Mapping out the yields of energy crops with data-driven global models, including climate and soil predictors

Siwar Saadaoui, David Makowski, Benoît Gabrielle, Thierry Brunelle

GCB Bioenergy

WILEY

Global Change Biology Bioenergy
BIOPRODUCTS FOR A SUSTAINABLE BIOECONOMY

RESEARCH ARTICLE OPEN ACCESS

Mapping Out the Yields of Energy Crops With Data-Driven Global Models, Including Climate and Soil Predictors

Siwar Saadaoui[1,2] | David Makowski[3] | Benoît Gabrielle[2] | Thierry Brunelle[1]

[1]CIRED, Nogent-sur-Marne, France | [2]INRAE, AgroParisTech, UMR ECOSYS, Université Paris-Saclay, Palaiseau, France | [3]INRAE, AgroParisTech, UMR MIA-PS, Université Paris-Saclay, Palaiseau, France

Correspondence: Siwar Saadaoui (siwar.saadaoui@agroparistech.fr) | Benoît Gabrielle (benoit.gabrielle@agroparistech.fr)

Received: 22 April 2025 | **Revised:** 24 July 2025 | **Accepted:** 8 August 2025

Funding: This study was supported by the Carma Chair and the French Agence Nationale de la Recherche CLAND (ANR-16-CONV-0003).

Keywords: bioenergy | energy crops | global maps | lignocellulosic crops | machine learning | modeling

ABSTRACT

Lignocellulosic crops such as Miscanthus, Eucalyptus, Poplar, Willow, and Switchgrass are gaining attention as promising feedstocks for renewable energy and carbon-mitigation strategies, especially on marginal lands. Assessing their global yield potentials requires models that go beyond climate drivers alone. Using a global dataset of 3963 yield observations for five species, we developed a high-resolution (5-arc-minute) modeling framework that augments climate with detailed soil and topographic predictors. Among seven machine learning algorithms, Random Forest, Extra Trees, and Gradient Boosting (GB) emerged as top performers. On an independent test set, the best model achieved a root mean square error (RMSE) of 4.8 t DM ha^{-1} $year^{-1}$ (across algorithms: 4.7–5.0 t DM ha^{-1} $year^{-1}$) and an R^2 of 0.67, a moderate error relative to the broad 4–19 t DM ha^{-1} $year^{-1}$ spatial yield range. After outlier handling via a two-phase cross-validation procedure, each model was applied globally under current climate and three future scenarios (SSP1-2.6, SSP2-4.5, and SSP5-8.5). Across scenarios (relative to the 1980–2000 baseline), median absolute yield changes over suitable land are modest (ca. 1–2 t DM ha^{-1} $year^{-1}$), yet localized hotspots show gains or losses up to 8 t DM ha^{-1} $year^{-1}$. Yields most often increase in presently cool, high-latitude areas and decrease in warmer/drier or edaphically constrained low-latitude regions. We additionally provide a "best-crop" map identifying where each species may offer the most favorable balance between yield and production cost, revealing pronounced geographic variation in suitability. Compared with alternative models based on coarser-resolution datasets, our approach generally yields more conservative estimates, likely reflecting the added constraint from soil and topographic predictors. These results underscore the importance of representing local environmental heterogeneity when predicting energy-crop productivity under climate change.

1 | Introduction

The transition to low-carbon energy systems hinges on bioenergy development in general and dedicated lignocellulosic crops in particular (Popp et al. 2017; Taylor et al. 2019). Species such as Miscanthus, Eucalyptus, Poplar, Willow, and Switchgrass offer several advantages: They can be grown on marginal lands without directly competing with food crops, require low amounts of chemical inputs while achieving high biomass yields (Winkler et al. 2020; Gopalakrishnan et al. 2011; Qin et al. 2015; W. Li et al. 2018). Integrated assessment models (IAMs), which were extensively used in the Intergovernmental Panel on Climate Change's Sixth Assessment Report (IPCC 2023), commonly incorporate estimates of bioenergy potentials to project global mitigation pathways. However, these models estimate the yields of lignocellulosic crops based on climatic drivers and neglect soil and topographic heterogeneities, which significantly

affect biomass productivity (Haberzettl et al. 2021; Pappas et al. 2015). Consequently, current yield estimates are likely biased for these crops (Calvin et al. 2021), raising concerns over the robustness of policy or investment decisions based on their outcomes.

Data-driven models offer an alternative approach to map global yields of lignocellulosic crops at large scales. Li et al. (2020) used machine learning models using mostly climate-related predictors at a coarse resolution (30 arc minutes) and one soil property (clay content). Although climate predictors are helpful for high-level assessments, they ignore spatial variability of soil and land characteristics, particularly in regions with complex soils, steep slopes, or fragmented landscapes (Haberzettl et al. 2021). Accounting for local edaphic constraints and topographic limitations is paramount to projecting crop yields and designing sustainable land-use strategies as these features can critically shape the viability of biomass crop cultivation.

Recent research suggests that integrating soil and topographic data can significantly improve yield estimates, revealing previously unrecognized "hot spots" of high productivity and areas at elevated risk of crop failure (Franz et al. 2020). However, global studies incorporating these variables at fine spatial scales are still scarce, partly due to data availability and computational constraints. In addition, local site conditions may interact with climate in complex ways. For example, improvements in temperature or precipitation might be offset by poor soils or steep terrain, underscoring the need for new modeling approaches.

Given these challenges, this study seeks to advance global yield assessments for five lignocellulosic crops by explicitly integrating detailed soil and topographic factors alongside climate data. Working at a 5 arc-minute resolution, we employ seven machine learning algorithms—Random Forest (RF), Extra Trees (ET), GB, LightGBM, SVM, Decision Tree, and MLP Regressor—to predict yields under both current and future climate scenarios (IPCC SSP1-2.6, SSP2-4.5, and SSP5-8.5). Following a multi-phase cross-validation procedure with outlier filtering, we select the top three performing models to generate high-resolution yield maps and "best crop" indices that factor in production costs. Finally, we compare our estimates to those of Li et al. (2020) and align our discussion with IAM-based yield potentials to assess the impact of finer-scale edaphic-topographic integration on yield forecasts.

2 | Materials and Methods

2.1 | Data Used to Train and Test the Models

We obtained global yield measurements for five lignocellulosic crops (Eucalyptus, Miscanthus, Poplar, Willow, and Switchgrass) from the dataset assembled by W. Li et al. (2018). This dataset includes 3963 field observations collected across 31 countries from 1980 to 2017, documenting observed yields and their respective coordinates. To assess the spatial representativeness of the 3963 field observations, each point was intersected with (i) the 1-km Köppen–Geiger climate raster (Beck et al. 2018) and (ii) the 100-m PROBA-V LC100 land-cover map (Copernicus Global Land Service 2019). Figure S1 displays the global distribution of the points, whereas Tables S1–S3 summarize their frequency across climate and land-cover classes. In addition to yield observations, we gathered spatialized climate, soil, and topographic data from various sources. We initially assembled 17 climate, soil, and topographic layers (Table S4) and computed pairwise Spearman coefficients between them. Whenever $|r|$ exceeded 0.65, the predictor with the clearer agronomic meaning, broader data coverage, or lower residual collinearity was retained (Table S5). Five layers (radiation, vapor pressure, silt, total nitrogen and potential evapotranspiration) were therefore removed, leaving 12 predictors: temperature, rainfall, wind speed, clay, sand, soil depth, slope, aspect, salinity, soil organic carbon stock (SOC), coarse fragments, and pH. A principal-component analysis confirmed that the first six components of this 12-variable subset explain 82.3% of total variance and reproduce the temperature- and moisture-driven gradients identified in the full matrix. The PCA based on the 12 retained predictors shows that PC1 (26.2%) and PC2 (20.9%) capture the same temperature and water-availability gradients observed in the complete set, further supporting the adequacy of the reduced predictor subset (Figures S2, S3). Removing redundant predictors resulted in a final set of 12 variables for subsequent model training (Table 1).

To supply explicit boundary conditions, we incorporated 167 zero-yield pixels (Figure 1), representing approximately 4% of the final training set. These pixels were randomly selected within the six Köppen–Geiger classes BWh, BWk, Dfc, Dfd, ET, and EF after masking a 50-km buffer around every yield-observation coordinate, thereby preventing any geographic overlap with the empirical data. The six classes were chosen because they encompass the two dominant biophysical limitations that permanently preclude large-scale cultivation of lignocellulosic crops: chronic water deficit and chronic cold. Hot-desert (BWh) and cold-desert (BWk) climates receive less than 200 mm of annual rainfall and routinely experience maximum temperatures above 40°C, conditions under which Miscanthus survival falls below 15% and Eucalyptus growth ceases (Clifton-Brown et al. 2019). Subarctic (Dfc) and extremely continental subarctic (Dfd) climates are cold-limited rather than water-limited; short frost-free seasons, low growing-degree days, and extreme winter cold constrain establishment and growth. In these conditions, yields of lignocellulosic crops such as switchgrass and willow decline markedly under frost and winter-kill risks; precipitation can compound these constraints when it falls below 400–600 mm $year^{-1}$. Tundra (ET) and perennial frost/ice (EF) climates feature mean annual temperatures around or under 0°C and fewer than 90 frost-free days, halting cambial activity in Poplar and preventing the establishment of any of the five target species (IPCC 2023). The random-selection workflow, implemented in QGIS, produced a comma-separated file listing the longitude, latitude and Köppen code of each zero-yield pixel (Table S1). A sensitivity analysis in which the zero-yield sample was halved to 84 points or doubled to 334 points altered the test-set RMSE by less than 0.05 t DM ha^{-1}, confirming that 167 zero-yield samples are sufficient to delineate the climatic frontier of unsuitability while keeping the zero-yield fraction well below the 5% threshold recommended to avoid class-imbalance overfitting in ensemble trees.

TABLE 1 | Variables used as predictors for mapping crop yields.

Variable	Description	Original resolution	Data source
Temperature	Mean annual temperature (°C)	5 arc minutes	WorldClim Fick and Hijmans (2017)
Rainfall	Mean annual rainfall (mm)	5 arc minutes	WorldClim Fick and Hijmans (2017)
Wind	Mean annual wind speed (m/s)	5 arc minutes	WorldClim Fick and Hijmans (2017)
Depth	Soil depth (cm)	250 m, resampled to 5 arc minutes	SoilGrids Poggio et al. (2021)
pH	Soil pH	250 m, resampled to 5 arc minutes	SoilGrids (Poggio et al. 2021)
SOC	Soil organic carbon stock (g/kg)	250 m, resampled to 5 arc minutes	SoilGrids Poggio et al. (2021)
CFVO	Volumetric fraction of coarse fragments (cm^3/dm^3) (> 2 mm)	250 m, resampled to 5 arc minutes	SoilGrids Poggio et al. (2021)
Clay	Clay fraction in soil (g/kg)	250 m, resampled to 5 arc minutes	SoilGrids Poggio et al. (2021)
Sand	Sand fraction in soil (g/kg)	250 m, resampled to 5 arc minutes	SoilGrids Poggio et al. (2021)
Salinity	Soil salinity (categorical: 0–4)	250 m, resampled to 5 arc minutes	Global Soil Salinity Maps Ivushkin et al. (2019)
Slope	Slope gradient (%)	5 arc minutes	HWSD Wieder (2014)
Aspect	Terrain aspect (°)	5 arc minutes	HWSD Wieder (2014)

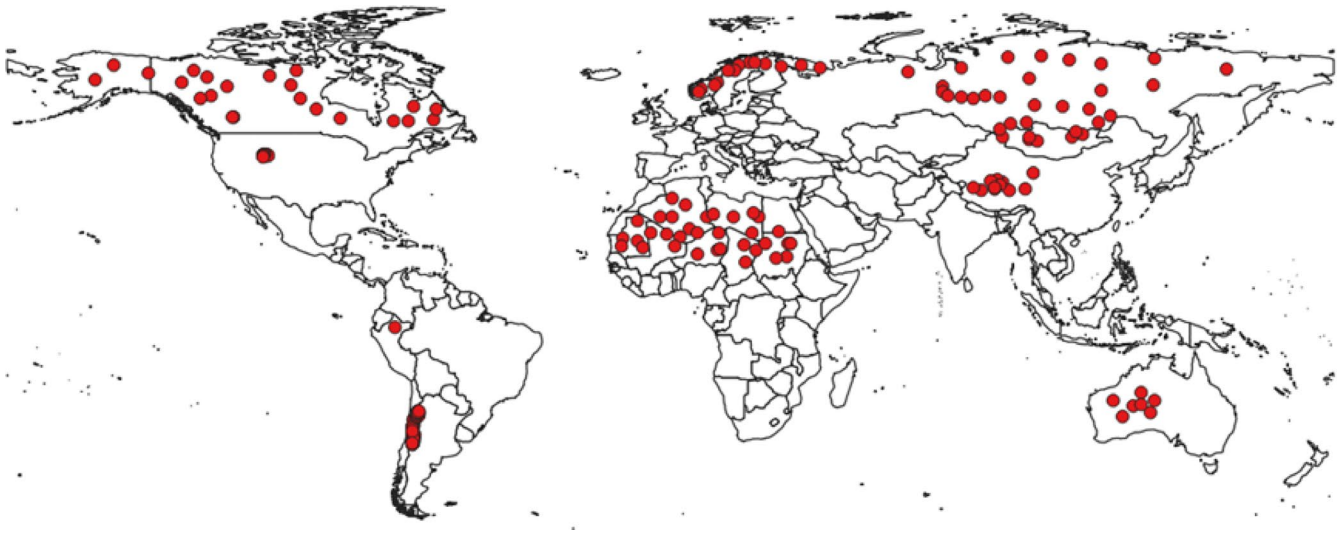

FIGURE 1 | Global distribution of zero-yield points. Map of the 167 pixels identified as unsuitable for lignocellulosic crop production based on the Köppen–Geiger climate classification. These pixels feature extreme environmental conditions where energy crops cannot be cultivated. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

2.2 | Machine Learning Models

Machine learning (ML) methods, especially tree-based algorithms, have shown strong predictive performance in agricultural yield forecasting due to their ability to handle complex, nonlinear relationships with minimal parameter tuning (Jeong et al. 2016; K. Suganthavalli 2024). In addition to tree-based algorithms, neural networks achieved competitive biomass and yield estimation accuracy in some contexts (Gopal and Bhargavi 2019). On the basis of these findings, we selected seven supervised regression models to predict lignocellulosic crop yields. Random Forest is an ensemble-based machine learning approach that leverages numerous decision trees to solve classification and regression problems (Liu et al. 2012; Abdulkareem and Abdulazeez 2021). Each tree is trained on a randomized subset of observations and features, and the final prediction

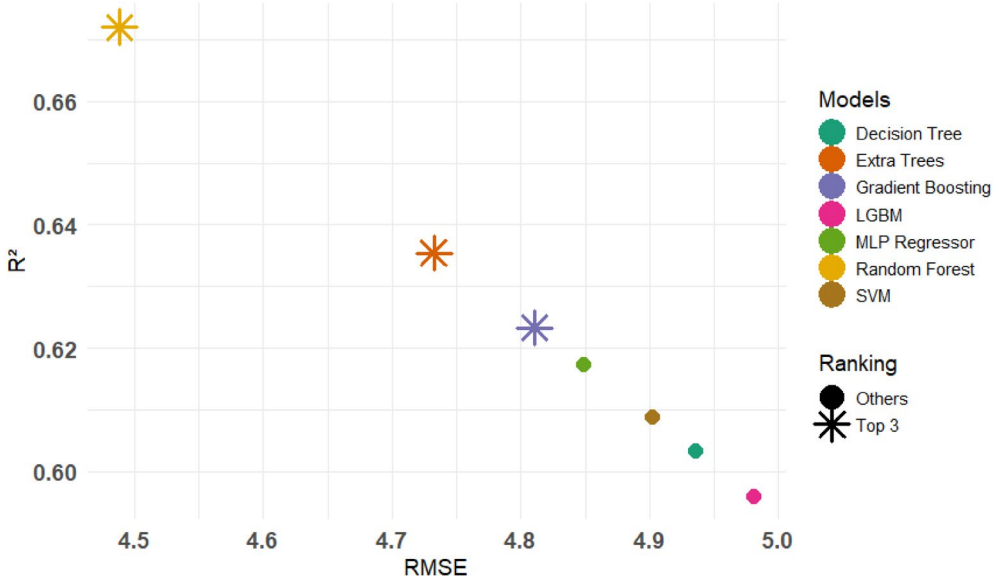

FIGURE 2 | Performance (RMSE (t ha^{-1}) and R^2 derived from the test dataset) of the models with their optimized hyperparameter values, with star symbols highlighting the top three performing models.

is generated by combining the outputs of all trees (Salman et al. 2024).

Gradient Boosting creates an ensemble of weak learners (usually decision trees) stage-wise, with new trees addressing errors of the previous ensemble (Friedman 2001). Studies have shown effectiveness in improving prediction accuracy, with reported rates ranging from 87.2% to 94.7% (Pradeep et al. 2023; Kumar et al. 2024). The ET model—also known as Extremely Randomized Trees—introduces additional randomness into the tree-building process (e.g., random thresholds) to reduce variance (Geurts et al. 2006). It is computationally efficient and often rivals RF in performance.

LightGBM Regressor (K. Suganthavalli 2024; Ke et al. 2017) is a gradient-boosting framework that uses a histogram-based approach to split features, improving efficiency and scalability. Its design allows it to handle large datasets with minimal memory usage, making it attractive for global-scale yield predictions. Support vector machine (SVM) works by plotting data points in a high-dimensional space and finding the optimal hyperplane that best separates different classes (Khatri et al. 2022). In agricultural yield modeling, SVMs can effectively handle linear and nonlinear relationships given appropriate kernel selection.

Single decision tree (DT) uses a hierarchical, rule-based approach to split the data into increasingly homogeneous subsets (Breiman et al. 1984). Although it is simple and interpretable, it can overfit if not adequately regularized.

Finally, multi-layer perceptron (MLP) regressor consists of one or more hidden layers of perceptrons, each applying a nonlinear activation function (Park 2016).

2.3 | Model Training

The same modeling procedure was employed with all seven algorithms. First, the yield dataset was partitioned into a training set (80%) and a test set (20%), holding the latter aside to provide an unbiased evaluation of model generalization. The hyperparameters of each model—referring to fixed settings governing how the algorithm learns from data, such as the maximum depth of the decision tree or the learning rate of neural networks—were systematically tuned to enhance predictive performance. We used a grid search, which involves evaluating all combinations of predetermined hyperparameter values in conjunction with k-fold cross-validation. During this procedure, the training set is split into k equally sized "folds" (here, five). In each iteration, one fold is set aside for validation, whereas the remaining folds are used for training. This process is repeated k times, ensuring each fold serves once as the validation set. The configuration of hyperparameters that minimized the average mean squared error across the five folds was selected for each model. After selecting the best hyperparameters, the model's residuals were computed (observed minus predicted yield) to detect and remove outliers (data whose absolute residuals exceeded one standard deviation of all training residuals). A second k-fold procedure was subsequently performed on the filtered training set to re-estimate performance metrics, namely, root mean square error (RMSE) and the coefficient of determination (R^2), and to verify that removing outliers improved overall predictive accuracy. Finally, each model was retrained again on the filtered training set using its optimal hyperparameter values, and the resulting models were evaluated on the previously isolated 20% test set. This approach ensured that all hyperparameter tuning, outlier filtering, and performance estimation steps were consistently applied to every model. The test dataset provided a fair, unbiased measure of generalization error.

TABLE 2 | Performances of the different models.

Model	RMSE and R^2 before filtering	RMSE and R^2 after filtering	Best hyperparameters	RMSE and R^2 of the test set	#Observations removed
Random forest	RMSE = 5 R^2 = 0.85	RMSE = 2.32 R^2 = 0.95	max_depth = 10, min_samples_split = 10, n_estimators = 100	RMSE = 4.80 R^2 = 0.67	79
Gradient boosting	RMSE = 5.07 R^2 = 0.74	RMSE = 2.26 R^2 = 0.90	learning_rate = 0.01, max_depth = 3, min_samples_leaf = 10, n_estimators = 300, subsample = 0.8	RMSE = 4.80 R^2 = 0.62	93
LGBM regressor	RMSE = 5.05 R^2 = 0.74	RMSE = 2.61 R^2 = 0.91	learning_rate = 0.01, max_depth = 1, min_child_samples = 40, n_estimators = 500, num_leaves = 31	RMSE = 4.98 R^2 = 0.59	84
Extra trees regressor	RMSE = 4.81 R^2 = 0.69	RMSE = 2.07 R^2 = 0.91	bootstrap = True, max_depth = 10, max_features = sqrt, min_samples_leaf = 2, min_samples_split = 10, n_estimators = 100	RMSE = 4.73 R^2 = 0.63	84
Decision tree	RMSE = 4.95 R^2 = 0.61	RMSE = 2.90 R^2 = 0.84	criterion = friedman_mse, max_depth = 10, max_features = sqrt, min_samples_split = 10, splitter = random	RMSE = 4.93 R^2 = 0.60	75
MLP regressor	RMSE = 5.34 R^2 = 0.73	RMSE = 2.43 R^2 = 0.91	mlp_activation = tanh, mlp_alpha = 0.01, mlp_hidden_layers_sizes = 50, mlp_learning_rate = constant, mlp_solver = adam	RMSE = 4.84 R^2 = 0.61	90
SVM	RMSE = 4.96 R^2 = 0.59	RMSE = 2.02 R^2 = 0.88	svr_C = 100, svr_epsilon = 0.5, svr_gamma = scale, svr_kernel = linear	RMSE = 4.90 R^2 = 0.60	92

Note: Model performances were assessed using two metrics (RMSE ($t\ ha^{-1}$) and R^2) by cross-validation (before and after filtering) and using an independent test dataset. The optimized hyperparameters and the number of observations removed after the filtering step are also presented.

Figure 2 illustrates the final performances of each model, through their values of RMSE (x-axis) and R^2 (y-axis) derived from the test dataset. The top three performers—RF, ET, and GB—are marked with a star as a symbol, reflecting their relatively lower RMSE values and higher R^2 scores than the other algorithms. Random Forest achieved the best balance overall (RMSE = 4.48 t ha^1, R^2 = 0.67), closely followed by ET (RMSE = 4.73 t ha^1, R^2 = 0.63) and GB (RMSE = 4.81 t ha^1, R^2 = 0.62). Although models such as LGBM, MLP, and SVM also performed reasonably well, their final R^2 values remained below 0.62.

Table 2 summarizes the results of the training set before and after outlier filtering, the test phase, the best hyperparameters identified through grid search, and the number of observations removed after outlier filtering. Although RF performed slightly better, ET and GB produced comparable results based on cross-validation and final test-set evaluations. Consequently, yield maps were produced using these three models to assess how different ensemble algorithms with similar performances compare in terms of yield projections.

2.4 | Model Prediction

After identifying the top three algorithms, we applied each model to a global dataset of environmental covariates to generate annual yield predictions for the five crops. Specifically, temperature, precipitation, wind speed, clay fraction, sand fraction, soil depth, slope, aspect, salinity, organic carbon store, coarse fragments, and pH were used as input predictors

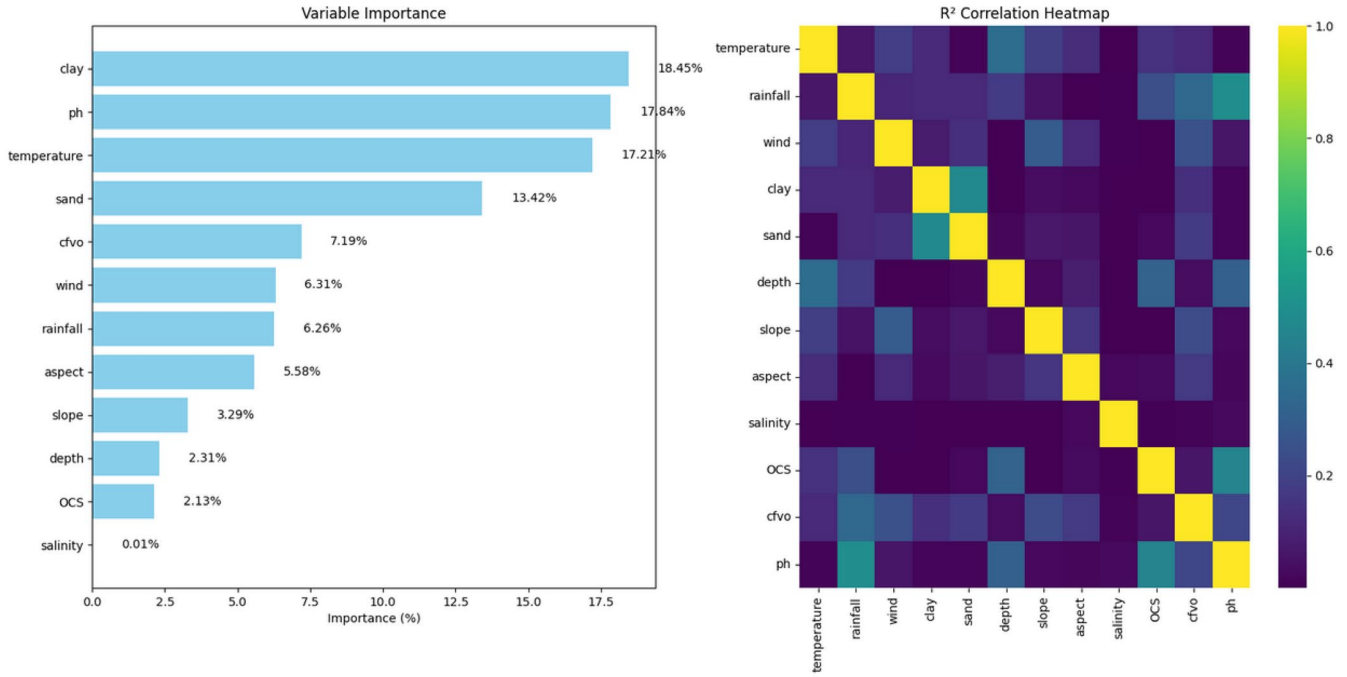

FIGURE 3 | Variable importances of predictors (left) and their correlations (right). On the left, each bar indicates the relative contribution of a given predictor to RF yield predictions (% of increase of RF prediction errors resulting from a random permutation of the predictor considered).

at a 5 arc-minute resolution. To ensure consistency with the training data structure, all predictor layers were flattened into a single feature matrix and encoded for crop type using dummy variables. Yearly predictions were then averaged to produce a mean annual yield for each grid cell. In addition, we used a binary raster mask to identify extreme climatic zones, assigning zero yields in those cells to remain consistent with zero-yield expectations. This entire procedure—year-by-year prediction followed by averaging—was repeated for each of the three algorithms and the five crops, leading to global average yield maps that account for spatial variability in environmental conditions and the biophysical constraints of lignocellulosic production.

In addition to generating yield estimates under current climatic conditions (average for 1980–2000), we applied the same modeling framework to future climate projections derived from the Shared Socioeconomic Pathway (SSP) scenarios—SSP1-RCP2.6, SSP2-RCP4.5, and SSP5-RCP8.5. These scenarios, developed in coordination with Representative Concentration Pathways (RCPs) in previous assessments (Van Vuuren et al. 2017), capture a broad range of possible greenhouse gas emissions and associated climate responses. All climate data for both the historical baseline (1980–2000) and the future period (2021–2100) were obtained from the WorldClim dataset and subsequently downscaled. By integrating the resulting temperature and precipitation layers, we produced spatial yield maps indicating how lignocellulosic crop productivity (averaged across the respective time windows) may shift under varying degrees of climate change.

In a subsequent step, "Best Crop" maps were generated by combining two types of spatial data for each crop: average yields and production costs. The latter include the main stages of biomass production—establishment, harvest, transport off the plot, and, in some cases, chipping, based on Domingues et al. (2022). Cost data were scaled to a resolution of 5 arc minutes. Yield and cost rasters were normalized to a scale of 0–1 using the formula $x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$, where min ($x$) and max ($x$) denote the minimum and maximum cost (or yield) values across all cells of the raster for the respective variable. This step ensures comparability between these otherwise heterogeneous datasets (Malczewski 2006). Next, we computed a composite index for each crop by subtracting the normalized cost from the normalized yield (composite index = normalized yield–normalized cost), thereby highlighting areas where high normalized biomass output coincides with relatively low normalized production cost. These composite indices were then stacked into a multi-layer raster, and for each pixel, the crop with the highest score was labeled as the "optimal" choice. Finally, we quantified the spatial distribution of optimal crops by calculating the percentage of pixels dominated by each species.

Regional production cost data used in this study were derived from Domingues et al. (2022) and spatialized by assigning each pixel within a region the corresponding regional cost. Detailed definitions of the production cost components and the spatialization method are provided in the Supporting Information (Table S5).

3 | Results

3.1 | Variable Importances and Correlation Among Predictors

Figure 3 provides two key perspectives on the model predictors: a bar chart showing each predictor's contribution to the final yield predictions of RF, and a heatmap illustrating how these predictors correlate with each other. From the bar

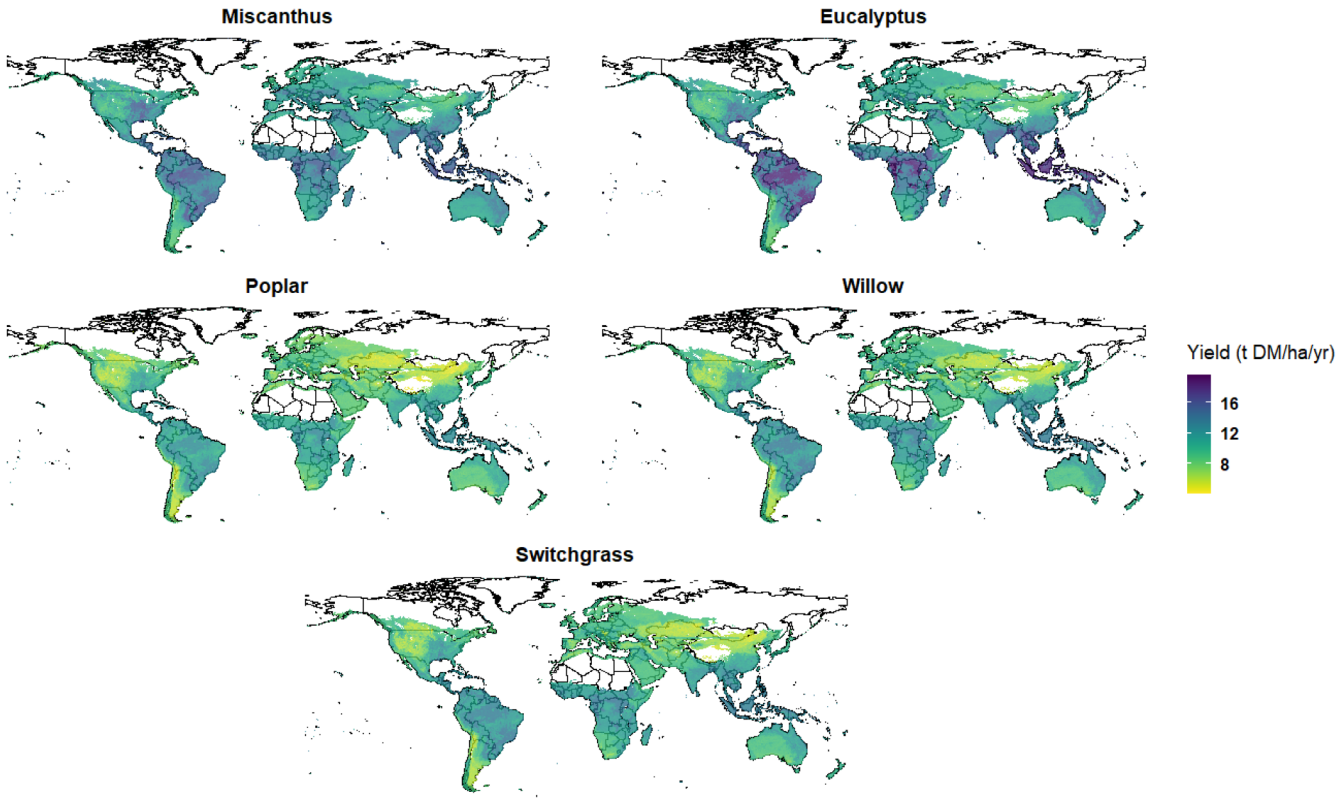

FIGURE 4 | Predicted global yield maps for five lignocellulosic crops under current climate conditions. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

chart, three variables emerge as particularly influential: clay content (18.45%), soil pH (17.84%), and temperature (17.21%). These findings align well with established agronomic literature, where clay-rich soil is known to enhance water retention and nutrient availability (Ye et al. 2019; Tahir and Marschner 2017), soil pH strongly affects microbial activity and nutrient solubility (Breugem et al. 2024), and temperature plays a defining role in plant growth and phenological development (Heggie and Halliday 2005).

On the contrary, salinity contributes very little (< 0.1%) to yield prediction despite a broader body of work indicating that salt stress can significantly lower crop productivity through osmotic and ionic stress (Alkharabsheh et al. 2021; Cushman 2001). This discrepancy likely stems from two factors in our analysis: Salinity was encoded as a categorical variable ([0, 1, 2, 3, 4]) rather than a continuous measure, and the observed field data show limited variability in salinity levels, which restricts the model's ability to pick up salinity-related effects. Meanwhile, the correlation heat map reveals low-to-moderate pairwise correlations, indicating an absence of strong multicollinearity among selected predictors. This result is related with the fact that highly correlated predictors were removed based on Spearman correlations before model training. In essence, the results highlight soil texture and chemical properties (clay and pH) alongside temperature as the principal drivers of lignocellulosic yield within our dataset. In contrast, factors like salinity have little impact here, mainly due to the narrow range of salt levels in the studied locations.

The training dataset is dominated by temperate Cfb (36%), continental Dfb (19%), and subtropical humid Cfa (18%) climates. Hot-desert (BWh) and polar tundra (ET) classes together account for < 1% of the observations. In terms of land cover, 36% of the points occur on cultivated land, 17% on herbaceous vegetation, and 14% on peri-urban areas; wetlands, shrublands, and permanent snow collectively represent under 2%. These figures confirm the wide coverage of the agro-climatic envelope where lignocellulosic crops are grown by our training dataset, while also pointing at under-represented environments.

3.2 | Global Yield Maps and Best-Crop Distribution in Actual Climate Conditions

Figure 4 presents the predicted yield maps (t DM ha^{-1} $year^{-1}$) of Miscanthus, Eucalyptus, Poplar, Willow, and Switchgrass under current climate conditions. In the maps, violet/blue shades indicate higher productivity, predominantly clustered in tropical and subtropical regions (e.g., Southeast Asia, Central Africa, and parts of South America). Some temperate areas, such as the eastern United States or western Europe, also exhibit high yields for certain species (e.g., Miscanthus and Switchgrass). Conversely, green/yellow hues signify more limited yields, typically found in arid or semi-arid zones (e.g., northern Africa, the Middle East, and interior Australia) or in colder/high-latitude regions (e.g., northern Canada and Russia) with shorter growing seasons.

Although each crop has its own climatic and edaphic preferences, their average yields follow this ascending order: Poplar (9.32 t DM $^{-1}$ $year^{-1}$), Switchgrass (9.97), Willow (10.00), Miscanthus (12.11), and Eucalyptus (12.20). These values span roughly 4–19 t DM $^{-1}$ $year^{-1}$ across the globe, illustrating how

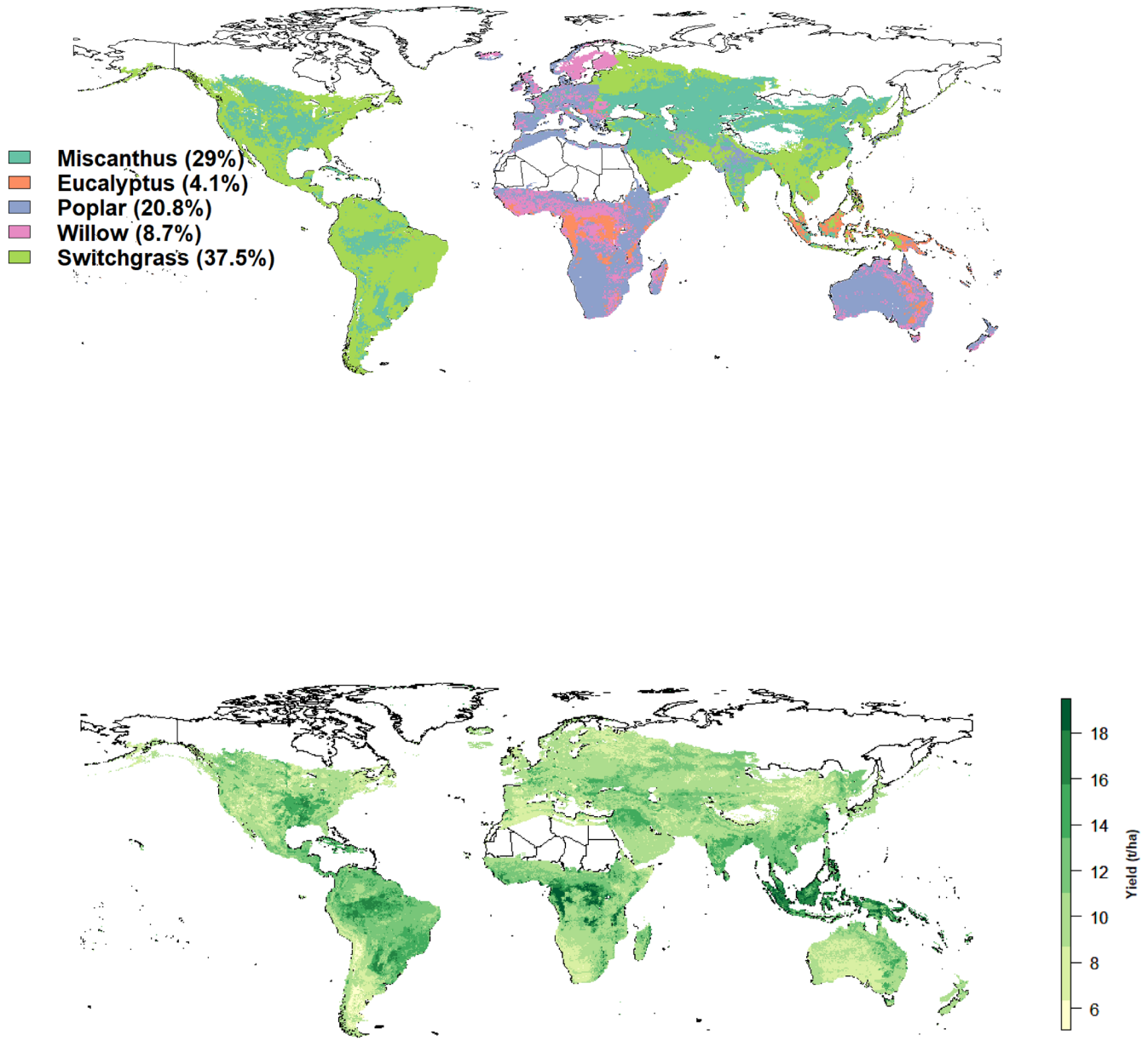

FIGURE 5 | Best crop map indicating the crop species leading to the best yield-cost balance (top) and resulting average yields (bottom). Map lines delineate study areas and do not necessarily depict accepted national boundaries.

local environmental conditions can substantially modulate productivity.

In Figure 5, a "best crop" map reveals which of the five species achieves the most favorable balance between yield and cost at each pixel. Overall, Switchgrass dominates 37.5% of the land area, followed by Miscanthus (29%), Poplar (20.8%), Willow (8.7%), and Eucalyptus (4.1%). The best crop average yield varies between 4 and 19 t DM ha^{-1} $year^{-1}$, with a global average yield of 10 t DM ha^{-1} $year^{-1}$. Although these proportions are partly consistent with the yield patterns seen in Figure 4, they are also shaped by the production costs illustrated in Figure 6.

A grassy crop is often selected in regions with moderate productivity (e.g., Switchgrass or Miscanthus) if the cost per gigajoule ($/GJ) is sufficiently low. Conversely, in areas of exceptionally high yield potential, wood energy crops like Eucalyptus may remain competitive despite relatively higher production costs, as larger biomass outputs offset the increased expense. Thus, the "best crop" distribution reflects both biophysical suitability and the cost differentials depicted in the new production cost maps.

3.3 | Global Yield Maps Under Future Climate Conditions

Figure 7 illustrates the predicted average yields of all five lignocellulosic crops (Miscanthus, Eucalyptus, Poplar, Willow, Switchgrass) under three future climate scenarios (SSP1-2.6, SSP2-4.5, and SSP5-8.5). At first view, the average yields show only small variations—around 1–2 t DM ha^{-1} $year^{-1}$—compared

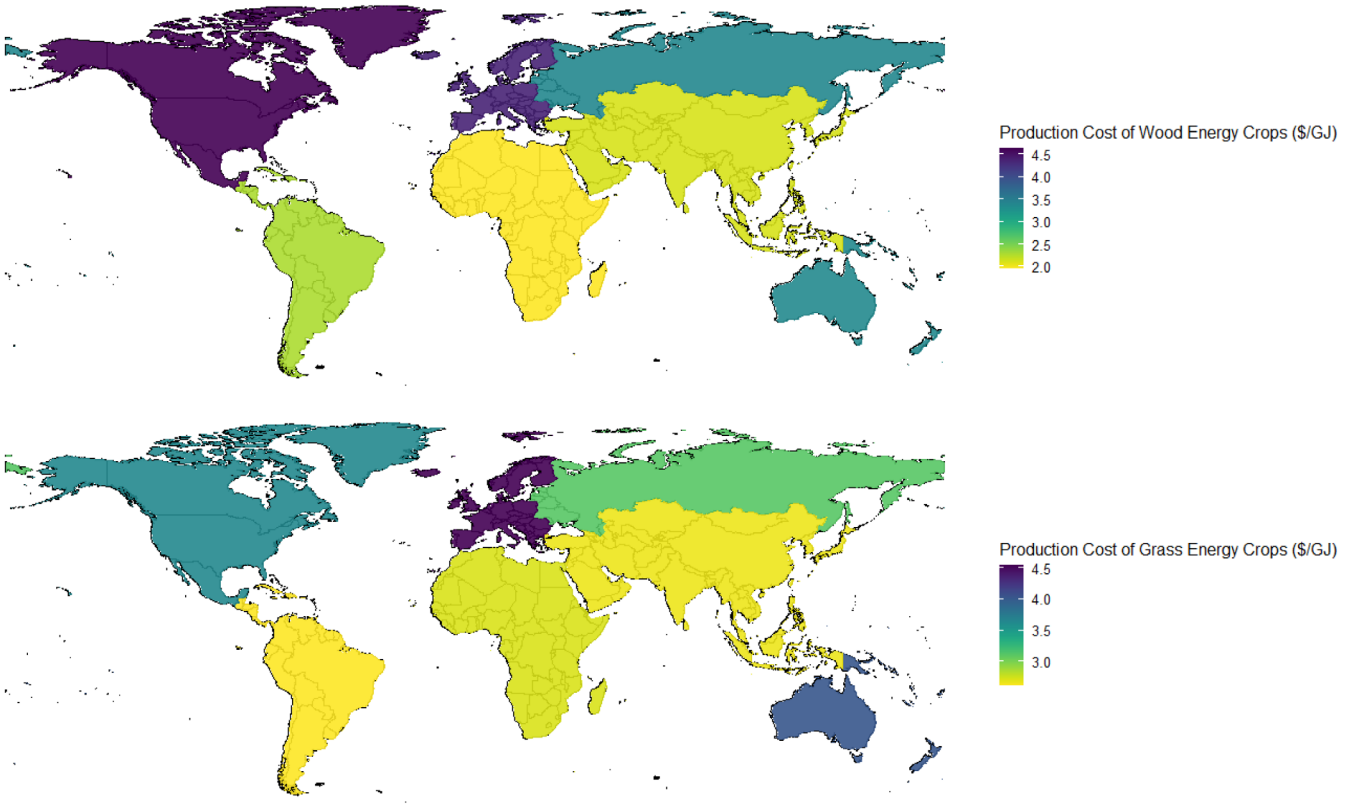

FIGURE 6 | Production costs of grassy and woody energy crops ($/GJ). (Domingues et al. 2022). Map lines delineate study areas and do not necessarily depict accepted national boundaries.

with predicted maps produced at current climate conditions, a finding likely attributed to the large spatial extent and relatively fine pixel resolution considered. To better capture regional nuances, Figure 8 provides difference maps (future minus current) for Miscanthus and boxplot illustrating the distribution of yield changes per scenario for Miscanthus. Similar figures are available in the Supporting Information for the other crops (Figures S4–S7).

Miscanthus exhibits changes ranging from −3 to +3 t DM $ha^{-1}year^{-1}$, with northern countries generally experiencing yield gains and southern regions showing reductions under climate change.

Eucalyptus shows higher yield change variability, from −5 to +7 t $ha^{-1}year^{-1}$, but demonstrates net positive outcomes in most areas except parts of sub-Saharan Africa and India. Willow, Poplar, and Switchgrass each display shifts between −6 and +6 t $ha^{-1}year^{-1}$, with negative average changes observed over a large share of the cropping area.

To further disentangle the individual impacts of climate variables on projected yield changes, we performed Partial Dependence Plot (PDP) analyses of the relationships between temperature and precipitation anomalies (ΔT and ΔP) and yield anomalies (ΔYield) for Miscanthus. Three climate scenarios were tested (SSP1, SSP2, and SSP5; see Figures S8–S10). The results reveal a nonlinear response of yield anomalies to temperature increases, with moderate warming generally leading to slight yield reductions. Conversely, precipitation changes showed a distinct unimodal relationship: moderate increases in precipitation positively influenced yields, whereas large deviations—either substantial increases or decreases—negatively impacted yields. These findings emphasize the complexity of climatic impacts on biomass crop productivity and the necessity of considering nonlinear responses when interpreting climate-driven yield changes.

Figure 9 presents scatter plots for each of the five lignocellulosic crops, comparing the yield predictions of RF vs. GB on the left and RF vs. ET on the right. As expected, across all crops, RF and GB predictions exhibit high correlations—often exceeding 0.90—indicating that their pixel-level yield predictions align closely, likely due to the relative similarity of the two algorithms. By contrast, the RF–ET correlations lie in the 0.50–0.60 range, suggesting a more pronounced spatial divergence, even though ET's global performance (RMSE, R^2) is comparable to the other ensemble models. This discrepancy can be traced to Extra Trees' randomized splitting of features, which can yield distinct local partitions despite ultimately achieving similar error metrics. Consequently, although both GB and ET perform well at a global scale, their respective internal mechanisms lead to different degrees of similarity in predicted yield patterns when compared to RF.

3.4 | Impact of Including Soil and Topographic Inputs for Yield Forecast

We repeated the entire training and prediction process without soil and topographic predictors to evaluate how soil and topographic information improves yield predictions compared with

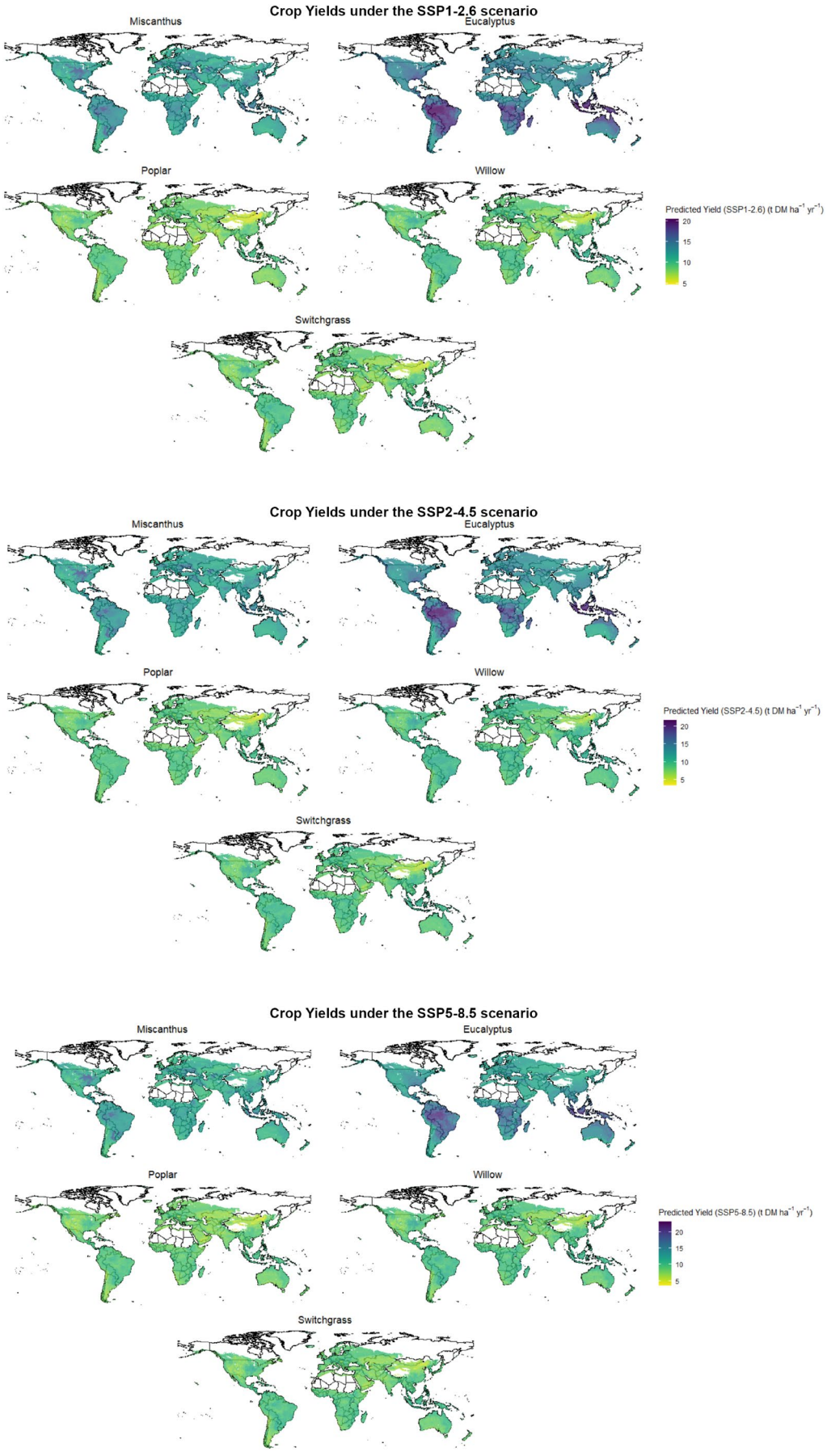

FIGURE 7 | Predicted global yield maps for five lignocellulosic crops under the SSP1-2.6, SSP2-4.5, and SSP5-8.5 scenarios. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

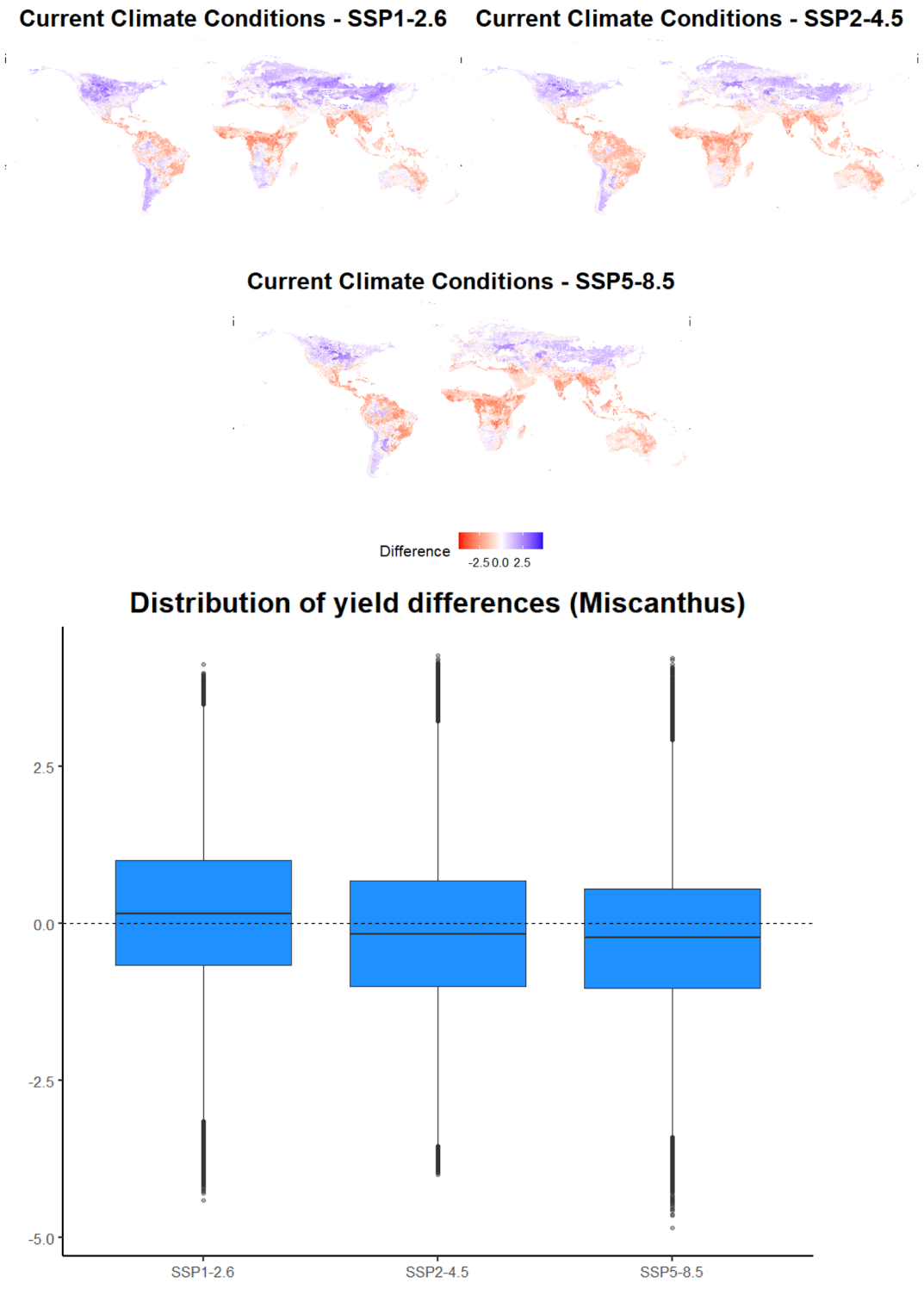

FIGURE 8 | Map of differences between Miscanthus average crop yield under current and future climate conditions (future–current), and boxplots showing the distributions of yield differences over all grid cells. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

models relying on climate predictors only, finding in RMSE = 4.5 and an $R^2 = 0.67$ (metrics of the test dataset). We then subtracted these climate-only yield maps from the yields predicted using the complete set of variables (i.e., including soil and topography). Figure 10 displays the resulting difference maps and boxplots summarizing the distribution of differences for each of the five crops.

The patterns reveal notable spatial disparities when transitioning from a climate-only model to one that also considers soil and topography. For Miscanthus, differences range from −8 to +8 t DM ha^{-1} $year^{-1}$, with most variations clustering around +2 t DM ha^{-1} $year^{-1}$. For Eucalyptus, yield prediction differences fall between −6 and +6 t ha^{-1}, showing pronounced yield reductions in parts of northern Latin America

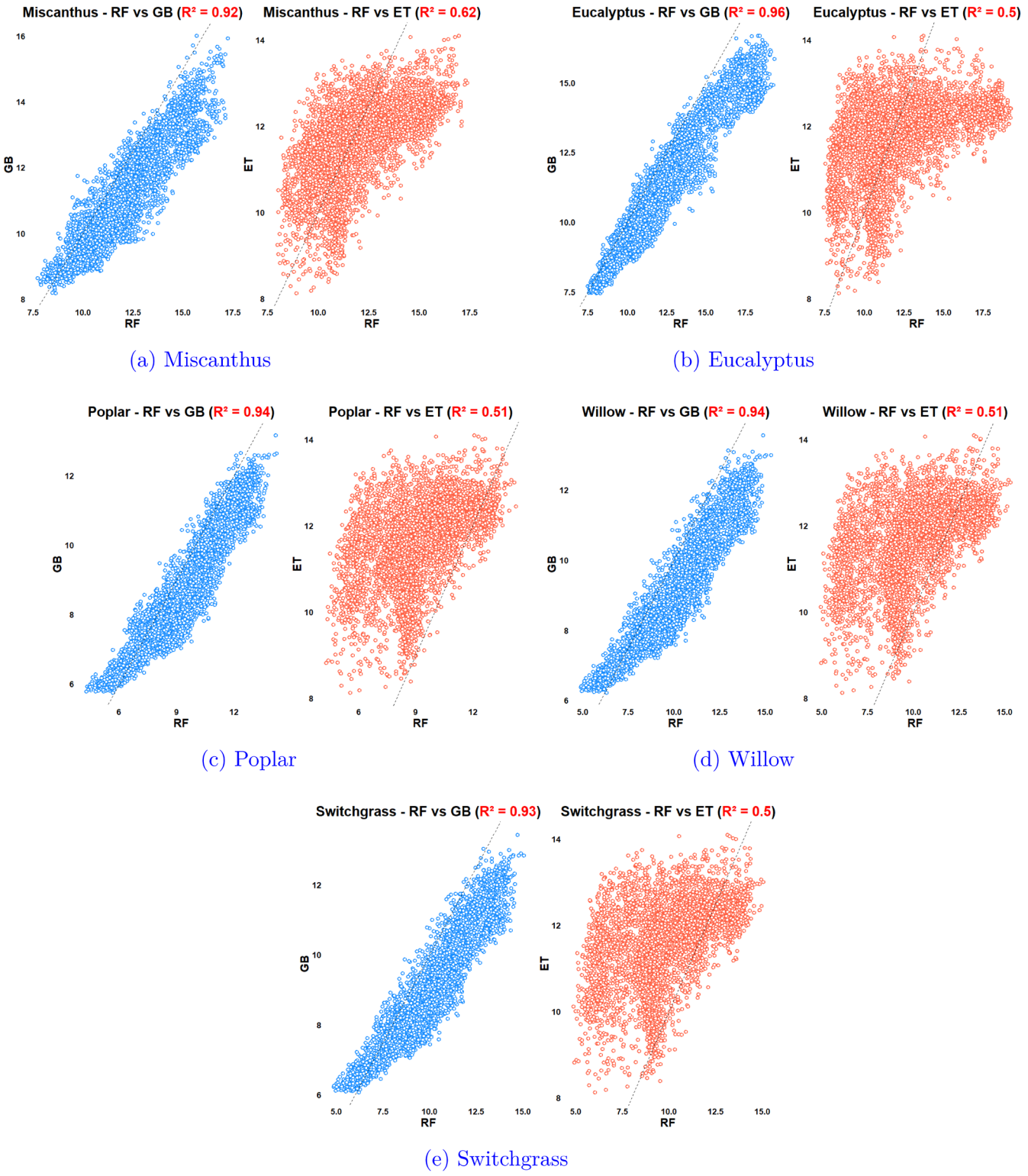

FIGURE 9 | Model-to-model scatter plots for lignocellulosic crop yield predictions. This Figure shows the pixel-level comparisons of predicted yields between Random Forest (RF) vs. Gradient Boosting (GB) and RF vs. Extra Trees (ET) for each of the five crops (units: t DM ha^{-1}).

under the climate-only scenario. For Poplar, differences span −2.5 to +6 t ha^{-1} $year^{-1}$, with an average difference of about +1 t ha^{-1} $year^{-1}$ in favor of the soil-inclusive model, whereas yield differences obtained for Willow and Switchgrass vary from −4 to +4 t ha^{-1} $year^{-1}$, exhibiting larger gains in certain regions of Latin America and Africa. Over all crop species, a global average comparison indicates that the model including soil predictors produces lower average yield estimates, with predicted yield values approximately 1–2 t ha^{-1} $year^{-1}$ lower than the predictions obtained with the model including climate predictors only. These findings underscore the substantial influence of soil properties and land characteristics on yield predictions, expanding on previous studies such as Li et al. (2020) that focused more narrowly on climate variables.

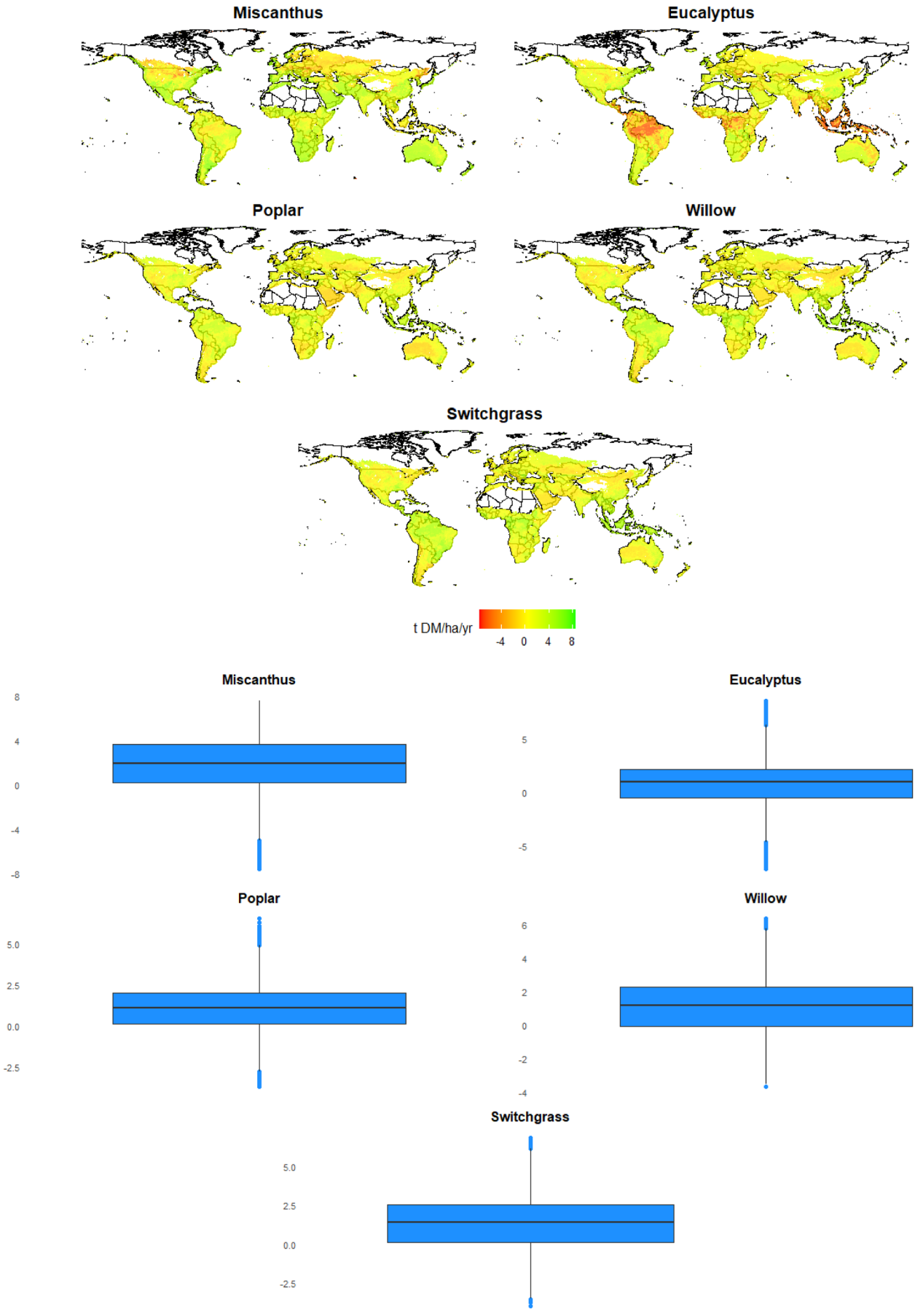

FIGURE 10 | Legend on next page.

FIGURE 10 | Differences in predicted yields by RF with and without soil and topographic predictors (predictions with–predictions without). This Figure compares yield predictions (in t ha^{-1} year^{-1}) of the full model (including soil and topographic variables) against those generated using only climate inputs as predictors. Each panel corresponds to a specific species. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

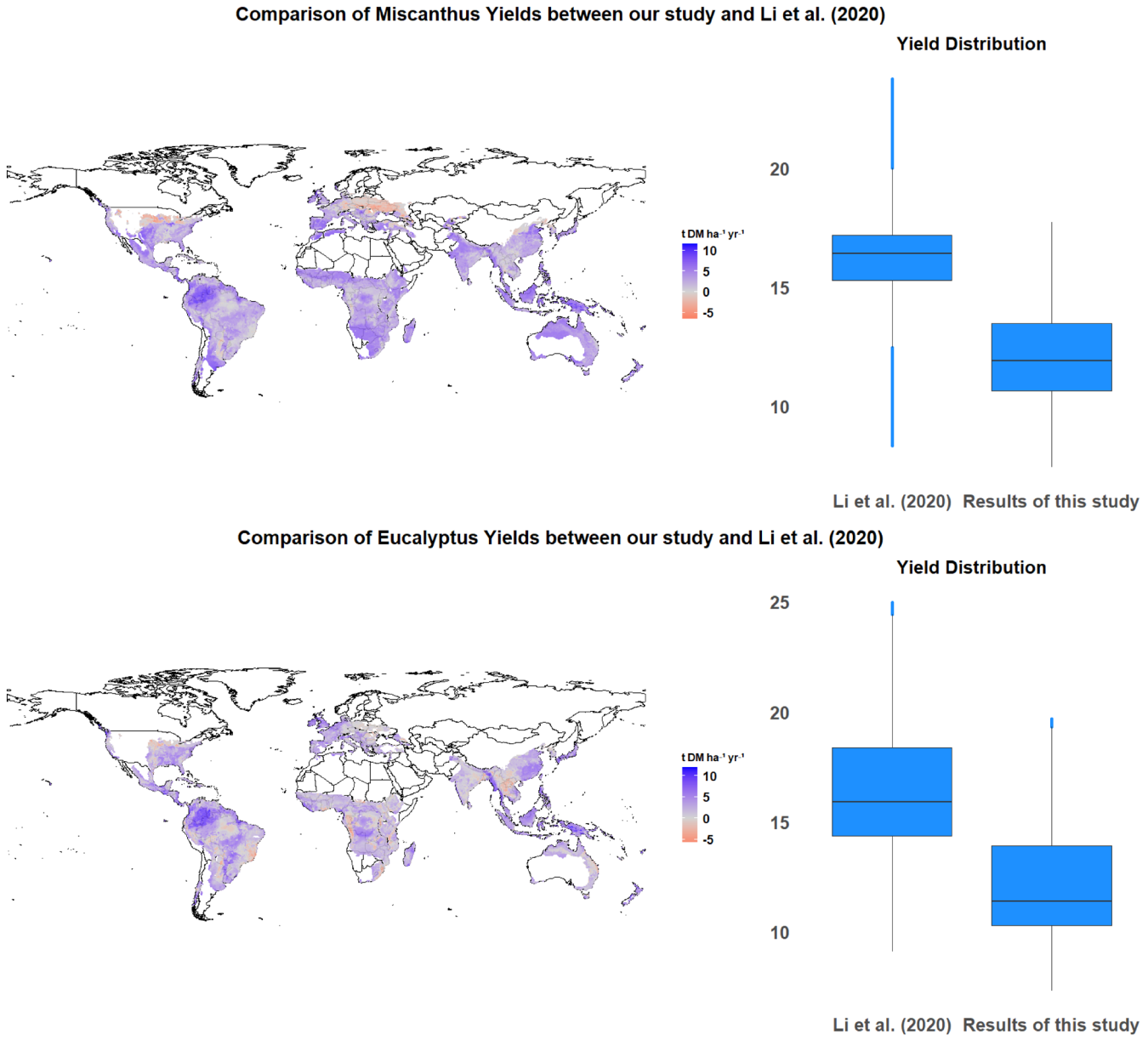

FIGURE 11 | Yield discrepancy maps and boxplots: Li et al. (2020) vs. this study. The maps illustrate the differences between the resampled yields from Li et al. (2020) and yield values predicted in the present work (Li et al. minus current study) for Miscanthus (top) and Eucalyptus (bottom). The boxplots (right) summarize the distribution of yield predictions obtained with both approaches for each crop. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

3.5 | Comparison With Li et al. (2020)'s Yield Maps

Figure 11 compares our global yield estimates (5 arc-minute resolution) with those reported by Li et al. (2020), originally produced at a 30 arc-minute scale and subsequently resampled to 5 arc minutes using bilinear interpolation for Miscanthus and Eucalyptus (see other crops in Supporting Information from Figures S11–S13). Regions outside the spatial extent covered by Li et al. (2020) were excluded for consistency, resulting in masked areas on the difference maps. Each panel presents a map of yield discrepancies (Li et al. minus our predictions), alongside a boxplot summarizing the overall distribution of these differences. Consistent with previous findings, the largest deviations appear for Eucalyptus and Miscanthus, with mean differences of approximately 4 t DM ha^{-1}, whereas Poplar, Willow, and Switchgrass show more moderate

discrepancies of about 2 t DM ha^{-1}. In most regions, Li et al.'s estimates exceed our own, reflecting the inclusion of detailed soil and topographic variables leading to more conservative predictions.

Our results concur with those of Li et al. (2020) on the fact that Miscanthus is the crop that most frequently achieves the best yields among the five crops ("best crop"). However, we note significant differences regarding the other crops. In particular, Eucalyptus appears less frequently as the best crop here (35.9% vs. 4.1% in this study), whereas Switchgrass and Poplar appear more frequently (0% and 1.6% vs. 37.5% and 20% in this study).

4 | Discussion

4.1 | The Impact of Soil Properties

Although climate variables such as precipitation and air temperature often emerge as primary drivers of biomass productivity (Cacho et al. 2023), soil characteristics and topography substantially influence the yield of crops (Haberzettl et al. 2021). Our analysis underscores the effect of these factors as properties such as soil clay and sand contents, pH, and gravel content ranked in the top 5 most influential variables in our yield modeling. This is consistent with a study emphasizing the effect of soil texture on the yield of biomass crops (Jiang and Thelen 2004). Soil texture and gravel content control water and nutrient availability to crops, representing the main limiting factors to biomass production since lignocellulosic crops are less affected by pests and diseases than conventional food crops (Gabrielle et al. 2014). Neglecting such edaphic and topographic factors thus entails a risk of overestimating yield potentials or producing misleading spatial projections of high-yielding zones.

Incorporating detailed soil and topographical variables alongside climate data markedly improved the predictive performance of the five models and altered the spatial pattern of predicted yields. Areas with fertile soils (e.g., higher organic matter or optimal pH) and favorable landscape positions (e.g., low slope or higher moisture retention capacity) achieved greater production potentials than areas with poorer soils under similar rainfall and temperature regimes. This result highlights that favorable climate conditions alone are insufficient for high biomass yields: the local soil environment ultimately constrains crop productivity (Kravchenko and Bullock 2000). By integrating these site-specific variables, the model better captures the heterogeneous yield potential across regions. In practical terms, our findings stress that robust yield assessments require fine-resolution soil and terrain information, not just regional climate averages, to pinpoint high-yielding areas accurately and decipher yield gaps in marginal lands. Notably, the influence of edaphic factors can appear sharply: in some regions, we observed that differences in soil and topographic conditions could cause yield projections to vary by as much as 8 t DM ha^{-1}. Such large local deviations underscore the dominant role that soil constraints or advantages can play in shaping biomass productivity (Haberzettl et al. 2021), which becomes even more critical under changing climate conditions where water availability and soil quality may exacerbate or mitigate yield losses.

Note that the gridded soil covariates used here represent long-term regional averages. They do not capture within the field variability, recent amendments, drainage or irrigation investments, or other management interventions that can alleviate edaphic constraints. Accordingly, the soil effects we report should be interpreted as broad biophysical limitations rather than management-specific yield response.

4.2 | The Effect of Machine Learning Models

To capture the complex, nonlinear interactions between climate, soil, and topography on yields, we employed several ML algorithms. In particular, we compared ensemble tree-based models—RF and ET—as well as a Gradient Boosting Machine (GBM). All algorithms achieved strong predictive accuracy, reflecting their capability to model high-dimensional relationships more effectively than traditional linear methods in this context (Onsree et al. 2022). Indeed, our findings align with similar agricultural studies where ensemble ML approaches outperformed simpler models on multi-factor yield prediction tasks (Cacho et al. 2023; Su et al. 2022). For example, in switchgrass yield modeling, tree-based ensembles attained coefficients of determination (R^2) on the order of 0.85–0.88, substantially higher than those of regression or untuned neural network models (Cacho et al. 2023). These results demonstrate the value of ensemble learning techniques for yield prediction, given their ability to handle interactions and nonlinear responses that would be challenging to capture with parametric or purely mechanistic models.

In our analysis, the performance differences among RF, ET, and GBM were relatively small, with all three methods proving highly effective on cross-validation and independent test data. Random Forest achieved the best overall accuracy on the test set (lowest RMSE and highest R^2, 0.67), closely followed by ET and GB (with R^2 values of 0.63 and 0.62, respectively). Extra Trees had a slight edge in validation accuracy, consistent with its design: by injecting additional randomness in splitting nodes, ET can reduce overfitting while maintaining high predictive power (Geurts et al. 2006). The GB model also performed competitively, indicating that its sequential error-correcting mechanism can capture yield patterns almost as well as the bagging-based methods. Overall, the three top ensemble models captured yield variability with a high degree of fidelity, effectively accounting for interactions (e.g., the effect of soil water-holding capacity on drought mitigation) and nonlinear effects that would be difficult to pre-specify in regression models. Moreover, the models' internal variable importance rankings corroborated domain expectations: key predictors of yield included annual and growing-season temperatures, cumulative precipitation, and topographic indices such as slope and elevation—factors well-known to influence crop growth (Haberzettl et al. 2021). The convergence of different algorithms toward a similar set of influential predictors bolsters confidence that the identified environment–yield relationships are robust and not the artifacts of a single modeling technique.

An important caveat is that the training data lack information on management intensity (fertilization, irrigation, and stand age/rotation), cultivar or clone differences, pest and disease

damage, and episodic stresses. Water limitation enters the model only indirectly via climatic means and soil attributes; no explicit soil–water balance or irrigation module is represented. Consequently, the model outputs should be viewed as biophysical yield potentials under average environmental conditions, not as realized farm yields or sustainably harvestable outputs.

Beyond aggregate performance metrics, we observed interesting spatial differences in the yield maps produced by each ML model. In particular, the RF and GBM yield maps were very strongly correlated with each other (pixel-level predictions often showing Pearson correlations exceeding 0.90 for all five crops), indicating that these two algorithms agreed in their projections of high- or low-yielding zones. The RF–ET correspondence was weaker, with correlation coefficients in the 0.50–0.60 range. In other words, the ET model sometimes produced divergent local yield estimates relative to RF and GBM, even though its overall error statistics were similar. Such patterns involving local discrepancies despite an overall agreement between ET and other ML models were already reported with hydrological models (Galelli and Castelletti 2013). This suggests that the extra randomness brought about by ET in constructing trees enhances spatial variability in pixel-by-pixel predictions. Crucially, these local differences did not translate into significant global performance gaps, pointing to nuance when applying ensemble models: models with similar global accuracy can yield different spatial output patterns. For our study, this warranted generating and examining maps from all three top-performing models and not relying on a single one. The high agreement between RF and GBM points to a stable consensus, whereas ET provides a plausible alternative scenario in certain locales. This multi-model comparison thus gave us insight into the uncertainty of spatial predictions stemming from model choice. Nonetheless, the general consistency among the ensemble methods regarding high-yielding zones and key drivers is reassuring. In summary, the use of ensemble ML provided a flexible and accurate framework for yield mapping, and the close performance of RF, ET, and GBM suggests that our conclusions are not overly dependent on one specific algorithm.

4.3 | Policy Implications

The high-resolution yield and production cost maps developed in this study have potential implications for bioenergy policy and land-use planning. First, our spatial analyses identifying the "best crop" for each pixel (balancing yield and production costs) offer valuable guidance for optimizing land allocation. We found pronounced geographical variations in crop suitability. Switchgrass and poplar came out as optimal choices in many regions where earlier assessments had favored eucalyptus (Li et al. 2020), which emerged as the best option in only a small percentage of areas (around 4%).

This discrepancy underscores the influence of incorporating soil and terrain constraints: earlier, coarse approaches from this perspective may have overestimated the dominance of woody species by assuming uniformly favorable conditions, despite heterogeneous soil conditions. Our models account for edaphic limitations and indicate a more diversified portfolio of optimal crops. For policymakers, this means that strategies relying on a single "silver bullet" energy crop over large areas (e.g., Miscanthus across northern Europe or Switchgrass in the mid-Western US) would be suboptimal. Different regions should specialize in the crops best suited to their local environmental conditions, also involving the second or third best crops according to our modeling, to design multi-feedstock scenarios. Diversifying biomass supply and crops is a common scheme for lignocellulosic biorefineries, fostering resilience from an agronomic perspective and facilitating logistics. Benefiting from multiple harvesting windows reduces storage costs, for instance (Wang et al. 2020). Tailoring incentives and extension efforts to promote the right crops in the right locations can improve biomass supply chains' overall productivity and cost-effectiveness (Hudiburg et al. 2016).

Including production cost estimates alongside crop yields provides a practical economic dimension to our maps that can inform policy and investment by highlighting zones where biomass can be the cheapest to source. These insights help prioritize areas for bioenergy development. Because our maps express biophysical potential, they should be best used as a first-pass screening layer to assess the performance of crop species and identify the best candidates to establish locally. Moving from screening to actual implementation requires overlaying additional information and criteria: land tenure and security; competing land uses and opportunity costs; proximity to infrastructure and processing capacity; water availability and allocation rules; environmental safeguards (soil erosion risk, biodiversity constraints); and socio-economic conditions that shape farmer adoption and labor availability. Without these layers, the yield-plus-cost maps should not be used to rank specific investment projects or to make binding land-use allocations. For example, a region where Miscanthus has both high yield and low production cost per ton would be a strong candidate for commercial projects. In contrast, another region might achieve higher yields with Willow but at a prohibitive cost, suggesting that resources would be better directed toward alternative crops or technological improvements. Policymakers could use this information in planning initiatives such as biomass cropping programs, infrastructure placement (e.g., biorefineries), and incentive structures (subsidies or supports for certain crops in particular locations). In addition, our approach of combining yield and cost into a single suitability criterion aligns with sustainable bioenergy deployment. It ensures that bioenergy systems are economically viable for farmers and producers, which is crucial for long-term adoption.

Another key implication of our findings relates to the effect of climate change on biomass potentials. Our projections indicated modest yield changes (around ±1–2 t DM ha^{-1}) in many regions by mid-century, assuming no change in management. Thus, lignocellulosic crops could remain a reliable component of our energy supply in the face of climate change. However, we also identified specific areas where yields could change dramatically, with pronounced gains or losses up to 8 t DM ha^{-1}. These larger shifts occurred in the margins of crops' suitability areas, in regions currently limited by suboptimal temperatures, water scarcity, or poor soils. Such regions may experience significant yield boosts if climate changes should alleviate one of the limiting factors (e.g., a longer growing season in cooler climates) or, conversely, sharp declines with further warming or drying in currently semi-arid areas. For policy and planning, the granularity

provided by our maps proves crucial. It indicates that adaptation measures may be needed in vulnerable hotspots (such as developing more drought-tolerant varieties or improving soil water retention) to safeguard yields. At the same time, other areas might become newly attractive for bioenergy cultivation and could be targeted for expansion.

5 | Limitations

Despite the advances achieved in this modeling work, several limitations should be acknowledged. First, there are inherent uncertainties in the data used to train the models or project yields. Models reflect the quality of the data they are trained on. Although we used an extensive global dataset of yield observations (Li et al. 2020), cultivar- or clone-level information is rarely reported in that dataset. Yet, each of the crop species or genus considered encompasses diverse genotypes or species (e.g., *Miscanthus sinensis* vs. *Miscanthus*×*giganteus*; multiclone willow mixtures), which differ in yield potential, phenology, tolerance to temperature and water stress, and pest/disease vulnerability. Because we could not systematically distinguish between cultivars, our models necessarily represent an average, species-level biophysical response. Local yields may therefore deviate from our mapped estimates depending on the specific varieties to be deployed. Users should interpret the results with this additional uncertainty in mind, and adjust them when reliable cultivar information becomes available.

Our results represent the environmental (biophysical) yield potential derived from the gridded climate and soil covariates. They do not account for site-specific management (inputs, planting density, and rotation length), genotype, pest and disease pressures, interannual climate extremes, water competition or irrigation constraints, or sustainability limits on residue removal and harvest intensity. Reported production cost layers are regional averages applied uniformly within each region and therefore do not reflect local variation in labor, input prices, or logistics. These simplifications should be kept in mind when interpreting the political relevance of our maps.

Regarding soil and topographic inputs, the global soil databases and terrain maps we leveraged may not capture their fine-scale variability or transient soil properties (such as nutrient dynamics, microtopography, or soil fertility; Poggio et al. 2021). Some soil parameters were likely measured or estimated with some degree of uncertainty, and important soil characteristics such as depth to hardpan or historical land degradation were not included due to a lack of global data. Similarly, the climate data we used represent long-term averages and do not explicitly account for interannual variability or extreme events. This means that our yield predictions correspond to expected averages rather than best-/worst-case outcomes. These data-related limitations could affect the accuracy of yield estimates and warrant a degree of local ground truthing (Deines et al. 2021).

Because hot-desert, wetland and alpine environments are under-represented in the training set, yield projections for these specific climate–land-cover combinations should be interpreted with caution. Importantly, these environments are intrinsically unfavorable for lignocellulosic crops; in hot-desert climates, synergistic heat and water stress can reduce switchgrass yields by up to 50% and impair downstream fermentation (Chipkar et al. 2022). In alpine or high-latitude cold zones, exposure to soil temperatures below −3°C to −7°C kills *Miscanthus*×*giganteus* rhizomes, preventing stand persistence (Sage et al. 2015). Consequently, these zones are unlikely to represent future expansion areas for biomass crops, and their limited presence in the training data has little practical impact on global projections.

6 | Outlook and Future Work

A critical avenue is the integration of remote sensing data into yield prediction frameworks (Ziliani et al. 2021). Advances in satellite observations offer an opportunity to regularly monitor vegetation conditions, which could be assimilated into models to improve both spatial detail and temporal responsiveness. For instance, remotely sensed indices of vegetation greenness, biomass, or moisture status have been successfully used to augment yield estimates for conventional crops (Franz et al. 2020). In the context of lignocellulosic crops, remote sensing could help update yield predictions in near real-time (e.g., detecting the effects of an ongoing drought or a particularly favorable growing season), and could also identify where crops are currently being grown, addressing data gaps in planting area. Future modeling efforts could incorporate satellite-derived variables such as NDVI/EVI (indicators of plant health), surface soil moisture, or land surface temperature as additional predictors to capture short-term variability and unforeseen stress events.

Updating the yield database with more recent data, especially under commercial and suboptimal soil or topographic conditions, would also be of added value. As the acreage of lignocellulosic crops expands, the data collected from farms could also be used for ground truthing purposes.

In conclusion, integrating fine-resolution soil and topographic predictors with climate allowed us to derive a relative suitability index. These results provide a transparent, spatial screening resource: They highlight where biophysical conditions are comparatively favorable, but cannot substitute for detailed, location-specific assessments incorporating management, sustainability constraints, land tenure, infrastructure, and broader socio-economic considerations. Combining our maps with such information will enable more realistic planning of regional crop portfolios and bioenergy supply chains under climate change.

Author Contributions

Siwar Saadaoui: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review and editing. **David Makowski:** conceptualization, methodology, supervision, writing – review and editing. **Benoît Gabrielle:** data curation, supervision, writing – review and editing. **Thierry Brunelle:** supervision, writing – review and editing.

Acknowledgements

We gratefully acknowledge financial support from the *Chaire CarMa* and the *Institut Convergence CLAND*.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The model outputs (GeoTIFF rasters; 5 arc-minute; EPSG:4326) are openly available at Zenodo: https://doi.org/10.5281/zenodo.16841962. The modeling code is available at Zenodo: https://doi.org/10.5281/zenodo.15212483. Secondary datasets used (WorldClim v2, SoilGrids 250 m, Global Soil Salinity, regridded HWSD, and Köppen–Geiger maps) can be obtained from their public repositories.

References

Abdulkareem, N., and A. Abdulazeez. 2021. "Machine Learning Classification Based on Random Forest Algorithm: A Review." *International Journal of Science and Business, IJSAB International* 5, no. 2: 128–142. https://doi.org/10.5281/zenodo.4471118.

Alkharabsheh, M., H. M. Alkharabsheh, M. F. Seleiman, et al. 2021. "Field Crop Responses and Management Strategies to Mitigate Soil Salinity in Modern Agriculture: A Review." *Agronomy* 11, no. 11: 2299. https://doi.org/10.3390/agronomy11112299.

Beck, H. E., N. E. Zimmermann, T. R. McVicar, et al. 2018. "Present and Future Köppen-Geiger Climate Classification Maps at 1-km Resolution." *Scientific Data* 5, no. 1: 180214. https://doi.org/10.1038/sdata.2018.214.

Breiman, L., J. Friedman, R. A. Olshen, et al. 1984. *Classification and Regression Trees*. 1st ed. Routledge. https://doi.org/10.1201/9781315139470.

Breugem, A., H. Kros, and W. De Vries. 2024. "Impacts of pH on Mechanisms and Rates of Carbon and Nitrogen Mineralisation: A Review." https://doi.org/10.18174/653235.

Cacho, J. F., J. Feinstein, C. R. Zumpf, et al. 2023. "Predicting Biomass Yields of Advanced Switchgrass Cultivars for Bioenergy and Ecosystem Services Using Machine Learning." *Energies* 16, no. 10: 4168. https://doi.org/10.3390/en16104168.

Calvin, K., A. Cowie, G. Berndes, et al. 2021. "Bioenergy for Climate Change Mitigation: Scale and Sustainability." *GCB Bioenergy* 13, no. 9: 1346–1371. https://doi.org/10.1111/gcbb.12863.

Chipkar, S., K. Smith, E. M. Whelan, et al. 2022. "Water-Soluble Saponins Accumulate in Drought-Stressed Switchgrass and May Inhibit Yeast Growth During Bioethanol Production." *Biotechnology for Biofuels and Bioproducts* 15, no. 1: 116 2731–3654. https://doi.org/10.1186/s13068-022-02213-y.

Clifton-Brown, J., A. Harfouche, M. D. Casler, et al. 2019. "Breeding Progress and Preparedness for Mass-Scale Deployment of Perennial Lignocellulosic Biomass Crops Switchgrass, Miscanthus, Willow and Poplar." *GCB Bioenergy* 11, no. 1: 118–151. https://doi.org/10.1111/gcbb.12566.

Copernicus Global Land Service. 2019. "Land Cover 100 m: Epoch 2015—Globe." https://land.copernicus.eu/global/products/lc.

Cushman, J. C. 2001. "Osmoregulation in Plants: Implications for Agriculture." *American Zoologist* 41, no. 4: 758–769. https://doi.org/10.1093/icb/41.4.758.

Deines, J. M., R. Patel, S.-Z. Liang, W. Dado, and D. B. Lobell. 2021. "A Million Kernels of Truth: Insights Into Scalable Satellite Maize Yield Mapping and Yield Gap Analysis From an Extensive Ground Dataset in the US Corn Belt." *Remote Sensing of Environment* 253: 112174. https://doi.org/10.1016/j.rse.2020.112174.

Domingues, J., C. Pelletier, and T. Brunelle. 2022. "Cost of Ligno-Cellulosic Biomass Production for Bioenergy: A Review in 45 Countries." *Biomass and Bioenergy* 165: 106583. https://doi.org/10.1016/j.biombioe.2022.106583.

Fick, S. E., and R. J. Hijmans. 2017. "WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37, no. 12: 4302–4315. https://doi.org/10.1002/joc.5086.

Franz, T. E., S. Pokal, J. P. Gibson, et al. 2020. "The Role of Topography, Soil, and Remotely Sensed Vegetation Condition Towards Predicting Crop Yield." *Field Crops Research* 252: 107788. https://doi.org/10.1016/j.fcr.2020.107788.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29, no. 5: 3451. https://doi.org/10.1214/aos/1013203451.

Gabrielle, B., L. Bamière, N. Caldes, et al. 2014. "Paving the Way for Sustainable Bioenergy in Europe: Technological Options and Research Avenues for Large-Scale Biomass Feedstock Supply." *Renewable and Sustainable Energy Reviews* 33: 11–25. https://doi.org/10.1016/j.rser.2014.01.050.

Galelli, S., and A. Castelletti. 2013. "Assessing the Predictive Capability of Randomized Tree-Based Ensembles in Streamflow Modelling." *Hydrology and Earth System Sciences* 17, no. 7: 2669–2684. https://doi.org/10.5194/hess-17-2669-2013.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning* 63, no. 1: 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Gopal, P. S., and R. Bhargavi. 2019. "A Novel Approach for Efficient Crop Yield Prediction." *Computers and Electronics in Agriculture* 165: 104968. https://doi.org/10.1016/j.compag.2019.104968.

Gopalakrishnan, G., C. Negri, and S. Snyder. 2011. "A Novel Framework to Classify Marginal Land for Sustainable Biomass Feedstock Production." *Journal of Environmental Quality* 40, no. 5: 1593–1600. https://doi.org/10.2134/jeq2010.0539.

Haberzettl, J., P. Hilgert, and M. Von Cossel. 2021. "A Critical Review on Lignocellulosic Biomass Yield Modeling and the Bioenergy Potential From Marginal Land." *Agronomy* 11, no. 12: 2397. https://doi.org/10.3390/agronomy11122397.

Heggie, L., and K. Halliday. 2005. "The Highs and Lows of Plant Life: Temperature and Light Interactions in Development." *International Journal of Developmental Biology* 49, no. 5–6: 675–687. https://doi.org/10.1387/ijdb.041926lh.

Hudiburg, T. W., W. W. Wang, M. Khanna, et al. 2016. "Impacts of a 32-Billion-Gallon Bioenergy Landscape on Land and Fossil Fuel Use in the US." *Nature Energy* 1: 15005. https://doi.org/10.1038/nenergy.2015.5.

IPCC. 2023. *Climate Change 2022—Mitigation of Climate Change: Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009157926.

Ivushkin, K., H. Bartholomeus, A. K. Bregt, A. Pulatov, B. Kempen, and L. de Sousa. 2019. "Global Mapping of Soil Salinity Change." *Remote Sensing of Environment* 231: 111260. https://doi.org/10.1016/j.rse.2019.111260.

Jeong, J. H., J. P. Resop, N. D. Mueller, et al. 2016. "Random Forests for Global and Regional Crop Yield Predictions." *PLoS One* 11, no. 6: e0156571. https://doi.org/10.1371/journal.pone.0156571.

Jiang, P., and K. D. Thelen. 2004. "Effect of Soil and Topographic Properties on Crop Yield in a North-Central Corn–Soybean Cropping System." *Agronomy Journal* 96, no. 1: 252–258. https://doi.org/10.2134/agronj2004.0252.

Ke, G., Q. Meng, T. Finley, et al. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA.

Khatri, S. S., A. Patidar, A. Kavitkar, et al. 2022. "Dietary Application Through Image Processing for Calorie Management." *International*

Journal of Advanced Research in Science, Communication and Technology 2: 594–596. https://doi.org/10.48175/IJARSCT-3821.

Kravchenko, A. N., and D. G. Bullock. 2000. "Correlation of Corn and Soybean Grain Yield With Topography and Soil Properties." *Agronomy Journal* 92, no. 1: 75–83. https://doi.org/10.2134/agronj2000.92175x.

Kumar, S., M. Sohail, and S. Jadhav. 2024. "Light Gradient Boosting Machine for Optimizing Crop Maintenance and Yield Prediction in Agriculture." *ICTACT Journal on Soft Computing* 15, no. 2: 3551–3555. https://doi.org/10.21917/ijsc.2024.0495.

Li, W., P. Ciais, D. Makowski, and S. Peng. 2018. "A Global Yield Dataset for Major Lignocellulosic Bioenergy Crops Based on Field Measurements." *Scientific Data* 5, no. 1: 180169. https://doi.org/10.1038/sdata.2018.169.

Li, W., P. Ciais, E. Stehfest, et al. 2020. "Mapping the Yields of Lignocellulosic Bioenergy Crops From Observations at the Global Scale." *Earth System Science Data* 12, no. 2: 789–804. https://doi.org/10.5194/essd-12-789-2020.

Liu, Y., Y. Wang, and J. Zhang. 2012. "New Machine Learning Algorithm: Random Forest." In *Information Computing and Applications*, vol. 7473, 246–252. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32.

Malczewski, J. 2006. "GIS-Based Multicriteria Decision Analysis: A Survey of the Literature." *International Journal of Geographical Information Science* 20, no. 7: 703–726. https://doi.org/10.1080/13658810600661508.

Onsree, T., N. Tippayawong, S. Phithakkitnukoon, and J. Lauterbach. 2022. "Interpretable Machine-Learning Model With a Collaborative Game Approach to Predict Yields and Higher Heating Value of Torrefied Biomass." *Energy* 249: 123676. https://doi.org/10.1016/j.energy.2022.123676.

Pappas, C., S. Fatichi, S. Rimkus, P. Burlando, and M. O. Huber. 2015. "The Role of Local-Scale Heterogeneities in Terrestrial Ecosystem Modeling." *Journal of Geophysical Research: Biogeosciences* 120, no. 2: 341–360. https://doi.org/10.1002/2014JG002735.

Park, S. 2016. "Approximate Bayesian MLP Regularization for Regression in the Presence of Noise." *Neural Networks* 83: 75–85. https://doi.org/10.1016/j.neunet.2016.07.010.

Poggio, L., L. M. de Sousa, N. H. Batjes, et al. 2021. "SoilGrids 2.0: Producing Soil Information for the Globe With Quantified Spatial Uncertainty." *Soil* 7, no. 1: 217–240. https://doi.org/10.5194/soil-7-217-2021.

Popp, A., K. Calvin, S. Fujimori, et al. 2017. "Land-Use Futures in the Shared Socio-Economic Pathways." *Global Environmental Change* 42: 331–345. https://doi.org/10.1016/j.gloenvcha.2016.10.002.

Pradeep, G., T. D. V. Rayen, A. Pushpalatha, et al. 2023. "Effective Crop Yield Prediction Using Gradient Boosting to Improve Agricultural Outcomes." In *2023 International Conference on Networking and Communications (ICNWC)*, 1–6. IEEE. https://doi.org/10.1109/ICNWC57852.2023.10127269.

Qin, Z., Q. Zhuang, and X. Cai. 2015. "Bioenergy Crop Productivity and Potential Climate Change Mitigation From Marginal Lands in the United States: An Ecosystem Modeling Perspective." *GCB Bioenergy* 7, no. 6: 1211–1221. https://doi.org/10.1111/gcbb.12212.

Sage, R. F., M. de Melo Peixoto, P. Friesen, and B. Deen. 2015. "C4bioenergy Crops for Cool Climates, With Special Emphasis on Perennial C4grasses." *Journal of Experimental Botany* 66, no. 14: 4195–4212. https://doi.org/10.1093/jxb/erv123.

Salman, H., A. Kalakech, and A. Steiti. 2024. "Random Forest Algorithm Overview." *Babylonian Journal of Machine Learning* 2024: 69–79 3006–5429. https://doi.org/10.58496/BJML/2024/007.

Su, Y., H. Zhang, B. Gabrielle, et al. 2022. "Performances of Machine Learning Algorithms in Predicting the Productivity of Conservation Agriculture at a Global Scale." *Frontiers in Environmental Science* 10: 812648. https://doi.org/10.3389/fenvs.2022.812648.

Suganthavalli, K. 2024. "A Resilient Forecasting Model for Sustainable Agriculture and Optimal Yield Production." *Nanotechnology Perceptions* 20: S11.72. https://doi.org/10.62441/nano-ntp.v20iS11.72.

Tahir, S., and P. Marschner. 2017. "Clay Addition to Sandy Soil Reduces Nutrient Leaching—Effect of Clay Concentration and Ped Size." *Communications in Soil Science and Plant Analysis* 48, no. 15: 1813–1821. https://doi.org/10.1080/00103624.2017.1395454.

Taylor, G., I. S. Donnison, D. Murphy-Bokern, et al. 2019. "Sustainable Bioenergy for Climate Mitigation: Developing Drought-Tolerant Trees and Grasses." *Annals of Botany* 124, no. 4: 513–520. https://doi.org/10.1093/aob/mcz146.

Van Vuuren, D. P., K. Riahi, K. Calvin, et al. 2017. "The Shared Socio-Economic Pathways: Trajectories for Human Development and Global Environmental Change." *Global Environmental Change* 42: 148–152. https://doi.org/10.1016/j.gloenvcha.2016.10.009.

Wang, Y., J. Wang, J. Schuler, D. Hartley, T. Volk, and M. Eisenbies. 2020. "Optimization of Harvest and Logistics for Multiple Lignocellulosic Biomass Feedstocks in the Northeastern United States." *Energy* 197: 117260. https://doi.org/10.1016/j.energy.2020.117260.

Wieder, W. 2014. "Regridded Harmonized World Soil Database v1.2." https://doi.org/10.3334/ORNLDAAC/1247.

Winkler, B., A. Mangold, M. von Cossel, et al. 2020. "Implementing Miscanthus Into Farming Systems: A Review of Agronomic Practices, Capital and Labour Demand." *Renewable and Sustainable Energy Reviews* 132: 110053. https://doi.org/10.1016/j.rser.2020.110053.

Ye, R., B. Parajuli, and G. Sigua. 2019. "Subsurface Clay Soil Application Improved Aggregate Stability, Nitrogen Availability, and Organic Carbon Preservation in Degraded Ultisols With Cover Crop Mixtures." *Soil Science Society of America Journal* 83, no. 3: 597–604. https://doi.org/10.2136/sssaj2018.12.0496.

Ziliani, M. G., B. Aragon, T. Franz, et al. 2021. "Target Food Security: Assimilating Ultra-High Resolution Satellite Images Into a Crop-Yield Forecasting Model." https://doi.org/10.5194/egusphere-egu21-12357.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** gcbb70078-sup-0001-Supinfo.pdf.