPP.

put space so we can effectively make inferences about whether relative orderings in input climate variable space are reflected in the GPP space as well.

We illustrate analysis based on the JSD in Figure 2 with four different possible use cases and how inferences can be made from them. Each sub figure shows the actual JSD in input (on the x-axis) and GPP (y-axis) space between three hypothetical models - A, B and C. The distances are then scaled by dividing all the distances in input space by the smallest such distance among all pairs of models. The distance in GPP space between that same pair of models is then used to scale all model pair distances in GPP space. This scaling allows us to effectively compare distances in input space vs GPP space. In subplot (a), we see that the relative ordering of distances between pairs of models is the same on both axes, the model pair A-B has the smallest distance in input space as well as GPP space while the model pair C-A has the largest distance in both these spaces. This provides some evidence that similarities or differences between pairs of models in the atmospheric forcing is also reflected in their GPP simulations. In (b), the distances in the atmospheric forcing are the same for all pairs of models but that's not the case in their GPP simulations where the distance between C-A is larger than the other pairs indicating possible differences in process representation across the models. In (c), the model pairs show larger differences in their input forcing but not in the simulated GPP space, indicating that despite having different climate, the models end up simulating very similar GPP values potentially differing in the processes involved in calculating GPP from these climate variables. Finally, in (d) we see another example for where proximity in input forcing does not translate to similar GPP simulations. In model pair A-B, differences lie more in simulated GPP than in the atmospheric forcing while the opposite is the case for model pairs C-A and B-C. We can thus use this analysis to attribute reasons for differences in GPP simulations between pairs of models.

The JSD measure was also used to determine how well the ML emulators estimate GPP by comparing the emulator estimated values with ESM simulations and we found that these distances tended to zero (results not shown). This further gives us confidence in our deployment of these ML emulators.

The ML emulators with Feature Selection, Jensen-Shannon Distance metric comparisons and more traditional analysis involving univariate statistics are all combined in our analysis of differences across ESMs in how they simulate GPP. Results from the analysis and a discussion on where the ML methods work well and where they don't is discussed in the next sections.

3 Results

In this section, we look at two key sets of results coming from the ML framework proposed in section 2.4. We first look at regional feature importances, that is, what the ML emulators determine to be the most relevant climate variable for GPP in a given region. We discuss results for regions in the JJA and DJF seasons as seen in Figures 3 and 4 but also provide results from the annual mean analysis for a more general overview in Supplementary Figure S2. We study the differences and similarities in GPP representation across pi-Control simulations in ESMs but due to the lack of observational datasets for this period, we use the literature on historical observations to guide our evaluation.

Our second set of results is from the comparison of relative distances between ESMs in the input climate space vs the GPP distribution space as described in Subsection 2.5 and shown in Figure 5. In our current analysis, we provide examples for how the JSD based comparisons can be useful as a tool to identify potential sources of differences in ESMs but leave more detailed region by region analysis for future work.

3.1 Model differences in relevant climate variables for GPP

Figures 3 and 4 show the most relevant climate variables for predicting GPP from two feature selection methods – Recursive Feature Elimination (RFE) and Permutation Importance (PI) in the first and second columns respectively. The RFE method's selection of best feature is considered the most relevant variable for GPP by the ML emulator and means that this variable is primarily responsible for estimating GPP. The PI method's selection on the other hand is more a measure of GPP's sensitivity to climate variables given the ML emulator. The most important climate variable could also be the variable GPP is most sensitive to, as in both methods could agree on the choice of climate variable(s) but differences are possible since the metrics involved are slightly different (low error vs best fit). ESM differences in the top features from the methods are considered an appropriate potential starting point for investigating divergence in GPP estimates from ESMs. We refer to the regions by their acronyms as defined in Iturbide et al. (2022) and are shown in Supplementary Figure S3 for reference.

Overall, all ESMs considered agree that temperature followed by precipitation are key variables for GPP for most of Europe, N.America and Asia. Over Africa and S.America, there is less of a consensus across ESMs and methods in accordance with previous analysis (Churkina & Running, 1998). Temperature is considered the most important variable for GPP in the Russian-Arctic (RAR) and Northern Europe (NEU) regions in JJA for most ESMs. Conditions of almost constant sunlight and water availability make temperature the key driver for GPP here. The northern N.American regions are a combination of arctic tundra and boreal forests and similarly show temperature as the main

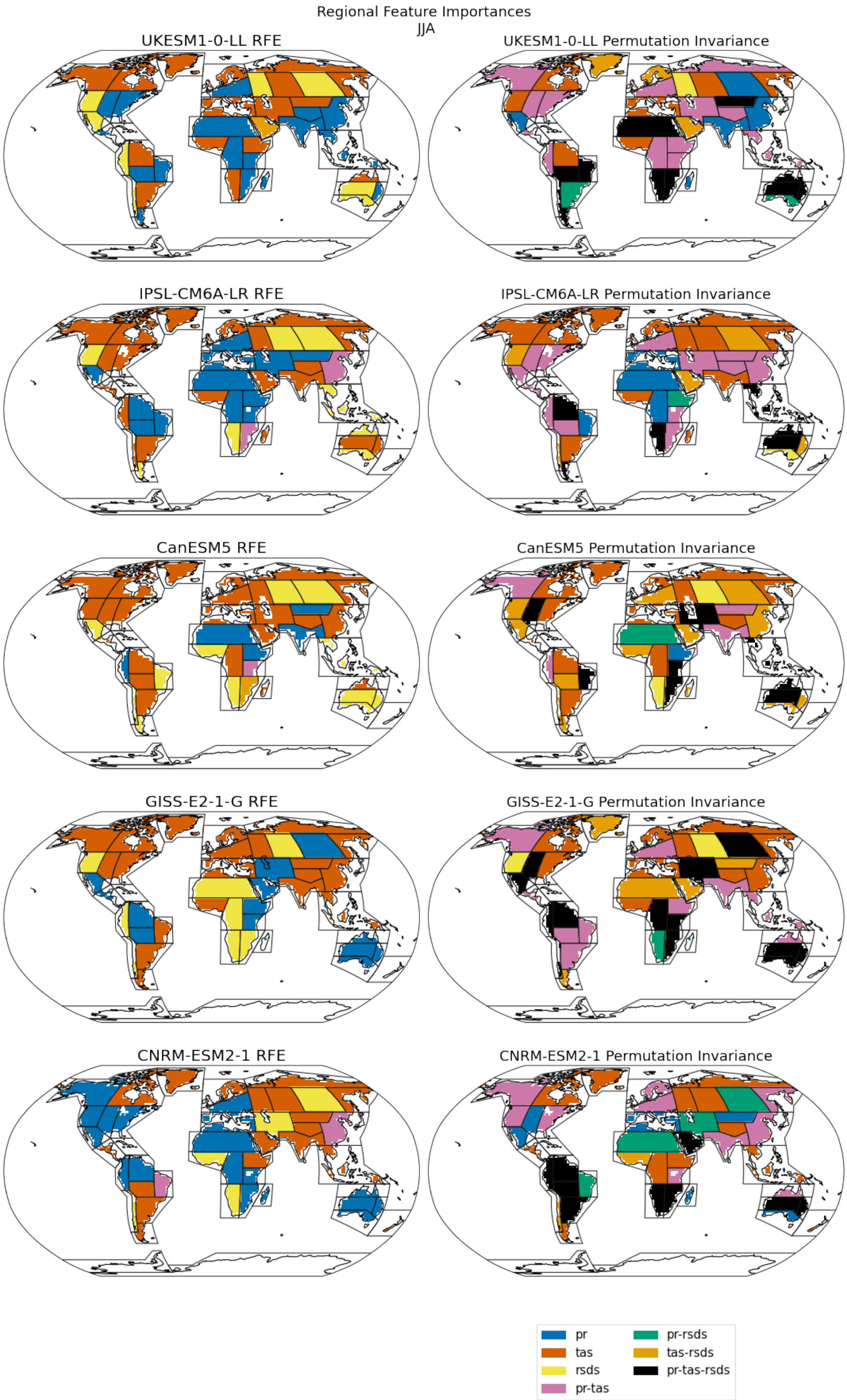


Figure 3. JJA feature importance from two methods - Recursive Feature elimination and Permutation Invariance for the IPCC regions defined in Iturbide et al. (2022).

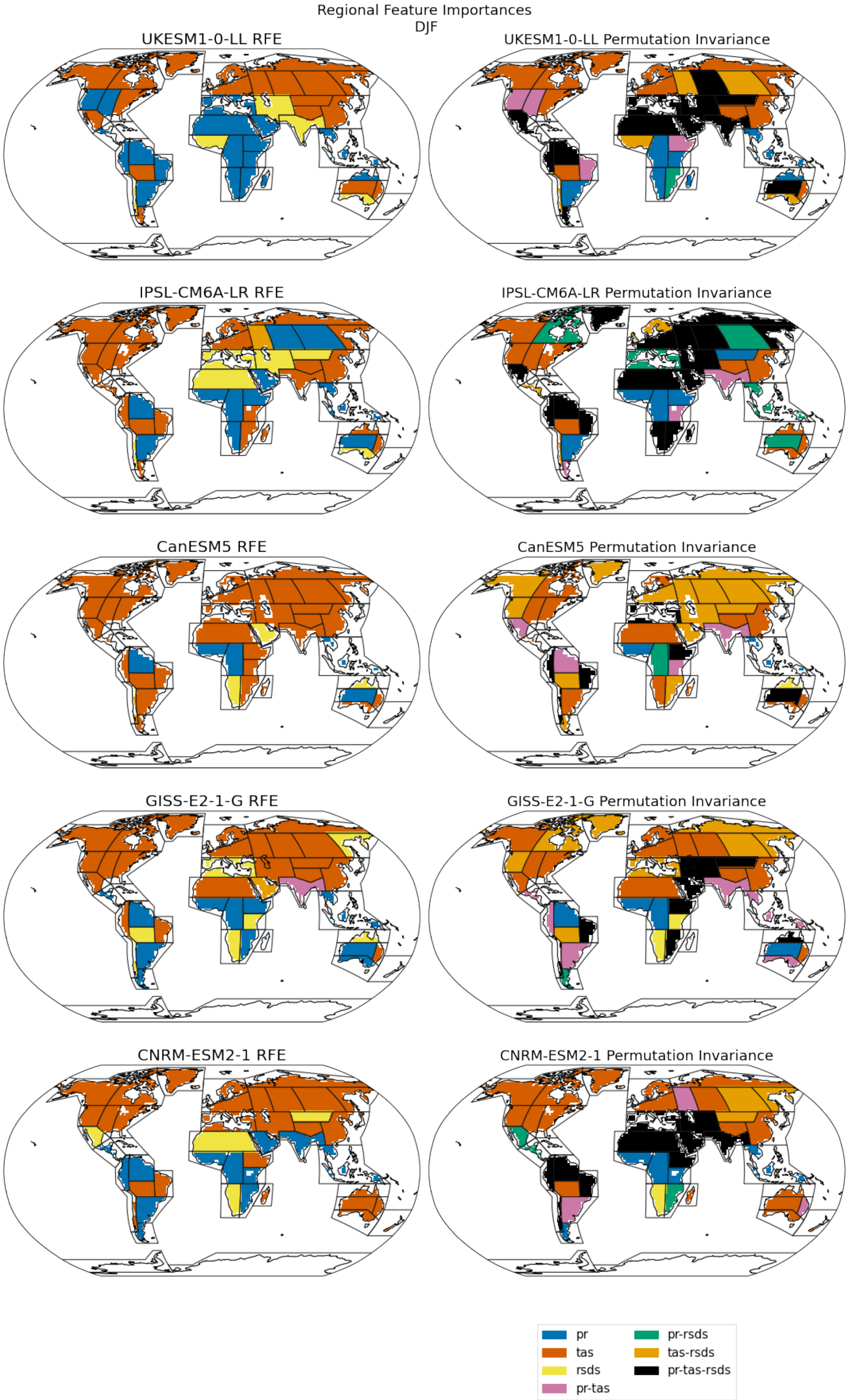


Figure 4. DJF feature importance from two methods - Recursive Feature elimination and Permutation Invariance for IPCC regions defined in Iturbide et al. (2022)

driving factor except for Northwestern North America (NWN) in CNRM-ESM2-1 where precipitation is determined as the key driver.

Boreal forest regions such as Eastern Europe (EEU), Western and Eastern Siberia (WSB, ESB) and the Russian Far East (RFE) show more divergence across ESMs with GPP being more dependent in both RFE and PI methods on temperature or radiation or both but in some instances (ESB for GISS-E2-1-G) on precipitation. In the central and eastern continental United States (CNA, ENA), UKESM1-0-LL and CNRM-ESM2-1 models consider precipitation to be most relevant for GPP while all other models find temperature more relevant. The variability in GPP is also dominated by a combination of these two variables as seen in the PI method. In the western north American region (WNA), radiation is seen as driving GPP except in CanESM5 (temperature) and CNRM-ESM2-1 (precipitation). In fact, precipitation seems to be most relevant for GPP in almost all N.American regions in the CNRM-ESM2-1 model and this can be considered as an indication that either the availability or the parameterization of this variable is important for GPP in this model more so than in others.

All ESMs in our study agree precipitation and temperature play a more important role than radiation in the Mediterranean region (MED), where radiation is largely available and a lack of rainfall or very high temperature is likely to influence vegetation more (Gea-Izquierdo et al., 2015). The CNRM1-ESM2-1 and IPSL-CM6A-LR are the two models that rank precipitation higher than temperature as an important feature. For the region covering the Indian subcontinent (SAS), precipitation is considered most important in the UKESM1-0-LL and CanESM5 models, consistent with previous studies (Varghese & Behera, 2019; Verma et al., 2022) while all three other models favor temperature as the key factor. In East Asia (EAS) temperature is considered the most important driver for GPP followed by precipitation and radiation in some regions (Yao et al., 2018; Bo et al., 2022) and all models except UKESM1-0-LL (precipitation) are in agreement.

In the DJF season, all models except CanESM5 consider precipitation most relevant for GPP in South East South America (SES) and all models agree that temperature is most relevant for Eastern Australia (EAU). We find the largest source of disagreement with regards to GPP drivers (looking at both DJF and JJA seasons) in regions where there is a significant presence of tropical forests such as Northern South America (NSA), Central-Africa (CAF), South-East Asia (SEA) and Northern Australia (NAU). We note radiation plays a role in some regions, possibly due to the lack of sufficient radiative energy available due to cloud cover which makes it hard to distinguish the relative importance between features. However almost all ESMs over a majority of these regions reference temperature and precipitation as key variables and from observational records we know that the two variables are strongly correlated in these regions (Nzabarinda et al., 2021; F. Zhang et al., 2022; Kanniah et al., 2011). Although precipitation appears most frequently as as the most important variable in determining GPP, especially using the RFE method of feature selection, in more than one instance all three features are considered relevant. This is consistent with results from previous studies using observations and non-ML approaches applied to finding GPP drivers (Churkina & Running, 1998; Kanniah et al., 2013; D. Wu et al., 2014). Another area where models lack consensus over the drivers is Southern Africa (ESAF and WSAF) for the DJF season. In reality, these areas are dominated by savannah, and are likely water limited but this is seen only in the UKESM1-0-LL model. Water limitaion effects on GPP in ESMs is typically modelled quite crudely, with uncertain parameterization (Harper et al., 2020) , and this is likely a significant source of disparity between the models.

3.2 Comparing differences in climate forcing vs GPP in model pairs

We compare ESM differences in the input feature space with their GPP distributions with the approach described in 2.5. In Figure 5 we show the comparative distances as a scatter plot to illustrate how we can potentially develop our hypotheses for quantifying and thus attributing differences in GPP to differences in climate forcing or process representation.

From the scatter plots in 5, we see differences across regions in how the pairwise model distances relate. If distances in input climate space between pairs of models translated to similar distances in GPP distributions, we would see the data points scattered along the diagonal unit slope line as seen in the NSA region. However this is not always the case, and we see more of a spread along the input space or x-axis (MED, RAR and somewhat also in SAS) where the plot indicates a spread in climate not quite seen in the simulated GPP and where relative differences in GPP are smaller than in input forcing . In other regions (SEA) however almost all pairs are above the unit slope line, which means that distances are larger in the GPP space.

We can use information from where there is a spread to investigate the likely causes underlying GPP divergence across models. In at least two regions (RAR and SAS), we notice that relative model distances with UKESM1-0-LL are greater in the y-axis even though such distances in the input space lie more or less in the middle range. This is an indication that the GPP simulated by UKESM1-0-LL is most different compared to other models even though not largely different in climate. In the SAS region for instance, the IPSL-CM6A-LR and UKESM1-0-LL models are closest in input space relative to other model pairs (seen as black colored letter I), and the CanESM5 model is identically distanced from both these models in the input space (seen as black and blue letters Ca). However, we see that in GPP space the UKESM1-0-LL distance with CanESM5 is more than the distance between CanESM5 and IPSL-CM6A-LR. Therefore one hypothesis worth investigating for this region is whether GPP process representation in IPSL-CM6A-LR and CanESM5 is similar in parameterization and different from UKESM1-0-LL. We would also include information from our feature importance results in 3 where we see that the two models differ in the variable considered most relevant for GPP (this is precipitation for UKESM1-0-LL, CanESM5 and temperature for IPSL-CM6A-LR). We argue that this type of analysis would be difficult to apply if we only consider univariate statistics as we show with examples in Supplementary Figure S4.

As a counter example, the ENA and to some extent the WSAF regions are examples of where it is not so clear how much of the difference in GPP to attribute to the influence of atmospheric forcing vs process representation from the scatter plot in Figure 5 due to close clustering in the relative distances.

4 Discussion

4.1 Choice of ML Approach for Evaluation

GPP is the largest individual carbon flux in the Earth System and changes to it have implications for the atmospheric carbon dioxide concentration, net carbon balance of the land surface and climate feedbacks (Friedlingstein et al., 2014). Interannual variability in GPP is influenced by changes in climate especially in hotspot regions such as tropical forests (O'Sullivan et al., 2020; Jung et al., 2011). Earth System Models provide the capability to simulate the Earth System's biogeochemical interactions and carbon cycle but global GPP estimates from ESMs vary greatly. For instance, in the five CMIP6 ESMs in our study, we found the global mean annual GPP to be in the range of 82-115 PgC year^{-1} for the pre-industrial period. The need to evaluate the carbon cycle in ESMs is thus critical for both better process representation and for modeling interactions with other components of the Earth System such as the atmosphere (Spafford & MacDougall, 2021; Reichler & Kim, 2008). Advances in Machine Learning and AI provides the algorithms that can help to facilitate evaluation of these complex interactions and uncover process based differences across ESMs (Huntingford et al., 2019). Our ap-

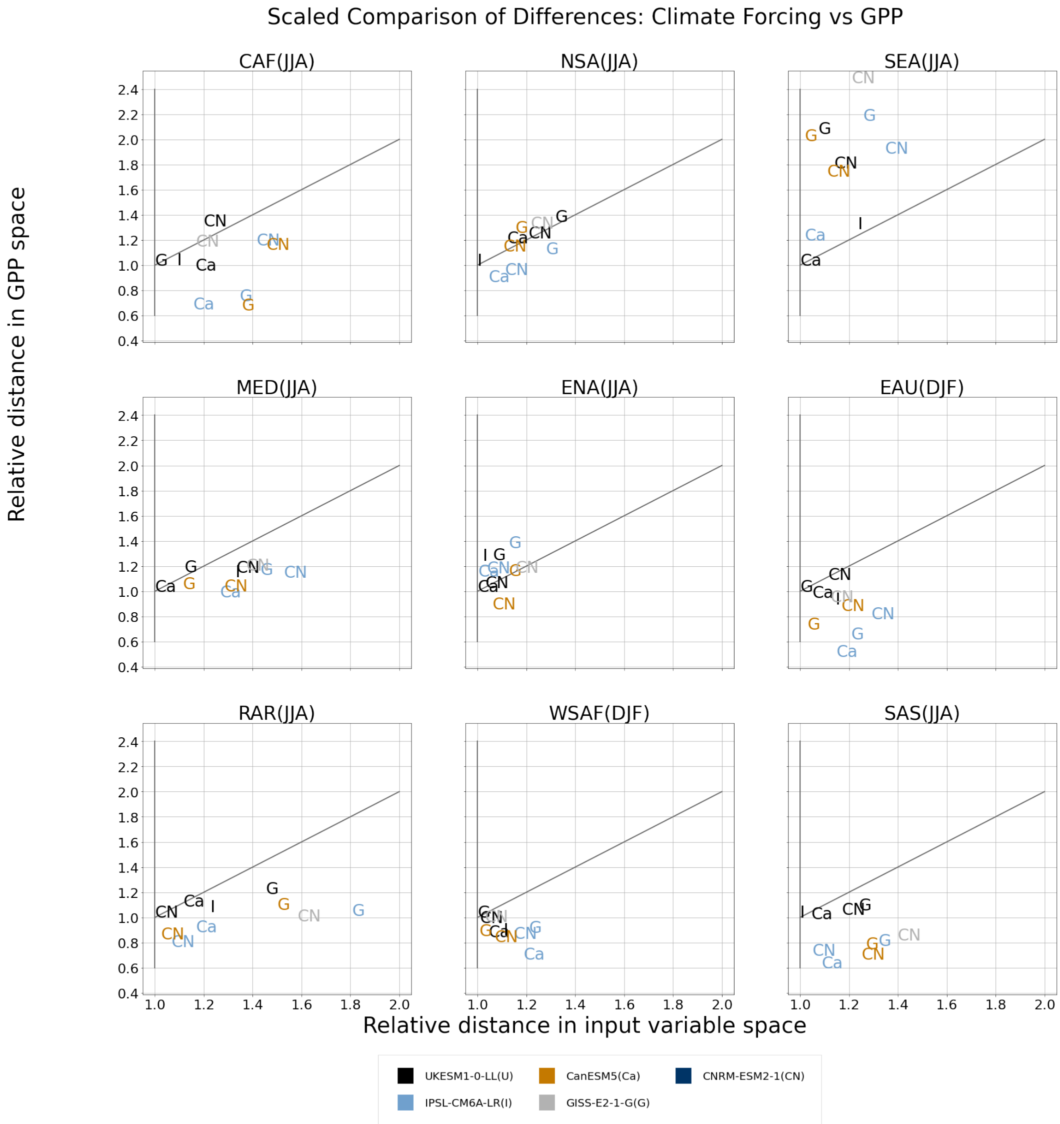


Figure 5. A comparison of relative distances in climate forcing and in GPP from different climate models is shown. Every model is referenced by both a color and an alphabet, the color and alphabet pairing tells us which pair of models are represented. Since the JSD is symmetric, there is only one colored symbol to show the distance between every pair of models. For this reason, there is no letter seen for the first model in the list, UKESM1-0-LL but its color (black) and letters for other models show the distance between UKESM1-0-LL and other models. For each region, the actual JSD values are scaled by factor that is the smallest distance in the input space across all pairs of models as seen in the x-axis and by the distance measure for that same pair in the GPP space as seen in the y-axis. This scaling follows from the description in Section 2 and Figure 2.

proach has been to start with the simplest ML models suited for our purpose. For this study, we build ML emulators with three input climate features to estimate GPP and for that emulator to be interpretable, which we demonstrate with our Feature Selection algorithms. Therefore, our ML emulators are not black boxes but can be interpreted in the context of physical and biogeochemical Earth System processes. We evaluated a choice of regression schemes before determining that Decision Trees best suited our task and further added better generalization capabilities with Boosting in the form of an Ensemble Learner with Adaboost. Such an emulator was capable of readily providing explanations on the modeled interactions between the atmospheric variables and GPP. At the same time, our framework is flexible enough for this emulator to be replaced with more complex ML algorithms such as Deep Architectures (LeCun et al., 2015) as we expand our suite of interacting variables for more nuanced evaluation of the carbon cycle. We further built robustness into our methods through rigorous cross validation and through the approaches outlined in Section 2.3 and demonstrate a reliable and adaptable framework that is also interpretable. With this framework, we were able to show regional similarities and differences in ESMs in the influence of key climate variables for GPP. Our emulator has the capability to capture non-linear relationships between the climate variables and GPP which can help to address limitations or complement more traditional approaches using correlations or calculated indices seen in the literature (O'Sullivan et al., 2020; Seddon et al., 2016).

The second component of our framework is a way to compare differences in climate variables influencing GPP with differences in processes estimating GPP in ESMs and we choose an algorithm based on the Jensen Shannon distance that is robust against small variations in distributions, allows a comparison of the joint input space with three variables and has bounds [0,1] to enable relative placement of distances. Also where a statistic such as a mean could be close for two different distributions, such as unimodal vs bimodal, the JSD will capture a difference in parameterization resulting in quite different distributions with similar means. Finally, our method enables a more flexible and less expensive way to perform this comparison where previously modeling experiments had to be conducted for similar analysis (Hardouin et al., 2022).

4.2 Application of ML framework for GPP Evaluation

The ML framework described in this paper can be used to identify areas of differences in GPP modeling in ESMs. For instance, from Figure 4 and Figure 3, we see that while models have overall agreement on what variables are important for certain regions (temperature and precipitation for the Mediterranean, South Asia, Eastern and Central North America; temperature and radiation in the tundra and boreal forest regions) differences exist in the which individual climate variable matters for a given ESM. Further the comparison using JSD gives us a starting point for whether these differences are more in the state of the climate influencing GPP or in the processing of these variables such as through parameterizations. This ML framework can serve as a guide to investigate probable reasons why differences in GPP modeling exist in ESMs in a computationally less expensive manner to actually running model simulations.

4.3 Limitations and Challenges

In our current study, we sample data uniformly from the spatio-temporal domain which does not capture sub-regional and sub-seasonal variability and trends. This limitation is mainly driven by the lack of availability of GPP data from CMIP6 ESMs at higher temporal resolutions for the pi-Control experiment. However, this is more a feature of the data used and our framework will allow us to experiment with different resolutions in data when available. The JSD approach provides a relatively inexpensive method, without actually having to run model simulations, to compare differences across models in GPP vs climate variables but in some regions such as Eastern North America (ENA) seen in Figure 5, it is harder to infer where the differences lie. Along with future work to develop this analysis, we also suggest that individual components of the ML framework as well as more traditionally considered descriptive statistics such as means and variability should all be used in a complementary fashion in the evaluation process so we can take insights from different modes of analysis. Finally, the three predictor variables were chosen because of their importance in determining the supply of water (precipitation), its loss through evapotranspiration (temperature) and the available energy for photosynthesis (shortwave radiation). We recognize the need to include a broader suite of variables for a more holistic evaluation of the carbon cycle which is possible to do with our framework.

5 Conclusions

This study demonstrates the potential of using interpretable ML approaches to investigate differences in GPP modeling across a selection of CMIP6 models and over land regions defined in the IPCC's Sixth Assessment Report and two seasons. In conclusion:

1. The relative importance of key climate drivers for GPP was identified across different regions and ESMs using Feature Selection Methods with interpretable ML emulators. We illustrate this with examples such as the Mediterranean region where all models agree that drought variables such as temperature or precipitation influence GPP more than radiation but models differ in which of the two variables is most relevant.
2. With a comparative distance metric based on the Jensen Shannon Distance, we are able to show that proximity or distance in climate between any two models does not necessarily translate to a similar proximity or distance in their estimated GPP distributions with the Russian Arctic (RAR) and Mediterranean regions (MED) as two such examples. We take this as evidence that process based differences exist across models and are at least partly responsible for differences in GPP estimates from ESMs.
3. Where the JSD method suggests divergence in GPP potentially due to process modeling, for instance in South Asia (SAS) between the UKESM1-0-LL, IPSL-CM6A-LR and CanESM5 models, the Feature Selection process can offer an explanation. In this case the UKESM1-0-LL and IPSL-CM6A-LR models differ in the key climate variable for GPP but the UKESM1-0-LL and CanESM5 models don't and a possible reason for this can be differences in parameterization or characteristics of this variable not considered in the input features.
4. There are some regions where models do not show a clear consensus on what climate variables matter the most or identify all three variables as equally important such as the tropics. Similarly our distance metric based comparison also presents cases where a direct inference on attributing GPP differences cannot be made, such as the Eastern North American (ENA) region. We identify these as regions of uncertainty to address in future work.

Data from the pre-industrial Control experiments served as a baseline for the development of this evaluation framework. In future work, additional climate drivers and characteristics such as sub-monthly variability will also be incorporated as possible causes for variations in GPP estimates from ESMs and analysis will be conducted with data from historical experiments and observations towards the goal of improving vegetation modeling in Earth System Models.

6 Open Research

Data from CMIP6 climate models is available for download on Earth System Grid Federation nodes and were downloaded and preprocessed using the open source software

ESMValTool v2.8.0 (doi:10.5281/zenodo.3401363) and ESMValCore v2.8.0 (doi:10.5281/zenodo.3387139). Code used to produce the results in this paper is available under the CC-BY license at the Github respository (https://github.com/rswamina/gpp-ml-eval-1-publish) which is currently private but will be made public once the manuscript has been accepted for publication.

Acknowledgments

RS and TQ are funded by the UK Research Infrastructure Natural Environment Research Council (UKRI-NERC) funded TerraFIRMA: Future Impacts, Risks and Mitigation Actions in a changing Earth system grant (NE/W004895/1). RA is funded by the National Centre for Earth Observation grant: NE/RO16518/1.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., … others (2015). Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, *53*(3), 785–818. doi: 10.1002/2015RG000483

Bo, Y., Li, X., Liu, K., Wang, S., Zhang, H., Gao, X., & Zhang, X. (2022). Three decades of gross primary production (gpp) in china: Variations, trends, attributions, and prediction inferred from multiple datasets and time series modeling. *Remote Sensing*, *14*(11), 2564. doi: 10.3390/rs14112564

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., … others (2020). Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. doi: 10.1029/2019MS002010

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140. doi: /10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi: doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Cart. *Classification and regression trees*.

Churkina, G., & Running, S. W. (1998). Contrasting climatic controls on the estimated productivity of global terrestrial biomes. *Ecosystems*, *1*(2), 206–215. doi: 10.1007/s100219900016

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., … others (2014). The earth system grid federation: An open infrastructure for access to distributed geospatial data. *Future Generation Computer Systems*, *36*, 400–417. doi: 10.1016/j.future.2013.07.002

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M., … others (2011). The joint uk land environment simulator (jules), model description–part 2: carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701–722. doi: 10.5194/gmd-4-701-2011

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, 146–158. doi: 0.1214/aop/1176996454

Delire, C., Séférian, R., Decharme, B., Alkama, R., Calvet, J.-C., Carrer, D., … others (2020). The global land carbon cycle simulated with isba-ctrip: Improvements over the last decade. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS001886. doi: 10.1029/2019MS001886

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of machine learning research*, *3*, 1265–1287. doi: 10.5555/944919.944973

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. doi: 10.48550/arXiv.1702

.08608
Drucker, H. (1997). Improving regressors using boosting techniques. In *Icml* (Vol. 97, pp. 107–115). doi: 10.5555/645526.657132
Dunkl, I., Lovenduski, N., Collalti, A., Arora, V. K., Ilyina, T., & Brovkin, V. (2023). Gross primary productivity and the predictability of co 2: more uncertainty in what we predict than how well we predict it. *Biogeosciences*, *20*(16), 3523–3538. doi: 10.5194/bg-20-3523-2023
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. doi: 10.5194/gmd-9-1937-2016
Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2018MS001453. doi: 10.1029/2018MS001453
Fisher, R. A., Koven, C. D., Anderegg, W. R., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., ... others (2018). Vegetation demographics in earth system models: A review of progress and priorities. *Global change biology*, *24*(1), 35–54. doi: 10.1111/gcb.13910
Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in cmip5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, *27*(2), 511–526. doi: /10.1175/JCLI-D-12-00579.1
Gea-Izquierdo, G., Guibal, F., Joffre, R., Ourcival, J., Simioni, G., & Guiot, J. (2015). Modelling the climatic drivers determining photosynthesis and carbon allocation in evergreen mediterranean forests using multiproxy long time series. *Biogeosciences*, *12*(12), 3695–3712. doi: 10.5194/bg-12-3695-2015
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. doi: 10.1145/3422622
Greenslade, M., Murphy, S., Treshansky, A., DeLuca, C., Guilyardi, E., & Denvil, S. (2014). The earth system (es-doc) project. In *Egu general assembly conference abstracts* (p. 12988).
Gültas, M., Düzgün, G., Herzog, S., Jäger, S. J., Meckbach, C., Wingender, E., & Waack, S. (2014). Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum jensen-shannon divergence and cuda programming. *BMC bioinformatics*, *15*(1), 1–17. doi: 10.1186/1471-2105-15-96
Gutiérrez, J. M., Jones, R., Narisma, G., et al. (2021). Ipcc interactive atlas. In *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change.* Cambridge University Press Cambridge. doi: 10.1175/BAMS-D-20-0256.1
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182. doi: 10.5555/944919.944968
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389–422. doi: 10.1023/A:1012487302797
Hardouin, L., Delire, C., Decharme, B., Lawrence, D. M., Nabel, J. E., Brovkin, V., ... others (2022). Uncertainty in land carbon budget simulated by terrestrial biosphere models: the role of atmospheric forcing. *Environmental Research Letters*, *17*(9), 094033. doi: 10.1088/1748-9326/ac888d
Harper, A. B., Williams, K. E., McGuire, P. C., Duran Rojas, M. C., Hemming, D., Verhoef, A., ... others (2020). Improvement of modelling plant responses

to low soil moisture in julesvn4. 9 and evaluation against flux tower measurements. *Geoscientific Model Development Discussions*, *2020*, 1–42. doi: gmd-14-3269-2021

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. doi: 10.1088/1748-9326/ab4e55

Iturbide, M., Fernández, J., Gutiérrez, J. M., Pirani, A., Huard, D., Al Khourdajie, A., ... others (2022). Implementation of fair principles in the ipcc: the wgi ar6 atlas repository. *Scientific data*, *9*(1), 629. doi: 0.5281/zenodo.3691645

James, G., Witten, D., Hastie, T., Tibshirani, R., James, G., Witten, D., ... Tibshirani, R. (2021). Linear regression. *An introduction to statistical learning: with applications in R*, 59–128. doi: 10.1007/978-1-0716-1418-1_3

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... Reichstein, M. (2019). The fluxcom ensemble of global land-atmosphere energy fluxes. *Scientific data*, *6*(1), 1–14. doi: 10.1038/s41597-019-0076-8

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., ... others (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, *116*(G3). doi: 10.1029/2010JG001566

Kanniah, K. D., Beringer, J., & Hutley, L. (2013). Exploring the link between clouds, radiation, and canopy productivity of tropical savannas. *Agricultural and Forest Meteorology*, *182*, 304–313. doi: 10.1016/j.agrformet.2013.06.010

Kanniah, K. D., Beringer, J., & Hutley, L. B. (2011). Environmental controls on the spatial variability of savanna productivity in the northern territory, australia. *Agricultural and Forest Meteorology*, *151*(11), 1429–1439. doi: 10.1016/j.agrformet.2011.06.009

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., ... others (2020). Giss-e2. 1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025. doi: 10.1029/2019MS002025

Kiang, N. (2012). *Description of the nasa giss vegetation dynamics model* (Tech. Rep.). Tech. rep., NASA.

Kim, D., Lee, M.-I., Jeong, S.-J., Im, J., Cha, D. H., & Lee, S. (2018). Intercomparison of terrestrial carbon fluxes and carbon use efficiency simulated by cmip5 earth system models. *Asia-Pacific Journal of Atmospheric Sciences*, *54*(2), 145–163. doi: 10.1007/s13143-017-0066-8

Koch, A., Hubau, W., & Lewis, S. L. (2021). Earth system models are not capturing present-day tropical forest carbon dynamics. *Earth's Future*, *9*(5), e2020EF001874. doi: /10.1029/2020EF001874

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, *19*(1). doi: 10.1029/2003GB002199

Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, *4*(3), 211–229. doi: 10.6029/smartcr.2014.03.007

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444. doi: doi.org/10.1038/nature14539

Levis, S. (2010). Modeling vegetation and land use in models of the earth system. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(6), 840–856. doi: 10.1002/wcc.83

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151. doi: 10.1109/18.61115

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, *45*(1), 1–40. doi: 10.1145/2379776.2379786

Molnar, C. (2020). *Interpretable machine learning.* Lulu. com.

Nishina, K., Ito, A., Falloon, P., Friend, A., Beerling, D., Ciais, P., ... others (2015). Decomposing uncertainties in the future terrestrial carbon budget associated with emission scenarios, climate projections, and ecosystem simulations using the isi-mip results. *Earth System Dynamics*, *6*(2), 435–445. doi: 10.5194/esd-6-435-2015

Nzabarinda, V., Bao, A., Xu, W., Uwamahoro, S., Jiang, L., Duan, Y., ... Long, G. (2021). Assessment and evaluation of the response of vegetation dynamics to climate variability in africa. *Sustainability*, *13*(3), 1234. doi: 10.3390/su13031234

on Climate Change, I. P. (2023). Global carbon and other biogeochemical cycles and feedbacks. In *Climate change 2021 – the physical science basis: Working group i contribution to the sixth assessment report of the intergovernmental panel on climate change* (p. 673–816). Cambridge University Press. doi: 10.1017/9781009157896.007

O'Sullivan, M., Smith, W. K., Sitch, S., Friedlingstein, P., Arora, V. K., Haverd, V., ... others (2020). Climate-driven variability and trends in plant productivity over recent decades based on three global products. *Global Biogeochemical Cycles*, *34*(12), e2020GB006613. doi: 10.1029/2020GB006613

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. doi: 10.5555/1953048.2078195

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106. doi: doi.org/10.1007/BF00116251

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–312. doi: 10.1175/BAMS-89-3-303

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., ... others (2020). Earth system model evaluation tool (esmvaltool) v2. 0–technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199.

Santini, M., Collalti, A., & Valentini, R. (2014). Climate change impacts on vegetation and water cycle in the euro-mediterranean region, studied by a likelihood approach. *Regional Environmental Change*, *14*(4), 1405–1418.

Sarkar, D. P., Shankar, B. U., & Parida, B. R. (2022). Machine learning approach to predict terrestrial gross primary productivity using topographical and remote sensing data. *Ecological Informatics*, *70*, 101697. doi: 10.1016/j.ecoinf.2022.101697

Schapire, R. E. (2013). Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37–52. doi: 10.1007/978-3-642-41136-6_5

Schimel, D. S., House, J. I., Hibbard, K. A., Bousquet, P., Ciais, P., Peylin, P., ... others (2001). Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, *414*(6860), 169–172. doi: 10.1038/35102500

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., & Reichstein, M. (2020). Constraining uncertainty in projected gross primary production with machine learning. *Journal of Geophysical Research: Biogeosciences*, *125*(11), e2019JG005619. doi: 10.1029/2019JG005619

Schwalm, C. R., Huntzinger, D. N., Michalak, A. M., Schaefer, K., Fisher, J. B., Fang, Y., & Wei, Y. (2020). Modeling suggests fossil fuel emissions have been driving increased land carbon uptake since the turn of the 20th century. *Scientific Reports*, *10*(1), 9059. doi: 10.1038/s41598-020-66103-9

Seddon, A. W., Macias-Fauria, M., Long, P. R., Benz, D., & Willis, K. J. (2016). Sensitivity of global terrestrial ecosystems to climate variability. *Nature*, *531*(7593), 229–232. doi: 10.1038/nature16986

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., ... others (2019). Evaluation of cnrm earth system model, cnrm-esm2-1: role of earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4182–4227. doi: 10.1029/2019MS001791

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., ... others (2019). Ukesm1: Description and evaluation of the uk earth system model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. doi: 10.1029/2019MS001739

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Spafford, L., & MacDougall, A. H. (2021). Validation of terrestrial biogeochemistry in cmip6 earth system models: a review. *Geoscientific Model Development*, *14*(9), 5863–5889. doi: gmd-14-5863-2021

Sun, Y., Frankenberg, C., Wood, J. D., Schimel, D., Jung, M., Guanter, L., ... others (2017). Oco-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. *Science*, *358*(6360), eaam5747. doi: 10.1126/science.aam57

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., ... others (2019). The canadian earth system model version 5 (canesm5. 0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. doi: 10.5194/gmd-12-4823-2019

Varghese, R., & Behera, M. (2019). Annual and seasonal variations in gross primary productivity across the agro-climatic regions in india. *Environmental monitoring and assessment*, *191*(10), 631. doi: 10.1007/s10661-019-7796-2

Verma, A., Chandel, V., & Ghosh, S. (2022). Climate drivers of the variations of vegetation productivity in india. *Environmental Research Letters*, *17*(8), 084023. doi: 10.1088/1748-9326/ac7c7f

Verseghy, D. (2012). Class–the canadian land surface scheme (version 3.6). *Environment Canada Science and Technology Branch Tech. Rep*, *176*.

Wu, D., Zhao, X., Zhao, W., Tang, B., & Xu, W. (2014). Response of vegetation to temperature, precipitation and solar radiation time-scales: A case study over mainland australia. In *2014 ieee geoscience and remote sensing symposium* (pp. 855–858). doi: 10.1109/IGARSS.2014.6946559

Wu, Z., Ahlström, A., Smith, B., Ardö, J., Eklundh, L., Fensholt, R., & Lehsten, V. (2017). Climate data induced uncertainty in model-based estimations of terrestrial primary productivity. *Environmental Research Letters*, *12*(6), 064013. doi: 10.1088/1748-9326/aa6fd8

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. doi: /10.1016/j.neucom.2020.07.061

Yao, Y., Wang, X., Li, Y., Wang, T., Shen, M., Du, M., ... others (2018). Spatiotemporal pattern of gross primary productivity and its covariation with climate in china over the last thirty years. *Global Change Biology*, *24*(1), 184–196.

Yu, T., Zhang, Q., & Sun, R. (2021). Comparison of machine learning methods to up-scale gross primary production. *Remote Sensing*, *13*(13), 2448. doi: 10.3390/rs13132448

Zampieri, M., Grizzetti, B., Toreti, A., De Palma, P., & Collalti, A. (2021). Rise and fall of vegetation annual primary production resilience to climate variability projected by a large ensemble of earth system models' simulations. *Environmental Research Letters*, *16*(10), 105001. doi: 10.1088/1748-9326/ac2407

Zarakas, C. M., Swann, A. L., Laguë, M. M., Armour, K. C., & Randerson, J. T. (2020). Plant physiology increases the magnitude and spread of the transient climate response to co2 in cmip6 earth system models. *Journal of Climate*, *33*(19), 8561–8578. doi: 10.1175/JCLI-D-20-0078.1

Zhang, F., Lu, X., Huang, Q., & Jiang, F. (2022). Impact of different era reanalysis data on gpp simulation. *Ecological Informatics*, *68*, 101520. doi: 10.1016/j.ecoinf.2021.101520

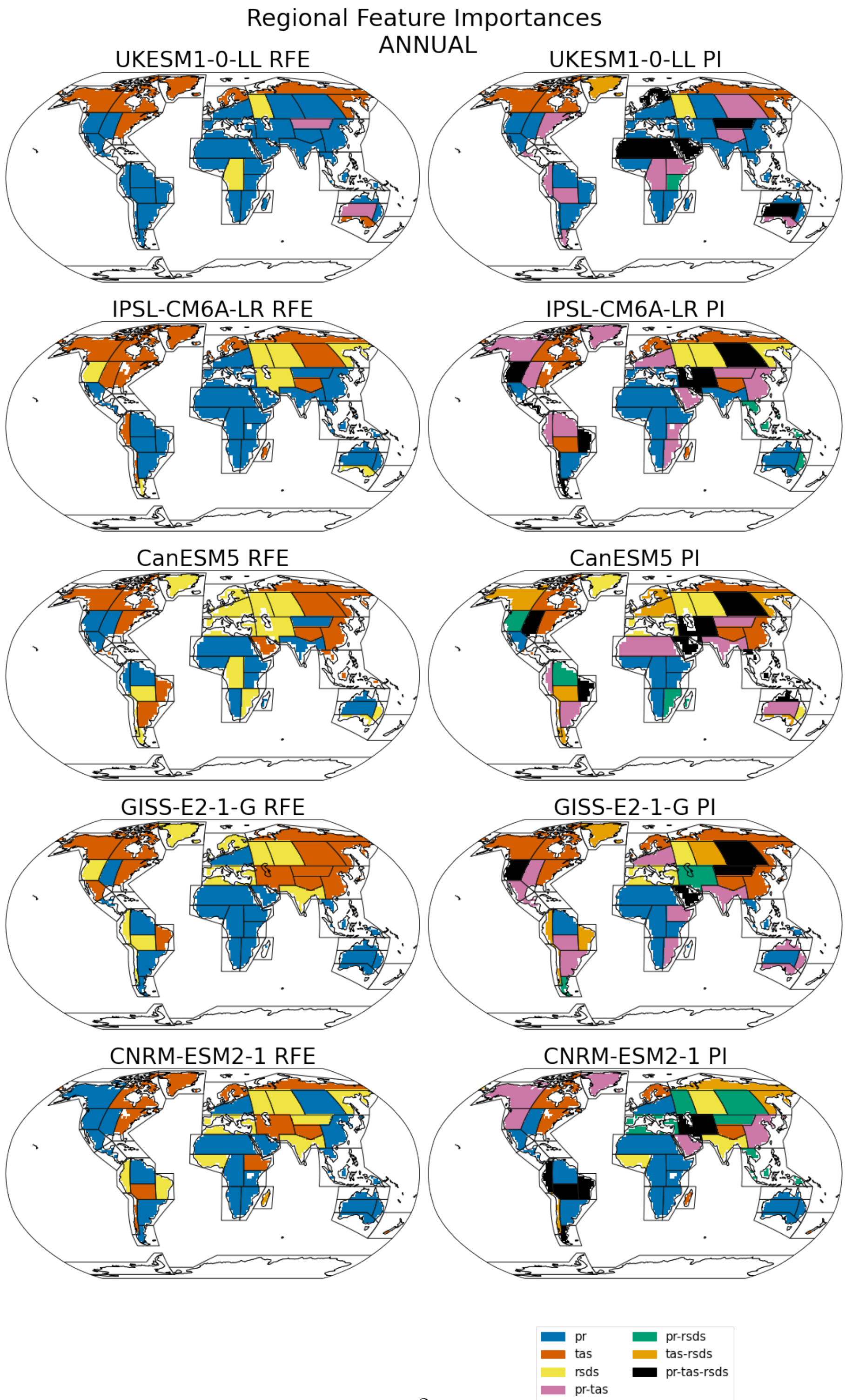
Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S., & Gentine, P. (2018). A global spatially contiguous solar-induced fluorescence (csif) dataset using neural networks. *Biogeosciences*, *15*(19), 5779–5800. doi: 10.5194/bg-15-5779-2018

Zhang, Z., Xin, Q., & Li, W. (2021). Machine learning-based modeling of vegetation leaf area index and gross primary productivity across north america and comparison with a process-based model. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002802. doi: 10.1029/2021MS002802

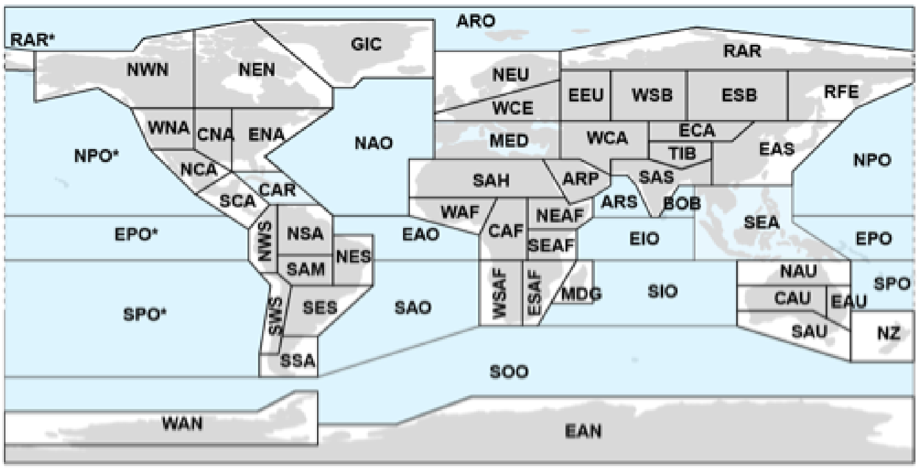
Evaluating Vegetation Modelling in Earth System Models with Machine Learning Approaches (Supplementary Figures for submission to the Journal of Advances in Modeling Earth Systems (JAMES))

Ranjini Swaminathan Tristan Quaife Richard Allan

S1: Gross Primary Productivity values estimate by the ML emulator for a selection of IPCC regions. Every column shows the difference between the ML emulator output and the GPP simulated by a given ESM. The RMSE error is shown at the top of each region alongwith the difference in area averaged mean between the ML emulator estimates and the ESM sumlated values. All units are in g/m2/month.

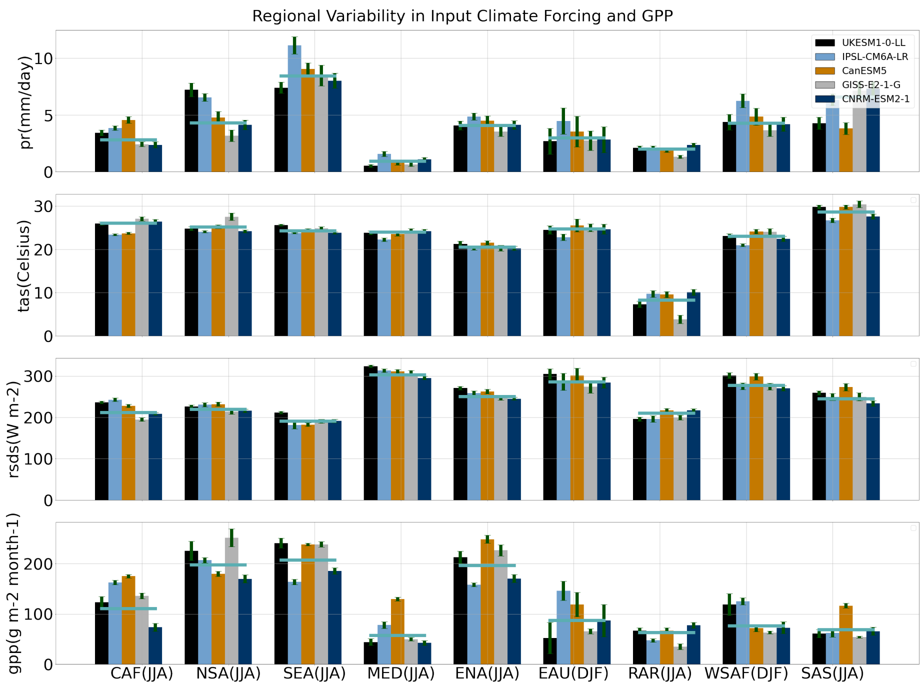


S2: Annual feature importance from two methods - Recursive Feature elimination and Permutation Invariance for IPCC regions.



1	GIC	Greenland/Iceland	21	SAH	Sahara	40	NAU	N.Australia
2	NWN	N.W.North-America	22	WAF	Western-Africa	41	CAU	C.Australia
3	NEN	N.E.North-America	23	CAF	Central-Africa	42	EAU	E.Australia
4	WNA	W.North-America	24	NEAF	N.Eastern-Africa	43	SAU	S.Australia
5	CNA	C.North-America	25	SEAF	S.Eastern-Africa	44	NZ	New-Zealand
6	ENA	E.North-America	26	WSAF	W.Southern-Africa	45	EAN	E.Antarctica
7	NCA	N.Central-America	27	ESAF	E.Southern-Africa	46	WAN	W.Antarctica
8	SCA	S.Central-America	28	MDG	Madagascar	47	ARO	Arctic-Ocean
9	CAR	Caribbean	29	RAR	Russian-Arctic	48	NPO	N.Pacific-Ocean
10	NWS	N.W.South-America	30	WSB	W.Siberia	49	EPO	Equatorial.Pacific-Ocean
11	NSA	N.South-America	31	ESB	E.Siberia	50	SPO	S.Pacific-Ocean
12	NES	N.E.South-America	32	RFE	Russian-Far-East	51	NAO	N.Atlantic-Ocean
13	SAM	South-American-Monsoon	33	WCA	W.C.Asia	52	EAO	Equatorial.Atlantic-Ocean
14	SWS	S.W.South-America	34	ECA	E.C.Asia	53	SAO	S.Atlantic-Ocean
15	SES	S.E.South-America	35	TIB	Tibetan-Plateau	54	ARS	Arabian-Sea
16	SSA	S.South-America	36	EAS	E.Asia	55	BOB	Bay-of-Bengal
17	NEU	N.Europe	37	ARP	Arabian-Peninsula	56	EIO	Equatorial.Indic-Ocean
18	WCE	Western&Central-Europe	38	SAS	S.Asia	57	SIO	S.Indic-Ocean
19	EEU	E.Europe	39	SEA	S.E.Asia	58	SOO	Southern-Ocean
20	MED	Mediterranean						

S3: IPCC AR 6 reference regions and their acronyms.



S4: A comparison of means and standard deviations of the climate variables or input forcings considered important for GPP. Each row shows the mean and standard deviation for a single variable with colored bars representing individual models grouped by regions. Vertical lines overlayed on the colored bars shows the standard deviation and the horizontal line shows the multimodel mean.